DOM and Visual Clues for Extracting Structured data from Web

Pathik B Shah, Asst. Prof. Hiren J Patel

[1]M.E Computer Engineer, Silver Oak College of Engineering & Technology, Gujarat, India.

[2]Assistant Professor, Computer/IT Department, Silver Oak College of Engineering & Technology, 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. In this paper defines a new approach which has removed LRU (Least Recently Used) web pages. This approach enables very accurate Alignment of multiple data records and removes unwanted web pages which is don't use for the longest time. 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.

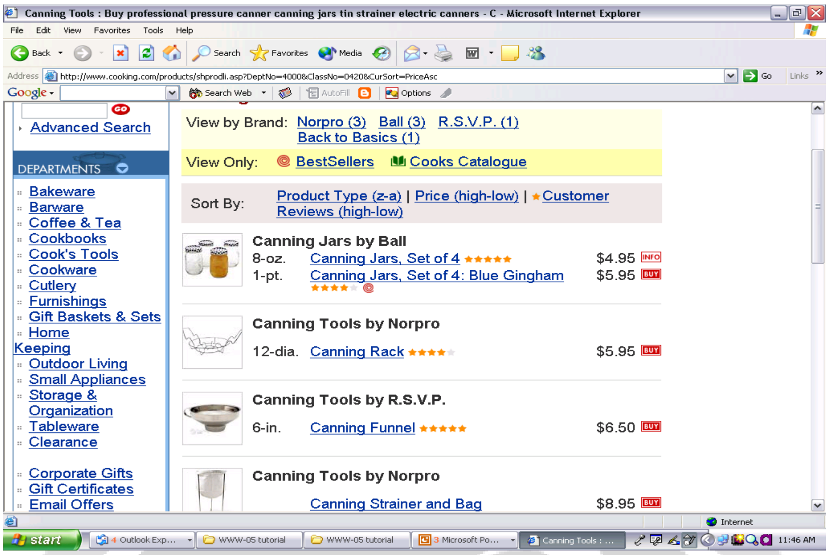
Keyword: - *Structured Data Extraction, Visual Clues, DOM tree, Web data mining, Web data extraction, Visual features for web pages.*

1. 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.

In this research paper DOM and Visual clues are most important method for extracting structured data from the web page. The Document Object Model (DOM) is a cross-platform and language independent convention for representing and interacting with objects in HTML, XHTML, and XML documents. The nodes of every document are organized in a tree structure, called the DOM tree.The usual first step is to build a DOM tree (tag tree) of a HTML page. New characteristics of web pages are Two-Dimensional Logical Structure and Visual Layout presentation.



Canning Tools : Buy professional pressure canner canning jars tin strainer electric canners - C - Microsoft Internet Explorer

Address http://www.cooking.com/products/shprodli.asp?DeptNo=4000&ClassNo=0420&CurSort=PriceAsc

- Bakeware
- Barware
- Coffee & Tea
- Cookbooks
- Cook's Tools
- Cookware
- Cutlery
- Furnishings
- Gift Baskets & Sets
- Home Keeping
- Outdoor Living
- Small Appliances
- Storage & Organization
- Tableware
- Clearance
- Corporate Gifts
- Gift Certificates
- Email Offers

View by Brand: Norpro (3) Ball (3) R.S.V.P. (1) Back to Basics (1)

View Only: BestSellers Cooks Catalogue

Sort By: Product Type (z-a) | Price (high-low) | ★Customer Reviews (high-low)

Canning Jars by Ball

8-oz.	Canning Jars, Set of 4 ★★★★★	$4.95 INFO
1-pt.	Canning Jars, Set of 4: Blue Gingham ★★★★☆	$5.95 BUY

Canning Tools by Norpro

12-dia.	Canning Rack ★★★★☆	$5.95 BUY

Canning Tools by R.S.V.P.

6-in.	Canning Funnel ★★★★★	$6.50 BUY

Canning Tools by Norpro

	Canning Strainer and Bag	$8.95 BUY

Fig -1 : An Example of Structure data from website.

