การศึกษาเปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ใช้เทคนิคการวัดความคล้ายคลึงเชิงมุม และเทคนิคการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนที่กำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์จากการจัดกลุ่มข้อมูล

นางสาวสุนันทา เปี่ยมพริ้ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

A COMPARISON STUDY OF THE EFFICIENCY OF INFORMATION RETRIEVAL SYSTERMS USING COSINE ANGLE AND EUCLIDEAN DISTANCE WHERE SIMILARITY FRAME IS GUIDED BY OUTPUT FROM CLUSTERING TECHNIQUE

Miss Sunantha Piempring

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science Program in Business Software Development

Department of Statistics

Faculty of Commerce and Accountancy

Chulalongkorn University

Academic Year 2007

หัวข้อวิทยานิพนธ์	การศึกษาเปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ใช้เทคนิคการวัดความคล้ายคลึงเชิงมุมและเทคนิคการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนที่กำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์จากการจัดกลุ่มข้อมูล
โดย	นางสาวสุนันทา เปี่ยมพริ้ง
สาขาวิชา	การพัฒนาซอฟต์แวร์ด้านธุรกิจ
อาจารย์ที่ปรึกษา	อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร. อรรณพ ตันละมัย)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. อัษฎาพร ทรัพย์สมบูรณ์)

..... อาจารย์ที่ปรึกษา
(อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน)

..... กรรมการ
(อาจารย์ ดร. บุรัชย์ ภัทรโกศล)

สุนันทา เปี่ยมพริ้ง : การศึกษาเปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ใช้เทคนิคการวัดความคล้ายคลึงเชิงมุมและเทคนิคการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนที่กำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์จากการจัดกลุ่มข้อมูล. (A COMPARISON STUDY OF THE EFFICIENCY OF INFORMATION RETRIEVAL SYSTERMS USING COSINE ANGLE AND EUCLIDEAN DISTANCE WHERE SIMILARITY FRAME IS GUIDED BY OUTPUT FROM CLUSTERING TECHNIQUE) อ. ที่ปรึกษา : อ. ดร.จันทร์เจ้า มงคลนาวิน, 183 หน้า.

วิทยานิพนธ์นี้เสนอการศึกษาเปรียบเทียบประสิทธิภาพของระบบการค้นคืนเอกสารเทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม และวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการประยุกต์ใช้ทฤษฎีการจัดกลุ่มข้อมูลแบบ K-mean Clustering กำหนดเงื่อนไข หรือกรอบความคล้ายคลึงในการเลือกเอกสารที่เป็นคำตอบ ถ้าเอกสารใดที่มีระยะห่างกับข้อสอบถามภายใต้กรอบความคล้ายคลึงที่กำหนดจะถูกค้นคืนออกมาแสดงต่อผู้ใช้ โดยได้ทดสอบกับชุดเอกสารนิตยสารไทม์ จำนวน 425 เอกสาร และข้อสอบถามจำนวน 83 ข้อสอบถาม โดยเปรียบเทียบประสิทธิภาพของระบบการค้นคืนเอกสารทั้ง 2 รูปแบบข้างต้น ด้วยค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค

จากผลการทดลองสรุปได้ว่า ระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม มีค่าประสิทธิภาพทั้ง 3 ค่ามากกว่าระบบการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน ผู้วิจัยตั้งข้อสังเกตว่าวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนอาจจะไม่เหมาะสมสำหรับนำมาใช้ในกระบวนการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ เมื่อทดสอบด้วยชุดเอกสารนิตยสารไทม์ เนื่องจากเป็นชุดเอกสารที่มีความหลายหลายของคำสูง

ผู้วิจัยจึงได้ศึกษาว่าการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-mean Clustering บนระยะห่างเชิงมุมมากำหนดเงื่อนไขในการเลือกเอกสารที่เป็นคำตอบ จะสามารถเพิ่มประสิทธิภาพของระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมได้หรือไม่ ผลการทดลองแสดงให้เห็นว่าเมื่อเปรียบเทียบกับระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม ประสิทธิภาพของระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมที่ใช้เทคนิคการจัดกลุ่มข้อมูลสามารถทำให้ค่าประสิทธิภาพความแม่นยำและค่าเฉลี่ยฮาร์โมนิคดีขึ้น แต่ค่าประสิทธิภาพความระลึกต่ำลง

