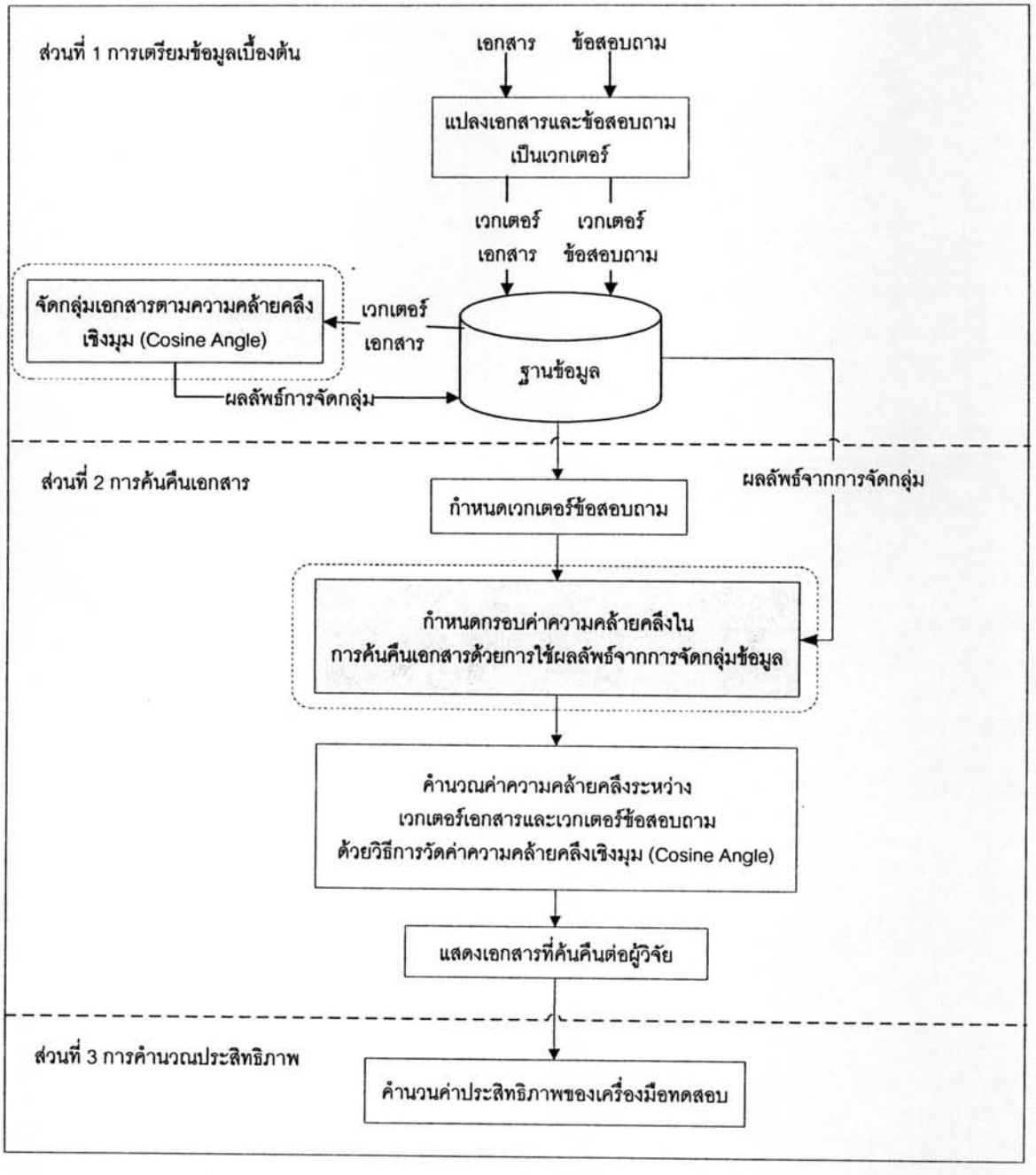
บทที่ 5

การศึกษาเชิงสำรวจ

ผู้วิจัยมีความต้องการศึกษาเพิ่มเติม ว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ตามความคล้ายคลึงเชิงมุมนั้น มีประสิทธิภาพการค้นคืนเอกสารแตกต่างกับการไม่ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) มากำหนดกรอบค่าความคล้ายคลึงการค้นคืนหรือไม่ โดยเรียกการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ตามความคล้ายคลึงเชิงมุม ว่า "เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3" โดยศึกษาเปรียบเทียบประสิทธิภาพการค้นคืนเอกสารด้วยค่าประสิทธิภาพ 3 ค่า คือ ค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ผลการวิเคราะห์ผลการทดลอง แสดงดังรายละเอียดต่อไปนี้

5.1 เครื่องมือทดสอบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ตามความคล้ายคลึงเชิงมุม

