ชื่อเรื่องวิทยานิพนธ์ เทคนิคการสร้างดัชนีเพื่อการค้นคืนด้วยวิธีพาเชียลลีสสแควร์

ผู้เขียน นายวสุวัฒน์ ทิพย์สังวาลย์

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.รัฐสิทธิ์ สุขะหุต

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการสร้างดัชนี สำหรับการค้น คืนข้อมูล โดยใช้เอกสารชีวสารสนเทศทางการแพทย์ซึ่งมีการจำแนกกลุ่มของเอกสารไว้แล้วทั้งหมด 2000 เอกสาร ทำการทดลองโดยแบ่งเอกสารออกเป็นสองชุดขนาดเท่ากัน ใช้วิธีการพาเชียลลีสสแควร์ เพื่อสร้างดัชนีเพื่อค้นคืนเอกสาร กระบวนการที่นำเสนอคือ การใช้วิธีการพาเชียลลีสสแควร์สำหรับสร้างดัชนี ดัชนีที่ได้จะมีขนาดเล็กกว่าขนาดดัชนีเดิม จากนั้น ทำการ ค้นคืนเอกสาร โดยใช้ ทฤษฎี ความคล้ายคลึงเชิงมุม หาความสัมพันธ์ระหว่างคำสำคัญที่ต้องการค้นคืนและดัชนี ทำการทดสอบประสิทธิภาพ โดยทำการวัดหาค่าความเที่ยงเฉลี่ย และจับเวลาที่ใช้ในการคำนวณเพื่อสร้างดัชนี แล้วนำมาเปรียบเ ทียบกับประสิทธิภาพของ ดัชนีที่สร้างจากวิธีการแอลเอสไอ ผลที่ได้โดยนำเอาผลประสิทธิภาพจากข้อมูลสองชุดมาหาค่าเฉลี่ยคือ ดัชนีตัวแทนเอกสารที่สร้างจากวิธีการพาเชียลลีสสแควร์ ให้ความเที่ยงเฉลี่ยมากที่สุด เมื่อทำการลดมิติข้อมูลลงเหลือร้อยละ 27.5 จากขนาดมิติเดิม โดยมีความเที่ยงเฉลี่ยร้อยละ 47.54 และใช้เวลาในการคำนวณเพื่อสร้างดัชนีเฉลี่ย 133.78 วินาที ในขณะที่ดัชนีตัวแทนเอกสารที่สร้างจากวิธีการแอลเอสไอมีค่าความเที่ยงเฉลี่ยสูงสุดเมื่อทำการลดมิติข้อมูลลงเหลือ ร้อยละ 65 โดยมีค่าความเที่ยงเฉลี่ยร้อยละ 47.46 และใช้เวลาเฉลี่ย 241.59 วินาที ในการคำนวณเพื่อสร้างดัชนี แสดงให้เห็นถึงความสามารถในการสร้างดัชนีที่เหมาะสมเมื่อนำมาเทียบกับวิธีการแอลเอสไอซึ่งเป็นวิธีการที่มีการนำเสนออย่างแพร่หลาย

Thesis Title Indexing Technique for Retrieval Using Partial Least Squares Method

Author Mr. Wasuwat Thipsungwan

Degree Master of Science (Computer Science)

Thesis Advisor Assistant Professor Dr. Rattasit Sukhahuta

ABSTRACT

In this paper, we present a technique for literature indexing and retrieving for information. The aim is to improve indexing efficiency, decrease time and increase the accuracy of data system. In the proposed method, Partial Least Squares (PLS) was used to reduce the dimensions of documents. This method was applied to index and calculate ranking weight by using cosine similarity technique. We tested from 2000 dataset of Bio-Medical information to perform an index, dataset has been divided into 2 groups, in the overall performance of the developed system by comparing PLS with LSI (Latent Semantic Indexing). The result show that PLS has average precision at 47.54% and spent 133.78 seconds to create index in 27.5% reduced dimension, meanwhile LSI can reduce dimension to 65% with average precision at 47.46% and spent 241.59 seconds to create index, proven that PLS has ability to use as method for indexing.