Data Extraction and Alignment in Web Databases

Mrs K.R.Karthika
M.Phil Scholar
Department of Computer Science
Dr N.G.P arts and science college
Coimbatore,India

Mr K.Kumaravel
Ph.D Scholar
Department of Computer Science
Govt., arts college (Autonomous)
Coimbatore,India

Dr A.Marimuthu
Associate Professor
Department of Computer Science
Govt., arts college (Autonomous)
Coimbatore,India

Abstract

Web databases comprises of large amount of data. There is a high demand for collecting data of interest from multiple web databases by the users. Query based interfaces are used to access the data from such databases. Data extraction from large portion of web databases is very tedious portion of many search engine processes. Without any consideration search engine returns result pages from the structural and unstructured databases based on a unique URL.A typical result page returned from such a web database has multiple query result records (QRR) and auxiliary information such as comments, advertisements and multiple sites pertaining to the advertisements. This paper defines the semantical representation of a concept as an entity, which means a data unit will be extracted from a structured or unstructured web page and are then organized into tables so that they can they be easily compared and aggregated to utilize the data. This paper aims to automatically extract wrappers, tags and values from a raw HTML file by employing Structural Semantic Combination of Tag and Value Similarity (SCTVS) method, which aims at automatically assigning meaningful labels to the data units in QRRs. It has a main objective of finding relative labels for QRR result set alignment. Optimal temporary storage is used to store the auxiliary information for enabling data extraction easier. In addition, a new method of data alignment called Priority alignment along with Data Summarization is used .Experimental results show that the current method yields high precision and accuracy compared with the existing methods.

Index Terms: Data Extraction, Data Summarization, Semantic Similarity, Priority Alignment

1. Introduction

Web databases generate web data relevant to user’s search query. Web applications such as data integration, Meta querying, Comparison shopping need data from multiple web databases. Such data encoded in the HTML files in database are either structured or semi structured. Data Extraction is the only means to obtain data from the database. Web information extraction involves two things namely extracting information from natural language text and extracting structured data from Web pages. This paper aims on extracting structured data from web pages. The main goal of web database data extraction is to extract QRR from the Query result pages and aligning them in tabular format based on the criteria that the tag and value are similar with labels assigned to the data units. Another goal is to remove any irrelevant information from the query result page. This paper introduces a method called SCTVS to extract data from the web database.

SCTVS improves data extraction accuracy in many ways like the earlier method used in [12]. It thoroughly analyses the relationship between tag and data units semantically. New techniques are proposed to handle the case when the QRRs are not contiguous in the dataset, it extracts the available data, aligns them in the tabular format, calculates priority and makes it outline for the fast data summarization. Auxiliary information in the result page is used for data labelling.

1. QRR extraction is done by means of the following steps:

a. Data region identification is used to identify the noncontiguous QRRs. Data region merge is used to merge the data regions from different pages.

b. Tag identification is used to identify the similarities between the web pages.

c. Tag classification is used to classify and prioritize the tags from web page sources.

2. A fresh method is proposed to align the data values in the identified QRRs, first pair wise then holistically and then finally prioritizing, 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. Nested-structure processing algorithm is proposed to handle any nested structure in the QRRs after the priority alignment. Unlike existing nested-structure processing algorithms that rely on only tag information, SCTVS uses similar tag and data value based on semantics to improve nested-structure processing accuracy.
4. The implementation of clustering-based data summarization technique helps to align data units into different groups so that the data units inside the same group have the same semantic. This approach considers other important features shared among data units, such as their priority based information. Finally the system integrates the interface schema over multiple web data bases with the common cluster to enhance data summarization.

2. Related Works

Information extraction from web pages is an active research area. Researchers have been developing various solutions from all kinds of perspectives. Web information extraction can be at the record level or data unit level.

Record level extraction generally involves identifying data regions that contain all the records, and then partitioning the data regions into individual records. Since the records within a data region are usually highly homogeneous and the data regions are often constructed simply by a list of records, the record level extraction is easier than the data unit level extraction.

Some recently proposed fully automatic extraction methods on record level extraction achieved satisfying performances. [5], [12].

On the other hand, the data unit level extraction is more complicated, and the performance of proposed fully automatic methods is not satisfactory.

K. Simon and G. Lausen [5] 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 tandem repeats and using visual context information. However, this technique lacks performance in few datasets.

