Survey Based on DOM and Visual Clues for Extracting Structure data from Web

M.E Computer Engineering Department, SOCET, Ahmedabad

Pathik B Shah,
M.E (C.E),
SOCET,
Ahmedabad,
Gujarat, India.

Prof. Hiren Patel
Assistant Professor,
Computer/IT Department,
SOCET,
Ahmedabad,
Gujarat, India.

ABSTRACT

This paper studies the problem of extracting data from a Web Page that contains several structured data records. The objective is to segment these data records, extract data items/fields from them and put the data in a database table. This paper proposes a new Method to perform the task automatically. It consists of two steps, (1) Identifying individual data records in a web page, and (2) Aligning and extracting data items from the identified data records. For Step 1, we propose a novel Document Object Model (DOM Trees). A technique based on tree matching. Removal of noise blocks is made from DOM trees. For step 2, we propose a method based on Visual Clues information Segment data records, which is more accurate than existing Methods. This approach enables very accurate Alignment of multiple data records. Experimental results using a large number of Web pages from diverse domains show that the proposed two-step technique is able to segment data records, align and extract data from them very accurately.

Keywords— Structured Data Extraction, Visual Clues, DOM tree, Web data mining, Web data extraction, Visual features for web pages.

I. INTRODUCTION

Most of the information on the World Wide Web shares the same template and has a structured HTML form as they are developed dynamically from the database. Extracting structured data from Web pages is the challenging problem. Many web sites have pages induced using the common template. For example, at Amazon site the author, title, comments, etc. are in presented the same way in all of its book pages. Web pages are produced by taking values from a database. So, there is a need to extract the values from the template generated web pages automatically. So, the ultimate goal of the proposed system is to provide unsupervised page-level data extraction approach to deduce matching schema for template pages. Also, there is a possibility that same web site can use variant templates, so we have to extract common schema.

Structured data refers to data expressed using the relational model. Structured data allows operations on domain-specific data elements. The Structured Web is that portion of Web information that could usefully be queried using a domain-sensitive representation. For example, a list of upcoming musical tour dates should be part of the Structured Web, but a poem would not be. An information extractor takes an unstructured input and emits a more-structured representation of the information - that is, the extractor adds domain-sensitivity to the representation. For example, an extractor transforms the unstructured textual representation of musical tour dates into a structured, more-domain specific, relational version.

II. LITERATURE REVIEW

Structured data extraction algorithms are based on the assumption that structured data is rendered regularly, usually in the form of records.

The extraction process is done in two stages:

1) Automatic annotation, which consists in recognizing instances of the input SOD's entity types in page content.
2) Extraction template construction, using the semantic annotations

From the previous stage and the regularity of pages the extraction process is done in five stages [10].

1) Type recognizer
2) Annotation and page sample selection
3) Wrapper generation
4) The output template
5) Stopping early the wrapper generation

Extracting data rely on an algorithm that performs page segmentation. The web page information can be obtained through the programming interface provided by browsers. In this paper the web pages transpose into a Visual Block tree and extract the visual information [3].

A Visual Block tree is actually a segmentation of a Web page. The root block represents the whole page, and each block in the tree corresponds to a rectangular region on the Web page [3]. Extraction of data records from deep web pages aims to discover the boundary of data records and extract them from the deep Web pages. The data records are the primary content of the deep Web pages and the data region is centrally located on these pages. The data region corresponds to a block in the Visual Block tree.

Ontology is just the standard description of the sharable and universal conception in a certain domain [4].

Proposes a three-step approach, including template generation, template detection and data extraction.

At the beginning, the training pages are clean and parsed into DOM trees for further process later

The DOM trees will be fed to tree clustering module which adapts to calculate similarities among the trees. DOM trees in the same cluster will be processed to generate a Seed Tree that represents all the trees in the cluster.

III. CHARACTERISTICS OF STRUCTURED DATA

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets.

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

Structured data has the advantage of being easily entered, stored, queried and analyzed. At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. Anything that couldn't fit into a tightly organized structure would have to be stored on paper in a filing cabinet.

IV. WEB DATA EXTRACTION TECHNIQUES

Web data extraction is using various techniques.

1) DEPTA is a two step approach. First step is to identify data record. Second step extracts data items using partial tree alignment method.

2) IEPAD is an information extraction system which applying pattern discovery techniques. It has three components, an extraction rule generator, pattern viewer and an extractor module. Extraction rule generator accepts input web page and generate extraction rules. Extraction rule generator includes a token translator, PAT tree constructor, pattern discoverer, a pattern validator and an extraction rule composer.

