WEB PAGE SEGMENTATION USING COMBINED VIPS

KHAINGWAHWAH LINN

UCSY, Myanmar
E-mail: khaingwah2linn@gmail.com

Abstract- The World Wide Web (WWW) serves a huge, widely distributed global information service. Web page usually contains various contents, which are relevant or irrelevant to the main topic of the web page. Retrieving of useful or relevant information in mass information has become the focus of information extraction research.Web page segmentation is an important technology for web driven application such as search engine and web browser. But it has to overcome a great amount of irrelevant information. This paper proposed a simple web page segmentation method and main content extraction system by combining DOM and Vision of the web pages.

Keywords- Web Page Segmentation, Vision based page segmentation algorithm, DOM.

I. INTRODUCTION

Web Mining consists of: web usage mining, web structure mining and web content mining. Mining extraction and integration of useful data, information and knowledge from web page content is called web content mining. Web Content Mining uses the ideas and principles of data mining and knowledge discovery to print more specific data.

Today there are increasing amount of data available on the web and almost 1.5 million web pages are being added daily. Therefore, the organization of these pages can't be easily searched. Besides, classification for efficient and accurate web search is very essential. A web page usually contains the number of noise which is not related to the main information of this page such as navigation bar, advertisements, related articles and so on. These noises on the web pages tend to problem mining the main content of these pages. Web page partition into semantic parts becomes an important issue in the information retrieval on web. The World Wide Web is focused on the ability for people to collaborate and share information online. The Web document usually contains several types of data, such as text, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data.

Web pages also contain different contents, which are relevant or irrelevant to the key topic. Relevant contents defined as informative content blocks.

II. RELATED WORKS

The simplest way to clean Web pages is to remove metadata and tags the source data. A number of approaches have been reported in the literature for extracting information from web pages. One of the main issues on the web is that the information contained on web sites is often mixed with irrelevant content. We proposed a main content extraction system by combining DOM and Vision of the web page. They receive data and meta data of a single web page ad input and they first output a content structure, usually in the form of semantic tree.

III. THEORY BACKGROUND

Many researchers have described in comparison the number of web page segmentation method such as FixedPS, DOMPS, VIPS and so on. In FixedPS, fixed-length passages are used to overcome difficulty of length normalization. A fixed length passage contains fixed number of continuous words. The main shortcoming of the fixed-length is that no semantic information is taken into account in the segmentation process. In DOMPS, provide each web page with a fine-grained structure, which illustrates not only the content but also the presentation of the page. This method tend to partition pages based on their predefined syntactic structure, i.e., the HTML tags, DOM is a linear structure, so visually adjacent blocks may be far from each other in the structure and departed wrongly. Moreover, tags such as <TABLE> and <P> are used not only for content presentation but also for layout structuring. Furthermore, DOM prefers more on presentation to content. Therefore, it is not accurate enough to discriminate different semantic blocks in a web page.

Vision-based Page Segmentation algorithm (VIPS) extracts the blocks structure by using some visual cues and tag properties of the nodes. Unlike DOM based page segmentation, a visual block can contain DOM nodes from different branches in the DOM structure with different granularities. Structure tags such as <TABLE> and <P> can be divided appropriately with the help of visual information and wrong presentation of DOM structure can be reorganized to a proper form. Therefore, VIPS can achieve a better content structure for the original web page. Because ofthese benefits, we use VIPS algorithm in our system for web page segmentation. This algorithm makes full use of page layout features and tries to partition the page at the semantic level as shown in figure 1 and 2. In this step,

we output content structure as shown in figure 3. Each node in the extracted content structure will correspond to a block of coherent content in the original page.

Vision Based Segmentation Algorithm (VIPS) aims to extract the block structure by using some visual cues and tag properties of the nodes. Visual Block Extraction is the basic and the most important part of the algorithm. Therefore, this process should be handled in detail.