Web Data Extraction techniques allow to gather a large amount of structured data continuously generated and disseminated by Web 2.0, Social Media and Online Social Network users. In the tree, each pair of tags is a node, and the nested tags within it are the children of the node. Noisy information such as Navigation, Decoration, Interaction, advertisement, etc… are removed by comparison of DOM tree structures.

2. LITERATURE REVIEW

Organized information extraction calculations depend on the supposition that organized information is rendered routinely, typically as records. The extraction procedure is done in two phases:

1)Automatic explanation, which comprises in perceiving occasions of the information SOD's element sorts in page content.
2)Extraction format development, utilizing the semantic comments

From the past stage and the consistency of pages the extraction procedure is done in five phases [10].

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

Removing information depends on a calculation that performs page division. The site page, data can be gotten through the programming interface gave by programs. In this paper the pages transpose into a Visual Block tree and concentrate the visual data [3].

A Visual Block tree is really a division of a Web page. The root piece speaks to the entire page, and every square in the tree relates to a rectangular district on the Web page [3]. Extraction of information records from profound website pages expects to find the limit of information records and concentrate them from the profound Web pages. The information records are the essential substance of the profound Web pages and the information area is midway situated on these pages. The information locale compares to a piece in the Visual Block tree.

Philosophy is only the standard portrayal of the sharable and all inclusive origination in a specific area [4]. Proposes a three-stage approach, including layout era, format recognition and information extraction.

Toward the starting, the preparation pages are perfect and parsed into DOM trees for further process later

The DOM trees will be sustained to tree grouping module which adjusts to ascertain likenesses among the trees. DOM trees in the same group will be prepared to produce a Seed Tree that speaks to all the trees in the bunch.

3. PROBLEM DEFINITION

For website pages made by some layout, discover blueprint that is the structure of pages and concentrate information.

Schema:
Schema of Web locales perceives the structure of Web webpage that distinguishes information of taking after sorts.

The following challenges occur during the previous approach:

1) Accuracy and robustness of Information Extraction System need to be improved.
2) The programs of information extraction rely on the structure of Web pages, which makes program can't be reused.
3) Suitable only for small number of blocks.
4) 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:

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

4. PROPOSED SYSTEM

For integrating information and for providing value added services like comparative shopping, finding the schema of the website has been a key step. Aim of our proposed work is a to find the schema of website and extract data.

System Overview:

Step 1 : Enter URL of website.

Step 2 : Extract each page URL of website and remove unwanted links and pages.

Step 3: Read log file of that website and enters data into database.

Step 4: Performed LRU to remove least recently used web pages from web pages list and update URL pages..

Step 5 : Build DOM (Document Object Model) tree of updated URL pages.

Step 6 : Defined page template type (Fixed / Variant) using CombPS algorithm.

Step 7 : Remove Noise from DOM tree pages.

Step 8 : Apply tree merging algorithm on Fixed page template.

Step 9 : Extract data by matching important blocks from variant template pages.

For extracting structured data define two new algorithm. LRU(Least Recent Used) Algorithm which is used to identify noises to improve efficiency of web mining and also removal of noises. Least Recent Used algorithm is less time consuming and less complex algorithm for web mining. And another is CombPS (Combined Page Segmentation) Algorithm, which is a combination of VIPS and FixedPS (Fixed-length Page Segmentation Approach). CombPS Algorithm takes advantage of both visual layout and length normalization

This proposed system is mainly solved a two problems of existing system which is Still have varying length problem, Labor intensive and Time consuming.

