A New Webpage Classification Model Based on Visual Information Using Gestalt Laws of Grouping

Zhen Xu(✉) and James Miller

Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada
{zxu3,jimm}@ualberta.ca

Abstract. Traditional text-based webpage classification fails to handle rich-information-embedded modern webpages. Current approaches regard webpages as either trees or images. However, the former only focuses on webpage structure, and the latter ignores internal connections among different webpage features. Therefore, they are not suitable for modern webpage classification. Hence, semantic-block trees are introduced as a new representation for webpages. They are constructed by extracting visual information from webpages, integrating the visual information into render-blocks, and merging render-blocks using the Gestalt laws of grouping. The block tree edit distance is then described to evaluate both structural and visual similarity of pages. Using this distance as a metric, a classification framework is proposed to classify webpages based upon their similarity.

Keywords: Webpage classification · Block tree · Gestalt laws of grouping · Normalized compression distance · Tree edit distance

1 Introduction

Webpage classification is becoming increasingly essential because it plays a substantial role in various information management and retrieval tasks, such as web data crawling and web document categorization [1]. Modern webpages, with much more abundant information, presents additional challenges to webpage classification [2]. Hence, traditional approaches that rely on text content cannot handle modern webpages. Nevertheless, people can get visual information directly. Specifically, people subconsciously follow the Gestalt laws of grouping for immediate content identification to perceive rich content [3]. Consequently, providing the machines with the visual features from webpages directly is a feasible way for them to "read and think" as people. Therefore, this paper proposes a methodology to evaluate webpage similarity by visual information using Gestalt laws of grouping, and classifies webpages in terms of their visual similarity.

J. Wang et al. (Eds.): WISE 2015, Part II, LNCS 9419, pp. 225–232, 2015.
DOI: 10.1007/978-3-319-26187-4_18

2 Related Work

To date, extensive work has been done on webpage classification [1]. In general, two major orientations are widely applied to explore webpage classification including treating webpages as images or trees.

In the first category, webpages are abstracted as images before computing their similarity. Recently, many scholars have focused their study on image similarity [4]. Liu et al. [5] proposed a feature-based image similarity measurement approach which uses image phase congruency measurements to compute the similarities between two images. Kwitt et al. [6] presented an image similarity model by using Kullback-Leibler divergences between complex wavelet sub band statistics for texture retrieval. Sampat et al. [7] put forward an image similarity method called the complex wavelet structural similarity. The theory behind it is that consistent phase changes in the local wavelet coefficients may arise owing to certain image distortions. Although image similarity techniques are very useful in searching for a similar image to the specified image, they are not suitable for webpage similarity assessment directly. This is because a specified webpage is an object embedded with a variety of elements and these elements can interact (such as overlap or partly overlap) with each other. It is, therefore, a different problem than pure image similarity assessment.

In the other category, a webpage is regarded as tree structured data. Thus, webpage similarity is studied through investigating tree similarity. With respect to tree structured data, a handful of tree distance functions are applied, such as tree edit distance [8], multisets distance [9], and entropy distance [10]. The tree edit distance is defined as the minimum cost of operations for transferring from one tree to another [11]. Tree edit distances can be further divided into different subcategories in terms of distinct mapping constraints. Mapping constraints include top-down, bottom-up, isolated subtree, etc. [12]. Müller-Molina et al. [9] propose a tree distance function with multisets, which are sets that allow repetitive elements. Based on multiset operations, they define a similarity measure for multisets. They did this by converting a tree into two multisets, with one multiset including complete subtrees and another consisting of all the nodes without children. Connor et al. [10] developed a bounded distance metric for comparing tree structures based on Shannon's entropy equations. Although the above achievements on tree similarity are significant, the theory cannot be used directly on webpage similarity research. The main reason is that the theme of tree similarity has always been structural similarity. However, our focus is on content similarity, in spite the obvious connection between structural and content similarity.

3 Render-Block Tree

Visual information of a webpage is retrieved and represented as the render-block tree by taking the webpage's DOM tree as a prototype instead of parsing sources code. This is because the DOM tree contains all information of a webpage, both textually and visually.

3.1 Render-Blocks

Each node of the render-block tree, i.e., a render-block, maps onto a DOM element. However, only visible DOM elements and their visible attributes are meaningful for analysis. Meanwhile, the semantic meaning of text in a webpage is not part of its visual features, so it is not considered. Hence, only text styles are of major concern. Properties of the render-block contain:

1. A render block always correlates to a DOM element;
2. A render block is always visible in the webpage;
3. A render block only contains visual features of corresponding DOM element.