Y. Zhai and B. Liu [13] 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 automatic Web data extraction is proposed. This method has the following drawbacks: When an object is very dissimilar to its neighboring object, DEPTA misses it. Another issue is few identified data records contain extra information and some of them miss part of their original data items.

H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu [3] presented a technique for producing wrappers automatically that can be used to extract search result records from dynamically generated result pages returned by search engines(web databases). The main problem with this method is it relies on the tag structure in the query result pages and suffers from poor results.

J.Wang and F.H. Lochovsky [10] presented a novel data extraction method, ODE (Ontology-assisted Data Extraction) that automatically extracts the query result records from the HTML pages based on the queries. It is necessary that the labels should appear in the query interfaces or query result pages within a domain for attribute labeling. If the query result records are arranged in two or more different formats in the query result pages, only one format will be identified as the query result section. Finally, the performance of ODE on certain types of query result pages is far from satisfactory.

P. Christen and K. Goiser [8] presented an overview of the issues involved in measuring data linkage and deduplication quality and complexity. It is shown that measures in the space of 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 single numerical values.The 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.

A.K.Elmagarmid, P.G.Ipeirotis, and V.S.Verykios [1] made a thorough analysis of the literature on duplicate record detection. The similarity metrics are commonly used to detect similar field entries. An extensive set of duplicate detection algorithms that can detect duplicate records in a database are covered. This technique yet suffers from lack of standardized, large scale benchmarking data sets.

S. Chaudhuri, V. Ganti, and R. Motwani [9] proposed two criteria namely compact set and sparse neighbourhood,that enable characterization of fuzzy duplicates more accurately compared with existing techniques. This method does not deal with blocking strategies.

M. Bilenko and R.J. Mooney [6] presented a framework for improving duplicate detection using trainable measures of textual similarity. Learnable text distance functions for each database field are employed to show that such measures are capable of adapting to the specific notion of similarity that is appropriate for the field’s domain. However, this method suffers from over fitting issues in few cases.

Weifeng Su, Jiying Wang, Frederick H. Lochovsky and Yi Liu [12] presented a paper “Combining Tag and Value Similarity for Data Extraction and Alignment ” for accurate data extraction from the databases. This method identifies the non contiguous data regions by a novel extraction and alignment method. However, this method suffers from some limitations by treating the optional attributes as auxiliary information and doesn’t handle when multiple data values are used.

3. QRR Extraction

Query Result Records are obtained from the result pages by identifying the data regions in the web pages followed by merging the data regions that exist in several pages. For a query result page, the Tag Tree Construction module constructs a tag tree for the page rooted in the <HTML> tag. Each node in the tag tree represents a tag in the HTML page, its children are tags enclosed inside it.

Every internal node n of the tag tree has a tag string t_{sn} and a tag path t_{pn}, which includes the tags from the root node to n. Next, the Data Region Identification module identifies all possible data regions in web pages that contain dynamically generated data, top down starting from the root node to n. The data region merge module helps to merge different data regions into one. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions. Query result set identification module helps to identify the merged region and finally QRR are extracted from this region.

Earlier works on web page segmentation were on free texts and motivated by raising performances of the information retrieval. In information retrieval, documents are extracted with the values of semantics of the web page to the queries:

$$sim(q,d) = \frac{\sum w_{q,i} * w_{d,i}}{(\sum w_{q,i}^2)^{\frac{1}{2}} (\sum w_{d,i}^2)^{\frac{1}{2}}}$$

In this paper, a data unit is a piece of text that semantically represents one concept of an entity. Data unit corresponds to the value of a record under an attribute. They differ from text node in which text node refers to a sequence of text surrounded by a pair of HTML tags.

There is a high demand for collecting data of interest from multiple web databases by the users.. For example, in a book comparison shopping system, a query interface collects multiple result records from different book sites to determine whether any two QRRs refer to the same book .In such cases, data extraction is the novel method to obtain vital data.

3.1 Data Region Identification

This step identifies data region with similar data records in the web pages.

Definition:

A generalized node (or a node combination) of length r consists of r (r≥1) nodes in the HTML tag tree with the following two properties:

1. The nodes all have the same parent and
2. The nodes are adjacent.

A data region is a collection of two or more generalized nodes with the following properties

1. The generalized nodes all have the same parent.
2. The generalized nodes are all adjacent.
3. Adjacent generalized nodes are similar.