การทำงานของเครื่องมือทดสอบรูปแบบที่ 3 แสดงได้ดังรูปที่ 5.1 ซึ่งส่วนที่ได้เพิ่มเติมจากการทำงานของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 แสดงภายในกรอบเส้นประสี่เหลี่ยม



รูปที่ 5.1 รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 3

จากรูปข้างต้นแสดงการทำงานเครื่องมือทดสอบการค้นคืนรูปแบบที่ 3 ซึ่งแบ่งการทำงานออกเป็น 3 ส่วน ได้แก่

ส่วนที่ 1 การเตรียมข้อมูลเบื้องต้น เป็นส่วนของการเตรียมข้อมูลเอกสาร และข้อสอบถามให้อยู่ในรูปแบบเวกเตอร์ ซึ่งเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 มีกระบวนการทำงานในส่วนนี้เหมือนกันกับเครื่องมือทดสอบการค้นคืนรูปแบบที่ 1 และ 2 (ดังในหัวข้อ 3.5.1) แต่ได้

เพิ่มส่วนการทำงานของการจัดกลุ่มเอกสารตามความคล้ายคลึงเชิงมุม (Cosine Angle) โดยก่อนการทำงานในส่วนที่ 2 เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ต้องเตรียมผลลัพธ์ของการจัดกลุ่มเอกสาร ซึ่งหลังจากการแปลงเอกสารให้อยู่ในรูปแบบเวกเตอร์ ผู้วิจัยจะนำเวกเตอร์ของเอกสารทั้งหมดที่อยู่ในฐานข้อมูลมาทำการจัดกลุ่มของเอกสารด้วยการใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-means Clustering ด้วยการวัดค่าความความเหมือนโคไซน์ และนำผลลัพธ์จากการจัดกลุ่มที่ได้เก็บลงในฐานข้อมูล

ส่วนที่ 2 การค้นคืนเอกสาร เป็นส่วนแสดงขั้นตอนการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถามเชิงมุม ซึ่งจะมีการทำงานที่เหมือนกับการค้นคืนเอกสารรูปแบบที่ 2 (ดังที่ได้กล่าวในบทที่ 3.5.2) แตกต่างกันในส่วนของการกำหนดเงื่อนไข (threshold) ในการแสดงเอกสารต่อผู้ใช้ ซึ่งเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 จะกำหนดเงื่อนไขในการค้นคืนด้วยค่าความเหมือนต่ำสุดเท่ากับค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าความเหมือนของทุกข้อสอบถามกับทุกเอกสาร ซึ่งจะเป็นค่าที่คงที่สำหรับทุกข้อสอบถาม แต่ในเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 นั้น จะกำหนดเงื่อนไข (threshold) หรือกรอบค่าความคล้ายคลึงของเอกสารที่จะถูกนำมาแสดง ด้วยการกำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) บนระยะห่างเชิงมุม ซึ่งจะถูกเปลี่ยนแปลงไปตามแต่ละกลุ่มของเอกสารที่ใกล้เคียงกับข้อสอบถาม

ส่วนที่ 3 การคำนวณประสิทธิภาพ เป็นส่วนของการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบ เพื่อพิจารณาว่าเครื่องมือทดสอบนั้นสามารถค้นคืนเอกสารมีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด ใช้วิธีการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบรูปแบบที่ 3 ด้วยค่าประสิทธิภาพการค้นคืน 3 ค่า คือ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)

5.1.1 เทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ตามความคล้ายคลึงเชิงมุม

ในขั้นตอนนี้เครื่องมือทดสอบการค้นคืนรูปแบบที่ 3 เป็นขั้นตอนการทำงานที่เหมือนกับเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 แตกต่างกันในส่วนของการวัดระยะห่างระหว่างเอกสาร ซึ่งเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 วัดระยะห่างระหว่างเอกสารด้วย ค่าความเหมือนระยะห่างยูคลิเดียน ในขณะที่เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 วัดระยะห่างระหว่างเอกสารด้วย ค่าความเหมือนโคไซน์

ขั้นตอนการทำงานของ เทคนิค K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุม แสดงได้ดังรูปที่ 5.2 เครื่องมือที่ใช้ในการพัฒนาเทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุม คือ ภาษาซี (C programming language) โดยใช้ตัวแบบโปรแกรม kmeans.c จาก Zhang (2005) แต่อย่างไรก็ตาม โปรแกรม kmeans.c เป็นโปรแกรมที่ช่วยในการจัดกลุ่มโดยใช้ระยะห่างยูคลิเดียน ผู้วิจัยได้ดาวน์โหลดโปรแกรมนี้ และทำการดัดแปลงให้สามารถจัดกลุ่มได้โดยใช้ระยะห่างเชิงมุม

