A SURVEY ON WEB DATA EXTRACTION TECHNIQUES

Ankitha B[1], Chaitra Rao B[2], Naveen Chandavarkar[3], Balasubramani R[4]

[1,2]NMAM Institute of Technology, Department of Computer Science and Engineering
ankitha.shetty14@yahoo.com, raochaitra12@gmail.com
[3]NMAM Institute of Technology, Department of Computer Science and Engineering
Chandavarkar@nitte.edu.in
[4]NMAM Institute of Technology, Department of Information Science and Engineering
balasubramani.r@nitte.edu.in

Abstract

Web data extraction is an important part of web data analysis. Web pages generally contain large amount of data. Extraction of the web pages is just the opposite of web page generation. In the present market trends, to maintain healthy competition, there is huge necessity for the extraction of web data efficiently. Extraction of the data manually is tedious. Extracting information from web pages for the searchable websites has been a key step for web information integration. Huge efforts have been made to extract the data. There are lot of techniques to extract the data. Data is extracted either by using the HTML structure of the web page or by creating a DOM tree.

Keywords: *DOM Tree, HTML Structure, Web Data Extraction, Pattern Mining, Data Alignment.*

----------***----------

1. INTRODUCTION

Internet contains a vast amount of information and it is growing very fast. Web pages are meant for visualization. Web pages generally contain information in the form of "hidden web". Hidden web contains the contents from the databases that can be searched on the web. These searches are presented in the form of web pages. Search engine results can be either static or dynamic. Static web pages display the same data as stored in them. Dynamic web pages are capable of displaying different contents for different users. Analysing the web pages helps in decision making as we can recognize the price details, product specification and trends in the market. So it is necessary to extract the data from the web pages. Web data extraction is the process of obtaining data from the web. Extraction of the web pages is the opposite of the web page generation.The extracted data from the web is transferred either to a database or to excel or to another application. The process of web data extraction can be divided into four parts: crawling of the dynamic web pages, extracting the data-rich section of the page by the removal of noisy data like advertisement and navigation panel, constructing wrappers from these data-rich section of the web pages and finally using these wrappers to extract data from each page.

In this paper, we mainly focus on the extraction of the data from the web pages, either by HTML structure or DOM tree structure. This paper includes various techniques of web data extraction like DeLa, DEPTA, ROADRUNNER, NET, FivaTech, EXALG, IEPAD, W4Fand XWRAP.

2. LITERATURE SURVEY

Devika K etal [1] has worked on various techniques for web data extraction. There are mainly two types of data extraction techniques, one is using wrapper induction and the other is automatic extraction. There are many extraction techniques that are briefly explained below.

DeLa (Data extraction and Label assignment)is used for extracting the data from the websites and assigning labels to them. Extraction is done in four steps, first is collecting the labels of the website form elements. Second step is to generate the regular expression using wrapper generator. Wrapper generation can be done in three ways i.e., data-rich section extraction, c-repeated pattern and optional attributes & disjunction. The third step is data alignment that takes place in two phases. They are data extraction and attribute separation. The fourth step is to assign labels to the data present in the columns of the table. Figure 1 shows the architecture of DeLa technique.

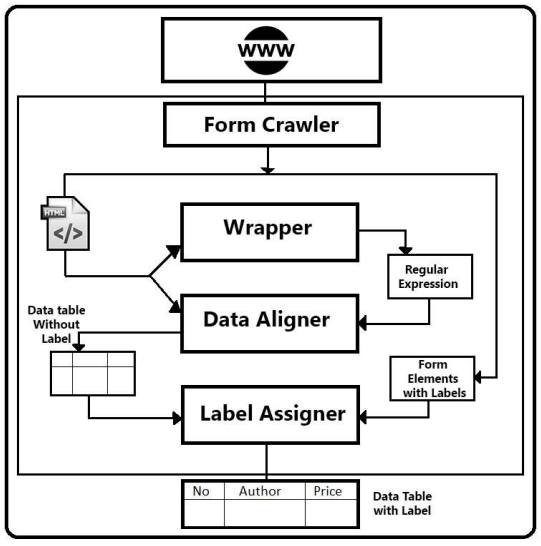

Figure 1. DeLa architecture

EXALG is used to extract the templates thattakes place in two phases. First stage generates equivalence classes and second stage is the analysis stage. In the first stage there are four steps that include differentiating the role of non-tag tokens, finding the equivalence classes, detection and removal of invalid LFEQs, differentiating the role of tag tokens. LFEQs refer to the equivalence classes that are large and contain tokens that appear in large number of pages. In the analysis stage, templates are constructed using the LFEQs. Figure 2 shows the two stages and their functionalities.

