Analysis of DOM Based Automatic Web Content Extraction

Bhavdeep Mehta
Department of Computer Engineering
D. J. Sanghvi College of Engineering
Mumbai, India
mehtabhavdeep@gmail.com

Shaikh Sakina Banu
Department of Computer Engineering
D. J. Sanghvi College of Engineering
Mumbai, India
shaikhsakinab@yahoo.com

Abstract— **The World Wide Web plays an important role while searching for information in the data network. This paper deals with research in the area of automatic extraction of textual and non-textual information. Developed method consists of two data types of extractions i.e. image and text data extraction. The extraction is performed using the concepts of Document Object Model (DOM) tree. The paper presents a series of data filters to detect and remove irrelevant data from the web page. Many web applications adopt AJAX to enhance their user experience. But AJAX has a number of properties making it extremely difficult for traditional search engines to crawl. The paper proposed an AJAX crawling scheme based on DOM and breadth-first AJAX crawling algorithm.**

keyword: DOM, Extraction images, Content Extraction.

I. Introduction

The World Wide Web is flooded with a lot of information.

Today, the Web can be partitioned into the Surface Web reached by common crawler-based techniques, and the rapidly growing Deep Web or Hidden Web, which consist of structured data hidden behind search forms. There was a need for a mechanism to extract the relevant data and separate it from the useless and irrelevant data. Thus wrappers were created for this purpose. The proposed wrapper uses Document Object Model (DOM) tree properties. Data records are structured data objects retrieved from a backend database and displayed in Web pages with some fixed templates. A group of data records that contains descriptions of a set of similar objects is typically presented in a contiguous region of a page using similar tags. Web pages contain multiple data records. The underlying data records are discovered. The data records are further extracted and the semi-structured HTML pages are then converted into DOM trees.

However, the multimedia documents consist of different Components (texts, images, sounds, videos, etc.), our approach is oriented to determine the relation between digital images and textual segments in web documents. The main idea is to establish automatically the correspondences between the images and their associated texts in order to ensure the optimal use of any heterogeneous resources.

In this paper, we proposed an AJAX crawling scheme based on DOM and breadth-first AJAX crawling algorithm. The AJAX crawl controller can utilize a set of DOM elements which capable of triggering AJAX events, to invoke corresponding JavaScript functions. Along with the invocation to AJAX event, the structure of DOM tree on the user interface is also changed. Because of the change of DOM tree representing the state change, so from the structure changes of DOM tree, we can build a State Transition Graph(STG). The STG contains a set of states and the possible transitions between them.

The rest of the paper is organized as follows:

Section II we will dicuss about all the techniques used for information retrieval

II. Textual Information Retrieval

The following section gives the proposed method of implementation of the wrapper to retrieve text information. An overview for VEDD (Visual Wrapper for Extraction of Data using DOM Tree) wrapper application is as follows:

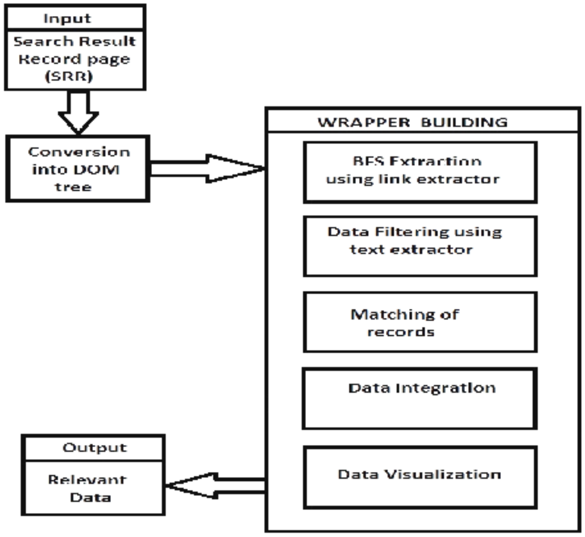


Figure 1. Diagrammatic representation of VEDD wrapper

A. Overview of VEDD

When a query is submitted to a search engine, the search engine returns dynamically generated result page containing the result records. The wrapper is divided into two main parts. The first part involves the parsing of the HTML web page and storing them as Document Object Model (DOM) tree. In the second part, the wrapper carries out the various extraction techniques to extract the relevant data records i.e. the SRRs (Search Record Results).

