Block-o-Matic: a Web Page Segmentation Algorithm and its Evaluation

Andrés Sanoja **Stéphane Gançarski**

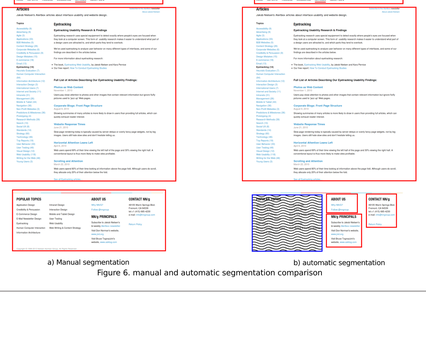
Laboratoire d'Informatique de Paris 6
Université Pierre et Marie Curie
4 place Jussieu, 75005
Paris, France

Abstract

In this paper we present our prototype for the web page segmentation called Block-o-matic (BoM) and its counterpart Manual-design of Blocks (MoB), for manual segmentation. The main idea is to evaluate the correctness of the segmentation algorithm. Build a ground truth database for evaluation can take days or months depending on the collection size, however we address our solution with our manual segmentation tool intended to minimize the time of ground truth generation. Both tools implements the same rules for segmentation, for the manual version allows to propose candidates blocks to assessor and for the automatic the block selection. We present our demonstration scenario with a collection of web pages organized in categories. After its annotation they are compared with the automatic segmentation version and it is given a score and a visual comparison.

Inportance of Evaluation

Search an analitical and general way to evaluate web page segmentation for those heuristic based algorithms



a) Manual segmentation b) automatic segmentation

Figure 6. manual and automatic segmentation comparison

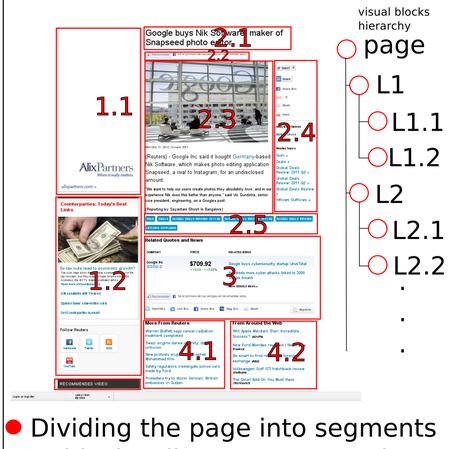
Why to Segment a Web Page?

Web Page segmentation is a technique used for dividing a web page into particular parts, not overlapping, called segments or blocks.

Describing the page as a unit

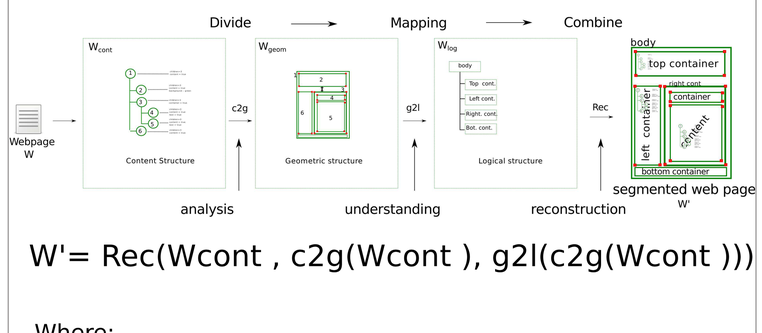
- Describe the content of a web page as a whole can lead to a not so relevant result : information not related to main content are included in the description
- Noisy content is included in the analysis

Describe the page based on detected visual blocks



- Dividing the page into segments or blocks allows to treat each one as a « subpage », most important blocks being representative of the whole page
- Noisy content can be excluded
- Blocks 2.1 and 2.2 are the most importants in the above example

Segmentation Model



$$W' = Rec(Wcont , c2g(Wcont), g2l(c2g(Wcont)))$$

Where:

W is the original Web page
W' is the segmented web page

Wcont is the content structure (DOM & CSS properties)
Wgeo is the geometric structure
Wlog is the logic structure

c2g function extracts the geometric structure (W geom) from the content structure (Wcont).

g2l function maps the geometric structure (Wgeom) into a logical structure (Wlog)

Rec is a function that builds the page (W') which is the original document (W) enriched with all structure, into an linear document in a XML-like style.

Ground Truth Construction Model

Let define the manual segmentation of a web page W as:

```
Wgeom = c2g(Wcont )
forall go in Wgeom
   createNewLogicObjectFrom(go)

then, |Wgeo| = |Wlog|
```

