Proposal of seam degree and content similarity for web page segmentation

Zeng, Jun
Graduate School of Information Science and Electrical Engineering, Kyushu University

Flanagan, Brendan
Graduate School of Information Science and Electrical Engineering, Kyushu University

Xiong, Qingyu
Graduate School of Software Engineering, Chongqing University

Wen, Junhao
Graduate School of Software Engineering, Chongqing University

他

出版情報：Proceedings - 2nd IIAI International Conference on Advanced Applied Informatics, IIAI-AAI 2013, pp.9-14, 2013-12-16
バージョン：
権利関係：

Proposal of Seam Degree and Content Similarity for Web Page Segmentation

Jun Zeng
Graduate School of Information Science and Electrical Engineering,
Kyushu University,
Fukuoka, Japan
zeng.j.000@s.kyushu-u.ac.jp

Brendan Flanagan
Graduate School of Information Science and Electrical Engineering,
Kyushu University,
Fukuoka, Japan
bflanagan.kyudai@gmail.com

Qingyu Xiong
Graduate School of software engineering,
Chongqing University,
Chongqing, China
xiong03@cqu.edu.cn

Junhao Wen
Graduate School of software engineering,
Chongqing University,
Chongqing, China
jhwen@cqu.edu.cn

Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University,
Fukuoka, Japan
hirokawa@cc.kyushu-u.ac.jp

***Abstract*— Page segmentation has received great attention in recent years. However, most research has been based on some pre-defined heuristics or visual cues which may be not suitable for large-scale page segmentation. In this paper, we proposed two parameters: seam degree and content similarity, to indicate the coherent degree of a page block. Instead of analyzing pre-defined heuristics or visual cues, our method utilizes the visual and content features to determine whether a page block should be divided into smaller blocks. We also proposed a principled page segmentation method using these two parameters. An experiment was conducted to determine the relationship between the two parameters and the number of segment results. The empirical results also show that our segmentation method can effectively segment a page into different semantic parts.**

Keywords-page segmentation; seam degree; content similarity; semantic segment.

I. INTRODUCTION

Web pages are typically designed for visual interaction. In order to support visual interaction, Web pages are designed to consist of multiple segments with different functionalities, such as: main content, navigation bar, menu list, advertisements, etc. Recent research has shown that Web pages can be sub-divided into smaller segments. This process is known as Web page segmentation. The goal of Web page segmentation is to break a large page into smaller segments, in which contents with coherent semantics are collected [8]. Web page segmentation can be very useful for different fields, for example: Web pages can be properly displayed or repurposed for mobile devices [2, 3, 14], duplicate Web pages can be detected [4, 5], information retrieval systems can use such implicit information to provide better search results [6, 7], etc.

Recognizing the importance of Web page segmentation, numerous previous works have proposed to solve this problem. These works can be roughly divided into two types: an HTML structure-based method and a visual heuristic-based method. An HTML structure-based method often transforms HTML code into a Document Object Model (DOM) tree or HTML tag tree, and divides pages based on their pre-defined syntactic structure. However, tags such as <TABLE> and <P> are used not only for content markup but also for layout structure presentation. It is difficult to obtain the appropriate segmentation granularity. Moreover, in many cases DOM prefers presentation over content and therefore it is not accurate enough to discriminate between different semantic segments in a web page [6]. Visual heuristic-based approaches rely on visual cues from browser renderings. Most of the vision-based methods focus on the location, size or font features of elements. However, most of these methods involve some set of heuristics. These heuristics typically utilize many features present on a Web page. While a heuristic approach might work well on small sets of pages, it isn't suitable for large-scale sets of pages [5].

In this paper we propose two parameters for Web page segmentation. The two parameters are Seam Degree (SD) and Content Similarity (CS). Seam Degree describes the seam degree of two adjoining blocks. Content Similarity describes the similarity of contents in two blocks. The two parameters can utilize the vision and content feature to describe the coherent degree of Web page blocks. A Web pages block may contain many smaller sub-blocks. The averaging coherent degree of sub-blocks can determine whether a block should be divided into smaller blocks. By adjusting the threshold of the two parameters, we can obtain a fine-grained page segmentation result. These two parameters do not depend on either pre-defined HTML syntactic structure or visual heuristics. We built a page segment system using the two parameters. Through empirical analysis we show that the page segment system can divide a Web page into appropriate semantic segments.

