Probabilistic confidence prediction in document clustering

Kristóf Csorba and István Vajk

In this paper we propose a novel technique to predict the confidence of the clustering of a document before performing the clustering procedure itself. If an application has time limitations for the document clustering and it has to cluster as much documents as possible, it can use the predicted confidence value to sort the documents and process the ones with high expected confidence first. This means that the system begins with the documents for which the probability of a certain result is high. If it has enough time, it can process the remaining documents as well, but only after the most beneficial ones. The base method consists of three key techniques of the feature-space based document clustering: term frequency quantizing, singular value decomposition and double clustering.

Term frequency quantizing removes the unnecessary variation of the term frequencies. This step was introduced through the idea, that the exact occurrence number of a frequent term in a document does not provide more information than a "frequent" status.

Singular value decomposition was already shown to be capable to capture semantic relationships between terms based on occurrence behavior similarities. SVD is used to reduce the feature space dimensionality before the clustering itself.

Double clustering means, that instead of clustering the feature vectors of the documents, the feature vectors of the individual terms are clustered to create termclusters first. Document clustering is then performed in a feature-space based on these term-clusters.

After these steps the document clustering is performed in the resulting feature-space by employing the k-means algorithm using the cosine distance measure. During the clustering beside the assigned cluster an additional confidence value is provided for every document to mark ambiguous cases.

Before the procedure above a *probabilistic model* of it can be employed to calculate the expected confidence value. Before employing, this *prediction method* uses an initial training phase: after executing the original clustering method on a training set, the internal state is analyzed to retrieve probability of internal decisions. After the training the predictions for the further documents are calculated with a single matrix multiplication. This prediction is faster, but therefore less accurate. The predicted results can be used to *reorder the documents* to process the easier cases first which helps timecritical applications to get more sure results in the available time.

In this paper we provide a description of the prediction method and experimental results in connection with absolute performance and about the correlation between the predicted and measured results.