Awajan A. 2015. Keyword Extraction from Arabic Documents using Term Equivalence classes. ACM Transactions on Asian and Low-Resource Language Information Processing. Volume 14, Issue 2, Article 7 (March 2015), 18 pages.

Abstract

The rapid growth of the Internet and other computing facilities in recent years has resulted in the creation of a large amount of text in electronic form, which has increased the interest in and importance of different automatic text processing applications, including keyword extraction and term indexing. Although keywords are very useful for many applications, most documents available online are not provided with keywords. We describe a method for extracting keywords from Arabic documents. This method identifies the keywords by combining linguistics and statistical analysis of the text without using prior knowledge from its domain or information from any related corpus. The text is preprocessed to extract the main linguistic information, such as the roots and morphological patterns of derivative words. A cleaning phase is then applied to eliminate the meaningless words from the text. The most frequent terms are clustered into equivalence classes in which the derivative words generated from the same root and the non-derivative words generated from the same stem are placed together, and their count is accumulated. A vector space model is then used to capture the most frequent N-gram in the text. Experiments carried out using a real-world dataset show that the proposed method achieves good results with an average precision of 31% and average recall of 53% when tested against manually assigned keywords.