การค้นคืนสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้

นางสาวศิรัตน์ ศิรนานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2549
ISBN 974-14-2059-5

INFORMATION RETRIEVAL USING ASSOCIATION RULES TOGETHER WITH RELEVANT FEEDBACK

Miss Sirat Sirananon

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 2006

ISBN 974-14-2059-5

หัวข้อวิทยานิพนธ์	การค้นคืนสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้
โดย	นางสาว ศิรัตน์ ศิรนานนท์
สาขาวิชา	การพัฒนาซอฟต์แวร์ด้านธุรกิจ
อาจารย์ที่ปรึกษา	อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะพาณิชยศาสตร์และการบัญชี
(ผู้ช่วยศาสตราจารย์ ดร. ดนุชา คุณพนิชกิจ)

คณะกรรมการสอบวิทยานิพนธ์

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

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

..... กรรมการ
(อาจารย์ ดร. พิมพ์มณี รัตนวิชา)

ศิรัตน์ ศิรนานนท์ : การค้นคืนสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้. (Information Retrieval using Association Rules together with Relevant Feedback)
อ. ที่ปรึกษา : อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน, 172 หน้า. ISBN 974-14-2059-5.

วิทยานิพนธ์นี้เสนอการทดสอบประสิทธิภาพของระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์และผลสะท้อนกลับจากผู้ใช้ โดยจะเปรียบเทียบกับระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์ของคำ ซึ่งในเทคนิคปริภูมิเวกเตอร์จะมีการแปลงเอกสารและข้อสอบถามให้อยู่ในรูปของเวกเตอร์ ส่วนเทคนิคกฎความสัมพันธ์เป็นเทคนิคของการทำเหมืองข้อมูล โดยหาความสัมพันธ์ของคำที่เกิดขึ้นพร้อมกันบ่อยครั้งในเอกสาร เพื่อเพิ่มคำที่มีความสัมพันธ์กับคำในข้อสอบถามก่อนนำไปใช้ดึงเอกสาร ส่วนเทคนิคผลสะท้อนกลับจากผู้ใช้คือเทคนิคที่ใช้ผลสะท้อนกลับจากผู้ใช้ในการปรับข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องกับข้อสอบถามมากยิ่งขึ้น

งานวิจัยนี้เป็นงานวิจัยเชิงทดลอง โดยใช้เอกสารนิตยสาร TIME จำนวน 425 เอกสารและข้อสอบถามจำนวน 83 ข้อสอบถามทดลองเปรียบเทียบประสิทธิภาพของระบบค้นคืนเอกสารโดยการคำนวณค่าเฉลี่ยฮาร์โมนิคของระบบค้นคืนเอกสารทั้ง 3 รูปแบบ ดังกล่าวข้างต้น

จากการวิเคราะห์ผลการทดลองสรุปได้ว่าระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์ของคำสามารถทำให้ประสิทธิภาพดีขึ้นกว่าการใช้เทคนิคปริภูมิเวกเตอร์ แต่เมื่อใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์และผลสะท้อนกลับจากผู้ใช้ทำให้ประสิทธิภาพของระบบการค้นคืนเอกสารมากกว่าการใช้เทคนิคปริภูมิเวกเตอร์เพียงอย่างเดียว แต่ต่ำกว่าการใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

ภาควิชา สถิติ ลายมือชื่อนิสิต ศิรัตน์ ศิรนานนท์
สาขาวิชา การพัฒนาซอฟต์แวร์ด้านธุรกิจ ลายมือชื่ออาจารย์ที่ปรึกษา
ปีการศึกษา 2549

4782391626 : MAJOR Business Software Development

KEY WORD: INFORMATION RETRIEVAL/ ASSOCIATION RULE / RELEVANT FEEDBACK

SIRAT SIRANANON : INFORMATION RETRIEVAL USING ASSOCIATION RULES TOGETHER WITH RELEVANT FEEDBACK. THESIS ADVISOR : JANJAO MONGKOLNAVIN, Ph.D., 172 pp. ISBN 974-14-2059-5.

This thesis presents an experimental study on using an information retrieval system that employs a vector space technique together with association rules and relevant feedback in comparison with a system that uses the vector space technique alone and a system that uses the vector space technique together with association rules. In vector space technique, documents and queries are transformed to be vectors, while association rules is a data mining technique that is used to find associations of words that appear in the same documents. The list of associated words can be used to extend the query vector before using it to retrieve a set of relevant documents. Such query vector can be refined further by applying relevant feedback which is a technique that adjusts the query vector according to user feedback on the list of documents which is the result from the first round of query. This is to make the query vector closer to the target documents.

In the thesis, the performance of the three information retrieval systems above is compared through Harmonic mean. The experiments were conducted on 425 documents and 83 queries from the TIME Magazine collection which is obtained from ftp://ftp.cs.cornell.edu/pub/smart/time.

The experimental results show that the information retrieval system that uses vector space with association rules has a best performance while the one using vector space together with association rule and relevant feedback shows a better performance than the one using vector space alone.

Department : Statistics Student's Signature : Sirat Siranonon

Field of Study : Business Software Development Advisor's Signature :

Academic Year 2006

กิตติกรรมประกาศ

วิทยานิพนธ์นี้จะสำเร็จลุล่วงและสมบูรณ์ไปได้ด้วยดีต้องขอกราบขอบพระคุณ อาจารย์ ดร. จันทร์เจ้า มงคลนาวิน อาจารย์ที่ปรึกษาวิทยานิพนธ์เป็นอย่างยิ่งที่ได้ให้คำแนะนำ และข้อคิดเห็นต่าง ๆ ตรวจแก้วิทยานิพนธ์ฉบับนี้อย่างละเอียด ตลอดจนแนวทางในการวิจัยด้วยดี ตลอดมา ขอขอบพระคุณอาจารย์ ผู้ช่วยศาสตราจารย์ ดร.อัษฎาพร ทรัพย์สมบูรณ์ และ อาจารย์ ดร.พิมพ์มณี รัตนวิชา กรรมการวิทยานิพนธ์ที่กรุณาเสียสละเวลาให้คำแนะนำ เพื่อแก้ไขรูปแบบและเนื้อหาวิทยานิพนธ์ฉบับนี้จนเสร็จสมบูรณ์ และขอบพระคุณอาจารย์ ดร. อรุณี กำลัง ที่ให้คำปรึกษาเกี่ยวกับการวิเคราะห์ข้อมูลในการทดลอง

ขอบคุณเพื่อน ๆ ที่ให้คำปรึกษาและความช่วยเหลือในด้านต่าง ๆ ซึ่งทำให้งานวิจัยเป็นไปได้อย่างราบรื่นตลอดจนกำลังใจที่มอบให้เสมอมา

สุดท้ายนี้ ผู้วิจัยใคร่กราบขอบพระคุณบิดา มารดาและครอบครัวที่คอยช่วยเหลือให้การสนับสนุนและคอยเป็นแรงกระตุ้นให้แก่ผู้วิจัยเสมอจนสำเร็จการศึกษา

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

ต้นฉบับไม่มีหน้า ฐ , ฑ

NO PAGE ฐ , ฑ IN ORIGINAL

สารบัญภาพ

หน้า

หน้า

บทที่ 1

ที่มาและความสำคัญของปัญหา

1.1 ความสำคัญของปัญหา

ปัจจุบันการขยายตัวทางด้านระบบสารสนเทศที่ใช้คอมพิวเตอร์มาช่วยในการทำงานได้มีอัตราเพิ่มสูงขึ้น ซึ่งส่งผลให้เกิดการเก็บข้อมูลชนิดข้อความในรูปแบบอิเล็กทรอนิกส์ (Electronic text document) เพิ่มขึ้นจำนวนมาก เนื่องจากเป็นรูปแบบข้อมูลที่ใช้กันอย่างแพร่หลาย เช่น หนังสือ วารสาร รายงานการประชุม เอกสารวิชาการ เป็นต้น และนับวันข้อมูลประเภทนี้จะเพิ่มปริมาณสูงขึ้นเรื่อย ๆ (Sullivan, 2001) ทำให้การค้นหาข้อมูลมีความลำบากมากยิ่งขึ้น ดังนั้นระบบการค้นคืนสารสนเทศประเภทนี้จึงเป็นสิ่งจำเป็น ในปัจจุบันได้มีการนำคอมพิวเตอร์มาประยุกต์ใช้งานร่วมกับการค้นคืนสารสนเทศอิเล็กทรอนิกส์ที่อยู่ในรูปแบบตัวอักษรหรือข้อความ ตัวอย่างเช่น โปรแกรมค้นคืนเว็บ (Web browser) โปรแกรมการทำงานสำหรับระบบงานแผนกช่วยเหลือ (Help desk system)[1] โปรแกรมค้นหาข่าว โปรแกรมค้นคืนเอกสารที่อยู่ในรูปแบบเอกซ์เอ็มแอล (XML : Extended Makeup Language) พจนานุกรมอิเล็กทรอนิกส์ ห้องสมุดอิเล็กทรอนิกส์และโปรแกรมค้นคืนเอกสารต่าง ๆ ในคลังเอกสาร เป็นต้น (Baeza-Yates and Ribeiro-Neto, 1999; Meadow et al., 2000; Chakrabarti, 2003; Weiss et al., 2005)

การค้นคืนข้อมูลชนิดข้อความ (Text) มักประสบปัญหาเรื่องผลลัพธ์ของการค้นคืนอาจไม่ตรงตามความต้องการของผู้ใช้ระบบ บ่อยครั้งผู้ใช้ระบบจะต้องเสียเวลามากในการค้นคืน เพื่อให้ได้มาซึ่งข้อมูลที่ตนเองต้องการ โดยเฉพาะอย่างยิ่งในกรณีผู้ใช้ขาดความรู้เกี่ยวกับข้อมูลที่ต้องการค้นหายิ่งทำให้ผู้ใช้เสียเวลามากขึ้นอีก ส่วนหนึ่งเนื่องมาจากการที่ผู้ใช้ไม่ทราบว่าคำใดเป็นคำที่เหมาะสมสำหรับการค้นหาข้อมูลที่ต้องการ ดังนั้นเมื่อผู้ใช้กรอกข้อสอบถาม (Query) ไม่เหมาะสมเข้ามาในระบบ ระบบจะค้นคืนข้อมูลที่ตรงกับข้อสอบถามของผู้ใช้ที่กรอกเข้ามา แต่ระบบจะไม่สามารถค้นคืนข้อมูลที่ตรงกับความต้องการของผู้ใช้ได้อย่างแท้จริง ทำให้ผู้ใช้ไม่ได้ผลลัพธ์ที่น่าพอใจในการค้นคืนข้อมูลแต่ละครั้ง

ปัญหาที่เกิดจากระบบค้นคืนข้อมูลส่วนใหญ่คือ ระบบจะค้นคืนเอกสารที่มีคำเหมือนกับคำในข้อสอบถาม (Query) เท่านั้น นอกจากนี้การดำเนินการดังกล่าวยังมีขีดจำกัดทางด้านการคำนวณ (Computation) คือ ระบบจะต้องไปเทียบคำทุกคำในเอกสารทั้งหมด อีกทั้งความซับซ้อน

---

[1] เป็นโปรแกรมที่รับเอกสารเกี่ยวกับปัญหาต่าง ๆ จากผู้ใช้ที่ส่งมายังระบบงานช่วยเหลือนี้ เพื่อให้ผู้ดูแลระบบแก้ไขปัญหาที่เกิดขึ้น

ของธรรมชาติทางภาษา (Natural Language Complexity) ยิ่งส่งผลให้ระบบไม่สามารถค้นคืนข้อมูลออกมาอย่างถูกต้องได้ เช่น เมื่อผู้ใช้กรอกคำว่า “Thailand” ระบบจะดึงข้อมูลเอกสารที่มีคำว่า “Thailand” ออกมา โดยที่บางเอกสารที่มีคำว่า “Thai” ปรากฏอยู่จะไม่ถูกดึงออกมาด้วย ซึ่งเอกสารเหล่านั้นอาจเป็นเอกสารที่ผู้ใช้ต้องการ ทำให้ผู้ใช้พลาดโอกาสในการได้ข้อมูลที่ถูกต้องได้ (Baeza-Yates and Ribeiro-Neto, 1999) โดยลักษณะของคำที่ทำให้ระบบค้นคืนข้อมูลส่วนใหญ่ไม่สามารถทำงานได้อย่างมีประสิทธิภาพสามารถสรุปได้ดังต่อไปนี้

(1) คำที่มีความหมายเหมือนกันแต่เขียนในรูปแบบต่างกัน เช่น คำว่า “home” มีความหมายเหมือนกับคำว่า “house” แต่เขียนอยู่ในรูปแบบที่ต่างกัน

(2) คำที่สะกดไม่เหมือนกันแต่มีความสัมพันธ์กัน (Correlation) เช่น ถ้าผู้ใช้กรอกข้อสอบถามคำว่า “basketball” เอกสารที่มีคำว่า “slam dunk” ปรากฏอยู่ควรจะถูกดึงขึ้นมาแสดงด้วย

(3) คำที่เป็นคำเดียวกันแต่ถูกแบ่งคำด้วยช่องว่าง ซึ่งอาจจะทำให้ระบบเข้าใจว่าคำนี้เป็นคำที่เป็นคำเดี่ยว 2 คำที่เป็นอิสระต่อกัน (Independence) เช่น คำว่า “United Nations” ระบบสารสนเทศจะค้นคืนเอกสารที่มีคำว่า “United” หรือ “Nations” คำใดคำหนึ่งปรากฏอยู่ออกมาด้วย อาจทำให้มีเอกสารที่มีคำว่า “United State of America” แสดงออกมา ซึ่งไม่ตรงกับความต้องการของผู้ใช้

คำประเภทดังที่กล่าวมาข้างต้นนั้นอาจทำให้ระบบค้นคืนสารสนเทศไม่สามารถแสดงผลลัพธ์ได้ตรงกับความต้องการที่แท้จริงของผู้ใช้ได้อย่างดีพอ (Baeza-Yates and Ribeiro-Neto, 1999; Chowdhury, 2004; Weiss et al., 2005) ดังนั้นในการกำหนดข้อสอบถาม (query formulation) คำ (term) ที่ปรากฏในข้อสอบถามจึงเป็นสิ่งสำคัญที่จะช่วยทำให้ระบบค้นคืนสารสนเทศมีประสิทธิภาพในการค้นคืนเอกสารมากขึ้น

ในปี ค.ศ.1974 เทคนิคการค้นคืนสารสนเทศ (Information Retrieval) มีการใช้คลังคำศัพท์ (Thesaurus) (Chowdhury, 2004) เป็นวิธีกำหนดหรือแนะว่าให้ใช้ศัพท์คำใดคำหนึ่งเป็นตัวแทนของกลุ่มคำที่มีรูปแบบต่างกันแต่มีความสัมพันธ์กัน เพื่อช่วยแก้ปัญหาคำที่มีความแปรผัน (Variant word)[1] ที่ไม่สามารถจัดการด้วยวิธีขั้นตอนธรรมดา (Simple algorithm) เทคนิคนี้จะหาคำเหมือนกันหรือมีความสัมพันธ์กันเพื่อช่วยในการกำหนดดรรชนี (Indexing) และการค้นหา (Searching) นอกจากนี้ยังมีการจัดลำดับชั้นในคลังคำศัพท์เพื่อที่จะขยายข้อสอบถาม (Query) ของผู้ใช้ไปทั้งทางด้านกว้าง (Broadening) คือ ขยายคำในข้อสอบถามให้มีความหมายกว้างขึ้นในระดับความหมายเท่ากัน เช่น คำว่า “Notebook” กับคำว่า “Personal Computer” โดยที่ทั้ง

---

[1] คำที่สะกดแตกต่างกันแต่สามารถใช้ในความหมายในเชิงเดียวกันได้ เช่นคำว่า “*post* the letter” และ ”*mail* the letter” (Robert, 1997)

“Notebook” และ “Personal Computer” เป็นชนิดของคอมพิวเตอร์เหมือนกัน และขยายคำในข้อสอบถามทางด้านลึก (Narrowing) คือ มีความหมายที่แคบลงไปในรายละเอียดของคำในข้อสอบถาม เช่นคำว่า “pet” กับคำว่า “dog” นั่นคือคำว่า “dog” จะมีความหมายที่แคบกว่าความหมายของคำว่า “pet” กล่าวคือคำว่าสุนัข (dog) นั้นจัดเป็นประเภทหนึ่งของสัตว์เลี้ยง (Pet) (Baeza-Yates and Ribeiro-Neto, 1999) ซึ่งเมื่อผู้ใช้ต้องการค้นคืนสารสนเทศ ในระบบสารสนเทศที่มีการใช้คลังคำศัพท์ ระบบค้นคืนสารสนเทศ (Information Retrieval System) จะค้นคืนเอกสารที่มีคำ (Term) ที่สะกดตรงกัน คำที่มีความหมายเหมือนกัน และคำที่มีความหมายใกล้เคียง โดยถูกจัดไว้อยู่ในกลุ่มความหมายพวกเดียวกันออกมาแสดงด้วย วิธีนี้จึงเป็นวิธีที่ช่วยกำหนดข้อสอบถาม (Query formulation) ในส่วนของการจัดการคำในข้อสอบถาม ทำให้การดึงข้อมูลมีประสิทธิภาพมากยิ่งขึ้น (Robert,1997; Yates and Neto,1999; Chowdhury, 2004)

