

การศึกษาเปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ใช้เทคนิคการวัดความคล้ายคลึงเชิงมุม
และเทคนิคการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนที่กำหนดกรอบ
ค่าความคล้ายคลึงด้วยผลลัพธ์จากการจัดกลุ่มข้อมูล

นางสาวสุนันทา เปี่ยมพริ้ง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

A COMPARISON STUDY OF THE EFFICIENCY OF INFORMATION RETRIEVAL SYSTEMS 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

Copyright of Chulalongkorn University

500721

หัวข้อวิทยานิพนธ์

การศึกษาเปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ใช้
เทคนิคการวัดความคล้ายคลึงเชิงมุมและเทคนิคการวัดความคล้ายคลึง
เชิงระยะห่างยูคลิเดียนที่กำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์
จากการจัดกลุ่มข้อมูล

โดย

นางสาวสุนันทา เปี่ยมพริ้ง


สาขาวิชา

การพัฒนาซอฟต์แวร์ด้านธุรกิจ

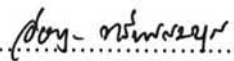
อาจารย์ที่ปรึกษา

อาจารย์ ดร. จันท์เจ้า มงคลนาวิน


คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัย
นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโทบริหารธุรกิจ

.....  คณบดีคณะพาณิชยศาสตร์และการบัญชี
(รองศาสตราจารย์ ดร. อรรณพ ตันละม้าย)

คณะกรรมการสอบวิทยานิพนธ์

.....  ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. อัสภาพร ทรัพย์สมบูรณ์)

.....  อาจารย์ที่ปรึกษา
(อาจารย์ ดร. จันท์เจ้า มงคลนาวิน)

.....  กรรมการ
(อาจารย์ ดร. ปุริชย์ ภัทรโกศล)

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

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

จากผลการทดลองสรุปได้ว่า ระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม มีค่าประสิทธิภาพทั้ง 3 ค่ามากกว่าระบบการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน ผู้วิจัยตั้งข้อสังเกตว่าวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนอาจจะไม่เหมาะสมสำหรับนำมาใช้ในกระบวนการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ เมื่อทดสอบด้วยชุดเอกสารนิตยสารไทม์ เนื่องจากเป็นชุดเอกสารที่มีความหลายหลายของคำสูง

ผู้วิจัยจึงได้ศึกษาว่าการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-mean Clustering บนระยะห่างเชิงมุมมากำหนดเงื่อนไขในการเลือกเอกสารที่เป็นคำตอบ จะสามารถเพิ่มประสิทธิภาพของระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมได้หรือไม่ ผลการทดลองแสดงให้เห็นว่าเมื่อเปรียบเทียบกับระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม ประสิทธิภาพของระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมที่ใช้เทคนิคการจัดกลุ่มข้อมูลสามารถทำให้ค่าประสิทธิภาพความแม่นยำและค่าเฉลี่ยฮาร์โมนิกดีขึ้น 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 : สุณันtha ปิเอมพริง
 Field of study : Business Software Development Advisor's signature : [Signature]
 Academic year 2007

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้จะสำเร็จลุล่วงไปได้ด้วยดีต้องขอกราบพระคุณ อาจารย์ ดร. จันทรเจ้า มงคลนาวิน อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างสูงยิ่ง อาจารย์ได้ให้คำแนะนำ ข้อคิดเห็นต่าง ๆ ที่แนะแนวทางในการวิจัยด้วยดีตลอดมา รวมทั้งตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียดจนสำเร็จเป็นวิทยานิพนธ์ฉบับสมบูรณ์นี้ และผู้วิจัยขอขอบพระคุณอาจารย์ ผู้ช่วยศาสตราจารย์ ดร. อัมภพร ทรัพย์สมบูรณ์ ประธานกรรมการวิทยานิพนธ์ และอาจารย์ ดร. บุรุษย์ ภัทรโกศล กรรมการวิทยานิพนธ์ที่กรุณาเสียสละเวลาให้คำแนะนำ ที่แนะสิ่งต่าง ๆ จนเนื้อหา วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ และขอบพระคุณอาจารย์ ดร.อรุณี กำลัง ที่ให้คำปรึกษา ที่แนะแนวทางในเรื่องของการวิเคราะห์ผลการทดลองทางสถิติ

