Classifying the Fields of Subjects Using Case-Based Reasoning

Hnin Yu Wai, Khin Aye Than
Computer University (Dawei)
Snowbabylove21@gmail.com

Abstract

Today text classification is a necessity due the very large amount of text documents that we have to deal with daily. Text classification is a task of assigning a text document into classes. In this thesis, the system will be implemented to classify the fields of subjects using case-based reasoning. This system includes two phases, training phase and classification phase. In these two phases, the system will perform the preprocessing step such as tokenize the document into individual word, remove the stop words and stemming the words as their root words (features). In training phase, the system uses Term-frequency – Inverse Document Frequency (TF/IDF) method to calculate the weight of terms (words) in the document. This weight is statistical measure which is used to evaluate how important a word in a collected document. In classification phase, the system uses the K-Nearest Neighbor Algorithm (K-NN) to classify the new document as the appropriate fields. K-NN algorithm will retrieve the similar case in the case base by applying Euclidean distance measure. Thus, the system will classify the new document as the appropriate fields based on the retrieve case.