นอกจากนี้ยังสามารถเพิ่มประสิทธิภาพคลังคำศัพท์ให้รู้จักคำที่ไม่ใช่คำเดี่ยว[1] อย่างเดียวด้วยวิธีการเชื่อมคำเดี่ยวให้เป็นคำเดียวกัน (Multiword Feature) ที่เรียกว่าคำผสม[2] โดยการขยายขนาดของพจนานุกรม (Dictionary) ให้ไม่เก็บคำเดี่ยวเพียงอย่างเดียว แต่ยังเก็บคำผสมอีกด้วย เช่น คลังคำศัพท์จะเก็บคำว่า ”search engine” ด้วย นอกจากจะเก็บคำว่า “search” กับ “engine” เท่านั้น ซึ่งวิธีนี้เป็นอีกวิธีหนึ่งที่จะช่วยให้ระบบพิจารณาคำที่ผู้ใช้กำหนดมาในข้อสอบถามอย่างมีประสิทธิภาพมากขึ้น (Weiss et al., 2005) เช่น เมื่อผู้ใช้ต้องการเอกสารที่เกี่ยวกับ “search engine” โดยจะกรอกคำว่า “search” เข้ามายังระบบ ระบบต้องไม่จะค้นคืนเอกสารที่มี ”search warrant” (หมายค้น) ปรากฏออกมาด้วย

แม้ว่าคลังคำศัพท์ช่วยเพิ่มประสิทธิภาพให้กับการจัดการคำในข้อสอบถาม คือสามารถดึงข้อมูลให้ตรงตามความต้องการของผู้ใช้มากขึ้น วิธีคลังคำศัพท์นั้นก็ยังคงทำงานได้ไม่ดีในกรณีที่ความสัมพันธ์ระหว่างคำนั้นยังไม่ถูกระบุอย่างถูกต้องหรือยากแก่การระบุ โดยเฉพาะอย่างยิ่งในกรณีที่ข้อสอบถามของผู้ใช้เป็นบริบทที่เป็นศัพท์เฉพาะในสาขาของผู้ใช้ นอกจากนี้วิธีนี้ยังไม่เหมาะกับระบบที่ต้องการการประมวลผลที่รวดเร็ว เนื่องจากก่อนที่ระบบนำข้อสอบถามไปเทียบความเหมือนกับเอกสารและต้องไปเทียบข้อสอบถามกับคำในคลังคำศัพท์เสียก่อน (Baeza-Yates and Ribeiro-Neto, 1999)

---

[1] คำที่เกิดจากการนำตัวอักษรมาเรียงต่อกันโดยไม่มีเว้นวรรคคั่นและจะต้องเป็นอิสระกับคำอื่น ๆ คือ ไม่มีความหมายร่วมกับคำอื่นด้วย (Weiss et al., 2005) เช่น boat water fish เป็นต้น

[2] คำที่ประกอบขึ้นมาจากคำเดี่ยวตั้งแต่ 2 คำขึ้นไปและมีความหมาย เช่น Personal Computer เป็นต้น

ปัจจุบันมักมีการนำเทคนิคเหมืองข้อมูล (Data Mining) มาประยุกต์ใช้กับข้อมูลเอกสารเพื่อการค้นคืนสิ่งที่ซ่อนเร้นอยู่ในเอกสารปริมาณมาก (Text Mining) เทคนิคที่พบว่ามีการใช้กันทั่วไปคือ เทคนิคการค้นพบกฎความสัมพันธ์ (Association Rule Discovery) ซึ่งเป็นการหาความสัมพันธ์หรือความใกล้ชิดกันของเซตข้อมูล (Item set) ที่เกิดขึ้นในธุรกรรม (Transactions) จำนวนมาก เช่น สำหรับข้อมูลรายการสินค้าในซุปเปอร์มาร์เกต เราอาจพบว่าในธุรกรรมที่เกิดขึ้นส่วนใหญ่มีการขายลูกพลัม มะเขือเทศและผักกาดพร้อมกันบ่อยครั้ง จากการค้นพบดังกล่าวนี้ทำให้เราสามารถพิจารณาถึงความสัมพันธ์ที่เกิดขึ้นจากความถี่ของการเกิดขึ้นพร้อมกันของรายการสินค้าทั้ง 3 ชนิดได้ ซึ่งเทคนิคดังกล่าวจะสามารถนำมาช่วยในการหาความสัมพันธ์ของคำที่เกิดขึ้นในกลุ่มของเอกสาร โดยจะพิจารณาถึงคำที่ปรากฏพร้อมกันในเอกสารส่วนใหญ่ แต่วิธีนี้จะไม่ครอบคลุมถึงไวยากรณ์ภาษา เนื่องจากระบบจะพิจารณาถึงความถี่ของคำที่เกิดพร้อม ๆ กัน โอกาสที่คำแต่ละคำจะมีความสัมพันธ์กันนั้นจะสูงขึ้นตามอัตราการเกิดของคำที่เกิดขึ้นพร้อมกัน ซึ่งอาจจะทำให้สามารถค้นพบคำที่มีความสัมพันธ์แต่ไม่สามารถกำหนดได้ด้วยนักภาษาศาสตร์หรือผู้เชี่ยวชาญทางด้านนั้น ๆ ได้ โดยเฉพาะเอกสารทางด้านธุรกิจที่จะใช้คำที่เป็นคำทั่ว ๆ ไป ไม่เหมือนกับเอกสารทางด้านวิทยาศาสตร์ที่ใช้คำที่เป็นศัพท์เทคนิค (Technical term) ที่มีการกำหนดความสัมพันธ์กันไว้แล้ว นอกจากนั้นวิธีนี้ยังสามารถจัดการกับข้อมูลที่มีปริมาณมาก ๆ ได้อีกด้วย (Ye, 2001)

จากงานวิจัยที่ผ่านมาเกี่ยวกับการใช้กฎความสัมพันธ์มาช่วยงานระบบค้นคืนสารสนเทศของข้อมูลประเภทเอกสารมีมากมาย ซึ่งสามารถยกตัวอย่างได้ดังนี้

1) การปรับปรุงข้อสอบถามด้วยวิธีการค้นหากฎความสัมพันธ์ เพื่อให้การค้นคืนเอกสารมีประสิทธิภาพมากขึ้น โดยการขยายคำในข้อสอบถามและค้นคืนเอกสารที่เกี่ยวข้องกับข้อสอบถามมาแสดงเป็นผลลัพธ์ (Delgado et al., 2002; Song et al., 2005)

2) การหาความสัมพันธ์ของคำเพื่อการจัดกลุ่มเอกสาร โดยเอกสารที่มีคำที่เกี่ยวข้องกันปรากฏอยู่จะถูกจัดอยู่ในกลุ่มเดียวกันเพื่อให้การค้นหาข้อมูลง่ายยิ่งขึ้น เนื่องจากเอกสารได้ถูกจัดเป็นหมวดหมู่ไว้ล่วงหน้าแล้ว เช่น เอกสารที่มีคำว่า “basketball” ปรากฏอยู่และเอกสารที่มีคำว่า “slam dunk” ถ้าระบบค้นพบว่าคำสองคำนี้มักจะเกิดขึ้นพร้อม ๆ กัน ระบบจะจัดเอกสารที่มีคำสองคำนี้ให้อยู่ในหมวดหมู่เดียวกัน เมื่อมีการค้นคืนเอกสารเกิดขึ้นนอกจากระบบจะค้นคืนเอกสารที่มีคำนั้นแล้ว ระบบจะค้นคืนเอกสารที่มีคำที่อยู่ในกลุ่มเดียวกันปรากฏอยู่ออกมาด้วย (Kou and Gardarin, 2002; Antonie and Zaiane, 2002; Zhuang and Dai, 2004)

3) การกำหนดคำสำคัญหรือดึงคำสำคัญที่มีความสัมพันธ์กันในเอกสาร โดยใช้การหาความสัมพันธ์ของคำที่ปรากฏในเอกสาร โดยจะมีวิธีการคือ กำหนดเอกสารที่อยู่กลุ่มเดียวกันและเอกสารที่ไม่อยู่กลุ่มเดียวกันไว้ก่อน ถ้าเอกสารที่อยู่กลุ่มเดียวกันปรากฏคำใด ๆ ด้วยความถี่มาก ๆ จะแสดงว่าคำเหล่านั้นมีความสัมพันธ์กัน (Matsumura et al., 2002; Matsuo and Ishizuka, 2003)

และแม้ว่าจะมีงานวิจัยที่จะขยายคำในข้อสอบถาม (Query expansion) ด้วยเทคนิคการค้นพบกฎความสัมพันธ์ (Association Rule Discovery) แต่ยังไม่พบว่ามีงานวิจัยใดที่ให้ผู้ใช้ระบบให้ผลสะท้อนกลับ (Feedback) ร่วมด้วยกับเทคนิคดังกล่าว เพื่อปรับปรุงค่าน้ำหนักของคำในข้อสอบถามโดยสะท้อนกลับมายังระบบ ซึ่งเป็นวิธีพัฒนาการขยายคำในข้อสอบถาม (Query expansion) โดยใช้ผลสะท้อนกลับที่เกี่ยวเนื่องกับความต้องการ (Relevant feedback) โดยได้เสนอออกมาช่วงกลางทศวรรษที่ 60 (Chowdhury, 2004) เป็นการขยายคำในข้อสอบถามโดยพิจารณาถึงความต้องการของผู้ใช้ โดยจะสามารถขยายคำหรือลดคำและปรับเปลี่ยนค่าน้ำหนักให้กับคำต่าง ๆ ที่ขยายและลดในข้อสอบถาม ระบบจะปรับปรุงข้อสอบถามของผู้ใช้ให้เข้าใกล้กลุ่มเอกสารที่ตรงกับความต้องการของผู้ใช้จากการที่ผู้ใช้ที่ป้อนผลสะท้อนกลับเกี่ยวกับเอกสารที่เกี่ยวเนื่องและไม่เกี่ยวเนื่องกับความต้องการกลับมา ทำให้ระบบจะสามารถค้นหาผลลัพธ์ที่ตรงกับความต้องการของผู้ใช้มากขึ้น เช่น เมื่อผู้ใช้กรอกข้อสอบถามคำว่า “mountain” เข้ามาเพื่อสืบค้นเอกสาร ระบบจะค้นคืนเอกสารออกมาแสดงต่อผู้ใช้ จากนั้นผู้ใช้จะให้ผลสะท้อนเกี่ยวกับความถูกต้องของเอกสารกลับมายังระบบ ระบบจะนำผลสะท้อนกลับนั้นไปปรับปรุงข้อสอบถามด้วยการพิจารณาคำในเอกสารที่ผู้ใช้กำหนดความถูกต้องเข้ามาโดยจะเพิ่มน้ำหนักให้แก่คำที่ปรากฏในเอกสารที่ถูกต้อง และลดค่าน้ำหนักของคำในเอกสารที่ไม่ถูกต้อง เมื่อได้ข้อสอบถามที่ปรับเปลี่ยนคำและค่าน้ำหนักแล้ว ระบบจะนำข้อสอบถามใหม่ที่ได้ไปค้นคืนเอกสารอีกครั้ง (Baeza-Yates and Ribeiro-Neto, 1999; Chowdhury, 2004)

ด้วยเหตุนี้งานวิจัยนี้จึงขอเสนอการทดสอบการค้นคืนเอกสารโดยการหาความสัมพันธ์ของคำ ตามวิธีของการทำเหมืองข้อมูลข้อความ (Text Mining) ด้วยเทคนิคที่เรียกว่าการค้นหากฎความสัมพันธ์ (Association Rule Discovery) ร่วมกับเทคนิคการใช้ผลสะท้อนกลับ (Relevant feedback) จากผู้ใช้เพื่อปรับค่าน้ำหนักของคำในข้อสอบถามว่าสามารถเพิ่มประสิทธิภาพการค้นคืนเอกสารให้ตรงตามความต้องการของผู้ใช้มากขึ้นหรือไม่

1.2 วัตถุประสงค์ของการวิจัย

1. เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร (Document Retrieval) ที่ใช้เทคนิคการค้นหากฎความสัมพันธ์กับการค้นคืนเอกสารที่ไม่ใช้เทคนิคการค้นหากฎความสัมพันธ์
2. เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร (Document Retrieval) ที่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ไม่ใช้ 2 เทคนิคดังกล่าว
3. เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร (Document Retrieval) ที่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคกฎความสัมพันธ์ของคำ

1.3 ขั้นตอนโดยสรุปของการทำวิจัย

1. ศึกษารายละเอียดเกี่ยวกับวิธีการของระบบการค้นคืนสารสนเทศที่มีอยู่ในปัจจุบัน
2. ศึกษารายละเอียดเกี่ยวกับวิธีของการทำเหมืองข้อมูล (Data Mining) ในการหากฎความสัมพันธ์ (Association Rule)
3. ศึกษาเกี่ยวกับวิธีการปรับปรุงข้อสอบถามจากผลสะท้อนกลับจากผู้ใช้
4. ออกแบบเครื่องมือทดสอบตามที่ศึกษา
5. กำหนดค่าน้ำหนักในการปรับปรุงข้อสอบถามเมื่อผู้ใช้ให้ผลสะท้อนกลับ (Relevant Feedback)
6. พัฒนาเครื่องมือทดสอบตามที่ได้ออกแบบไว้
7. ทดสอบการทำงานของเครื่องมือที่พัฒนา
8. ประเมินผลการทำงานของเครื่องมือ
9. วิเคราะห์ผลการทดลองและสำรวจข้อมูลเพิ่มจากผลการทดลอง
10. จัดทำเอกสารสรุปงานวิจัย และข้อเสนอแนะ

1.4 ตัวแปรที่ศึกษา

1. **ตัวแปรอิสระ (Independent variables)** มีจำนวนสองตัวแปร ได้แก่

เทคนิคการค้นคืนเอกสารในงานวิจัยนี้สนใจเทคนิค 2 เทคนิคร่วมกันคือ เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Discovery) และเทคนิคการให้ผลสะท้อนกลับ (Relevant Feedback) เนื่องจากผู้วิจัยต้องการศึกษาการปรับปรุงประสิทธิภาพของระบบค้นคืนเอกสารให้

ตรงความต้องการผู้ใช้มากขึ้น ด้วยการใช้เทคนิคทั้ง 2 เทคนิคร่วมกันดังกล่าว ดังนั้นงานวิจัยนี้จะศึกษาเปรียบเทียบประสิทธิภาพการค้นคืนของระบบค้นคืนเอกสารทั้งหมด 3 รูปแบบดังนี้

1) ระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้
2) ระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ
3) ระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

2. ตัวแปรตาม (Dependent variables) คือ

ประสิทธิภาพของระบบการค้นคืนเอกสาร คือ ระบบสามารถค้นคืนเอกสารที่ตรงความต้องการของผู้ใช้ออกมาได้มากที่สุด โดยมีเอกสารที่ไม่ตรงกับความต้องการของผู้ใช้น้อยที่สุด โดยงานวิจัยนี้จะวัดประสิทธิภาพโดยการหาค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) (รายละเอียดกล่าวในบทที่ 2)

3. ตัวแปรอื่นที่เกี่ยวข้อง มีจำนวน 3 ตัวแปร ได้แก่

3.1 เอกสาร คือ บทความในฐานข้อมูลของระบบค้นคืนเอกสารในงานวิจัยนี้
3.2 ข้อสอบถาม คือ คำหรือประโยคที่ผู้วิจัยกรอกเข้ามายังระบบเพื่อต้องการค้นคืนเอกสารที่เกี่ยวเนื่องกับคำในข้อสอบถามที่กรอกเข้าไป
3.3 ความถูกต้องระหว่างเอกสารและข้อสอบถาม
3.4 ความถูกต้องของผลสะท้อนกลับจากผู้ใช้
3.5 เครื่องมือพัฒนาระบบค้นคืนเอกสาร

1.5 ขอบเขตของการวิจัย

1. เอกสารที่จะนำมาใช้ที่จะพัฒนาขึ้นในการศึกษาจะเป็นบทความของนิตยสารไทม์ (TIME Magazine) ในปี 1963 เท่านั้น ซึ่งเป็นบทความภาษาอังกฤษ
2. การทดสอบประสิทธิภาพการค้นคืนเอกสารนี้จะไม่ครอบคลุมถึงคำกำกวม (Ambiguity) เช่นคำว่า “apple” ซึ่งสามารถแปลความหมายได้ทั้งเป็นชื่อสินค้าคอมพิวเตอร์หรือผลไม้
3. การทดสอบประสิทธิภาพการค้นคืนเอกสารจะไม่พิจารณาถึงไวยากรณ์ภาษา
4. เครื่องมือที่พัฒนาจะรับผลสะท้อนกลับจากผู้ใช้ในการแสดงผลลัพธ์ครั้งแรกมาเพื่อปรับปรุงข้อสอบถามเพียงครั้งเดียวเท่านั้น

5. เครื่องมือทดสอบเทคนิคการค้นคืนเอกสารนี้สร้างขึ้นเพื่อทดสอบกับชุดเอกสารและชุดข้อสอบถามของไทม์ (TIME Collection) เท่านั้น

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้ที่ต้องการพัฒนาระบบค้นคืนเอกสาร สามารถนำงานวิจัยนี้ไปเป็นแนวทางสำหรับประยุกต์ใช้เข้ากับระบบค้นคืนเอกสารที่พัฒนา เพื่อปรับปรุงประสิทธิภาพให้ดียิ่งขึ้น
2. ผลการทดสอบประสิทธิภาพระบบค้นคืนเอกสารที่ได้สามารถเป็นแนวทางในการนำเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการใช้เทคนิคกฎความสัมพันธ์ของคำร่วมกับผลสะท้อนกลับจากผู้ใช้ ที่เสนอในงานวิจัยนี้ไปเลือกใช้ได้อย่างเหมาะสม
3. ผู้ที่ต้องการพัฒนาระบบค้นคืนเอกสารสามารถนำเทคนิคการใช้กฎความสัมพันธ์ของคำไปประยุกต์ใช้กับเอกสารที่ไม่สามารถระบุความสัมพันธ์กันได้

บทที่ 2

วรรณกรรมที่เกี่ยวข้อง

2.1 บทนำ

ในส่วนนี้จะกล่าวถึงทฤษฎีที่สำคัญ งานวิจัยต่าง ๆ ที่นำมาประยุกต์ใช้ในงานวิจัยและข้อจำกัดของงานวิจัยในอดีต งานวิจัยนี้มีวัตถุประสงค์ที่จะเพิ่มประสิทธิภาพระบบค้นคืนเอกสารให้ตรงกับความต้องการของผู้ใช้มากขึ้น โดยประยุกต์เทคนิคของความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้ ดังนั้นจึงมีทฤษฎีที่เกี่ยวข้องดังต่อไปนี้

