Retrieve Main Content using Vision-base Web Page Segmentation with Gomory-Hu Tree

Khaing Wah Wah Linn
UCSY, Myanmar

ABSTRACT
The world wide (www) serves a huge, widely distributed global information services. A huge amount of data have been accumulated and stored on the web. The information on Web is usually presented via Hypertext Markup Language (HTML) to make its perception easier for humans. Web pages usually contain various contents, which are relevant or irrelevant to the main topic. Irrelevant contents are called noise. A web page usually contains the number of noise which is not related to the main information of the page such as navigation bar, advertisements, and related articles and so on. Noise on the web pages tends to problem mining the main content of these pages. This paper is proposed wed page segmentation using Gomory-Hu tree based Vision-based Page Segmentation (VIPS) algorithm.

General Terms
Web Content Mining

Keywords
Web Page Segmentation, Vision-based Page Segmentation, Gomory-Hu tree

1. 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."[2]. Web Content Mining extracts or mines useful information or knowledge from web page content. 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. Web page usually contains several types of data such as text, image, audio, video, metadata and hyperlink. Some of them are semi-structured such as HTML document. Some are a more structured data like table but most of the data is unstructured test data. The unstructured characteristic of Web data forces the Web content mining towards a more complicated approach. Web pages also contain different contents, which are relevant or irrelevant to the key topic. Relevant contents are defined as informative content blocks, whereas the irrelevant contents are defined as clutters. Clutters tend to mislead search engines, or initiate an artificially high link-based ranking for specific target pages. Apart from the useful information on the Web, it usually has such information as navigation panels, copyright notices, banner ads, etc. which are called noisy information. Although these information items are useful for human viewers and necessary for the Web site owners, they can lead to poor results and also seriously harm automated information collection and Web data mining, e.g. Web page clustering, Web page classification, and information retrieval. It is also important to distinguish valuable information from noisy data within a single Web page. Several algorithms have been developed to do above works.

There are several applications of web page segmentation. (1) Ranking: Passage retrieval is a research topic with long history in the IR community which addresses the shortcomings of whole-document ranking. (2) Duplicate detection: Identical web content information is presented in different web page layouts. (3) Content Extraction: web page layout is segmented informative and non-informative content on a web page; removing noise might also increase classifier performance. The input of informative content extraction is the DOM tree of the Web page. The final output of the extraction is the informative data without the cluttering view.

Traditional approaches of web page segmentation usually use heuristic rules or machine learning algorithms either by visual analysis or interpreting the meaning of tag structures. While these approaches might work well on some kind of pages, they still have problems.

In the sense of human perception, it is always the case that people view a web page as different semantic objects rather than a single object. Some research efforts show that users always expect that certain functional part of a web page (e.g.

Navigational links, advertisement bar) appears at certain position of that page. Actually, when a web page is presented to the user, the *spatial and visual cues* can help the user to unconsciously divide the web page into several semantic parts. Therefore, it might be possible to automatically segment the web pages by using the spatial and visual cues.

2. RELATED WORK
Appropriate webpage structure mainly composed of navigation, hyperlink and the main text, which can help to expand search engine collection and speed up searching effectively. A number of approaches have been reported in the literature for extracting information from web pages.

3. BACKGROUND THEORY
3.1 Introduction to VIPS and Web Page Segmentation
Many researcher 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 tends 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. There are three main parts in Vision-based Page Segmentation algorithm (VIPS), such as Visual block extraction, Separator detection and Content structure construction.

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. Structural 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 of these 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. [1]

3.2 Gomory-Hu tree

A *Gomory-Hu tree* (also known as a *cut tree*) is an O(n)-space data structure which represents the pair wise edge connectivity of all pairs of vertices in an undirected graph. More precisely, it is a weighted tree T on V, with the property that the pairwise edge connectivity between any two vertices s and t in the graph equals the minimum weight of an edge on the unique s-t path in T . Further, the partition of the vertices produced by removing this edge from T is a minimum s-t cut in the graph, i.e. a cut of cardinality equal to the s-t edge connectivity. An undirected graph has at least one Gomory-Hu tree, but it might not be unique; on the other hand, examples by Bencz´ur[Ben95] show that Gomory-Hu trees need not exist for directed graphs. Gomory-Hu trees have many applications in multi-terminal network flows.

Theoretically Gomory-Hu algorithm is an optimal algorithm. However, since it partitions the graph following the minimum cut criteria, it tends to generates outliers, very small sub graph or singleton vertices.

