International Journal of Hybrid Information Technology
Vol.8, No.2 (2015), pp.247-256
http://dx.doi.org/10.14257/ijhit.2015.8.2.23

Research and Implementation of View Block Partition Method for Theme-oriented Webpage

Lv Fang, Huang Junheng, Wei Yuliang and Wang Bailing*

Harbin Institute of Technology at Weihai, Shandong, 264209
{huangjh, wbl@hitwh.edu.cn}

Abstract

A semantic block is treated as a unit while analyzing the webpage. First, we implement the VTPS algorithm to partition a webpage into semantic blocks. Then, we propose an algorithm to extract the spatial and content features, and then construct the feature vector for each block. Based on these vectors, the SVM learning algorithm is applied to train and classify the various theme-oriented webpage blocks. At last, the classification experiments show the efficiency of this method.

Keywords: *Semantic Block, VTPS Algorithm, Feature Vector, Classification*

1. Introduction

In the Internet age, Web has become an important means for people to obtain information online. With the rapid increase of information spreading on the Web, an effective method for users to discern the useful information from the junk is in great demand. Different information inside a web page has different importance for different applications. Therefore, the technology of page segmentation which is useful in extracting information from the webpage has gain more attention. Web page is radically distinct from traditional text due to its dynamic and informal natural. Meanwhile webpage has extensive content performance and interaction features. Visually, the message of webpage can be conveyed through its presentation format. The layout features of webpage are much more important than the content features [1].The technology of page segmentation is facilitated by visual features, such as information extraction [2], information retrieval [3], information storage, web page classification, and web adaption. However, most of the vision-based page segmentation methods lack of generality for they are designed for a special application.

The main contributions of this paper are: 1) the introduction of the related technology; 2) the VTPS algorithm is proposed to partition a web page into semantic blocks; 3) a theme-oriented webpage partition model is proposed to automatically assign different function areas in the webpage. This model takes into account both spatial features and content features.4) do experiments to test the model.

ISSN: 1738-9968 IJHIT

2. Related Work

2.1. Document Object Model Structure

In general, the structure of an HTML document is composed of kinds of labels and components. Meanwhile, the order they appear in the document is the same as its display order. DOM tree is a tree topology structure which is obtained by parsing the HTML document. The DOM tree is good at describing data of a semi-structured nature such as HTML document. The DOM tree structure can accurately describe the relative position and hierarchical relationships between the tree nodes. From the visual angel, Web can be viewed as the set of visual blocks which were defined by recursion, a visual block can be seen as the set of smaller visual blocks. Each node in the DOM tree is a component object (such as element, attributes, text) of the HTML documents. The root node of DOM tree is HTML document, all of the body text, picture, hyperlinks and tag are leaf nodes. We can complete all kinds of HTML document processing by operating DOM nodes.

2.2. Page Segmentations based on Layout

One of the efficient and widely researched page segmentation which is based on the layout of webpage is VIPS(Vision-based Page Segmentation) [4].By detecting useful visual cues based on DOM structure, a tree-like vision-based content structure of web page is obtained. Visual cues such as font, color and size are used to detect blocks. The three steps of the algorithm are shown as follows: 1) extract visible blocks. The webpage is divided into several separate semantic blocks recursively; 2) detect the separator bar. Finding out the visual vertical and horizontal lines of the webpage which are used to divide the webpage; 3) recreate the content structure. The segmentation will not stop unless the coherence between each block is bigger than the threshold.

VIPS excels in both an appropriate partition granularity and coherent semantic aggregation. However, the complexity of VIPS is high and it is difficult to ensure the consistency and integrity of the heuristic rules.

Many researcher are finding plenty of inspiration in VIPS. [5] have proposed a CTVPS which will reduce the time and space complexities obviously. However CTVPS does not suitable for “div+css” layout which is very popular now. Most of the page segmentation are based on special application, [2] proposed a webpage page segmentation to receive the subject information by delete the blocks which have no relevance to the subject. This method isn’t generic, even though the retrieval performance is improved.

2.3. Block Importance Model

Though page segmentation take one step ahead to look down into the structure of a webpage instead of treating it as a unit, they do not differentiate the function of the blocks in

a page and still treat them uniformly. The block importance model treats a semantic block instead of a web page as a unit.

