Web Document Segmentation for Better Extraction of Information: A Review

Hassan F. Eldirdiery
Faculty of Computer Science and Information Technology
University of Science and Technology, Omdurman, Sudan

A. H. Ahmed
Faculty of Computer Science and Information Technology
Al-Neelain University, Khartoum, Sudan

ABSTRACT
This paper reviews the problem of web page segmentation. According to the recent studies, there exist different approaches used to segment the web page into multiple blocks. Segmentation of web document is an essential step for many applications, such as text classifications, clustering, extraction of information and searching. The study provided full description for each approach and showed its contribution to the work area of research. Also the paper discusses the variance between these approaches, explaining the benefits and limitations of each one. In addition to that it explores most of the effective algorithms those based on these approaches and explains the application area of each algorithm.

Keywords
Web page segmentation, DOM tree, Information Extraction.

1. INTRODUCTION
The content of the internet, which composed of web documents, is full of valuable information for users demand and many other applications, like information retrieval, web page clustering, classification, mobile web content adaptation …etc. Web pages designed typically for visual interactions. So every page visually composed of many segments. But the process of identifying the distinct parts of the web page is becoming too hard for some applications those get benefits out of the content of web pages. Since the web page content can be rendered by the browser into many sections, each is for different purpose. For example the web page may contains a section that represents the main article, a section that contains an advertisement banner, a section for navigation menus, links and abstract of related web pages …etc. However, the underlying source code of each page is not coded in such a way to differentiate between those segments.

There are several methods used to segment a web document into fragments. These methods can discriminate informative from non-informative content on a webpage; they can also identify the different types and classifications of information. The ability of distinguishing between different types of information assumed very useful in web ranking and web data mining applications. Consider as instance a multiword query whose terms match across different segments in a page; this information can clearly be useful in adjusting the relevance of the page to the query. Also appropriately labeling the segments for example into informative and non-informative, can improve the precision of web mining tasks like duplicate detection. Identification of segments also plays a vital role in displaying web pages on screen-space constrained devices such as smart phones and PDAs.

The web page segmentation problem has been addressed from many different perspectives. Some researches addressed the problem by analyzing the DOM (Document Object Model) structure of the HTML page, either by rendering and visual layout analysis or by interpreting or learning the meaning and importance of tag structures in some way. However, the number of possible DOM layout patterns is virtually infinite, which inescapably leads to errors when moving from training data to the real web data. In another perspective the problem addressed through segmenting the web page visually depending on the rendering of the web page using a browser. The main idea of segmentation makes full use of the layout feature of the web page. In some other approaches, the segmentation of the web document studied through investigating the actual retrieved element of document – the text, where the process of segmentation based on the low level properties of text. These approaches addressed the concept of block density, which is defined as the number of words within a portion of web document. Recently appeared some other approaches used to emerge more approaches together in order enhance the process of segmentation.

In this paper we review the methods and approaches that used to segment the web page into multiple parts or sections. Our main objective is to enrich and provide full literature about this area for other researchers.

The paper is organized as follows: In section 2 produces the Web page segmentation approaches. Section 3 presents the DOM-based segmentation approach. Section 4 covers the Vision-based segmentation approach. Section 5 discusses the Text-based approach. Section 6 explains the hybrid approach and section 7 concludes the work.

2. WEB PAGE SEGMENTATION APPROACHES
The web document could be segmented into regions or blocks using various different methods. The methods of segmentation could be classified according to the following approaches:

- DOM-based approach
- Vision-based approach
- Text-based approach
- Hybrid approach.

3. DOM-BASED APPROACH
The Document Object Model (DOM) is an application programming interface (API) [1] for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated.

In the DOM-based segmentation approach, an HTML

[1] www.W3C.org

document is represented as DOM tree, which provides a useful structure for a web page, but often not accurate enough to identify different semantic blocks in a web page.

Bar-Yossef et al. [1] used a technique based on segmentation of the web page into pagelets. A pagelet is a self-contained logical region within a page that has a well defined topic or functionality. The web document can be decomposed into one or more pagelets, corresponding to different topics or functions those appear in the web document. For example a web document of a portal can be partitioned into multiple pagelets; search window pagelet on the top side, copyright pagelet on the bottom side, the navigation bar pagelet on the top of search window and so forth. The main assumption of their approach is that pagelets are the more appropriate unit for information retrieval.

To segment the web page, the technique defines HTML elements (the "tags") as the pagelets. Any element in a parse tree of a page is a pagelet when none of its children contains at least k hyperlinks; and none of its ancestor elements is a pagelet. When specific element contains k links, it is likely represent an independent idea or topic; otherwise it is topically integrated to its parent. In their implementation, the value of k is 3.