1) สุ่มเลือกเอกสารเท่ากับจำนวนกลุ่มที่ต้องการแบ่ง เพื่อนำมาเป็นจุดศูนย์กลาง (center) ของแต่ละกลุ่มเอกสารในรอบแรก
2) หาขอบเขต (boundaries) ระหว่างกลุ่มเอกสาร โดยขอบเขตของกลุ่มเอกสารที่อยู่ติดกันคือกึ่งกลางระหว่างมุมของจุดศูนย์กลางของกลุ่มเอกสารทั้งสอง
3) กำหนดกลุ่มให้กับเอกสารแต่ละเอกสาร โดยพิจารณาจากตำแหน่งที่อยู่ของเอกสารนั้นว่าทำมุมอยู่ภายใต้ขอบเขตของกลุ่มเอกสารใด
4) คำนวณจุดศูนย์กลางของกลุ่มทุกกลุ่มใหม่ วนการทำงานซ้ำไปยังข้อ 2
5) ถ้าจุดศูนย์กลางของกลุ่มเอกสาร หรือขอบเขตระหว่างกลุ่มเอกสารไม่เปลี่ยนแปลงแล้วก็จะหยุดการทำงาน

รูปที่ 5.2 รูปแสดงขั้นตอนการทำงานของเทคนิค K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุม

5.1.2 ผลการทดลองเทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ตามความคล้ายคลึงเชิงมุม

ผลลัพธ์จากการพัฒนาเครื่องมือเทคนิคการจัดกลุ่มเอกสารแบบ K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุมนั้น จะระบุค่าสถิติต่าง ๆ ดังนี้

- Frequency of Cluster แสดงจำนวนเอกสารภายในกลุ่มนั้น ๆ
- Nearest Cluster แสดงกลุ่มเอกสารที่อยู่ใกล้กับกลุ่มเอกสารนั้น ๆ
- Average Similarity within Cluster แสดงมุมเฉลี่ยภายในกลุ่มเอกสารนั้น ๆ

- Similarity to Nearest Cluster แสดงมุมที่กระทำกันระหว่างกลุ่มเอกสารที่อยู่ใกล้กัน ซึ่งมุมที่กระทำระหว่างกลุ่มนี้ ผู้วิจัยศึกษาวิธีการคำนวณมุมระหว่างกลุ่ม 2 วิธีการด้วยกัน คือ
    1) Centroid Linkage วิธีการนี้จะคำนวณจากมุมระหว่างจุดศูนย์กลางของทั้งสองกลุ่ม
    2) Single Linkage วิธีการนี้จะคำนวณมุมที่น้อยที่สุดที่กระทำระหว่างเอกสารของทั้งสองกลุ่ม
- Minimum Similarity from Cluster Seed แสดงมุมกว้างที่สุดระหว่างเอกสารกับ จุดศูนย์กลางของกลุ่มเอกสารนั้น ๆ

ผู้วิจัยได้ทดลองกำหนดกลุ่มให้กับเอกสารที่ต้องการแบ่งเป็น 5, 10, 15, 20, 25 และ 30 กลุ่ม ด้วยเครื่องมือการจัดกลุ่มเอกสารแบบ K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุมนั้น โดยผลการทดลองที่ได้แสดงได้ดังตารางที่ 5.1

ตารางที่ 5.1 ตารางแสดงผลการจัดกลุ่มเอกสารด้วยค่าความคล้ายคลึงเชิงมุม และกำหนดจำนวนกลุ่มที่แตกต่างกัน

Cluster (K)	Average Similarity within Cluster	Similarity to Nearest Cluster		Minimum Similarity from Cluster Seed
		Centroid Linkage	Single Linkage	
5	0.2463	0.3929	0.4693	0.0771
10	0.3377	0.1290	0.3008	0.2845
15	0.3559	0.2757	0.3177	0.1670
20	0.4051	0.2392	0.3091	0.2109
25	0.3996	0.2720	0.3020	0.2538
30	0.4310	0.2914	0.3235	0.2540

* หมายเหตุ ค่าที่แสดงในตารางแสดงด้วยค่าความเหมือนโคไซน์

การพิจารณาเลือกจำนวนกลุ่มเอกสาร (K) ที่เหมาะสมสำหรับในงานวิจัยนี้ สามารถพิจารณาด้วยการวิเคราะห์ค่าสถิติต่าง ๆ ที่ได้จากผลการทดลองการจัดกลุ่มด้วยความคล้ายคลึงเชิงมุม ดังตารางข้างต้น โดยหลักเกณฑ์ในการวิเคราะห์ลักษณะของค่าสถิติต่าง ๆ แสดงได้ดังนี้