VIPS Algorithm aims to find visual blocks in a web page in three steps which are: (1) visual block extraction, (2) visual block separation and (3) content structure construction. The most important one of these three steps are visual block extraction, since the other two steps depend on the first step. If the construction of block structure of the web page is false, the separators and content structure construction is false too. Therefore, visual block separation needs more attention; it is the root of the algorithm.

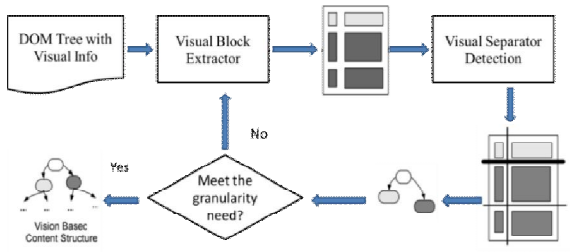

Fig. 1 VIPS algorithm

A. Visual block Extraction

Extracting visual blocks consists of a top-down tree. This step aim at finding all appropriate visual blocks contained in the current sub-page. Every node in the DOM tree can present a visual block. For each extracted node that represents a visual block, its DoC value is set according to its intra visual difference. This process is iterated until all appropriate nodes are found to represent the visual blocks in the current sub-page.

```
Visual Block Extraction
Algorithm DivideDomtree(pNode, nLevel)
{
IF (Dividable (pNode, nLevel)== TRUE)
{
   FOR EACH child OF pNode{
DivideDomtree (child, nLevel);
}
}ELSE {
        Put the sub-tree (pNode) into the pool as a
block;
}
}
```

The visual cues used in are tag, color, text and size cues. For color cue, background color is the main criteria. The size cues of the nodes are compared to a threshold, comparing the node size to the whole page or a part of the page. Note that, this threshold is undefined. The aim is to reach to a predefined pDoC value, by comparing it to the DoC value obtained in each turn in the algorithm. If it finds a DoC value greater than pDoC value, it stops executing the rules defined.

B. Visual Separator Detection

The second step is separator detection. Separator is defined as horizontal and vertical line or rather rectangular area which doesn't intersect any of previously detected blocks. The algorithm is initialized by a single separator covering the whole page. Separators are always detected for a particular level of visual block tree. Then for each block it performs detection of its relation to each existing separator:

If the block is fully covered by area of a separator, divide this separator into two
If the block partially intersects a separator, the intersection is eliminated.
If the block fully intersects with a separator (i.e. covers entire height of a horizontal separator or entire width of vertical separator), remove the separator entirely.
The algorithm is finished by removing separators at the edges of document.

After all blocks are extracted, they are put into a pool for visual separator detection.Separators are horizontal or vertical lines in a web page.

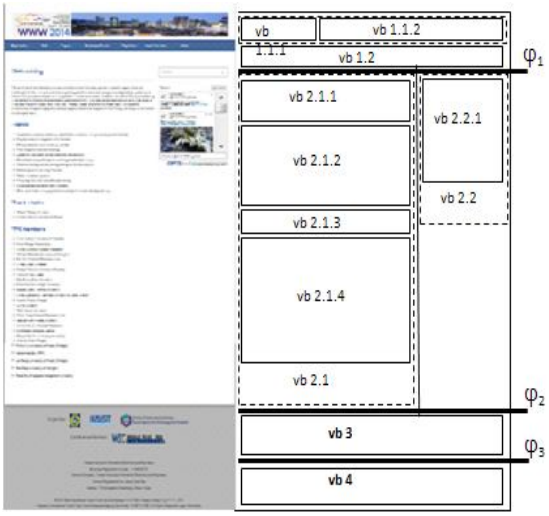

Fig. 2 Layout Structure and Vision based Content Structure of an example page

C. Content Structure Construction

The final step of VIPS is content structure construction. The construction process starts from the separators with the lowest weight and the blocks beside these separators are merged to form new virtual blocks. We have to check whether blocks meet granularity requirement. This process iterates till separators with maximum weights are met as shown in figure 3.