Given an undirected, weighted graph *G. .V; E; c.* a cut-tree

T . .V; F;w. is a tree with edge-set *F* and capacities *w* that

fulfills the following properties.

1. Equivalent Flow Tree: For any pair of vertices *s; t* 2 *V,*

f .s; t. in *G* is equal to *fT .s; t.*

2. Cut Property: A minimum *s-t* cut in *T* is also a minimum

cut in *G*.

Here, *f .s; t.* is the value of a maximum *s-t* flow in *G*, and *fT .s; t.*is the corresponding value in *T*. [8]

3.3 Informative Content Extraction

Informative content extraction is the process of determining the parts of a web page which contain the main textual content of this document. A human user nearly naturally performs some kind of Informative Content Extraction when reading a web page by ignoring the parts with additional non-informative contents, such as navigation, functional and design elements or commercial banners. Therefore, it is difficult for a human to determine the main content of a document in automatic way. Some applications in the field of information Retrieval and Information Extraction, web mining and Text Summarization use Informative Content Extraction to process the raw data in order to improve accuracy.

4. PROPOSED SYSTEM

4.1 Cleaning the Web Page

Most web pages are not well-formed documents. They contain invalid tag structure such as there is an opening tag with no corresponding closing tag and vice versa. Some HTML tags are nested in wrong order and also some tags are mixed up. In order to construct the DOM tree of the input web page correctly, HTML file needs to be well-formed. Therefore these invalid tag structures are needed to be cleaned before processing them. The CleanHTML Method for cleaning web page to construct the proper DOM tree effectively is as follows. [11]

```
CleanHTML(HTMLpage)
    For each HTML tag in page
        Begin
            If the tag is missing or mismatched
                Then detect and correct this tag
            Else if the tags are nested in the wrong order
                Then correct the tag order
            Else if there are tags lacking close '>'
                Then fix this case
            End if
        End
End for
```

4.2 Content Structure Construction

We are describing our proposed system "Retrieve Main Content using Vision-base Web Page Segmentation with Gomory-Hu tree" as shown in figure 1. First of all, we take a web page as an input of our system and this page is segmented by VIPS algorithm. After this process, the number of different blocks is achieved. These content blocks are not only main content of that page but also noise such advertisement, navigation bar, and so on. Therefore, we find the similarity weight to retrieve main contents of that page using Gomoty-hu tree. This experimental web page is processed through our system to output the main content of this page.

For this particular system, we took web pages from Information Technology education web sites, including *en.wikipedia.org, www.informingscience.us/icarus/*. These are web pages which contain not only news/ articles concerning but also noisy data. These web pages are segmented using VIPS algorithm with Gomory-Hu tree and the performance of this algorithm as shown in figure 2 and 3.

