High Prioritized Data Region Ranking Technique and Multiple Clustered Tags Analyzing Method for Extracting and Aligning Requisite Data

S. Shiva Shankar, M. Gopi

Abstract— High prioritized ranking technique and multiple clustered tags analyzing method for extracting and aligning requisite data. Web data extraction plays a vital role in extracting requisite data in order to compare and align. Multiple Clustering of tagged data mining technique used to place data elements into related groups without advance knowledge of the group definitions. Establishment of an updated method called a novel data extraction and alignment method called HPDRRT (Which prioritize the data region once the web query content is extracted) and MCTA (Similar data and value regions are clustered) for analyzing the multiple data region alignment enhances the efficiency of the data extraction and alignment. Record alignment algorithm has been introduced in order to perform efficient contemporary alignment method based on comparison of data and value content wise alignment method and web similarity parent-child tag wise alignment. The problem of spanning multiple clustered tags in the multiple data regions for grouping the data from multiple data regions have been rectified and enhanced so as to fasten the process of extraction and aggregation.

Keywords— Data Extraction, Data and value regions, High prioritized ranking technique, multiple clustered tag, data content comparison, similarity parent-child wise alignment.

I. INTRODUCTION

Prioritizing the data on Web databases comprises of several web data ranking methods. Page can easily be ranked by using several page ranking algorithms, but giving rank to the data Comparison of data regions, grouping of multiple clustered tags, Data-value content wise tag alignment method and parent child tag wise similarity method will give an updated version for the use of different kind of web data multiple region comparison and web data – value related comparison applications.

We employ the following four-step method, called HPDRRT, MCTA, Comparison of data and value content wise alignment method, web similarity parent-child tag wise alignment method.

1. Web data Record extraction process identifies the data record region and extracts the needed record.

2. Web data Record alignment process aligns the data values of the web data record into a table so that data values for the same attribute are aligned into the same table column. Compared with existing data value content wise extraction methods, High prioritized data region ranking technique improves data extraction accuracy in three ways.

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

   a. An ascent web data region identification method is proposed to identify the non-contiguous web data record (WDR) that have the same parents according to their tag similarities.

   b. A fusion of extracted web data record method is proposed to combine different data regions that contain the WDRs (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 comparison of data and value wise alignment and second method is in association with web similarity parent-child tag wise alignment 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 contiguously clustered data region processing algorithm is proposed to handle any nested structure in the WDRs after the parent-child wise alignment.

Unlike existing nested-structure processing algorithms that rely on only tag information, the HPDRRT method uses the prioritized data region for analysing the extracted web data tag and value tag in order to improve the nested structure extraction accuracy.

2. OVERVIEW

In this paper the problem is focused on extracting the data from multiple data regions based on the parent-child tag and multiple clustered tag method.

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 WDRs into a table.

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 Data Region Ranking Method

Data region Ranking method is an important part of data region optimization which is in related to the clustering of prioritized data and its regions. Data region ranking method has been implemented in association with grouping of several data regions and to rank them to find out the most similarity gained data region. After finding out the data regions it has been given the most preference and set it as a parent for the other data regions. Then the second most similarity data regions have been given the second most similarity gained data region and it is set as child data region. Spanning around the multiple data regions the formation of graph- tag method will provide the most similarity data regions and those clustered most similarity data regions and its values will be grouped and aggregation in to the table.

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 data region ranking and clustering algorithms have not been introduced. Enhancement of these algorithms will give an extra value to the web pages as per the data region ranking algorithm and similar pages will be clustered in order to extract the needed information immediately which reduces the time consumption.

3. DESIGN CONSIDERATION

HPDRRT method and MCTA methods have been implement to rectify the problems that have been found out in CTVS method.

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.

Parent- child tags similarities have been found out in order to group the similar data. In MCTA (multiple clustered tag analysing method) the tags of one data region will be combined and analysed with the tags of other data regions. The tag similarity of the data regions will be examined by parent-child tag method. After examining the tag similarity of all the regions data value alignment and parent child wise alignment method will be performed in order to place the data in to the table containing similar data in to the rows and columns.

