Abstract

This paper presents a novel technique of document clustering based on frequent concepts. The proposed technique, FCDC (Frequent Concepts based document clustering), a clustering algorithm works with frequent concepts rather than frequent items used in traditional text mining techniques. Many well known clustering algorithms deal with documents as bag of words and ignore the important relationships between words like synonyms. The proposed FCDC algorithm utilizes the semantic relationship between words to create concepts. It exploits the WordNet ontology in turn to create low dimensional feature vector which allows us to develop a efficient clustering algorithm. It uses a hierarchical approach to cluster text documents having common concepts. FCDC found more accurate, scalable and effective when compared with existing clustering algorithms like Bisecting K-means, UPGMA and FIHC.

Reference


CherryPicker Coreference resolution Tool. Available at http://www.hlt.utdallas.edu/~altaf/cherrypicker.html

Index Terms

Computer Science  Data Mining

Key words

Document clustering  Clustering

algorithm

Frequent Concepts based Clustering

WordNet