- Average Similarity within Cluster เป็นค่าสถิติแสดงค่าความเหมือนโคไซน์เฉลี่ยภายในกลุ่มนั้น ๆ ถ้าค่าสถิตินี้มากจะแสดงว่าเอกสารภายในกลุ่มมีความเหมือนกันมาก และเอกสารภายในกลุ่มทำมุมใกล้กัน
- Similarity to Nearest Cluster เป็นค่าสถิติแสดงระยะห่างค่าความเหมือนโคไซน์เฉลี่ยของกลุ่มเอกสารที่อยู่ใกล้กัน ถ้าค่าสถิติระยะห่างความเหมือนโคไซน์ระหว่างกลุ่มเอกสารทั้ง 2 มีค่าน้อย แสดงว่ากลุ่มเอกสารทั้ง 2 กลุ่มทำมุมห่างกัน
- Minimum Similarity from Cluster Seed เป็นค่าสถิติแสดงค่าความเหมือนโคไซน์น้อยสุดระหว่างเอกสารกับจุดศูนย์กลางของกลุ่มเอกสารนั้น ๆ ถ้าเอกสารในกลุ่มมีการกระจุกตัว ค่าสถิตินี้ควรมีค่ามาก นั่นคือเอกสารในกลุ่มทำมุมระหว่างกันน้อยหรือมีค่าความคล้ายคลึงกันมาก

ดังนั้น จากขั้นตอนการพิจารณาเลือกจำนวนกลุ่มของเอกสาร (K) ที่เหมาะสมตามผลการทดลองที่ได้จากเครื่องมือการจัดกลุ่มเอกสารแบบ K-means Clustering ด้วยการวัดความคล้ายคลึงเชิงมุมข้างต้นนั้น ผู้วิจัยจึงกำหนดจำนวนกลุ่มของเอกสาร (K) สำหรับการจัดกลุ่มเอกสารเป็น 20 กลุ่ม เพราะมีค่าสถิติทั้ง 4 ค่าที่เหมาะสมมากกว่าการกำหนดจำนวนกลุ่มค่าอื่น ๆ

ตารางที่ 5.2 ตารางแสดงผลการทดลองกำหนดจำนวนกลุ่มของเอกสาร 20 กลุ่ม

Cluster->20					
Cluster	Frequency of Cluster	Nearest Cluster	Average Similarity within Cluster	Similarity to Nearest Cluster	Minimum Similarity from Cluster Seed
1	45	16	0.31321	0.326298	0.119001
2	17	8	0.432405	0.217783	0.107297
3	13	17	0.380917	0.202111	0.242257
4	9	17	0.353162	0.201661	0.226232
5	2	16	0.679131	0.076005	0.455869
6	3	11	0.595404	0.156277	0.563763
7	2	17	0.704148	0.080718	0.552562
8	24	1	0.371805	0.299533	0.154805
9	12	11	0.495925	0.184965	0.153235
10	13	15	0.346091	0.337312	0.162736

Cluster	Frequency of Cluster	Nearest Cluster	Average Similarity within Cluster	Similarity to Nearest Cluster	Minimum Similarity from Cluster Seed
11	41	20	0.352122	0.228561	0.066526
12	6	1	0.470111	0.158916	0.380814
13	22	8	0.335572	0.293353	0.108405
14	8	15	0.396802	0.245397	0.24277
15	57	10	0.270573	0.337312	0.08123
16	34	1	0.231323	0.326298	0.118854
17	51	20	0.242125	0.349915	0.083628
18	18	17	0.401719	0.205907	0.155839
19	28	17	0.42055	0.206484	0.078349
20	20	17	0.309587	0.349915	0.164668

ผลการทดลองจัดกลุ่มเวกเตอร์เอกสารนิตยสารไทม์ (TIME Magazine) จำนวน 425 เอกสาร โดยกำหนดจำนวนกลุ่มเอกสารที่ต้องการจัดกลุ่ม 20 กลุ่ม ได้ผลการทดลองดังตารางที่ 5.2 จากผลการทดลองจะพบว่าค่าสถิติ Minimum Similarity from Cluster Seed จะแสดงมุมไกลที่สุดระหว่างเอกสารกับจุดศูนย์กลางของกลุ่มเอกสารนั้น ๆ หรือก็คือ ค่าความเหมือนน้อยสุดระหว่างเอกสารกับจุดศูนย์กลางของกลุ่มเอกสาร ดังนั้น ผู้วิจัยจึงนำค่า Minimum Similarity from Cluster Seed ในแต่ละกลุ่ม มากำหนดกรอบค่าความคล้ายคลึงของการค้นคืน และผู้วิจัยจะเรียกค่า Minimum Similarity from Cluster Seed ว่า "รัศมีความคล้าย (RS)"

ผู้วิจัยนำรัศมีความคล้าย (RS) นี้มากำหนดกรอบค่าความคล้ายคลึงในการค้นคืนจากข้อสอบถามไปเป็นรัศมีความคล้าย (RS) โดยมีจุดศูนย์กลางเป็นข้อสอบถาม ซึ่งเวกเตอร์เอกสารใดทำมุมกับเวกเตอร์ข้อสอบถามภายในรัศมีความคล้ายที่กำหนด เวกเตอร์เอกสารนั้นจะถูกค้นคืนออกมาแสดงต่อผู้วิจัย

