Abstract

Continuous expansion of digital libraries and online news, the huge amount of text documents is existing on the web. Consequently the need is to organize them. Text Categorization is an active analysis field can be used for organizing text document. Text categorization is the process of assigning documents with predefined categories that are associated with their contented.

CAWP algorithm is designed for Text Categorization. But this algorithm does not present the best results for large datasets. K-means Clustering with Support Vector Machine approach is used to enhance the results. K-means group the data into a number of clusters follow which it uses as training samples for Support Vector Machine in each cluster to divide the new sample data efficiently. The experiment performed on 20Newsgroups dataset, K-means with SVM provides better results than CAWP algorithm in terms of F-measure.
A Novel Text Categorization Approach based on K-means and Support Vector Machine


Index Terms

Computer Science  Software Engineering

Keywords

Classification, K-means, SVM, Document Categorization, Text mining