Abstract

Term weighting is one of the very important step in text mining. This step is applied in order to give a value/weight to terms contained in a document. The weight given to a term depends on the method that is used for the weighting. Most of the existing weighting methods work under the assumption that the whole data set is available and static, this fact significantly limits the use of these schemes in applications where continuous data streams must be analyzed in real-time.

One way to handle this problem is by using TF-ICF term weighting method, it does not require term frequency information from other documents within the set and thus, it can process document streams in linear time.

In this final exam, some term weighting methods like TF-ICF, TF-IDF, and ATC are compared each other by seeing the output of the text classification performance with dynamic document stream. Some parameters that will be used as a measurement for comparing the text classification performance are F-measure, ROC and weighting time. In order to test the output of the weighting result, C4.5 used as its classifier.

Based on the test result, can be concluded that TF-ICF can produce classification results that are comparable quality as TF-IDF and ATC, and it significantly faster than those method.

Keywords: *term, term weighting, dynamic document stream, TF-ICF, TF-IDF, ATC.*