2.2 เทคนิคการค้นคืนสารสนเทศ

เทคนิคการค้นคืนสารสนเทศมีจุดประสงค์หลักของการค้นคืนสารสนเทศคือ ต้องการค้นคืนสารสนเทศให้ตรงกับข้อสอบถามที่ผู้ใช้ป้อนเข้ามาให้มากที่สุดและมีสารสนเทศที่ไม่ตรงความต้องการให้น้อยที่สุด (Baeza-Yates and Ribeiro-Neto, 1999) การบวนการพื้นฐานของการค้นคืนสารสนเทศทั่วไปมีกระบวนการดังนี้

1. กำหนดข้อมูลเพื่อระบุขอบเขตข้อมูลที่ผู้ใช้ต้องการค้นหาจากระบบ เพื่อเก็บข้อมูลที่กำหนดนี้เข้าสู่ระบบ
2. วิเคราะห์ความหมายของข้อมูลนั้น ๆ
3. แสดงความหมายของข้อมูลในฐานข้อมูลให้อยู่ในรูปแบบที่สามารถเปรียบเทียบกับข้อสอบถามของผู้ใช้
4. วิเคราะห์ข้อสอบถามของผู้ใช้และแสดงข้อสอบถามในรูปแบบที่เหมาะสมกับข้อมูลในฐานข้อมูล
5. เปรียบเทียบข้อสอบถามกับข้อมูลในฐานข้อมูล
6. ดึงข้อมูลที่ตรงความต้องการ
7. ปรับระบบโดยใช้ผลสะท้อนกลับของผู้ใช้ (ขั้นตอนนี้อาจจะมีหรือไม่มีก็ได้)

จากกระบวนการพื้นฐานของระบบค้นคืนสารสนเทศสามารถนำมาประยุกต์ใช้ในการสร้างระบบค้นคืนเอกสารในงานวิจัยนี้ได้ โดยการสร้างระบบค้นคืนเอกสาร แบ่งเป็นขั้นตอนได้ดังต่อไปนี้

1. เมื่อได้เอกสารที่ต้องการเก็บเข้าสู่ระบบแล้วจะจัดเก็บเอกสารโดยดึงคำสำคัญออกจากเอกสารเก็บลงดรรชนี
2. กำหนดรูปแบบของเอกสารและข้อสอบถามในการค้นคืนเอกสาร
3. ขยายเวกเตอร์ข้อสอบถามโดยใช้เทคนิคกฎความสัมพันธ์ (Association Discovery)
4. เปรียบเทียบความเหมือนระหว่างเอกสารและข้อสอบถามออกมาแสดงต่อผู้ใช้
5. ตั้งค่าความเหมือนต่ำสุดในการค้นคืนเอกสาร
6. ปรับปรุงข้อสอบถามจากผลสะท้อนกลับจากผู้ใช้ เพื่อนำข้อสอบถามนั้นไปค้นคืนเอกสารอีกครั้ง เพื่อให้ระบบค้นคืนเอกสารค้นคืนเอกสารออกมาเกี่ยวเนื่องกับความต้องการของผู้ใช้มากยิ่งขึ้น
7. วัดประสิทธิภาพของระบบค้นคืนเอกสารว่าระบบสามารถค้นคืนเอกสารที่เกี่ยวเนื่องกับความต้องการของผู้ใช้ออกมาได้มากเพียงใด

จากขั้นตอนข้างต้นนั้นสามารถนำเทคนิคต่าง ๆ ที่เกี่ยวข้องมากำหนดรูปแบบของระบบค้นคืนเอกสาร โดยในแต่ละขั้นตอนมีเทคนิคที่เกี่ยวข้องดังต่อไปนี้

2.3 การดึงคำสำคัญออกจากเอกสารเก็บลงดรรชนี

เมื่อได้ชุดเอกสารมาแล้ว จะแปลงแต่ละเอกสารให้อยู่ในรูปแบบที่ระบบสามารถทำงานได้ โดยมีรายละเอียดดังต่อไปนี้

2.3.1 การกำจัดคำยกเว้น (Elimination of stop words) (Baeza-Yates and Ribeiro-Neto, 1999)

คำยกเว้น เป็นคำที่ปรากฏบ่อยมาก ๆ ในเอกสารและไม่สามารถใช้ในการแยกแยะเอกสารได้ เช่น คำว่า “the” , “a” , “and” และอื่น ๆ ซึ่งการตัดคำที่เป็นคำยกเว้นนั้นจะเป็นการลดคำศัพท์ที่เก็บในระบบด้วย งานวิจัยนี้กำหนดใช้คำยกเว้นจากสมาร์ท (SMART)[1] โดยมีรายการคำยกเว้น 571 คำ ดังแสดงในภาคผนวก ข

2.3.2 การลดรูปคำ (Stemming)

บ่อยครั้งที่ผู้ใช้กำหนดข้อสอบถามมาเพื่อค้นคืนเอกสารหรือในแต่ละเอกสารในฐานข้อมูล คำที่อยู่ในข้อสอบถามหรือเอกสารจะปรากฏอยู่ในเอกสารในรูปแบบต่าง ๆ เช่น คำนามที่อยู่ในรูปแบบของพหูพจน์ คำกริยาที่ตามหลังด้วย “ing” คำกริยาที่อยู่ในรูปของอดีต เป็นต้น คำที่อยู่ใน

[1] ระบบค้นคืนเอกสารที่พัฒนาโดย Cornell University กว่า 60 ปี โดยการใช้แบบจำลองปริภูมิเวกเตอร์ ซึ่งมีจุดประสงค์หลักเพื่อกำหนดกรอบงาน (Framework) ของระบบค้นคืนเอกสารในการทำวิจัย (Williamson and Lesk, 1971).

รูปแบบที่หลากหลายเหล่านี้จะถูกปรับเปลี่ยนให้อยู่ในรูปแบบเดียวกัน ด้วยวิธีการลดรูปคำ (Stemming) โดยการตัดส่วนที่เป็น prefix หรือ suffix ออก ประโยชน์ในการลดรูปคำจะช่วยเพิ่มจำนวนเอกสารที่ตรงตามความต้องการให้ถูกค้นคืนออกมามากขึ้น (Chakrabarti, 2003) โดยทั่วไปพบว่าขั้นตอนวิธีที่เป็นที่นิยมคือขั้นตอนวิธีของพอร์ทเตอร์ (Porter algorithm) (Porter, 1980). เนื่องจากเป็นวิธีที่เรียบง่ายและได้ผลดี (Baeza-Yates and Ribeiro-Neto, 1999) โดยวิธีนี้มีความคิดที่จะประยุกต์ใช้ชุดของกฎที่กำหนดขึ้นมาเพื่อเติมหรือปรับเปลี่ยนส่วนที่อยู่ท้ายของคำต่าง ๆ ในเอกสาร เช่น การเปลี่ยนคำที่อยู่ในรูปพหูพจน์เป็นเอกพจน์จะตัด s ที่อยู่ท้ายคำออก แสดงด้วยกฎ $s \rightarrow \phi$ วิธีของขั้นตอนวิธีของพอร์ทเตอร์นั้นสามารถช่วยให้ผลลัพธ์การลดรูปคำ (Stemming) ที่มีประสิทธิภาพและสามารถทำงานได้อย่างรวดเร็วอีกด้วย (Baeza-Yates and Ribeiro-Neto, 1999) (ขั้นตอนวิธีของพอร์ทเตอร์มีรายละเอียดดังแสดงในภาคผนวก ค)

2.3.3 การกำหนดดรรชนี (Baeza-Yates and Ribeiro-Neto, 1999; Chowdhury, 2004)

เมื่อชุดข้อมูลมีขนาดใหญ่มาก ๆ หรือมีเอกสารจำนวนมาก จำเป็นต้องเก็บดรรชนี เพื่อช่วยให้การค้นหามีความรวดเร็วยิ่งขึ้น ซึ่งรูปแบบของดรรชนีนั้นมีหลากหลายรูปแบบ ในที่นี้จะยกตัวอย่าง 3 รูปแบบ ดังต่อไปนี้

1) แฟ้มซิกเนเจอร์ (Signature file)

เป็นการค้นหาแบบแฮช (Hashing) โดยจะแปลงข้อความให้อยู่ในรูปแบบลำดับบิต คือ การกำหนดเลขลำดับ 0 หรือ 1 มาเรียงต่อกันแทนเป็น 1 ตัวอักษร เช่น ตัวอักษร “A” แทนด้วยลำดับ ”00000001” และข้อสอบถามที่เข้ามาจะถูกแปลงให้อยู่ในรูปแบบลำดับบิตเช่นกัน ในการเปรียบเทียบข้อสอบถามกับคำ จะคำนวณโดยใช้ฟังก์ชันแฮช แต่ค่าที่ได้จากการคำนวณมีโอกาสที่มีคำตอบมากกว่า 1 ที่ชี้ไปยังเอกสารต่างที่กัน ทำให้คำตอบที่ได้อาจไม่ถูกต้อง วิธีนี้ไม่เหมาะกับเอกสารขนาดใหญ่และเอกสารที่ไม่สามารถแบ่งคำได้ เช่น เอกสารที่มีข้อความเป็นลำดับโปรตีน (DNA) เป็นต้น

2) ต้นไม้ซัฟฟิก (Suffix tree)

การทำงานจะสร้างแถวลำดับ (Suffix array) เพื่อลดขนาดของเอกสาร และเป็นวิธีที่เปลี่ยนลำดับของตัวชี้ (Pointer) ไปยังเอกสารแบบสุ่มได้อย่างมีประสิทธิภาพ วิธีนี้จะทำให้การค้นหารวดเร็ว เนื่องจากช่วงเวลาในการค้นหาจะเปรียบเทียบคำที่ต้องการค้นหากับแถวลำดับซัฟฟิกบนเอกสาร ซึ่งจะมีโครงสร้างเป็นแบบต้นไม้นั่นคือเป็นลำดับชั้นลงไปเรื่อย ๆ โดยโครงสร้างนี้ทำให้การค้นหารวดเร็ว เนื่องจากไม่ต้องค้นหากับคำทั้งหมดในแถวลำดับนั่นคือสามารถตัดกิ่งที่ไม่ตรงตามความต้องการออกไปได้

3) **แฟ้มผกผัน** (Inverted file)

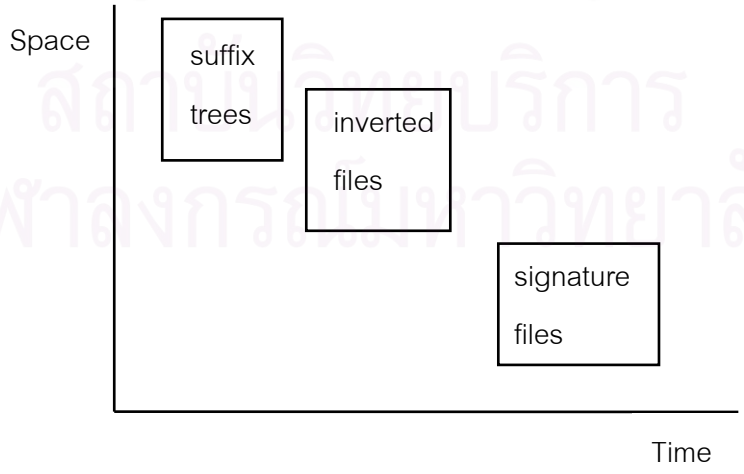
เป็นโครงสร้างที่มีหลักการคือจะเก็บเอกสารและตำแหน่งของคำที่เกิดขึ้นในเอกสารนั้นด้วย โดยจะแบ่งเป็น 2 ตารางคือตารางเก็บเอกสาร และตารางคำศัพท์ที่เก็บคำศัพท์ที่ปรากฏในเอกสารและตำแหน่งของคำนั้นในเอกสาร ดังรูปที่ 2.1 (Baeza-Yates and Ribeiro-Neto, 1999) วิธีนี้เป็นวิธีที่มีโครงสร้างรูปแบบง่าย ๆ แต่จะใช้พื้นที่ในการจัดเก็บมาก เนื่องจากต้องเก็บตารางดรรชนีที่ชี้ไปยังเอกสารด้วย

Text

ตำแหน่ง	1	6	9	11	17	19	24	28	33
คำศัพท์	This	is	a	text	.	A	text	has	many	word

Vocaburary	Co-occurrence
many	28
text	11,19
word	33

รูปที่ 2.1 รูปแสดงการสร้างแฟ้มผกผัน



รูปที่ 2.2 การเปรียบเทียบพื้นที่ในการเก็บดรรชนีและเวลาในการค้นคืนของดรรชนีทั้ง 3 เทคนิค (Baeza-Yates and Ribeiro-Neto, 1999)

จากรูปที่ 2.2 แสดงว่าเทคนิคการกำหนดดรรชนีแบบซัฟฟิกทรี (Suffix tree) เสียพื้นที่มากที่สุด แต่จะใช้เวลาในการค้นคืนน้อยที่สุด ส่วนเทคนิคการกำหนดดรรชนีแบบแฟ้มซิกเนเจอร์ (Signature file) จะใช้พื้นที่น้อยที่สุดและใช้เวลาในการค้นคืนมากที่สุด ส่วนเทคนิคการกำหนดดรรชนีแบบแฟ้มผกผัน (Inverted file) จะเสียพื้นที่มากกว่าเทคนิคการกำหนดดรรชนีแบบแฟ้มซิกเนเจอร์ (Signature file) แต่จะใช้เวลาในการค้นหามากกว่าเทคนิคการกำหนดดรรชนีแบบซัฟฟิกทรี (Suffix tree)

2.3.4 **คลังคำศัพท์ (Thesaurus)** (Baeza-Yates and Ribeiro-Neto, 1999)

รูปแบบพื้นฐานของคลังคำศัพท์คือเป็นที่เก็บรวบรวมคำที่มีความสำคัญและแต่ละคำที่เก็บไว้นั้นจะเก็บคำที่มีความสัมพันธ์กันไว้ด้วย ซึ่งมีจุดประสงค์เพื่อจัดเก็บคำศัพท์มาตรฐานเพื่อการทำดรรชนีและการค้นหา อีกทั้งยังช่วยให้ผู้ใช้กำหนดคำในข้อสอบถามให้เหมาะสมและจัดกลุ่มคำให้อยู่ในรูปแบบลำดับชั้นในทางด้านกว้าง (Broadening) และด้านลึก (Narrowing) การจัดการคำในลักษณะนี้เพื่อการทำดรรชนีมีวัตถุประสงค์เพื่อ

- สามารถทำให้กรอบความคิดของดรรชนีมีค่าเป็นมาตรฐาน
- ลดดรรชนีที่ไม่มีคุณภาพออกไปนั่นคือ ลดดรรชนีที่ไม่สามารถใช้ในการค้นคืนข้อมูลออกมาให้ผู้ใช้อย่างถูกต้องได้
- เพื่อกำหนดดรรชนีที่มีความหมายชัดเจนและค้นคืนเอกสารด้วยการใช้ความหมายมากกว่าการสะกดของคำศัพท์

ส่วนประกอบของคลังคำศัพท์นี้จะประกอบด้วย 2 ส่วนด้วยกันคือ

1) **คำดรรชนี** (index)

คำที่ปรากฏในคลังคำศัพท์นี้จะแสดงถึงแนวความหมายที่สามารถถ่ายทอดไปยังความคิดที่ต้องการสื่อความหมาย โดยอาจจะเป็นคำเดี่ยว กลุ่มของคำ วลี แต่ส่วนใหญ่จะเป็นคำเดี่ยวและคำนาม

2) **ความสัมพันธ์ระหว่างคำ**

เซตของคำที่มีความสัมพันธ์กับคำที่เก็บอยู่ในคำดรรชนีจะประกอบไปด้วยคำที่มีความหมายเหมือนกันและคำที่มีความหมายใกล้เคียงกัน ความสัมพันธ์ที่กล่าวมานี้จะได้มาโดยการเกิดขึ้นพร้อมกันของคำในเอกสารต่าง ๆ ซึ่งจะอยู่ในรูปแบบของลำดับชั้นและสามารถแสดงการขยายคำตามความหมายทางด้านกว้าง (Boardening Term: BT) และทางด้านลึก (Narrowing Term: NT) อย่างไรก็ตามความสัมพันธ์ของคำไม่ได้จำเพาะแค่ด้านกว้างและด้านลึกเท่านั้น ยังมีความสัมพันธ์ที่ไม่ได้อยู่ในรูปแบบของลำดับชั้น (Relation Term: RT) ซึ่ง

ความสัมพันธ์แบบนี้จะกำหนดยากมาก ต้องอาศัยผู้เชี่ยวชาญที่มีความสามารถเกี่ยวกับเรื่องนั้น ๆ โดยเฉพาะมากำหนด (Baeza-Yates and Ribeiro-Neto, 1999) เช่นคำว่า “rubber” มีความสัมพันธ์กับคำว่า “elasticity” ซึ่งคำว่า “elasticity” เป็นคำเฉพาะในสาขาพอลิเมอร์

จากคุณสมบัติของคลังคำศัพท์นั้นสามารถนำมาเก็บคำศัพท์ที่มีความหมายความสัมพันธ์กันได้ในรูปแบบต่าง ๆ ไม่ว่าจะเป็นทางด้านกว้าง (Boardening Term: BT) ทางด้านลึก (Narrowing Term: NT) และความสัมพันธ์ที่ไม่ได้อยู่ในรูปแบบของลำดับชั้น (Relation Term: RT) โดยในงานวิจัยนี้จะเก็บคำศัพท์ที่สัมพันธ์กันในเชิงที่ไม่ได้อยู่ในรูปแบบลำดับชั้น (Relation Term: RT) ที่หาได้จากการค้นหาความสัมพันธ์ของคำด้วยเทคนิคการค้นหากฎความสัมพันธ์ เนื่องจากการค้นหาความสัมพันธ์ของคำที่ได้จากเทคนิคนี้เป็นความสัมพันธ์ที่ไม่ได้พิจารณาถึงความหมายของคำ แต่พิจารณาความถี่ในการเกิดร่วมกันของคำ รายละเอียดของเทคนิคนี้จะกล่าวต่อไปในหัวข้อเทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Discovery)