The rest of the paper is organized as follows: Related works are reviewed in Section II. Notation and problem description are introduced in Section III and Section IV. The seam degree and content similarity are described in Section V and Section VI. A segmentation method is proposed in Section VII. Empirical analysis and result are reported in Section VIII. Finally, conclusion and future work are given in Section I.

II. RELATED WORK

In the past few years, there has been plenty of work on automatic Web page segmentation. A Good survey of works on Web information extraction can be found in [1]. The page segmentation solutions roughly fall into two categories: HTML structure-based approaches and vision-based approaches.

A. HTML structure-based Approaches

HTML source code is often transformed into DOM tree or tag tree. D. Chakrabarti et al. [4] proposed a graph-theoretic approach to Web page segmentation. Their approach is based on formulating an appropriate optimization problem on weighted graphs, where the weights can determine whether two nodes in the DOM tree should be placed together or apart in the segmentation. However, this algorithm needs data learning and this could be an issue in the overall automation of the process. X. Liu et al. [11] proposed a Gomory-Hu Tree based Web page segmentation algorithm. The algorithm firstly extracts vision and structure information from a web page to construct a weighted undirected graph, whose vertices are the leaf nodes of the DOM tree and the edges represent the visible position relationship between vertices. Then it partitions the graph with a Gomory-Hu tree based clustering algorithm. G. Hattori et al. [12] proposed a Web page segmentation method which utilized both content-distance and page layout information. The content-distance depends on the relative HTML tag hierarchy, and layout analysis is only based on the HTML tag. However the layout information of an HTML tag does not always correspond to the actual layout of a Web page.

B. Vision-based Approaches

Vision-based approaches rely on visual cues from browser renderings. Most of the vision-based methods focus on the location, size or font features of elements. D. Cai et al. [10] proposed a Vision-based Page Segmentation (VIPS) algorithm to divide a web page into semantic segments. They consider that each DOM node corresponds to a block. Each node is assigned a value (Degree of Coherence) to indicate how coherent the content is in the block. However, the VIPS algorithm depends on the visual cues, which are only fit for a small set of Web pages. H. Guo et al. [13] proposed to use visual renderings of the web page provided by Mozilla. The authors indicate that information about spatial locality is most often used to cluster, or draw boundaries around groups of items in a web page, while information about presentation style similarity is used to segment or draw boundaries between groups of items. P. Xiang et al. [14] proposed that a web page is considered as a composition of basic visual blocks and separators. Therefore, their algorithm focuses on first identifying the blocks and then discovering the separators between these blocks.

Besides the two major categories, there are several other methods [4, 8, 10]. Due to paucity of space we don't introduce these methods. Our work is closed to VIPS. However VIPS utilizes the visual cues which cannot be suitable for every page. Moreover, these visual cues cannot correctly indicate the difference between different semantic segments. Instead of analyzing the visual cues, we utilize seam degree and content similarity to indicate how coherent the content is in the block.

III. NOTATIONS

A web page is made up of finite blocks. We also call these blocks visual block or block for short. We consider a visual block as a visible rectangular region on a web page. The definition of a visual block is as follows:

Definition III-1: Visual block $VB = (E, R)$, where E is an Element object that is defined by the HTML DOM based on W3C standard, and R represents the visible rectangular region where VB is displayed in the web page.

According to W3C standard, the Element object of the DOM represents an element in the HTML document. The details of Element object can be found in the official W3C website [16]. The Element object not only contains the attributes of an HTML element, such as "tagName", "id", "value" etc., but also contains the properties defined by the DOM, such as "childNodes", "nextSibling", etc. Besides the DOM Element, the other parameter is the visual rectangular region $R = (x, y, w, h)$ as shown in Figure 1. Here x is the horizontal coordinate, y is the vertical coordinates of top-left point of visual block, w is the width, and h is the height of the visual block. Sometimes they are written as $Rx(VB)$, $Ry(VB)$, $Rw(VB)$ and $Rh(VB)$ when only one parameter needs to be mentioned. According to the definition of visual block, not all HTML elements have their corresponding visual blocks. The elements that are not visible such as <head>, <script>, <meta>, etc. and the elements whose "display" property is "none" or "hidden" property is "true" are not considered as a visual block in this paper.