ขอบคุณเพื่อน ๆ ทุกคนที่ให้ความช่วยเหลือ ให้คำปรึกษาแนะนำวิธีการแก้ปัญหาต่าง ๆ และให้กำลังใจตลอดมา ขอขอบคุณพี่ศิริรัตน์ สำหรับความช่วยเหลือและคำปรึกษาในการทำ วิทยานิพนธ์ตลอดมา ขอขอบคุณพี่กุลยา และพี่ ๆ เจ้าหน้าที่ฝ่ายเทคนิคบริษัท SAS ประเทศไทยที่ให้ คำปรึกษาเกี่ยวกับการใช้โปรแกรม SAS Enterprise Miner 5.1 เป็นอย่างดี

ท้ายนี้ ผู้วิจัยใคร่กราบขอบพระคุณบิดามารดา และครอบครัวที่ให้กำลังใจยาม ท้อแท้ คอยให้ความช่วยเหลือ ให้การสนับสนุนแก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญ.....	ช
สารบัญตาราง.....	ฎ
สารบัญภาพ.....	ฏ
บทที่	
1 ที่มาและความสำคัญของปัญหา	
1.1 ความสำคัญของปัญหา.....	1
1.2 วัตถุประสงค์ของการวิจัย.....	7
1.3 ขั้นตอนโดยสรุปของการวิจัย.....	7
1.4 ตัวแปรที่ศึกษา.....	8
1.5 ขอบเขตของการวิจัย.....	9
1.6 ประโยชน์ที่คาดว่าจะได้รับ.....	9
2 วรรณกรรมที่เกี่ยวข้อง	
2.1 เทคนิคการค้นคืนสารสนเทศ.....	10
2.2 การสกัดคำสำคัญออกจากเอกสาร.....	13
2.2.1 การตัดคำ.....	13
2.2.2 การกำจัดคำยกเว้น.....	13
2.2.3 การลดรูปคำ.....	13
2.3 การกำหนดดรรชนี.....	14
2.3.1 เพิ่มผกผัน.....	14
2.4 การกำหนดรูปแบบการค้นคืนเอกสารและข้อสอบถาม.....	15
2.5 การให้ค่าน้ำหนักของคำในข้อสอบถาม.....	20
2.6 การค้นหาเอกสารที่ตรงกับข้อสอบถามของผู้ใช้.....	21
2.7 การแบ่งกลุ่มเอกสารเทคนิค K-means Clustering.....	24

บทที่		หน้า
2.8	การกำหนดค่าความเหมือนในการค้นคืนเอกสารต่อผู้ใช้.....	25
2.9	การวัดประสิทธิภาพระบบค้นคืนเอกสาร.....	26
2.10	งานวิจัยที่เกี่ยวข้อง.....	28
3 ระเบียบวิธีวิจัย		
3.1	แผนแบบการทดลอง.....	34
3.1.1	ตัวแปรต้น.....	34
3.1.2	ตัวแปรตาม.....	35
3.1.3	ตัวแปรควบคุม.....	35
3.2	สมมติฐานงานวิจัย.....	39
3.3	แนวทางการทำวิจัย.....	40
3.4	ภาพรวมการทำงานของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร.....	41
3.5	องค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร.....	43
3.5.1	ส่วนที่ 1 การเตรียมข้อมูลเบื้องต้น.....	45
3.5.2	ส่วนที่ 2 การค้นคืนเอกสาร.....	47
3.5.3	ส่วนที่ 3 การคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบเทคนิคการ ค้นคืนเอกสาร.....	52
3.6	การทดสอบประสิทธิภาพของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร	52
3.7	ความถูกต้อง (Validity) และความน่าเชื่อถือ (Reliability).....	53
3.8	กรอบการวิเคราะห์ข้อมูล (Data Analysis Framework).....	54
4 ผลการทดลองและบทวิเคราะห์		
4.1	การกำหนดค่าตัวแปรอื่น ๆ ที่เกี่ยวข้องกับการทดลอง.....	56
4.1.1	การพิจารณาเลือกจำนวนกลุ่มของเอกสารที่เหมาะสม.....	56
4.1.2	การกำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์ที่ได้จากเทคนิคการ จัดกลุ่มข้อมูล (Clustering).....	64
4.2	ผลการทดลองประสิทธิภาพการค้นคืนเอกสาร.....	64
4.2.1	การวัดประสิทธิภาพการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1 ด้วยการกำหนดกรอบค่าความคล้ายคลึงที่ เหมาะสม.....	65