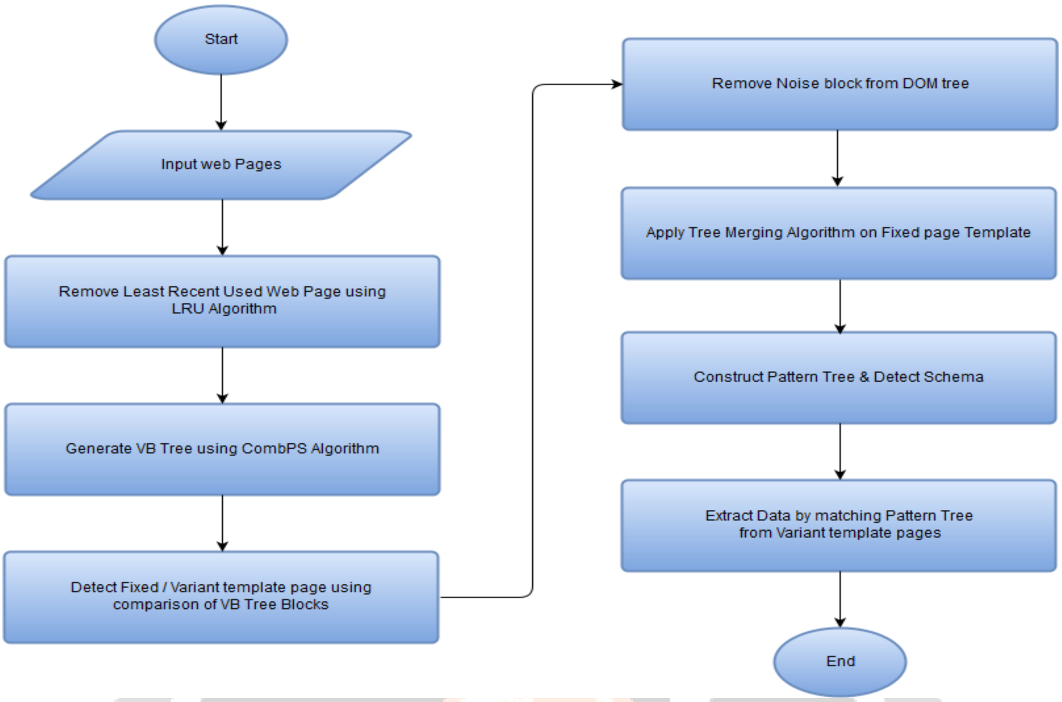

Start
Input web Pages
Remove Least Recent Used Web Page using LRU Algorithm
Generate VB Tree using CombPS Algorithm
Detect Fixed / Variant template page using comparison of VB Tree Blocks
Remove Noise block from DOM tree
Apply Tree Merging Algorithm on Fixed page Template
Construct Pattern Tree & Detect Schema
Extract Data by matching Pattern Tree from Variant template pages
End

Fig -2 : System Architechture

Step 1 : Input web pages.

First enter URL of website for given input web pages.

Step 2 : Extract web pages.

Extract each page URL of website and remove unwanted links and pages. When enter the url of web page and press the First Button (Extract Web Page and Remove Unwanted Pages) then list of extracted web pages is arrived and unwanted link of web pages are removed from this operation for example facebook and twitter web page of appropriate website..

Step 3 : Read log file.

Web Log File is defined as the each and every interaction of user with the web will be recorded and stored in the text file. Read log file of that website and enters data into database.

Step 4: Performed LRU

LRU is defined as Least Recently Used Web pages, remove the web pages where it's not used in recent times that means which is not useful for important data. Before LRU performed, there are many rows of web pages where user searches on the websites but its frequency rate is different. After LRU performed there are only small number of rows of web pages where user search on the websites.

Step 5 : Build DOM (Document Object Model) tree.

DOM provides a hierarchy structure for every web page. DOM is still a kind of linear structure & usually enables to represent the semantic structure of a web page. The Document Object Model (DOM) is a cross-platform and language independent convention for representing and interacting with objects in HTML, XHTML, and XML documents. The nodes of every document are organized in a tree structure, called the DOM tree. Finally build DOM (Document Object Model) tree of updated URL pages.

Step 6 : Defined page template type.