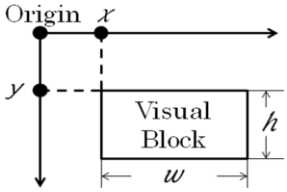


Figure 1. The absolute coordinate and size of a visual block

Definition III-2: For two given visual blocks $VB_1 = (E_1, R_1)$ and $VB_2 = (E_2, R_2)$, if E_1 is a child node of E_2, then VB_1 is the child of VB_2.

Definition III-3: If a visual block $VB = (E, R)$ does not have any children, then VB is a leaf visual block, denoted $VB : leaf$.

IV. PROBLEM DISCUSSION

The purpose of our work is to break a large page into smaller segments, in which contents with coherent semantics are collected.

Figure 2. An example of different semantic segment.

A Web page can be considered as a large block, which consists of several child blocks with different semantics, such as: main content, navigation bar, menu list, advertisements, etc. These segments have different functions and visual characteristics. For example, in a news site, a long text may be the main content; a link list may be the related news list; a big picture may an advertisement, etc. Figure 2 shows the example. We can utilize the visual and content difference to indicate how coherent the child blocks are. If the coherent degree of child blocks is high, then the block should not be divided, otherwise it should be divided further. Therefore the issue of page segmentation can be seen as an issue of calculating the coherent degree of child blocks in each block. In this paper, we introduce two parameters to describe the coherent degree, they are: the Seam Degree and Content Similarity. In the next section, we will introduce the two parameters in detail.

V. SEAM DEGREE

A. The Seam Degree of Two Adjoining Blocks

The seam degree is used to describe how close two adjoining blocks are. First, we give the definition of adjoining blocks.

Definition V-1: For two given visual blocks VB_1 and VB_2, let's assume that $Ry(VB_1) + Rh(VB_1) \le Ry(VB_2)$. If the following conditions are satisfied:

(1) $Max\{Rx(VB_1), Rx(VB_2)\} \le Min\{Rx(VB_1)+Rw(VB_1), Rx(VB_2)+Rw(VB_2)\}$;

(2) There is NOT a VB_i which is between VB_1 and VB_2.

We define VB_1 and VB_2 are adjoining in the vertical direction. Similarly, we can also define two visual blocks are adjoining in the horizontal direction (we skip over it here). If VB_1 and VB_2 are adjoining in the vertical direction or horizontal direction, we define VB_1 and VB_2 are adjoining blocks. Figure 3 shows two examples of adjoining blocks.

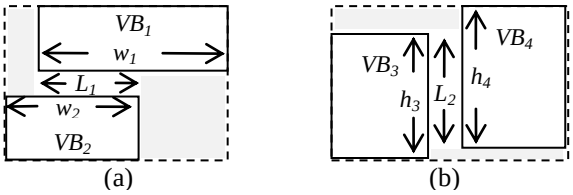


Figure 3. Two examples of adjoining blocks.

In Figure 3, VB_1 and VB_2, VB_3 and VB_4 are adjoining blocks. The dotted rectangles are the minimum rectangles that cover the two blocks in Figure 3 (a) and Figure 3 (b). L_1 and L_2 are the seam length of the two adjoining blocks, w_i is the width of VB_i, and h_i is height of VB_i. Intuitively, we consider VB_3 and VB_4 are closer than VB_1 and VB_2. This is because VB_3 and VB_4 can almost fill up the minimum rectangle, but VB_1 and VB_2 cannot fill up it. The gray regions indicate the areas that are not filled up in Figure 3. It is known that each segment has a corresponding rectangle appearing in the page. In other words, VB_3 and VB_4 are more likely to be a segment, but VB_1 and VB_2 cannot be considered as a segment. We utilize seam degree to describe the visual coherent degree. The definition of seam degree is given as follows:

Definition V-2: For two given visual blocks VB_1 and VB_2, if VB_1 and VB_2 are adjoining in vertical direction. The seam degree $SD(VB_1, VB_2)$ can be calculated as in formula (1):

$$SD(VB_1, VB_2) = \frac{SeamLength(VB_1, VB_2)^2}{Rw(VB_1) \times Rw(VB_2)} \quad (1)$$

where $SeamLength(VB_1, VB_2)$ represents the seam length of VB_1 and VB_2, and $Rw(VB_i)$ represents the width of VB_i. Similarly, if VB_1 and VB_2 are adjoining in horizontal direction, The seam degree $SD(VB_1, VB_2)$ can be calculated as in formula (2):

$$SD(VB_1, VB_2) = \frac{SeamLength(VB_1, VB_2)^2}{Rh(VB_1) \times Rh(VB_2)} \quad (2)$$

where $Rh(VB_i)$ represents the height of VB_i.