The transformation from DOM elements to render-blocks only takes into account visible DOM elements, text content, and CSS attributes. The visibility of a DOM element is decided by its tag name, size, and styles. Elements with certain tag names, sizes, or styles are invisible, such as the elements with the tag name of `SCRIPT` or `TITLE`, width or height of 0, `display` style of `none`, etc. Most texts are displayed in a webpage, but some are not, such as texts of `IMG` elements. Visible CSS attributes refer to three sets of "front end" styles, namely, text styles (`font`, `color`, etc.), paragraph styles (`direction`, `list-style`, etc.), and background styles (`background-color`, `border-width`, etc.). On the contrary, the "back end" styles are not drawn by the browser, such as `margin`, `cursor`, etc. They are ignored during the transformation. Additionally, geometry information, such as top, left, width, and height, is kept during transformation. DOM elements keep their own offset positions inherited from their parent elements, but we convert them into absolute positions.

3.2 Tree Hierarchy

The render-block tree takes the DOM hierarchy as a prototype to illustrate the visual (render) layout. However, due to the flexibility of CSS, the DOM hierarchy sometimes is not consistent with the rendered layout. For example, a child DOM element by default overlaps its parent that is at the left top of the webpage, but a `float` command can move it onto a third element located at the right bottom. Therefore, in order to eliminate the inconsistency, the render-block tree hierarchy must be modified so that it always follows the rendered layout.

To construct a render-block tree, we manipulate nodes as follows:

1. Take the `BODY` render-block as the root node.
2. From the root node on, for every render-block, append all child render-blocks according to their corresponding DOM hierarchy.
3. If any render-block is completely located inside any of its sibling render-blocks, move it downward so that it becomes a child of that sibling. However, sibling nodes that geometrically overlap each other are acceptable in a render-block tree, and they are still considered as siblings.
4. If a parent DOM element is invisible or empty, then it has no corresponding render-block; however, its child DOM elements may have a block. In this condition, these child render-blocks shall become children of the render-block which is related to this parent DOM element's first visible parent element.

4 Semantic-Block Tree

The semantic-block tree shares the same hierarchy as the render-block tree. However, the nodes of this tree, i.e., the semantic-blocks, are achieved by merging semantically correlated render-blocks with Gestalt laws of grouping.

4.1 Interpreting Gestalt Law of Simplicity

Although the content of a DOM element can be further split, we do not split it in order to follow the Gestalt law of simplicity. As shown in Fig. 1a (the homepage of "google.ca"), the middle image above the search box contains multiple content (i.e., "GOOGLE" serves as the newspaper title and the three columns are utilized as texts, images, and animations, respectively), but when we read the whole webpage, we treat it as one large image instead of the aforementioned separated ones.

(a) (b)

Fig. 1. Homepage of "Google.ca" and "Twitter.com"

4.2 Interpreting Gestalt Law of Closure

It is evident that an upper render-block will cover a lower one visually, leading to "incomplete" display of the latter. In this case, however, the lower render-blocks are still perceived as complete because of the Gestalt law of closure. As shown in Fig. 1b (the homepage of "twitter.com"), the upper right part of the background image is covered by two log-in boxes, but the image is still regarded as a complete rectangle (although we cannot see what is exactly covered). That is, the render-block remains as a complete rectangle.

4.3 Interpreting Gestalt Law of Proximity

In webpages, the size of the render-blocks cannot be ignored. They are grouped by distance in the Gestalt law of proximity. To measure distances between two non-zero-area render-blocks, a normalized Hausdorff distance (NHD) is employed. Consider two render-blocks R_1 and R_2:

1. For any point r_1 in R_1 and r_2 in R_2, the distance between them is the length of the corresponding line segment:

$$\|r_1 - r_2\| = \sqrt{(x_{r_1} - x_{r_2})^2 + (y_{r_1} - y_{r_2})^2}; \tag{1}$$

2. For any point r_1 in R_1, the distance between itself and any point in R_2 is the infimum of distances between r_1 and all points in R_2:

$$d(r_1, R_2) = \inf_{r_1 \in R_2} \|r_1 - r_2\|; \tag{2}$$

3. The Hausdorff distance (HD) from R_1 to R_2 ($hd_{1,2}$) is the supremum of distances between all points in R_2 and all points in R_1:

$$hd_{1,2} = \sup_{r_1 \in R_1} d(r_1, R_2) = \sup_{r_1 \in R_1} \inf_{r_1 \in R_2} \|r_1 - r_2\|; \tag{3}$$

