3737107 SCCS/M : MAJOR : COMPUTER SCIENCE : M.Sc. (COMPUTER SCIENCE)
KEY WORDS : DOCUMENT / RETRIEVE / ALIGNMENT / RELEVANT / ENGLISH
PRACHYA YODPRASIT : DOCUMENT ALIGNMENT WITH THESAURUS-LIKE DICTIONARY WORDLIST. THESIS ADVISOR : DAMRAS WONGSAWANG, Ph.D., SUKANYA PHONGSUPHAP, Ph.D. 64p. ISBN 974-664-488-2

Document Alignment is one of the most useful searching tools in information retrieval system. Part or whole of a document can be used as an input text enquiry key to perform searching for relevant documents in the collection instead of user specified key words. Neither manually extracting significant keywords from a document nor looking for other keywords having the same semantics is needed to form a query. Document clustering is one of the most important applications of such a kind of searching. The main problem of document alignment is the automatic significant keyword extraction. Many approaches and methods have been currently proposed to get as many keywords as possible and as much relevancy as possible. However, the efficiency of the retrieval in terms of precision and recall is still not good enough for some special types of documents. In this research we proposed the keywords extracting and filtering scheme, called DAEDW (Document Alignment with English Dictionary Wordlist), for document alignment by using a thesaurus-like dictionary. After stopwords were filtered out, they were put into two groups; dictionary words and non-dictionary words. The non-dictionary words are usually appeared to be some specific names which had a higher significance of keywords. We also took advantage of the thesaurus-like dictionary to get synonymous keywords. A collection of documents consisted of news and articles which were simulated to test with the proposed alignment scheme. We found that DAEDW could increase the efficiency of retrieval in both precision and recall. Furthermore, the document ranking output was also improved. This thesis presented DAEDW in details including analysis and the computer implementation. Experimental results are discussed and future developments are also suggested.