AUTOMATIC QUERY EXPANSION FOR THAI TEXT RETRIEVAL

SATIT SRISWANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE MASTER DEGREE IN COMPUTER SCIENCE
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1998

ISBN 974-661-726-5

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.

3937009 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)
คำสำคัญ : การสืบค้นข้อมูลภาษาไทย / การขยายคำสั่งสอบถามแบบอัตโนมัติ

สาธิต ศรีสว่าง : การขยายคำสั่งสอบถามแบบอัตโนมัติสำหรับการสืบค้นข้อมูลภาษาไทย (Automatic Query Expansion for Thai Text Retrieval) คณะกรรมการควบคุมวิทยานิพนธ์ : คำรัส วงศ์สว่าง, Pd.D., ศุภชัย ตั้งวงศ์ศานต์, Ph.D., 73 หน้า. ISBN 974-661-726-5

ระบบการสืบค้นข้อมูลถูกสร้างขึ้นเพื่อตอบสนองความต้องการของผู้ใช้ระบบในการค้นหาเอกสารที่ต้องการ ปัญหาที่พบบ่อยในระบบการสืบค้นข้อมูลก็คือ คำสั่งสอบถามประกอบด้วยคำที่ไม่อยู่ในดัชนีของระบบ ดังนั้นถึงแม้จะมีเอกสารที่ต้องการอยู่ในระบบ ผู้ใช้ระบบอาจไม่ได้รับเอกสารตามที่ต้องการ นอกจากนี้ผู้ใช้ระบบส่วนใหญ่ไม่มีทักษะในการเลือกใช้คำค้นหาที่ดี ทำให้คำสั่งสอบถามมีลักษณะคลุมเครือ

การใช้โครงสร้างข้อมูลสำหรับการขยายคำสั่งสอบถามจึงถูกนำเสนอขึ้นเพื่อแก้ไขปัญหาข้างต้น งานวิจัยส่วนใหญ่มุ่งเน้นที่ระบบการสืบค้นข้อมูลภาษาอังกฤษ และได้มีการนำมาใช้ในปัจจุบัน อย่างไรก็ตามยังไม่มีงานวิจัยใดที่ศึกษาในระบบการสืบค้นข้อมูลภาษาไทย ดังนั้น งานวิจัยฉบับนี้จึงทำการศึกษาการใช้พจนานุกรมซึ่งเป็นโครงสร้างข้อมูลประเภทหนึ่งสำหรับการขยายคำสั่งสอบถามในระบบการสืบค้นข้อมูลภาษาไทย โดยทำการสร้างระบบจำลองเพื่อวัดผลประสิทธิผลในการสืบค้นข้อมูลเทียบกับระบบการสืบค้นทั่วไป จากการทดลองพบว่า ระบบที่นำเสนอนี้ให้ค่า recall และค่า precision ที่สูงกว่าระบบการสืบค้นทั่วไป นอกจากนี้ยังได้ทำการศึกษาถึงปัจจัยต่างๆ ที่มีผลต่อการสืบค้น ได้แก่ จำนวนคำที่เพิ่มเข้ามาเพื่อขยายคำสั่งสอบถาม, เปอร์เซนต์ของเอกสารที่เป็นผลมาจากการสืบค้นในครั้งแรก ซึ่งจะถูกนำมาใช้ในการกำหนดคำค้นหาที่ดีในคำสั่งสอบถาม, ค่า K ใน weighting function และการกำหนด threshold ของค่าความสัมพันธ์ระหว่างคำสั่งสอบถามกับเอกสาร ผลการทดลองแสดงให้เห็นว่า การเพิ่มจำนวนคำที่เพิ่มเข้ามาเพื่อขยายคำสั่งสอบถาม ทำให้การเพิ่มขึ้นของค่า recall สูงขึ้น แต่การเพิ่มขึ้นของค่า precision ลดลง การใช้ 10% ของเอกสารที่เป็นผลมาจากการสืบค้นครั้งแรกในการกำหนดคำค้นหาที่ดีให้ผลดีกว่าการใช้ 20% ของเอกสาร และค่า K ที่เหมาะสมใน weighting function คือ 0.5 นอกจากนี้ยังพบว่าการกำหนด threshold ของค่าความสัมพันธ์ระหว่างคำสั่งสอบถามกับเอกสารที่สูงสามารถเพิ่มค่า precision ของระบบได้ แต่อย่างไรก็ตาม การเพิ่มขึ้นของค่า recall จะลดลง ท้ายที่สุดงานวิจัยฉบับนี้ยังได้ศึกษาถึงข้อเสียของระบบที่นำเสนอนี้ ตลอดจนแนวทางการวิจัยในอนาคต