DEEP WEB CONTENT EXTRACTION USING VISUAL APPROACH

Sumedha K. Chumble[1], Sayali P. Badhan[2]

[1] *PG Student, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University, Maharastra, India*
sumedha.chumble@gmail.com
[2] *Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education and Research, Savitribai Phule Pune University, Maharastra, India*
sayalibadhan@gmail.com

Abstract - *The Contents of World Wide Web has tremendous amount of data stored in web databases which can be searched through web query interfaces. The web pages searched by search engines are called surface web which need not be accessed through web databases while deep web can be accessed only by websites interfaces. Deep web pages are complicated in structure therefore extracting data from these web pages is a major problem. Solutions to this problem usually depend on web-page-programming language. Web pages are designed using HTML and it is frequently updating. For better presentation of web pages, more and more presentation techniques are used by web page designers. This makes structure of web page more complicated. Previous systems for deep web data extraction have many limitations. First, they are HTML dependent as they analyze HTML source code of web pages. Second, they are not able to handle increasingly complex structure of HTML source code. This leads to seek a different and efficient approach for deep web data extraction and to overcome limitations of previous works by using visual features. Visual features of web pages can be used for deep web data extraction. The visual system obtains visual representation of a given deep web page and converts it into Visual Block Tree. This Visual Block Tree helps to identify data region which contains the useful information to be extracted. After removing noise blocks a filtered data region is further processed to extract data records. Finally Visual wrapper gets generated for web database to which a given deep web page belongs. This improves efficiency of deep web data extraction as compared to previous method.*

Key Words: Deep web, Web mining, Visual Block Tree, Web data extraction, Wrapper generation.

1. INTRODUCTION

Deep web contains more valuable information than surface web. But making use of such consolidated information needs much efforts since the web pages are generated for visualization and not for data exchange. The World Wide Web is the large repository of information available to users who accesses the contents of the web. It has large number of web databases and these web databases can only be searched via web query interfaces. The web pages resulted by search engines are said to be surface web which can be accessed without accessing web databases. The surface web is basically static which might be is linked with other pages and deep web refers to those contents web cannot be indexed directly by the general search engine. Deep web can be accessed only by websites interfaces. So it is inaccessible to search engines. Thus the deep web is a complex entity that contains information from a variety of source types and it can be called as mixture of different file types and media. It is much more than static, self-contained Web pages. These web pages are dynamically generated and size of deep web is far larger than the surface web. The Deep Web contains information that resides in web databases which are behind portals. Deep web pages are dynamically-generated in response to a query through a web site's search interface and contain results to query which nothing but data records returned from web database. Thus deep web pages are complex in structure. Therefore a major research challenge is raised to unlock the contents of deep web.

Previous systems for deep web data extraction have limitations such as, they are HTML dependent because they are based on analyzing HTML source code of deep web pages and they are incapable of handling increasing complexity of HTML source code of web pages. Thus, the aim is to design a deep web data extraction system which will improve performance and quality of web data extraction using visual approach. Visual features of web pages can be used for deep web data extraction to improve extraction process.

1.1 Literature Survey

There are number of approaches for extracting information from web pages. In this section, we briefly

review previous approaches based on the degree of automation in Web data extraction.

1.1.1 Manual Approaches

The earliest approaches are the manual approaches in which languages were designed to assist programmer in constructing wrappers to identify and extract all the desired data items. Some of the best known tools that adopt manual approaches are Minerva, TSIMMIS, and Web-OQL. They have low efficiency and are not scalable.

1.1.2 Semi Automatic approaches

Semi Automatic techniques can be classified into sequence based techniques and tree-based techniques. The former approaches such as WIEN, Soft-Mealy and Stalker, represent documents as sequences of tokens or characters and generate delimiter based extraction rules through a set of training examples. The latter approaches, such as W4F and XWRAP parse the document into a hierarchical tree (DOM tree), based on which they perform the extraction process. These approaches require some manual efforts, for example, labeling some sample pages. So, they are labor-intensive and time-consuming.

1.1.3 Fully Automatic approaches

In order to improve the efficiency and reduce manual efforts, automatic approaches have been developed. Some representative automatic approaches are RoadRunner[2], IEPAD[3] and DEPTA[4].

1.2 Web Data Extraction Approaches

Different algorithms have been proposed earlier for deep web data extraction.