4. The Hausdorff distance [13] *between* R_1 *and* R_2 is the maximum value between the HD from R_1 to R_2 ($hd_{1,2}$) and the HD from R_2 to R_1 ($hd_{2,1}$):

$$HD(R_1, R_2) = \max\{hd_{1,2}, hd_{2,1}\}; \tag{4}$$

5. The normalized Hausdorff distance (NHD) is calculated by adding a normalizing factor f to HD:

$$NHD(R_1, R_2) = \max\left\{\frac{hd_{1,2}}{f_{R_1}}, \frac{hd_{2,1}}{f_{R_2}}\right\}. \tag{5}$$

The normalizing factor f can be the width, height, or diagonal distance of the render-block, depending on their relative position. As shown in Fig. 2, the surrounding region of R_0 is split by dashed lines. The normalizing factor f is calculated as: the height of R_2 (R_2 locates in the north/south region of R_0); the width of R_3 (R_3 locates in the west/east region of R_0); or the diagonal of R_4 (R_4 covers corner regions of R_0).

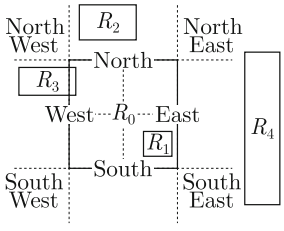

Fig. 2. NHD normalizing factor

4.4 Interpreting Gestalt Law of Similarity

The render-block similarity is divided into three parts: foreground similarity, background similarity, and size similarity. Due to most render-blocks being rectangles, shape similarity is not considered. Background similarity compares both the color and the image; foreground similarity includes textual and paragraph styles; and size similarity checks if the two render-bocks share the same width or height.

The CIE-Lab color space provides standard color difference. RGB colors are obtained directly from CSS and are translated into CIE-Lab colors [14].

Normalized compression distance (NCD) [15] is employed to calculate the similarity between two images x and y (in CIE-Lab color space) as shown in (6), where C calculates the compressed length of corresponding input.

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}. \tag{6}$$

4.5 Interpreting Reminder Terms of Gestalt Laws

The Gestalt law of common fate refers to the motion trend. In a group of render-blocks, if they are not placed on the same path, then the off-path render-blocks share no common fate with the other blocks. The Gestalt law of continuity is interpreted as alignment. If any of the four sides (left, right, top or bottom) of two render-blocks are aligned, then they are continuous.

The Gestalt law of symmetry tells that different but symmetrical render-blocks should be merged, however, there are very few webpages containing such instances. Hence, this law is not interpreted. Also, the Gestalt law of past experience is not considered because it refers to high level semantics and requires external knowledge.

5 Webpage Similarity Classification Model

Because a webpage can be ultimately represented by a semantic-block tree (or simply block tree), and each block contains all the visual information, visual similarity between two webpages can be reflected by block tree similarity. That is, visual similarity is evaluated by block tree edit distances.

5.1 Block Tree Edit Distance

Let T be a block tree, $|T|$ be the size of it, and t_i be its ith node. Two different block trees can then be denoted by T^p and T^q. The tree edit distance (TED) is then defined as the minimum cost of editing operations ("insert", "delete", and "relabel") when shifting from T^p to T^q [8]. This reflects the structural similarity between T^p and T^q by mapping node pairs.

Webpage similarity includes both structural and content similarity (visual similarity). To compare visual similarity between two webpages, a block tree edit

distance (B-TED) is introduced. By encoding the content of each block into its label, the mapping procedure in TED calculation compares the blocks by their visual information. Same as comparing background images in Sect. 4.4, visual similarity between two blocks are evaluated by NCD. If they are not similar, then a "relabel" operation is needed.

People always see a subtree of a block rather than itself because it is overlaid by its descendants (if there is any). To simulate this, the content for encoding shall not be that of the block itself but of the complete subtree. In B-TED calculation, it can be achieved simply by encoding the screen capture image of a block.

5.2 Classification Using a Naive Bayes Classifier

The model adopts a naive Bayes classifier for classification. Through reading the feature vectors of two webpages whose categories are known, the classifier learns the connections between the features and categories. The feature vector contains three components: the block trees of the two webpages, and the B-TED value between them. The category variable is a Boolean; and its value is either `T` indicating the two webpages are similar or `F` indicating they are different. Details of this model are illustrated in Fig. 3.

