Abstract

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

1. Grune, D, and M, Huntjens, Detecting copied submissions in computer science
workshops, Vakgroep Informatica, faculteit Wiskunde & Informatica, Vrije Universiteit,
AMSTERDAM, 1989.
3. Clough, P., Old and new challenges in automatic plagiarism detection, Department of
Information Studies, University of Sheffield, 2003.
Software Report, Computer Assisted Assessment Centre, University of Luton, Luton, UK.
5. Kang, N., A. Gelbukh and S. Han, PPChecker: Plagiarism Pattern Checker in Document
7. Poinçot, P., S. Lesteven and F. Murtagh, Comparison of Two “Document Similarity
Algorithms”, In String Processing and Information Retrieval, 7th. International Symposium on,
9. S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, “Basic Local Alignment
10. I. Yang, C. Huang and K. Chao, “A fast algorithm for computing a longest common
249-253.
13. Y. Liu and L. Liang, “A Dual-method Model for Copy Detection”, IEEE, IAT Workshops,
Detection Techniques”, Intern. Conference on Computational Science, Amsterdam, Holand,
Calculation Using the Multilingual Thesaurus EUROVOC, Springer-Verlag Berlin Heidelberg,
2002.
March 2015.


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.


26. Mohamed Elhadi and Amjad Al-Tobi, Part of Speech (POS) Tag Sets Reduction and Analysis using Rough Set Techniques, Twelfth International Conference on Rough Sets, Fuzzy Sets, Data Mining & Granular Computing RSFDGrC 2009, Indian Institute of Technology, Delhi, India, December 16-18, 2009


Index Terms

Computer Science  
Pattern Recognition

Keywords

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