2.4 การกำหนดรูปแบบการค้นคืนเอกสารและข้อสอบถาม

เทคโนโลยีที่นำมาใช้ในการค้นคืนเอกสารนั้นมีอยู่มากมาย ซึ่งวิธีที่ง่ายและใช้กันเป็นส่วนมากคือ การค้นคืนเอกสารโดยใช้คำสำคัญ (Keyword search) โดยวิธีนี้จะนำคำในข้อสอบถามของผู้ใช้มาเปรียบเทียบกับคำในเอกสารทีละคำ หากพบว่าเอกสารใดมีคำคำนั้นปรากฏ ระบบจะค้นคืนเอกสารออกมาแสดง แต่วิธีนี้ไม่ได้คำนึงถึงคุณภาพของผลลัพธ์ที่ได้ ดังนั้นเอกสารที่แสดงออกมาส่วนใหญ่จึงไม่ตรงกับความต้องการของผู้ใช้ ต่อมาวิธีการค้นคืนข้อมูลจึงถูกปรับปรุงให้ดียิ่งขึ้นด้วยวิธีการต่าง ๆ เช่น แบบจำลองความน่าจะเป็น (Probabilistic Model) แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) เป็นต้น (กฤษณี อริยชาญศิลป์, 2545)

แบบจำลองความน่าจะเป็น (Probabilistic Model) เป็นรูปแบบการค้นคืนที่มีจุดประสงค์ในการหาความน่าจะเป็นที่เอกสารตรงกับข้อสอบถาม โดยแบบจำลองนี้จะสร้างบนสมมติฐานที่ว่าค่าความน่าจะเป็นของความเกี่ยวเนื่องกับความต้องการของผู้ใช้จะสามารถคำนวณได้จากข้อสอบถามและเอกสารเท่านั้น ซึ่งจะค้นคืนเอกสารที่มีความน่าจะเป็นที่เอกสารตรงกับข้อสอบถามมาก ๆ ออกมา

ระบบค้นคืนเอกสารในงานวิจัยนี้จะค้นคืนเอกสารโดยใช้เทคโนโลยีแบบจำลองปริภูมิเวกเตอร์ (Vector Space Model) เนื่องจากเทคโนโลยีนี้จะช่วยการค้นคืนระบบสารสนเทศให้ดียิ่งขึ้นโดยจะแบ่งออกเป็น 3 ขั้นตอน ดังนี้

1) การกำหนดเวกเตอร์เอกสารและข้อสอบถาม

แบบจำลองเวกเตอร์เป็นแบบจำลองที่ถูกคิดขึ้นโดย Gerard Salton (Silva et al., 2004) โดยเทคนิคการค้นคืนสารสนเทศนั้นได้แสดงคำ (Term) เอกสาร (Document) และข้อสอบถาม (Query) ให้อยู่ในรูปแบบเวกเตอร์ (Vector) ดังนี้

ถ้าระบบมีจำนวนคำ t คำ ดังนั้นเวกเตอร์ของเอกสารและข้อสอบถามมี t มิติ สามารถนิยามดังนี้

กำหนดให้ d_j คือ เอกสารที่ j

q คือ ข้อสอบถาม

$k_{i,j}$ คือ คำ i ในเอกสารที่ j

$k_{i,q}$ คือ คำ i ในข้อสอบถาม q

$$d_j = \sum_{i=1}^{t} k_{i,j} \quad \text{หรือ} \quad d_j = (k_{1,j}, k_{2,j}, \ldots, k_{t,j}) \qquad (2.1)$$

เช่นเดียวกับเวกเตอร์ข้อสอบถาม

$$q = \sum_{i=1}^{t} k_{i,q} \quad \text{หรือ} \quad q = (k_{1,q}, k_{2,q}, \ldots, k_{t,q}) \qquad (2.2)$$

จากคำนิยาม 2.1 และ 2.2 แต่ละตำแหน่งมิติของเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามจะมีค่าก็ต่อเมื่อ

- ถ้าในเวกเตอร์ข้อสอบถามหรือเอกสารมีคำใดปรากฏอยู่ ที่ตำแหน่งมิติของคำนั้นในเวกเตอร์จะมีค่าเท่ากับค่าน้ำหนักของคำในเอกสารนั้น ๆ (จะกล่าววิธีการให้น้ำหนักคำในหัวข้อถัดไป)
- ถ้าในเวกเตอร์ข้อสอบถามหรือเอกสารไม่มีคำใดปรากฏอยู่ ที่ตำแหน่งมิติของคำนั้นในเวกเตอร์จะมีค่าเท่ากับ 0

เช่น กำหนดให้ระบบมีคำอยู่ 5 คำคือ a b c d และ e รูปแบบเซตจะแสดงได้ดังนี้ {a, b, c, d, e} ถ้าในเอกสาร d_1 ปรากฏคำ “a” และ ”c” ดังนั้นเวกเตอร์ของเอกสาร d_1 = {1, 0, 1, 0, 0} ในที่นี้สมมุติให้น้ำหนักคำมีค่าเท่ากับ 1 ทั้งหมด

2) การให้ค่าน้ำหนักคำ (Baeza-Yates and Ribeiro-Neto, 1999)

วิธีการประมาณค่าความสำคัญของคำ โดยที่การให้น้ำหนักคำที่เหมาะสมที่นิยมใช้มี 2 วิธีด้วยกัน ดังนี้

- **ความถี่ของคำ** (Term Frequency : tf)

ความถี่ในการปรากฏของคำสามารถเป็นสิ่งที่บ่งบอกถึงความสำคัญของคำคำนั้นที่มีต่อเอกสารหนึ่ง ๆ เอกสารที่มีคำที่ผู้ใช้ต้องการจะเป็นเอกสารที่มีประโยชน์ต่อผู้ใช้สูงกว่า เช่น ถ้าผู้ใช้ต้องการว่า “computer” เอกสารที่ใช้คำว่า “computer” 10 ครั้ง จะเป็นประโยชน์กว่าเอกสารที่มีคำว่า “computer” เพียงครั้งเดียว โดยที่ค่า tf หาได้จากสมการที่ (2.3)

กำหนดให้ $freq_{i,j}$ คือ ความถี่ที่คำ k_i ปรากฏ ในเอกสาร d_j

$\max freq_{i,j}$ คือ ความถี่ของคำใด ๆ ในเอกสาร d_j ที่มากที่สุด

$$tf_{i,j} = \frac{freq_{i,j}}{\max freq_{l,j}} \qquad (2.3)$$

- **ความถี่ของเอกสารแบบผกผัน** (Inverse Document Frequent : idf)

เนื่องจากคำทั่วไปที่ปรากฏในเอกสารบ่อยครั้งแต่ไม่สามารถทำให้เอกสารแตกต่างกับเอกสารอื่นได้ โดยเป็นคำที่ใช้บ่อยครั้งในแต่ละเอกสารจะมีความสำคัญน้อยกว่าคำที่ใช้น้อยหรือคำที่ใช้เฉพาะบางเอกสาร ตัวอย่างเช่น การพบคำว่า “the” ในเอกสาร แม้ว่าจะพบคำว่า “the” บ่อยในเอกสารก็ตามแต่คำนี้ไม่มีความสำคัญสำหรับเอกสารนั้น ๆ โดยค่า idf หาได้จากสมการที่ (2.4)

กำหนดให้ N คือ จำนวนเอกสารทั้งหมดในระบบ

n_i คือ จำนวนเอกสารที่มีคำ k_i ปรากฏ

$$idf_i = \log \frac{N}{n_i} \qquad (2.4)$$

ค่าความถี่ของคำ tf นั้นจะพิจารณาแค่ความถี่ของคำ แต่สำหรับคำที่ปรากฏอยู่ในเอกสารบ่อยครั้ง แต่ไม่สามารถที่จะแยกเอกสารที่ตรงกับความต้องการออกจากเอกสารที่ไม่ตรงกับความต้องการได้นั้น จะได้ค่าน้ำหนักออกมาอย่างไม่เหมาะสม ดังนั้นจึงพิจารณาค่าความถี่แบบผกผัน idf ร่วมด้วยสำหรับแก้ปัญหาคำประเภทเหล่านี้ เพื่อให้การให้น้ำหนักของคำมีประสิทธิภาพมากขึ้น

จากค่าความถี่ของคำ (tf) และค่าความถี่แบบผกผัน (idf) สามารถช่วยให้การให้ค่าน้ำหนักของคำในเอกสารมีความเหมาะสมยิ่งขึ้นจากที่กล่าวมาข้างต้น ดังนั้นในการให้ค่าน้ำหนัก

ของคำในเอกสารสามารถคำนวณ โดยใช้ค่าความถี่ของคำ (tf) และค่าความถี่แบบผกผัน (idf) มาคำนวณร่วมกันดังสมการที่ 2.5

กำหนดให้ $w_{i,j}$ คือ ค่าน้ำหนักของคำ i ในเอกสารที่ j

$tf_{i,j}$ คือ ความถี่ของคำ i ในเอกสาร j

idf_i คือ ค่าความถี่ผกผันของคำ i ของเอกสารทั้งหมด

$$\text{Document term weight } (w_{i,j}) = tf_{i,j} \times idf_i \qquad (2.5)$$

การคำนวณค่าน้ำหนักของคำในข้อสอบถามข้างต้น เป็นการพิจารณาโดยใช้ค่า tf-idf เช่นกัน จากค่าความถี่ (tf) ของคำจะมีค่าตั้งแต่ 0 ถึง 1 ดังนั้นจึงกำหนดให้ข้อสอบถามมีค่าความถี่ของคำพื้นฐานเป็น 0.5 (ค่ากลาง) และบวกค่าความถี่ของคำโดยพิจารณา 0 ถึง 0.5 เท่านั้น เนื่องจากข้อสอบถามที่ผู้ใช้กรอกเข้ามานั้นถือเป็นคำที่มีนัยสำคัญมาก ดังนั้นค่าความถี่ของคำควรจะมีน้ำหนักไม่ต่ำกว่าครึ่งหนึ่ง (Yates and Neto,1999) จากนั้นนำค่าความถี่ที่ได้คูณด้วยค่าความถี่ของเอกสารแบบผกผัน (idf) ดังนั้นสมการในการคำนวณค่าน้ำหนักของคำในข้อสอบถามจะแสดงดังสมการที่ 2.6

กำหนดให้ $w_{i,q}$ คือ ค่าน้ำหนักของคำลำดับที่ i ในข้อสอบถาม

N คือ จำนวนเอกสารทั้งหมดในระบบ

n_i คือ จำนวนเอกสารที่มีคำ k_i ปรากฏ

$freq_{i,q}$ คือ ความถี่ที่คำ k_i ปรากฏ ในข้อสอบถาม q

$$\text{Query term weight } (w_{i,q}) = \left(0.5 + \frac{0.5\, freq_{i,q}}{\max_l freq_{l,q}}\right) \times \log\frac{N}{n_i} \qquad (2.6)$$

2.5 **เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Discovery)** (Ye, 2001)

เทคนิคของการทำเหมืองข้อมูล (Data mining) ที่ใช้หาความสัมพันธ์ของรายการในธุรกรรมที่เกิดขึ้น คือ การค้นหากฎความสัมพันธ์ (Association Rule Discovery) ซึ่งเป็นเทคนิคที่หาความสัมพันธ์หรือความใกล้ชิดระหว่างเซตรายการ (Item sets) ที่ปรากฏในธุรกรรมแต่ละธุรกรรม (Transaction) ที่เกิดขึ้นคือเซตรายการหนึ่งเซต (Item set) ตัวอย่างเซตรายการ (Item sets) ในรูปที่ 2.3 (Ye, 2001) กำหนดให้ Jane Austen แทนด้วย “A” Agatha Christie แทนด้วย “C” Sir Arthur Conan Doyle แทนด้วย “D” Mark Twain แทนด้วย “T” และ P. G. Wodehouse

แทนด้วย “W” ดังตาราง Distinct Database Items และในตาราง Database จะประกอบไปด้วย 6 ธุรกรรมคือ ในธุรกรรมที่ 1 จะปรากฏ ACTW พร้อมกัน ในธุรกรรมที่ 2 จะปรากฏ CDW พร้อมกัน ธุรกรรมที่ 3 จะปรากฏ ACTW พร้อมกัน ธุรกรรมที่ 4 จะปรากฏ ACDW พร้อมกัน ธุรกรรมที่ 5 จะปรากฏ ACDTW พร้อมกัน และธุรกรรมที่ 6 จะปรากฏ CDT พร้อมกัน ตารางสุดท้ายคือตาราง All Frequent Itemsets แสดงอัตราการเกิดเซตรายการในธุรกรรม (Transaction) ที่เกิดขึ้น จากตารางนี้แสดงว่าการเกิดรายการ “C” มีอัตราการเกิดเท่ากับ 100% ในธุรกรรมทั้งหมดนั่นคือเกิดในทุกธุรกรรม รายการ “W” และรายการ ”CW” มีอัตราการเกิดเท่ากับ 83% ในธุรกรรมทั้งหมด รายการ ”A” รายการ ”D” รายการ ”T” รายการ ”AC” รายการ ”AW” รายการ ”CD” รายการ ”CT” และรายการ ”ACW” มีอัตราการเกิดเท่ากับ 67% ในธุรกรรมทั้งหมด รายการ ”AT” รายการ ”DW” รายการ ”TW” รายการ ”ACT” รายการ ”ATW” รายการ ”CDW” รายการ ”CTW” และรายการ ”ACTW” มีอัตราการเกิดเท่ากับ 50% ในธุรกรรมทั้งหมด

Distinct Database Items

Jane Austen	Agatha Christie	Sir Arthur Conan Doyle	Mark Twain	P.G. Wodehouse
A	C	D	T	W

Database

Transaction	Items
1	ACTW
2	CDW
3	ACTW
4	ACDW
5	ACDTW
6	CDT

All Frequent Itemsets

Minimum Support = 50%

Support	Itemsets
100%	C
83%	W,CW
67%	A,D,T, AC,AW,CD,CT,ACW
50%	AT,DW,TW,ACT,ATW, CDW,CTW,ACTW

รูปที่ 2.3 รูปแสดงการสร้างเซตความถี่รายการ

กฎความสัมพันธ์นั้นประกอบด้วย 2 เซตรายการที่เรียกว่า สิ่งที่เกิดก่อน (Antecedent) และสิ่งที่ตามมา (Consequent) บ่อยครั้งสิ่งที่ตามมานั้นมักจะถูกจำกัดให้มีเพียงหนึ่งรายการ (Single item) บ่อย ๆ แต่นั่นไม่เสมอไป กฎจะถูกแสดงโดยใช้ “ $\rightarrow$ ” (ลูกศร) ที่นำจากสิ่งที่มาก่อนไปสู่สิ่งที่ตามมาเช่น {A,B,C} $\rightarrow$ {B} กฎความสัมพันธ์จะแสดงถึงความใกล้ชิดระหว่างเซต

รายการของสิ่งที่มาก่อนและเซตรายการของสิ่งที่ตามมาหรือแสดงว่าเมื่อเกิดสิ่งที่เกิดก่อนมักเกิดสิ่งที่ตามมานั้นด้วย นอกจากนั้นยังสามารถระบุระดับความใกล้ชิดของเซตรายการได้จากความแข็งแรงของกฎความสัมพันธ์ โดยใช้สถิติพื้นฐานจากความถี่ (Frequency-based statistics) มี 3 ค่าดังต่อไปนี้

1) **ค่าสนับสนุน** (Support) คือ อัตราส่วนของธุรกรรมที่มีทั้งส่วนที่มาก่อนและส่วนที่ตามมาการคำนวณของค่าสนับสนุน (Support) ของกฎความสัมพันธ์สามารถคำนวณได้ดังต่อไปนี้ เช่น สำหรับกฎความสัมพันธ์ $A \rightarrow C$

กำหนดให้ D คือ ฐานข้อมูลของธุรกรรม

N คือ จำนวนธุรกรรมในฐานข้อมูลของธุรกรรม

I คือ เซตรายการอื่น ๆ

X คือ เซตรายการที่ต้องการคำนวณ

$$\text{Support (X)} = \frac{\left|\{I \mid I \in D \wedge I \supseteq X\}\right|}{N}$$

หรือ $$\text{Support (X)} = \frac{\text{จำนวนธุรกรรมในฐานข้อมูล D ที่มี X ปรากฏ}}{\text{จำนวนธุรกรรมทั้งหมดในฐานข้อมูล D}}$$

$$\text{Support } (A \rightarrow C) = \text{Support } (A \cup C)$$

2) **ค่าความเชื่อมั่น** (Confidence) คือ อัตราส่วนของธุรกรรมที่มีส่วนที่มาก่อน แล้วจะมีส่วนที่ตามมานั้นด้วย การคำนวณค่าความเชื่อมั่น (Confidence) ของกฎความสัมพันธ์ จะสามารถคำนวณได้ดังต่อไปนี้ เช่น สำหรับกฎความสัมพันธ์ $A \rightarrow C$

$$\text{Confidence } (A \rightarrow C) = \text{Support } (A \cup C) / \text{ Support } (A)$$

