This paper reports on work performed to investigate the use of a combined Part of Speech (POS) tagging and a minimum edit operations algorithm to determine the level of similarity between pairs of Arabic text documents. The level of similarity can be used as an indication of duplication in full or in part of the document's content. Text is first converted into POS tags that are then fed to the string similarity algorithm to determine the similarity of pairs of documents. A normalized score is calculated and used to rank documents. Documents ranked higher than some selected threshold are considered similar and can be near or complete duplicate. The performed experiments compare results based on the use of a set of selected common subsequences that are the results of translation of text into a sequence of syntactical units. The strings are first produced using full-text (FULL). These are further refined to produce a REDUCED; where repeated consecutive characters are reduced to a single character and a number, and more refined to produce a UNIQUE string; where all repeating characters are replaced by a single character. Syntactical features of the text were used as a structural representation of the documents' content. Results obtained from the experiments using the
FULL, the REDUCED and the UNIQUE POS-strings showed a clear advantage over the use of the plain text in terms of reduced string size while maintaining the same discrimination power. In particular the unique (most-reduced) string has shown quite comparable results to the reduced, the full and the actual text string.

References


24. Mohamed Elhadi and Amjad Al-Tobi Webpage Duplicate Detection Using Combined 2009 World Congress on Computer Science and Information Engineering (CSIE 2009), March 31 - April 2, 2009, Los Angeles/Anaheim, USA.


Index Terms

Computer Science

Pattern Recognition

Keywords

Arabic text processing; syntactical structures; document similarity; reduction; edit-based string similarity; copy detection.