$SD(VB_1, VB_2)$ is between 0 and 1. Since the seam degree is based on the visual information of blocks, it can indicate the visual coherent degree of adjoining blocks.

B. The Averaging Seam Degree of Adjoining Child Blocks in a Block

If a block has child blocks, the averaging seam degree of adjoining child blocks can indicate the visual coherent degree of the content in the block. For a given visual block VB, the set of child blocks in VB is $Child(VB) = \{b_1, b_2, \ldots, b_n\}$. If two child blocks are adjoining, we count 1 pair. Let us assume that there are n pairs of adjoining child blocks. The averaging seam degree $AvgSD(VB)$ can be calculated as in formula (3);

$$AvgSD(VB) = \frac{\sum SD(b_i, b_j)}{n} \quad (3)$$

where b_i and b_j are adjoining child blocks.

$AvgSD(VB)$ degree is also between 0 and 1. If it is closer to 0, the visual coherent degree of child blocks is lower. If it

is closer to 1, the visual coherent degree of child blocks is higher.

VI. CONTENT SIMILARITY

A. The Content Vectors of a Block

As mentioned before, segments with different semantics always have different types of contents. For example, a navigation bar has a list of short link text; an advertisement has a big picture; a user registration form has some text boxes, pull-down menus, buttons, etc. If the contents of two blocks are similar, the two blocks have a high content coherent degree. We introduce the Content Similarity to describe the content coherent degree. We roughly classify the contents into four categories:

(1) **Text Contents (TC)**: all the text falls into this category, except the text that contains a hyper link.
(2) **Link Text Contents (LTC)**: the text that contains a hyper link can be classified into this category.
(3) **Image Contents (IMC)**: this category contains pictures, photos, icons, etc.
(4) **Input Contents (INC)**: this category includes elements that can accept user input, such as: text box, radio button, pull-down menus, etc.

For a given *VB*, the content set is $C = \{c_1, c_2, \ldots, c_n\}$. First, the contents are classified into the four categories mentioned above. Then four types of content sets can be obtained, denoted $TC = \{tc_1, tc_2, \ldots, tc_o\}$, $LTC = \{ltc_1, ltc_2, \ldots, ltc_p\}$, $IMC = \{imc_1, imc_2, \ldots, imc_q\}$, and $INC = \{inc_1, inc_2, \ldots, inc_r\}$. Obviously, *TC*, *LTC*, *IMC* and *INC* are the subsets of *C*. If one of the content subsets is ∅, it means that *VB* does not contain the contents of the corresponding type. We use $Area(c_i)$ to represent the area of the corresponding block of c_i. If c_i is a text content or link text content, we approximately calculate the area as in formula (4):

$$Area(c_i) = Length(c_i) \times FontSize(c_i)^2 \quad (c_i \in TC \cup LTC) \tag{4}$$

where $Length(c_i)$ represents the length of text or link text, $FontSize(c_i)$ represents the font size of text or link text.

According to the area of contents, the four content subsets can be sorted from large to small area. By utilizing the sorted content subsets, four content area vectors can be obtained, denoted V_{tc}, V_{ltc}, V_{imc} and V_{inc}. The values of elements in the four vectors are the areas of corresponding contents. After the content vectors are determined, the content similarity of two blocks can be calculated.

B. The Content Similarity of Two Blocks

If the content vectors of two given blocks are determined, the similarity of each content area vector can be calculated. There are many algorithms to calculate the similarity of two vectors, of which the cosine similarity is a simple and efficient algorithm [15]. Here we take the vector of the text content as an example to explain the calculation of cosine similarity. For two given blocks VB_1 and VB_2, their text content area vectors are $V_{tc\_1} = (u_1, u_2, \ldots, u_m)$ and $V_{tc\_2} = (v_1, v_2, \ldots, v_n)$. Let us assume that $V_{tc\_1} \neq \emptyset$, $V_{tc\_1} \neq \emptyset$, and $n > m$. Because the cosine similarity requires that the two vectors must have the same number of elements, we need to add $(n-m)$ elements whose value are 0 into V_{tc_1}, denoted $V'_{tc\_1} = (u_1, u_2, \ldots, u_m, u_{m+1}, \ldots, u_n)$. The cosine similarity of V'_{tc_1} and V_{tc_2} can be calculated as in formula (5):