3) **ค่าลิฟท์** (Lift) เป็นสัดส่วนของความถี่ของสิ่งที่มาก่อน (Antecedent) แล้วเกิดสิ่งที่ตามมา (Consequent) ในธุรกรรมเทียบกับการเกิดสิ่งที่ตามมา (Consequent) ในธุรกรรมทั้งหมด โดยเป็นการบอกค่าสัดส่วนของค่าความถี่ที่ปรากฏและความถี่ที่คาดหวัง ซึ่งความถี่คาดหวังนั้นเป็นความถี่ในการเกิดขึ้นของสิ่งที่มาก่อนและสิ่งที่ตามมาในกรณีที่ทั้งสองสิ่งเป็นอิสระต่อกันหรือไม่มีความสัมพันธ์ต่อกัน ดังนั้นถ้าค่าลิฟท์ (Lift) มีค่ามากกว่า 1 แสดงว่าสิ่งที่มาก่อนมี

ความสัมพันธ์กับสิ่งที่ตามมาและถ้ามีค่าเท่ากับ 1 แสดงว่าสิ่งที่มาก่อนไม่มีความสัมพันธ์กับสิ่งที่ตามมา

$$\text{Lift}(A \rightarrow C) = \text{confidence}(A \rightarrow C) \,/\, \text{support}(C)$$

ตัวอย่าง พิจารณาความสัมพันธ์ {a}$\rightarrow${b} ถ้าค่าสนับสนุนของ {b} = 0.4 และค่าความเชื่อมั่นของ {a}$\rightarrow${b} = 0.67 ดังนั้นค่าลิฟท์ (Lift) ของ {a}$\rightarrow${b} = 0.67 / 0.4 = 1.675 จะแสดงว่าการเกิดรายการ a แล้วเกิดรายการ b จะพบมากกว่าการเกิดรายการ b อย่างเดียวเป็น 1.675 เท่า ในทางตรงข้าม พิจารณาอีกตัวอย่างที่มีค่าความเชื่อมั่นเท่ากัน ความสัมพันธ์ {a}$\rightarrow${c} ถ้าค่าสนับสนุนของ {c} = 0.6 และค่าความเชื่อมั่นของ {a}$\rightarrow${c} = 0.67 ดังนั้นค่าลิฟท์ (Lift) ของ {a}$\rightarrow${c} = 0.67 / 0.6 = 1.117 จะแสดงว่าการเกิดรายการ a แล้วเกิดรายการ c จะพบมากกว่าการเกิดรายการ c อย่างเดียวเป็น 1.117 เท่า จาก 2 กฎความสัมพันธ์ข้างต้นจะแสดงว่า a มีความสัมพันธ์กับการเกิดของ b มากกว่า c นั่นเอง

การค้นหากฎความสัมพันธ์ (Association Rule Discovery) จะค้นหาความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นมากกว่าค่าที่ผู้ค้นหากฎกำหนด ในการสร้างกฎความสัมพันธ์เป็นการรวมตัว (Combination) ของสิ่งที่มาก่อน (Antecedent) และสิ่งที่ตามมา (Consequent) ประเมินค่าสนับสนุนและค่าความเชื่อมั่น และตัดความสัมพันธ์ที่ไม่น่าสนใจ คือมีค่าไม่ถึงค่าสนับสนุนและค่าความเชื่อมั่นที่ผู้ค้นหากฎตั้งไว้ โดยการตั้งค่านั้นขึ้นอยู่กับความเหมาะสมของแต่ละงานที่จะใช้สร้างกฎความสัมพันธ์ โปรแกรมที่สามารถค้นหากฎความสัมพันธ์ (Association Rule Discovery) ได้นั้นมีหลากหลายโปรแกรม เช่น โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ (SAS Enterprise Miner) โปรแกรมไมน์เซต (MineSet) โปรแกรมอีซีไมน์เนอร์ (EasyMiner) เป็นต้น

2.6 การค้นคืนเอกสารที่ตรงกับข้อสอบถามของผู้ใช้

การเปรียบเทียบความเหมือนระหว่างเวกเตอร์เอกสารในระบบและเวกเตอร์ข้อสอบถามที่ผู้ใช้กรอกเข้ามา สามารถคำนวณระดับความเหมือนระหว่างเวกเตอร์ 2 เวกเตอร์ โดยมีวิธีการคำนวณ 4 วิธีด้วยกัน (Chowdhury, 2004) แสดงได้ดังนี้

กำหนดให้มีเวกเตอร์เอกสาร d_j คือเอกสารลำดับที่ j ใด ๆ และเวกเตอร์ข้อสอบถาม q ดังตัวอย่างที่กล่าวมา

- **การหาค่าความเหมือนวิธีไดซ์ (Dice coefficient)** สมการของ Dice coefficient แสดงได้ดังสมการที่ 2.7

$$sim(d_i, q) = 2\frac{\left[\sum_{i=1}^{t} w_{i,j} \times w_{i,q}\right]}{\sum_{i=1}^{t} w_{i,j} + \sum_{i=1}^{t} w_{i.q}} \quad (2.7)$$

- **การหาค่าความเหมือนวิธีเจคคาร์ด (Jaccard coefficient)** สมการของ Jaccard coefficient แสดงได้ดังสมการที่ 2.8

$$sim(d_i, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j} + \sum_{i=1}^{t} w_{i.q} - \sum_{i=1}^{t} \left(w_{i,j} w_{i.q}\right)} \quad (2.8)$$

- **การหาค่าความเหมือนวิธีโคซายน์ (Cosine coefficient)** สมการของ Cosine coefficent แสดงได้ดังสมการที่ 2.9

$$sim(d_j,q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^{\ 2}} \times \sqrt{\sum_{i=1}^{t} w_{i.q}^{\ 2}}} \quad (2.9)$$

- **การหาค่าความเหมือนวิธีโอเวอร์แล็บ (Overlab coefficient)** สมการของ Overlab coefficient แสดงได้ดังสมการที่ 2.10

$$sim(d_i, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\min\left(\sum_{i=1}^{t} w_{i,j} + \sum_{i=1}^{t} w_{i.q}\right)} \quad (2.10)$$

ซึ่งวิธีของการหาค่าความเหมือนวิธีเจคการ์ด (Jaccard coefficient) และการหาค่าความเหมือนวิธีโคไซน์ (Cosine coefficient) มีลักษณะการวัดที่เหมือนกันโดยช่วงค่าความเหมือนจะมีค่าตั้งแต่ 0 ถึง 1 และ 2 วิธีนี้จะคำนวณค่าความเหมือนสำหรับเวกเตอร์ที่แต่ละมิติไม่มีค่าติดลบ (Chowdhury, 2004) ทั้ง 4 วิธีเป็นวิธีการคำนวณค่าความเหมือนที่ผลการค้นคืนไม่แตกต่างกัน ในงานวิจัยนี้จึงเลือกใช้การหาค่าความเหมือนวิธีโคซายน์ (Cosine coefficient) ในการคำนวณค่าความเหมือนในการค้นคืนเอกสาร เนื่องจากเป็นวิธีที่เรียบง่ายและใช้กับระบบค้นคืนเอกสารส่วนใหญ่ (Chowdhury, 2004; Qin et al., 2004)

2.7 การตั้งค่าความเหมือนต่ำที่สุดในการค้นคืนเอกสาร

ในขั้นตอนของการพัฒนาระบบค้นคืนเอกสารจะต้องตั้งค่าความเหมือนต่ำที่สุดในการค้นคืนเอกสาร เมื่อเอกสารและข้อสอบถามใดมีค่าความเหมือนมากกว่าค่าความเหมือนต่ำสุดที่ตั้งไว้ ระบบจะค้นคืนเอกสารนั้นออกมาแสดง ซึ่งการตั้งค่านี้จะเป็นการตั้งค่าตามความเหมาะสมของระบบค้นคืนเอกสารแต่ละระบบ ไม่มีรูปแบบการคำนวณที่กำหนดไว้ (Baeza-Yates and Ribeiro-Neto, 1999)

จากงานวิจัยของ Udomchaiporn Akadej ได้ตั้งค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) ลบกับค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) (Udomchaiporn, 2005) แต่ในงานวิจัย Udomchaiporn ได้เสนอไว้ว่าการตั้งค่าความเหมือนสามารถตั้งได้ตามความเหมาะสมกับระบบค้นคืนเอกสารนั้น ๆ โดยจะสามารถตั้งไว้ที่ค่าเฉลี่ย (Mean) หรือค่าเฉลี่ยบวกกับค่าเบี่ยงเบนมาตรฐาน (Mean + Standard Deviation) หรือมากกว่านี้ได้ตามความเหมาะสม

2.8 การปรับปรุงข้อสอบถามจากผลสะท้อนกลับจากผู้ใช้

การปรับปรุงข้อสอบถามสำหรับแบบจำลองปริภูมิเวกเตอร์เป็นการเปลี่ยนแปลงคำในข้อสอบถามและเปลี่ยนแปลงน้ำหนักของคำในแต่ละมิติของเวกเตอร์ข้อสอบถาม โดยเมื่อผู้ใช้ระบุเอกสารที่เกี่ยวเนื่องกับคำในข้อสอบถามและเอกสารที่ไม่เกี่ยวเนื่องกับคำในข้อสอบถาม เอกสารที่เกี่ยวเนื่องกับความต้องการจะมีคำที่เกี่ยวข้องปรากฏอยู่ ข้อสอบถามใหม่ที่เปลี่ยนแปลงได้นั้นจึงต้องมีความคล้ายหรือใกล้ชิดกับเอกสารที่ถูกระบุว่าเกี่ยวเนื่องกับความต้องการนั่นเอง ซึ่งสูตรในการคำนวณเวกเตอร์ข้อสอบถามใหม่แสดงได้ดังสมการที่ 2.11 (Baeza-Yates and Ribeiro-Neto, 1999)

กำหนดให้

$\vec{q}_m$ คือ เวกเตอร์ข้อสอบถามที่กำหนดขึ้นใหม่

$\vec{q}$ คือ เวกเตอร์ข้อสอบถามเริ่มต้น

$|D_r|$ คือ จำนวนเอกสารที่เกี่ยวเนื่องกับความต้องการ

$|D_n|$ คือ จำนวนเอกสารที่ไม่เกี่ยวเนื่องกับความต้องการ

$\vec{d}_j$ คือ เวกเตอร์ของเอกสารที่ j

D_r' คือ เซตของเอกสารที่เกี่ยวเนื่อง (Relevant Documents) ในจำนวนเอกสารที่ค้นคืนได้ทั้งหมด

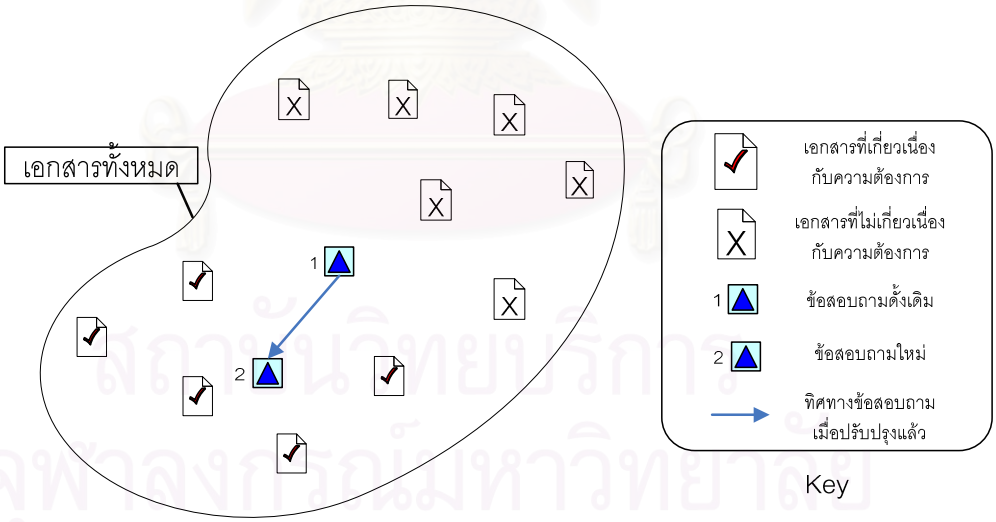
D_n' คือ เซตของเอกสารที่ไม่เกี่ยวเนื่อง (Non-Relevant Documents) ในจำนวนเอกสารที่ค้นคืนได้ทั้งหมด

α, β, γ คือ ค่าคงที่สำหรับการปรับค่า

$$\vec{q}_m = \alpha\vec{q} + \frac{\beta}{|D_r|}\left(\sum_{\forall \vec{d}_j \in D_r'} \vec{d}_j\right) - \frac{\gamma}{|D_n|}\left(\sum_{\forall \vec{d}_j \in D_n'} \vec{d}_j\right) \qquad (2.11)$$

การกำหนดสมการรอคชิโอ (Rochio) (Baeza-Yates and Ribeiro-Neto, 1999) ดังสมการที่ 2.11 นั้น เวกเตอร์ข้อสอบถามเดิมนั้นจะประกอบไปด้วยข้อมูลที่สำคัญ โดยที่ค่าคงที่ α, β, γ จะสามารถคำนวณได้จากการปรับเปลี่ยนค่าไปแต่ละครั้งที่ทดลองค้นคืนเอกสาร จนกว่าการค้นคืนจะออกมาตรงตามความต้องการของผู้ใช้มากที่สุด ซึ่งโดยปกติข้อมูลที่อยู่ในเอกสารที่เกี่ยวเนื่องกับความต้องการจะมีความสำคัญมากกว่าข้อมูลที่อยู่ในเอกสารที่ไม่เกี่ยวเนื่องกับความต้องการ ดังนั้นโดยปกติแล้วค่าคงที่ γ จะมีค่าน้อยกว่า ค่าคงที่ β (Baeza-Yates and Ribeiro-Neto, 1999)

โดยข้อดีของการให้ผลสะท้อนกลับด้วยวิธีนี้คือมีความเรียบง่าย กล่าวคือ การปรับเปลี่ยนค่าน้ำหนักถูกคำนวณมาโดยตรงจากเอกสารที่ถูกดึงขึ้นมาและให้ผลลัพธ์ที่ดีนั่นคือ ปรับเปลี่ยนเวกเตอร์ข้อสอบถามเดิม โดยมีผลกระทบกับค่าน้ำหนักคำของเวกเตอร์ข้อสอบถามเดิม ณ ตำแหน่งมิติใด ๆ ทำให้ข้อสอบถามใหม่ที่ได้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องมากขึ้นดังรูปที่ 2.4



รูปที่ 2.4 รูปแสดงผลที่เกิดจากการดำเนินการปรับปรุงข้อสอบถามจากผลสะท้อนกลับ

งานวิจัยของ Iwayama ที่วิจัยเกี่ยวกับความถูกต้องของการให้ผลสะท้อนกลับและการจัดกลุ่มเอกสาร (Iwayama, 2000) และงานวิจัยเกี่ยวกับผลกระทบเมื่อให้ผลสะท้อนกลับ (Buckley et al., 1994) ได้กำหนดให้ค่าน้ำหนักของข้อสอบถามเดิมมีค่าเท่ากับ 8 ค่าน้ำหนักของกลุ่ม

เอกสารที่เกี่ยวเนื่องมีค่าเท่ากับ 16 และค่าน้ำหนักของกลุ่มเอกสารที่ไม่เกี่ยวเนื่องมีค่าเท่ากับ 4 นั่นคือจากสูตรของ Rochio ดังสมการที่ 2.11 การตั้งค่า α, β, γ จะมีค่าเท่ากับ 8, 16, 4 ตามลำดับ ซึ่งทำให้ประสิทธิภาพการค้นคืนเอกสารนั้นดีขึ้น โดยคำนวณจากค่าความเรียกคืนและค่าความถูกต้อง ดังนั้นผู้วิจัยจึงได้นำค่าน้ำหนักที่กำหนดโดยงานวิจัยทั้งสองมาใช้กำหนดค่า α, β และ γ ในงานวิจัยของผู้วิจัยนี้

2.9 การวัดประสิทธิภาพระบบค้นคืนเอกสาร

การวัดประสิทธิภาพของระบบค้นคืนเอกสารสามารถคำนวณได้โดยใช้ค่าเรียกคืน (Recall) ค่าความถูกต้อง (Precision) ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ซึ่งค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) เป็นค่าที่คำนวณจากค่าเรียกคืนและค่าความถูกต้องมาเฉลี่ยกัน

มีงานวิจัยมากมายที่ใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) วัดประสิทธิภาพของระบบค้นคืนเอกสาร เช่น งานวิจัยการขยายคำในข้อสอบถามโดยรวมกฎความสัมพันธ์กับสิ่งที่ศึกษาร่วมกับเทคนิคการค้นคืนสารสนเทศ (Song et al., 2005) งานวิจัยการจัดกลุ่มเอกสารทางเว็บโดยใช้เซตรายการที่มากที่สุด (Zhuang and Dai, 2004) และงานวิจัยการปรับปรุงเอกสารโดยใช้ฐานความรู้เวิร์ดเนต (ของพิลาวัลย์ พลับรู้การ และกฤษณะ ไวยมัย, 2547) เป็นต้น ซึ่งสามารถคำนวณได้ดังนี้

ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision)

วิธีการวัดประสิทธิภาพระบบค้นคืนเอกสารในงานวิจัยนี้จะวัดด้วยค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) โดยที่ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง ซึ่งวิธีการวัดดังกล่าวเป็นวิธีการวัดผลการทดลองในห้องปฏิบัติการส่วนใหญ่ (Baeza-Yates and Ribeiro-Neto, 1999) คือการวัดที่รวมทั้งค่าเรียกคืนและค่าความถูกต้อง ซึ่งค่านี้จะมีค่าอยู่ในช่วง 0 ถึง 1 ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องจะมีค่าเป็น 0 เมื่อเอกสารที่ค้นคืนมาได้นั้น ไม่มีเอกสารใดเกี่ยวเนื่องกับข้อสอบถามเลยและจะมีค่าเป็น 1 เมื่อทุกเอกสารในที่ค้นคืนมาได้เป็นเอกสารที่เกี่ยวเนื่องกับข้อสอบถามทั้งหมด สมการในการคำนวณค่านี้แสดงได้ดังสมการที่ (2.12)