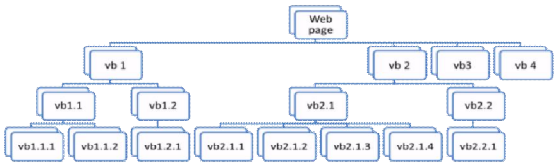

Fig. 3 Content Structure Construction of an example page

IV. SYSTEM ARCHITECTURE

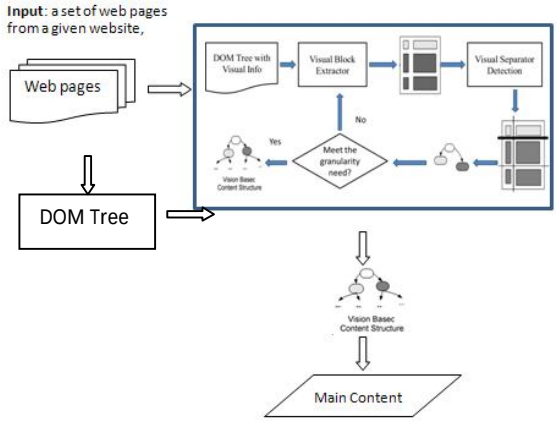

Fig. 4 Architecture of proposed system

There are describing proposed system as shown in figure 4. It takes a web page as an input of this system at first and this pageis segmented by DOM plus Vision based algorithm. Firstly, web page that enters to the proposed system is constructed DOM structure then vision based content structure by using visual cues of web pages. There are three main steps: visual block extraction, separator detection and content structure construction. When content structure is constructed, we have to check whether blocks meet granularity requirement.

The granularity requirement is represented by PDoC and the general rule for meeting the granularity requirement is that DoC>PDoC. If the block doesn't meet the requirement, it returns to step one with root node being that visual block. After this process, vision based content structure tree is retrieved.Finally, this system retrieves the main content of this web page by using proposed equation.

V. MAIN CONTENT EXTRACTION

In this system, main content of web pages are considered using following equation (1).

$$Weight_B_i = \frac{l_i \times w_i \times fs_i}{d(toB_i, toP) + d(coB_i, coP)} \quad (1)$$

If $Weight_B_i > Weight_B_{i+1}$

Then $MainContentWeight_B_i$ ←

Where

l_i= Length of block i

w_i= weight of block i

fs_i= font size of block i

$d(toB_i, toP)$= distance between top of block and top of page

$d(coB_i, coP)$= distance between center of block and center of page

Equation (1) shows the main content extraction of web page. The weight of the Block i containing in the web content blocks is evaluated and if the weight of the block i is greater than the weight of the next block i+1, then the weight of the block i is the main content of the block.

VI. SYSTEM ANALYSIS AND PERFORMANCE

Let us consider an example shown in figure 2. At the first round of block extraction (visual block) VB1, VB2, VB3 and VB4 will be extracted and put into the pool. Its DOM tree structure is shown in figure 5, which is a part of the whole web page. The experimental results show that our algorithm improves upon the other two with much higher precision and recall, and its running time is far lower than that two other.

