Visual Structure–based Web Page Clustering and Retrieval

Paul Bohunsky
Database and Artificial Intelligence Group
Vienna University of Technology, Austria
bohunsky@dbai.tuwien.ac.at

Wolfgang Gatterbauer
Computer Science and Engineering
University of Washington, WA, USA
gatter@cs.washington.edu

ABSTRACT

Clustering and retrieval of web pages dominantly relies on analyzing either the *content* of individual web pages or the *link* structure between them. Some literature also suggests to use the *structure* of web pages, notably the structure of its DOM tree. However, little work considers the visual structure of web pages for clustering. In this paper (*i*) we motivate visual structure-based web page clustering and retrieval for a number of applications, (*ii*) we formalize a visual box model-based representation of web pages that supports new metrics of visual similarity, and (*iii*) we report on our current work on evaluating human perception of visual similarity of web pages and applying the learned visual similarity features to web page clustering and retrieval.

Categories and Subject Descriptors: H.3.3: Information Search and Retrieval

General Terms: Design, Theory, Human Factors

1. INTRODUCTION

Clustering and similarity search help targeted information retrieval by organizing and understanding available information. Various user independent similarity functions have been proposed in the context of web page clustering that fall into two main categories [2]: *content-based* metrics that compare textual content of web pages, and *link-based* metrics that analyze the hyperlink structure between web pages.

A third, less-known category explores *structure-based* similarity between web pages. The majority of the proposed methods compute structural similarity using features derived from HTML code or DOM tree representation of web pages [1, 5, 11]. Only little work has been done to compare web pages based on their *visual* structure [6], and those approaches often rely on image processing techniques based on screenshots of web pages [10]. From a web user's point of view, however, the visual structure of a web page is more discriminating than the structure of its source code: The fundamental reason is that the process of rendering a web page is a *non-injective, and hence "lossy" mapping* from a one-dimensional code fragment into a two-dimensional arrangement, where the same visual appearance can be generated by very distinct HTML code fragments. As a concrete example, any web table generated with a ⟨TABLE⟩ tag can be re-created to exact visual identity with a code fragment of ⟨DIV⟩ tags[1]. With ever more complex web pages and more available HTML options to create the same design, structural similarity as perceived by web users can be only reliably determined from a web page's visual rendering.

[1]See example: `http://gatterbauer.name/tables/DIV_table.html`

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

Applications. Three application scenarios will benefit from visual-based clustering of web pages: (*1*) *Content similarity.* Web pages that look similar have been shown to often address similar topics [10]. Hence, efficient visual similarity metrics add to existing ones and can improve current voting-based clustering algorithms. (*2*) *Preprocessing.* Clustering web pages with similar structures helps information extraction algorithms that specialize on certain kinds of data structures to automatically harvest knowledge from the Web. With an observed paradigm shift in search from whole web pages to smaller subunits [7], the granularity of visual structure analysis can be adapted to allow comparison at the web object level; (*3*) *Exploration.* Searching for web pages that look similar to a specified example are a new and fun way to explore the Web, similar to StumbleUpon [9]. As an analogy, imagine a hypothetical system that lets you search for people who do not have the same hobbies (content), are not related (links), and do not share the same genetical traits ("structure" of genotype), but who rather just look similar to you (visual appearance).

2. VISUAL REPRESENTATION

An important prerequisite for developing visual similarity metrics is a *structural representation* of the 2D visual appearance of web pages that captures the notion of visual appearance and that can be efficiently calculated on large scale. Similarity metrics on the DOM tree have the advantage of being efficient to calculate. In such approaches, the structure of a web page P is commonly represented by its *DOM tree* $T = \langle N, E\rangle$ where $N = N_t \cup N_e$ with N_t representing the set of *text nodes*, N_e the set of *element nodes* and $E \subset N_e \times N$ the directed edges. Each text node $n_t \in N_t$ contains one non-empty string, $n_t = \langle s\rangle$, and each element node $n_e \in N_e$ a non-empty list of property-value attributes with an obligatory label declaration, $n_e = \langle \mathbf{a}\rangle$.

