Tactful Management Research Journal
Vol. 2 , Issue. 4 , Jan 2014
ISSN :2319-7943
Impact Factor : 0.119 (GIF)
ORIGINAL ARTICLE

VARIOUS APPROACHES FOR DEEP WEB DATA EXTRACTION

SHILPA DESHMUKH[1] , NEHA CHOPADE[2] , P. P KARDE[3] AND V.M. THAKARE[4]

[1]Assistant Professor SIES College of Management Studies Nerul, Navi Mumbai .
[2]Associate Professor SIES College of Management Studies Nerul, Navi Mumbai .
[3]Department of Information Technology, Government Polytechnic, Professor, Amravati, India.
[4]Professor & Head, Department of Computer Science, SGBAU, Amravati University, Amravati, India.

Abstract:

World Wide Web has vast useful Web databases which are difficult to extract relevant data from various sources. The number of Web databases has reached 50 millions according to a recent survey. These web databases can be searched through their web query interfaces. The web pages resulted are said to be surface web which can be accessed by search engines without accessing web databases and deep web refers to the web page that is not indexed by the general search engine. Deep web can be accessed only by websites interfaces. So it is inaccessible to search engines. So extracting data from deep page is critical problem.This paper studies some deep web data extraction techniques. A different way for deep web data extraction to overcome limitations of previous works is using visual approach. Visual features of deep web pages are used as primary concern to extract contents from deep web pages. It includes both data record extraction and data item extraction. Visual wrapper gets generated for web database to which a given deep web page belongs.

KEYWORDS:

Various Approaches , techniques , web databases , Deep Web Data Extraction.

INTRODUCTION

The World Wide Web is the promising field available to access the web contents. It has huge number of web databases and these web databases can be searched through their web query interfaces. The web pages which are results without accessing web databases are known as surface web which can be accessed by search engines. The surface web is static and is linked with other pages and deep web is web page that is not indexed by the general search engine. Deep web can be accessed only by websites interfaces. So it is inaccessible to search engines.

A large part of Deep web comprises of online structured domain specific databases that are accessed using web query interfaces. Web pages in the Deep Web are dynamically generated in response to a query through a web site's search form and often contain rich content. Data Extraction from Web even those web sites with some static links that are "crawlable" by a search engine often have much more information available only through a query interface. Unlocking this vast deep web content is a major research challenge.

Deep Web are dynamically generated in response to a query through a web site's search form and often contain rich content. Data Extraction from Web even those web sites with some static links that are "crawlable" by a search engine often have much more information available only through a query interface. Unlocking this vast deep web content is a major research challenge. Deep web pages have complex structure therefore extracting data from these web pages is critical problem. Web pages are designed using HTML and HTML is frequently evolving to newer versions. Previous systems for deep web data extraction

Please cite this Article as : SHILPA DESHMUKH[1] , NEHA CHOPADE[2] , P. P KARDE[3] AND V.M. THAKARE[4] , VARIOUS APPROACHES FOR DEEP WEB DATA EXTRACTION : Tactful Management Research Journal (Jan ; 2014)

have some limitations such as web-page-programming language dependency. First, they are HTML dependent because they are based on analyzing HTML source code of deep web pages. Second, they are not capable of handling ever increasing complexity of HTML source code of web pages. This motivates to seek a different way for deep web data extraction and to overcome limitations of previous works by using visual approach. Visual features of web pages can be used for deep web data extraction. The vision based 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 and data items. By considering the visual features, the web page programming language dependent problems can be solved. This approach aims at automatically adapting the information extraction knowledge previously learned from a source web site to a new unseen site, at the same time, discovering previously unseen attributes.

By considering the visual features, the web page programming language dependent problems can be solved. This approach aims at automatically adapting the information extraction knowledge previously learned from a source web site to a new unseen site, at the same time, discovering previously unseen attributes. The four step strategy is employed for the extraction. They are given as:

(1) Taking a sample deep Web page from a Web database, obtain its visual representation; transforming it into a Visual Block tree.
(2) From the visual block tree extract the data records.
(3) Then, the data item separation and align the data items of same semantic together.
(4) Generate Visual wrappers for the resulted web database of the sample deep web pages.
This improves efficiency of deep web data extraction.

Related Work

1Manual Approach:- In this approach users program a wrapper for each Web site using general programming languages . Language can be Perl or any special-designed languages. These tools need the user to have enough computer and programming backgrounds, hence it becomes expensive. This manual approach utilizes various tools. Some of the tools are:

Minerva: This tool uses the grammar in EBNF style, for each document, a set of productions is defined. This tool attempts to combine advantage of a declarative grammar based approach with features typical for procedural programming language by incorporating an explicit exception – handling mechanism inside the grammar[1].

web OQL: This tool is a declarative query language capable of locating selected pieces of data in the HTML pages. This tool originally aims at performing queries like SQL over the web[2].

TSIMMIS: This tool basically meant for semi structured data . It includes wrappers which can be configured through specification files written by the user. Specification files are composed by a sequence of commands that define extraction steps. An extractor based on the specification file parses an html page to locate the interesting data and extract them. TSIMMIS provides two important operators: split and case. The split operator is used to divide the input list element into individual elements. The case operator allows the user to handle the irregularities in the structure of the input pages[3].

