AN INTELLIGENT CATEGORIZATION TOOL FOR MALAY RESEARCH ARTICLES

GRANT NO: RAG0008-TK-2012

Mohd Norhisham Bin Razali, Rayner Alfred and Chin Kim On

Final Report

Faculty of Computing and Informatics
Universiti Malaysia Sabah
Malaysia

2015

AN INTELLIGENT CATEGORIZATION TOOL FOR MALAY RESEARCH ARTICLES

Abstract

Unlabeled research articles published in Malay language are becoming increasingly common and available in Malaysia. Thus, the task of manually indexing these research articles is difficult and time consuming. In order to facilitate research activities that depend on research resources written in Malay language, these research articles must be categorized or indexed efficiently so that appropriate and relevant domains of knowledge can be recommended to researchers in Malaysia. There are not many researches conducted to efficiently categorize Malay research articles. The task of categorizing Malay research articles is more complex compared to the task of categorizing English research articles due to the complexity of Malay language and thus categorizing Malay research articles represents a major contemporary challenge. Malay text documents are often represented as high-dimensional and sparse vectors, by using Malay words as features, which consist of a few thousand dimensions and a sparsity of 95 to 99% is typical. Determining the appropriate number of categories for large amount of Malay documents is also difficult and time consuming task due to the sparsity of the documents. Related documents may be grouped into different clusters, if there are too many number of categories assigned to these documents. On the other hand, unrelated documents may be clustered into the same cluster, if there are too few number of categories assigned to these documents. This research addresses issues that involve improving several pre-processing processes that affect the performance of the clustering process. These pre-processing processes include stemming, part-of-speech tagging and named-entity recognition. In this work, the effects of improving all these pre-processing processes will be investigated. It is anticipated that by improving the clustering results, it will also improve the mapping of Malay and English clusters obtained from the bilingual clustering. Hence, by increasing the mapping percentage for the bilingual clusters, a more robust clustering algorithm can be developed for clustering bilingual documents. As a result, by increasing the mapping percentage for the bilingual clusters, a more robust clustering algorithm can be developed for clustering bilingual documents. In this study, a genetic algorithm (GA) is also proposed to be implemented in order to determine the set of terms that can be used in clustering bilingual documents with more effective.