$$Cos(V'_{tc\_1}, V_{tc\_2}) = \frac{\sum_{i=1}^{n} u_i \times v_i}{\sqrt{\sum_{i=1}^{n} (u_i)^2} \times \sqrt{\sum_{i=1}^{n} (v_i)^2}} \tag{5}$$

If both V'_{tc_1} and V_{tc_2} are ∅, $Cos(V'_{tc\_1}, V_{tc\_2})$ is ill-formed. In this case, we define the $Cos(V'_{tc\_1}, V_{tc\_2})$ to be zero. Similarly, the cosine similarity of other content area vectors (including V_{ltc}, V_{imc} and V_{inc}) can also be determined.

Additionally, the four types of contents have different weight in VB_1 and VB_2. Also, we take the text content as an example to explain the calculation of weight. For two given blocks VB_1 and VB_2, their text content area vectors are $V_{tc\_1} = (u_1, u_2, \ldots, u_m)$ and $V_{tc\_2} = (v_1, v_2, \ldots, v_n)$. The weight of text content can be calculated as in formula (6):

$$Weight(Tc) = \frac{\sum_{i=1}^{m} u_i + \sum_{j=1}^{n} v_j}{Area(VB_1) + Area(VB_2)} \tag{6}$$

where the $Area(VB_i)$ represents the total area of all contents in VB_i. It means that the greater area of the corresponding type of contents is, the higher its weight will be.

After the cosine similarity and weight of each content area vector are determined, the content similarity $CS(VB_1$ and $VB_2)$ of VB_1 and VB_2 can be calculated as in formula (7):

$$CS(VB_1, VB_2) = \sum Weight_i \times Cos_i \tag{7}$$

where $Weight_i$ represents the weight of four types of contents, and Cos_i represents the cosine similarity of four types of contents area vectors.

$CS(VB_1, VB_2)$ is between 0 and 1. Since the content similarity is based on the content information of blocks, it can indicate the content coherent degree of blocks.

C. The Averaging Content Similarity of Adjoining Child Blocks in a Block

Similar to the averaging seam degree, if a block has child blocks, the averaging content of adjoining child blocks can indicate the content coherent degree of the child blocks in the block. It should be noted that only the content similarity of adjoining child blocks is considered. For a given visual block *VB*, the set of child blocks in *VB* is $Child(VB) = \{b_1, b_2, \ldots, b_n\}$. If two child blocks are adjoining, we count 1 pair. Let us assume that there are *n* pairs of adjoining child blocks. The averaging seam degree $AvgCS(VB)$ can be calculated as in formula (8);

$$AvgSD(VB) = \frac{\sum CS(b_i, b_j)}{n} \tag{8}$$

where b_i and b_j are adjoining child blocks.

$AvgCS(VB)$ is also between 0 and 1. If it is closer to 0, the content coherent degree of child blocks is lower. If it is closer to 1, the content coherent degree of child blocks is higher.

VII. PAGE SEGMENTATION BASED ON SEAM DEGREE AND CONTENT SIMILARITY

Based on the seam degree and content similarity, we propose a page segmentation method. In order to divide a page into segments, a page needs to be transformed into a DOM tree. The nodes that will not appear in the Web page should be pruned, such as the nodes whose tags are <SCRIPT>, <META>, <STYLE>, etc, and the nodes whose height or width is zero. Also, we need to get the visual information of each node by utilizing the APIs of browsers. This is because the DOM nodes do not contain the absolute coordinate. In this way, the corresponding block of each DOM node can be determined.

For a given block, the averaging seam degree and content similarity of its adjoining child blocks are calculated. We introduce two thresholds α and β to determine whether the node should be divided or not. The segmentation algorithms are as follows:

Step 1: For a given block, if the averaging seam degree and content similarity of its adjoining child blocks is less than α, then the block should be divided.

Step 2: For a given block, if the averaging seam degree and content similarity of its adjoining child blocks is greater than α, and the averaging content similarity of its adjoining child blocks is less than β, then the block should be divided.

