ABSTRACT

The technique presented in this report is a automatic top-down, tag-tree independent approach to detect the content structure of a given web page using its visual cues. It simulates how a user understands web layout structure based on his visual perception.

The whole process of content structure extraction is mainly consisting of three steps.

1) Visual block extraction
2) Separator detection
3) Content structure construction