Abstract

A page of website usually contain a lot of information content in each of information block that shown.Sometimes , news content shown in the page of website not purely giving relevent information with the core of news e.g., navigation bars, copyright, user guide, links, synopsis and also advertisement. The information blocks that is not the main content or irrelevant information in web pages is called *web pages noise.*

On this final project, *Style Tree* technique will be used to get general presentation style (layout) and actual content from web pages using pages sampling. First, web pages will be modelled with DOM tree , then building Style Tree with join DOM structure to capture the common structure and spliting information block in a website. The Information that is captured will be used to measure and evaluate the importance of each node until giving a weight to each feature in each content block. The weighting result will be used to measure the performance of *preprocessing* by *classification* to get *F-measure* score.

Keyword: *Style Tree, DOM*, *weighting, identification noise, web mining.*