Dynamic Web Page Segmentation on Heterogeneous Sites Using Tag Path Clustering

N.S.Arunkarthik[#1], Dr.S.Karthik[#2], K.R.Nivethitha[#3], S. Kannan[#4]

[#] *Department of Computer Science and Engineering, SNS College of Technology, Coimbatore, India*

[1]n.s.arunkarthik@gmail.com

[2]profskarthik@gmail.com

[3]niviknce@gmail.com

[4]6kannan6@gmail.com

***Abstract*— There are millions of websites in the World Wide Web that are actually designed for large screen devices. When these websites are viewed on small screen devices, they cannot be easily browsed with limited interface. By dynamically segmenting the web pages we can divide the contents of the site based on the web layout. It is possible to extract the web layout structure from the DOM pattern of the web page. We proposed a web page partition procedure by clustering of the source using their tags for noise pruning the contents of a web page. The tag based page partitioning procedure can help to improve the accuracy across heterogeneous sites by comparing the occurrence of similar objects reappearance key pattern with that of individual segment visual signals. By noise pruning, the segmented data can be displayed with hyperlinks on small screen devices that can render the user preferable information.**

Keywords- DOM (Document object Model), Reappearance key Patterns, Visual signals, Noise pruning.

I. INTRODUCTION

In the recent era of electronic gadgets, the use of smart phones has become very popular. With the aid of these portable hand-held computers, it is possible to access the internet very easily for various purposes. The websites in the World Wide Web are actually coded for the large screen devices. So when these websites are displayed on the mobile devices they cannot fit in perfectly. In order to have a better access to the websites through smart phones the users are required to pinch zoom or drag along the web pages. This persists in even higher end mobile devices and gives an annoying user experience. By segmenting the web page into logical blocks we shall categorize the web data and prune out all the noisy information. Fig.1,shows an example of page segmentation where the segmented blocks are marked with boxes.

This method can enhance the technique of information extraction and prune all the noisy and unwanted blocks and focus only on the informative content of the web page and facilitates the improved display of useful information on the small screen mobile devices. Most of the websites do not maintain a standardized approach but develop web pages.

Figure 1.Example of Web page segmentation into logical blocks.

In spite of many successful approaches, most of the vision based techniques highly rely on heuristic rules that make it more complex to cope with a dynamic web layout where the information of visual data is periodically getting changed. To resolve these issues, this papers proposes a novel technique of segmenting the web pages by analyzing the tag pattern in the DOM tree structure of the web page. These HTML tag patterns are called as tag paths. Even if the web pages do not maintain a structural layout, we are still able to partition them with the use of these tag paths across heterogeneous sites.

II. Related Work

Guohua Hu and Qingshan Zhao [3] proposed a general idea to eliminate the noisy data from the web page by using Site Style Tree (SST). It was able to filter out the contents stored in multi-node by resolving the content extraction problem from semi structured websites. Although this approach can produce noise free data of a web page t here is still not a method that meets the requirements of information collection in all kinds of fields.

Vision-based Page Segmentation algorithm [2,6] to extract the semantic structure for a web page. Such semantic structure is a hierarchical structure in which each node will correspond to a block. It was very helpful for applications such as web adaptation, information retrieval and extraction but these cause the retrieval performance relatively low although better than the baseline. Peifeng Xiang [13] introduced an effective automatic segmentation method which combining pattern analysis and visual separators. The basic idea is that a page's semantic structure is largely reflected by repeated continuous patterns and visual separators that coincide with human's visual perception. Debnath [7] formally defined Web page blocks and devise a new algorithm to partition an HTML page into constituent Web page blocks. They proposed four new algorithms, Content Extractor, Feature Extractor, K-Feature Extractor, and L-Extractor.

[4] Robust web page segmentation method calculates "content-distances," a measure which expresses the strength of connections between content elements of the Web page based on the structural depth of HTML tags, and segments Web pages at the positions where the calculated content-distance exceeds a predefined threshold. Since this method does not explicitly utilize HTML grammar for analysis of the Web page, it is robust to frequently observed HTML grammar errors, thus, is capable of segmenting any Web page. Although segmentation accuracy is high, it was incapable to handle separate Web page components. Lim [14] assumed that the <TABLE> tag is widely used to make the structure of a Web page, and proposed a method primarily using the <TABLE> tag to extract blocks from a Web page [10, 8]. However, this method cannot be applied to those Web pages without the <TABLE> tags. The web page tag analysis method predefines the content tags that has more relevant information and finds the block contents by measuring the distance between the tags[1],[2],[10]. The template-based method builds a template with extracting rules based on regular expressions [11, 12].