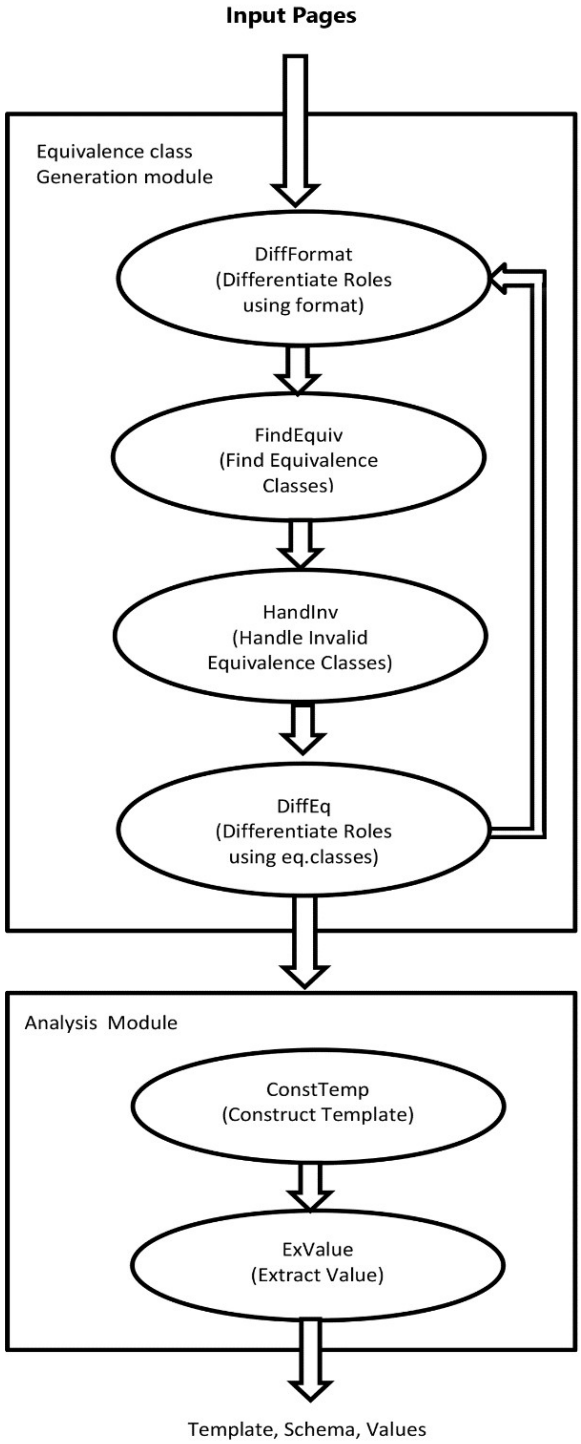

Figure 2. Stages of EXALG

NET (Nested data Extraction based on Tree matching) is a technique for extraction of data items from data records. It can also extract data from nested data records. The first step in this technique is building the tag tree that is done based on containment check i.e. check for nested data records. The second step is to identify the data records and then extract the data from them. This techniques uses post-order traversal for traversing the tag tree to find the nested records at the lower level. This ensures that the nested data records are found. For this, two algorithms traverse() and output() are used. [6] Before designing the algorithm, the extraction task is simplified by the following two observations

- The group of data records that contains the descriptions of a list of similar objects are represented in a contiguous region and these are then formatted using HTML tags.

- The similar data records placed in a region are represented in a tag tree under one parent node.The traverse() algorithm finds the nested records bycalling itself recursively only if the depth of the sub-tree from the current node level is greater than or equal to three as it unlikely to have a data record with only one level of tag. It uses simple tree matching algorithm. The output() algorithm will output the extracted data to the user in the relational table. Linked list data structure is used here to point to the next item in the list.