3) Viper It is a fully automated information extraction tool. This technique is based on assumption that the web page contains at least two consecutive data records which exhibits some kind of structural and visible similarity. ViPER is able to extract relevant data with respect to user's visual perception of the web page.

4) Road runner defines data extraction problem as "given a set of sample HTML pages belonging to the same class, find the nested type of the source data set and extract the source data set from which the pages have been generated".

5) EXALG performs template extraction in two stages. First stage is Equivalence Class Generation Stage (ECGM) and second is analysis stage.

6) FivaTech is a page-level web data extraction technique. Data extraction is performed in two modules. First module takes DOM trees of web pages as input and merges all DOM trees into a structure called fixed/variant pattern tree. In the second module template and schema are detected from fixed/variant pattern tree.

Comparison of web data extraction techniques is define in below table 1:

Techniques	Type of record considers	Extraction method	Single page/ Multiple pages
DEPTA	Flat	Partial tree alignment	Single
IEPAD	Flat	Wrapper induction	single
ViPER	Flat	Visual Perception based	Single

RoadRunner	Flat and nested	Wrapper induction	Multiple
EXALG	Flat	Equivalence class generation	Multiple
FivaTech	Flat	Tree merging and schema detection	Multiple

V. METHODS USED FOR EXTRACTING STRUCTURE DATA

1) *DOM Tree Based Approach*

Threshold and data filters to detect and remove irrelevant & repetitive information from the web page. The data filters will also be used to further improve the similarity check of data records. Our system will be able to extract 75%-80% user relevant content by eliminating noisy content from the different structured web pages like blogs, forums, articles etc. in the dynamic environment [6].

The basic structure of a Web page is DOM (Document Object Model) structure.

- The DOM structure of a Web page is a tree structure, where every HTML tag in the page corresponds to a node in the DOM tree.
- The Web page can be segmented by some predefined structural tags.
- Two nodes in the DOM tree have the same parent.
- The two nodes might not be more semantically related to each other than to other nodes.

The DOM tree structure fails to correctly identify the semantic relationships between different parts.

2) Visual Clues Concept

Vision-based approach is a web page Programming language independent approach is proposed. This approach utilizes the visual features of the web pages to extract data from deep web pages, including data record extraction and data item extraction. Again, we also propose a new evaluation measure revision to capture the human effort needed to produce the exact extraction of data [3].

Visual Clues are four step strategy.

a) Visual information transforms Visual Block tree.
b) Extract data record from Visual Block tree.
c) Partition the data record into data items & align the data items of same semantic together.

3) Partial Tree Alignment

This method to perform the task automatically. It consists of two steps, (1) identifying individual data records in a page, and (2) aligning and extracting data items from the identified data records. For step 1, we propose a method based on visual information to segment data records, which is more accurate than existing methods. For step 2, we propose a novel partial alignment technique based on tree matching. Partial alignment means that we align only those

data fields in a pair of data records that can be aligned (or matched) with certainty, and make no commitment to the rest of the data fields. This approach enables very accurate alignment of multiple data records. Experimental results using a large number of Web pages from diverse domains show that the proposed two-step technique is able to segment data records, align and extract data from them very accurately.

4) Decision Tree

According to the datasets by information extraction, a decision tree constantly updated data to update the decision tree, and then generate the understandable rules. The experiment proves that it is feasible to realize the Web information extraction based on the decision tree [11].

Decision Tree Technology, which has three courses.

a) Construct wrapper.
b) Decision tree building process, a rough decision tree will be constructed based on an algorithm.
c) Decision tree refinement process and to automatically extract knowledge or rules.

The advantages and disadvantages of the Structure data extraction from web methods are as follows:

Method	**Advantages**	**Disadvantages**
DOM Tree Based Approach	• Relatively simple to modify the data structure. • Robust API for the DOM tree. • Load the complex XML document • DOM fails to obtain significant improvement over baseline.	• Suitable for small data structure (document). • Store the entire document in memory. It is called memory intensive. • Only based on tags rely on the DOM structure to partition. • DOM seems to be the worst & unstable method.
Visual Clues Approach	• Visual Clues can help the user to divide the web page into several semantic parts. • Top-Down partition the web page based on separator. • Very good and stable. • Extract semantic structure of web page to some extent, base on visual perception.	• Web page programming language dependent. • Low efficiency, not scalable. • Labour intensive, Time consuming.
Partial Tree Concept	• Uses visual information to find the data records. • Visual information is utilized to infer the structural relationship among tags and to construct a tag tree.	• Computation time for constructing the tag tree is overhead. • It fails to identify some of the data records. • The tag tree can be built

		correctly only as long as the browser is able to render the page correctly.
Decision Tree	• It is very easy that the path from root node to leaf node contents the classification rules.	• Lower predictive capability. • Prune operation will get smaller tree

VI. ISSUE IN PREVIOUS APPROACH

The following challenges occur during the previous approach:

- Accuracy and robustness of Information Extraction System need to be improved.
- The programs of information extraction rely on the structure of Web pages, which makes program can't be reused.
- Suitable only for small number of blocks.
- The algorithm requires a large amount of memory space, very large amount of information will be lead to slow.