The main purpose behind the segmentation of the web page is to detect template among web site. This method of segmentation relies on the tags of HTML.

However, the drawback from their method that is the tag almost contains attributes that represents the visual layout of the content. Also the second problem is how to determine the value of the parameter k which distinguishes the pagelet element from none pagelet element.

Chakrabarti et al. [2] formulated a method that makes use of weighted graphs. Where the nodes of the graph are the nodes of DOM tree and the weight of the edge simply identify whether to put the end point in the same segment or in different one.

Their method developed based on two formulations: correlation clustering and energy-minimizing graph cuts. The edge weights basically determine the quality of segmentation. The quality of segmentations depends on edges weights. They build a learning model for edge weighting from manually labeled data.

In their method the weights capture if two nodes in the DOM tree should be placed together or apart in the segmentation. The decision of placing two nodes together in the same segment depends on some features of the nodes (i.e. visual features like background color font size and content features like average size of sentences). They proposed an objective function that assigns to each node a segment label. To estimate the weight of each feature, their method uses a machine learning tools. The particular method used for learning depends on the objective function. The machine learning uses a set of manually labeled web pages. Where, each DOM node is assigned a segment ID.

However, considering some features to determine the weight of the node is one of the drawbacks of this approach. Also the second drawback is using heuristic rules which take a significant amount of time to execute the algorithm.

Yi et al. [3] proposed an approach aims to find the noisy portions in a given web site. Their method used a tree similar to a DOM tree. It based on stile tree structure. To construct the desired tree, the method uses the information of visual layout and the actual contents of pages within a site. Then it scans set of pages and every distinct node in a page will be added to the structure of the tree. However, the process of building the tree requires scanning more than 400 web pages.

Debnath et al. [4] proposed an approach that partition the web page into small coherent blocks based on their contents. In their method the block defined as a portion of web page enclosed within open-tag and it is matching close-tag, where open and close tag belong to an ordered list of tags that includes tags like <HR>, <TR>, <P> and <UL>.

The main objective of their method is to define informative blocks and none informative blocks in the web page. They extract the informative blocks from the rest of the blocks depending on specific features of these blocks. To partition a web page into blocks, they rely on some heuristic rules and basis of HTML tables. However, the content of modern pages can be found in other different structures of HTML.

Wang et al. [5] proposed a method that detect template on web page through using segmentation of the web document. The segmentation process contains two steps: 1) A web page is divided into multiple blocks. They choose some html tags that usually determine the page layout as separators, these html tags are <TABLE>, <DIV>, etc. 2) Then each block is further divided into text segments by html tags, process instructions, and html comments. However, the modern web pages can be developed with different structures.

Kolcz and Yih [6] proposed a method that segments the web document into blocks based on element nodes of DOM tree. In their method, the element node of DOM consider as a valid block if it satisfies two requirements: 1) the node should corresponds to specific predefined HTML tags such as {div, td, tr, table, etc.}. 2) It should have sufficient text content (i.e. the length of normalized text content is at least 40 characters and at least 3 unique words). The main objective of web document segmentation was to detect templates in web pages.

Xiao. et al. [7] proposed an algorithm that aims to identify two types of blocks: links blocks and content blocks. Their work developed in order to adapt web page browsing in mobile cell devices. But their algorithm focuses only on the TABLE tags.

Ahmadi. et al. [8] suggested a simple model for the web page, which consists of top, menus, main content and button part. Their method used heuristic rules to identify the blocks of the main content. The process starts by grouping the elements of HTML into four groups: structure, formatting, header and separator, then it uses this set to advise the heuristic rules in order to identify the blocks of the main content.

Vineel [9] developed unsupervised algorithm which uses DOM tree with a mining approach. They used content size and entropy of the nodes to detect repetitive patterns. The content size defined as the textual content of the node, while the entropy measures the local patterns of the node.

Kang et al. [10] proposed an approach works to recognize the repetitive tag patterns in the DOM tree structure. Their objective is to correct the process of segmentation to better fit the devices with small screens.

Rajkumar et al. [11] presented a new method that segments web pages based on either reappearance based scheme, by recognizing reappearance tag patterns from the DOM tree structure of a web page. Based on the detection of tag patterns, it generates implicit nodes to segment the nested block. If it contains reappearance tag in tag pattern means, it

will segment based on reappearance based segmentation. Otherwise it will segment based on web layout information. From that segmented block hyperlink is displayed on the mobile first and then user select hyperlinks based on his area of interest. The interested information alone is displayed to the user.

