QUERY EXPANSION USING LOCAL CONTEXT ANALYSIS

APICHART PHUNCHAROENPONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2001

ISBN 974-04-4780-3

3937017 SCCS/M: MAJOR : COMPUTER SCIENCE : M.Sc.(COMPUTER SCIENCE)
KEY WORDS : INFORMATION RETRIEVAL / QUERY EXPANSION / LOCAL CONTEXT ANALYSIS / WEIGHTED INVERSE DOCUMENT FREQUENCY
APICHART PHUNCHAROENPONG: QUERY EXPANSION USING LOCAL CONTEXT ANALYSIS. THESIS ADVISORS: DAMRAS WONGSAWANG Ph.D., SUPACHAI TANGWONGSAN Ph.D., 81 p. ISBN 974-04-4780-3

Information Retrieval (IR) is concerned with locating documents that are relevant to the user's information needs or queries from a large collection of documents. A fundamental problem for information retrieval is word mismatch. A query is usually a short and incomplete description of the underlying information needed. The users of IR systems and the authors of the documents often use different words to refer to the same concept. Most users have no skill in selecting good search terms.

This thesis investigated text analysis technique, Local Context Analysis using Weighted Inverse Document Frequency (LCAWIDF), and applied a query expansion technique for Thai text retrieval. We simulated the test environments and compared the results between the LCAWIDF and the original model, Local Context Analysis (LCA) system. The experimental results show that LCAWIDF provides precision and recall improvement over the LCA system. Factors that affect retrieval improvement were studied. These factors are: the number of expanding terms, the percentage of top ranked documents, the number of passages used and passage size. We found that when increasing the number of expanding terms, the recall improvement increased while the precision improvement decreased. The appropriate value of percentage of top ranked documents is 30%. Optimal retrieval efficiency was found when over ten passages at 300 words each, were used.

3937017 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

อภิชาติ พันธ์เจริญพงศ์ : การขยายคำสอบถามโดยวิธีการวิเคราะห์บริบทท้องถิ่น (QUERY EXPANSION USING LOCAL CONTEXT ANALYSIS) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D., ศุภชัย ตั้งวงศ์ศานต์, Ph.D., 81 หน้า ISBN 974-04-4780-3

ระบบการสืบค้นข้อมูลถูกสร้างขึ้นเพื่อตอบสนองความต้องการของผู้ใช้ระบบในการค้นหาเอกสารที่ต้องการ ปัญหาที่พบบ่อยในระบบการสืบค้นข้อมูล คือ คำสอบถามประกอบด้วยคำที่ไม่อยู่ในดัชนีของระบบ อันเนื่องมาจากการระบุข้อความที่ต้องการสืบค้นนั้นไม่สมบูรณ์ หรือ ข้อความสั้นเกินไป ซึ่งผู้ใช้ระบบการสืบค้นข้อมูล มักจะใช้คำสอบถาม แตกต่างกับคำที่ผู้เขียนใช้ในการเขียนเอกสาร แต่ทั้งสองคำนั้นมีความหมายเดียวกัน โดยส่วนมากแล้วผู้ใช้ระบบสืบค้นข้อมูล จะไม่มีทักษะที่ดีในการเลือกคำสอบถามที่ใช้ในการค้นหา

วิทยานิพนธ์ฉบับนี้ได้แนะนำเทคนิคการขยายคำสอบถามโดยวิธีการวิเคราะห์บริบทท้องถิ่นในระบบการสืบค้นข้อมูลภาษาไทย โดยทำการสร้างระบบจำลองเพื่อวัดผลประสิทธิผลในการสืบค้นข้อมูลเปรียบเทียบกับระบบการสืบค้นข้อมูลต้นแบบ จากการทดลองพบว่า ระบบที่นำเสนอนี้ให้ค่า precision และ ค่า recall ที่สูงกว่าระบบการสืบค้นข้อมูลต้นแบบ นอกจากนี้ยังได้ทำการศึกษาถึงปัจจัยต่าง ๆ ที่มีผลต่อการสืบค้นข้อมูล ได้แก่ จำนวนคำที่เพิ่มเข้ามาเพื่อขยายคำสอบถาม เปอร์เซนต์ของเอกสาร และ จำนวนของเอกสารที่เป็นผลมาจากการสืบค้นข้อมูลในครั้งแรก ซึ่งจะถูกนำมาใช้ในการกำหนดคำค้นหาข้อมูลที่ดีในคำสอบถาม และ ขนาดของเอกสาร ผลการทดลองแสดงให้เห็นว่า การเพิ่มจำนวนคำที่เพิ่มเข้ามาเพื่อขยายคำสอบถาม ทำให้การเพิ่มขึ้นของค่า recall สูงขึ้น แต่การเพิ่มขึ้นของค่า precision ลดลง การใช้เอกสารมากกว่า 10 เอกสาร หรือ ประมาณ 30% ของเอกสารที่เป็นผลมาจากการสืบค้นข้อมูลครั้งแรก เป็นค่าที่เหมาะสมในการกำหนดคำค้นหาที่ดี และ ขนาดของเอกสารที่ให้ค่าการสืบค้นที่มีประสิทธิภาพ คือ 300 คำ