2Semi-Automatic Approach: This approach uses the HTML – aware tools. The semi-automatic technique is broadly classified into text-based and sequence based technique. It rely on inherent structural features of HTML documents for accomplishing data extraction and grouping. The documents are transformed into parsing tree before processing. Some tools of this approach are W4F [4], XWRAP [5]. These are briefly summarized as follow:

XWRAP: is an important HTML –aware tool for semi automatic construction of wrappers. The tool features a component library that provides basic building blocks for wrappers, and a user friendly interface to makes wrapper development task easy. Here the wrapper generation process is classified into two phases: structure analysis and source -specific xml generation. In the first phase, XWRAP fetches, cleans up, and generates a tree-like structure for the page. Then the system identifies regions, semantic tokens of interest and useful hierarchical structures of sections of the page by interacting with users through object (record) and element extraction steps. In the second phase, the system generates a XML template file based on the content tokens and the nesting hierarchy specification, and then constructs a source-specific XML

generator. XWRAP can be classified as supervised WI systems for no rule writing is necessary; however, it requires users' understanding of the HTML parse tree, the identification of the separating tags for rows and columns in a table, etc[4].

World Wide Web Wrapper Factory: This is a toolkit for the construction of wrappers. It is the java toolkit for building wrappers. The wrapper development process consists of three independent layers. They are: Retrieval layer, Extraction layer, and Mapping layer. In the retrieval layer, a to-be processed document is retrieved (from the Web through HTTP protocol), cleaned and then fed to an HTML parser that constructs a parse tree following the Document Object Model (DOM). In the extraction layer, extraction rules are applied on the parse tree to extract information and then store them into the W4F internal format called Nested String List (NSL). In the mapping layer, the NSL structures are exported to the upper-level application according to mapping rules. Extraction rules are expressed using the HEL (HTML Extraction Language), which uses the HTML parse tree (i.e. DOM tree) path to address the data to be located. This tool kit classifies the wrapper development process in three phases: first, the user describes how to access the document, second, he describes what pieces of data to extract, and third, he declares what target structure to use for storing the data extracted[5].

3Automatic Approach: The automatic approaches are primarily on text-based and tag-structured based approach. This approach uses tools that each tool will perform their functions separately. They do not combine their process to give whole result. Each process is independent of their functions. Though this approach is automatic, it has some limitations. The tools used by this approach are Roadrunner [6], IEPAD [7], DEPTA [8].

Roadrunner : It is a tool that explores the inherent features of HTML documents to automatically generate wrappers. By comparing HTML structure of web pages of same “page class”, generating a result of schema for the data contained in the pages. The unique feature of this tool is that no user intervention is requested[6].

IEPAD: This tool generalizes the extraction pattern from the unlabelled web pages. If a web page contains multiple homogenous data records to be extracted, they are rendered using the same template which provides good visualization. The center star algorithm is applied for the alignment of multiple strings[7].

DEPTA : 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 substring 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 substrings with starting tags having the same parent in the HTML tag tree (similar to HTML DOM tree but only tags are considered). The insight is that data records of the same data region are reflected in the tag tree of a Web page under the same parent node. Thus, irrelevant substrings do not need to be compared together as that in suffix-based approaches. Furthermore, the substring comparison can be computed by string edit distance instead of exact string match when using suffix trees where only completely similar substrings are identified[8].

Visual Based Appraoch

Deep web pages always arrange the data records and data items with visual regularity. These visual features of the data record and data items are be utilized to extract deep web data automatically. This visual based approach [9] employs a four-step strategy is described as follows:

1.Obtaining Visual representation: Given a sample deep web page from a Web database, visual representation is obtained and transformed it into visual Block Tree. 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 [9].

The leaf blocks are the blocks that cannot be segmented further, and they represent the minimum semantic units, such as continuous text or images. Above figure a) shows a popular presentation structure of deep Web pages and b) gives its corresponding Visual Block tree. VIPS algorithm [10] will be employed to transform a Web page into a Visual Block tree, which will enable extraction of, the visual information. Figure1 first a) shows the presentation structure of deep web page and figure1 b) shows obtained Visual Block Tree. The designers often deploy different types of information with distinct visual characteristics to make the information on Web pages easy to understand.