In first component, VEDD wrapper converts and stores a given HTML web page from a search engine into a DOM tree. In the second component, VEDD undergoes different stages of extraction of data records. VEDD reduces the list by removing irrelevant data records in each data region using filtering techniques.

B. Phase1: DOM tree

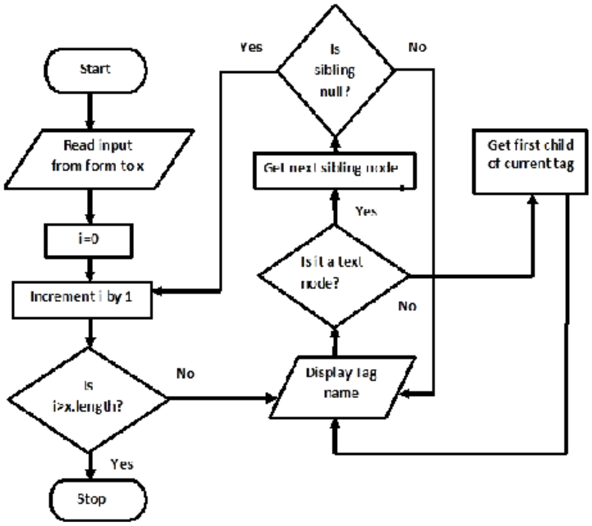


Figure2. Flowchart for DOM tree creation

The DOM tree approach is the most effective way to identify the HTML tags or codes before the web data extraction. In VEDD wrapper, a DOM tree is created as a reference to gauge the structure of the page and is especially useful for unstructured or semi-structured source code.

C) Phase 2:VEDD extraction module

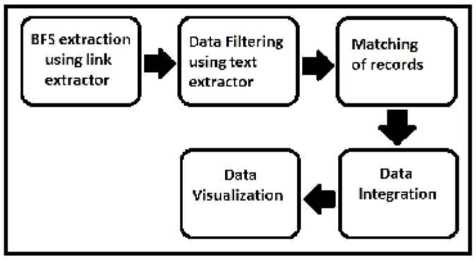

BFS extraction using link extractor
Data Filtering using text extractor
Matching of records
Data Integration
Data Visualization

Figure 3:Extraction Process

In the following extraction process, a DOM tree is created first, followed by the BFS (Breadth First Search) extraction process. After carrying out this process, filtering stages are performed followed by the final extraction of potential relevant data regions separated from the irrelevant ones.

Stage 1: BFS Extraction using link extractor:

After the DOM tree is created BFS is carried out. Usually potential data records in a DOM tree can be identified as those which lie in the same level in the tree and which have a common parent tag and recurring sequence of HTML tags.

Stage 2: Data Filtering using text extractor:

The next stage is a filtering stage called the text extractor stage [16]. In this stage the text which forms the title of the hyperlink in the SRRs, as well as their descriptions are extracted.

Stage 3: Matching of records (Similarity Filter):

After all the relevant links, link titles and record descriptions have been extracted from each of the search result pages of different search engines, there are possibilities of the results containing some redundant records.

Stage 4: Data Integration:

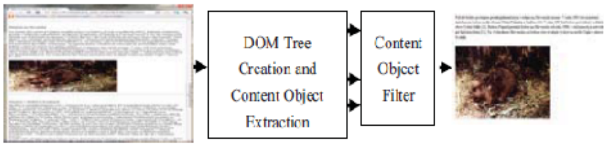
The purpose of Data Integration is to find the easiest way to combine and integrate data records from different search engines.