1.2.1 Roadrunner

It explores the features of HTML to automatically generate the wrappers. It compares the HTML tag structure of two or more sample web pages belonging to a same page class. Based on that comparison it generates a schema for the data contained in those web pages. From this schema, a grammar is inferred which can recognize instances of the attributes identified for this schema in the pages of the same class. All the extraction process is based on an algorithm that compares the tag structure of the sample pages. This algorithm generates regular expressions that handle structural mismatches found between the two structures. So, the algorithm discovers structural features such as tuples, lists and variations, So, the process of data extraction is fully automatic and no user intervention is required.

1.2.2 EXALG

The input to EXALG is a set of web pages created from the unknown template T and the values to be extracted. EXALG deduces this unknown template T and uses it to extract the set of values from the web pages encoded using the same template T as an output. EXALG detects the unknown template T using the two techniques namely differentiating roles and equivalence classes.

1.2.3 DEPTA

DEPTA is known as Data Extraction based on Partial Tree Alignment. Like IEPAD, DEPTA can be only applicable to web pages that contain two or more data records in a data region. However, instead of discovering repeat sub string based on suffix trees, which compares all suffixes of the HTML tag strings (as the encoded token string described in IEPAD), it compares only adjacent sub strings with starting tags having the same parent in the HTML tag tree (similar to HTML DOM tree but only tags are considered). DEPTA is limited to handle nested data records. So, a new algorithm, NET, is developed to handle such data records by performing a post order traversal of the visual-based tag tree of a Web page and matching subtrees in the process using a tree edit distance method and visual cues. DEPTA conducts the mining process from single Web pages, while RoadRunner do the analysis from multiple Web pages.

1.2.4 IEPAD

IEPAD is one of the first IE systems that generalize extraction patterns from unlabeled Web pages. This method exploits the fact that if a Web page contains multiple (homogeneous) data records that are to be extracted, they are often rendered regularly using the same template for good visualization. Thus, repetitive patterns can be discovered if the page is well encoded. Therefore, learning wrappers can be solved by discovering repetitive patterns. IEPAD uses a data structure called PAT trees which is a binary suffix tree to discover repetitive patterns in a Web page. Since such a data structure only records the exact match for suffixes, IEPAD uses center star algorithm to align multiple strings which start from each occurrence of a repeat and end before the start of next occurrence. Finally, a signature representation is used to denote the template to comprehend all data records.

2. VISUAL APPROACH

The Visual approach utilize visual representation of web page for extracting the contents. As the web page is usually displayed in a two-dimensional form, users has to browse the contents of the web page. This explores a new

research direction in which visual features can be used to extract deep web contents. It can also use some non visual information which includes same type of font, symbols and data types. Since the web pages consist most of text and images, web page layout and font are considered as visual information. The fonts are determined by its size, face, color, frame, etc., these visual features are important for identifying useful data in the pages. Thus, the visual features such as position, layout, appearance and content are used to extract data from web pages. The visual approach considers that the given deep web page contains useful data in a special region which is called data region. Using visual approach system extracts data records from data region on deep web pages. Deep web pages belonging to the same site are given as an input to the system and system generates visual wrapper for the web page.
Visual approach has following steps as shown in figure1

