Visual Structure–based Web Page Clustering and Retrieval

Paul Bohunsky[1] and Wolfgang Gatterbauer[2]

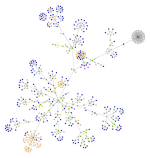
1: Database and Artificial Intelligence Group
Vienna University of Technology, Austria
bohunsky@dbai.tuwien.ac.at

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

Three existing and a new approach to clustering and navigating the Web

There are currently three dominant approaches to clustering and example-based web page retrieval. We propose a 4th approach.

1. Link structure



Link graph

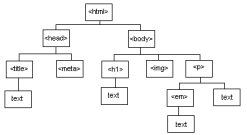
2. Content

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

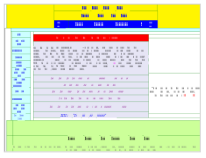
tf-idf

3. Document structure



DOM tree

4. Visual structure



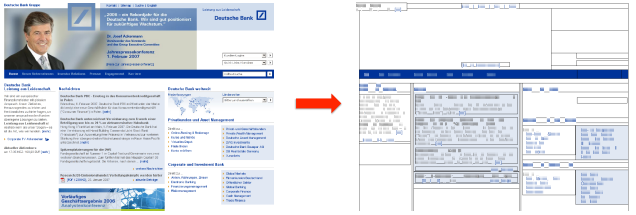
Visual box model

Why a new 4th approach?

- User's perspective: For a web user, the visual appearance is more discriminating than the document structure, e.g. TABLE vs DIV tables: http://gatterbauer.name/tables/DIV_table.html
- Similar appearance hints at similar content (language independent?)
- Preprocessing: for automated information extraction approaches targeting visual structures
- Exploration: Looking for visually similar web pages is a fun and new way to explore the Web.

How to visually represent a web page?

Remove pictures, replace text, use the visual box model:

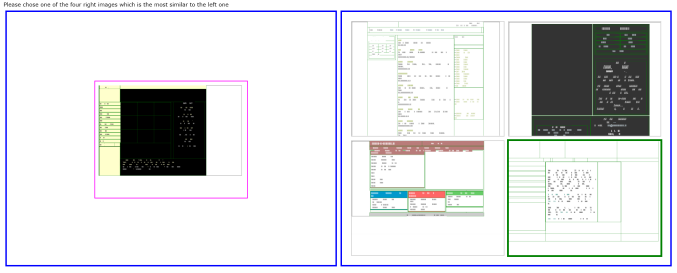


How to evaluate and measure human perception of visual similarity?

Consider an asymmetric version of a two player verification game (similar to ESP):
- Player 1 sees one input image (the visual box model) and chooses one of four other images she considers most similar to the input
- Player 2 sees the chosen image and chooses among the other four images the most similar

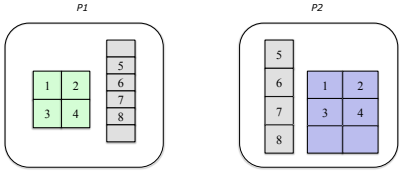
Goals: (*i*) to learn human judgment of visual similarity, (*ii*) training and evaluation set for our visual edit distance

For a simple demo, visit: http://pbit.at/vsc/



How to measure visual edit distance?

Consider a bipartite matching (*BM*) between the visual box models of two web pages [indicated by numbers]



Calculate the visual edit distance as the weighted sum of all differences (*j*):
- adjacency and alignment violations (*V*) [e.g. 4 and 8 adjacent in P1, not P2]
- transformations of box groupings (*T*) [e.g. dimension and color of group 1-4]
- missed matches (*M*) [e.g. boxes without matches and numbers]

$\sum_j w_{t(j)}$

j ... violation, transform or miss
$t(j)$ ... type of difference
$w_{t(j)}$ ... weight of this difference type

Find the bipartite matching with the minimal visual edit distance

$$\min_{BM} \sum_{j \in V \cup T \cup M} w_{t(j)}$$

Two caveats: (*i*) hard metrics, (*ii*) learning of weights required

Key points

1. We propose a 4th and yet unexplored approach to clustering and navigating the "Visual Web" (contrast with the Linked, Syntactic or Semantic Web): Our focus is on the visual appearance of web pages.
2. We propose to capture the visual appearance of a web page by removing images, replacing text with repeated letters and focusing on the visual box model, and define a general visual edit distance between two visual box models.
3. We propose a two player verification game whose purpose is to learn the feature weights of our similarity measure.

Open point: can human perception of visual similarity be captured simply in terms of low level visual features? Or do we need to build a measure that reasons in terms of more abstract intermediate concepts?

Reference: Paul Bohunsky and Wolfgang Gatterbauer. Visual Structure-based Web Page Clustering and Retrieval. In *Poster proceedings WWW 2010*.

Links (also in the paper):
Visual similarity comparison demo: http://pbit.at/vsc/
Example DIV table: http://gatterbauer.name/tables/DIV_table.html

Partially supported by a DOC scholarship from the Austrian Academy of Sciences, by FIT-IT contract FFG 813567 from the Austrian Federal Ministry of Transport Innovation and Technology, by grant 819563 from the Austrian Forschungsfoerderungsgesellschaft FFG, and by NSF grant IIS-0915054.