3937009 SCCS/M : MAJOR :COMPUTER SCIENCE : M.Sc. (COMPUTER SCIENCE)
KEYWORDS : INFORMATION RETRIEVAL / AUTOMATIC QUERY EXPANSION/ SIMILARITY THESAURUS
SATIT SRISWANG : AUTOMATIC QUERY EXPANSION FOR THAI TEXT RETRIEVAL. THESIS ADVISORS : DAMRUS WONGSAWANG Ph.D., SUPACHAI TANGWONGSAN Ph.D. 73 p. ISBN 974-661-726-5

An information retrieval system is constructed to satisfy a user's query information need by identifying the documents in a document collection that contain the desired information. A common problem of most information retrieval systems is user's query formulation. A user may construct his (her) query using terms different from the ones indexed in the system. So, the user may get no result although there are some related documents in the collection. Furthermore, most users have no skill in selecting good search terms. Therefore, a query may be formulated vaguely and few related documents are retrieved.

Using an information structure for automatic query expansion is introduced to solve these problems. Most researches has focused on English text retrieval systems and some of them have been implemented and currently in use. However, for Thai text IR system, query expansion has not been investigated. In our work, we introduced the use a similarity thesaurus as an information structure. The model called IRTQE (Information Retrieval with Thesaurus Query Expansion) using similarity thesaurus for query expansion in Thai text retrieval has been proposed and its performance has been investigated. We simulated the test environments and compared the results between the IRTQE and the traditional information retrieval system. We found that IRTQE provides recall and precision improvement over the traditional system at any database size. Furthermore, we also studied the factors that affect retrieval improvement. These factors are the number of expanding terms, the percentage of top ranked documents to determine the good search terms in the original query, the value of K in weighting function, and the threshold value of similarity between query and document. The experimental results show that when increasing the number of expanding terms, the recall improvement is increased while the precision improvement is decreased. For our test environment, the appropriate value of percentage of top ranked documents and K in weighting are 10% and 0.5, respectively. Since a larger number of expanding terms produces less precision, the threshold value of similarity between query and document can be used to enhance the precision. We also found that using a high threshold value of similarity between query and document increases the precision, but decreases the recall improvement. Finally, some drawbacks of IRTQE are discussed and some suggestions are presented.