Stage 5: Data Visualization:

Data Visualization is the last stage in the creation of VEDD wrapper, in which the extracted data is visually presented to the users.

III. NON-TEXTUAL INFORMATION RETRIEVAL

Image analysis is the extraction of meaningful information from images. Images analysis can be defined as identifying a person, animals, buildings and other objects in image. It is based on digital image processing techniques. For doing this processing pipeline is used as follows:



In first stage, the article body is extracted. The web HTML page or collections of HTML pages are processed by the DOM. The output of DOM is tree of various content objects. The DOM objects are analyzed to extract the content blocks that contain the article text body. In the next step, content sub-blocks in the article block are further analyzed to eliminate easy-to identify unwanted blocks such as lists of links.

A) Hybrid Segmentation Algorithm

Hybrid methods are created by combining two or more image segmentation algorithms. In first step, the hybrid algorithm is applied to image filtering using Mean Shift following the image dividing into segments by applying the Mean Shift. In the second step, the image is split into segments using Mean Shift algorithm. In final step, the similar small segments are combined into bigger segments, through Belief Propagation.

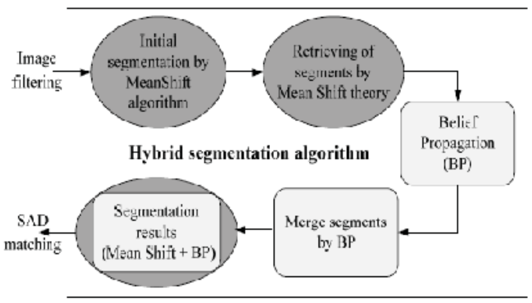


Image filtering
Initial segmentation by MeanShift algorithm
Retrieving of segments by Mean Shift theory
Belief Propagation (BP)
Hybrid segmentation algorithm
SAD matching
Segmentation results (Mean Shift + BP)
Merge segments by BP

Fig. 3: Hybrid segmentation algorithm

B) SIFT algorithm

Scale Invariant Feature Transform (SIFT) is an algorithm in computer vision to detect and describe local features in images . The SIFT descriptor extracts from an image a collection of frames or keypoints. SIFT descriptor is coarse descriptor of edges found in the image. In first step, the locations of potential interest points are computed in the image by detecting the maximal and minimal sets of Difference of Gaussian (DoG) filters applied at different scales around the image. Then, these locations are refined by discarding points of low contrast and an orientation is assigned to each key point based on local image features .Finally, a local feature descriptor is computed at each key point. This descriptor is based on the local image gradient, transformed according to the orientation of the key point to provide orientation invariance.

C) Segmentation of image and text

For better understanding of segmentation of text and image in the web page ,the schematic diagram is given as:

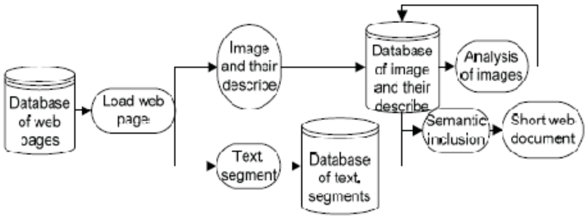


Fig 4: schematic diagram for separating image and text for extracting short web document

The goal of this method is to obtain the semantic description of web image followed by a brief web document. This method is based on non-textual information of web documents analysis. The process of semantic description is based on four steps. In the first step, the web page with identification of images and textual segments are loaded by Document Object Model . Next, the obtained information are stored into different databases. In third step,the analysis of non-textual information's based on feature extraction and SVM classification is made. The results of semantic descriptions are stored into database. Finally, the semantic inclusion of digital images and textual segments is implemented.

