International Journal of Computer Science and Engineering (IJCSE)
ISSN(P): 2278-9960; ISSN(E): 2278-9979
Vol. 3, Issue 2, Mar 2014, 27-34
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

# CONTENT EXTRACTION USING DOCUMENT OBJECT MODEL AND NATURAL LANGUAGE PROCESSING FOR WEB

**AAKRITI AGARWAL, SHAILEY CHHEDA, KRIMA SHAH & MEERA NARVEKAR**

Department of Computer Engineering, Mumbai University, DJSCOE, Mumbai, Maharashtra, India

## ABSTRACT

Web pages often contain clutter such as pop-up advertisements, unnecessary images and extraneous links around the body of an article that distract a user from actual content and may reduce effects of many advanced web applications. Often this noisy content is combined with the main content leaving no clean boundaries between them. This noisy content as a result makes the problem of information harvesting from web pages much harder. Most approaches to removing clutter or making content more readable involve changing font size or removing HTML which takes away from a webpage its inherent look. Unlike 'Content Reformatting', which aims to recreate the webpage in a more convenient form, our solution directly addresses 'Content Extraction', an approach that does not require previous knowledge of website templates. For higher accuracy in content extraction, the analyzing software needs to act as a human user and understand content in natural language along with HTML DOM analysis in order to eliminate noisy content. In this paper, a combination of HTML DOM analysis and Natural Language Processing (NLP) techniques for automated extractions of main article of interest with associated images from web pages has been described**.**

**KEYWORDS:** Content Extraction, DOM, NLP