MODERN HIGH-EFFICIENCY AUTOMATIC CATEGORIZATION ALGORITHM FOR ARABIC CORPUS

ABD ELZIZ HASSAN KHARSANI[1], SAMANI A. TALAB[2] & AWAD H. ALI[3]

[1]Department of Computer Science, Salman Bin Abdulaziz University, Saudi Arabia

[2,3]Department of Computer Science, University of Neelain, Sudan

ABSTRACT

Text categorization is the process of classifying documents into a predefined set of categories based on the content of the documents. Many studies have discussed the topic, but in principle there are many obstacles to automatizing the categorization process. This paper describes a hybrid commercial preprocessing stemming algorithm to improve the accuracy of the stemming method. The effectiveness of different methods in the Arabic text categorization process is evaluated, and the most suitable methods for the Arabic language are chosen. We achieve improvements of about 96% in the classification process by human expert evaluation. This tool is found to be essential in automatizing the categorization process, because Arabic versions are needed for development and usage purposes.

KEYWORDS: Machine Learning, Stemming Approaches, Bilateral Names, Stop Words, The Root, Document Indexing, The Classifier, Text Categorization