5.2 ผลการทดลอง

5.2.1 การวัดประสิทธิภาพการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ด้วยการกำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน

เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 นั้น จะนำค่ารัศมีความคล้าย ดังที่ได้กล่าวในหัวข้อ 5.1 มากำหนดกรอบค่าความคล้ายคลึงของการค้นคืน ดังนั้นผู้วิจัยได้ทดลองเปรียบเทียบ

ประสิทธิภาพการค้นคืนด้วยการกำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน คือ กำหนดกรอบค่าความคล้ายคลึงด้วยค่ารัศมีความคล้าย และค่ารัศมีความคล้ายหารสอง

ซึ่งผลการทดลองค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) จากการกำหนดกรอบค่าความคล้ายคลึงด้วยค่ารัศมีความคล้าย (RS) และค่ารัศมีความคล้ายหารสอง (RS/2) แสดงในภาคผนวก จ และสามารถสรุปได้ดังตารางที่ 5.3

ตารางที่ 5.3 ตารางสรุปค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ที่กำหนดกรอบค่าความคล้ายคลึงที่ต่างกัน

ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)	กรอบค่าความคล้ายคลึงด้วยค่ารัศมีความคล้าย (RS)	กรอบค่าความคล้ายคลึงด้วยค่ารัศมีความคล้ายหารสอง (RA/2)
ค่าเฉลี่ย	0.4056	0.3271
ค่าเบี่ยงเบนมาตรฐาน	0.2849	0.2413

ตารางที่ 5.3 แสดงให้เห็นว่า ค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคที่ได้จากการทดลองเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ด้วยการกำหนดกรอบค่าความคล้ายคลึงค่ารัศมีความคล้าย (RS) มีค่ามากกว่าการกำหนดกรอบค่าความคล้ายคลึงค่ารัศมีความคล้ายหารสอง (RS/2)

ดังนั้น ผู้วิจัยจึงกำหนดกรอบค่าความคล้ายคลึงในการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ด้วยค่ารัศมีความคล้าย (RS) และดำเนินการทดลองเปรียบเทียบค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) กับเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ต่อไป

5.2.2 ผลการทดลองเปรียบเทียบประสิทธิภาพของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2

ค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ที่ได้จากการทดลองแสดงได้ดังตารางที่ 5.4

ตารางที่ 5.4 ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิค (Harmonic mean), ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2

ลำดับข้อสอบถาม	ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)		ค่าความแม่นยำ (Precision)		ค่าความระลึก (Recall)	
	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2
1	0.4828	0.4118	0.3182	0.2593	1.0000	1.0000
2	0.1481	0.1111	0.0800	0.0588	1.0000	1.0000
3	0.1818	0.1154	0.1034	0.0625	0.7500	0.7500
4	0.2439	0.2222	0.1389	0.1250	1.0000	1.0000
5	0.2703	0.2326	0.1563	0.1316	1.0000	1.0000
6	0.4500	0.3913	0.2903	0.2432	1.0000	1.0000
7	0.2857	0.2000	0.1667	0.1111	1.0000	1.0000
8	0.3636	0.2353	0.2222	0.1333	1.0000	1.0000
9	0.7500	0.6364	0.7500	0.5000	0.7500	0.8750
10	0.8333	0.5217	0.8333	0.3529	0.8333	1.0000
11	0.6667	0.2353	1.0000	0.1333	0.5000	1.0000
12	0.8750	0.5600	0.7778	0.3889	1.0000	1.0000
13	0.6000	0.3158	0.4286	0.1875	1.0000	1.0000
14	0.0000	0.3333	0.0000	0.2000	0.0000	1.0000
15	0.7273	0.5714	0.6667	0.4444	0.8000	0.8000
16	0.5000	0.2727	0.4000	0.1579	0.6667	1.0000
17	0.4444	0.2353	0.2857	0.1333	1.0000	1.0000
18	0.0000	0.3333	0.0000	0.2000	0.0000	1.0000
19	0.5556	0.4167	0.3846	0.2632	1.0000	1.0000