The essence of the block importance model advantages lies in the ability the most important content from less important and noisy information. The model could automatically analyze the information in a webpage and assign importance measures for different regions in the web page. Any application involved web page will be facilitated by the block importance model, such as information retrieval [6] web page classification [9] and web adaption [8].

The block Importance Model in [9] use VIPS to segment a page into semantic blocks. The spatial and content features of each semantic block are extracted to construct its feature vector to present the block. Based on these features, leaning algorithm, such as SVM and neural network, are applied to train various block importance models.

3. Vision-based Webpage Segmentation Principle

To facilitate the description of the Page Segmentation (VTS) principle. Define it as follows,

Definition 1 Segments Partition: It is the course that divides the page elements which have close semantics into the set of semantic blocks with appropriate granularity.

Definition 2 Block Tree: The set of semantic blocks obtained from segmentation algorithm is a hierarchy tree, here we call it Semantic Tree, or Block Tree. We can define the semantic block of the Block Tree as a range of non-overlapping semantic blocks recursively.

Definition 3 Page Consistency: The fashion and purpose of any element object in single page are consistent.

The procedure of vision-based segmentation algorithm, are shown as follows:

1. The HTML tag, attributes and spatial information will be saved in DOM tree by using the Dom parser. These attributes information comes from the HTML element objects, and the spatial information comes from the initial coordinates and size of the semantic block.
2. The layout information, attributes information and spatial information of the DOM node are taken into account in designing the heuristic rules based on Page Consistency. The core ideas of these heuristic rules are combine and cut the lower semantic blocks layer by layer. The Block tree is received by Bottom-up dynamically adjusted to the tree structure.
3. The HTML tag, attributes and spatial information will be saved in DOM tree by using the Dom parser. These attributes information comes from the HTML element objects, and the spatial information comes from the initial coordinates and size of the semantic block.
4. The layout information, attributes information and spatial information of the DOM node are taken into account in designing the heuristic rules based on Page Consistency. The core ideas of these heuristic rules are combine and cut the lower semantic blocks

layer by layer. The Block tree is received by Bottom-up dynamically adjusted to the tree structure.

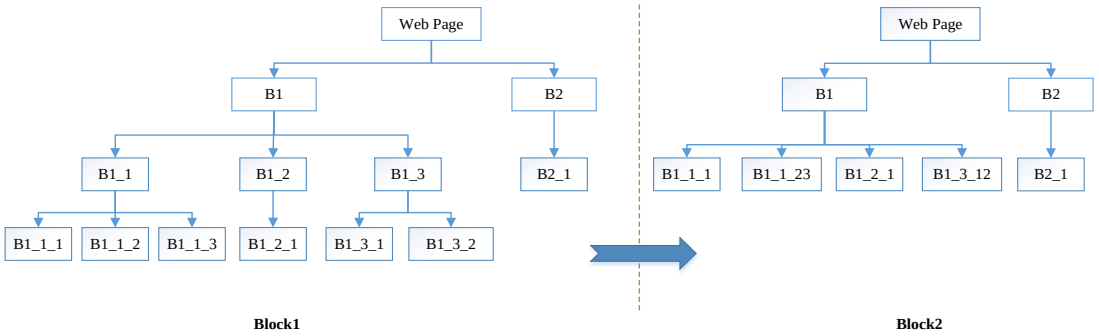

Figure 1. Sketches of the Dynamically Adjustment

Figure 2 shows the schematics of the dynamically adjustment course.

To facilitate the description of heuristic rules. Define it as follows,

Definition 4 Text Node: The elements in DOM tree which was used to describe text content. For instance, "SPAN", "INPUT", "A", "TEXTAREA", "LABEL","EM", "BR", "STRONG","BUTTON", "B", and so on.

Definition 5 Distance: The distance between the Center Coordinate of the brother semantic blocks.

Definition 6 Boundary Diffusion: The minimum rectangular boundary which could cover all the semantic block region.

Definition 7 Combine Rule: First, they are adjacent brother nodes; second, they are Text Node or they have the same tag name; Third, the Distance between them are close.