3.1 Web Query Page and Multiple Clustered Tag Extraction

After finding out the data region by using HPDRRT and MCTA method multiple tags are clustered to form a same data region and then the content will be compared with the data in the other multiple clustered 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 Multiple clustered tags and Data Region Identification

Similarity tags have been found out and it has been analysed by the parent-child tag method. Parent child tag method has a tag-similarity alphabet repetition method to find out the similar data and values that have been occurring in the data regions and those similarity data and values will be aggregated by the method called Content similarity aggregation method and it will be aligned in the table based on the parent-child tag wise alignment and Data- value content wise alignment method.

Data similarity Tags from one data regions will be compared with the other tags from other data regions. If the data tags are similar then it will be clustered. The process of comparison of data tag will not be stopped in the same data region it will be spanned and the process will be prolonged to the multiple data regions to find out all the clustered tags and all the similar data regions.
3.1.2 Web data record Identification and Alignment

Web data record identification method will be performed by automatic data identification method so as to find out the web data record. Comparing two regions and after identifying it the similar values will be arranged based on data-value content wise similarity method and parent-child tag wise similarity method.

3.1.3 Identification of Multiple Clustered Data Region and merge method

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 0.6, which is a threshold used to judge whether two records are similar in regions.

<table>
<thead>
<tr>
<th>WDR</th>
<th>Image</th>
<th>Nokia</th>
<th>3G</th>
<th>Color</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>302</td>
<td>3G</td>
<td>Blue</td>
<td>$55</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>502</td>
<td>3G</td>
<td>Blue</td>
<td>$56</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>325</td>
<td>3G</td>
<td>Red</td>
<td>$57</td>
</tr>
</tbody>
</table>

Table 1: Content data – value alignment

3.1.4 Prioritized data region formation method

Most frequently spanned data –value similarity content will be given the weight age so as to select the high ranked data regions. The most high ranked data region will preference to be a parent and other similarity data regions will be selected as child.

3.1.5 Extraction of multiple data similarity tags from multiple regions

Rectifying the problems of finding out the tags in same data region in existing CTVS method we have proposed a method called extraction of multiple data similarity tags which extracts the similarity tags from multiple data regions.

3.1.6 Web data result 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 WDR 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 WDRs.

3.2 Alignment of Query Result Record

WDR alignment is performed by data –value content similarity wise alignment and parent child tag similarity wise alignment rectifying the problems found in existing CTVS method which has the methods of pair wise and holistic alignment

3.2.1 Data-value Content similarity wise alignment

Extraction has been done from multiple data regions by HPDRRT. If the values are same based on the data tag then it will be aligned.

3.2.2 Parent-Child Similarity tag wise Alignment

Parent-child similarity tag wise alignment has been performed by examining all the similarity clustered tags from multiple data regions. Tags which are similar from multiple data regions will be aligned in to the table.

<table>
<thead>
<tr>
<th>Tag name</th>
<th>Parent</th>
<th>Child</th>
<th>Data</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;P&gt;</td>
<td>Engine CC specification</td>
<td>Engine cc</td>
<td>1</td>
<td>Red</td>
</tr>
<tr>
<td>&lt;Head&gt;</td>
<td>Bike</td>
<td>Pulsar</td>
<td>1</td>
<td>Blue</td>
</tr>
<tr>
<td>&lt;Head&gt;</td>
<td>Bike</td>
<td>SZR</td>
<td>1</td>
<td>Blue</td>
</tr>
<tr>
<td>&lt;Head&gt;</td>
<td>Bike</td>
<td>Honda</td>
<td>1</td>
<td>Blue</td>
</tr>
</tbody>
</table>

Table 2 - Parent- Child wise similarity

3.2.3 K – Means x data region fusion clustering method

K – Means x data region fusion clustering method has been associated in order to aggregate the multiple data regions to be clustered based on the similarity found in the multiple clustered tags.

