Abstract

Currently being developed clustering techniques, such as techniques using single-word representation of items, representing a text document as a "bag of words" in which a document is seen as a set of words. In this representation there is no order between words or sentences are considered because each word is considered stand alone without any connection to one another so not exactly in the cluster labeling results.

The above problems can be handled using Clustering Based On Frequent Word Sequences (CFWS). High dimensional data can be addressed by reducing the terms that are not frequent.Labeling of clusters was done by tracing "word sequences" in each document.

The result of this clustering algorithm is visualized in the form of a hierarchical tree. According to the experiments, clusters generated by the algorithm CFWS has represented cluster description quality news content.

Keywords: clustering, frequent word sequences, CFWS, F-Measure, purity.