Visual box model. In contrast, we represent a web page's structure by a *2D visual box model* $C_b = \langle V, X_E\rangle$ where $V = V_w \cup V_e$ with V_w representing the set of *visualized words*, V_e the set of *visualized element nodes* (*VENs*) and X_E is a minimum double topological cell grid superimposed on the four coordinates $\mathbf{x} = \langle x_1, x_2, y_1, y_2\rangle$ of all VENs as rendered by a chosen rendering algorithm or browser b [3, 4]. Each VEN $v_e \in V_e$ contains coordinates together with a feature-value list of computed style attributes $\mathbf{a} = \langle a_1, ..., a_{|\mathbf{a}|}\rangle$: $v_e = \langle \mathbf{x}, \mathbf{a}\rangle$. Each visualized word $v_w \in V_w$ additionally contains one non-empty string s: $v_w = \langle \mathbf{x}, \mathbf{a}, s\rangle$.

Visual edit distance. Given a bipartite matching between the nodes of two web pages, we define visual edit distance as the weighted sum of (*i*) *adjacency and alignment violations*, e.g. two boxes adjacent in one page are not adjacent in the other; (*ii*) *transformations of box groupings*, e.g. a group of adjacent nodes is scaled between the two

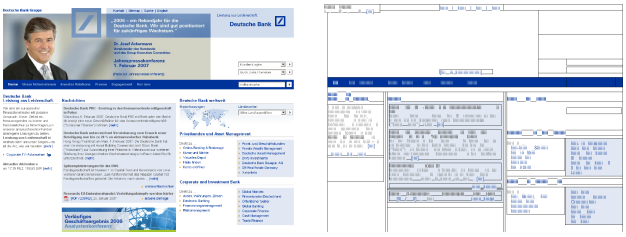

Figure 1: Left: Screenshot of an example web page. Right: Its visual-box representation in VisBox.

pages; and (*iii*) *missed matches*, i.e. the number of boxes that remain unmatched. The visual edit distance is then the minimum distance over all Bipartite Matchings (BM): $\min_{BM} \sum_{j \in V \cup T \cup M} w_{t(j)}$, where $t(j)$ is the type of Violation, Transformation or Missed match, and $w_{t(j)}$ its weight.

The big innovation of a box-based visual structure representation over computer vision approaches relying on *image processing* is the explicitness by which we can reference and compare individual characteristics of web pages and the resulting facility in constructing similarity measures. The advantage over DOM tree-based similarity measures is that visual similarity actually measures the perceived similarity by the user in contrast to HTML code or DOM tree similarity (see example from footnote above). However, our current focus and main *technical challenge* is the computational complexity of our similarity measure; we need better approximation heuristics to make our measure tractable.

3. CURRENT AND FUTURE WORK

VisBox. We have been implementing VisBox, as extension to the Firefox browser that can be controlled from Java via XPCOM bindings. Figure 1 shows an example web page and the rendered visual structure representation of VisBox. In this case we replaced the actual content with pseudo words, notably the letter *i*. We are currently developing adaptations of our strict model that allow to capture the essence of visual structure in a simplified model focusing on the visual appearance, not the content.

Page selection. The dominant *conceptual challenge* of our approach is that many web pages gain a considerable fraction of their visual appearance from structures contained in images, which we cannot explicitly model. Comparing to image processing approaches, this aspect is largely compensated given that the accuracy of image processing and, particularly, object detection in images is neither reliable nor fast. An approach that works well on a subset of web pages can be supplemented with plain image analysis in a hybrid approach at a later stage. For this reason, we currently focus on web pages for which images cover a small fraction of the overall web page area. For evaluation we build upon WebPageDump [8], a tool we previously developed to evaluate diverse web archiving approaches: We calculate the byte difference between screenshots of the original web pages and of our visual structure representation, and retain those pages for which the difference is smaller than a cut-off threshold:

$$\frac{\text{web page area covered by images}}{\text{total web page area}} \sim \text{byte difference} < \text{threshold} \ .$$

Those web pages are converted into our simplified visual representation for further analysis (Fig. 1).