Alcic et al. [12] investigate the problem of segmentation from a clustering point of view by using distance measures for content units based on their DOM, geometric and semantic properties.

4. VISION-BASED APPROACH

Visual approaches segment the web page from the browser-side perspective as it is rendered. They used to partition the page into separators, such as lines, whitespace and images, and content and build a content-structure out of this information. They take into account visual features such as background color, styling, layout, font size and type and location on the page. But in order to render the page we need access to a browser engine, which complicates the implementation of an algorithm. And also it requires external resources such as CSS files and images in order to work correctly.

Cai et al. [13] proposed an algorithm based on the visual layout of the web page. Their method named as vision based page segmentation algorithm VIPS. VIPS aims to extract the semantic structure of a web page based on its visual presentation. Such semantic structure is a tree structure; each node in the tree corresponds to a block. Each node will be assigned a value (Degree of Coherence) to indicate how coherent of the content in the block based on visual perception, the bigger is the DoC value, the more coherent is the block.. The VIPS algorithm makes full use of page layout structure.

The segmentation process in VIPS has three steps: block extraction, separator detection and content structure construction. The method of extraction visual blocks from DOM tree basically depends on many heuristics rules and cues.

Kovacevic et al. [14] proposed another approach that based on the layout of a web page. In this approach a web page generally separated into 5 regions: top, down, left, right and center. They define a virtual screen (VS) that defines a coordinate system for specifying the positions of HTML objects inside Web pages. The VS is a rectangle with a predefined width and an infinite height both measured in pixels. The VS is set to correspond to the page display area in a maximized browser window on a standard monitor with resolution of 1024x768 pixels. Width of the VS is set to be 1000 because when vertical scroll bars from browser are removed, that quantity is usually left for rendering the page. Obviously pages are of different length and so theoretically height can be infinite. Top let comer of the VS represents the origin of the VS coordinate system.

They used a parser to parse HTML file of the web page, to extract two types of data elements – tags and data. They used these pair of data with a predefined heuristic rules to build a tree that represents the HTML structure of the web page.

Song et al. [15] proposed a learning method that automatically assigns importance weights to hierarchically arranged segments in web pages, termed as blocks. The method requires blocks in sample web pages to be labeled by users based on their judgment of the importance of each block. To partition the web page, they used Vision-based Page Segmentation algorithm (VIPS) according to the content coherence by analyzing the visual layout of the page.

Lei et al. [16] presented a method that based on VIPS. Their work used to utilize VIPS algorithm with a traditional method to overcome the shortage in DOM based methods in order extract the content of the Web page.

Burget et al. [17] proposed an algorithm that uses heuristics. Their algorithm has four steps: 1. detecting the visual blocks on the web page, 2. guessing the purposes of the detected blocks, 3. text line detection to join the same areas in the same line and 4. Block detection to detect the larger areas with the same visual style of blocks.

Xiao et al. [18] presented a method that uses the VIPS algorithm to identify the blocks in a web page. They provide the user with a display of the web page, then the user click the desired region to retrieve the relevant sub-page.

Yan et al. [19] proposed a multi-cue algorithm which simulates the process of user-perception. This method uses different types of information, like visual information (background color, font size), some non-visual information (tags), text information and link information.

A. Zhang et al. [20] presented a method which segments the web page based on semantic block headers detection using visual and structural features of the pages. their work aims to enhance browsing of web pages through mobile devices.

Saad. et al. [21] uses VIPs algorithm to identify if the change in the page is important for archiving or not. Their work aims to enhance the efficiency of web page archiving. In order to achieve this target, the developed method detects only the important changes between the versions of pages.

Akpinar and Yesilada[22] improved the older version VIPS algorithm. The authors focus on improving the first phase of VIPS[7], the visual block extraction. Their method divides the HTML tags into nine classes rather than three in the older version. Then they define new separation rules for those classes based on visual cues and tag properties of the nodes.

5. TEXT-BASED APPROACH

In text-based, the algorithms looks only to the textual contents of the page. They analyze certain textual features like text-density or link-density of some parts of the page. These approaches rely on quantitative linguistics basis, which declare that, statistically, text portion with similar features are likely to belong together [23]. However the optimal similarity threshold depends on the wanted granularity and needs to be determined experimentally.

Kohlschutter et al. [24] proposed an approach that based on methods from Quantitative Linguistics field. This approach is called Densitometric approach. They utilize the notion of text-density as a measure to identify the individual text segments of a web page. They developed a model that used the low-level properties of text. The number of words within a portion of a text consider as a good feature to segment a document. The field of Quantitative Linguistics full of worthy statistically measures to identify structural patterns in plain text documents, in particular for identifying subtopics [23].