Step 3: For a given block, if it does not satisfy the Rule 1 and Rule 2, then the block should not be divided.

If a given block that does not contain any child block, we define both its averaging seam degree and content similarity are one. If a given block that contains only one child block, we define both its averaging seam degree and content similarity are zero. Our method is a top-down algorithm. We calculate the averaging seam degree and content similarity from the root block. If a block should be divided, its child blocks will be checked further. If a block should not be divided, it will be pushed into a segment array and its child blocks will not be checked any more. Finally, all the segments can be determined.

VIII. EXPERIMENT AND ANALYSIS

We submitted 10 queries to Google, from which we randomly collected 10 pages from the search results as test pages. We set the thresholds α and β to be 0 to 1 respectively, where the step is 0.1, and obtained a set of 121 α and β pairs { (0, 0), (0, 0.1), …, (1, 0.9), (1, 1)}. Using the 121 threshold pairs, we use our method to segment each page 121 times. Each time we recorded the segment numbers of each page, denoted by n_i . For a given Web page, the set of segment number is $\{n_1, n_2, \ldots, n_{121}\}$. Since the segment number of each page is different, the set of segment numbers should be normalized as in formula (9):

$$Normalize(n_i) = \frac{n_i}{Max\{n_1, n_2, ..., n_{121}\}} \quad (9)$$

where $Max\{n_1, n_2, \ldots, n_{121}\}$ represents the largest value of the set.

We calculated the average results of the normalized segment number of the 100 pages, denoted $\{N_1, N_2, \ldots, N_{121}\}$. Figure 4 shows the coordinate graph of these results.

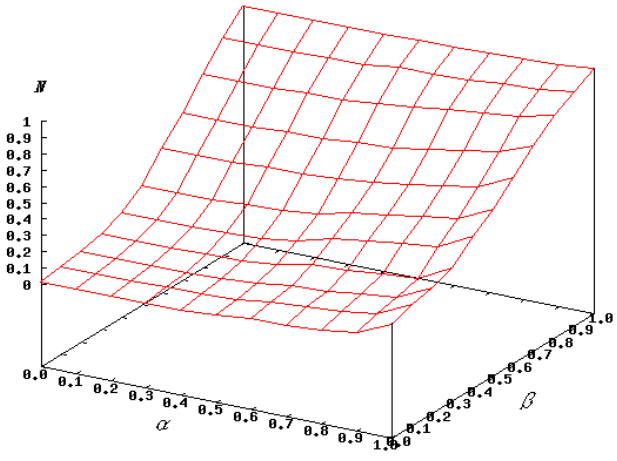


Figure 4. The coordinate graph of results.

According to Figure 4, the following inferences can be drawn:

(1) According to section VII, the blocks whose average seam degrees are less than α will be divided. Thus, if α=0, the step 1 in Section VII will be invalid. In other words, only the average content similarity is effective to determine whether divide a block or not. In this case, β and the normalized segment number are approximately proportional. We can infer that the averaging content similarities of all the blocks of are approximate uniform distribution.

(2) According to section VII, if α=1, most of the blocks should be divided, and the normalized segment number should be close to one no matter how β changes. However, that was not the case. β and the normalized segment number are still approximately proportional. We can infer that the averaging seam degrees of most blocks are one.

(3) If α is constant, the curve increases steeply along with β. Conversely, if β is constant, the curve increases gradually along with α. We can infer that the averaging content similarity plays a main role to determine whether a block should be divided or not, and the average seam degree plays a supplementary role.

Based on the three inferences above, we let α be 0.9 and β be 0.8, and segment the 100 pages using our method. Figure 5 shows some examples of the page segmentation results, and the red rectangle represents the segments.

These examples show that our method is effective to segment a web page into different semantic parts.

