# Informative Content Extraction for Web Page using Text Density and Vision-based Page Segmentation (VIPS) Algorithm Integration

**Ei Phyu Phyu Mon, Yuzana**
*University of Computer Studies, Yangon*
*eiphyuphyumon@ucsy.edu.mm,*
*yuzana.yzn@gmail.com*

## Abstract

*Web pages consist of not only actual content, but also other elements such as branding banners, navigational elements, advertisements, copyright etc.Irrelevant content in the Web page is treated as noisy content. This noisy content is typically not related to the main subjects of the webpages. A method is necessary to extract the informative content and discard the noisy content from Web pages. This system is used an integration of textual and visual importance features to extract the informative contents from Web pages. Initially a web page is converted into Document Object Model (DOM) tree. For each node in the DOM tree, textual and visual importance is calculated. Textual importance and visual importance is combined to form hybriddensity.DensitySumis calculated and used in content extraction algorithm to extract the informative content from Web pages. The algorithm is tested with various web domains and styles of web pages. Performance of web content extraction is obtained by calculating precision and recall.*