4. ULTIMATE GOAL: HUMAN PERCEPTION OF VISUAL SIMILARITY

One ultimate goal of our work is to study and evaluate human perception of visual web page similarity [12, 14]. Whereas our current focus is on developing efficient algorithms to calculate visual box edit distances (see above), we have a clear experimental setup in mind.

Building upon the set of previously selected pages whose screenshots are well represented by our visual structure, we implement an asymmetric version of a two player verification game similar to the ESP game [13]. In each cycle, player 1 sees one screenshot as input and chooses one out of five other screenshots which she seems most similar to the input. Player 2 sees this last screenshot as input and chooses the most similar among the other five screenshots. If player 2 chooses the original input to player 1, they agreed on the visual similarity between these two images and they earn a lot of points. Preventing possible forms of collusion, these two pieces of evidence from both directions and from two distinct individuals is strong evidence of mutual agreement of visual similarity of two among six web pages. This game is repeated over many cycles and at each cycle the set of six web pages is chosen using collaborative filtering over the previously learned features. Our goal with this game and implied test method is to get a good impression of what people consider as visually similar web pages, expressed in our visual structure representation. Finally, the structure of the game will also allow us to evaluate the predictive nature of our learned similarity features by inserting web pages into the game and comparing our prediction of most similar documents with the actual choice of human players.

Acknowledgement. Research supported in part by a DOC scholarship from the Austrian Academy of Sciences, by FIT-IT contract FFG 813567 from the Austrian Federal Ministry for Transport, Innovation and Technology, by project grant 819563 from the Austrian Forschungsförderungsgesellschaft FFG, and by NSF grant IIS-0915054.

5. REFERENCES

[1] V. Crescenzi, P. Merialdo, and P. Missier. Clustering web pages based on their structure. *Data Knowl. Eng*, 54(3):279–299, 2005.

[2] D. Dhyani, W. K. Ng, and S. S. Bhowmick. A survey of web metrics. *ACM Comput. Surv*, 34(4):469–503, 2002.

[3] W. Gatterbauer and P. Bohunsky. Table extraction using spatial reasoning on the CSS2 visual box model. In *Proc. 21st AAAI*, pp. 1313–1318. AAAI Press, July 2006.

[4] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krüpl, and B. Pollak. Towards domain independent information extraction from web tables. In *Proc. 16th WWW*, pp. 71–80. ACM, May 2007.

[5] S. Joshi, N. Agrawal, R. Krishnapuram, and S. Negi. A bag of paths model for measuring structural similarity in web documents. In *Proc. 9th SIGKDD*, pp. 577–582. ACM, Aug. 2003.

[6] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic. Recognition of common areas in a web page using visual information: a possible application in a page classification. In *Proc. 2nd ICDM*, pp. 250–257. IEEE, Dec. 2002.

[7] Z. Nie, Y. Ma, S. Shi, J.-R. Wen, and W.-Y. Ma. Web object retrieval. In *Proc. 16th WWW*. ACM, May 2007.

[8] B. Pollak and W. Gatterbauer. Creating permanent test sets of web pages for information extraction research. In *Proc. 33rd SOFSEM: Theory and Practice of Computer Science*, vol. II, pp. 103–115, Jan. 2007.

[9] StumbleUpon. http://www.stumbleupon.com.

[10] Y. Takama and N. Mitsuhashi. Visual similarity comparison for web page retrieval. In *Proc. 4th Intern. Conf. on Web Intelligence (WI)*, pp. 301–304. IEEE, Sept. 2005.

[11] A. Tombros and Z. Ali. Factors affecting web page similarity. In *Proc. 27th ECIR*, pp. 487–501. Springer, Mar. 2005.

[12] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[13] L. von Ahn and L. Dabbish. Designing games with a purpose. *Commun. ACM*, 51(8):58–67, 2008.

[14] A. L. Zobrist and W. B. Thompson. Building a distance function for gestalt grouping. *IEEE Trans. Computer*, 24(7):718–728, July 1975.