กำหนดให้ F(j) คือ ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของ r(j) และ P(j)

r(j) คือ ค่าเรียกคืนของเอกสารที่ j ในลำดับ (Ranking)

P(j) คือ ค่าความถูกต้องของเอกสารที่ j ในลำดับ (Ranking)

j คือ หมายเลขเอกสารซึ่งอยู่ในลำดับที่ j

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} \qquad (2.12)$$

จากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) มีการคำนวณหาค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) โดยค่าทั้ง 2 นี้มีรายละเอียดดังต่อไปนี้

ค่าเรียกคืน (Recall) เป็นสัดส่วนของเอกสารที่เกี่ยวเนื่องกับความต้องการ (relevant document) ที่ถูกดึงออกมาแสดงเทียบกับเอกสารที่เกี่ยวเนื่องทั้งหมดในระบบดังสมการที่ 2.13

กำหนดให้ |Ra| คือ จำนวนเอกสารเกี่ยวเนื่องตามความต้องการที่ค้นคืนออกมาได้

|R| คือ จำนวนเอกสารเกี่ยวเนื่องกับความต้องการที่อยู่ในฐานข้อมูลทั้งหมด

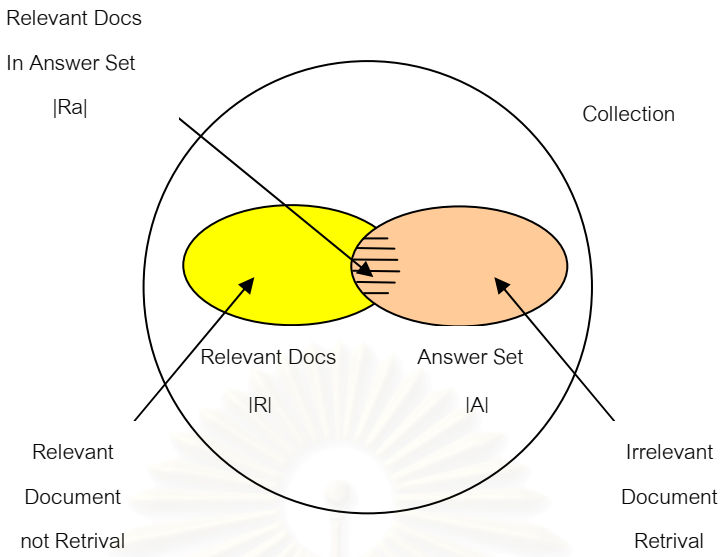
$$\text{Recall} = \frac{|Ra|}{|R|} \qquad (2.13)$$

ค่าความถูกต้อง (Precision) เป็นสัดส่วนของเอกสารที่ถูกดึงขึ้นมาแล้วเกี่ยวเนื่องกับความต้องการ (Relevant) เทียบกับเอกสารที่ถูกดึงออกมาแสดงทั้งหมดดังสมการที่ 2.14

กำหนดให้ |Ra| คือ จำนวนเอกสารเกี่ยวเนื่องตามความต้องการที่ค้นคืนออกมาได้

|A| คือ จำนวนเอกสารทั้งหมดที่ค้นคืนออกมา

$$\text{Precision} = \frac{|Ra|}{|A|} \qquad (2.14)$$



รูปที่ 2.5 รูปแสดงเซตของเอกสารที่เกี่ยวข้อง และเซตของคำตอบในชุดข้อมูลหนึ่ง ๆ (Baeza-Yates and Ribeiro-Neto, 1999)

2.10 งานวิจัยที่เกี่ยวข้อง

งานวิจัยชิ้นนี้เป็นการพัฒนาระบบการค้นคืนสารสนเทศที่เป็นเอกสาร ในอดีตมีงานวิจัยที่เกี่ยวข้องกับระบบการค้นคืนเอกสารคือ งานวิจัยที่สร้างระบบการค้นคืนโดยพิจารณาการขึ้นต่อกันระหว่างคำในแบบจำลองปริภูมิเวกเตอร์ (Silva et al., 2004) โดย Silva และคณะจะหากฎความสัมพันธ์โดยใช้ขั้นตอนวิธี Aprior Algorithm แล้วใช้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) เพื่อคัดแยกกฎความสัมพันธ์ที่มีคุณภาพออกมา เมื่อได้กฎความสัมพันธ์ที่มีคุณภาพออกมาแล้ว จะนำกฎความสัมพันธ์ที่ได้นี้มาขยายคำในเอกสารและข้อสอบถามก่อนจะนำมาเทียบความเหมือนระหว่างเอกสารและข้อสอบถาม เพื่อที่จะค้นคืนเอกสารที่มีความใกล้เคียงกับข้อสอบถามออกมาแสดงต่อผู้ใช้ ซึ่งในงานวิจัยของ Silva และคณะในปี 2004 ใช้เทคโนโลยีแบบจำลองปริภูมิเวกเตอร์แทนรูปแบบของเอกสารและข้อสอบถาม โดยจะเปลี่ยนรูปแบบของข้อสอบถามและเอกสารให้อยู่ในรูปแบบของเวกเตอร์ที่มีแต่ละมิติของเวกเตอร์จะเป็นน้ำหนักของคำที่อยู่ในตำแหน่งนั้น ซึ่งแต่ละคำที่มีอยู่ในระบบจะแสดงอยู่ในรูปแบบของเวกเตอร์เช่นกัน

แต่ละคำ k_i ถูกแสดงด้วยเวกเตอร์ t มิติ ซึ่ง t คือจำนวนคำทั้งหมดในที่เก็บคำศัพท์ แบบจำลองปริภูมิเวกเตอร์กำหนดให้เวกเตอร์ k_i แสดงเซตของคำ k_i เซต โดยรูปแบบเซตของเวกเตอร์คำทั้งหมดจะเขียนได้ดังนี้ $\{k_1, k_2,\ldots, k_t\}$ โดยที่ $k_1 = \{1,0,0,\ldots,0\}$, $k_2 = \{0,1,0,\ldots,0\}$, …, $k_t = \{0,0,0,\ldots,1\}$ นั่นคือ ที่ตำแหน่งมิติของเวกเตอร์คำนั้นจะแทนด้วยคำแต่ละคำในระบบ เช่น

กำหนดให้ระบบมีคำอยู่ 3 คำคือ a b และ c รูปแบบเซตจะแสดงได้ดังนี้ {a, b, c} โดยที่ a = (1,0,0), b = (0,1,0) และ c = (0,0,1)

แต่ละเวกเตอร์ของเอกสารหรือข้อสอบถามนั้นอาจมีคำที่สัมพันธ์กันในกฎความสัมพันธ์ ตัวอย่างเช่น ถ้ามีกฎความสัมพันธ์ $k_i \rightarrow k_j$ นั่นคือถ้าปรากฏคำ "k_i" แล้วจะปรากฏคำ "k_j" ด้วยค่าความเชื่อมั่น c_{ij} ที่ปรับค่าให้อยู่ระหว่าง 0 ถึง 1 มุมระหว่างเวกเตอร์ระหว่าง 2 คำนี้ ซึ่งแต่เดิมคือ 90 องศาจะถูกปรับให้มีมุม $\theta = 90\left(1 - c_{ij}\right)$ จากเวกเตอร์คำ k_i จะเปลี่ยนเป็นเวกเตอร์ใหม่ k_i'โดยที่แต่ละมิติที่ r ของเวกเตอร์คำใหม่ k_i'นี้จะมีค่าเป็น a_r ซึ่งสามารถคำนวณได้ดังนี้

$$a_r = \sin(\theta_{ij}) \leftrightarrow r = i$$
$$a_r = \cos(\theta_{ij}) \leftrightarrow r = j$$
$$a_r = 0 \leftrightarrow r \neq i \text{ และ } r \neq j$$

จากนั้นนำเวกเตอร์ k_i'ไปปรับค่าเวกเตอร์เอกสารและข้อสอบถาม และเมื่อนำเวกเตอร์ข้อสอบถามกับเวกเตอร์เอกสารมาเปรียบเทียบความเหมือนจะได้ค่าความเหมือนค่าใหม่ ทำให้ผลลัพธ์รายการเอกสารที่เกี่ยวเนื่องกับข้อสอบถามนั้น ๆ เปลี่ยนไป ซึ่งผลการทดลองของระบบการค้นคืนนี้สามารถให้ค่าความถูกต้องที่ดีขึ้นกว่าระบบการค้นคืนข้อมูลที่ไม่ได้ปรับเปลี่ยนค่าน้ำหนักของคำ จากงานวิจัยที่กล่าวมาผู้วิจัยมีข้อคิดเห็นว่าการปรับเปลี่ยนน้ำหนักที่เอกสารนั้นทำให้ระบบต้องเสียเวลาในการประมวลผลเพื่อปรับเปลี่ยนค่าน้ำหนักของเอกสารแต่ละเอกสารก่อนจะนำมาเปรียบเทียบความเหมือน ดังนั้นจึงควรที่จะปรับเปลี่ยนน้ำหนักที่ข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่ถูกต้อง ซึ่งจะทำให้ช่วยลดเวลาในการทำงานของระบบได้มากขึ้น

งานวิจัยของ Silva และคณะปี 2004 หาความสัมพันธ์ด้วยขั้นตอนวิธีอะเพียวรี (Aprior Algorithm) ซึ่งจะทำให้มีข้อจำกัดในด้านการประมวลผล เนื่องจากขั้นตอนวิธีนี้ถ้าจำนวนเอกสารมีมากจะทำให้การประมวลผลนั้นใช้เวลาสูงมาก เช่นเดียวกับงานวิจัยการขยายคำในข้อสอบถามโดยรวมกฎความสัมพันธ์กับสิ่งที่ศึกษาร่วมกับเทคนิคการค้นคืนสารสนเทศ (Song et al., 2005) ได้ใช้ขั้นตอนวิธีอะเพียวรี (Aprior Algorithm) ในการหาความสัมพันธ์เพื่อการขยายคำในข้อสอบถามเช่นกัน แต่ในงานวิจัยของ Song และคณะในปี 2005 จะใช้เทคนิคของเวิร์ดเน็ต

(Wordnet)[1] ซึ่งเป็นวิธีคลังคำศัพท์ชนิดหนึ่งมาช่วยในการขยายคำในข้อสอบถามด้วย โดยที่การทำงานของระบบค้นคืนนี้จะมีขั้นตอนดังนี้

1) ระบบจะค้นคืนเอกสารที่เกี่ยวเนื่องออกมาก่อนด้วยระบบการค้นคืนสารสนเทศของเลอมัว (Lemur) [2] ซึ่งเป็นโปรแกรมค้นคืนเอกสารโดยการเปรียบเทียบข้อสอบถามกับเอกสารที่มีคำในข้อสอบถามออกมา
2) นำเอกสารที่ดึงออกมาเลือกคำสำคัญโดยใช้วิธีพีโอเอสแทคกิง (POS-Tagging) ซึ่งเป็นวิธีการตัดคำโดยพิจารณาถึงชนิดของคำว่าทำหน้าที่อะไรในประโยค เช่น คำนาม คำกริยา คำขยาย เป็นต้น ให้น้ำหนักแต่ละคำด้วยค่าความถี่ของคำและค่าความถี่แบบผกผัน (tf-idf) และการลดรูปคำ (Stemming)
3) ขยายข้อสอบถามโดยกฎความสัมพันธ์ที่หาความสัมพันธ์ด้วยขั้นตอนวิธีอะเพียวรี (Aprior Algorithm) ร่วมกับการใช้เวิร์ดเน็ต (Wordnet)

การทดลองการประเมินผลโดยการใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องแทนประสิทธิภาพของระบบ ซึ่งผลที่ได้ออกมาแสดงให้เห็นว่าระบบที่ออกแบบสามารถช่วยเพิ่มประสิทธิภาพในการค้นคืนเอกสารได้ดีกว่าวิธีที่ไม่ใช้เวิร์ดเน็ต (Wordnet) ร่วมด้วย แต่เนื่องจากระบบการค้นคืนดังกล่าวใช้เทคนิคที่มาช่วยเพิ่มประสิทธิภาพการค้นคืนที่หลังจากการค้นหาจากระบบเลอมัว (Lemur) มาก่อน ซึ่งอาจจะทำให้ระบบนั้นเสียเอกสารที่ตรงกับความต้องการของผู้ใช้ไปตั้งแต่ขั้นตอนการใช้ระบบเลอมัว (Lemur) แล้ว

ในการหากฎความสัมพันธ์นั้นยังสามารถนำไปช่วยปรับปรุงการกำหนดดรรชนีให้มีความเหมาะสมกับเอกสารมากยิ่งขึ้นอีกด้วย โดยจะหากฎความสัมพันธ์ด้วยขั้นตอนวิธีโคลส (Close algorithm) ซึ่งจะช่วยเพิ่มประสิทธิภาพในการหากฎความสัมพันธ์จากเอกสารจำนวนมาก โดยมีงานวิจัยที่ใช้ขั้นตอนวิธีนี้นั่นคือ งานวิจัยของ Cherfi และคณะ (2006) เสนอวิธีการทำเหมืองข้อมูลเอกสารโดยใช้การดึงกฎความสัมพันธ์ออกมา (Cherfi et al., 2006) ซึ่งในงานวิจัยของ Cherfi และคณะขั้นตอนแรกจะสร้างดรรชนี จากนั้นระบบจะค้นหากฎความสัมพันธ์ด้วยขั้นตอนวิธีโคลส (Close algorithm) แล้วดึงความสัมพันธ์ที่มีคุณภาพโดยพิจารณาจากการตั้งค่าขั้นต่ำของค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ร่วมกับค่าอื่นอีก 5 ค่าในการคัดเลือกกฎความสัมพันธ์ คือ การวัดค่าความน่าสนใจ (Interesting measure) เป็นวิธีการวัดค่าลิฟท์ (Lift)

---

[1] เป็นระบบอ้างอิงคำศัพท์ออนไลน์ (Online Lexical Reference) ที่เป็นภาษาอังกฤษ โดยจะค้นหาคำที่มีความสัมพันธ์กับคำที่กำหนด (Cognitive Laboratory, 2005)

[2] โปรแกรมค้นคืนเว็บ (Web Browser) ที่ดำเนินการโดย Yahoo (Yahoo, 2005)

ของ IBM การวัดค่าความเชื่อมั่น (Conviction measure) การวัดความขึ้นต่อกัน (Dependency measure) การวัดค่าความแปลกใหม่ (Novelty measure) และค่าความพึงพอใจ (Satisfaction measure)

เมื่อได้กฎความสัมพันธ์ที่มีคุณภาพออกมาแล้ว หากผู้เชี่ยวชาญพิจารณาแล้วพบว่าผลที่ได้จากการหาความสัมพันธ์ยังมีบางความสัมพันธ์ที่ไม่เป็นจริงและการเรียงลำดับความแข็งแรงของความสัมพันธ์นั้นยังไม่ถูกต้องจะทำให้ความสัมพันธ์นั้นไม่ถูกพิจารณา จากความสัมพันธ์ที่ได้รับการตรวจสอบแล้วสามารถนำไปปรับปรุงดรรชนีให้มีความเหมาะสมยิ่งขึ้นได้ เช่น กฎความสัมพันธ์ “mycobacterium tuberculosis” → ”tuberculosis” คือ เมื่อมีคำว่า “mycobacterium tuberculosis” แล้วจะมีคำว่า “tuberculosis” เกิดขึ้นด้วย ผู้เชี่ยวชาญกล่าวว่าดรรชนี “tuberculosis” ไม่เกี่ยวข้องและสามารถทำให้การแปลความหมายเป็นไปในทางที่ผิด (Cherfi et al., 2006)

งานวิจัยของ Cherfi และคณะเป็นงานวิจัยค้นหาความสัมพันธ์ของคำเพื่อปรับปรุงดรรชนี ถ้าดรรชนีถูกกำหนดอย่างเหมาะสมแล้วจะทำให้การค้นคืนมีประสิทธิภาพมากยิ่งขึ้น การปรับปรุงดรรชนีและกฎความสัมพันธ์นั้นจะปรับปรุงโดยการให้ผลสะท้อนกลับจากผู้เชี่ยวชาญเป็นผู้คัดกรองความถูกต้องของกฎความสัมพันธ์และค่าดรรชนีอีกครั้งหนึ่ง

นอกจากนี้ยังมีงานวิจัยที่หากฎความสัมพันธ์ของคำเพื่อสร้างข้อสอบถามในการค้นหา (Qin et al., 2004) Qin และคณะได้เสนอแบบจำลองในการค้นคืนสำหรับแก้ไขปัญหาของข้อสอบถามที่ยังไม่ดีพอ โดยยึดหลักการของความสัมพันธ์ระหว่างคำในคลังเอกสาร ระบบจะมีกระบวนการการทำงานแบ่งออกเป็น 4 ส่วนดังนี้คือ

1) ส่วนต่อประสาน (Interactive interface) ทำหน้าที่ดังนี้
   - เป็นส่วนที่รับข้อสอบถามจากผู้ใช้
   - เป็นส่วนที่แสดงผลลัพธ์การสืบค้นในแต่ละครั้งที่ผู้ใช้สืบค้น
   - เป็นส่วนที่ให้ผู้ใช้เลือกผลลัพธ์ของการสืบค้นที่เกี่ยวเนื่องหรือปรับเปลี่ยนข้อสอบถาม
   - เป็นที่รวมผลสะท้อนจากผู้ใช้แล้วส่งผลสะท้อนนั้นไปยังส่วนการปรับปรุงกฎความสัมพันธ์ (Association rule maintenance module)
2) ส่วนการบำรุงรักษากฎความสัมพันธ์ (Association rule maintenance module) มีหน้าที่ดึงความสัมพันธ์ของคำจากเอกสารที่เป็นผลสะท้อนกลับจากผู้ใช้ แล้วรวมผล

สะท้อนกลับเหล่านั้นเข้ากับเซตของกฎความสัมพันธ์เพื่อสร้างข้อสอบถามใหม่ โดยผู้ใช้สามารถปรับเปลี่ยนค่าน้ำหนักของข้อสอบถามใหม่