บทที่	หน้า
4.2.2 การวัดประสิทธิภาพการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1 เฉพาะเอกสารที่มีคำในข้อสอบถามปรากฏ เท่านั้น.....	71
4.2.3 การเปรียบเทียบประสิทธิภาพของเครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1* และเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2.....	78
4.3 การวิเคราะห์ผลการทดลองทางสถิติ.....	84
4.3.1 การวิเคราะห์ข้อมูลค่าความแม่นยำ (Precision).....	85
4.3.2 การวิเคราะห์ข้อมูลค่าความระลึก (Recall)	88
4.3.3 การวิเคราะห์ข้อมูลค่าเฉลี่ยฮาร์โมนิค (Harmonic mean).....	90
4.4 สรุปผลการทดลอง.....	93
4.5 อภิปรายผลการทดลอง.....	94
5 การศึกษาเชิงสำรวจ	
5.1 เครื่องมือทดสอบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัด ความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบค่าความคล้ายคลึงที่ กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ตามความ คล้ายคลึงเชิงมุม.....	98
5.1.1 เทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ตามความ คล้ายคลึงเชิงมุม.....	100
5.1.2 ผลการทดลองเทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ตามความคล้ายคลึงเชิงมุม.....	101
5.2 ผลการทดลอง.....	104
5.2.1 การวัดประสิทธิภาพการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 3 ด้วยการกำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน.....	104
5.2.2 ผลการทดลองเปรียบเทียบประสิทธิภาพของเครื่องมือทดสอบการค้น คืนเอกสารรูปแบบที่ 3 และ รูปแบบที่ 2.....	105
5.3 สรุปผลการวิเคราะห์การศึกษาเชิงสำรวจ.....	111
5.4 อภิปรายผลการศึกษาเชิงสำรวจ.....	111

บทที่	หน้า
6 สรุปผลการวิจัย	
6.1 การทดลองและลักษณะของข้อมูลที่ใช้ทดสอบการค้นคืนเอกสาร.....	113
6.2 สรุปผลการวิจัย.....	113
6.3 การนำงานวิจัยไปประยุกต์ใช้.....	117
6.3.1 การนำงานวิจัยไปใช้ในเชิงทฤษฎี.....	117
6.3.2 การนำงานวิจัยไปใช้ในเชิงประยุกต์.....	118
6.4 ข้อจำกัดของงานวิจัย.....	118
6.5 แนวทางการศึกษาต่อเนื่อง.....	119
รายการอ้างอิง.....	120
ภาคผนวก.....	
ภาคผนวก ก ตัวอย่างเอกสารและข้อสอบถาม.....	125
ภาคผนวก ข รายการคำยกเว้น (Stop words list).....	129
ภาคผนวก ค ขั้นตอนวิธีของพอร์เตอร์ (Porter's Algorithm).....	135
ภาคผนวก ง การออกแบบการทำงานของเครื่องมือทดสอบ.....	139
ภาคผนวก จ สรุปผลค่าประสิทธิภาพความแม่นยำ (Precision), ความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ของ เครื่องมือทดสอบการค้นคืน.....	160
ภาคผนวก ฉ สรุปผลค่าเฉลี่ยที่ปรากฏในแต่ละกลุ่มและผลการทดลองกำหนดกลุ่ม ให้กับข้อสอบถาม.....	172
ภาคผนวก ช การวิเคราะห์ผลการทดลองทางสถิติของเครื่องมือทดสอบการค้นคืน เอกสารรูปแบบที่ 1 และรูปแบบที่ 1*.....	174
ประวัติผู้เขียนวิทยานิพนธ์.....	183