Yonghyun Hwang [5] proposed two general solutions to this disparity exist, manual and automatic reauthoring. In manual reauthoring, Web authors prepare multiple versions of a Web page targeted to resource profiles of various platforms, including the Wireless Application Platform. Although this approach can produce high-quality pages for specific devices, it assumes a Web author will both be available to reauthor the pages.

Yossef, [15] proposed a method to identify frequent templates of web pages and pagelets to perform data cleaning in hypertext corpora. Lin. [16] developed a method which partitions a page into several content blocks and according to the entropy value, a method is proposed to dynamically select the entropy threshold that partitions blocks into either informative or redundant. Informative content blocks are distinguished parts of the page, whereas redundant content blocks are common parts.

The featured DOM [9] tree approach detects and eliminates the local noised in a web page. The work did not give the efficient method to find the contents of the main block. In order to solve that problem, Tag Path algorithm is proposed.

III. Proposed Method

The proposed method of detecting the tag path of the web page identifies the key patterns. The data records that constitute a Web page are typically represented using an HTML code template. Thus, they often have a similar appearance and are visually aligned. The Tag Path Clustering (TPC) procedure follows four phases after the process of DOM tree formation. The HTML tag path is extracted from the source and the DOM tree is constructed. In the first phase of TPC the visual signals from the web page are extracted. In the second phase the reappearance between the tags are found by the similarity measurement. The third phase is the segmentation, in which the web page contents are segmented and from which the information are extracted for display by pruning all the noisy data. In the fourth phase the extracted useful information are displayed along with the hyperlinks on to the small screen devices. The users allowed to choose the necessary information from them based on their interest.

A. Extraction of web elements:

The web pages is taken from any website and stored in a folder. Then select a web page for segmentation using browse button. The content of the web page is displayed to the user. The tags are extracted from the web page. Less meaningful tags such as <a>, <b>, <script>, etc are removed.

B. DOM tree Construction:

DOM or Document Object Model is a tree structure representing a nesting of elements within the page and it is often used for traversing the web page. In DOM, an element is referred as a node. The steps involved in constructing DOM tree are as follows. A DOM tree is constructed, after removing less meaningful tags such as <a>,<b>,<script>,etc .

C. Extracting Visual Signals:

The visual signals that are rendered in the web pages are displayed in the form of layout and font customizations with the aid of HTML tags. A given hyperlink tag can have different appearances when it follows different paths in the DOM tree. For each tag occurrence, there is an HTML tag path, containing an ordered sequence of ancestor nodes in the DOM tree. A Web page can be viewed as a string of HTML tags, where only the opening position of each HTML tag is considered. Each HTML tag maps to an HTML tag path. An example is shown in Table 1.

TABLE 1: TAG PATHS FOR HTML TAGS

HTML Code	Pos	Tag Path
<html>	1	html
<body>	2	html/body
Nuclear	3	html/body/h1
<table>	4	html/body/table
<tr>	5	html/body/table/tr
<td> Fission</td>	6	html/body/table/tr/td
</tr>	NA	NA
<tr>	7	html/body/table/tr
<td> Fusion </td>	8	html/body/table/tr/td
</tr></table></body> </html>	NA	NA

Roughly speaking, each tag path defines a unique visual pattern. Our goal is to mine the visually repeating information in the Web page using this simplified representation. An inverted index characterizing the mappings from HTML tag paths to their locations in the HTML document can be built for each Web page, as shown in Table 2. Each indexed term in the inverted index, i.e., one of the unique tag paths, is defined to be a visual signal.

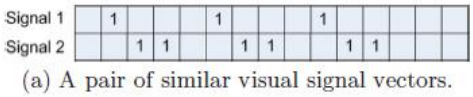
TABLE 2: EXTRACTING VISUAL SIGNALS

Unique Tag Path	Position	Visual Signal Vector
html	1	[1,0,0,0,0,0,0,0]
html/body	2	[0,1,0,0,0,0,0,0]
html/body/h1	3	[0,0,1,0,0,0,0,0]
html/body/table	4	[0,0,0,1,0,0,0,0]
html/body/table/tr	5,7	[0,0,0,0,1,0,1,0]
html/body/table/tr/td	6,8	[0,0,0,0,0,1,0,1]

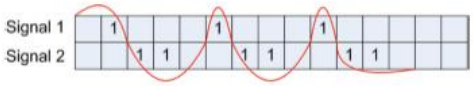
D. Similarity Measurement:

A similarity function will capture all the data that seem to show a likelihood behavior. For instance in a web page a table can have its HTML coding with contents placed in them at a regular interval of table syntax such as <TD>,</TD>,<TR>,</TR>, etc. A critical factor in finding the clustering performance is the choice of similarity function.