ลำดับข้อสอบถาม	ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)		ค่าความแม่นยำ (Precision)		ค่าความระลึก (Recall)	
	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2
20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	0.2667	0.1250	0.1538	0.0667	1.0000	1.0000
22	0.0000	0.2000	0.0000	0.1111	0.0000	1.0000
23	1.0000	0.4000	1.0000	0.2500	1.0000	1.0000
24	0.2500	0.0909	0.1429	0.0476	1.0000	1.0000
25	0.1333	0.1000	0.0714	0.0526	1.0000	1.0000
26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
27	0.0000	0.2222	0.0000	0.1250	0.0000	1.0000
28	0.2500	0.2424	0.3333	0.1429	0.2000	0.8000
29	0.5714	0.2105	0.4000	0.1176	1.0000	1.0000
30	0.5333	0.2941	0.4000	0.1724	0.8000	1.0000
31	0.0000	0.3478	0.0000	0.2500	0.0000	0.5714
32	0.1053	0.0400	0.0556	0.0204	1.0000	1.0000
33	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
34	1.0000	0.2857	1.0000	0.1667	1.0000	1.0000
35	0.4000	0.1818	0.2500	0.1000	1.0000	1.0000
36	1.0000	0.5000	1.0000	0.3333	1.0000	1.0000
37	0.5000	0.2222	0.5000	0.1429	0.5000	0.5000
38	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
39	0.8571	0.5000	0.7500	0.3333	1.0000	1.0000
40	0.4615	0.3913	0.7500	0.2432	0.3333	1.0000
41	0.5455	0.2703	0.6000	0.1613	0.5000	0.8333

ลำดับข้อสอบถาม	ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)		ค่าความแม่นยำ (Precision)		ค่าความระลึก (Recall)	
	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2
42	0.0000	0.0526	0.0000	0.0270	0.0000	1.0000
43	0.5000	0.1739	0.3333	0.0952	1.0000	1.0000
44	0.5000	0.2667	0.3333	0.1538	1.0000	1.0000
45	0.0000	0.2353	0.0000	0.1379	0.0000	0.8000
46	0.6667	0.5667	0.7333	0.4048	0.6111	0.9444
47	0.2500	0.3571	0.5000	0.2273	0.1667	0.8333
48	0.0000	0.1667	0.0000	0.0909	0.0000	1.0000
49	0.8000	0.5161	0.8571	0.3478	0.7500	1.0000
50	0.5000	0.0833	0.3333	0.0435	1.0000	1.0000
51	0.4000	0.1579	0.5000	0.0857	0.3333	1.0000
52	0.6667	0.0500	1.0000	0.0263	0.5000	0.5000
53	0.6667	0.1905	0.5000	0.1053	1.0000	1.0000
54	0.2000	0.0769	0.1111	0.0400	1.0000	1.0000
55	0.6667	0.4490	0.6667	0.2973	0.6667	0.9167
56	0.3333	0.0909	0.2000	0.0476	1.0000	1.0000
57	0.2222	0.2222	0.1429	0.1250	0.5000	1.0000
58	0.6316	0.5185	0.5455	0.3684	0.7500	0.8750
59	0.5714	0.1600	0.4000	1.0000	1.0000	0.0870
60	0.4444	0.1818	0.2857	1.0000	1.0000	0.1000
61	0.6429	0.6842	0.6923	0.8667	0.6000	0.5652
62	0.5000	0.2222	0.3333	1.0000	1.0000	0.1250
63	0.6875	0.4400	0.5238	1.0000	1.0000	0.2821