The heuristic rules will be applied on all the semantic blocks of the Block tree from bottom to up. During this process, the information of lower block will be transfer upward. As Figure 1 shows, the structure of Block1 will be adjusted into Block2 after a level of recursion. Block B1_1 in Block1 has three child nodes, they areB1_1_1, B1_1_2 and B1_1_3.block B1_1_2 and B1_1_3 match to the Combine Rules, then they will be combined into block B1_1_23 based on Boundary Diffusion.B1_1_1 and B1_1_23 which do not meet the Combine Rule, will replace their farther node. Block B1_2 which has only one child node will be replaced by B1_2_1 directly. Similarly, block B1_3 will be replaced by its only child B1_3_12, which was received by combing B1_3_1 and B1_3_2.

4. Vision-based Theme-oriented Webpage Partition Model

Based on the requirement of current research about theme-oriented webpage, the semantic blocks are divided into five categories: navigation area, subject area, recommendation area, advertising area, copyright area.

1. Navigation Area is generally located at top of the page with small area. The link to link text ratio and the number of link to region area ratio are always bigger.
2. Subject Area is generally located around the center axis of the page with bigger area. The number of text to region area is always bigger.
3. Relevant recommendation Area is distributed in the lower and middle or in the sidebar part of the page. The link to link text ratio and the number of link to region area ration are always smaller.
4. Advertising Area has no set positions, contains no link, no text and no picture.
5. Copyright Area is generally located at the bottom of the page. The link to link text ratio and the number of link to region area ratio are always bigger.

4.1. Extracting Text features of Blocks

Definition 8 Text Features: The attributes information in semantic blocks of Block Tree, which was used to describe the information of its links, text, pictures.

Text Features is (pimg_area, pa_area, p_textbox, p_a_text, p_a_btext, p_text_btext).

1. pimg_area denotes image area ratio: $pimg\_area$

$$pimg\_area = img\_area / block\_area \quad (1)$$

2. pa_area denotes the area of link text ratio(400 is the average area of a Chinese characters)

$$pa\_area = 400 * a\_text / block\_area \quad (2)$$

3. p_textbox denotes ratio of the number of textbox in the block to the number in the whole page:

$$p\_textbox = b\_num\_textbox / page\_num\_textbox \quad (1)$$

4. p_a_text denotes the ratio of the number of links to the length of link text :

$$p\_a\_text = a\_num / a\_text \quad (2)$$

5. p_a_btext denotes the link text ratio:

$$p\_a\_btext = a\_text / block\_text \quad (3)$$

6. p_text_btext denotes the plain text ratio:

$$p\_text\_btext = text / block\_text \quad (6)$$

4.2. Extracting Visual Features of Blocks

Definition 9 Visual Features: The spatial information in semantic blocks in Block Tree, such as the position of the block center, the size of the block.

The same block has a quite difference rate in different pages. At present, there are two ways to construct visual feature vectors: absolute visual features and relative visual features. The absolute features only take into account page spatial features, while relative features combined with the size of the display window.

4.2.1. Extracting Absolute Visual Features: Absolute Spatial Features denotes: (p_center_x, p_center_y_ab, p_height_ab, p_width,p_area_ab)

1. center horizontal:

$$p\_center\_x = (left + width/2)/page\_width \tag{4}$$

2. absolute center ordinate:

$$p\_center\_y\_ab = (top + heigth/2)/page\_height \tag{8}$$

3. width ratio:

$$p\_width = width/page\_width \tag{5}$$

4. absolute height ratio:

$$p\_height\_ab = height/page\_height \tag{6}$$

5. absolute area ratio:

$$p\_area\_ab = hight * width/(page\_width * page\_height) \tag{11}$$

4.2.2. Extracting Relative Visual Features: Relative visual features denotes: (p_center_x, p_center_y_re, p_height_re, p_width,p_area_re)

1. p_center_y_re denotes relative height ratio(window_height is the height of the display window)

$$p\_height\_re = height/window\_height \tag{12}$$

2. relative area ratio:

$$p\_area\_re = height * width/(page\_width * window\_height) \tag{7}$$

3. p_center_y_re denotes relative center ordinate ratio(HeaderHeight means the height of page header, FooterHeight means the height of page footer)

$$p\_center\_y\_re = \begin{cases} center\_y/2 * HeadHeight & if\ center\_y < HeaderHeight \\ 0.5 & otherwise \\ 0.5 + (page\_height - center\_y)/(2 * FooterHeight) & \\ & if\ center\_y > FooterHeight \end{cases} \tag{8}$$

5. Model Measurement and Analysis

In order to test the VTS algorithm and VTP models, we choose 110 theme-oriented pages from Tencent, Netease, Xinhua, PD Online, Xinhua BEIJING. All the pages were partitioned into 6964 semantic blocks. According to artificial visual and block content finish manual tagging. Three fifth of the annotated blocks are used as training set and two fifth are testing set. The radial basis kernel function was used in SVM machine learning to learn the models. The higher precision model of block forecast was built by optimizing parameters.

