UMS

UNIVERSITI MALAYSIA SABAH

ENHANCING DOCUMENT CLUSTERING BY INTEGRATING SEMANTIC BACKGROUND KNOWLEDGE AND SYNTACTIC FEATURES INTO THE BAG OF WORDS REPRESENTATION

FINAL REPORT

GRANT NO: SLB0010-TK-1/2011

RAYNER ALFRED, SURAYA ALIAS DAN ASNI TAHIR

SYNOPSIS

The basic Bag of Words (BOW) representation generally used in text documents clustering or categorization loses important syntactic and semantic information contained in the documents. When the texts contain a lot of stop words or when they are of a short length this may be particularly problematic. In this research, we study the contribution of incorporating syntactic features [and semantic background knowledge into the representation in clustering texts corpus. We investigate the quality of clusters produced when incorporating syntactic and semantic information into the representation of text documents by analyzing the internal structure of the cluster using the Davies-Bouldin index (DBI). In this research, we compare the quality of the clusters produced when four different sets of text representation used to cluster texts corpus. These text representations include the standard BOW representation, the standard BOW representation integrated with syntactic features, the standard BOW representation integrated with semantic background knowledge and finally the standard BOW representation integrated with both syntactic features and semantic background knowledge. This research helps the understanding on how the quality of documents clustering can be improved by enriching the classic bag of words representation with additional background information.