International J. of Engg. Research & Indu. Appls. (IJERIA).
ISSN 0974-1518, Vol.2, No.I (2009), pp 131-138

SET THEORETICAL APPROACH FOR MINING WEB CONTENT THROUGH OUTLIERS DETECTION

G. POONKUZHALI, K. THIAGARAJAN AND K. SARUKESI

Abstract

Today, a primary challenge in information retrieval and web mining is to discover interesting data from web documents. Thus, developing user-friendly and automated tools for organizing and retrieving required information from the web documents has been on a higher demand. Most of the existing web mining algorithms have concentrated on finding frequent patterns while neglecting the less frequent ones that are likely to contain the outlying data. This paper refers to outliers present on the web as web outliers to distinguish them from traditional outliers. This paper proposes new algorithm for mining web content using clustering technique and mathematical set formulae such as subset, union, intersection etc for detecting outliers. Then the outlying data is removed from the original web content to get the required web content by the user. Also, the removal of outliers improves the quality of the results from the search page.

Keywords: clustering, web contents, web mining, web outliers