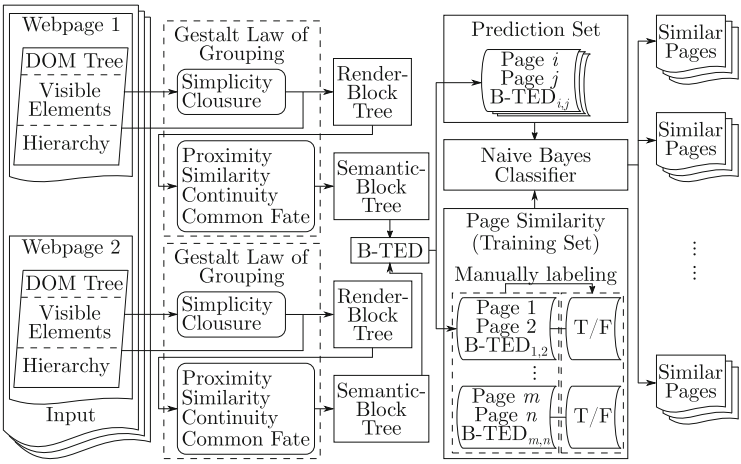

Fig. 3. Classification model

6 Conclusion

In this paper, a novel approach to evaluate webpage similarity is proposed. A render-block tree model is introduced to represent a webpage visually, and

a semantic-block tree model is then retrieved by interpreting and applying the Gestalt laws of grouping. During interpretation, a normalized Hausdorff distance is introduced to evaluate proximities; the CIE-Lab color space and its color difference are used to find color similarities; and the normalized compression distance is employed to calculate image similarity. A classification model is finally proposed to evaluate webpage similarity. Block tree edit distance can be applied to recognize both structural and visual similar-ity of webpages.

Acknowledgment. The authors give thanks to China Scholarship Council (CSC) for their financial support.

References

1. Qi, X., Davison, B.D.: Web page classification: features and algorithms. J. ACM **41**(2), 12:1–12:31 (2009)
2. Wei, Y., Wang, B., Liu, Y., Lv, F.: Research on webpage similarity computing technology based on visual blocks. SMP **2014**, 187–197 (2014)
3. Wertheimer, M.: Laws of organization in perceptual forms (1938)
4. Rohlfing, T.: Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans. Med. Imaging **31**(2), 153–163 (2012)
5. Liu, Z., Laganière, R.: Phase congruence measurement for image similarity assessment. Pattern Recogn. Lett. **28**(1), 166–172 (2007)
6. Kwitt, R., Uhl, A.: Image similarity measurement by kullback-leibler divergences between complex wavelet subband statistics for texture retrieval. ICIP 2008, pp. 933–936 (2008)
7. Sampat, M.P., Wang, Z., Gupta, S., Bovik, A.C., Markey, M.K.: Complex wavelet structural similarity: a new image similarity index. IEEE Trans. Image Process. **18**(11), 2385–2401 (2009)
8. Shahbazi, A., Miller, J.: Extended subtree: a new similarity function for tree structured data. IEEE Trans. Knowl. Data Eng. **26**(4), 864–877 (2014)
9. M"uller-Molina, A.J., Hirata, K., Shinohara, T.: A tree distance function based on multisets. In: Chawla, S., Washio, T., Minato, S., Tsumoto, S., Onoda, T., Yamada, S., Inokuchi, A. (eds.) PAKDD 2008. LNCS, vol. 5433, pp. 87–98. Springer, Heidelberg (2009)
10. Connor, R., Simeoni, F., Iakovos, M., Moss, R.: A bounded distance metric for comparing tree structure. Inf. Syst. **36**(4), 748–764 (2011)
11. Cording, P. H.: Algorithms for Web Scraping (2011). [PDF] http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/6183/pdf/imm6183.pdf
12. Zhai, Y., Liu, B.: Structured data extraction from the web based on partial tree alignment. IEEE Trans. Knowl. Data Eng. **18**(12), 1614–1628 (2006)
13. Chaudhuri, B.B., Rosenfeld, A.: A modified Hausdorff distance between fuzzy sets. Inf. Sci. **118**(1), 159–171 (1999)
14. Johnson, G.M., Fairchild, M.D.: A top down description of SCIELAB and CIEDE2000. Color Res. Appl. **28**(6), 425–435 (2003)
15. Li, M., Chen, X., Li, X., Ma, B., Vitányi, P.M.: The similarity metric. IEEE Trans. Inf. Theory **50**(12), 3250–3264 (2004)