Abstract

In a term based clustering technique with the vector space model, the issue of high dimensional vector space due to the number of words used always appears. This causes the clustering performance drops because the distance among the points tends to have the same value. The reduction of dimension by decreasing the number of words can be done by stemming. Stemming was used as term selection to reduce the many terms generated on preprocessing. The utilization of algorithm of enhance confix stripping stemmer reduced the terms that must be processed of 199.358 terms resulted from 108 text documents, became 5.476 terms result of the stemming. This reduction would speed up the process and saved the storage media. The evaluation by utilizing clustering was done using confusion matrix. The accuracy of experiment increased.

References

1. Sharma, D.: Improved stemming approach used for text processing in information
retrieval system. Master of Engineering in Computer Science & Engineering, Thapar University, Patiala (2012)


Index Terms

Computer Science
Algorithms
Keywords

Stemming, Clustering, Confusion Matrix, Enhance Confix Stripping Stemming