Abstract

In this paper we are proposing a text clustering method based on a well-known Word Sense Disambiguation (WSD) algorithm, the Lesk algorithm, to classify textual data by doing highly accurate Word Sense Disambiguation. The clustering of text data is thus primarily based on the context or meaning of the words used for clustering. The Lesk algorithm is used to return the sense identifiers for the words used to classify the text files by looking up the senses of a word in a Knowledge-Base similar to the English WordNet (enriched with more informative columns or fields for each synset [synonym set] of the English WordNet database), so as to greatly increase the chances of contextual overlap, thereby resulting in high accuracy of proper sense or context identification of the words. The proposed scheme has been tested on a number of heterogeneous text document datasets. The clustering results and accuracies, obtained using the proposed scheme, have been compared with the results obtained using the K-means clustering algorithm on the Vector Space Models generated for all the heterogeneous textual datasets. Experimental results show that our algorithm performs much better than the Vector Space Model (VSM) and K-means based approach. The technique will thus help the users much better in searching for meaningful contextual information from a highly diversified datasets.
collection of textual information, which is a key task of the information overload problem.

References


Index Terms

Computer Science

Artificial Intelligence

Keywords
Text Clustering using a WordNet-based Knowledge-Base and the Lesk Algorithm

Lesk Synset Wsd Knowledge Base K-means Vector Space Model Context Wordnet