I. CONCLUSION AND FUTURE WORK

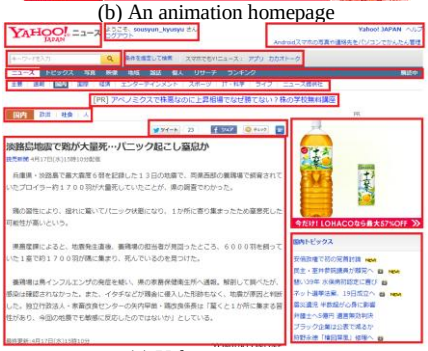
In this paper, we proposed two parameters seam degree and content similarity to indicate the coherent degree of a page block. The seam degree is based on the visual information of blocks, therefore it can indicate the visual coherent degree of adjoining blocks. The content similarity is

based on the content information of blocks, therefore it can indicate the content coherent degree of blocks. Instead of analyzing pre-defined heuristics or visual cues, our method utilized the visual and content coherent degree to determine whether a page block should be divided into smaller blocks. We also proposed a page segmentation method using these two parameters. An experiment was conducted to determine the relationship between the two parameters and the number of segment result. The empirical results also show that our segmentation method is effective to segment a page into different semantic parts.

However our method cannot identify recurrent blocks. For example, in the search result page of Amazon, each product record has an independent semantic. Since they have similar contents, they are probably not divided into different segments. In the future, we are planning to solve this problem and improve the segmentation results.

(a) A blog search result page

(b) An animation homepage



(c) Yahoo news page

Figure 5. Several examples of segmentation results.

REFERENCES

[1] Yesilada, Yeliz, “Web Page Segmentation: A Review”. Technical Report. University of Manchester and Middle East Technical University Northern Cyprus Campus, 2011. (Unpublished)

[2] X. Yin and W.S. Lee. “Using link analysis to improve layout on mobile devices”. In Proceedings of the Thirteenth International World Wide Web Conference, pages 338–344, 2004.

[3] X. Xie, G. Miao, R. Song, J. Wen, and W. Ma. “Efficient browsing of web search results on mobile devices based on block importance model,” In Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications, pages 17–26, 2005.

[4] C. Kohlschutter and W. Nejdl. “A densitometric approach to web page segmentation,” In Proceeding of the 17th ACM conference on Information and knowledge management, CIKM ’08, pages 1173–1182, 2008.

[5] D. Chakrabarti, R. Kumar, and K. Punera. “A graph-theoretic approach to webpage segmentation,” In WWW’08: Proceeding of the 17th international conference on World Wide Web, pages 377–386, 2008.

[6] A. Madaan, W. Chu, S. Bhalla, "VisHue: Web Page Segmentation for an Improved Query Interface for MedlinePlus Medical Encyclopedia," Databases in Networked Information Systems, Vol. 7108, pp 89-108, 2011.

[7] K. S. Kuppusamy, G. Aghila, "Multidimensional web page segment evaluation model," Journal of Computing, Vol. 3, Iss. 3, pp.24-27, 2011.

[8] P. Xiang, X. Yang, and Y. Shi, "Web page segmentation based on gestalt theory," In Multimedia and Expo 2007 IEEE International Conference (ICME), pp. 2253-2256, 2007.

[9] D. Cai, S. Yu, J. Wen, and W.g Ma. "VIPS: a vision based pagesegmentation algorithm". Technical Report MSR-TR-2003-79, Microsoft Research, 2003.

[10] H. Sano, R. M. E. Swezey, S. Shiramatsu, T. Ozono, and T. Shintani, "A Web Page Segmentation Method by using Headlines to Web Contents as Separators and its Evaluations," International Journal of Computer Science and Network Security, Vol. 13, No. 1, pp.1-6, 2013.

[11] Xinyue Liu, Hongfei Lin, and Ye Tian, "Segmenting webpage with Gomory-Hu tree based clustering," Journal of software, Vol. 6, No. 12, 2011.

[12] G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya, "Robust Web Page Segmentation for Mobile Terminal Using Content-Distances and Page Layout Information," in Proceedings of the 16th international conference on World Wide Web (WWW '07), pp.361-370, 2007.

[13] H. Guo, J. Mahmud, Y. Borodin, A. Stent, I.V. Ramakrishnan, "A general approach for partitioning web page content based on geometric and style information," in Proceedings of the International Conference on Document Analysis and Recognition, pp. 929-933, 2007.

[14] P. Xiang and Y. Shi. "Recovering semantic relations from web pages based on visual cues," In Proceedings of the 11th international conference on Intelligent user interfaces(IUI’06), pp.342–344, 2006.

[15] S. Amit, "Modern Information Retrieval: A Brief Overview," Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 24, No. 4, pp.35–43, 2001.

[16] http://www.w3.org/standards/techs/dom#w3c_all, 2012