3.1.1 Data Region Merge

Two data regions can be treated similar if the records they contain are similar. Similarity between tag strings are measured by similarity between any two records from two data regions. The data region identification step identifies several data regions in a query result page. In general, data records span several data regions, QRRs with different parents in the HTML tag tree. Before identifying all the QRRs in a query result page, it is necessary to determine whether any data regions should be merged.

3.2 Tag Identification

Tag identification eliminates a number of the most common HTML tags typically found in web pages. It places tags in classes depending on their semantic connotation. This helps to identify and examine similarities and regions between web pages. And it also used to eliminate common structured tags and value

3.3 Tag Classification

This step is used to classify tags from a page source, which will help to identify the prioritized regions. For example a string or tag can be prioritized by its region i.e. in which class the tag flows in.

<div class="rlp-h"> <b class="sample"></b>

<h3>**Camera Type**</h3></div>

3.3.1 Semantic similarity

It proposes the method used for measuring the similarity between two records. This module helps to make the page count and text snippets value to make the final values. The matching is totally based on the correlation between the words.

4. QRR Alignment

The purpose of data alignment is to put the data units of the same concept into one group so that they can be annotated holistically. Two data units belong to the same concept are determined by how similar they are based on the features described in following sections. In this system, the similarity between two data units' d1 and d2 is a weighted sum of the similarities of the five features between them.

Sim (d1, d2) = w1 * Sim(d1; d2)

Where d1 is the extracted dataset 1 and d2 is the second dataset.

QRR alignment is performed by a novel 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. Priority alignment aligns the data values based on the priority.
4. Nested structure processing identifies the nested structures that exist in the QRRs.
5.

4.1 Pair wise QRR Alignment

This method aligns data values between every pair of QRR's. During the pair wise alignment, it is required that the data value alignments must satisfy the following three constraints:

1. Same record path constraint.
2. Unique constraint.
3. No cross alignment constraint in the data value alignments during the pair wise alignment of query result records.

4.2 Holistic Alignment

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. If we consider each data value in the QRRs as a vertex and each pair wise alignment between two data values as an edge, the pair wise alignment set can be viewed as an undirected graph. This alignment method is equivalent to that of finding connected components in an undirected graph. Each connected component in the graph represents a table column and the connected data values from different records are aligned vertically. Even though there are many algorithms for finding connected components in the Graph Theory literature, we need to consider two application constraints that are specific to the holistic alignment problem as follows:

1. Vertices from the same record are not allowed to be included in the same connected component as they are considered to come from two different attributes of the record. If two vertices from the same record break this constraint, a breach path exists between the two.

2. Connected components are not allowed to intersect each other. If C1 and C2 are two connected components, then vertices in C1 should be either all on the left side of C2 or all on the right side of C2, and vice versa (i.e., no edge in C1 cuts across C2, and no edge in C2 cuts across C1).

4.3 Priority alignment

Priority alignment in the proposed method contains the following constraints

1. Extracted data will be preprocessed- the pre processing stage eliminates duplicate values and labels.
2. Checks with the training dataset.
3. Based on the attribute priority the table will be updated.

4.4 Nested Structure Processing

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

Existing techniques and methods rely only on HTML tags to identify the nested structures. They may incorrectly identify a plain structure as a nested one. To overcome this problem, SCTVS uses both the HTML tags and the data values to identify the nested structures. Given an aligned table, a nested column comprises at least two ordered sets representing the data values that are generated by repetitive parts in the template.

This method employs nested structure identification algorithm [12] to identify the nested structures. Given n records with a maximum of m data values and a maximum tag string length of l, the time complexity of the nested structure processing algorithm is O (nl2m2).

Compared with the earlier nested structure procedures used in existing papers, the procedure used in this SCTVS method yields many advantages. This method first aligns the data records and then processes the nested structures.

5. Data Summarization

Data summarization technique uses clustering algorithms. This first identifies all data units in the QRRs and then organize them into different groups with each group corresponding to a different concept.

Grouping data units of the same semantic can help identify the common patterns and features among these data units. These features are the basis of labelling technique for data summarization.

A tag node corresponds to an HTML tag surrounded by angular brackets starting with "<" and ending with ">" in HTML source whereas a text node is the text outside the tags "<" and ">."Consider the following example

<b> stereo type </b>

Text nodes are the visible elements on the webpage and data units are located in the tags. Tags are not always identical to data units. Since the annotation is at the data unit level, this needs to identify data units from tag structure.

