An Efficient Technique for Tag Extraction and Content Retrieval from Web Pages

S.Sathya M.Sc¹, Dr. B.Srinivasan M.C.A., M.Phil, M.B.A., Ph.D., ²

¹Mphil Scholar, Department of Computer Science, Gobi Arts and Science College, Gobi, Tamil Nadu, India.
² Associate professor, Department of Computer Science, Gobi Arts and Science College, Gobi, Tamil Nadu, India

ABSTRACT

In web database contains a large amount of information that is in the form of structured objects which are called as data records. In web databases to automatically extracting data records that are encoded in the query result page. These data records are important because these are present the essential information of their host pages, e.g., lists of products or services. A query result page contains not only the actual data, but also other information, such as navigational panels, advertisements, comments, information about hosting sites. The goal of web database data extraction is to remove any irrelevant information from the query result page, extract the query result records from those page, and align the extracted query result record (QRR) from the page, and align the extracted query result records into a table such that data values belonging to the same attribute are placed into the same table column. The proposed technique is able to handle both the attribute based and content based values are retrieved from the web pages in structured and unstructured data.

Keywords

Web data records, data region identification, record alignment, wrapper, information integration.

I. INTRODUCTION

Nowadays web content is mainly formatted in HTML pages, automatic data extraction is necessary. A web database contains the structured data or semi-structured data, encoded in HTML pages. To extract the query result page by using the combining tag and value similarity methods are used. A large amount of information on the web is presented in regularly structured objects. A list of such objects in a Web page often describes a list of similar items, e.g., a list of products or services. In this paper, we call them data records. Mining data records is useful because it allows us to integrate information from multiple sources to provide value-added services. In this paper a novel and more effective method to mine data records in a web page automatically.

The algorithm is called MDR (Mining Data Records in Web pages). Structured databases on the web, which return structured objects with attribute-value pairs (e.g., a book source like amazon.com returns books with author, title, etc. Thus, our focus essentially distinguishes un-structured databases, which provide data objects as unstructured data (e.g., text, images, audio, video). To extract the structured data from web pages it is very useful to carry complex queries over the data [1]. To extract a query result record (QRR) from a query result page 2 methods are used.

1. Record extraction identifies the Query result record (QRR) in query result page p and it includes data region identification and actual segmentation step.

2. Record alignment aligns the data values of the QRR in p into a table the data values for the same attribute are aligned into the same table column.
3. Compared with existing data extraction methods, proposed method improves data extraction accuracy in two ways.

a. 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, such as a comment, recommendation, or advertisement.

b. While the method in can find all data regions containing at least two QRRs in a query result page using data mining techniques, almost all other data extraction methods, assume that the QRRs are presented contiguously in only one data region in a page.

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 noncontiguous data region problem.

Hence, the pairwise alignment is reduced to finding a value-to-value alignment with the maximal data value similarity score under some constraints. After all pairs of records are aligned, a holistic alignment is performed, by viewing the pairwise alignment result as a graph and finding the connected components from the graph.

The databases are often also referred to as the hidden or invisible Web: The perception naturally arises: Since such information cannot be accessed directly through static URL links, they are only available as responses to dynamic queries submitted through the query interface of a database [12]. Because current crawlers cannot effectively query databases, such data are invisible to traditional search engines, and thus remain largely hidden from users.

Structured databases on the Web, which return structured objects with attribute-value pairs (e.g., a Book source like amazon.com returns books with author, title, etc.) [2]. Thus, our focus essentially distinguishes unstructured databases, which provide data objects as unstructured media (e.g., texts, images, audio, and video). We believe such distinction is both desired and necessary: First, such structured or “relational” data are traditionally of greater interest to the database community. Second, structured sources necessarily imply different paradigms and techniques from unstructured sources [16].

The goal of web database data extraction is to remove any irrelevant information from the query result page, extract the query result records (referred to as QRRs in this paper) 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.

A web database responds to a user query with the relevant data, either structured or semi-structured, embedded in HTML pages (called query result pages in this paper) [13]. To utilize this data, it is necessary to extract it from the query result pages. Automatic data extraction is very important for many applications, such as meta querying, data integration and data warehousing, that need to co-operate with multiple web databases [10]. Only when the data are extracted and stored in a database can they be easily compared and aggregated using traditional database querying techniques.

1. The layout of the QRR and the auxiliary information in the result page’s HTML tag trees (i.e., DOM) trees [3].

a. Data region identification method is used to identify the non-contiguous QRRs that have the same parents according to the tag similarities.

b. Merge method is used combine different data regions that contain the QRRs with or without the same parent into a single data region.

2. A novel method is used to identify the data values in the QRRs as pairwise and then holistically. The data values are arranged in same data type and similar data values in result of the query.