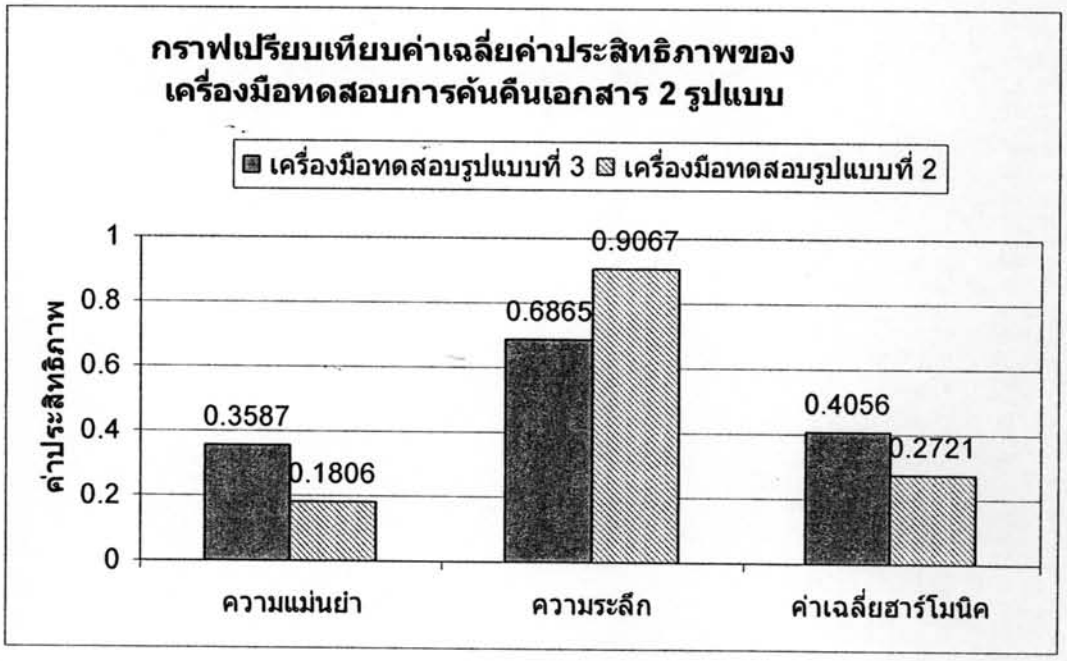
ลำดับข้อสอบถาม	ค่าเฉลี่ยฮาร์โมนิค (Harmonic mean)		ค่าความแม่นยำ (Precision)		ค่าความระลึก (Recall)	
	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2	เครื่องมือทดสอบรูปแบบที่ 3	เครื่องมือทดสอบรูปแบบที่ 2
64	0.4444	0.1905	0.2857	1.0000	1.0000	0.1053
65	0.1250	0.0870	0.0667	1.0000	1.0000	0.0455
66	0.2222	0.1429	0.1250	1.0000	1.0000	0.0769
67	0.6000	0.3750	0.4286	1.0000	1.0000	0.2308
68	0.4000	0.3429	0.3333	0.7500	0.5000	0.2222
69	0.9630	0.7879	0.9286	1.0000	1.0000	0.6500
70	0.0000	0.1333	0.0000	1.0000	0.0000	0.0714
71	0.4000	0.5714	0.5000	0.6667	0.3333	0.5000
72	0.1667	0.0667	0.0909	1.0000	1.0000	0.0345
73	0.4000	0.1538	0.2500	1.0000	1.0000	0.0833
74	0.2500	0.1905	0.1429	1.0000	1.0000	0.1053
75	0.0000	0.0833	0.0000	1.0000	0.0000	0.0435
76	0.5000	0.3030	0.3636	1.0000	0.8000	0.1786
77	0.0000	0.0833	0.0000	1.0000	0.0000	0.0435
78	0.3077	0.1538	0.1818	1.0000	1.0000	0.0833
79	0.5000	0.0952	0.3333	1.0000	1.0000	0.0500
80	0.3333	0.4571	0.5714	0.4706	0.2353	0.4444
81	0.1176	0.0364	0.0667	0.5000	0.5000	0.0189
82	0.5455	0.3704	0.5000	1.0000	0.6000	0.2273
83	0.2857	0.1212	0.2000	1.0000	0.5000	0.0645

จากตารางที่ 5.4 สามารถสรุปผลการทดลองค่าเฉลี่ยฮาร์โมนิค, ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วย

วิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) บนระยะห่างเชิงมุม ด้วยค่ารัศมีความคล้าย (RS) และวิธีการวัดความคล้ายคลึงเชิงมุมที่ไม่ได้ใช้เทคนิคการจัดกลุ่มข้อมูลเข้าร่วม ดังตารางที่ 5.5 และเปรียบเทียบค่าเฉลี่ยค่าประสิทธิภาพของการค้นคืนเอกสารดังรูปที่ 5.3

ตารางที่ 5.5 ตารางสรุปผลการทดลองของค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค ของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2

		เครื่องมือทดสอบการค้นคืนเอกสาร	
		รูปแบบที่ 3	รูปแบบที่ 2
ค่าความแม่นยำ	ค่าเฉลี่ย	0.3587	0.1806
	ค่าเบี่ยงเบนมาตรฐาน	0.3040	0.1661
ค่าความระลึก	ค่าเฉลี่ย	0.6865	0.9067
	ค่าเบี่ยงเบนมาตรฐาน	0.3855	0.2198
ค่าเฉลี่ยฮาร์โมนิค	ค่าเฉลี่ย	0.4056	0.2721
	ค่าเบี่ยงเบนมาตรฐาน	0.2849	0.1917



รูปที่ 5.3 รูปแสดงกราฟเปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค ระหว่างเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 2

ตารางที่ 5.5 แสดงให้เห็นว่า ค่าเฉลี่ยค่าความแม่นยำของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 3 มากกว่าเครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 เท่ากับ 0.1781 ค่าเฉลี่ยค่าความระลึกของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 3 น้อยกว่าเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 เท่ากับ 0.2202 และค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 3 มากกว่าเครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 เท่ากับ 0.1335 ดังนั้น ค่าเฉลี่ยค่าความแม่นยำ และค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 โดยเทคนิควิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบความคล้ายคลึงค่ารัศมีความคล้าย (RS) มีค่าเฉลี่ยมากกว่าเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 แต่ค่าเฉลี่ยของค่าความระลึกของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 มีค่าเฉลี่ยน้อยกว่าเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2

