World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:05, 2016

Using Closed Frequent Itemsets for Hierarchical Document Clustering

Authors : Cheng-Jhe Lee, Chiun-Chieh Hsu

Abstract : Due to the rapid development of the Internet and the increased availability of digital documents, the excessive information on the Internet has led to information overflow problem. In order to solve these problems for effective information retrieval, document clustering in text mining becomes a popular research topic. Clustering is the unsupervised classification of data items into groups without the need of training data. Many conventional document clustering methods perform inefficiently for large document collections because they were originally designed for relational database. Therefore they are impractical in real-world document clustering and require special handling for high dimensionality and high volume. We propose the FIHC (Frequent Itemset-based Hierarchical Clustering) method, which is a hierarchical clustering method developed for document clustering, where the intuition of FIHC is that there exist some common words for each cluster. FIHC uses such words to cluster documents and builds hierarchical topic tree. In this paper, we combine FIHC algorithm with ontology to solve the semantic problem and mine the meaning behind the words in documents. Furthermore, we use the closed frequent itemsets instead of only use frequent itemsets, which increases efficiency and scalability. The experimental results show that our method is more accurate than those of well-known document clustering algorithms.

Keywords : FIHC, documents clustering, ontology, closed frequent itemset
Conference Title : ICISET 2016 : International Conference on Information Science, Engineering and Technology
Conference Location : Rome, Italy
Conference Dates : May 02-03, 2016

Open Science Index, Computer and Information Engineering Vol:10, No:05, 2016 publications.waset.org/abstracts/41381.pdf