USING THE PASSAGE-BASED AND TERM-WEIGHT RETRIEVAL IN FULL-TEXT

NALINRAT WITSAWAKITTI

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

ISBN 974-665-680-5

3936991SCCS/M:MAJOR:COMPUTER SCIENCE; M.Sc.(COMPUTER SCIENCE)
KEY WORDS : INFORMATION RETRIEVAL / FULL-TEXT/ AUTOMATIC INDEXING / PASSAGE-BASED RETRIEVAL/TERM-WEIGHTED RETRIEVAL

NALINRAT WITSAWAKITTI: USING THE PASSAGE-BASED AND TERM-WEIGHT RETRIEVAL IN FULL-TEXT. THESIS ADVISOR: DAMRAS WONGSAWANG, Ph.D. 156 p. ISBN 974-665-680-5

Deciding what keywords and topics for documents selected in an index is a very difficult task if the decision-maker is not a specialist in that area. To solve this problem, full-text indexing is introduced and keywords of documents are taken from every word in the documents. Although this method is easy for index creating, efficiency in retrieval of terms of recall and precision may decrease since indexes contain too many words which are not keywords. This thesis proposed a method to improve full-text indexing which can be applied to Thai documents efficiently by using passage-based retrieval and term-weight retrieval. (Passage-based and Term-weight Retrieval (PTR) model)

To verify and prove the model software was developed to simulate the PTR model. The document structures were divided into three levels because all documents collected in databases were academic articles. The contents were usually separated into three passages: title passage, abstract passage and body passage. Experiments were classified into three groups depending on size of database collection containing 100 documents, 300 documents and 500 documents. Structure of documents contained title-body passage, title-abstract passage, Thai language and English language documents. Types of queries used Thai only, English only, or both Thai and English.

Suitable body weight values were less than threshold value and abstract and title weight value were near or greater than threshold value. Effective retrieval of PTR increased by 14%-47% (compared to term-weighted retrieval model); different sizes of database collection did not effect the suitable weight value. Different categories of documents have an effect to suitable weight values if documents contain only title and abstract passages. A query into Thai and English language resulted in better on effective retrieval than query with only Thai or English language. Future work could be done to improve the capability of string searching, queries searching by Boolean Expressions and to improve the weighting system.

3936991 SSCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม.(วิทยาการคอมพิวเตอร์)

นลินรัตน์ วิศวกิตติ : การใช้โครงสร้างของเอกสารร่วมกับการถ่วงน้ำหนักสำหรับการสืบค้นทั้งเอกสาร (USING PASSAED-BASED AND TERM-WEIGHTED RETRIEVALE IN FULL-TEXT) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D. ธันวดี สุเนตรนันท์ , Ph.D. 156 หน้า. ISBN 974-665-680-5

การกำหนดคำสำคัญของเอกสารเพื่อใช้เป็นอินเด็กซ์ในการค้นหาเป็นงานที่ยากถ้าผู้กำหนดไม่ใช่ผู้เชี่ยวชาญในด้านที่เกี่ยวข้องกับเนื้อหาของเอกสาร ดังนั้นจึงได้มีการเสนอให้ใช้คำทุกคำที่ปรากฏอยู่ในเนื้อหาของเอกสารเป็นคำสำคัญเพื่อแก้ปัญหา แม้ว่าวิธีการดังกล่าวจะง่ายแต่พบว่าทำให้ประสิทธิภาพในการสืบค้นต่ำ เนื่องจากมีอินเด็กซ์มากเกินไปและอินเด็กซ์บางตัวอาจจะไม่ใช่คำสำคัญที่แท้จริงของเอกสาร งานวิจัยนี้จึงนำโครงสร้างของเอกสารและการถ่วงน้ำหนักมาใช้สร้างแบบจำลอง เพื่อช่วยในเพิ่มประสิทธิภาพในการสืบค้นเอกสารภาษาไทย เรียกว่า Passage-based and Term-weight Retrieval (PTR)

งานวิจัยนี้จะทำการทดลองเพื่อค้นหาค่าถ่วงน้ำหนักที่เหมาะสมสำหรับแต่ละชนิดของโครงสร้างเอกสาร เนื่องจากเอกสารที่นำมาทดลองเป็นบทความทางวิชาการ ในงานวิจัยจึงได้ทำการแบ่งโครงสร้างของเอกสารออกเป็น 3 ระดับได้แก่ ชื่อเรื่อง บทคัดย่อ และ เนื้อหา โดยทำการหาค่าถ่วงน้ำหนักในสภาพแวดล้อมที่แตกต่างกัน 3 ลักษณะ คือ จำนวนเอกสาร รูปแบบของเอกสาร และ รูปแบบของคำที่ใช้ในการสืบค้น

จากการวิจัยพบว่า ค่าถ่วงน้ำหนักของเนื้อหาควรจะมีค่าต่ำ ส่วนบทคัดย่อและชื่อเรื่องควรมีค่าที่สูง ประสิทธิภาพของการสืบค้นเพิ่มขึ้น 14%-47% เมื่อเทียบกับการสืบค้นที่ใช้การถ่วงน้ำหนักเพียงอย่างเดียว รูปแบบของเอกสารไม่มีผลกระทบต่อค่าถ่วงน้ำหนักยกเว้นเอกสารที่มีโครงสร้างเพียงชื่อเรื่องและบทคัดย่อ และการค้นหาด้วยคำภาษาไทยและภาษาอังกฤษจะมีประสิทธิภาพการสืบค้นดีกว่าการใช้ภาษาใดภาษาหนึ่งเพียงภาษาเดียว งานวิจัยนี้สามารถนำไปสู่การพัฒนางานในด้านต่างๆ เช่น การปรับปรุงวิธีการในการค้นหาข้อความที่ใช้ในตัวแบบจำลองให้ดีขึ้น สามารถใช้นิพจน์บูลลีนในการค้นหาได้ และการประยุกต์ใช้กับเอกสารอื่นๆ