Jordan University of Science and Technology

Design and implementation of automatic indexing for information retrieval with Arabic documents

Authors: Ismail Hmeidi1, Ghassan Kanaan2and Martha Evens2,*

Abstract: Abstract We have put together a corpus of 242 abstracts of Arabic documents using the Proceedings of the Saudi Arabian National Conferences as a source. All these abstracts involve computer science and information systems. We also designed and built an automatic information retrieval system from scratch to handle Arabic data. The system was implemented in the C language using the GCC compiler and runs on IBM/PCs and compatible microcomputers. We have implemented both automatic and manual indexing techniques for this corpus. A long series of experiments using measures of recall and precision has demonstrated that automatic indexing is at least as effective as manual indexing and more effective in some cases. Since automatic indexing is both cheaper and faster, our results suggest that we can achieve a wider coverage of the literature with less money and produce as good results as with manual indexing. We have also compared the retrieval results using words as index terms versus stems and roots, and confirmed the results obtained by Al-Kharashi and Abu-Salem with smaller corpora that root indexing is more effective than word indexing. ? 1997 John Wiley & Sons, Inc.