Page template type (Fixed / Variant) is defined using CombPS algorithm. CombPS algorithm is a combination of VIPS (VIsion based Page Segmentation Approach) and FixedPS (Fixed-length Page Segmentation Approach). CombPS Algorithm takes advantage of both visual layout and length normalization. Advantage of VIPS is solves the problems of noisy information and multi-topics and FixedPS can deal with the variant document length problem.

Page template type is defined as in the website shows that each web pages are concern to each other in semantic (Fixed Template) or Variant Template. Page template Type shows that webpage is Fixed or Variant.

Step 7 : Remove Noise from DOM tree pages.

Noise blocks are removed from the DOM tree. Noise Block is defined as a area which is common or not used in every web page. Due to of Noice block slow the extraction of structured data. The web page contains two types of blocks that are noise blocks and data rich blocks.

Noise blocks are the blocks of the web pages that contain data such as advertisements, navigational panels. Data rich blocks are blocks that contain data in which users are interested. Removing noise blocks from the pages is important as it improves the efficiency of the extraction algorithm and helps us to get accurate results. Noice block are must be removed from the web pages before processing the whole page of data. Header, Footer, Sidebar area is same in all the web pages, so that area blocks is defined as noise block and that are removed from DOM tree pages.

Step 8: Apply tree merging algorithm on Fixed page template.

Fixed page template is defined as which page structures are same as some web page. To detect the schema of website, all input DOM trees are combined to form fixed/variant pattern tree. Fixed-variant pattern tree is traversed from the root to downward.

Step 8: Construct pattern tree and Detect schema

In this progression, outline is surmised which recognizes fundamental sort, set sort, discretionary sort and tuple sort. Here, settled variation design tree is navigated from the root to descending and hubs are set apart as k-request or k-tuple. It is not important to label it as 1-tuple if hubs are with stand out kid and not set apart as set or discretionary sort.

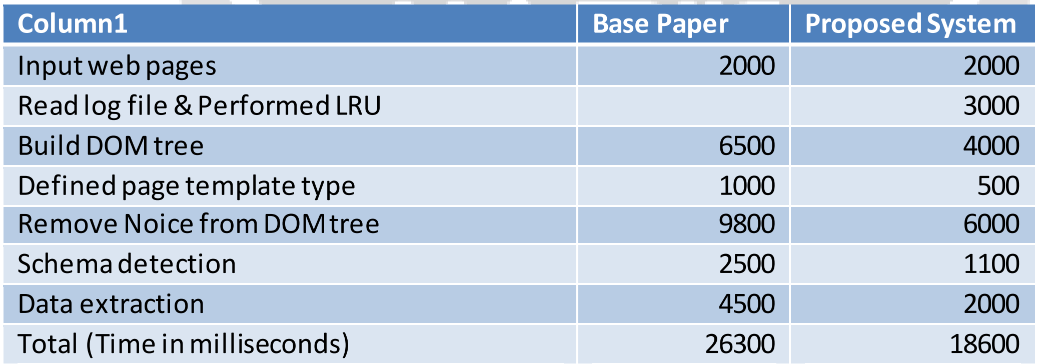
Step 9 : Extract data by matching important blocks from variant template pages.

Finally data is extracted which is in variant template. Data is extracted in fixed template in which division, that division data is searched into variant template and find that data to extracted.

A few sites use variation presentation formats for showing records. Additionally, different sites use diverse documentations for depicting items, e.g., some will utilize cost and some will utilize Rs. To depict the measure of item.

5. EXPERIMENTAL RESULT

For another measure of the experiment, time-based comparison is conducted. We use a system Intel Core i5-6200U with 12GB RAM for the experiment. The Third column shows the time consumed by the proposed system and the second column shows the time consumed by existing system in the table I.