3) ตัวสร้างข้อสอบถาม (Query constructor) ส่วนนี้จะใช้กฎความสัมพันธ์และข้อมูลที่ผู้ใช้ป้อนเข้ามาสร้างข้อสอบถาม
4) ตัวประมวลผลข้อสอบถาม (Query processor) ทำหน้าที่ในการคำนวณค่าน้ำหนักความเหมือนระหว่างข้อสอบถามและเอกสารโดยใช้การหาค่าความเหมือนวิธีโคซายน์ (Cosine coefficient)

การทำงานของระบบนี้จะเริ่มจากผู้ใช้ป้อนข้อสอบถามเข้ามายังระบบ จากนั้นระบบจะค้นคืนเอกสารที่เกี่ยวเนื่องกับความต้องการออกมาและเรียงลำดับเอกสารที่เกี่ยวเนื่องกับความต้องการของผู้ใช้จากมากไปน้อย เมื่อผลลัพธ์เอกสารที่ค้นคืนและคำที่มีความสัมพันธ์กับคำในข้อสอบถามของผู้ใช้แสดงออกมายังหน้าจอ เพื่อให้ผู้ใช้สามารถให้ผลสะท้อนเอกสารที่เกี่ยวเนื่องกับความต้องการของผู้ใช้กลับมายังระบบและจากคำที่มีความสัมพันธ์ที่แสดงทางหน้าจอ ผู้ใช้สามารถเพิ่มคำหรือลดคำในข้อสอบถาม เพื่อใช้ค้นคืนเอกสารออกมาอีกครั้ง ระบบจะนำเอกสารเหล่านั้นมาหากฎความสัมพันธ์ใหม่โดยพิจารณาร่วมกับกฎความสัมพันธ์ที่มีอยู่แล้ว เพื่อนำไปขยายคำของข้อสอบถามก่อนนำมาค้นคืนเอกสารใหม่ให้ตรงกับความต้องการของผู้ใช้มากขึ้น

ผลการทดลองการขยายคำโดยใช้กฎความสัมพันธ์การค้นคืนของระบบมีค่าความถูกต้อง (Precision) ที่ดีกว่าการปรับเปลี่ยนข้อสอบถามด้วยวิธีรอคชิโอ (Rochio Algorithm) ซึ่งเป็นวิธีที่ให้ผลสะท้อนกลับสำหรับวิธีแบบจำลองปริภูมิเวกเตอร์

งานวิจัยที่หาความสัมพันธ์ของคำเพื่อการขยายข้อสอบถามอีกงานหนึ่งคือการค้นพบคำที่สัมพันธ์กันโดยใช้กฎความสัมพันธ์ (Relation between Terms Discovery by Association Rules) ของ (Haddad et al., 2000) งานวิจัยของ Haddad และคณะจะเสนอวิธีการขยายคำในข้อสอบถามโดยใช้เทคนิคของการทำเหมืองข้อมูลที่เรียกว่ากฎความสัมพันธ์ที่ทดลองกับเอกสารภาษาฝรั่งเศส นั่นคือ เมื่อคัดกฎความสัมพันธ์ที่มีคุณภาพออกมาแล้ว โดยการตั้งค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ต่ำสุดไว้ ถ้ากฎความสัมพันธ์ใดมีค่าสนับสนุนและค่าความเชื่อมั่นมากกว่าค่าที่ตั้งไว้ แสดงว่ากฎนั้นมีคุณภาพ แล้วนำกฎนั้นมาขยายคำในข้อสอบถาม การทดลองของระบบนั้นจะทดลองกับข้อมูลเอกสาร ซึ่งผลการค้นคืนจะวัดด้วยค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ซึ่งผลการศึกษาพบว่าวิธีขยายคำนี้จะสามารถเพิ่มประสิทธิภาพในการค้นคืนเอกสาร

จากงานวิจัยของ Qin และคณะ ในปี 2004 ที่เปรียบเทียบวิธีการขยายคำในข้อสอบถามโดยการใช้เทคนิคการค้นหากฎความสัมพันธ์กับการปรับปรุงข้อสอบถามด้วยวิธีของ Rochio (Qin et al., 2004) ซึ่งประสิทธิภาพของการใช้วิธีของรอคชิโอ (Rochio) ให้ประสิทธิภาพที่ดีด้วยเช่นกัน และยังไม่มีงานวิจัยใดที่นำเทคนิคการค้นหากฎความสัมพันธ์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ ดังนั้นผู้วิจัยจึงสนใจที่จะนำมาปรับปรุงระบบค้นคืนเอกสารโดยเทคนิคดังกล่าวร่วมกัน เพื่อให้ระบบสามารถค้นคืนเอกสารที่ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น

บทที่ 3

ระเบียบวิธีวิจัย

3.1 บทนำ

ในบทนี้จะกล่าวถึงแนวทางของการทำวิจัย แผนแบบการทดลอง (Experimental Design) การทดสอบสมมติฐาน การทำงานของเครื่องมือทดสอบประสิทธิภาพของการค้นคืนเอกสารรูปแบบต่าง ๆ ที่งานวิจัยกำหนด รวมทั้งแสดงขั้นตอนวิธีการพัฒนาเครื่องมือทดสอบและการทดสอบประสิทธิภาพของการค้นคืนเอกสาร ประเด็นของความเชื่อถือได้ (Reliability) ความถูกต้อง (Validity) และกรอบการวิเคราะห์ข้อมูล (Data Analysis Framework) ดังรายละเอียดต่อไปนี้

3.2 แผนแบบการทดลอง

การศึกษานี้มีวัตถุประสงค์ประสงค์ในการทดลองเพื่อศึกษาเปรียบเทียบประสิทธิภาพการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ นอกจากนี้จะเปรียบเทียบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

จากวัตถุประสงค์งานวิจัยที่กล่าวมาข้างต้น ผู้วิจัยจึงเลือกใช้แผนแบบการทดลองแบบกลุ่มก่อนทดสอบและหลังทดสอบ (One Group Pretest - Posttest Design) ซึ่งเป็นแผนแบบการทดลองที่เหมาะสมกับการทดลองที่ต้องการวัดค่าตัวแปรตามของกลุ่มตัวอย่างก่อนถูกกระตุ้นเทียบกับค่าของตัวแปรตามของกลุ่มตัวอย่างหลังจากถูกกระตุ้นโดยการให้ทรีตเมนต์ (Treatment) นั่นคือ ค่าประสิทธิภาพของระบบค้นคืนเอกสารว่ามีค่าแตกต่างกันอย่างไร โดยกำหนดให้มีตัวแปรในการทดลองเปรียบเทียบการค้นคืนเอกสารในรูปแบบที่ต้องการทดสอบ ดังต่อไปนี้

3.2.1 ตัวแปรต้น

ตัวแปรต้นเป็นตัวแปรที่ผู้วิจัยต้องการศึกษา ซึ่งงานวิจัยนี้สนใจว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ (Association rule) และเทคนิคผลสะท้อนกลับจากผู้ใช้ (Relevant feedback) และการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำนั้นสามารถเพิ่มประสิทธิภาพของระบบค้น

คืนได้หรือไม่ ดังนั้นตัวแปรต้นของการศึกษาในครั้งนี้จะเป็นการเปรียบเทียบการค้นคืนเอกสาร 3 รูปแบบดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วมด้วย
3) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ โดยใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

จากการค้นคืนเอกสารทั้ง 3 รูปแบบ ผู้วิจัยจะเรียกการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้ด้วยคำว่า"การค้นคืนเอกสารรูปแบบที่ 1" ส่วนการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำเรียกว่า"การค้นคืนเอกสารรูปแบบที่ 2" และเรียกการค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ด้วยคำว่า"การค้นคืนเอกสารรูปแบบที่ 3"

3.2.2 ตัวแปรตาม

เนื่องจากงานวิจัยนี้สนใจเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารโดยการใช้เทคนิคดังที่กล่าวในหัวข้อตัวแปรต้น ดังนั้นการเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารจะพิจารณาจากความถูกต้องในการค้นคืนเอกสารที่เกี่ยวเนื่องกับความต้องการของผู้ใช้และเอกสารที่ไม่เกี่ยวเนื่องกับความต้องการของผู้ใช้ออกมามากเพียงใด โดยจะวัดประสิทธิภาพของระบบค้นคืนเอกสารจากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ซึ่งรายละเอียดและวิธีการคำนวณค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ได้กล่าวไว้ในบทที่ 2

3.2.3 ตัวแปรควบคุม

ตัวแปรที่ผู้วิจัยจะควบคุมในการสร้างเครื่องมือทดสอบการค้นคืนเอกสารในรูปแบบต่าง ๆ เพื่อให้ผลการทดลองที่เกิดขึ้นนั้นเกิดจากเทคนิคการค้นคืนเอกสารที่ต้องการทดสอบอย่างแท้จริง โดยจะกำหนดให้ในการทดลองการค้นคืนเอกสารทั้ง 3 รูปแบบมีการควบคุมตัวแปรต่าง ๆ เหมือนกันทุกรูปแบบ ซึ่งจะประกอบด้วยตัวแปร ดังต่อไปนี้

1) เอกสาร

เอกสารนี้เป็นหน่วยตัวอย่างของการทดลองในงานวิจัยครั้งนี้ ซึ่งผู้วิจัยหวังว่าจะทดลองระบบกับเอกสารภาษาอังกฤษทุกเอกสาร แต่ในทางปฏิบัตินั้นไม่สามารถนำเอกสารประเภทนั้นมา

ทั้งหมดได้ ดังนั้นผู้วิจัยจึงเลือกใช้เอกสารจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) จำนวน 425 เอกสาร เป็นหน่วยตัวอย่างในการพัฒนาระบบค้นคืนเอกสาร ซึ่งเป็นเอกสารที่เกี่ยวกับข่าวสารทั่วไป ดังนั้นผู้วิจัยจึงหวังว่าเอกสารนี้มีความใกล้เคียงกับลักษณะของเอกสารทางธุรกิจ ฐานข้อมูลนี้เป็นฐานข้อมูลมาตรฐานที่สร้างโดยมหาวิทยาลัยคอแนล (Cornell University) โดยสร้างมาเพื่อเป็นหน่วยทดสอบระบบค้นคืนเอกสารในงานวิจัยการค้นคืนเอกสาร (Smart Collection, 1963) และเป็นฐานข้อมูลเอกสารที่ใช้ในการทดสอบงานวิจัยทางด้านการค้นคืนเอกสารมากมาย (Dumais, 1991; Lee et al., 1997; Rauber and Merkl, 1999; Rauber and Merkl, 2000)

2) **ข้อสอบถาม**

ในการทดสอบการค้นคืนเอกสารจะต้องใช้ข้อสอบถามเพื่อค้นคืนเอกสารที่เกี่ยวเนื่องกับข้อสอบถามออกมาแสดง โดยจำนวนข้อสอบถามที่จะนำมาทดลองนี้ ผู้วิจัยหวังว่าจะทดลองระบบกับข้อสอบถามที่ได้จากคำทุกคำที่มีอยู่ในระบบ แต่ในความเป็นจริงจำนวนคำที่มีอยู่ในระบบมีจำนวนมากและเนื่องจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) ได้กำหนดข้อสอบถามไว้สำหรับทดสอบระบบค้นคืนเอกสารจำนวน 83 ข้อสอบถาม (Smart Collection, 1963) ดังนั้นผู้วิจัยจึงเลือกข้อสอบถามดังกล่าวเป็นหน่วยตัวอย่างในการทดสอบระบบค้นคืนเอกสารที่พัฒนาขึ้นทั้ง 3 รูปแบบ

3) **ความถูกต้องระหว่างเอกสารและข้อสอบถาม**

ฐานข้อมูลนิตยสารไทม์ (TIME Collection) มีการกำหนดกลุ่มเอกสารที่ถูกต้องในแต่ละข้อสอบถามทั้ง 83 ข้อสอบถามไว้แล้ว ดังนั้นผู้วิจัยจะทราบจำนวนเอกสารที่เกี่ยวเนื่องในการค้นคืนเอกสารแต่ละครั้ง ทำให้สามารถวัดค่าประสิทธิภาพของระบบค้นคืนเอกสารโดยการคำนวณหาค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องได้

4) **ความถูกต้องของผลสะท้อนกลับจากผู้ใช้**

เนื่องจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) ได้มีการกำหนดกลุ่มเอกสารที่ถูกต้องสำหรับข้อสอบถามทั้ง 83 ข้อสอบถามไว้ ดังนั้นการให้ผลสะท้อนกลับผู้วิจัยจึงกำหนดให้ผลสะท้อนกลับมีความถูกต้องเสมอตามที่กำหนดมาในฐานข้อมูลนิตยสารไทม์ (TIME Collection)

5) **เครื่องมือที่ใช้ในการพัฒนาเครื่องมือทดสอบระบบค้นคืนเอกสาร**

งานวิจัยนี้จะต้องใช้เครื่องมือในการพัฒนาเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบ ดังต่อไปนี้

- **โปรแกรมทีเอ็มจี (TMG)** : A MATLAB Toolbox for generating term-document matrices from text collections (Dimitrios and Gallopoulos, 2005)

จากที่กล่าวในบทที่ 3 มาแล้วว่าวิธีการวิจัยที่ใช้เทคนิคต่าง ๆ มาสร้างเวกเตอร์ให้กับเอกสารและข้อสอบถาม ผู้วิจัยได้นำโปรแกรมทีเอ็มจี (TMG) เวอร์ชั่น 2.0R3.0 มาสร้างเวกเตอร์ให้กับเอกสารและข้อสอบถาม ซึ่งโปรแกรมทีเอ็มจี (TMG) นี้เป็นโปรแกรมที่สร้างโดย Dimitrios และ Gallopoulos โดยได้รับลิขสิทธิ์เมื่อปี 2005

โปรแกรมทีเอ็มจี (TMG) จะต้องทำงานบนโปรแกรมแมทแล็บเวอร์ชั่น 6.5 (MATLAB version 6.5) โปรแกรมทีเอ็มจี (TMG) จะสร้างเวกเตอร์ให้กับเอกสารและข้อสอบถาม โดยที่แต่ละมิติของเวกเตอร์จะเป็นตำแหน่งของคำต่าง ๆ ด้วยเทคนิคตามที่ผู้วิจัยกำหนดไว้นั่นคือ การลดรูปคำ (Stemming) การตัดคำยกเว้น (Stop word) ที่ต้องการให้ตัดออกและไม่นำมาพิจารณาในการสร้างเวกเตอร์ และสามารถกำหนดวิธีการคำนวณค่าน้ำหนักคำในแต่ละมิติของเวกเตอร์ซึ่งในงานวิจัยนี้ให้น้ำหนักด้วยค่าความถี่และค่าความถี่ของเอกสารแบบผกผัน (tf-idf) โดยรายละเอียดการใช้งานโปรแกรมนี้ได้แสดงในภาคผนวก ฉ

- **โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1)** (SAS and all other SAS Institute Inc., 2005)

โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ที่ออกแบบมาเพื่อวิเคราะห์ข้อมูลและการทำเหมืองข้อมูล ซึ่งโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) มีส่วนของการวิเคราะห์ข้อมูลด้วยเทคนิคต่างๆ หนึ่งในนั้นคือการทำเหมืองข้อมูลด้วยเทคนิคการค้นหากฎความสัมพันธ์ (Association Discovery) โดยที่ในส่วนนี้สามารถค้นหาข้อมูลต่าง ๆ ต่อไปนี้

- กราฟของกฎความสัมพันธ์ที่เรียงด้วยค่าความเชื่อมั่น (Confidence)
- เส้นกราฟทางสถิติของค่าลิฟท์ (Lift) ค่าความเชื่อมั่น (Confidence) ค่าความเชื่อมั่นที่คาดหวัง (Expected Confidence) และค่าสนับสนุน (Support) ของกฎความสัมพันธ์ (Association Rules)
- แสดงกราฟค่าความเชื่อมั่นที่คาดหวัง (Expected Confidence) เปรียบเทียบกับค่าความเชื่อมั่น (Confidence)
- ตารางแสดงรายละเอียดกฎความสัมพันธ์
- กราฟความสัมพันธ์ของกฎความสัมพันธ์ที่ค้นพบ

- **ไอไอเอส** 5.0 (IIS 5.0)

ไอไอเอส (IIS: Internet Information Services) เป็นส่วนโปรแกรม (Component) ให้บริการด้านเซิฟท์เวอร์ (Server) ในรูปแบบต่างๆของอินเทอร์เน็ต (Internet) ที่รวมทั้ง ไฮเปอร์เทกซ์เซิร์ฟเวอร์ (Hypertext Transfer Protocol server) และไฟล์ทรานสเฟอร์โปรโตคอล เซิร์ฟเวอร์ (File Transfer Protocol server) ซึ่งสามารถทำให้ระบบที่ผู้วิจัยพัฒนาจากภาษาพีเอช พี (PHP) สามารถทำงานได้ ซึ่งไอไอเอส (IIS) นี้รวมมากับระบบปฏิบัติการไมโครซอฟท์วินโดว์เอ็น ที (Windows NT) วินโดว์ 2000 (Windows 2000) และวินโดว์เอกซ์พีโปรเฟสชันนอล (Windows XP Professional) (สมประสงค์ ธิตินิลนิธิ, 2545) โดยสามารถติดตั้งส่วนโปรแกรม (Component) โดยการไปติดตั้งเพิ่มในส่วนโปรแกรมของวินโดว์ (Component) อื่น ๆ ของวินโดว์ (Windows)

- **พีเอชพี** (PHP)

