ISSN:2320-0790

An International Journal of Advanced Computer Technology

DATA EXTRACTION AND ALIGNMENT USING TAGS AND VALUE SIMILARITY

Mrs. S. Padmavathi (M.Sc., M.Phil, B.Ed.)[1], K.Tamilselvi[2],
Master of Philosophy in Computer Science[1],
Marudupandiyar College

ABSTRACT: Web databases generate query result pages based on a user's query. Automatically extracting these data from query result pages is very important for many applications, such as data integrations, which needs to cooperate with multiple web databases. This system presents a novel data extraction and alignment method called DATVS that combines both tag and value similarity. DATVS automatically extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in the query result pages and then aligning the data segmentation QRRs into a table, in which the data values from the same each attributes the put into the same column. Specifically, This propose new techniques to handle the case when the QRRs is not contiguous, which may be due to presence of an auxiliary information, such a comment, recommendation or advertisement and for handling they any nested structure that may exist in the QRRs. The new system is a design and the new record alignment algorithm that aligns the attributes in a record and first pair wise and they holistically, by combines the tag and data value similar information. Experimental results show that DATVS achieves high precision and outperforms existing state-of-the-art data extraction methods.

Keywords: Data Extraction, QRRs, HTML DOM, Value Similarity

I. INTRODUCTION

Web databases generate query result pages based on a user's query. Automatically extracting the data's from the query result pages is very important for many applications, such as the data integration, which need to cooperate with multiple web databases. We present a novel data extraction and alignments method called DATVS that combines both tag and value similarity. DATVS automatically extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in this query result pages and then aligning the segmented QRRs into a table, in any the data values from the same attributes are put into the same column. Specifically, these propose new techniques have to handle the case. When the QRRs are not contiguous, this may be due to the presence of auxiliary information, such as comment recommendation/advertisement, and for handling for any nested structure that may exists in the QRRs.

Object similarity is to support as focus on the role of extreme values in object matching and its termed hyper matching. Importance weights are first introduced to the matching and variations formulated by objects that do not share all the same attributes. Objects can be both possess the same or different extreme valued attributes [1]. To segment object from the web images are using logo detection. This method consists of a three steps. In the first step the logos are located from the original image by SIFT matching. Under the logo location and the object shapes model, the second steps extract the object boundary from the images. In the third steps, we use the objects boundary to model the object appearance, which is then used in the MRF based the segmentation method to finally achieves the object segmentation. To cope with the shape variations, affine transform of the shape model is considered [2]. Automatically extracting the data from these query result pages is very important for many applications, such as the data integrations, which need to cooperate with multiple web databases. The data values from the same attribute are put into the same column. Specifically, we proposed the new techniques to handle the case. When the QRRs are not contiguous, this may be due to the presence of the auxiliary information, such as comments, recommendations or advertisements, and for handling for an any nested structures that may exist in the QRRs [3]. The Internet, it is desirable to interpret and the extract useful information from the Web. One of the major challenges in Web interface interpretations is to discover the semantic structures and underlying a web interface. Many heuristics approach has been

developed to and discover the groups semantically related interface objects. The spatial graph grammar (SGG) is selected to perform the semantic grouping and interpretation of segment screens object. Instead of an analyze the HTML source code and be apply to an efficient image processing technologies to recognize atomics interface object from the screenshots of an interface and produces a spatial graph [4]. To improve and achieves the efficiency and accuracy of automatic wrappers is able to check the similarity of data records and to detects the correct data regions with the higher precision data and using the semantic properties of these data records. The advantages of these methods is it can extract three types of data records has named by single section data records, multiple section data records and loosely structured data records; it also provides options for an aligning iterative and disjunctive data item [5]. The adaptation of a general search computing frameworks for exploratory search over the web data is suggests by specify the location and web based data services. The result is conceptual model of geographical entities, the spatial function of operating on them, and a special purposes exploratory interface that lets users search combinations of georeferenced objects directly on a map [6]. A new Web data extraction approach, called FiVa Tech to the problem of page-level data extraction. We formulate the page generation models using a encoding scheme based on tree templates and scheme, which organize the data by their parent node in the DOM trees. FiVa Tech contains two phases: phase I is merging input DOM trees to construct the fixed/variant pattern tree and phase II is schema and template detection based on the pattern treen [7]. The framework for adapting information, extraction wrappers with a new attribute discovery via Bayesian learning. A generative model for the generation of text fragments related to the attributes and layout format in Web pages is designed to harness the uncertainty. Bayesian learning and EM techniques are employed in our framework for tackling the wrapper adaptation and new attribute discovery tasks [8]. In general, the desired information is embedded in the deep Web pages in the form of data records returned by Web databases. The visual information of Web pages can help us implement Web data extraction. Based on observation of a large numbers of deep Web pages it's identified by a set of interesting the common visual features that are useful for deep Web data extraction. The main trait of this vision-based approach is that it primarily utilizes the visual features of deep Web pages [9]. A new approach is to extract structured data from web pages. Although the problems have been studied by the several researchers, existing techniques either are inaccurate or make several assumptions. Novel partial tree alignment technique to align corresponding data fields of multiple data records. Empirical results using a large number of Web pages demonstrated the effectiveness of the proposed technique [10].

