DEVELOPMENT OF A PARALLEL CLUSTERING OF BILINGUAL CORPORA BASED ON REDUCED TERMS

LEOW CHING LEONG

THESIS SUBMITTED IN FULFILLMENT OF THE DEGREE OF MASTER OF SCIENCE

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2015

ABSTRACT

Document clustering is a process that groups a set of documents based on their similarities. There are several studies related to document clustering. However, with the current technology, clustering bilingual text documents provides more benefits to users. There are several advantages when clustering bilingual corpus. It helps in verifying the classification and constraints of languages. Other than that, it also helps in eliminating the biased language-specific usages. However, not many works conducted that are related to clustering bilingual documents found, especially for Malay text articles. The quality of clustering bilingual text documents is highly influenced by the quality of the bag-of-word presentation of Malay text articles presented to the clustering algorithm. Hence, the aim of this study is to investigate the effects of reducing terms used in clustering bilingual text articles in English and Malay on the quality of clustering results. 500 news articles for both languages are retrieved manually from Bernama archieve and TheStar website. In order to achieve this, there are three outlined objectives. The first objective of this study is to improve the stemming process for Malay language by increasing the efficiency of stemming Malay words. By improving this stemming process (0.5% error rate), the number of terms is also reduced and increases the quality of clustering results. The bag-of-word representation for Malay documents can also be improved by identifying the entities found in the text articles. By identifying the named-entity that exists in the Malay text articles, a better bag of words representation of text articles can be obtained by reducing the terms based on the named-entity recognition. The F-Measure obtain is 94.72%. Next, the second objective of this paper is to design an experimental setup that studies the effects of using different clustering linkages coupled with various proximity measurement techniques in clustering bilingual documents on the quality of clustering results. The clustering linkages include the single, complete, average and centroid linkages and the proximity measurement techniques include the cosine similarity and extend Jaccard. Based on the findings obtained, the average linkage shows ideal clustering results compared to the other clustering linkages even though the single linkage shows a lower Davies-Bouldin Index (DBI) value. This is because the standard deviation of the number of documents for all clusters is low. Not only that, this study also shows that the extend Jaccard coefficient produces a better clustering results compared to the cosine similarity. Finally, the third objective of this study is to investigate the effects of reducing the set of terms considered in clustering English and Malay documents. A Genetic Algorithm (GA) will be implemented to reduce the number of terms used. A set of relevant terms will be selected based on the GA based terms selection process. The parallel mapping percentages show an improvement when the number of terms reduced using the GA with different mutation rate.