Column1	Base Paper	Proposed System
Input web pages	2000	2000
Read log file & Performed LRU		3000
Build DOM tree	6500	4000
Defined page template type	1000	500
Remove Noice from DOM tree	9800	6000
Schema detection	2500	1100
Data extraction	4500	2000
Total (Time in milliseconds)	26300	18600

TABLE I. TIME-BASED COMPARISON

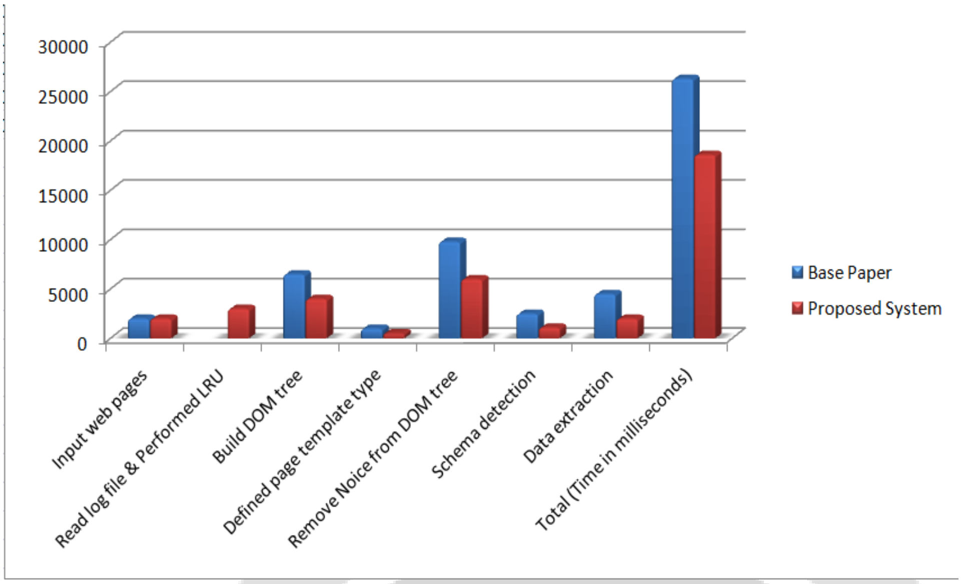


Fig -3 : Time Analysis for tested sites

As shown in above fig. 3, the graph represents the time analysis for the www.stellanstechnosoft.com sites given in the table. Here, X-axis represents the step of data extraction on which experiment is done and Y-axis represents time in milliseconds..Blue bar represents time required by existing system for extracting data from web pages removing noise blocks and orange bar represents time required by proposed system for extracting data from same web pages after removing noise blocks.

6. CONCLUSIONS

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. With the use of LRU (Least Recent Used) Algorithm to identify noises to improve efficiency of web mining. And also removal of noises. The Least Recent Used algorithm is less time consuming and less complex algorithm for web mining.We have introduced an algorithm CombPS algorithm is a combination of VIPS and FixedPS (Fixed-length Page Segmentation Approach). CombPS Algorithm takes advantage of both visual layout and length normalization. Using the Document Object Model and Visual Clues Techniques which improve the performance and scalability of the extracting structured data.These techniques are based on HTML structure, same technique identifies the data record without extracting data field, and some are based on visual information to extract data. This research gives an idea of the standard format of extracting structured data from web.

7. REFERENCES

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

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

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

[4] Web Data Extraction, Applications and Techniques: A Survey, Emilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd.

[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] Nora Derouiche, Bogdan Cautis, Talel Abdessalem,"Automatic Extraction of Structure Web Data with Domain Knowledge" IEEE 2012, Paris, France, 28th International Conference on Data Engineering.

[9] Shinde Shantaji Krishna, Joshi shashank Dattatraya, "Schema Inference and Data Extraction from Templatized Web Pages, IEEE 2015, ICPC.

[10] 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.

[11] Punam Bajaj, Payal Joshi, Anchal Garg, "Enhancement in DOM Tree to Reduce Time and Complexity – A Proposal", IJIRSET-2014.

[12] eDng Cai1* Shipeng Yu2* Ji-Rong Wen* Wei-Ying Ma, "Block-based Search Web" Microsoft Research Asia, Tsinghua University, Beijing, China.