Text density is a measure for the number of words within a particular 2-dimentional area [24]. Then the block b_x density $p(b_x)$ could then be formulated as follows:

$$p(b_x) = \frac{Number\ of\ tokens\ in\ b_x}{Number\ of\ lines\ in\ b_x}$$

This definition of text density relies on simple special property. In order to compute the density of a text you do not need to perform much effort in grammatical or lexical analysis to the text, you need only to get the number of words within the text.

Ruijie et al. [25] proposed a method that analyzes semantically the content of the web page in order to extract the textual contents. They used string match method to segment the retrieved words. To detect similarity between pages, they used vector space model with TF-IDF method for computation of features weights.

6. HYBRID APPROACH

In this approach the developed algorithms solve the problem of web page segmentation taking into account the limitations that derived from using one approach. The methods which follow this approach emerge multi approaches together in order to get benefits from most of the information which provided from each approach. This process results in a significance improvement in the problem of web page segmentation.

Kreuzer. et al.[26] proposed an algorithm which combines a plain structural approach with a rendering-based approach. They make use of DOM tree with addition to the information of visibility and dimensions of each of the tree.

Zhang and Deng [27] established a block tree model by combining DOM tree and visual characteristics of web content and a statistical learning method using neural networks. The work aims to remove noisy information out of web pages in order to enhance the applications related to them-based pages.

Wang1. and Liu1.[28] presented an algorithm, which segments the web document using both characteristics of the web page; the structure and the text attribute. Their method built based on generalized hidden Markov model. The structure of the web page provides information like, the color of the links, font size of headline, background color of text, while the text attribute provides words, which can be used to divide the different content of text.

Safi. et al.[29] developed a method which emerge the vision-based approach, DOM-based approach and Graph-based approach. Then they used a cluster algorithm to group the closest blocks into one zone (block). The purpose of the proposed is to transform the semantic of symbols in these zones, or blocks, or HTML elements into vibrations with different frequencies and amplitudes. The proposed method serves and helps Visually Impaired People by using Vibro-Tactile Access on Touch-Screen Devices.

Sanoja. et al.[30] built a web page segmentation framework which combines DOM tree and vision-based approaches. They implement a modified version of VIPS in order to enhance the precision of the extracted visual block from a human-side perception.

7. CONCLUSIONS

In this paper we studied the problem of web document segmentation. The paper discussed the various approaches that concern with the problem of web page segmentation. The review explores four approaches and explained the differences between them. Also it presents most of the algorithms that based on each approach and the application area related to each algorithm. The results of this review showed the importance of web page segmentation as an essential step to various applications that related to it. Also emerging different approaches of segmentation together properly will increase the accuracy of segmentation, since there will be best utilization of the different properties of each approach.

8. REFERENCES

[1] Z. Bar-Yossef and S. Rajagopalan. 2002. Template detection via data mining and its applications. In proceedings of the International Conference on the World Wide Web. ACM Press, pp. 580-59.

[2] D. Chakrabarti, R. Kumar, K. Punera. 2008. A Graph-Theoretic Approach to Webpage Segmentation. In Proceeding of the 17th international conference on World Wide Web.ACM Press, pp. 377-386.

[3] Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating Noisy Information in Web Pages for Data Mining, SIGKDD, ACM Press, pp. 296-305.

[4] S. Debnath, P. Mitra, and C.L. Giles. 2005. Automatic Extraction of Informative Blocks from Web pages. In ACM Symposium on Applied Computing. ACM, pp. 1722-1726.

[5] Yu Wang, Bingxing Fang, Xueqi Cheng, Li Guo, Hongbo Xu. 2008. Incremental Web Page Template Detection. WWW. ACM, pp. 1247-1248.

[6] Aleksander Kolcz and Wen-tau Yih. 2007. Site-Independent Template-Block Detection. PKDD. Springer , pp. 152–163.

[7] Yunpeng Xiao, Yang Tao, and Qian Li. 2008. Web page adaptation for mobile device. In Proceeding of the14th conference on Wireless Communications, Networking and Mobile Computing. IEEE, pp. 1-5.

[8] Hamed Ahmadi and Jun Kong. 2008. Efficient web browsing on small screens. In Proceedings of the working conference on Advanced visual interfaces. ACM, pp. 23–30.

[9] G. Vineel. 2009. Web page dom node characterization and its application to page segmentation. In Proceedings of the 3rd IEEE international conference on Internet multimedia services architecture and applications. IMSAA'09, NJ, USA, pp. 1-6.

