Int. J. Human-Computer Studies 110 (2018) 95–114

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Layout-based computation of web page similarity ranks

Ahmet Selman Bozkir*, Ebru Akcapinar Sezer

Department of Computer Engineering, Hacettepe University, Ankara 06800, Turkey

ARTICLE INFO

Key words:
Web page layout
Layout similarity
Similarity ranking
Bag of features
Spatial pyramid matching
Histogram of oriented gradients

ABSTRACT

In this paper, we propose a ranking approach which considers visual similarities among web pages by using structure and vision-based features. Throughout the study, we aim to understand and represent the web page visual structure as in the way people do by focusing on the layout similarity through the wireframe design. The conducted study is composed of two parts. In the first part, structural similarities are analyzed with the proposed concept of "layout components" along with visual inspection of DOM trees. In this way, five types of structural layout components are proposed and revealed. Moreover, whitespaces are also utilized since they are important visual cues in the visual perception of web pages. In the second part, a computer-vision based method named histogram of oriented gradients (HOG) is employed to reveal local visual cues in terms of edge orientations. Following the feature extraction phases, extracted feature histograms are mapped on spatial information preserving multilevel and multi-resolution bag of features representation method named spatial pyramid matching. In this way, three goals were achieved: (1) the visual layout of web pages were mapped and compared in a multi-resolution schema; (2) the intermediate process of visual segmentation was removed; and (3) efficient and easily comparable web page layout signatures were generated. We also conducted a questionnaire study covering 312 subjects. This helped us to create a benchmark dataset involving similarity scores collected from individuals. So far, there exists no web page layout similarity ranking oriented corpus in the literature. Our suggested approach achieved a remarkable ranking performance at top-5 and top-10 retrieval results. According to the findings of the comparative study, our approach outperforms some structure and vision-based studies in the literature. With this achievement, web pages could be employed as a query item to find other, similar web pages by taking into consideration that they are web pages, instead of images or anything else.

1. Introduction

Definition and detection of similarities among web pages – the most widely used and ultimate information conveyors of the new era – has become an important issue since web-oriented information extraction, retrieval and mining have evolved. In fact, similarity searching in web pages may be handled within the scope of different domains (e.g. cognitive, psychology and computer sciences) with different point of views. To date, various properties affecting web page similarity have been addressed in the literature, including *textual content* (Tombros & Ali, 2005; Kleinberg, 1999), *link structure* (Tombros & Ali, 2005; Dehmer et al., 2006), *semantic content* (Joshi & Liu, 2009), *document structure* (Bartik, 2012) and *appearance/visual layout* (Hara et al., 2009; Bartik, 2012; Joshi & Liu, 2009; Kudelka et al., 2010; Alpuente & Romero, 2010; Zhang et al., 2013). Briefly, text-based similarity studies aim to seek content matching, whereas document structure-oriented studies explore the similarity of tags, links and DOM structure in HTML documents. Furthermore, visual layout-based approaches are based on identification and detection of partial or global visual similarities between web pages. Michailidou et al. (2008) have investigated the strong relation between the perception of complexity, aesthetic appearance and clarity regarding web pages and organization of structural layout. As a result, regarding the web pages, they found that layout directly affects the human perception in various forms. Therefore, it can be said that the layout of a web page constitutes an entry point towards the human perception thus constituting a utilizable source of information. As a result, visual layout itself has been subjected to scrutiny in many studies for different purposes, such as visual block importance learning (Song et al., 2004), page segmentation and information extraction (Kang & Choi, 2008; Cao et al., 2010), efficient web search (Cai et al., 2004) anti-phishing (Rosiello et al., 2007; Zhang et al., 2013), visual similarity comparison (Song, 2011) and efficient web page archiving (Law et al., 2012).

Web pages do not only function as information conveyors but they also constitute one of the essential and unignorable parts of corporate identity. It is a well-known fact that, well-decorated web design and the quality of a user-interface provide more credits in e-commerce and

* Corresponding author.
E-mail address: selman@cs.hacettepe.edu.tr (A.S. Bozkir).

https://doi.org/10.1016/j.ijhcs.2017.10.008
Received 13 September 2016; Received in revised form 18 October 2017; Accepted 24 October 2017
Available online 26 October 2017
1071-5819/

positive first influence (Möller et al., 2012; Robins & Holmes, 2008). Reinecke et al. (2013) point out the growing number studies related to web site aesthetics regarding to its economic effects. Likewise, the concerns related to the visual appeal of web pages (e.g. credibility and economic impact) have elevated the value of novelty and development of new ways of user interaction in web page design. However, this situation also led to an increase in layout plagiarism of web pages. For saving money and time, some website developers prefer to copy and paste the source codes of a target website or make reverse engineering in order to construct exact or similar visual layout of the original web page. Furthermore, competitors have a tendency to be inspired by or steal someone else's design if it is gaining success. Apart from inspiration, plagiarism provides an unlawful profit and damages corporate identity. It should also be noted that, design plagiarism is not always a conscious decision of web developers. Web page design guidelines generate a trend towards uniformity in design. Nevertheless, this situation may cause web page designers to be accused of plagiarism (Martine & Rugg, 2005). In order to cope with this problem, in member governments, the DMCA DMCA (2016) (Digital Millennium Copyright Act) provides a copyright protection for original design and content owners. Such that, in case of proven layout or content plagiarism, hosting companies become responsible to take down the illegal website.