5.1. Evaluation Standard of Model

Classical measures, such as precision, recall and Micro-F1, are borrowed to evaluate the block models. For the overall performance, Micro-F1of each partition is provided.

The number of blocks which belong to the corresponding category returned by the model is a. The number of blocks which was misjudged into the corresponding category returned by the model is b. The number of blocks which was not returned correctly by the model is c.

1. Precision:

$$P(i) = \frac{a}{a+b} \tag{9}$$

2. Recall:

$$R(i) = \frac{a}{a+c} \tag{10}$$

3. In order to show the comprehensive effect of precision ratio and recall ratio, we introduce Micro-F1:

$$F_1(i) = \frac{2 * P(i) * R(i)}{P(i) + R(i)} \tag{11}$$

5.2. Testing and Analysis

5.2.1. Features vs. All Features: We build a model which uses text features and relative visual features. To measure the impacts of spatial features and content features respectively, we also build a model which only uses text content features to represent blocks. We also use SVM with RBF kernel to train the model.

In Figure 2 compares the performance of the model with the one using all features. It is not surprising to see that the model only using text features can achieve good performance. When spatial features are added, there is also a significant increase on performance. The

average accuracy of the model which only uses text features is 83.65 percent, while the average accuracy of the model with all features achieves 90.66 percent. It proves that Spatial Features do provided some complementary information to text features to measure block category.

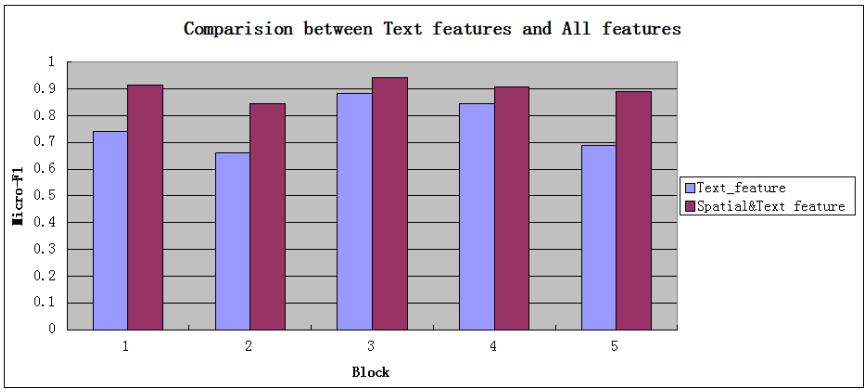

Figure 2. Comparison between Text and All features

5.2.2. Relative Visual Features vs Absolute Visual Features: Relative Visual features are used in above experiments. Here we compare all of the three kinds of spatial features: absolute, relative and comprehensive. The comprehensive features are combined by (p_center_x, p_center_y_ab, p_height_ab, p_width, p_area_re). Three Partition Models are trained based on the three kinds of spatial features respectively. The result comparison is shown in Figure 3. The three model have different performance in the five categories. For navigation area and subject area the performance of model using relative spatial features is much worse than the other two. And the performance of the model using comprehensive features is slightly higher than the other two.

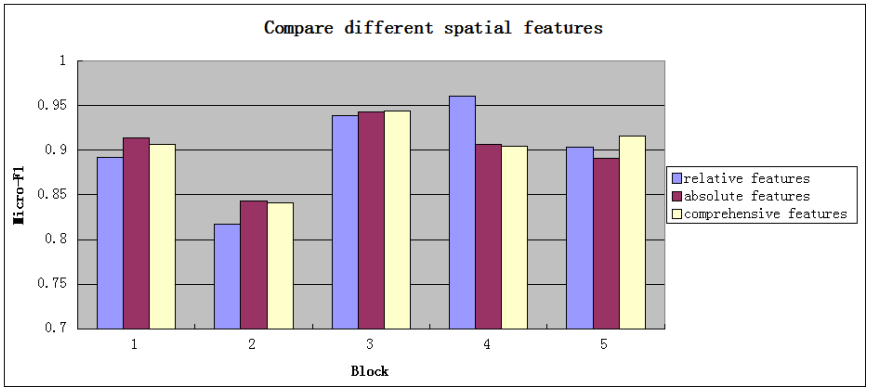

Figure 3. Comparison of Three Kinds of Spatial Features

Since long pages are dominant in our labeled data, and all of the theme-oriented webpages are filled with lots of recommendation and advertisement, the overall performance of relative

features is not very big. In the VTP Model, the advertising areas are the set of semantic blocks which has no text features .That is the reason why spatial features has a great impact on advertising area.

