International J. of Engg. Research & Indu. Appls. (IJERIA).
ISSN 0974-1518, Vol.3, No. II (May 2010), pp 199-210

SEMANTIC CLUSTERING WITH FEATURE SELECTION FOR TEXT DOCUMENTS

M. THANGAMANI AND P. THANGARAJ

Abstract

Text documents are the unstructured databases that contain raw data collection. The clustering techniques are used group up the text documents with reference to its similarity. The feature selection techniques are used to improve the efficiency and accuracy of clustering process. The feature selection is done by eliminate the redundant and irrelevant items from the text document contents. Statistical methods are used in the text clustering and feature selection algorithm. The cube size is very high and accuracy is low in the term based text clustering and feature selection method. The semantic clustering and feature selection method is proposed to improve the clustering and feature selection mechanism with semantic relations of the text documents. The proposed system is designed to identify the semantic relations using the ontology. The ontology is used to represent the term and concept relationship. The synonym, meronym and hypernym relationships are represented in the ontology. The concept weights are estimated with reference to the ontology. The concept weight is used for the clustering process.

The system is implemented in two methods. They are term clustering with feature selection and semantic clustering with feature selection. The performance analysis is carried out with the term clustering and semantic clustering methods. The accuracy and efficiency factors are analyzed in the performance analysis.

Keywords: Clustering, Text mining, Ontology, Feature selection