International J.of Multidispl.Research & Advcs. in Engg.(IJMRAE),
ISSN 0975-7074, Vol. 2, No. II, July 2010, pp. 117-122

DETECTION OF DUPLICATE AND NEAR DUPLICATE WEB DOCUMENTS

M. KIRUTHIKA

Abstract

Very large amount of web documents are swarming the web making the search engines less appropriate to the users. These web documents may have many duplicates and near duplicates i.e. variants derived from the same original web document due to which additional overheads are created for search engines by which their performance and quality is significantly affected. In this paper, mainly the problems faced due to duplicates is discussed. Also, existing approaches which has been tried to solve the above problem have been highlighted. In the end, a discussion about the need to address this problem is mentioned.

© Ascent Publication House: http: //www.ascent-journals.com