IEPAD is a technique used to extract information based on pattern discovery techniques. It consists of three components- extraction rule generator, pattern viewer and an extractor module. Extraction rule generator is used to generate the extraction rules. It includes token translator, PAT tree constructor, pattern discoverer, pattern validator and an extraction rule composer. Token translator generates tokens from the input webpage. Every token is represented by a binary code. PAT tree constructor receives the binary file to construct the PAT tree. Pattern discoverer finds out the repetitive patterns called as maximal repeats. Pattern validator generates the candidate patterns by removing the undesired patterns from the maximal repeats. Extraction rule composer composes the extraction rules in regular expression. Pattern viewer is a graphical user interface (GUI) that shows the discovered repetitive patterns. Extractor module extracts the desired information from the pages.

Vinayak B. Kadam et al [2] in his survey has mentioned that extraction of web data depends on the structure of the web page i.e., HTML structure or DOM tree structure. A few more data techniques for web data extraction have been mentioned in this survey paper.

W4F (Wysiwyg web wrapper factory) is used to generate the web wrappers using Java. It consists of three independent steps-retrieval, extraction and mapping step. Retrieval step is used to construct a parse tree using a parser that accepts the retrieved document as its input. Extraction step is used to extract the information from the parse tree using extraction rules. Extracted data is mapped in the mapping step.

XWRAP data extraction technique has four stages- structure normalization, data extraction, code generation, program testing and packaging. Structure normalization refers to reading the input pages, cleaning bad and ill formatted tags and converting the pages into the syntactic token tree. The data extraction module generates the extraction rules. Code generator generates the code from the extraction rules and the program validator and packaging stage validates the codes.

ViNTs (Visual information and Tag structure) is a technique for wrapper generation that uses visual as well as tag structure features. Data value similarity lines are identified using the visual data value similarity. Wrapper generation takes place by the combination of data value similarity lines and the tag structure regularity of the HTML page. The visual and the non-visual features are used to obtain the various extraction rules.

Kristina Lerman et al [3] from the University of Southern California has briefly explained about another web data extraction technique ROADRUNNER. It is a techniques used to generate wrapper for pages of the same class. It can extract data from nested pages also. This technique is efficient for web pages having similar structure. It compares two pages of the website and generates the wrapper in the form of Union-Free Regular Expression. The first page is considered to be the wrapper and this is matched with the rest of the pages, one-by-one and based on the mismatches the wrapper generation takes place. The two kinds of mismatches are tag mismatch and string mismatch. String mismatch is replaced by a general symbol in the wrapper. Tag mismatch can be either by the optional fields or by iterators. In either case regular expressions are used in the wrapper to represent the generalized data.

Mohammed Kayed et al [4] in his paper has overviewed FivaTech technique. It is used for extracting the data at the page level. It is performed in two stages. The first stage converts the web pages into a DOM tree and merges all the DOM trees into fixed or variant pattern tree. In the second stage these fixed or variant pattern trees are used for detecting the templates of the website. The first stage consists of four steps.

1. Peer-node recognition: In this step, comparison between the two nodes with same tag name is performed to check if they are peer sub-tree. Peer nodes are the nodes with same tag name but different functions.
2. Matrix alignment: In this step, the nodes in the peer matrix are aligned to obtain the aligned nodes list. This step recognizes the leaf nodes corresponding to data items.
3. Pattern mining: this step is used to detect the repetitive patterns by taking the aligned nodes list as input. For detecting the repetitive patterns, occurrence of the pattern for the first time is deleted for further mining of longer repeats. The output of this step is a list of nodes without any repetitive patterns. In this, we detect every repetitive pattern or tandem repeat and merge them by deleting all the occurrences except for the first one. This is because the data that is structured are nested.
4. Optional node merging: This step recognizes the optional nodes, the nodes that disappear in some column of the matrix. Then it groups the nodes according to their occurrence vector. Grouping of the optional nodes depends on the following rules:

- Set of adjacent optional nodes having the same occurrence vectors are grouped as optional.
- The set of adjacent optional nodes having opposite occurrence vector are grouped as disjunction.
- If the optional node is a fixed node, it is grouped as nearest non-fixed nodes.

