Automatic Document Classification /Categorization by Subject using Clustering Techniques (k-means algorithm)

This article reports the findings of an empirical study about Automated Text Clustering applied to scientific articles and newspaper texts in Brazilian Portuguese, the objective was to find the most effective computational method able to cluster the input of texts in their original groups. The study covered four experiments, each experiment had four procedures: 1. Corpus Selections (a set of texts is selected for clustering),2. Word Class Selections (Nouns, Verbs and Adjectives are chosen from each text by using specific algorithms), 3.Filtering Algorithms (a set of terms is selected from the results of the preview stage, a semantic weight is also inserted for each term and an index is generated for each text), 4. Clustering Algorithms (the clustering algorithms Simple K-Means, sIB and EM are applied to the indexes). After those procedures, clustering correctness and clustering time statistical results were collected. The sIB clustering algorithm is the best choice for both scientific and newspaper corpus, under the condition that the sIB clustering algorithm asks for the number of clusters as input before running (for the newspaper corpus, 68.9% correctness in 1 minute and for the scientific corpus, 77.8% correctness in 1 minute). The EM clustering algorithm additionally guesses the number of clusters without user intervention, but its best case is less than 53% correctness. Considering the experiments carried out, the results of human text classification and automated clustering are distant; it was also observed that the clustering correctness results vary according to the number of input texts and their topics.

Text documents are one of the richest sources of data for businesses: whether in the shape of customer support tickets, emails, technical documents, user reviews or news articles, they all contain valuable information. Document classification or document categorization is a problem in information science and computer science. The task is to assign a document to one or more classes or categories, which can be done algorithmically using varying AI/Machine-Learning algorithms, including Clustering Techniques.