Table 1: Sample QRR's

Manufacture	**Model**	**Class**	**Year**	**Price**
HONDA	Accord lx	4door	2004	$11800
HONDA	Accord	4door	2005	$12134
HONDA	Accord lx	4door	2000	$10000
HONDA	Accord	4door	2001	$11000
HONDA	Accord lx	4door	2003	$10800

The proposed enhanced data alignment algorithm is based on the assumption that attributes appear in the same order across all QRRs on the same result page, although the QRRs may contain different sets of attributes (due to missing values). This is true in general because the QRRs from the same WDB are normally generated by the same template program. Thus, it can conceptually consider the QRRs on a result page in a table format where each row represents one QRR and each cell holds a data unit (or empty if the data unit is not available).

Each table column, in the proposed work, is referred to as an alignment group, that contains at most one data unit from each QRR. An alignment group contains all the data units of one concept and no data unit from other concepts.

The goal of alignment is to move the data units in the table so that every alignment group is well aligned and the order of the data units within every QRR is preserved. The following are the steps involved in the enhanced alignment for data summarization.

Step 1: Merge text nodes. This step detects and removes decorative tags from each Search Result Record to allow the text nodes corresponding to the same attribute (separated by decorative tags) to be merged into a single text node.

Step 2: Align text nodes. This step aligns text nodes into groups so that eventually each group contains the text nodes with the same concept (for atomic nodes) or the same set of concepts (for composite nodes).

Step 3: Split (composite) text nodes. This step aims to split the data values in composite text nodes into individual data units. Splitting is carried out based on the text nodes that exist in the same group holistically. A group whose "values" need to be split is called a composite group.

Step 4: Align data units. This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept.

A tag path of a text node is a sequence of tags traversing from the root of the QRR to the corresponding node in the tag tree.

6. Conclusion

This paper presented a novel method for extraction called SCTVS, for extracting structured data from a collection of query result pages. SCTVS first discovers the semantic data regions, identifies the semantic labels from the QRR and merges the data region. Finally it aligns the data values in QRR by the following methods: pair wise, holistic, Priority basis and nested structure processing. QRR's with similar tag and value will be stored in the tables. Optimal temporary storage technique is used to store the auxiliary information for future reference. Data summarization is used to enhance the alignment and summarize the description of the attributes with same semantic structures.

SCTVS extracts desired data from various QRR pages. The experiments on several collections of QRR, drawn from many well-known data rich sites, indicate that SCTVS is extremely good in extracting and aligning the data from the web page sources. Another desirable feature of the proposed system is that it does not completely fail to extract any data. The impact of the failed assumptions is limited to a few attributes. SCTVS prone to provide higher accuracy compared with the existing methods.

7. Future Enhancement

This system provides higher precision in extracting structured and semi-structured data from result pages in the web databases. In future, techniques for Crawling, Indexing and providing Querying support for unstructured pages can be done along with the above process to yield proven accuracy.

8. References

[1] 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.

[2] 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.

[3] 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.

[4] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th World Wide Web Conf., pp. 187-196, 2003.

[5] 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.

[6] M. Bilenko and R.J. Mooney, "Adaptive Duplicate Detection Using Learnable String Similarity Measures," Proc. ACM SIGKDD, pp. 39-48, 2003.

[7] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," White Paper, BrightPlanet Corporation, http://www.brightplanet.com/resources/details/deepweb.html, 2001.

[8] 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.

[9] S. Chaudhuri, V. Ganti, and R. Motwani, "Robust Identification of Fuzzy Duplicates," Proc. 21st IEEE Int'l Conf. Data Eng., pp. 865-876, 2005.

[10] 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.

[11] Weifeng Su, Jiying Wang, and Fredrick H.Lochovsky, "Record Matching Over Query Results from Multiple Web Databases", IEEE Trans. Knowledge and Data Eng., vol. 22, no. 4, April 2010.

[12] Weifeng Su, Jiying Wang, Frederick H. Lochovsky, Member, IEEE Computer Society, and Yi Liu, "Combining Tag and Value Similarity for Data Extraction and Alignment ," IEEE transactions on knowledge and data engineering, vol. 24, no. 7, July 2012

[13] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment," IEEE Trans.