Abstract

A web page usually contains various types of content such as navigation, decorations, and other parts that are not associated with the core information from these web pages. On the other hand, sometimes the user actually requires only core information from these pages. From this came the need for a system that can extract information from a web page.

Users see a web page through a web browser and get a 2D representation that have a lot of visual cues to help distinguish different parts of the page. Web designers usually organize content from a web page so that it is easy to understood by the user. Therefore, the content-related content semantically usually placed in one group and web pages are divided into regions for different content using a visual differentiator such as line, font size, color, etc. Same type content would normally be displayed with a similar visual form as well. These visual cues will be used for identification and data extraction processes.Visual Based Page Segmentation Method will use visual cues from the web page to extract data from these web pages.

Phase analysis and test results provide proves that appropriate pattern of visual cues can be used to create a system of information extraction from web pages although there is still some noises.

Keywords: web page, extraction, visual cues, pattern, noise.