II. QRR Methodology

This Article focuses on the problem of automatically extracting data records that are encoded in the query result pages generated by web databases. The goal of web database data extraction is to remove any irrelevant information.

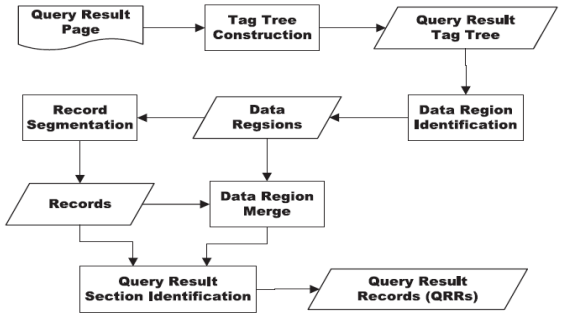

Fig.1 QRR record alignment

QRR Extraction

A query result page, the Tag Tree Construction module first constructs a tag tree for the page rooted in the <HTML> tag. 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 to n, which includes the tags from the root to n. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions

Data Region Identification

In this paper proposes a new method to perform the task automatically which is more effective than machine learning and semi-automated system. The proposed methods consist of two steps,

(1) Identifying the individual data records in a page.

(2) Aligning and extracting data items from identified data records.

RECORD SEGMENTATION

To illustrate the record segmentation algorithm, assumes that in Region 1 of the artificial tag tree in Fig.2, nodes 3, 6, 8, and 10 are similar and nodes 4, 7, and 9 are similar, while the region 2 and the node 12 and 13 is similar. Record segmentation first finds tandem repeats within a data region. For Ex, region 1 in Fig.3 can be represented as ABBABA if we use the character A to represents the element of the similar node set {3, 6, 8, 10} and B to represent an element of the similar node set {4, 7, 9}. In these case there are

two type tandem repeats AB and BA. Similarly the Region 2 in Fig. 2 can be represented as CC, which contains only one tandem repeat, C.

II. ALGORITHMS

VIPS:
VIPS (vision based page segmentation algorithm) is an automatic top down the tag tree independent approach to detect web content structure. VIPS algorithm is to transform a deep web page into a visual block tree. The leaf blocks are the blocks that cannot be segmented further and they represents the minimum number of semantic units, such as continuous texts or images. These block tree is constructed by using DOM (document object model) tree.

DOM TREE
In VIPS algorithm we will use DOM tress to find out the visual block tree. The Document Object Model (DOM) is a cross platform and language independent conventions for representing and interacting with the objects in HTML, XHTML & XML documents.

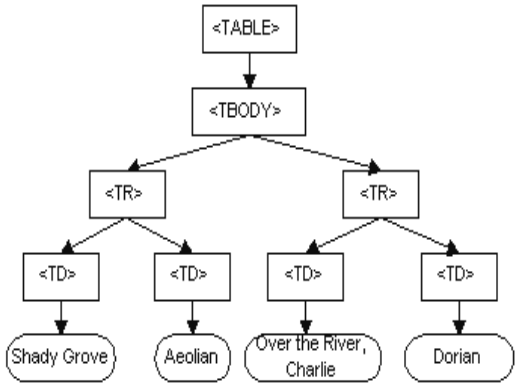

Fig 2. DOM Tree Structure

The Document Object Model can be used with any programming languages. In order to provide the precise language independent specification of the Document Object Model interfaces, various other IDLs could have been used OMG IDL does not imply a requirement to use a specific object binding runtime.
The DOM is a programming API for documents. It is based on an object structure that closely resembles the structure of the documents it models, However, the DOM does not specify that documents must be implemented as tree or grove nor does it specifies how they relationship among the objects be implemented. The DOM is a logical model that may be implemented by any convenient manner.