IV. USING AJAX AND DOM

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 DOM is used to construct additional internal structures used to display the page in the browser window. The nodes of every document are organized in a tree structure, called the DOM tree. The top most node in the DOM tree is the Document object. Each node has zero or more children. The DOM is the way JavaScript sees the browser state and the HTML page it contains. The state changes in AJAX applications are dynamically represented through the runtime changes on the DOM. That is to say search engine capable of crawling and indexing AJAX applications, can make use of this run-time dynamic DOM of the AJAX application. An AJAX application as not only a simple page identified by an URL, but also as a series of states and transitions.

CONCLUSION

This paper gives a brief idea for extraction of textual and non textual information using various techniques based on DOM tree structure. Thus we conclude that, the frequency of occurrence of data records in SRR pages is usually more than three. Large number of tags can be possible in DOM trees.
The BFS approach used in the beginning is feasible as it helps structure unstructured and semi-structured SSR pages which in turn further simplifies the extraction process. In this paper, new method of web document description based on image analysis has been proposed. we proposed hybrid segmentation algorithm to SIFT feature extraction process. This map should help with more accurately description of non-textual information. Thus all the methods in this paper are successful in retrieving the information with the help of DOM tree properties.

REFERENCES

[1] Jer Lang Hong, Deep Web Data Extraction, Systems Man and Cybernetics (SMC), 2010 IEEE International Conference, pp.3420-3427.

[2] Liu, B. and Zhai, Y., NET – A System for Extracting Web Data from Flat and Nested Data Records, WISE 2005, pp. 487-495.

[3] Jer Lang Hong, Eu-Gene Siew, and Simon Egerton, ViWER- Data Extraction for Search Engine Results Pages using Visual Cue and DOM Tree, Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference, March 2010, pp. 167-172.

[4] Jürgen Dorn and Tabbasum Naz, Structuring Meta-search Research by Design Patterns, .World Wide Web Conference, Canada, May2007.

[5] Wanjing Zhang, Wrapper Mill: A Tool for Generating and Managing Wrappers for Search Engines, 2007

[6] B. Shahnaz, “Advanced E-access Content Filimage System: Synthesis

File cards through Automatic Images-Captions Web-Pages Extraction,” Innovations in Information Technology, 2006 , vol., no., pp.1-4, Nov. 2006.

[7] L. P. Florence, “Image and Text Mining Based on Contextual Exploration from Multiple Points of View,” Twenty-Fourth International FLAIRS Conference, 2011, Palm Beach, Florida, 18-20 May.

[8] Sh. Behnami, Filimage System: Web’s Images and Texts Automatic Extraction, The World Scientific and Engineering Academic Society (WSEAS), Izmir, Turkey, 2004.

[9] R. HE, Y. ZHU, “A hybrid image segmentation approach based on Mean Shift and fuzzy C – Means,” Asia – Pacific Conference on Information Processing, [Online], 2009.

[10] S. GUAN, R. KLETTE, “Belief Propagation on edge image for stereo analysis of image sequences,” In Proceedings Robot Vision, LNCS 4931, 2006, p. 291 - 302.

[11] L. SIQIANG, L. WEI, “Image segmentation based on the Mean-Shift in the HSV space,” 2007, 26th Chinese Control Conference, [Online], p. 476-479.

[12] Ali Mesbah, Engin Bozdag, and Arie van Deursen, "Crawling AJAX by Inferring User Interface State Changes," Proc. Eighth International Conference on Web Engineering, 2008, pp. 122-134.

[13] Nick Matthijssen, and Andy Zaidman, "FireDetective: Understanding Ajax Client/Server Interactions," Proc. ICSE'11. Waikiki, Honolulu, HI, USA, 2011, pp. 998-1000.

[14] Cristian Duda, Gianni Frey, Donald Kossmann, Reto Matter, and Chong Zhou, "AJAX Crawl: Making AJAX Applications Searchable," Proc. IEEE International Conference on Data Engineering, 2009, pp. 78-89.

[15]Document Object Model. [Online]. Available: http://en.wikipedia.org wiki/Document_ Object_Model M. Young, The Technical Writer's Handbook.