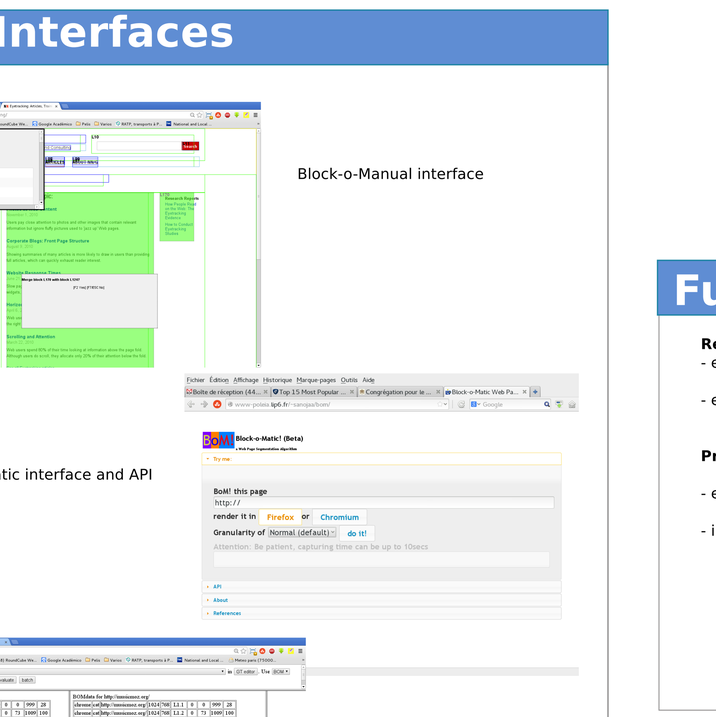
The user can perfom the same actions used by the g2l function, those are:

- Select a logic block
- Select the parent of a logic block
- Delete a logic block
- Add a logic block
- Modify the attributes of a logic block
- Merge two logic blocks

The result of annotation is a Block Record (BR) as:

- browser
- url
- document width
- document height
- block x
- block y
- block width
- block height

General Process



Segmentation Evaluation Model

G is the ground truth segmentation
P is the proposed segmentation

We constructs a bipartite graph with G and P as set of nodes

wt is the tolerance parameter

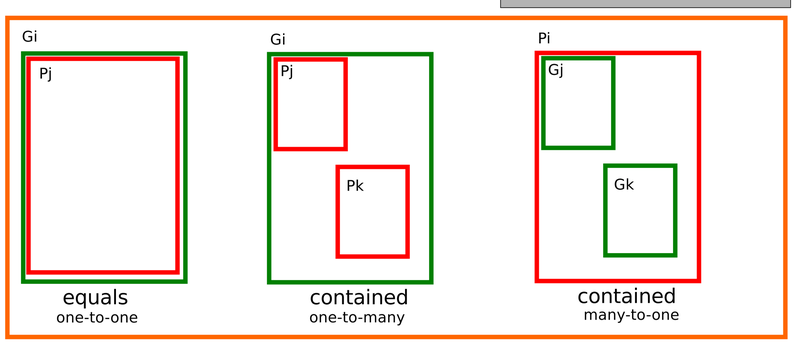
Each pair of node (one in G and other in P) can have:
- perfect matching
- oversegmented matching (P nodes contained in G nodes)
- undersegmented matching (G nodes contained in P nodes)

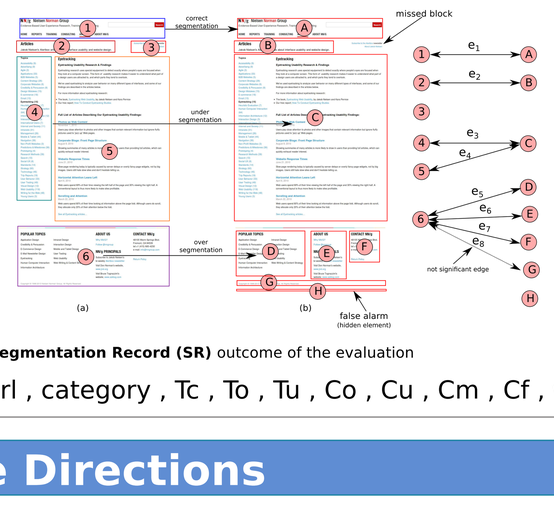
Metrics:

Tc : total correct blocks (one-to-one matching)
To : total oversegmented blocks (1-to-many mathing)
Tu : total undersegmented blocks (many-to-1 matching)
Co : oversegmented blocks (in G)
Cu : undersegmented blocks (in P)
Cm : missed blocks (G nodes with no arc to P)
Cf : false alarms (P nodes with no arc to G)

Prec = Tc / |G|

Score = Prec / (Cm + Cf)





Segmentation Record (SR) outcome of the evaluation

url , category , Tc , To , Tu , Co , Cu , Cm , Cf , wt , |G| , |P|

Technical Specifications

User Interfaces

Block-o-Manual interface

Block-o-Matic interface and API

Evaluation enviroment

Future Directions

Research
- enhace content structure for better detection of geometric objects

- enhance text processing in web pages

Practical issues

- enhace usability of manual tool

- integrate the three tools as browsers extensions

References

E. Bruno, N. Faessel, H. Glotin, J. Le Maitre, and M. Scholl. Indexing by permeability in block structured web pages. In Proc. of the 9th ACM symposium on Document engineering, DocEng '09, pages 70–73, New York, NY, USA, 2009. ACM.

D. Cai, S. Yu, J.-R. Wen, and W.-Y. Ma. Extracting content structure for web pages based on visual representation. In Fifth Asia Pacific Web Conference (APWeb'03), 2003.

Shafait, F. ; Image Understanding & Pattern Recognition Res. Group, German Res. Center for Artificial Intell., Kaiserslautern ; Keysers, D. ; Breuel, T.M.

M. B. Saad and S. Gançarski. Using visual pages analysis for optimizing web archiving. In Proc. of the 2010 EDBT/ICDT Workshops, EDBT '10, pages 43:1–43:7, New York, NY, USA, 2010. ACM.