In our case, the similarity function captures how likely two visual signals belong to the same data region. Figure 2(a) shows a pair of visual signals that are highly likely to belong to the same data region. Their positions are close to each other, and they interleave with each other. Every occurrence of visual signal 1 is followed by two occurrences of visual signal 2. The segmented visual signal vectors are shown in Figure 2(b).



(a) A pair of similar visual signal vectors.



(b) Segmented visual signal vectors.

Figure 2: Example pair of visual signals that appear regularly.

E. Segmentation:

The visual signals are grouped and categorized from which we shall segment into useful data contents and noisy data contents as in Figure 3. A pattern matching algorithm is used for segmentation. Key patterns are used to

build implicit nodes by separating the subseries containing key patterns. The noisy data such as font tags, other customization tags, image tags are analyzed and pruned. The main block contents are now logically segmented and grouped vividly as noisy and useful data.

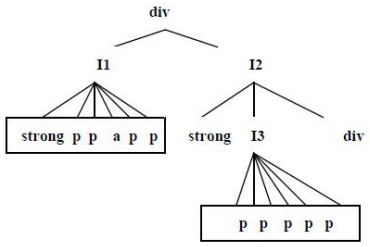


Figure 3: Generating Virtual nodes by using key patterns

F. Display hyperlink in mobile device

After segmentation process, the web page is divided into number of blocks. Informative blocks are determined by evaluating the amount of information in the block, which can be done by assigning an importance weight to each node considering the number of reappearance of node patterns in a Web page. For simple calculation, normalize the importance weight for each node by the maximum number of repetitions. From that informative segmented block hyperlink is created and displayed on the mobile device.

IV. Experimental Results

We experimentally tried the TPS concept for the dynamic web page segmentation by selecting the web pages which were collected from heterogeneous sites that were stored in a folder. Figure 4 shows the extraction of data from the stored location.

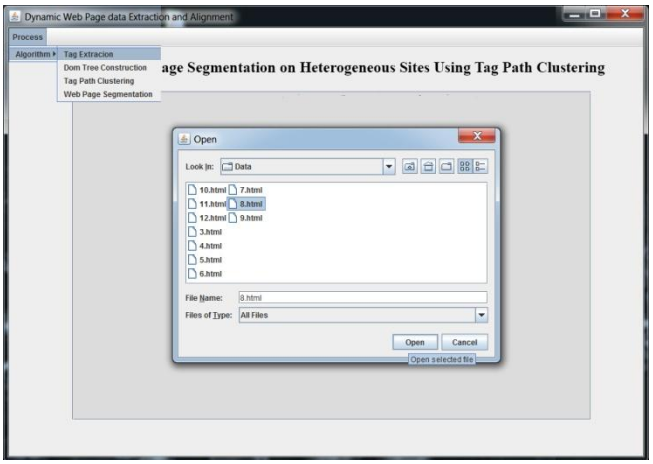


Figure 4: Extracting the web data source

The web page composes various layouts with predefined logical blocks as shown in Figure 5. Each logical block has their own content which is related to the user search and may also have certain noisy data such as external links and advertisements. Initially the entire web page is parsed and DOM structure of the web page is developed.

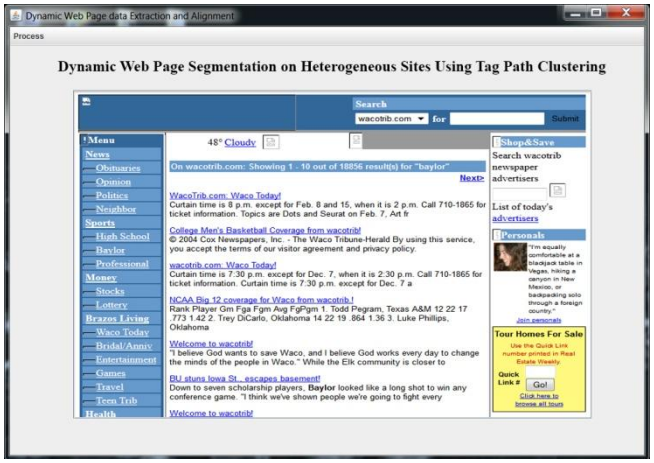


Figure 5: Display of the web page

The DOM tree is constructed for the web page according to the order of blocks in which they are nested in. The DOM structure shown in Figure 6 gives the hierarchical pattern in which the source code is organized and gives the entire vision of the web page.