HTML DOM

The Document Object Model (DOM) is a programming API for HTML and XML documents. It defines the logical structures of document and the way a document is accessed and manipulated. Anything found in an HTML or XML document can be accessed and changed, deleted or added the document Object Model, with a few exceptions in particular, the DOM interfaces the internal subset and external subset have been not yet specified. The DOM is a programming API for the documents. It is based on object structures that closely and resembles the structure of a documents and it models. For instance, consider method of this table has taken from an HTML document. In this we will take a sample html code and converted into a DOM tree.

Architecture

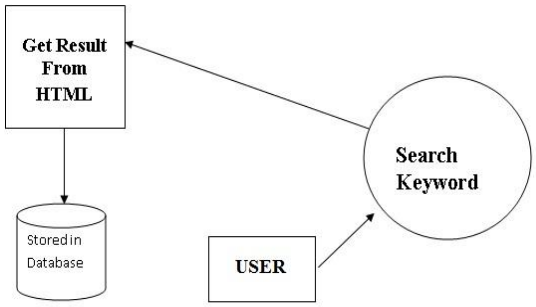

Fig 3. Architecture Diagram

IV. Performance & Evaluations

Similarity value between f1i and f2j each QRR includes two kinds of information: the text string for the ith value and the tag path for the ith value. During the pair wise alignment, we require that the data value alignments must satisfy the following three constraints:

1. Same record path constraint. The record path of a data value f comprises the tag from the root of the record to the node that contains f in the tag tree of the query result page. Each pair of these matched values should have the same tag path. Hence, if f1i has a different tag path with f2j, then sij is assigned a small negative value to prevent the pair of values from being aligned.

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

3. No cross alignment constraint. If f1i is matched to f2j, then there should be no data value alignment between f1k and f2l such that $k < i$ and $l > j$ or $k > i$ and $l < j$ [2].

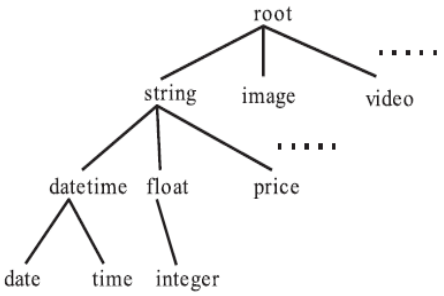

Fig 4. Data type tree

Holistic Alignment:

The two data values as an edge, the pair wise alignment set can be viewed as undirected graph. Thus our holistic alignment problem is equivalent to that of finding connected components in an undirected graph.

Each connected components of this graph represents a table columns inside in which the connected data values are different records are aligned vertically.

Evaluation Metrics

Two sets of evaluation metrics are used to compare the performance. The first set is a record level and includes the precision level and recall metrics are defined as C_c is the count of correctly extracted and aligned the QRRs. C_e is the count of extracted QRRs and C_r is the actual count of the QRRs in the query result pages. The number of QRRs is different query result pages its varies from the few to hundreds. Consequently pages with many QRRs will dominate the record level metrics. To deals with this problem where the C_p is the count of correctly extracted pages which means that all the QRRs in the pages are correctly extracted and it's aligned, N_a is the count of all the pages from which QRRs are extracted.