5.3 สรุปผลการวิเคราะห์การศึกษาเชิงสำรวจ

จากการวิเคราะห์เพิ่มเติมในส่วนการวัดประสิทธิภาพการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูลบนระยะห่างเชิงมุม เมื่อทดสอบและวิเคราะห์ผลการทดลองแล้วสรุปได้ว่า ค่าความแม่นยำ (Precision) ของการค้นคืนเอกสารรูปแบบที่ 3 มากกว่าการค้นคืนเอกสารรูปแบบที่ 2 เท่ากับ 98.61% ค่าความระลึก (Recall) ของการค้นคืนเอกสารรูปแบบที่ 3 น้อยกว่าการค้นคืนเอกสารรูปแบบที่ 2 เท่ากับ 24.29% และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ของการค้นคืนเอกสารรูปแบบที่ 3 มากกว่าการค้นคืนเอกสารรูปแบบที่ 2 เท่ากับ 49.06%

5.4 อภิปรายผลการศึกษาเชิงสำรวจ

จากการศึกษาเพิ่มเติมเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ซึ่งแตกต่างกับเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ในส่วนของการกำหนดเงื่อนไข (threshold) ในการแสดงเอกสารต่อผู้ใช้ โดยเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 จะกำหนดเงื่อนไขในการแสดงเอกสารด้วยค่าที่คงที่สำหรับทุกข้อสอบถาม แต่เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 นั้น จะกำหนดเงื่อนไข หรือกรอบค่าความคล้ายคลึงของเอกสารที่จะถูกนำมาแสดง ด้วยการกำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) บนระยะห่างเชิงมุม ซึ่งจะถูกเปลี่ยนแปลงไปตามแต่ละกลุ่มของเอกสารที่มีความใกล้เคียงกับข้อสอบถาม

ผลการทดลองและวิเคราะห์ข้อมูลสรุปได้ว่า การค้นคืนเอกสารรูปแบบที่ 3 มีประสิทธิภาพค่าความแม่นยำ และค่าเฉลี่ยฮาร์โมนิคดีกว่าการค้นคืนเอกสารรูปแบบที่ 2 แต่มีประสิทธิภาพค่าความระลึกที่ต่ำกว่า ซึ่งหมายความว่า เอกสารที่ถูกค้นคืนมีความเกี่ยวเนื่องกับความต้องการมาก แต่ไม่ได้เป็นเอกสารที่เกี่ยวเนื่องกับความต้องการทั้งหมด อาจจะมีสาเหตุมาจากการค้นคืนเอกสารรูปแบบที่ 3 กำหนดเงื่อนไขการแสดงเอกสาร หรือกรอบค่าความคล้ายคลึงการค้นคืนที่จะถูกเปลี่ยนแปลงไปตามลักษณะกลุ่มของเอกสารที่ข้อสอบถามมีความคล้ายคลึง ในขณะที่การค้นคืนเอกสารรูปแบบที่ 2 กำหนดกรอบการค้นคืนด้วยค่าความเหมือนต่ำสุดที่คงที่สำหรับทุกข้อสอบถาม ซึ่งเอกสารใดที่มีค่าความคล้ายคลึงกับข้อสอบถามมากกว่าค่าความเหมือนต่ำสุด เอกสารนั้นก็จะเป็นผลลัพธ์ของการค้นคืน ทำให้เอกสารถูกค้นคืนออกมาเป็นจำนวนมาก และมีความเป็นไปได้ที่เอกสารที่เกี่ยวเนื่องกับความต้องการจะถูกค้นคืนออกมาเป็นจำนวนมากด้วย

ทั้งนี้ ผู้วิจัยตั้งข้อสังเกตว่า ปรากฏการณ์ดังกล่าว อาจเกิดจากการที่เทคนิคการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) มีความแม่นยำในการระบุเอกสารที่มีความคล้ายคลึงกับข้อสอบถามอยู่แล้ว กล่าวคือ เวกเตอร์เอกสารที่มีความคล้ายคลึงกับเวกเตอร์ข้อสอบถามมากที่สุดจะอยู่ใกล้กับเวกเตอร์ของข้อสอบถามมากที่สุด เวกเตอร์ของเอกสารที่มีความคล้ายคลึงกับข้อสอบถามในลำดับที่ 2 ก็จะอยู่ห่างจากเวกเตอร์ของข้อสอบถามเป็นลำดับถัดไป ดังนั้น การกำหนดกรอบค่าความคล้ายคลึงที่เปลี่ยนแปลงไปตามรัศมีความคล้ายของกลุ่มเอกสาร ส่งผลให้กรอบค่าความคล้ายคลึงแคบลง เอกสารที่เป็นผลลัพธ์ของการค้นคืนจึงมีปริมาณน้อยลง และเป็นเอกสารที่มีความคล้ายคลึงกับข้อสอบถามมาก จึงอาจส่งผลให้ค่าความแม่นยำสูงขึ้น แต่ค่าความระลึกต่ำลง