สารบัญตาราง

ตาราง		หน้า
ตารางที่ 2.1	ตารางแสดงความถี่ของค่าในชุดเอกสาร.....	20
ตารางที่ 2.2	ตารางแสดงค่า idf ของค่าในชุดเอกสาร.....	20
ตารางที่ 4.1	ตารางแสดงผลการจัดกลุ่มเอกสารด้วยการกำหนดจำนวนกลุ่มที่แตกต่าง กัน.....	57
ตารางที่ 4.2	ตารางแสดงจำนวนกลุ่มที่มีความเหมาะสมสำหรับการแบ่งกลุ่มเอกสาร นิตยสารพิมพ์	58
ตารางที่ 4.3	ตารางค่าสถิติของกลุ่มเมื่อกำหนดจำนวนกลุ่มเท่ากับ 15 กลุ่ม	59
ตารางที่ 4.4	ตารางค่าสถิติของกลุ่มเมื่อกำหนดจำนวนกลุ่มเท่ากับ 20 กลุ่ม	60
ตารางที่ 4.5	ตารางค่าสถิติของกลุ่มเมื่อกำหนดจำนวนกลุ่มเท่ากับ 25 กลุ่ม	61
ตารางที่ 4.6	ตารางแสดงลักษณะของกลุ่มตามจำนวนเอกสารที่ปรากฏ.....	62
ตารางที่ 4.7	ตารางแสดงจำนวนเอกสารในกลุ่มและจำนวนกลุ่มของลักษณะของกลุ่ม ต่างๆ.....	62
ตารางที่ 4.8	ตารางแสดงร้อยละของการกระจายตัวของกลุ่มเอกสารในกลุ่มขนาดต่างๆ...	63
ตารางที่ 4.9	ตารางสรุปค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ของเครื่องมือทดสอบ การค้นคืนเอกสารรูปแบบที่ 1 ที่กำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน...	65
ตารางที่ 4.10	ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิก (Harmonic mean), ค่าความ แม่นยำ (Precision) และค่าความระลึก (Recall) ของเครื่องมือทดสอบการ ค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2	66
ตารางที่ 4.11	ตารางสรุปผลการทดลองของค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ย ฮาร์โมนิกของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2	70
ตารางที่ 4.12	ตารางสรุปผลค่าเฉลี่ยฮาร์โมนิก ของเครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1 ภายใต้กรอบค่าความคล้ายคลึงที่ต่างกัน และด้วยการพิจารณา ค้นคืนเฉพาะเอกสารที่มีค่าในข้อสอบถามปรากฏเท่านั้น.....	72
ตารางที่ 4.13	ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิก (Harmonic mean), ค่าความ แม่นยำ (Precision) และค่าความระลึก (Recall) ของเครื่องมือทดสอบการ ค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 1*.....	73
ตารางที่ 4.14	ตารางสรุปผลการทดลองของค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ย	

