Enrichment of BOW Representation with Syntactic and Semantic Background Knowledge

Abstrak

The basic Bag of Words (BOW) representation, that is generally used in text documents clustering or categorization, loses important syntactic and semantic information contained in the documents. When the text document contains a lot of stop words or when they are of a short length this may be particularly problematic. In this paper, we study the contribution of incorporating syntactic features and semantic 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 (DBI) index. This paper studies and compares 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. Based on the experimental results, it is shown that the quality of clusters produced is improved by integrating the semantic and syntactic information into the standard bag of words representation of texts corpus.