At this point, the necessity of a layout-plagiarism detection mechanism comes into prominence. Likewise, without any notice, original design owners cannot know plagiarizer web sites. In this regard, a search engine which investigates and indexes the visual similarities among the web pages is required. Last decade has witnessed considerable amount of studies aiming to utilize detection of visual similarities between web pages for different purposes such as phishing detection or page archiving. Nonetheless, according to our best knowledge, there exists no study which attempts to retrieve visually similar web pages when a web page is queried. Consequently, the main target of this study is to develop an approach that ranks web pages according to their visual similarities by considering the concordance to the visual similarity judgment of human. Coupled with this idea, it is aimed to fill the gap and enable the web page layout itself to be used as a query item. Hence, without any necessity of notice, web page designers, companies and legal template owners will be able to check the plagiarizer's web pages by querying the genuine web pages.

Regarding the problems stated above, we suggest that research should be conducted along three lines. The first line of our work starts with the question of "how can we represent a web page in a way that is discriminative and in concordance with human perception?". In response to this question, we utilized the *wireframes* as a perceptional representation schema. Ramon et al. (2016) define the wireframes as a layout-oriented, initial design and rapid development environment. Note that, wireframes deal with how the content elements are spatially placed on the page regardless of colors that will be used. According to the Reinecke et al. (2013) and Michailidou et al. (2008), the visual layout has a greater impact on the human perception than color. Two web pages having similar or identical layout structures may completely differ in color. Therefore, in this study, we are dealing with visual similarity purely from the point of view of layout-based similarity. This led to the central idea of our proposal, which aims to understand and represent the visual structure of the web pages by revealing its wireframe design, since it is the place where the initial design was born. As a result, our approach has been designed in order to capture layout-based visual similarities without addressing the color-based features.

The second line of our proposal deals with the computational aspect of visual similarity between web pages. Most of the works in the literature leverage either (1) structural information located in DOM trees (Cai et al., 2003; Rosiello et al., 2007) or (2) vision-based features (Law et al., 2012) in order to extract representative features for visual comparison. Likewise, these two types of approaches have their positive and negative consequences. The former one has the comfort of direct access to HTML content, vision-based methods have found widespread usage, especially in visual similarity-based phishing detection studies since they have the ability to analyze the whole web page screenshot. However, vision-based methods may pose limitations to an accurate exploration of layout structure due to the fact that web pages involve more compositional differences than ordinary photographs (Reinecke et al., 2013). Therefore, we combined the strengths of both methodologies, as done in Law et al. (2012). Through structure-based analysis, we first rendered the web page due to the aforementioned difficulties and classified the leaf nodes along with whitespaces into five types of visual words (text, static images, animations, form elements and whitespaces) in a manner of bag-of-features (BoF) representation. In this way, we extracted the wireframe design of the web pages and revealed the CSS box-model according to exact layout appearance, regardless of any details such as color, text size and font face. In the vision-based part, we employed Histogram of Oriented Gradients descriptors (HOG) (Dalal & Triggs, 2005) to reveal the visual cues of web pages by considering the edge directions and intensity distributions of gradients on web page screenshots. With the use of HOG descriptors, we extracted orientation bins over page regions in order to use them as descriptive visual features. Following the feature extraction phases, we employed a special kind of BoF method named spatial pyramid match (SPM) (Lazebnik et al., 2006) which enables comparison of the similarity between two feature sets considering their spatial arrangements of embedded features. As stated by Lazebnik et al. (2006), SPM increasingly divides the entire 2D feature space into equally sized cells over multi-level pyramids and produces histograms of local features in each cell. These concepts are expanded upon in Section 3. With the use of SPM, extracted structure and vision-based features were embedded into separate SPM pyramids in order to generate efficient and easily comparable page layout signatures for structure and vision-based schemes, respectively. Further, layout based similarities have been investigated in different levels of detail, such as *blurred, overview, inspiration* and *exact match*.

The third line of our research deals with the evaluation of the proposed approach. According to the best of our knowledge, there exists no corpus or dataset that was built for ranking evaluation of page layout similarity related studies. Therefore, we designed a novel corpus containing 40 web pages in 4 groups and conducted a questionnaire study covering 312 participants to generate a ground truth dataset. Following the statistical significance tests, rankings of our proposed approach have been examined against the rankings based on average human perception acquired by the questionnaire study. In the next stage, achieved performance results were measured via average normalized rank (ANR) metric which considers relevancy and ranking.

As a consequence, the following contributions have been carried out in this paper:

- Besides other studies, we propose a novel approach for ranking web pages according to their layout similarity;
- Wireframing is suggested as a representation scheme for measuring layout similarity;
- It is being enabled to use the page layout as a query item for web page retrieval;
- By using this approach, the need of visual segmentation is removed;
- We generated a valid and statistically verified dataset via a questionnaire in order evaluate web page visual similarity oriented studies. This dataset can be further used for various purposes;
- Comprehensive experiments based on the generated dataset show that wireframe representation along with structural and visual features embedded in SPM are highly reasonable in web page visual similarity comparison due to their remarkable ranking performance.

The rest of this paper is organized as follows. Section 2 briefly introduces the related work on web page visual similarity-based studies. Section 3 presents the foundations of our methods. Section 4 details the proposed approach in all its aspects. Section 5 reports the experimental setup along with corpus generation and performance evaluation. Section 6 provides a discussion on obtained results and system parame-