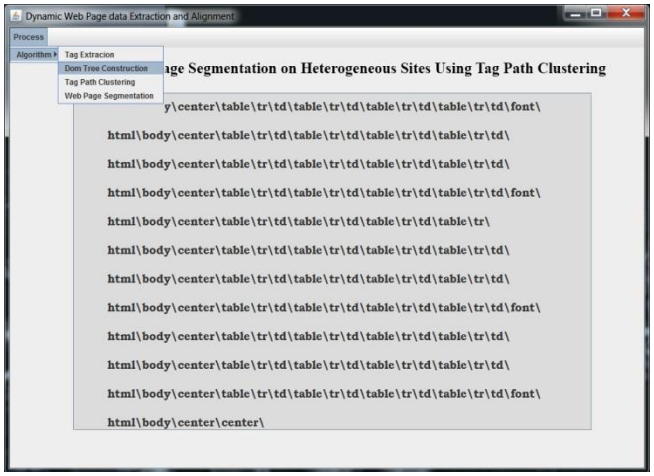


Figure 6: DOM Tree Construction

The Tag path clustering is then applied on the DOM to collect the visual signals and similarity measurements, based on which the web page is segmented depending of the useful and noisy data. The segmented block as shown in

Figure 7 is then sent along with the hyperlinks to display them on the mobile device.

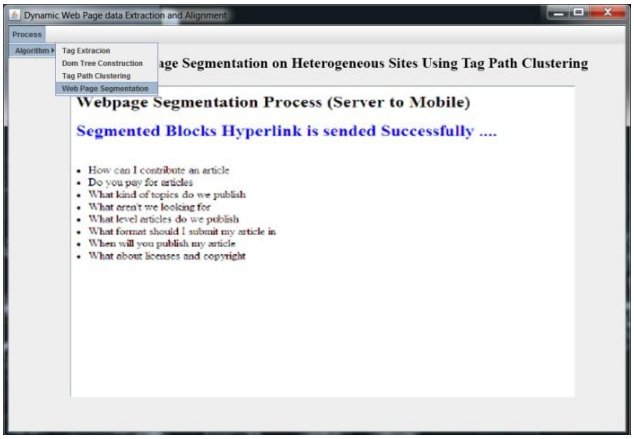


Figure 7: Sending the segmented data to the mobile device

The bandwidth is saved, when we display the segmented blocks to the mobile device. If a web page contains a total volume of 240KB and if the user wants to see this page on small screen devices, there may be some difficulties in finding their area of interest. When the user accesses the same page with our system, he receives only list of content blocks as shown in Figure 8(a). Getting such content blocks will results in lower KB of bandwidth such as 70. Suppose that user interested only in seeing first block alone in web page means, he selects that category from content blocks and it is alone displayed on mobile as shown in Figure 8(b), for each content block, there must be around 12.7 Kb.

Table 3 shows consumption of band width results for segmented web blocks in a web page.

TABLE 3: BANDWIDTH CONSUMPTION

Websites	With Interface	Without Interface	% Savings
wacotrib.com	650 KB	12.3 KB	98.10
metasearchengine.com	210 KB	65 KB	76
merges.net	40 KB	5 KB	90.17

The display of the small screen device can be much utilized by only displaying the contents that are related to the context and thus the noisy data are eliminated. This gives the user a much convenient interface so that they need not want to scroll along or navigate on the display of the small screen mobile devices. Also the visibility of the contents is getting improvised and gives the user a more options to select from the displayed contents.

