World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:08, 2015

A Quantitative Evaluation of Text Feature Selection Methods

Authors : B. S. Harish, M. B. Revanasiddappa

Abstract : Due to rapid growth of text documents in digital form, automated text classification has become an important research in the last two decades. The major challenge of text document representations are high dimension, sparsity, volume and semantics. Since the terms are only features that can be found in documents, selection of good terms (features) plays an very important role. In text classification, feature selection is a strategy that can be used to improve classification effectiveness, computational efficiency and accuracy. In this paper, we present a quantitative analysis of most widely used feature selection (FS) methods, viz. Term Frequency-Inverse Document Frequency (tf□idf), Mutual Information (MI), Information Gain (IG), CHISquare (□x2), Term Frequency-Relevance Frequency (tf□rf), Term Strength (TS), Ambiguity Measure (AM) and Symbolic Feature Selection (SFS) to classify text documents. We evaluated all the feature selection methods on standard datasets like 20 Newsgroups, 4 University dataset and Reuters-21578.

Keywords : classifiers, feature selection, text classification
Conference Title : ICADMA 2015 : International Conference on Advanced Data Mining and Applications
Conference Location : Barcelona, Spain
Conference Dates : August 17-18, 2015