UMS
UNIVERSITI MALAYSIA SABAH

# DEVELOPMENT OF A GENETIC-BASED HIERARCHICAL AGGLOMERATIVE CLUSTERING TECHNIQUE FOR PARALLEL CLUSTERING OF BILINGUAL CORPORA BASED ON REDUCED TERMS

FINAL REPORT

GRANT NO: FRG0225-TK-1/2010

RAYNER ALFRED, JASON TEO, CHUNG SENG KHEAU

# SYNOPSIS

In this project, we report on our work on applying Hierarchical Agglomerative Clustering (HAC) to a large corpus of documents where each appears both in Malay and English. We cluster these documents for each language and compare the results both with respect to the content of clusters produced. On the data available, the results of clustering one language resemble the other, provided the number of clusters required is relatively small. Further, we study the effects of changing the method used to compute the inter-clusters distance that includes single link, complete link and average link distance between clusters. Finally, we describe an experiment employing a genetic algorithm to fine-tune the individual term weights in order to reproduce more closely a predefined set of clusters.