Abstract

Text categorization is the process of classifying documents into a predefined set of categories based on its contents of keywords. Text classification is an extended type of text categorization where the text is further categorized into sub-categories. Many algorithms have been proposed and implemented to solve the problem of English text categorization and classification. However, few studies have been carried out for categorizing and classifying Arabic text. Compared to English, the Arabic text classification is considered as a very challenging due to the Arabic language complex linguistic structure and its highly derivational nature where morphology plays a very important role. This paper proposes a new method for Arabic text classification in which a document is compared with pre-defined documents categories based on its contents using the TF. IDF method (Term Frequency times Inverse Document Frequency) measure, then the document is classified into the appropriate sub-category using Chi Square measure.

References

- A. Farghaly, and K. Shaalan, "Arabic Natural Language Processing: Challenges..."
Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements


Index Terms

Computer Science
Text Processing
Keywords

Text Categorization  Text Classification  Term Frequency  Inverse Document Frequency  Chi Square.