The following limitations are faced during the extracting structured data:

- Webpage programming language-dependent, or more precisely, HTML-dependent [3].
- Difficult to scale web page collections with a large and complex schema [2].
- Performance depends on the coverage of the training webpage for a set of web page embedding similar structured data.

VII. SYSTEM OVERVIEW

Step 1: Noise blocks like advertisements are removed from DOM trees. [13]

Step 2: Take two web pages as input

Step3: For each page, VB tree is constructed segmenting web page.

Step 4: Fixed/variant template Pages are detected by comparing blocks in VB trees.

Step 5: For fixed template pages, DOM trees of pages are combined by using tree merging algorithm [1], which consists of following steps.

1. Identification of peer nodes
2. Alignment of matrix
3. Repetitive Pattern Mining
4. Merging of optional nodes

Step 6: Pattern tree is constructed and schema is detected.

Step 7: By matching pattern tree and HTML tree, data are extracted.

Step 8: Data is extracted from variant template pages.

VIII. MOTIVATION

Many web sites contain large sets of pages generated using a common template or layout. This dataset is a real-world web page collection used for research on the automatic extraction of structured data (e.g., attribute-value pairs of entities) from the Web. We hope it could serve as a useful benchmark for evaluating and

comparing different methods for structured web data extraction [8].

IX. USE OF EXTRCTING STRUCTURE DATA

Such pages contain most of structured data on the Web. Extracted structured data can be later integrated and reused in a very big range of applications, such as price comparison portals, business intelligence tools, various mashups and etc. It encourages industry and academics to seek automatic solutions.

Many approaches to extracting data from the Web have been designed to solve specific problems and operate in ad-hoc domains. Other approaches, instead, heavily reuse techniques and algorithms developed in the field of Information Extraction.

X. CONCLUSIONS

In this paper we survey some web data extraction methods. This research aims to study about the method, which is combining tags and value similarities using DOM and Visual Clues. We survey on different techniques of data extraction from web document to extract information.

We have introduced an algorithm CombPs that can automatically extract structured data from the web page using the Document Object Model and Visual Clues Techniques which improves the performance and scalability of the extracting structured data.

These techniques are based on HTML structure, some technique identifies the data record without extracting data field, and some are based on visual information to extract data. This paper gives an idea of the standard format of extracting structured data from web.

XI. REFERENCES

[1] S. Krishna and J. Dattatraya, " Schema Inference and Data Extraction from Templatized Web Pages" 2015 IEEE International Conference on Pervasive Computing (ICPC).

[2] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.

[3] Satish J. Pusdekar and Shaikh. Phiroj Chhaware, "Using Visual Clues Concept for Extracting Main Data from Deep Web Pages" ,IEEE 2014.

[4] Haikun Hong, Xiaoxin Chen, Guoshi Wu, Jing Li, "Web Data Extraction Based on Tree Structure Analysis and Template Generation" IEEE 2010.

[5] DOM Tree Based Approach for Web Content Extraction, Bhavdeep Mehta and Meera Narvekar IEEE 2015.

[6] Li LIU, Junfang SHI and Xinrui LIU, "Web Information Extraction Algorithm based on Ontology and DOM Tree" IEEE 2010.

[7] Extraction Of Flat And Nested Data Records From Web Pages, P.S Hiremath, Siddu P. Algur, IEEE 2010.

[8] From One Tree to a Forest: a Unified Solution for Structured Web Data Extraction Qiang Hao, Rui CAI, Yanwei Pang, Lei Zhang, Microsoft Research Asia, Beijing 100080, P.R. China 2011.

[9] Nora Derouiche, Bogdan Cautis, Talel Abdessalem, "Automatic Extraction of Structured Web Data with Domain Knowledge", T´el´ecom ParisTech - CNRS LTCI Paris, France, 2012 IEEE

28th International Conference on Data Engineering

[10] Chen Hong-ye, "Method of Web Information Extraction Based on Decision Tree", School of Information and Electronic Engineering, Zhejiang University of Science and Technology Hangzhou 310023, 2009 International Forum on Information Technology and Applications.