Multiple data regions have been associated by finding the fusion value of the similar data and aggregating them to form similar parent-child web data region group.

Dual data regions are associated in the fusion algorithms in the terms X and Y. Number of clusters to be formed in the K has to be spanned with the regions X and Y. Multiple similar clustered tags are then found which will be resulted in grouping the multiple data regions from the different and similar parents.
Data tag extraction and multiple cluster examining method will give the accurate results of fasten data extraction and web data tag comparison method.

Finding out the similar web pages based on the parent-child comparison method has been implemented in order to know about the similarity of the multiple data regions.

Clustered data regions will be formed then tags from the multiple parent-child comparisons would reflect that the different data regions are aggregated based the given data value.

Hence by specifying the data value which is similar to that high prioritized value would have been given in order to extract the weighted data region.

### 3.2.4 Data Value Similarity Calculation

Data-values of one data region will be compared with other data region.

Case 1: If (Parent1==Parent2) then child tag will be checked, else (Child1!=child2) of two data regions will be checked for finding out similarity tags.

Case 2: If(Parent1!=Parent 2) but (Child1=Child2)

Case 3: If(Parent1==Parent2) but (Child1!=Child 2)

Case 4: If both the parent and child tags are not similar then it will not be taken in to the data region group.

If both the parent and child are similar in two data regions then it will be ranked high and take as a high prioritized data region.

### 3.2.5 Nested Structure Processing

Parent child similarity tag wise and data value alignment constrains a data value in a WDR to be aligned to at most one data value from another WDR.

If a WDR 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 WDRs that are generated by nested structures

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 data region and K means x fusion clustering methods the above problems will be solved accordingly.

2. In an HPDRRT and MCTA 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.

### 4. EVALUATION SETUP

We have performed different tests to assess the performance of HPDRRT and MCTA algorithm. We now present the experimental results for HPDRRT AND MCTA
over five data sets and compare Updated CTVS with ViNTs, and DeLa.

We have chosen ViNTs and DeLa to compare with HPDRRT because both have been shown to perform very accurate data extraction and implementations of both are available to us. An HPDRRT is implemented in JAVA and C++.

We have also used data region ranking method in order to bring out the expected data region from multiple data regions 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.

Both contiguous and non contiguous regions have been handled by HPDRRT methods.

The problem of handling different data from multiple data regions also have been handled and rectified by HPDRT and MCTA method.

5. CONCLUSION

Implementation of HPDRRT and MCTA method will provide efficient data region formation and aggregation of data from multiple regions.

Multiple clustering of tags will be performed in order to find out the similar tags from multiple data regions.

Data-Value content similarity wise alignment method and Parent-Child similarity tag wise alignment methods give proper formation of data regions and alignment of data. Handling of data from same and multiple data regions will be processed by HPDRRT and MCTA method.

6. FUTURE ENHANCEMENTS

Data region ranking method has been introduced based on giving the weight to the particular data regions for speeding up the extraction of data regions and its tags.

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 WDRs in a data region to be non-contiguous.

Parent-child similarity tag wise alignment method and Data-value content wise similarity methods are proposed. Experiments on five sets show that HPDRRT and MCTA is generally more accurate than current state-of-the-art methods.

Although HPDRRT has been shown to be an accurate data extraction method, it still suffers from some limitations. First, it requires at least two WDRs in the query result page.

Future enhancements of Data region ranking method and K means x fusion clustering methods and the time consumption in retrieving the data has been reduced accordingly future enhancement as an application, the data region ranking method or algorithm has been implemented also if a campaign of exchanging links to increase handling of similar data regions with multiple clustered tags and handling of multiple data regions with multiple clustered tags have been implemented so as to reduce the problems that have been found out in CTVS method.

Enhancement of Parent-child wise similarity and data-value content wise similarity alignment method give exact alignment of similar data in to the table and rectifies problems found in existing methods.

REFERENCES


