Abstract

The large amount of documents which must be handled needs automatic organizing. A popular approach to clustering documents is the vector space model, which represents texts, usually generated from the set of terms contained in the documents.

The clustering based on the document-term frequency matrix suffers from noise caused by the frequent use of different words with similar meanings. These semantic relations (like synonyms) need to be handled.

The method described in this final project uses Latent Semantic Indexing (LSI) technique combined with double clustering to reduce the dimension of the vector space. In this way, the clustering is performed in a space with fewer dimensions and reduced noise.

K-means is a simply cluster analysis methodology. It can’t capture noise. Using K-means, which is added with LSI and double clustering technique, noise can be captured and reduced. When noise is reduces, the performance like purity, recall, precision and F-measure of *K-means* can be increased.

Keywords: clustering, Latent Semantic Indexing (LSI), K-means