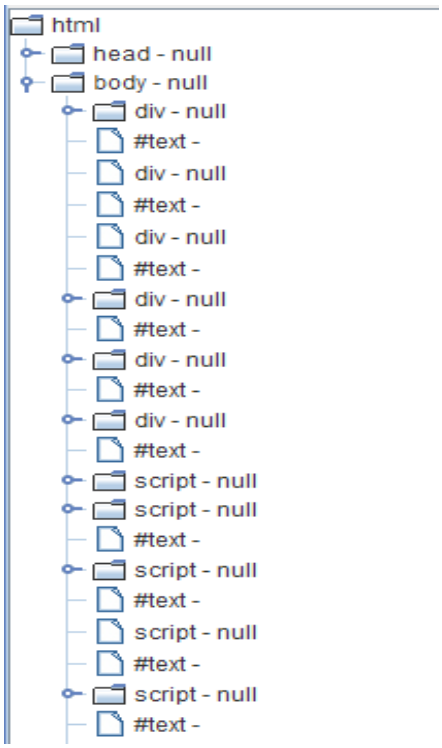

Fig. 5 DOM tree structure of web page

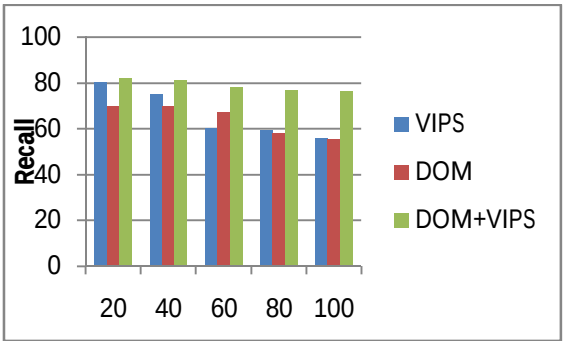

Fig. 6 Recall comparison

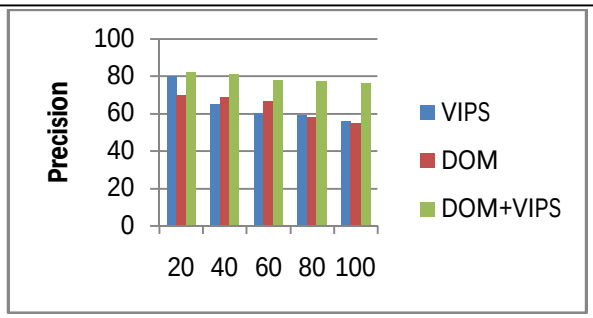

Fig. 6 Recall comparison

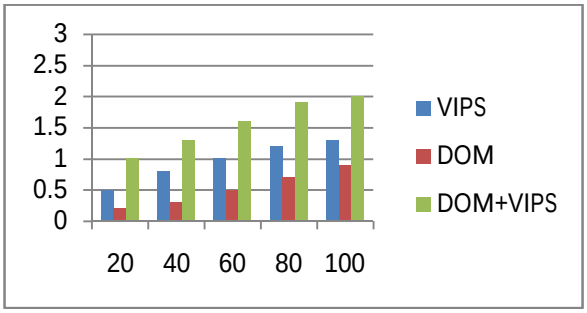

Fig. 8 running time comparison

CONCLUSION

Many researchers have been developed several approaches for extract main content from web pages. Most of the approaches based on the only DOM tree. A commercial web page typically contained many information blocks that is not related with the main content block. These noisy blocks can seriously drop web data accuracy. In this system, web pages are segmented by combining DOM tree with Vision based Page Segmentation algorithm. The overall goal of this research is to search main content of web page and to support the user query. For this particular system, we took web pages from many web sites, including www.informingscience.us/icarus/, www.wikipedia.org and so on.

REFERENCES

[1] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Vips: a vision based page segmentation algorithm. Technical Report MSR-TR-2003-97, Microsoft Research, 2003.

[2] Elgin Akpinar and YelizYesilada, "Vision Based Page Segmentation: Extended and Improved Algorithm", Middle East Technical University, Ankara, Turkey.

[3] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Extraction Content Structure for Web Pages based on Visual Representation", Microsoft Research Asia, China.

[4] SweSweNyein. "Mining Contents in Web Page Using Cosine Similarity", University of Computer Studies, Yangon, Myanmar.

[5] AmitChauhan, HimanshuUniyal, Dr. Bhasker Pant,, "Cleaning Web Pages for Relevant Text Extraction and Text Categorization".

[6] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Block-based web search" Microsoft Research Asia, China.

★★★