Abstract

Clustering is an important task gives good results with information retrieval (IR), it aims to automatically put similar documents in one cluster. Stemming is an important technique, used as feature selection to reduce many redundant features have the same root in root-based stemming and have the same syntactical form in light stemming. Stemming has many advantages it reduces the size of document and increases processing speed and used in many applications as information retrieval (IR). In this paper, we have evaluated stemming techniques in clustering of Arabic language documents and determined the most efficient in pre-processing of Arabic language, which is more complex than most other languages. Evaluation used three stemming techniques: root-based Stemming, light Stemming and without stemming. K-means, one of famous and widely clustering algorithm, is applied for clustering. Evaluation depends on recall, precision and F-measure methods. From experiments, results show that light stemming achieved best results in terms of recall, precision and F-measure when compared with others stemming.
Stemming Effectiveness in Clustering of Arabic Documents

- Illhoi Yoo, Xiaohua H. "Semantic Text Mining and its Application in Biomedical Domain." 2006
- Sawaf H, Zaplo J. and Ney H. "Statistical Classification Methods for Arabic News Articles;" Presented at the Arabic Natural Language Processing Workshop; 2001 July; Toulonse, France.
- Elkourdi M, Bensaid A, Rachidi T. "Automatic Arabic Document Categorization Based on the Naive Bayes Algorithm;" Proceedings of COLING 20th Workshop on Computational Approaches to Arabic Script-based Languages; 2004 Aug; Geneva, Switzerland.

- Duwairi R. "A Distance-based Classifier for Arabic Text Categorization;" International Conference on Data Mining (DMIN05); 2005 Jun; Las Vegas, USA.


Index Terms

Computer Science

Pattern Recognition

Keywords

Arabic text clustering Stemming light stemming K-means