International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 3, Issue. 4, Jul. - Aug. 2013 pp-2344-2346 ISSN: 2249-6645

A Novel Data Extraction and Alignment Method for Web Databases

Sravan Kumar Teegala[1], Ch.N.Santhosh Kumar[2], V. Sitha Ramulu[3]
[1]*M. Tech, Swarna Bharathi Institute of Science & Technology, Khammam, A. P., India*
[2]*Assoc. Professor, Dept. of CSE, Swarna Bharathi Institute of Science & Technology, Khammam, A.P., India.*
[3]*Assoc.Professor, Dept. of IT, Swarna Bharathi Institute of Science & Technology, Khammam, A.P., India.*

ABSTRACT**:** *Online databases, also called web databases, comprise the deep web tag and value. Compared with WebPages in the surface web, which can be accessed by using a unique URL, pages in the deep web are dynamically generated in response to a user query submitted through the query interface of a web database. Upon receiving a user's query, a web database returns the relevant data, either structured or semi structured, encoded in the HTML pages. Many web applications, such as data integration, met querying and comparison shopping, need the data from multiple web databases. For these applications to further utilize the data embedded in HTML pages, the automatic data extraction is necessary. Only when the data are extracted and organized in a structured manner, such as tables, can they be aggregated and compared. Hence, accurate data extraction is vital for these applications to perform process correctly. This paper focuses on the problem of automatically extracting data records that are encoded in the query result pages generated from web databases.*

KEYWORDS: *Data extraction, HTML, Query, Tag tree.*

I. INTRODUCTION

World Wide Web(WWW) is a powerful source of information. Search engines are very important tools for users to get the desired information on the web. Not only web users but many web applications also need to interact with the search engines. For decision making many business applications have to depend on the web in order to aggregate information from different web sites. By analyzing and summarizing web data we can find the latest market trends, price details, product specification etc. Manual data extraction is time consuming and leads to errors. In this context automatic web information extraction plays an important role. Example of web information extraction are: Extract competitor's price list from the web page regularly to stay ahead of competition, extract data from the web pages and transfer it to another application, extract people's data from the web pages and put it in a database.

Automatic data extraction plays an important role in processing results provided by the search engines after submitting the query by user. Wrapper is an automated tool which extracts Query Result Records (QRRs) from HTML pages returned by the search engines. Automated extraction is easier with the web sites having web service interfaces like Google and Amazon. But it's difficult for those that support B2C i.e. business to customer applications which does not have the web service interfaces. Normally Search engine result consists of query independent contents (static contents), query dependent contents (dynamic contents), while some contents are affected by many queries but independent of the content of specific query (semi-dynamic). As the web evolved web page creation process changed from manual to a more dynamic procedure using the complex templates. Many web pages are not created in advance, but are generated dynamically by querying the database server and sending the results to a predefined page structure. Automatic data extraction is very important for many applications, such as data integration , meta-querying and comparison shopping, that need to co-operate with multiple web databases to collect data from multiple sites and provide services.

II. RELATED WORK

In [1], the authors presented a novel data extraction method, ODE (Ontology-assisted Data Extraction), which automatically extracts the query result records from the HTML pages. To label attributes it is necessary that the labels appear in the query interfaces or query result pages within the domain. If the query result records are arranged into two or more different formats in the query result pages, then only one format will be identified as "query result section". Finally, the performance of ODE on certain types of query result pages is far from satisfactory only. In [2], the authors presented an overview of the issues involved in measuring data linkage and de duplication quality and complexity. It is shown that measures in the space of the record pair comparisons can produce deceptive accuracy results. It is recommended that the quality be measured using the precision-recall or F-measure graphs rather than using single numerical values, and that quality measures that include the number of true negative matches should not be used due to their large number in the space of record pair comparisons.

In [3], the authors made a thorough analysis of the literature on duplicate record detection. The similarity metrics that are commonly used to detect the similar field entries and an extensive set of duplicate detection algorithms that can detect approximately duplicate records in a database are covered. The lack of standardized, large scale benchmarking data sets can be a big obstacle as it is almost impossible to convincingly compare new techniques with the existing ones. In [4], the authors studied the problem of structured data extraction from arbitrary Web pages. In this paper, a novel and effective technique, called DEPTA, to perform the task of Web data extraction automatically is proposed. This method has the following drawbacks- When an object is very dissimilar to its neighboring objects, DEPTA misses it. This also causes a few identified data records to contain an extra information or to miss part of their original data items.

In [5], the authors presented a technique for automatically producing wrappers that can be used to extract search result records from dynamically generated result pages returned by search engines. The main problem with this method is its reliance on the tag structure in the query result pages, due to which it suffers from very poor results. In [6], the authors addressed the problem of unsupervised Web data extraction using a fully-automatic information extraction tool called ViPER. The tool is able to extract and separate data exhibiting recurring structures out of a single Web page with high accuracy by identifying the tandem repeats and using visual context information. However, this technique lacks overall performance in few datasets.

III. PROPOSED WORK

In this section, we describe the methods that are use in the process of Data extraction and Alignment. There is method for automatically extracting data records that are encoded in the query result pages generated by the web databases.