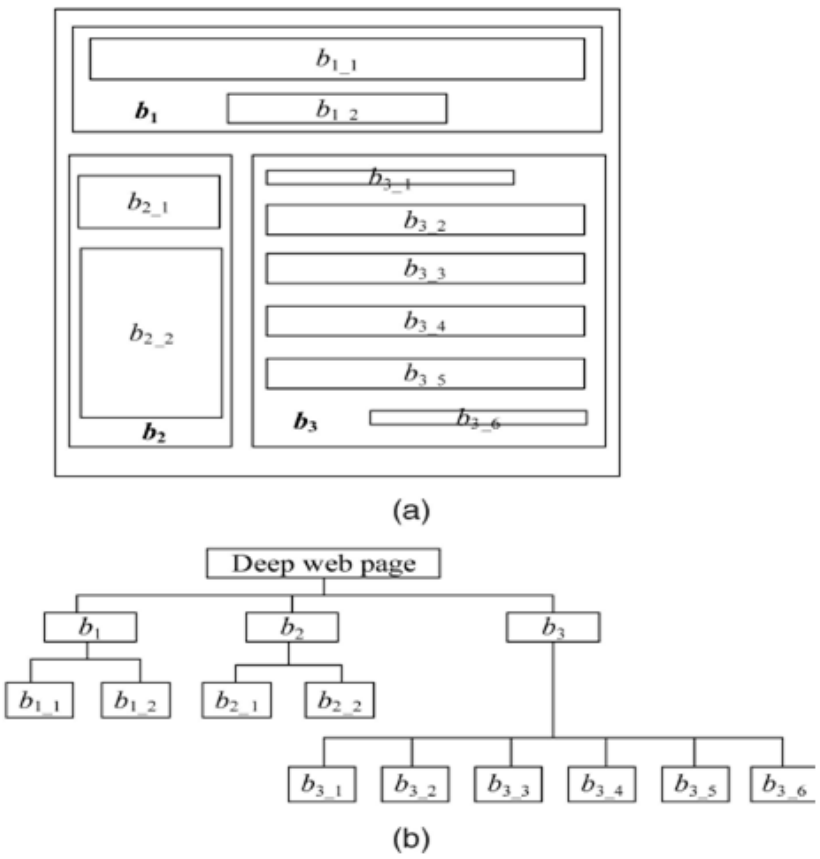

Fig1: a)The Presentation Structure and b) its Visual Block tree.

Four types of features will be studied:

1.Position features: indicate the location of the data region on a deep Web page.
2.Layout features: indicate how the data records in the data region are typically arranged.
3.Appearance features: capture the visual features within the data records.
4.Content features: hint the regularity of the contents in data records.

1.Extraction of data records and data items: This aims to discover the boundary of data records and extract them from deep web pages. It will try to accomplish:
a.All data records in the same as well as all given data regions.
b.For each extracted record no data item should missed and no incorrect data item is included.
Regions are located first instead of extracting data record directly from the deep web page and extract data records from the data region.

Data records consist of group of data item and some static template texts. Our approach will focus segmentation of records into data items and aligning the data items of the same semantic together.

2.Query Translation and Analysis: aims to implement correlation mining approach also aim to develop a correlation measure for positive and negative correlation. Co-occurrence pattern of attributes are investigated across sources to match the schema. Unlike most schema matching work matches two schemas at a time, all the schema are matched at the same time using m:n correlation.

3.Generation of Visual wrapper : This has two components:
a.Visual Data Record Wrapper: first locates the data region in the Visual Block tree, and then extracts the data records from the child blocks of the data region. Our wrapper aims to find the first block of each record and the last block of the last data record.

b.Visual Data Item Wrapper: Such data alignment algorithm is employed which groups data item from different data records into columns or attributes such that data item under the same column have the same semantic.

CONCLUSION

This discussion focuses on the deep web data extraction problem including data record extraction and data item extraction. First, we surveyed previous works on web data extraction and their inherent limitations. A new visual approach is introduced to achieve deep web data extraction. This vision-based approach is intended to solve the HTML–dependent problem. This approach employs the extraction of structured data using visual features, providing more efficiency. The primary steps in this approach are building visual block tree, extraction of data records and data items and the construction of visual wrappers.

REFERENCES

1.C.-H. Chang, C.-N. Hsu, and S.-C. Lui, “Automatic Information Extraction from Semi-Structured Web Pages by Pattern Discovery,” Conf. Data Eng., 2000
2.D. Cai, S. Yu, J. Wen, and W. Ma, “Extracting Content Structure for Web Pages Based on Visual Representation,” Proc. Asia Pacific
3.G.O. Arocena and A.O. Mendelzon, “WebOQL: Restructuring Documents, Databases, and Webs,” Proc. Int'l Conf. Data Eng. (ICDE), pp. 24-33, 1998.
4.http://db.cis.upenn.edu/DL/www8.pdf (accessed on 10th Oct 13)
5.J. Hammer, J. McHugh, and H. Garcia-Molina, “Semistructured Data: The TSIMMIS Experience,” Proc. East-European Workshop Advances in Databases and Information Systems (ADBIS), page 1-8, 1997.
6.L. Liu, C. Pu, and W. Han, “XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources,” Proc. Int'l
7.V. Crescenzi and G. Mecca, “Grammars Have Exceptions,” Information Systems, vol. 23, no. 8, pp. 539-565, 1998
8.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 Web Conf. (APWeb), pp. 406-417, 2003.
9.Wei Liu & Xiaofeng Meng ViDE: A Vision – Based Approach for Deep Web Data extraction in proceeding of IEE transaction (vol. 22 no. 3) pp. 447-460 March 2010.
10.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

SHILPA DESHMUKH
Assistant Professor SIES College of Management Studies Nerul, Navi Mumbai

NEHA CHOPADE
Associate Professor SIES College of Management Studies Nerul, Navi Mumbai