Abstract

Due to tremendous increase in number of documents, clustering of such document is difficult one. Document Clustering is the process of grouping related documents from the large collection of database. The mining of such related documents from the database which are unlabelled is a challenging one. To overcome this process, clustering is used to filter the unlabelled documents from the large collection of database. In this paper, a new concept is introduced for the document clustering by using k-means Enhanced Approach algorithm [1] with the Dictionary Defined Lexical Analyzer (DDLA). Basically K-Mean algorithm clusters the numeric values efficiently. But with the inclusion of DDLA the characters, words and sentences can also be clustered. Based on the weights, documents are clustered [7] by using bisecting k-means algorithm [1, 2] and topic detection method. The discovery of meaningful labels for the document is based on semantic similarity [8]. The efficient clustering of unlabeled documents with enhanced K-Mean algorithm and DDLA is one of the techniques which make clustering in an easiest way.

References
Effective K-Means Document Clustering using Dictionary Defined Lexical Analyzer (DDLA)

- Korean Text Extraction by Local Color Quantization and K-means Clustering, Anh-Nga Lai*, KeonHee Park, Manoj Kumar, GueeSang Lee*Department of Computer Science, Chonnam National University, 500-757 Gwangju, Korea ltanhnga@gmail.com, gslee@chonnam.ac.kr
- Fast Document Clustering Based on Weighted Comparative Advantage, Jie Ji Intelligent System Lab The University of Aizu Aizuwakamatsu, Fukushima, Japan d8102102@u-aizu.ac.jp
- A Comparison of Document Clustering Techniques, Michael Steinbach, George Karypis. Department of Computer Science University of Minnesota Technical Report #00-034 steinbac, karypis, kumar@cs.umn.edu Vinin Kumar
- Document Clustering in Correlation Similarity Measure Space, Taiping Zhang; Yuan Yan Tang; Bin Fang; Yong Xiang Knowledge and Data Engineering, IEEE Transactions on Volume: 24 ,,2012

Index Terms

Computer Science Information Sciences
Keywords

Clustering  K-Means Enhanced Approach Algorithm  Lexical Analyzer  Defined Dictionary DDLA.