	ฮาร์โมนิกของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 1*	77
ตารางที่ 4.15	ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิก (Harmonic mean), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1* และรูปแบบที่ 2	79
ตารางที่ 4.16	ตารางสรุปผลการทดลองของค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิกของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1* และรูปแบบที่ 2	83
ตารางที่ 4.17	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าความแม่นยำ	86
ตารางที่ 4.18	ตารางแสดงค่าสถิติทดสอบค่าความแม่นยำของการค้นคืนเอกสาร	87
ตารางที่ 4.19	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าความระลึก	88
ตารางที่ 4.20	ตารางแสดงค่าสถิติทดสอบค่าความระลึกของการค้นคืนเอกสาร	90
ตารางที่ 4.21	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าเฉลี่ยฮาร์โมนิก	91
ตารางที่ 4.22	ตารางแสดงค่าสถิติทดสอบค่าเฉลี่ยฮาร์โมนิกของการค้นคืนเอกสาร	93
ตารางที่ 4.23	ตารางแสดงตัวอย่างการคำนวณค่าความเหมือนระหว่างเวกเตอร์เอกสารและข้อสอบถามในแต่ละมิติ	95
ตารางที่ 5.1	ตารางแสดงผลการจัดกลุ่มเอกสารด้วยค่าความคล้ายคลึงเชิงมุม และกำหนดจำนวนกลุ่มที่แตกต่างกัน	102
ตารางที่ 5.2	ตารางแสดงผลการทดลองกำหนดจำนวนกลุ่มของเอกสาร 20 กลุ่ม	103
ตารางที่ 5.3	ตารางสรุปค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ที่กำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน	105
ตารางที่ 5.4	ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิก (Harmonic mean), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2	106
ตารางที่ 5.5	ตารางสรุปผลการทดลองของค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิกของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2	110
ตารางที่ จ.1	ตารางแสดงค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และ	

	ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) เครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1 เมื่อกรอบค่าความคล้ายคลึงที่ต่างกัน.....	160
ตารางที่ จ.2	ตารางแสดงค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และ ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) เครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 1* เมื่อกรอบค่าความคล้ายคลึงที่ต่างกัน.....	164
ตารางที่ จ.3	ตารางแสดงค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และ ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) เครื่องมือทดสอบการค้นคืนเอกสาร รูปแบบที่ 3 เมื่อกรอบค่าความคล้ายคลึงที่ต่างกัน.....	168
ตารางที่ ฉ.1	ตารางแสดงจำนวนค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานที่ปรากฏในแต่ละ กลุ่ม.....	172
ตารางที่ ฉ.2	ตารางแสดงผลการจัดกลุ่มให้กับทุกข้อสอบถาม.....	173
ตารางที่ ช.1	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่า ความแม่นยำ.....	175
ตารางที่ ช.2	ตารางแสดงค่าสถิติทดสอบค่าความแม่นยำของการค้นคืนเอกสาร.....	176
ตารางที่ ช.3	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่า ความระลึก.....	177
ตารางที่ ช.4	ตารางแสดงค่าสถิติทดสอบค่าความระลึกของการค้นคืนเอกสาร.....	179
ตารางที่ ช.5	ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าเฉลี่ย ฮาร์โมนิค.....	180
ตารางที่ ช.6	ตารางแสดงค่าสถิติทดสอบค่าเฉลี่ยฮาร์โมนิคของการค้นคืนเอกสาร.....	181

สารบัญภาพ

ภาพประกอบ	หน้า
รูปที่ 2.1	รูปแสดงเป้าหมายหลักในการค้นคืนสารสนเทศ..... 10
รูปที่ 2.2	รูปแสดงกระบวนการพื้นฐานระบบการค้นคืนสารสนเทศ..... 11
รูปที่ 2.3	รูปแสดงการทำงานทั้งหมดของระบบการค้นคืนเอกสาร..... 12
รูปที่ 2.4	รูปแสดงการสร้างแฟ้มผกผัน..... 15
รูปที่ 2.5	รูปแสดงเวกเตอร์สเปซของระบบมิติ 3 มิติ..... 17
รูปที่ 2.6	รูปแสดงตัวอย่างชุดเอกสารในระบบ..... 19
รูปที่ 2.7	รูปแสดงการทำมุมระหว่างเวกเตอร์เอกสารและข้อสอบถาม..... 22
รูปที่ 2.8	รูปแสดงแนวคิดของเทคนิค K-means Clustering..... 24
รูปที่ 2.9	รูปแสดงขั้นตอนการทำงานของเทคนิค K-means Clustering..... 25
รูปที่ 2.10	รูปแสดงเซตของเอกสารที่เกี่ยวข้องเนื่องและเซตของเอกสารที่ค้นคืน..... 27
รูปที่ 3.1	รูปแสดงภาพรวมของเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 2 รูปแบบ..... 42
รูปที่ 3.2	รูปแสดงองค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร..... 44
รูปที่ 3.3	รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 1..... 48
รูปที่ 3.4	รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 2..... 50
รูปที่ 3.5	รูปแสดงขั้นตอนการทดสอบประสิทธิภาพของเครื่องมือทดสอบทั้ง 2 รูปแบบ..... 53
รูปที่ 4.1	รูปแสดงกราฟเปรียบเทียบค่าความแม่นยำ ค่าความระลึกลับ และค่าเฉลี่ย ฮาร์โมนิค ระหว่างเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 และ รูปแบบที่ 2..... 70
รูปที่ 4.2	รูปแสดงกราฟเปรียบเทียบ ค่าความแม่นยำ, ค่าความระลึกลับ และค่าเฉลี่ย ฮาร์โมนิค ระหว่างเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 และ รูปแบบที่ 1* 77
รูปที่ 4.3	รูปแสดงกราฟเปรียบเทียบ ค่าความแม่นยำ, ค่าความระลึกลับ และค่าเฉลี่ย ฮาร์โมนิค ระหว่างเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1* และ รูปแบบที่ 2..... 83
รูปที่ 5.1	รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 3..... 99
รูปที่ 5.2	รูปแสดงขั้นตอนการทำงานของเทคนิค K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุม..... 101