ภาควิชา สถิติ ลายมือชื่อนิสิต สุนันทา เปี่ยมพริ้ง

สาขาวิชา การพัฒนาซอฟต์แวร์ด้านธุรกิจ ลายมือชื่ออาจารย์ที่ปรึกษา

ปีการศึกษา 2550

4882289426 : MAJOR BUSINESS SOFTWARE DEVELOPMENT
KEY WORD: INFORMATION RETRIEVAL /COSINE ANGLE /EUCLIDEAN DISTANCE /CLUSTERING

SUNANTHA PIEMPRING : A COMPARISON STUDY OF THE EFFICIENCY OF INFORMATION RETRIEVAL SYSTEMS USING COSINE ANGLE AND EUCLIDEAN DISTANCE WHERE SIMILARITY FRAME IS GUIDED BY OUTPUT FROM CLUATERING TECHNIQUE. THESIS ADVISOR : JANJAO MONGKOLNAVIN,Ph.D., 183 pp.

The thesis presents a comparison study of the efficiency between the vector space model information retrieval system using cosine angle technique and the one using Euclidean distance technique together with K-means clustering where K-means clustering is used to guide the threshold for retrieving answer documents. The experiments were conducted on the TIME Magazine collection which consists of 425 documents and 83 queries. The performance of the two information retrieval systems is compared through the use of Precision, Recall and Harmonic mean measurement.

The experimental results show that the performance of the information retrieval system using cosine angle technique is significantly better than those using Euclidean distance technique in all three measurements. It was observed that the Euclidean distance technique may be unsuitable for comparing the similarity in the TIME Magazine collection where the variation in words is extremely high.

Thus, the exploratory study was conducted to further investigate whether the use of the cosine angle technique together with K-mean clustering can improve the efficiency of the traditional cosine angle information retrieval system or not. The results show that the information retrieval system using the cosine angle together with K-mean clustering has higher Precision and Harmonic mean than those without K-mean clustering technique, but has lower Recall.

Department : Statistics Student's signature : สุนันทา เปี่ยมพริ้ง
Field of study : Business Software Development Advisor's signature :
Academic year 2007

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้จะสำเร็จลุล่วงไปได้ด้วยดีต้องขอกราบพระคุณ อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูงยิ่ง อาจารย์ได้ให้คำแนะนำ ข้อคิดเห็นต่าง ๆ ชี้แนะแนวทางในการวิจัยด้วยดีตลอดมา รวมทั้งตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียดจนสำเร็จเป็นวิทยานิพนธ์ฉบับสมบูรณ์นี้ และผู้วิจัยขอขอบพระคุณอาจารย์ ผู้ช่วยศาสตราจารย์ ดร. อัษฎาพร ทรัพย์สมบูรณ์ ประธานกรรมการวิทยานิพนธ์ และอาจารย์ ดร. บุรัชย์ ภัทรโกศล กรรมการวิทยานิพนธ์ที่กรุณาเสียสละเวลาให้คำแนะนำ ชี้แนะสิ่งต่าง ๆ จนเนื้อหาวิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ และขอบพระคุณอาจารย์ ดร.อรุณี กำลัง ที่ให้คำปรึกษา ชี้แนะแนวทางในเรื่องของการวิเคราะห์ผลการทดลองทางสถิติ

ขอบคุณเพื่อน ๆ ทุกคนที่ให้ความช่วยเหลือ ให้คำปรึกษาแนะนำวิธีการแก้ปัญหาต่าง ๆ และให้กำลังใจตลอดมา ขอบคุณพี่ศิรัตน์ สำหรับความช่วยเหลือและคำปรึกษาในการทำวิทยานิพนธ์ตลอดมา ขอบคุณพี่กุลยา และพี่ ๆ เจ้าหน้าที่ฝ่ายเทคนิคบริษัท SAS ประเทศไทยที่ให้คำปรึกษาเกี่ยวกับการใช้โปรแกรม SAS Enterpise Miner 5.1 เป็นอย่างดี

ท้ายนี้ ผู้วิจัยใคร่กราบขอบพระคุณบิดามารดา และครอบครัวที่ให้กำลังใจยามท้อแท้ คอยให้ความช่วยเหลือ ให้การสนับสนุนแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

บทที่ หน้า

บทที่ หน้า

สารบัญตาราง

สารบัญภาพ