3837384 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc. (COMPUTER SCIENCE)
KEY WORDS : INFORMATION RETRIEVAL / AUTOMATIC INDEXING / LATENT SEMANTIC INDEXING
THEERADEJ TUNPAIBOON : THAI AUTOMATIC INDEXING USING LATENT SEMANTIC INDEXING. THESIS ADVISORS : CHONCHANOK VIRAVAN, Ph.D., DAMRAS WONGSAWANG, Ph.D. 101 P. ISBN 974-664-218-9.

Traditional techniques for Thai text retrieval systems have problems of automatic indexing and retrieval effectiveness. The first problem is caused by the lack of a delimiter between Thai words, leading to problems of unknown words, prefix and suffix matching. Also, query results often contain too much irrelevant information. The second problem, traditional search techniques use only pattern matching to index, causing undesirable results of low recall and low precision.

In Thai automatic indexing, we added algorithm called compound generation to resolve the first problem. It uses threshold compound value in generating new indexes. In order to improve recall and precision to mitigate the second problem, we applied a Latent Semantic Indexing model with the Thai text retrieval. We call this method TLSI. It can retrieve a document by its semantic rather than pattern matching.

TLSI was tested on a Thai corpus collection from The Thai Junior Encyclopedia. This corpus was composed of 185 files with the total size of 9.79 MB after passing word segmentation. The comparison quality between the compound generation index and typical index shows that quality of index improves 6.34% to 23.44%. To measure the performance of TLSI, we tested with the query set containing 100 queries. We studied the factors that affect retrieval effectiveness. These factors are the relative change of matrix and similarity threshold between query and document. The experimental results showed that recall of TLSI increases as percent of relative change of matrix increase but the recall decreases as similarity threshold increase. The precision increases as similarity threshold increase but decreases as percent of relative change of matrix increase. Furthermore, we also compared the results of TLSI and inversion technique. We found that TSLI helped overcome the problem of low precision. Although experimental results showed that percentage of recall reduces a little bit, the percentage of precision improves significantly. In best case, the percentage of precision improves 81.82%.