รูปที่ 5.3	รูปแสดงกราฟเปรียบเทียบค่าความแม่นยำ ค่าความระลึกล และค่าเฉลี่ยฮาร์โมนิค ระหว่างเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2.....	110
รูปที่ ง.1	รูปแสดงแผนภาพการไหลของข้อมูลบริบท (Context Diagram) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering).....	140
รูปที่ ง.2	รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 1 (Data Flow Diagram Level 1) ของการค้นคืนเอกสารรูปแบบที่ 1.....	141
รูปที่ ง.3	รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของ Process 1 เก็บข้อมูลคำที่ปรากฏในเอกสาร, ข้อสอบถาม...	142
รูปที่ ง.4	รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของ Process 6 คำนวณค่าความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม.....	143
รูปที่ ง.5	รูปแสดงแผนภาพการไหลของข้อมูลบริบท (Context Diagram) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle).....	144
รูปที่ ง.6	รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 1 (Data Flow Diagram Level 1) ของการค้นคืนเอกสารรูปแบบที่ 2.....	145
รูปที่ ง.7	รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของ Process 1 เก็บข้อมูลคำที่ปรากฏและสร้างเวกเตอร์.....	146
รูปที่ ง.8	รูปแสดงขั้นตอนการทำงานหน้า 1 ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)	148
รูปที่ ง.9	รูปแสดงขั้นตอนการทำงานหน้า 2 ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)	149

รูปที่	ง.10	รูปแสดงขั้นตอนการทำงานหน้าที่ 1 ของการค้นคืนเอกสารที่ใช้เทคนิค ปริภูมิเวกเตอร์ด้วยวิธีวัดความคล้ายคลึงเชิงมุม (Cosine Angle).....	151
รูปที่	ง.11	รูปแสดงขั้นตอนการทำงานหน้าที่ 2 ของการค้นคืนเอกสารที่ใช้เทคนิค ปริภูมิเวกเตอร์ด้วยวิธีวัดความคล้ายคลึงเชิงมุม (Cosine Angle).....	152
รูปที่	ง.12	รูปแสดงแผนภาพเชิงแนวคิด (Conceptual Diagram).....	153
รูปที่	ง.13	รูปแสดงแผนภาพเชิงกายภาพ (Physical Diagram).....	154
รูปที่	ง.14	รูปแสดงหน้าจอแรก สำหรับเลือกข้อสอบถามที่ต้องการทดสอบ.....	158
รูปที่	ง.15	รูปแสดงหน้าจอยืนยันการเลือกข้อสอบถามที่ต้องการทดสอบ.....	158
รูปที่	ง.16	รูปแสดงหน้าจอผลการค้นคืนเอกสาร.....	159