Figure 1 shows the framework for QRR extraction.

A. Tag Tree Construction: The first step is constructing a tag tree for the page rooted in the <HTML> tag. Each node represents a tag in the HTML page and its children are tags enclosed in it. Each internal node n of the tag tree has a tag string "tsn", which includes the tags of n and all tags of n's descendants, and a tag path tpn, which includes the tags from the root to n.

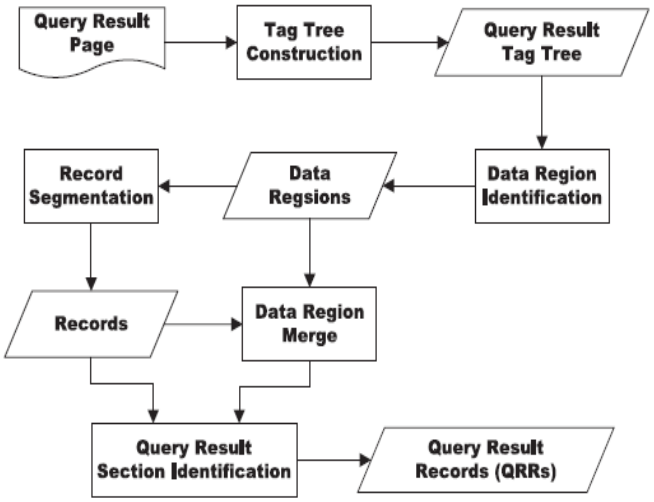

Figure 1: QRR Extraction framework

B. Data Region Identification: Similar data records of the same parent node are grouped as a "data region". It deals with the non- contiguous data records as well. Here we propose a new method that considers the auxiliary information leading to accurate data extraction. For this we need a temporary file which buffers the attributes as well as their values. These are further filtered which used to identify the exact data regions.

C. Record Segmentation: In a tag tree[4], the tandem repeats within a data region is initially found out. If only one repeat is found out then it corresponds to a record. In case of multiple repeats then any one has to be selected. Heuristics for record segmentation are as follows:

- Within a data region, if any auxiliary information is encountered, the tandem repeat that stops is the correct one since the auxiliary information cannot be inserted in the middle of the record.
- If the above two heuristics are failed to be used, then the tandem repeat that starts the data region is selected.

D. Data Region Merge: The data region identification step may identify various data regions in a query result page. Moreover, the actual data records may span several data regions identified. In the websites we examined, 12% had QRRs with different parents in the HTML tag tree. Thus, before we can identify all the QRRs in a query result page, we need to determine whether any of the data regions identified should be merged. Given any two data regions, we treat them as similar if the segmented records they contain are very similar. The similarity between any two records from two data regions is measured by using the similarity of their tag strings. The similarity between the two data regions is calculated as the average record similarity.

E. Query Result Section Identification: Even after performing the data region merge step, there may still be multiple data regions in the query result page. However, we assume that at most one of the data regions contains the actual QRRs. Three heuristics are used to identify this data region, called as the query result section:

- The query result section usually occupies a large space in the query result page
- The query result section is usually located at the center of the query result page

- Each QRR usually contains more raw data strings than the raw data strings in other sections.

F. ***QRR Alignment:*** The data values that belong to the same attribute generally show similarity in the data values and may include similar strings. Data value similarity [7] is calculated between every pair of values .The pairwise alignment determines whether the paired data values belong to the same attribute on the basis of calculated the data value similarity. Similarity of the record path is a constraint .The alignment of the data values between two QRRs must be unique. There should not be any cross alignment as well. After the pairwise alignment all data values of the same attribute are put in to the same table column globally by using of holistic alignment. This is similar to finding the connected components in an undirected graph . The vertices from the same record are not included in the same component. If any vertices breach this constraint, then the breach path is to be established. However the connected components are not allowed to intersect with each other. Finally a nested processing is needed to handle the attributes that having multiple values.

IV. CONCLUSION

In this project a novel data extraction method is proposed to automatically extract QRRs from a query result page. This method employs two steps for this task: The first step identifies and segments the QRRs to improve the existing techniques by allowing the QRRs in a data region to be non- contiguous. The second step is used to align the data values among the QRRs. An alignment method is used in which the alignment is performed in three consecutive steps: pair wise alignment, holistic alignment, and nested structure processing.

REFERENCES

[1]. W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2, article 12, p. 35, 2009.
[2]. P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, F. Guillet and H. Hamilton, eds., vol. 43, pp. 127-151, Springer, 2007.
[3]. A. K. Elmagarmid, P.G.Ipeirotis, and V.S.Verykios, "Duplicate Record Detection: A Survey", IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 1-16, Jan.2007.
[4]. 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.
[5]. H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. 14th World Wide Web Conf., pp. 66-75, 2005.
[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]. Weifeng Su, Jiying Wang, Frederick H.Lochovsky and Yi Liu, "Combining Tag and Value Similarity for Data Extraction and Alignment,"IEEE Transactions on Knowledge and Data Engineering,vol. 24, no. 7, pp. 1186-1199, 2012.