# Clustering bilingual documents using various clustering linkages coupled with different proximity measurement techniques

## Abstract

With the rich data on the web, a documents clustering task for monolingual documents is insufficient in order to produce an efficient information retrieval system. A Multilingual Document Clustering (MDC) had been introduced and it is one of the most popular trends in the area of natural language processing (NLP). In this paper, the effects of applying different clustering linkages coupled with different proximity measurements on the clustering bilingual Malay-English documents in parallel are investigated. A Hierarchical Agglomerative Clustering (HAC) has been implemented and applied in clustering bilingual Malay-English documents. Several different linkages are used in the HAC method that includes Single, Complete, Centroid and Average linkages. Not only that, the cosine similarity and the extend Jaccard coefficient are also applied in order to investigate a proper proximity measurement that can be coupled with the different type of clustering linkages used for clustering bilingual news articles written in English and Malay. The HAC method coupled with the average linkage can be considered to produce reasonable clustering results even though the average DBI is a bit high. Now only that, the study also shows that the extend Jaccard coefficient proximity measurement can produce a better clustering results compared to the cosine similarity.