Yanhong Zhai et al [5] in his paper has given a brief detail about DEPTA. DEPTA is an extraction technique based on the partial tree alignment. The first step in this techniques is to divide the web pages to identify the data records. This step is an improvement of MDR technique. MDR algorithm is based on two observations.

1. The data records that contains the set of similar objects are represented in contiguous regions of the page and are formatted using the same sequence of HTML tags.
2. Based on this observation, the tag tree is formed by the HTML tags in a web page. In this tree, the set of similar data records are formed by the child sub-trees of the same parent node.The steps in data record extraction are:

Step 1: building an HTML tag tree of theweb page.
Step 2: Data regions are mined in the web page using the tag tree. Data region is an area that contains a list of similar data records. Instead of finding the data records directly, first we find the data regions. Data regions can be found by comparing the tag strings of the individual nodes and combination of multiple nodes that are adjacent.
Step 3: From these data regions data records are identified.

3. COMPARISON

FivaTech is compared with DEPTA as they have the same 2 tasks i.e. data alignment and frequent pattern mining. In DEPTA repetitive patterns are first found and the alignment is done later. As a result, there is a possibility of loss of data. On the other hand, in FivaTech first data alignment takes place and later the frequently repetitive patterns are found. This avoids the loss of data. DEPTA can handle single data records whereas FivaTech can handle both single and multiple data records. FivaTech and EXALG use template and data tokens. Most of the above listed techniques consider only flat data records except for NET and ROADRUNNER.

CONCLUSION

In this paper, we study the various techniques to extract the data from the web pages either by HTML structure or by DOM tree structure. Some of the techniques use DOM tree to extract the data and some other techniques use HTML tag structure for the extraction i.e. by generating regular expression wrappers.

REFERENCES

[1]. Devika K, SubuSurendran. "An Overview of Web Data Extraction Techniques", volume 2, issue 4.
[2]. Vinayak B. Kadam, Ganesh K. Pakle. "A Survey on HTML Structure Aware and Tree Based Web Data Scrapping Technique", volume 5.
[3]. Kristina Lerman, University of Southern California."Automatic wrapper generation and data extraction.
[4]. Mohammed Kayed and Chia-Hui Chang. "FivaTech: Page level web data extraction from template pages", IEEE transactions on knowledge and data engineering, volume 22, no.2,2010.
[5]. Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment", Proc Int'l Conf. World Wide Web (WWW-14), 2005, pp. 76-85.
[6]. B. Liu and Y. Zhai, "NET- A System for extracting web data from flat and nested data records", Proc. WISE'05 Proceedings of the 6th international conference on web information systems engineering, 2005.

BIOGRAPHIES

Ankitha B is pursuing Bachelors of Engineering in Computer Science and Engineering in NMAM Institute of Technology, Nitte (an autonomous institution affiliated by VTU Belgaum). Her major work is in data mining and web automation. She is a member of CSI.

Chaitra Rao B is pursuing Bachelors of Engineering in Computer Science and Engineering in NMAM Institute of Technology, Nitte (an autonomous institution affiliated by VTU Belgaum). Her major work is in web extraction and data mining. She is a member of CSI.

Mr. Naveen Chandavarkar obtained his B.E in Computer Science and Engineering from VTU in 2008 and M.Tech in Computer Science and Engineering in 2012. He is pursuing his Ph.D in Computer Science and Engineering from VTU in the field of Opinion Mining. Mr. He is having 6 years of teaching experience, and sis a life member of Indian Society of Technical education.

Dr.Balasubramani obtained his BE in Electronics and Communication from Madurai Kamaraj University in 1990 and M.Tech in Information Technology from AAI-Deemed University, Allahabad in 2005. He obtained his Ph.D., in Information Technology from Vinayaka Missions University, Salem in 2011 for his research in the area of Digital Image Processing. Dr.Balasubramani is having 26 years of professional experience (12 years in industry & 14 years in teaching). Presently he is working as the Professor and Head in the Dept. of ISE at NMAMIT, Nitte.