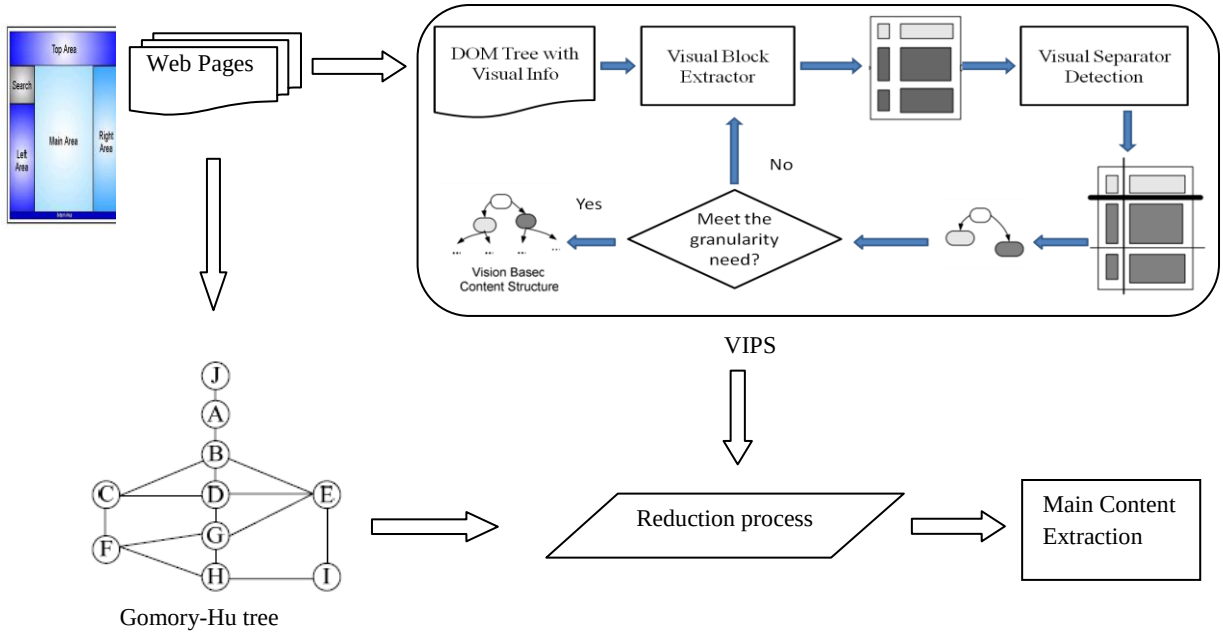

Figure 1: Vision –based Page Segmentation algorithm with Gomory-Hu tree

5. EVALUATION OF THE PROPOSED SYSTEM

For main content extraction of web page, it needs to segment the web page into semantic blocks correctly. VIPS with Gomory-Hu tree construction, we did 150 pages with 3854 blocks. The data set is crawled automatically and the blocks are manually segmented. We compared our algorithm (VIPS combine with Gomory-Hu tree) with simple VIPS algorithm.

The experimental results show that our approach more improves precision and recall than simple VIPS algorithm.

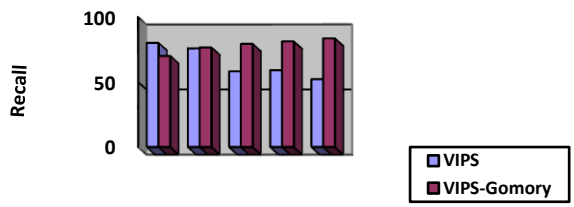

Figure 2. Recall Comparison

Table1. Recall Comparison

	20%	40%	60%	80%	100%
VIPS	80	75.8	58.2	59	52.1
VIPS-Gomory	70	76.5	79.3	81.2	83.5

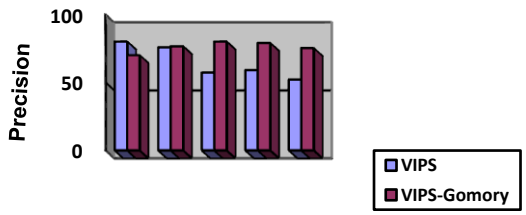

Figure 3. Precision comparison

Table2. Precision comparison

	20%	40%	60%	80%	100%
VIPS	80	75.8	57.2	59	52.1
VIPS-Gomory	70	76.5	80	79	75.1

5.1 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 paper we have addressed the web page segmentation problem by proposing an algorithm based on Gomory-Hu tree. The algorithm uses vision and structure information from a web page to construct a weighted undirected graph, and then it partitions the graph with the Gomory-Hu tree based clustering algorithm.

6. ACKNOWLEDGMENTS

I would like to express my gratitude to my teacher without his consistent and illuminating instruction; this paper could not be finished so quickly. Then thanks to the University of Computer Studies, Yangon which provide us. Thanks to all contributors to this paper.

7. REFERENCES

[1] Cai, D.,Yu, S., Wen, J.R., Ma, W.Y., “VIPS: A vision-based segmentation algorithm”. 2003.

[2] Elgin Akpinar and Yeliz Yesilada, “Vision Based Page Segmentation: Extended and Improved Alorithm”, 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] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[5] Amit Chauhan, Himanshu Uniyal, Dr.Bhasker Pant, "Cleaining Web Pages for Relevant Text Extraction and Text Categorization", Graphic Era University, India.

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

[7] Swe Swe Nyein, "Mining Contents in Web Page Using Cosine Similarity", University of Computer Studies, Yangon, Myanmar.

[8] Xinyue Liu, 2011 "Segmenting Webpage with Gomory-Hu Tree Based Clustering", Dalian University of Technology, Dalian, China.

[9] Han Fengjiao, Zhou Zhurong, 2012, "Semantics-based Extraction of Webpage Main Text", Chongqing.

[10] Aihua Zhang, Jiwu Jing, Le Kang, Lingchen Zhang, "Precise web page segmentation based on semantic block headers detection", University of science and technology, China.

[11] Chaw Su Win, "Informative Content Extraction By using Eifce", IJSTR, 2013