3. A new “nested-structure processing” algorithm is used to handle the nested structure in the QRRs after the holistic alignment [4].

The rest of the paper organized as follows: section 2 describes the mining data regions, section 3 describes the data extraction methodology, section 4 describes the query result record alignment and section 5 describes module description, section 6 includes experimental results finally concludes the paper and outlines possible future work.

II. MINING DATA REGIONS

This step mines every data region in a web page that contains similar data records. Instead of mining data records directly is hard, first mine generalized nodes in a page. A sequence of adjacent generalized nodes forms a data region. From each data region, to find the actual data records. The generalized node in the HTML tag tree with the following properties [8]:

![Diagram](http://www.ijcttjournal.org)
1. The nodes all have the same parent
2. The nodes are adjacent.

A data region is a collection of two or more generalized nodes.

2.1 Identifying Data Records

A generalized node may contain one or more data records.

<table>
<thead>
<tr>
<th>Row 1</th>
<th>Object 1</th>
<th>Object 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row 2</td>
<td>Object 3</td>
<td>Object 4</td>
</tr>
</tbody>
</table>

Figure 2: Each row with more than one data record

Fig 2 shows a data region that contains two table rows have been identified as generalized nodes. If a generalized node contains two or more data records, these data records must be similar in terms of their tag strings. Identifying data records from each generalized node is relatively easy because these are nodes together with their sub trees at the same level as the generalized node, or nodes at a lower level of the tag tree [5].

III. METHODOLOGY

Query Result page

Query Result Record

Record Segmentation

Aligned Query Result Table

Data Value Alignment

Label Alignment

Query Result Table

Figure 3: Data Extraction Process

Query result page, the Tag Tree Construction module first constructs a tag tree for the page rooted in the <HTML> tag. Each node represents a tag in the HTML page and its children tags are enclosed it. Fig 3 shows Data Region Identification identifies all possible data region, which usually dynamically generated data, top down starting from the root node. The record segmentation module identifies data regions into data records according to the tag patterns in the data region. Data region Merge module merges the data regions containing similar data records.

Query Result Section Identification module selects one of the merged data regions as the one that contains the Query Result Record. Data value Alignment aligns the data values from multiple records that belong to the same attribute so that these are arranged into a table. Label assignment assigns a suitable, meaningful label (i.e., an attribute name) to each column in an aligned table [15].

- HTML tags convey little semantic information since their main purpose is to facilitate the rendering of data [14].
- Some data may contain embedded tags, which may confuse the wrapper generators making them even less reliable[6].
- Nested-structure processing algorithms that rely on tag information.
- Employ wrapper induction can perform poorly when the format of a query result page changes [9].

IV. QUERY RESULT RECORD ALIGNMENT

Query Result Record (QRR) aligns the data values in 3 methods such as pairwise, holistic, and nested structure processing algorithm [7].

4.1 Pairwise QRR Alignment

This method aligns the data values in a pair of QRRs to provide the evidence for how the data values should be aligned among all the QRRs. These method may contain the attribute usually have the same data type and may contain similar strings. Each QRR includes the information: text string for the ith value and the tag path for the ith value. The data value alignments satisfy the 3 constraints:

a. Same record Path constraint: The record path of a data value comprises the tag from the root of the record to the node that contains in the tag tree of the query result page.

b. Unique constraint: Each data value can be aligned to at most one data value from the other QRR.

c. No cross alignment constraint: If the data value of the record path is matched if there is no cross alignment methods are not used.

4.2 Holistic Alignment
This method aligns globally among all QRRs to construct a table in which all data values of the same attribute are aligned in the same table column. This method is used to identify that the finding of the connected components in an undirected graph. Each connected component of the graph represents a table column inside which the connected data values from different data records are aligned vertically. These method problems are:

a. Vertices from the same record are not allowed to be included in the same connected component are considered to come from two different attributes of the record. If two vertices from the same record breach this constraint, a path must exist between the two, which called as breach path.

b. Connected components are not allowed to intersect with each other. A connected component in an undirected graph is a maximal sub graph such that for any two vertices in this sub graph.

V. MODULE DESCRIPTION

Here the modules which are used and how these modules are worked and described

1. Search (USER)

Web databases generate query result pages based on a user’s query. Automatically extracting the data from these query result pages is very important for many applications, such as data integration, which need to cooperate with multiple web databases. First, user can enter the URL of search engine then login if the user gives correct username, password then it transfer to search web page. In the search page user can give the keyword query to search information.

2. Construct tag and Identify data Regions

Given a query result page, the Tag Tree Construction module first constructs 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 inside it. Each internal node of the tag tree has a tag string, which includes the tags of n and all tags of n descendants, and a tag path, which includes the tags from the root to n. Next, the Data Region Identification module identifies all possible data regions, which usually contain dynamically generated data, top down starting from the root node. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions.