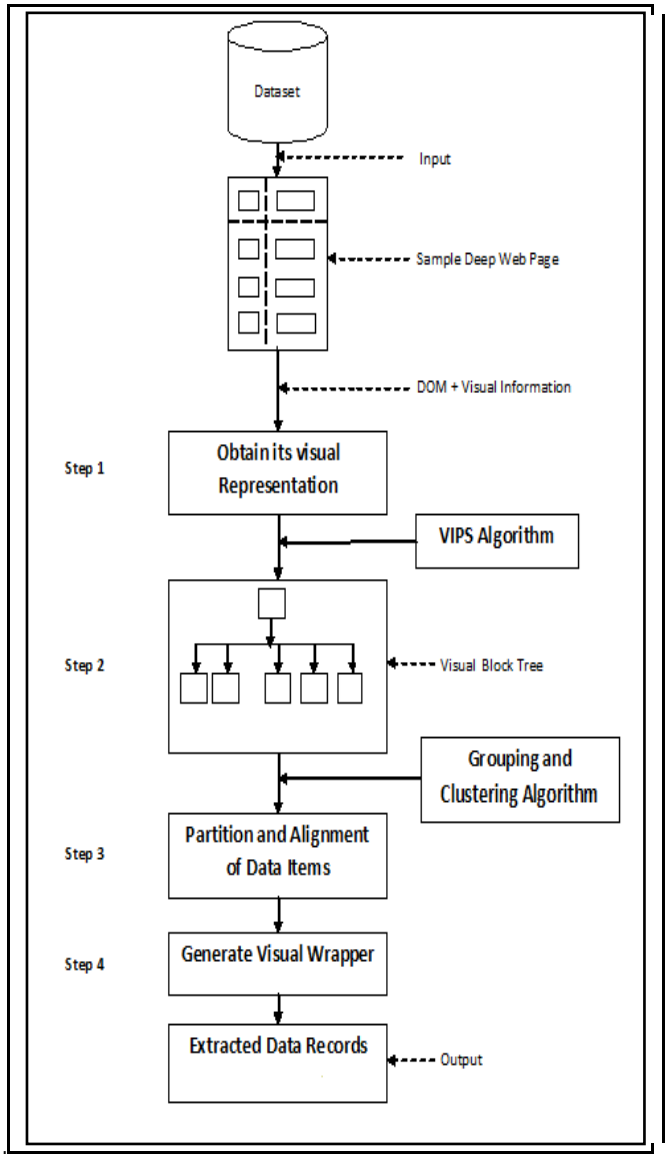

Fig -1: System Architecture

I. Sample Dataset
This system basically works on sample deep web pages. Sample dataset is used as an input to the system which is picked up from www.completeplanet.com which is the largest current deep web repository.

II. Visual Representation
It this part, visual representation is obtained and visual information is stored to be used in further part.

III. Visual block Tree generation
Using visual information from previous steps and VIPS algorithm, given web page is converted into Visual block tree. It is then processed further to get interested portion that is data region from which blocks are extracted.

IV. Partition and alignment
After all computations it generates data records which are clustered and regrouped together. These are then put into structured format.

V. Wrapper generation
Using visual features and extraction carried out in previous step, wrapper is generated for web database to which web page belongs.

VI. Extracted data
Finally it gives results in structured format which containing data records.

3. CONCLUSIONS
In this paper, an efficient visual approach for web data extraction is proposed which considers the fact that important data on web page is displayed in a special region called data region. The visual approach obtains visual representation of a given deep web page and converts it into Visual Block Tree. This Visual Block Tree helps to identify data region which contains the useful information to be extracted. After removing noise blocks a filtered data region is further processed to extract data records. Finally Visual wrapper gets generated for web database to which a given deep web page belongs. The web data extraction is improved by using visual information of deep web page.

REFERENCES

[1] Wei Liu, Xiaofeng Meng, Member, IEEE, and Weiyi Meng ,"ViDE: A Vision-Based Approach for Deep Web Data Extraction", IEEE Trans. Knowledge and Data Engg.,2010.

[2] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites," Proc. Int"l Conf. Very Large Data Bases (VLDB), pp. 109-118, 2001.

[3] Chang, C.-H. and Lui, S.-C., IEPAD: Information extraction based on pattern discovery. Proceedings of the Tenth International Conference on World Wide Web (WWW), Hong-Kong, pp. 223-231,2001.

[4] Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. Int'l World Wide Web Conf. (WWW), pp. 76-85, 2005.

[5] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. Int'l Conf. Data Eng. (ICDE), pp. 611-621, 2000.

[6] D.Cai, S. Yu, J. Wen, and Ma (2003), W. VIPS: A vision based page segmentation algorithm. Microsoft Technical Report MSR-TR-2003-79.

[7] J. Hammer, J. McHugh and H. Garcia-Molina, "Semi structured Data: The TSIMMIS Experience," Proc. East-European Workshop Advances in Databases and Information Systems (ADBIS), pp. 1-8, 1997.

[8] D. Cai, S. Yu, J. wen, and W. Ma (2003), "Extracting Content Structure for web Pages Based on Visual Representation," Proc. Asia Pacific web Conf. (APweb), pp. 406-417.

[9] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions," Proc. Conf.Information and Knowledge Management (CIKM), pp. 381-388, 2005.

[10] H. Zhao, W. Meng, and C.T. Yu, "Automatic Extraction of Dynamic Record Sections from Search Engine Result Pages," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 989-1000, 2006.

BIOGRAPHIES

Sumedha K. Chumble is pursuing M.E. (Computer) at K. K. Wagh Insitute of Engineering Education and Research from Savitribai Phule Pune University ,Nasik,Maharastra,India.