5.3. Application of Webpage Partition Model

Content extraction is used to test the performance of the VTP Model. First of all, partitioned the sample webpages into semantic blocks by VTS algorithm. Then extract the absolute features of each semantic block. Using Absolute Features VTP Model to tag these semantic blocks. All of the text comes from the blocks that has been identified as subject area make up the subject text. Jaccard Similarity method is used to calculate the text similarity between subject text and text in the body. The result is used to evaluate the VTP Model.

We chose 20 theme-oriented webpages to test the Partition Model. The average of text similarity is 95.5%.It is easy to see the block model has great performance in text extraction. The report time in body text is filtered out, some of the text blocks which belongs to recommendation area were chosen, are the reason why there are errors.

6. Conclusion

We implemented a VTPS algorithm to partition a web page into multiple semantic blocks. Based on the partition, we proposed an algorithm to extract feather from each blocks. After combining these algorithms, Then SVM learning algorithm is applied to train and classify partition models based on different kinds of features.

In our experiments, the average accuracy of the best partition model can achieve 91.27%.Visual features have significant effects on the performance of the partition models. Different areas are affected by different features on different levels. The text similarity achieved 95.5% when we apply this partition model to text extraction. In conclusion, the partition method can be used as a universal model for variety of web applications.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 6137117 7 and No. 61170262). We would like to thank the anonymous reviewers for their helpful com ments.

References

[1] X. Xiao, Q. Luo, X. Xie, W.-Y. Ma, “A Comparative Study on Classifying the Functions of Web Page Blocks. In proceeding of: Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management”, **(2006)** November 6-11; Arlington, Virginia, USA

[2] A. Z. Wen and J. F. Xu, “The research on vision-based Web page information extraction algorithm”, Microcomputer & Its Applications, vol. 3, **(2010),** pp. 38-41.

[3] S. Brin and L. Page, "The anatomy of a Large-Scale Hypertextual Web Search Engine", In the proceedings of the 7th International World Wide Web Conference, **(1998)** April14, Brisbane, Australia.
[4] D. Cai, S. P. Yu and J. R. Wen, "VIPS: a vision-based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79, **(2003).**
[5] G. Le, J. Zhang and X. Tian, "Improvement and Implementation of VIPS algorithm", computer systems & applications, vol. 4, **(2009),** pp. 65-69.
[6] S.-H. Lin and J.-M. Ho, "Discovering Information Content Blocks from Web Documents, In the proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining", **(2002)** Edmonton, Canada.
[7] Kovacevic, M. Dilligenti, M. Gori and V. Milutinovi, "Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification", in the proceedings of 2002 IEEE International Conference on Data Mining, **(2002)** December, Maebashi City, Japan.
[8] G. Suhit, K. Gail, N. David and G. Peter, "DOM-based Content Extraction of HTML Documents", In the proceedings of the Twelfth World Wide Web conference,**(2003)** May, Budapest, Hungary.
[9] R. Song, H. Liu, J.-r. Wen and W.-Y. Ma, "Learing Block Importance Model for Web Pages", Proceeding of the 13th international conference on World Wide Web", **(2004)** New York, USA.

Authors

Lv Fang, she is a second-year graduate student from Harbin Institute of technology at Weihai, majoring in computer science and technology. Her research is mainly on information security and webpage analysis.

Junheng Huang, He is working for Harbin Institute of Technology (abstract as HIT) as an associate professor. His research is mainly on social network, data mining, artificial intelligence and bioinformatics.

Wei Yuliang, He has been currently involved in doing his Ph.D at Harbin Institute of Technology (abstract as HIT). His research is mainly on information security and network security.

Bailing Wang, He is working for Harbin Institute of Technology (abstract as HIT) as a professor. He got the Ph.D. degree from HIT in 2006. His research is mainly on information security, network security, parallel computing.