3. Novel Data Extraction & Alignment

A novel method is proposed to align the data values in the identified QRRs, first pairwise 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. Both tag structure similarity and data value similarity are used in the pair wise alignment process.

First to combine tag structure and data value similarity to perform the alignment. We observe that the data values within the same attribute usually have the same data type, and similar data values in many cases, because they are the result for the same query. Hence, the pairwise alignment is reduced to finding a value-to-value alignment with the maximal data value similarity score under some constraints. After all pairs of records are aligned, a holistic alignment is performed, by viewing the pairwise alignment result as a graph and finding the connected components from the graph.

4. Nested Structure Processing Algorithm

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, CTVS uses both tag and data value similarity information to improve nested structure processing accuracy [8].

5. Extracted QRRs:

QRR alignment is performed by a novel three-step data alignment method that combines tag and value similarity.

1. Pair wise QRR alignment aligns the data values in a pair of QRRs to provide the evidence for how the data values should be aligned among all QRRs.
2. Holistic alignment aligns the data values in all the QRRs.
3. Nested structure processing identifies the nested structures that exist in the QRRs.
VI. EXPERIMENTAL RESULTS

E-COMM contains 100 E-commerce deep websites to take 3 popular domains such as mobile, laptop and book. Each domain contains 15-20 websites. For each website, to create five result pages for use as training pages by submitting the query. Compared with the another dataset or another website to find that the QRRs in E-COMM have more complex structures usually contains more nested levels and more optional attribute in the pages HTML tag tree, which reduces the data extraction[11]. The user wants to search the information about “mobile” product means to give the mobile name or mobile model in web browser the information should be displayed in “content based” that means to click the URL link and view the information. Our proposed method is used to remove any irrelevant information such as panels, advertisements, etc. Only the extracted query can be stored in QRRs. The another method is “attribute based” method the user have to specify the attributes such as mobile :< name>, <price>, <color> in HTML tag that the attributes are matches in query result record segmentation that the values are stored in QRRs.

Evaluation Metrics

Two sets of evaluation metrics are used to compared. The first is at the record level and includes the precision and recall metrics defined as

\[
Pr = \frac{Cc}{Ce} \\
Rr = \frac{Cc}{Cr}
\]

Where Cc is the count of correctly extracted and aligned QRRs, Ce is the count of extracted QRRs, and Cr is the actual count of QRRs in the query result pages.

The number of QRRs in different query result pages varies from a few to hundreds. Consequently, pages with many QRRs will dominate the record level metrics. To use a page-level metric, namely page-level precision defined as,

\[
Pp = \frac{Cp}{Na}
\]

Where cp is the count of correctly extracted pages, which means that all the QRRs in the pages are correctly extracted and aligned, Na is the count of all the pages from which QRRs are extracted. To assume that each input page contains at least two QRRs and data extraction is performed on all input pages.

<table>
<thead>
<tr>
<th>Classes</th>
<th>results</th>
<th>results</th>
<th>Resul t</th>
<th>Resu lts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site Description</td>
<td>#s</td>
<td>#w</td>
<td>#s</td>
<td>Extr</td>
</tr>
<tr>
<td>Amazon.com Cars by brand</td>
<td>21</td>
<td>1</td>
<td>21</td>
<td>Yes</td>
</tr>
<tr>
<td>Amazon.com Music best sellers</td>
<td>20</td>
<td>-</td>
<td>20</td>
<td>No</td>
</tr>
<tr>
<td>Buy.com Product sub categories</td>
<td>20</td>
<td>1</td>
<td>20</td>
<td>Yes</td>
</tr>
<tr>
<td>Buy.com Product information</td>
<td>10</td>
<td>1</td>
<td>10</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table 1: Experimental results
Table 1 contains the different websites the user wants to search information the following elements are retrieved: 1. Class a short description of each class and the number of samples considered for that class 2. Results obtained from the matching i.e., number of wrappers (#w) [17] created by the system, number of samples matching each wrappers (#s), outcome of the data extraction process (extr) i.e., whether it was possible to actually extract a dataset from the pages.

VII. CONCLUSION

The efficient data extraction method is used a novel data extraction method and nested processing algorithm. It takes only a few minutes to extract several Web pages. Comparing our algorithm to others, our approach is quick and effective. Though our work has involved primarily Internet Information Resources, The novel extraction method used QRRs alignment as pairwise, holistic, and nested structure processing algorithm. So, the data extraction method is more accurate. In our research the content based retrieval and attribute based retrieval methods are used. The limitations are: the query result page has more than one data region that contains the QRRs and the different data region are not similar to each other, so this method will select only one data region the other data regions are not selected.

REFERENCES