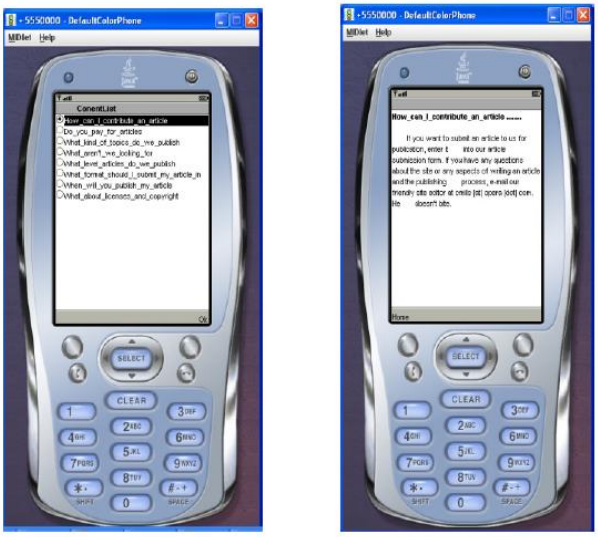


Figure 8 (a) Content items; Figure 8 (b) First content block of the site

Figure 8. Extracted block as accessed on sun java wireless tool kit

Table 4 shows the website that are used for the evaluation of the experiment.

TABLE 4: WEBSITES USED FOR EVALUVATION

S. No	Websites
1	Amazon
2	wacotrib.com
3	metasearchengine.com
4	merges.net
5	eBay

V. CONCLUSION

This paper proposed web page segmentation method for heterogeneous websites by tag path clustering method. Previous methods for Web page segmentation are not flexible in a dynamic Web environment because they largely relied on heuristic rules generated by exploiting structural tags and visual information inherent in a page. Heterogeneous tag path clustering method is suitable for dynamic web environment by recognizing reappearance tag patterns extracted from the DOM tree structure of a web page. Based on the recognitions of tag patterns, it generates

implicit nodes to segment the nested block correctly. From that segmented block hyperlink is displayed on the mobile device first and then user select hyperlinks based on his area of interest. The interested information alone is displayed to the user.

REFERENCES

[1] A. Arasu and H. Garcia-Molina, "Extracting structured data from web page," *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, pp.337–348, 2003.

[2] Cai,Shipeng Yu,"VIPS: a Vision-based Page Segmentation Algorithm", *IEEE Transactions on computer and information science*, vol. 25,no. 2,2008.

[3] Guohua Hu,"Study to Eliminating Noisy Information in Web Pages based on Data Mining" *IEEE Transactions on knowledge and Data Engineering* vol.18,no 9,2009.

[4] Hattori .G., Hoashi K. , Matsumoto K. , and Sugaya F., "Robust web page segmentation for mobile terminal using content distances and page layout information," *IEEE Transaction on World Wide Web*, vol. 53, no. 2, pp. 796-803, 2007.

[5] Yonghyun Hwang etal," Structure-Aware Web Transcoding for Mobile DeviceIEEE Transactions on Internet computing,2003.

[6] Pyungkwan Ko, Sanggil Kang,"Web Page Dependent Vision Based Segmentation for Web Sites", *IEEE Transactions on computer and information science*, Vol 25,No 2,2008.

[7] Sandip Debnath, Mitra. P, " Automatic Identification of Informative Sections of Web Pages", *IEEE Transactions on knowledge and Data Engineering* , vol. 17, no. 9, 2008.

[8] Riadh May .H, Akram M. Othman,"Web Content Adaptation system,*IJCA*,Vol.23,No.9, 2010.

[9] Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew, *"Eliminating Noisy Information in Web Pages using featured DOM tree,"* International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868, Volume 2– No.2, May 2012

[10] Xiang P.F, . "Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens", *IEEE Transactions on* Web Intelligence, vol. 23, no. 3, 2006.

[11] Zhigang Hua "Web Browsing on Small-Screen Devices: A Multiclient Collaborative Approach" IEEE Transactions on Pervasive computing, 2006.

[12] Yonghyun Hwang, Jihong Kim," A Block Gathering Based on Mobile Web Page Segmentation Algorithm", *IEEE Intl. Conf. on Trust,security,privacy in computing and communication*, Publication Year: 2011 , Page(s): 1425 – 1430.

[13] Peifeng Xiang ," Effective Page Segmentation Combining Pattern Analysis and Visual Separators for Browsing on Small Screens.", IEEE Transactions on web intelligence,2006.

[14] Lee .W, Kang .S, Lim .S, "Adaptive hierarchical surrogate for searching web with mobile devices," *IEEE Trans. Consumer Electron.*, vol. 53, no. 2, 2007.

[15] Ziv Bar-Yossef, Sridhar Rajagopalan,"Template Detection via Data Mining and its Applications", Proceedings of the 11th international conference on World Wide Web, pp 580-591, 2002.

[16] Shian-Hua Lin, Jan-Ming Ho, "Discovering informative content blocks from Web documents", Proceedings of ACM SIGKDD'02, July 2002.