[10] J. Kang, J. Yang, and J. Choi. 2010. Repetition-based web page segmentation by detecting tag patterns for small-screen devices. IEEE Transactions on Consumer Electronics. IEEE, pp.980–986.

[11] K. Rajkumar and V. Kalaivani, 2012. Dynamic web page segmentation based on detecting reappearance and layout of tag patterns for small screen devices, 2012 International Conference on Recent Trends In Information Technology. IEEE, pp. 508-513.

[12] S. Alcic and S. Conrad. 2011. Page segmentation by web content clustering. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics. WIMS '11, New York, NY, USA, pp. 1-24.

[13] D. Cai, S. Yu, J Wen, W. Ma. 2003. VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report (MSR-TR-2003-79).

[14] M. Kovacevic, M. Diligenti, M. Gori, V. 2002. Milutinovic. Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification. In Proceedings of the 2002 IEEE

International Conference on Data Mining, ICDM, pp. 250-257.

[15] R. Song, H. Lui, J.-R. Wen, and W.-Y. Ma. 2004. Learning block importance models for web pages. In proceedings of the International Conference on the World Wide Web. ACM Press, pp. 203-211.

[16] Fu Lei, Meng Yao, Yu Hao. 2009. Improve the Performance of the Webpage Content Extraction using Webpage Segmentation Algorithm. In proceedings of International Forum on Computer Science-Technology and Applications. Chongqing, China, pp. 323-325.

[17] Radek Burget and Ivana Rudolfova. 2009. Web page element classification based visual features. In 2009 First Asian conference on Intelligent Information and Database Systems. IEEE, pp. 67–72.

[18] Xiangye Xiao, Qiong Luo, Dan Hong, and Hongbo Fu. 2005. Slicing*-tree based web page transformation for small displays. In Proceedings of the 14th ACM international conference on Information and knowledge management. CIKM '05, New York, NY, USA, ACM, pp. 303-304.

[19] H. Yan and M. Miao. 2009. Research and implementation on multi-cues based page segmentation algorithm. International Conference on Computational Intelligence and Software Engineering, 2009. CiSE 2009, pp. 1-4.

[20] Zhang, J. Jing, L. Kang, and L. Zhang. 2010. Precise web page segmentation based on semantic block headers detection. IEEE, pp. 63–68.

[21] Myriam Ben Saad and Stephane Ganc¸arski. 2010. Using visual pages analysis for optimizing web archiving. In Proceedings of the 2010 EDBT/ICDT Workshops, , New York, NY, USA, ACM, pp. 1-43.

[22] E. Akpinar and Y. Yesilada. 2012. Vision based page segmentation: Extended and improved algorithm. eMINE Technical Report Deliverable 2 (D2), Middle East Technical University, Ankara, Turkey.

[23] M. A Hearst. Multi-paragraph segmentation of expository text. 1994. In proceedings of the 32nd annual meeting on Association for Computational Linguistics. Morristown, NJ, USA, pp. 9 -16.

[24] C. Kohlschutter, W. Nejdl. 2008. A Densitometric Approach to Web Page Segmentation. In Proceeding of the 17th ACM conference on Information and knowledge management. ACM Press, pp. 1173-1182.

[25] LI Ruijie, YANG Weidong and JIANG Haowei. 2010. Based on semantic web similarity. IEEE.

[26] Robert Kreuzer, Jurriaan Hage and Ad Feelders. 2014. A Quantitative Comparison of Semantic Web Page Segmentation Approaches. Technical Report, Utrecht University, Utrecht, The Netherlands.

[27] Y. Zhang and K. Deng. 2010. Algorithm of web page purification based on improved DOM and statistical learning. 2010 International Conference on Computer Design and Applications (ICCDA).

[28] Jing Wang1 and Zhijing Liu. 2009. A Novel Method for the Web page Segmentation And Identification. In Proceedings of the 2009 International Conference on Computer Engineering and Technology. IEEE.

[29] Waseem SAFI, Fabrice Maurel, Jean-Marc Routoure, Pierre Beust and Gaël Dias. 2014. A Hybrid Segmentation of Web Pages for Vibro-Tactile Access on Touch-Screen Devices. In Proceedings of the 25th International Conference on Computational Linguistics. Dublin, Ireland, pp. 95-102.

[30] Andres Sanoja and Stephane Gancarski. 2014. Block-o-Matic: A web page segmentation framework. In Proceedings of 2014 International Conference on Multimedia Computing and Systems, Marrakech. pp. 595-600.