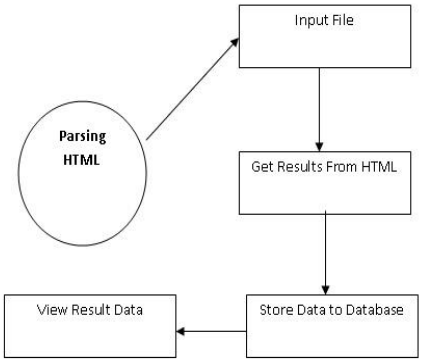

Fig 5. Data Flow Diagram of System Design

V. Experimental Results

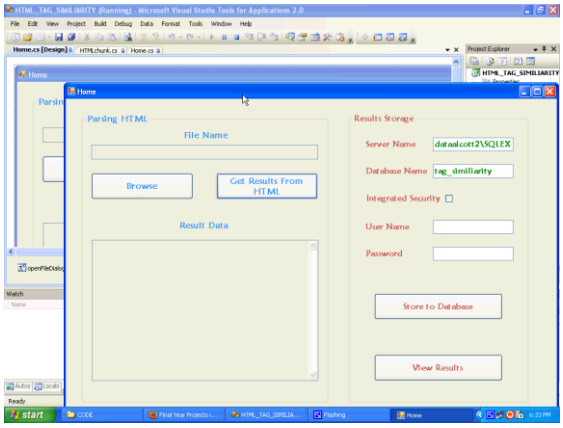

Fig 6. Home page & request server name, DB name

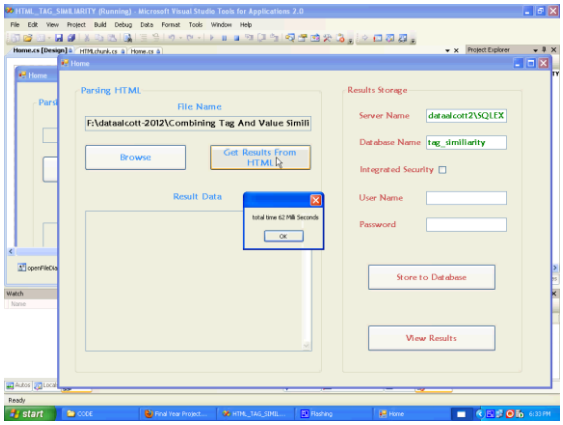

Fig 7. Select a URL Data base

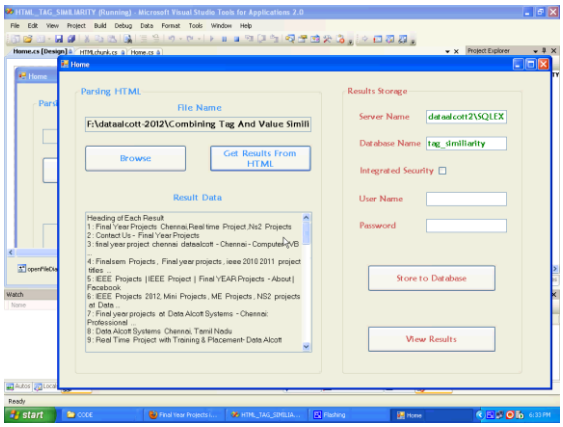

Fig 8. Getting a Result Data

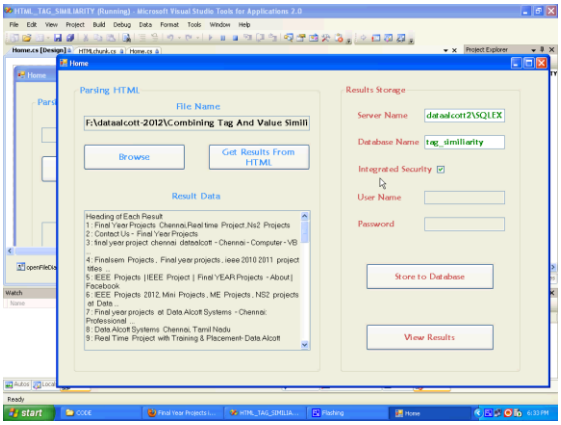

Fig .9 View Result Page

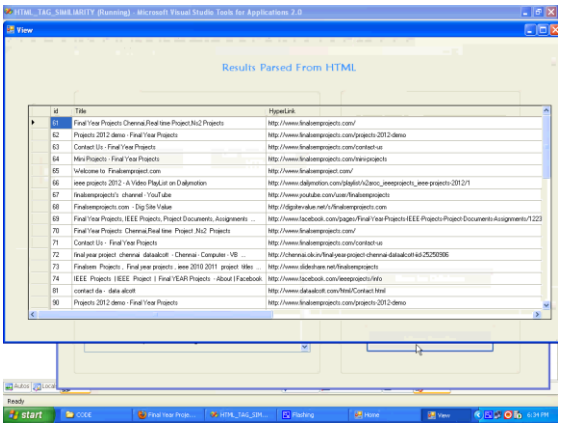

Fig 10 Total Html Result

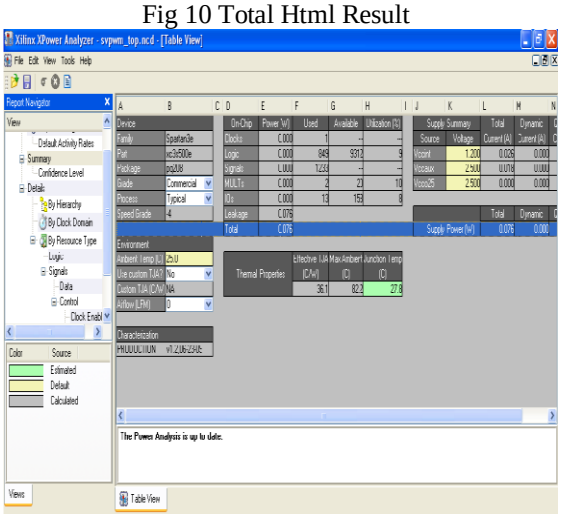

Fig: 11 Result parsed from html page

Testing and Integration

The purpose of testing method is to discover the errors. Testing is the process of an trying to discover every conceivable fault or weakness in a work product. It provides to a way to checks the functionality of component, sub-assemblies, and assemblies is a finished product. It is the process of an exercising the software with in the intent of ensuring the software system meets its requirement and user expectation and does not fails in unacceptable manners. Therefore a various types of test method have processed. In each test type addresses is a specific testing requirements. The following methods are types of testing methods,

1. System Test

2. White Box Testing

3. Black Box Testing

4. Integration testing

VI. Results & Discussions

The performance of the data extraction methods is a compared method in three different ways. Generally data set evaluation presents the performance of and the first three data sets, in which exhibits a variety of properties have been used in previous work by others. The other two evaluations focus on the specific properties of a query result pages. Non contiguous method QRR evaluations compares the performance method for query result pages in which the QRRs are contiguous and non contiguous. Nested structures evaluation of compares the performance for a query result pages with and without a

VII. CONCLUSION

