AUTOMATIC THESAURUS CONSTRUCTION WITH TERM CONTEXT AND SYNTACTIC ANALYSIS FOR THAI TEXT RETRIEVAL

KANYARAT LAIRUNGRUANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2003

ISBN 974-04-3367-7

การสร้างธิซอรัสแบบอัตโนมัติโดยใช้การวิเคราะห์บริบทของคำและไวยกรณ์สำหรับการค้นคืนเอกสารภาษาไทย (AUTOMATIC THESAURUS CONSTRUCTION WITH TERM CONTEXT AND SYNTACTIC ANALYSIS FOR THAI TEXT RETRIEVAL)

กัญญารัตน์ ไหลรุ่งเรือง 4137568 SCCS/M

วท.ม.(วิทยาการคอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D. , ชมทิพ พรพนมชัย,Ph.D.

บทคัดย่อ

ในระบบการสืบค้นข้อมูล เพื่อใช้ในการค้นหาเอกสารที่ต้องการ ผู้ใช้มีหน้าที่ที่จะต้องระบุคำสั่งสอบถามเข้าไปในระบบ เพื่อให้ระบบทำการค้นคืนข้อมูลที่เกี่ยวข้อง แต่ปัญหาที่พบบ่อยคือ คำสั่งสอบถามที่ผู้ใช้ระบุนั้นไม่ตรงกับคำที่เป็นดัชนีของระบบ ดังนั้น แม้จะมีเอกสารที่ต้องการในระบบ แต่ผู้ใช้อาจจะไม่ได้รับเอกสารตามที่ต้องการ นอกจากนี้ผู้ใช้ระบบอาจไม่มีทักษะในการเลือกคำค้นที่ดี ทำให้คำสั่งสอบถามมีลักษณะคลุมเครือ เพื่อแก้ปัญหาข้างต้นนี้ วิธีการขยายคำสั่งสอบถามจึงได้ถูกนำเสนอขึ้น

วิทยานิพนธ์ฉบับนี้ได้แนะนำวิธีการสร้างพจนานุกรมคำคล้ายแบบอัตโนมัติโดยการวิเคราะห์โครงสร้างทางไวยากรณ์ของประโยคสำหรับการสืบค้นข้อมูลภาษาไทย เพื่อใช้ในการขยายคำสั่งสอบถามโดยทำการสร้างระบบจำลอง เพื่อวัดประสิทธิภาพในการสร้างพจนานุกรมและการนำพจนานุกรมที่สร้างขึ้นมาไปใช้ เปรียบเทียบกับระบบการสร้างพจนานุกรมแบบเก่า จากการทดลองพบว่า ระบบที่นำเสนอนี้ มีค่าความถูกต้องของพจนานุกรมมากกว่าระบบเก่า และยังให้ค่า precision และค่า recall ที่สูงกว่าระบบต้นแบบ นอกจากนี้ยังได้ทำการศึกษาถึงปัจจัยที่มีผลในการสร้างพจนานุกรม ได้แก่ การกำหนดค่า threshold ของค่าความใกล้เคียงระหว่างคำ 2 คำ ผลการทดลองแสดงให้เห็นว่า การกำหนดค่า threshold ของค่าความใกล้เคียงระหว่างคำ 2 คำ ในช่วง 0-0.5 สามารถเพิ่มค่า precision และค่า recall ของระบบได้ และท้ายสุด งานวิจัยฉบับนี้ยังได้ศึกษาถึงแนวทางการวิจัยในอนาคต

98 หน้า. ISBN 974-04-3367-7

AUTOMATIC THESAURUS CONSTRUCTION WITH TERM CONTEXT AND SYNTACTIC ANALYSIS FOR THAI TEXT RETRIEVAL.

KANYARAT LAIRUNGRUANG 4137568 SCCS/M

M.Sc.(COMPUTER SCIENCE)

THESIS ADVISOR : DAMRAS WONGSAWANG, Ph.D. , CHOMTIP PORNPANOMCHAI Ph.D.

ABSTRACT

In information retrieval, algorithm based search engines are used to locate and select relevant documents from a corpus of documents on the internet. A user typically submits queries consisting of little more than two or three words related to the topic of interest, especially in a Web environment, where collections tend to be enormous and queries tend to be short, with users rarely taking time to modify their searches . On the other hand, most users have no skill in selecting good search terms. A user may use vocabulary for submitting a query that is different from the one indexed in the system. So, they will not get any results although there are some related documents in the collection. Therefore, if a query is formulated vaguely only a few related documents are returned. To solve this problem, the concept of query expansion using similarity is introduced. It is reasonable to assume that, by expanding query terms with additional related terms drawn from a thesaurus such word relations would offer the most systematic method for offering synonyms, homonyms and term relationship. Assuming that a query term is a good discriminator for finding a relevant document, a related term will likely be useful as a discriminator as well.

In this thesis, we introduced an automatic thesaurus construction. The model called ATCSA (Automatic Thesaurus Construction with Term Context and Syntactic Analysis for Thai text Retrieval) using term context and simple syntactic analysis for constructing a thesaurus for Thai text retrieval. We simulated the test environments and compared the results between ATCSA and the traditional system. We found that ATCSA provides results more accurately than the traditional system. Furthermore, we also studied the factors that effect retrieval improvement. These factors are the cluster-weight threshold for selecting the similarity term from the similarity thesaurus. We found that using a low threshold value of similarity between two terms increases the precision and recall. Finally, improvements of the model are proposed.

KEYWORDS : THAI TEXT RETRIEVAL / SIMILARITY THESAURUS / SYNTACTIC / TERM CONTEXT

98 p. ISBN 974-04-3367-7