It is presented a novel data extraction method, the CTVS to automatically extract the QRRs from a query result page. The CTVS employs two steps for this task. The first step identifies and segments of the QRRs. We improve method on existing techniques by allowing to the QRRs in a data region to be non contiguous. The second step aligns the data values among the QRRs. A novel alignment method is a proposed in which the alignment is performed in three consecutive steps: pair wise alignment, holistic alignment, and nested structures are processing. Experiments on five data sets shows that CTVS is generally more accurate than current state-of-the-art methods.

REFERENCES

[1] Ronald R. Yager and Frederick E. Petry, ” Hyper matching: Similarity Matching With Extreme Values” IEEE Transactions On Fuzzy Systems, Vol. 22, No. 4, August 2014.

[2] Fanman Meng, Hongliang Li, Guanghui Liu, and King Ngi Ngan, “From Logo to Object Segmentation” IEEE Transactions On Multimedia, Vol. 15, No. 8, December 2013.

[3] 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, July 2012.

[4] Jun Kong, Omer Barkol, Ruth Bergman, Ayelet Pnueli, Sagi Schein, Kang Zhang, and Chunying Zhao" Web Interface Interpretation Using Graph Grammars" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 42, No. 4, July 2012.

[5] Jer Lang Hong" Data Extraction for Deep Web Using Word Net" IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 41, No. 6, November 2011.

[6] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, and Silvia Quarteroni" A Framework for Integrating, Exploring, and Searching Location-Based Web Data" Published by the IEEE Computer Society 2011.

[7] Mohammed Kayed and Chia-Hui Chang, Member," FiVaTech: Page-Level Web Data Extraction from Template Pages", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 2, February 2010.

[8] Tak-Lam Wong and Wai Lam," Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 4, April 2010.

[9] Wei Liu, Xiaofeng Meng, and Weiyi Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 3, March 2010.

[10] Yanhong Zhai and Bing Liu," Structured Data Extraction from the Web Based on Partial Tree Alignment", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 12, December 2006