ภาษาพีเอชพี (PHP) คือภาษาที่ทำให้ข้อมูลถูกเปลี่ยนแปลงโดยอัตโนมัติตาม เงื่อนไขต่าง ๆ ที่ผู้เขียนกำหนด (Dynamic Language) และเป็นภาษาประเภท สคริปต์ (Script) ที่สามารถติดต่อกับผู้ใช้ได้ (กิตติ ภักดีวัฒนะกุล และคณะ, 2545) ซึ่งงานวิจัยนี้ใช้เครื่องมือ EditPlus ในการพัฒนาระบบค้นคืนเอกสารโดย ใช้ภาษาพีเอชพี (PHP)

- SQL Server 2000

เป็นโปรแกรมฐานข้อมูลที่ใช้เก็บข้อมูลภายในองค์กรต่าง ๆ ซึ่งนิยมใช้กันทั่วไป โดยเป็นฐานข้อมูลเชิงสัมพันธ์ (Relational Database) ของบริษัทไมโครซอฟท์ที่เป็นรุ่นถัดมาของ เอสคิวแอลเซิร์ฟเวอร์ (SQL Server) โดยจะสนับสนุนภาษาเอสคิวแอล (SQL) ที่สามารถสอบถาม (Query) วิเคราะห์ (Analyze) ตลอดจนจัดการข้อมูลผ่านเว็บ ด้วยการสนับสนุนภาษาเอกซ์เอ็ม แอล (XML) ช่วยในการจัดการข้อมูลทั้งแบบโอแอลทีพี (OLTP: Online Transaction Processing) และโอแอลเอพี (OLAP: Online Analytical Processing) เป็นไปได้อย่างง่ายดาย มีประสิทธิภาพ สูงสุดในการจัดเก็บข้อมูลและวิเคราะห์ข้อมูล (สมพร จิวรสกุล, 2545) อีกทั้งยังจัดการฐานข้อมูล เชิงสัมพันธ์ที่สนับสนุนการทำ “Two phased Commit” (Tight Consistency) เพื่อรักษา เสถียรภาพของข้อมูลระหว่างเซิร์ฟเวอร์ (Server) หลาย ๆ ตัว จากความสามารถด้านต่าง ๆ เหล่านี้ทำให้เอสคิวแอลเซิฟท์เวอร์ 2000 (SQL Server 2000) ใช้ได้กับธุรกิจทั้งขนาดเล็กขนาด กลางและขนาดใหญ่ (บัณฑิต จามรภูติ, 2541)

3.3 สมมติฐานงานวิจัย

จากวัตถุประสงค์ของงานวิจัยคือ ผู้วิจัยต้องการทดสอบว่าการใช้เทคนิคกฎความสัมพันธ์ของคำร่วมกับผลสะท้อนกลับจากผู้ใช้ และเทคนิคการใช้กฎความสัมพันธ์ของคำสามารถเพิ่มประสิทธิภาพของระบบค้นคืนได้หรือไม่ ดังนั้นงานวิจัยนี้จึงต้องการศึกษาประสิทธิภาพของระบบค้นคืนเอกสารทั้ง 3 รูปแบบตามที่กำหนดไว้แล้วในหัวข้อตัวแปรต้น โดยจะตั้งสมมติฐานไว้ดังนี้

กำหนดให้ μ_1 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 3

1) วิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบว่ามีความแตกต่างกันหรือไม่

H_0 : $\mu_1 = \mu_2 = \mu_3$

H_1 : ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารอย่างน้อย 1 คู่มีค่าไม่เท่ากัน

หากผลการทดสอบสมมุติฐานพิสูจน์ที่ออกมาปฏิเสธ H_0 แสดงว่าต้องมีค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารอย่างน้อย 1 คู่ไม่เท่ากัน ซึ่งผู้วิจัยต้องการทราบต่อไปว่าค่าเฉลี่ยของการค้นคืนเอกสารคู่ใดไม่เท่ากัน ดังนั้นผู้วิจัยจึงเปรียบเทียบค่าเฉลี่ยของการค้นคืนเอกสารทีละคู่ทั้ง 3 รูปแบบ โดยตั้งสมมติฐานดังต่อไปนี้

1) เปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 และการค้นคืนเอกสารรูปแบบที่ 2 ผู้วิจัยเห็นว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เพียงอย่างเดียวจะค้นคืนเอกสารที่มีคำที่อยู่ในข้อสอบถามเพียงอย่างเดียวเท่านั้น แต่หากใช้เทคนิคกฎความสัมพันธ์ของคำนั้นร่วมด้วยจะค้นคืนเอกสารที่มีคำที่มีความสัมพันธ์กับข้อสอบถามนั้นด้วย นอกจากจะค้นคืนเพียงคำที่อยู่ในข้อสอบถามอย่างเดียว ดังนั้นผู้วิจัยจึงคาดว่าการใช้เทคนิคการใช้กฎความสัมพันธ์ดีกว่าการไม่ใช้ จึงตั้งสมมติฐานไว้ ดังนี้

H_0: $\mu_2 \leq \mu_1$

H_1: $\mu_2 > \mu_1$

2) เปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 และการค้นคืนเอกสารรูปแบบที่ 3 ผู้วิจัยเห็นว่าการใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้ นอกจากจะค้นคืนเอกสารที่สัมพันธ์กับคำในข้อสอบถามที่ผู้วิจัยคิดว่าน่าจะสามารถเพิ่มประสิทธิภาพให้กับการค้นคืนเอกสารตามที่กล่าวมาในข้อที่ 1 แล้ว ยังสามารถให้ผู้ใช้ให้ผลสะท้อนกลับ โดยให้ผู้ใช้กำหนดเอกสารที่เกี่ยวเนื่องและเอกสารที่ไม่เกี่ยวเนื่องกลับมา เพื่อปรับข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องมากยิ่งขึ้น ดังนั้นผู้วิจัยจึงคาดว่าการใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้ดีกว่าไม่ใช้เทคนิคทั้งสอง จึงตั้งสมมติฐานไว้ ดังนี้

H_0: $\mu_3 \leq \mu_1$

H_1: $\mu_3 > \mu_1$

3) เปรียบเทียบค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 2 และการค้นคืนเอกสารรูปแบบที่ 3 ผู้วิจัยเห็นว่าข้อดีของการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้นั้นจากที่กล่าวมาในข้อ 2 เมื่อใช้ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ ซึ่งผู้วิจัยคิดว่าน่าจะมีประสิทธิภาพดีกว่าการใช้เทคนิคการใช้กฎความสัมพันธ์ของคำเท่านั้นจึงตั้งสมมติฐานไว้ ดังนี้

H_0: $\mu_3 \leq \mu_2$

H_1: $\mu_3 > \mu_2$

3.4 แนวทางการทำวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงทดลอง (Experimental Research) เนื่องจากเป็นการทดลองประสิทธิภาพของระบบค้นคืนเอกสารด้วยการใช้เทคนิคต่าง ๆ โดยในงานวิจัยนี้สนใจเทคนิคการใช้กฎความสัมพันธ์ของคำ (Association Rule) ที่ใช้ร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ (Relevant Feedback) มาช่วยเพิ่มประสิทธิภาพให้กับระบบค้นคืนเอกสาร ซึ่งจะควบคุมตัวแปรอื่น ๆ ให้เหมือนกันนั่นคือ เครื่องมือทดสอบและเอกสาร เพื่อให้ตัวแปรควบคุมที่กำหนดนั้นมีผลกระทบกับตัวแปรตามน้อยที่สุดและผลของงานวิจัยจะได้เป็นผลที่เกิดขึ้นจากการเปลี่ยนตัวแปรต้นอย่างแท้จริง นั่นคืองานวิจัยจะทดลองว่าการค้นคืนเอกสารจะมีประสิทธิภาพเปลี่ยนแปลงไป

อย่างไรเมื่อใช้เทคนิคการค้นคืนเอกสารแตกต่างกัน โดยทดลองสร้างเครื่องมือเพื่อทดสอบประสิทธิภาพของการค้นคืนเอกสาร ดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วมด้วย
3) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ โดยใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

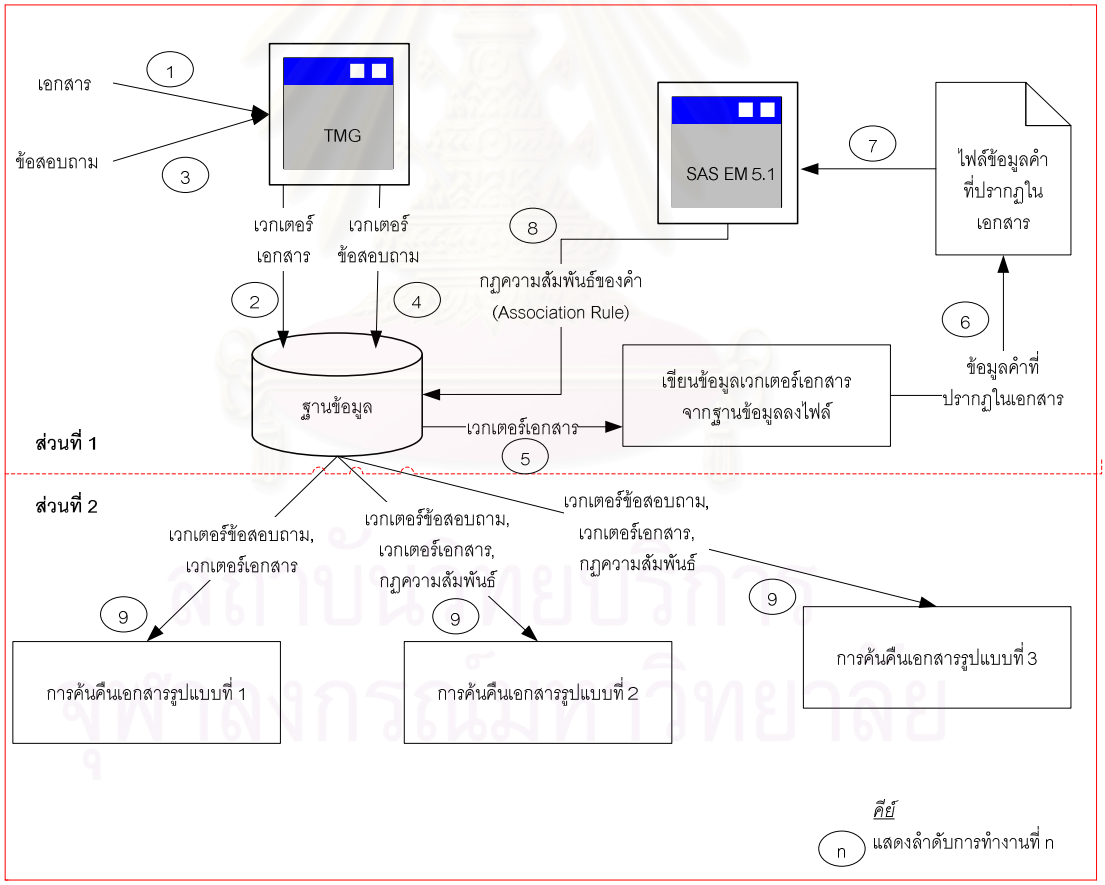
งานวิจัยนี้ได้สร้างระบบค้นคืนเอกสารออกเป็น 3 รูปแบบดังกล่าว เนื่องจากผู้วิจัยสนใจว่าการใช้เทคนิคการหากฎความสัมพันธ์ร่วมกับการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้และการใช้เทคนิคกฎความสัมพันธ์นั้นสามารถเพิ่มประสิทธิภาพของระบบค้นคืนได้หรือไม่ ดังนั้นในการสร้างระบบค้นคืนเอกสารจึงต้องสร้างระบบค้นคืนเอกสารที่ไม่ใช้ 2 เทคนิคดังกล่าวเป็นกลุ่มควบคุมเพื่อเป็นกลุ่มเปรียบเทียบกับระบบค้นคืนเมื่อใส่ทรีตเมนต์ (Treatment) นั่นคือเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

3.5 ภาพรวมการทำงานของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

ตามที่ได้กล่าวมาในหัวข้อแนวทางการทำวิจัยแล้วว่าผู้วิจัยได้พัฒนาเครื่องมือทดสอบเทคนิคระบบค้นคืนเอกสาร 3 รูปแบบ การออกแบบการทดลองของประสิทธิภาพการค้นคืนเอกสารรูปแบบต่าง ๆ นั้นมีภาพรวมดังรูปที่ 3.1 ซึ่งการทดสอบการค้นคืนเอกสารนี้จะแบ่งการสร้างเครื่องมือทดสอบเป็น 2 ส่วน โดยส่วนที่ 1 เป็นส่วนของการแปลงเอกสารให้เป็นเวกเตอร์เอกสาร การแปลงข้อสอบถามให้เป็นเวกเตอร์ข้อสอบถาม และการนำเวกเตอร์เอกสารไปค้นหาคำที่มีความสัมพันธ์กันเก็บลงฐานข้อมูล เพื่อเตรียมข้อมูลไว้ก่อนจะนำข้อมูลเหล่านี้ไปทำการทดสอบต่อในเครื่องมือทดสอบการค้นคืนเอกสารที่ผู้วิจัยพัฒนาต่อไปในส่วนที่ 2 ซึ่งเป็นส่วนที่ผู้วิจัยพัฒนาเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบต่าง ๆ นั่นคือ เทคนิคปริภูมิเวกเตอร์ เทคนิคกฎความสัมพันธ์ของคำ และเทคนิคการใช้ผลสะท้อนกลับของผู้ใช้

การทำงานโดยภาพรวมดังรูปที่ 3.1 ส่วนที่ 1 ขั้นตอนแรกจะนำเอกสารเข้าโปรแกรมทีเอ็มจี (TMG) เพื่อสร้างเวกเตอร์เอกสารเก็บลงฐานข้อมูล จากนั้นจะนำข้อสอบถามเข้าโปรแกรมทีเอ็มจี (TMG) เพื่อสร้างเวกเตอร์ข้อสอบถามเก็บลงฐานข้อมูล ในการสร้างกฎความสัมพันธ์โดย

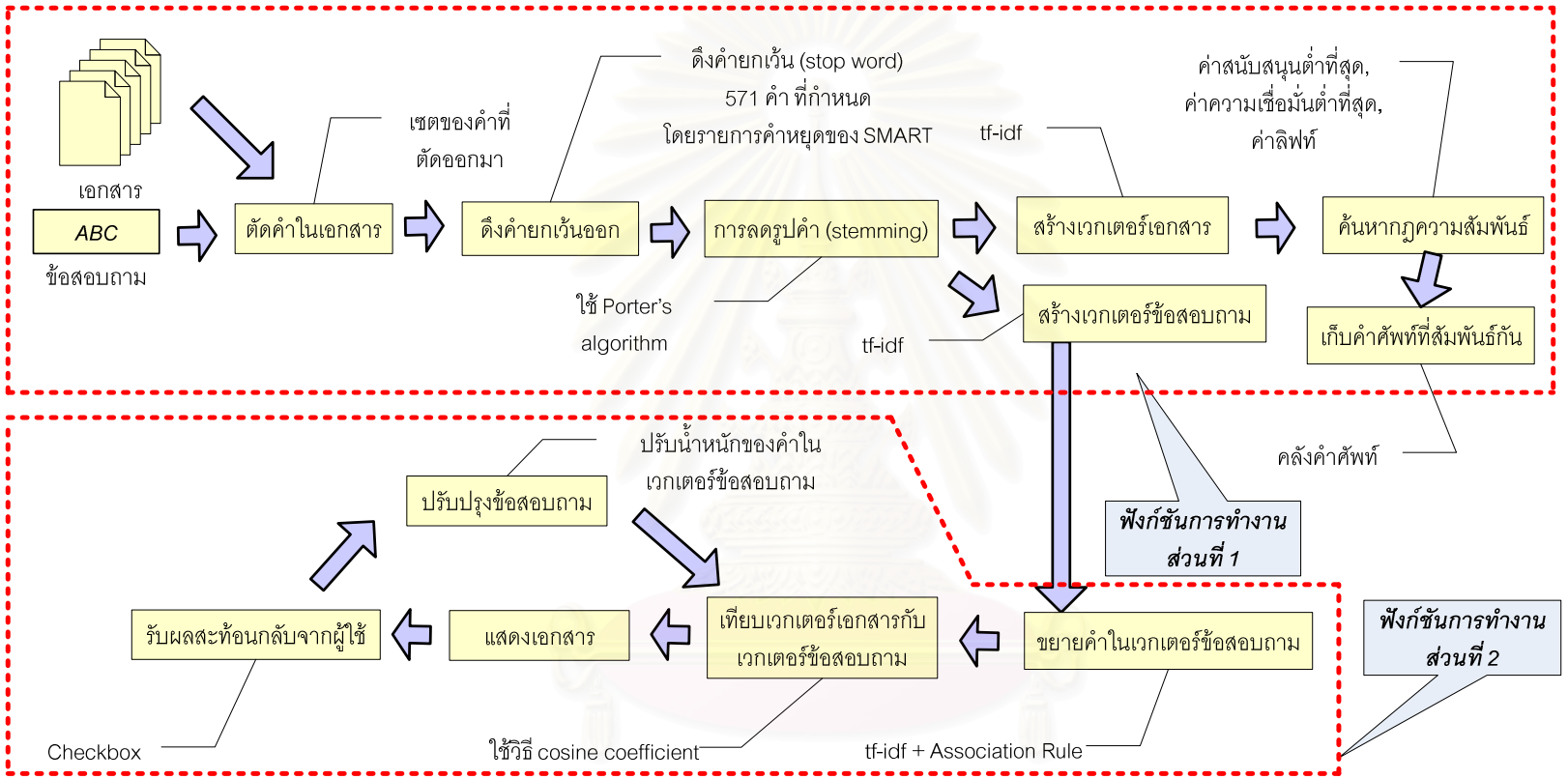
โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) จะนำข้อมูลเวกเตอร์เอกสารในฐานข้อมูลออกมาเขียนลงแฟ้มข้อมูล ก่อนนำแฟ้มข้อมูลนั้นเป็นข้อมูลนำเข้าให้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) เพื่อค้นหากฎความสัมพันธ์ของคำ เมื่อได้กฎความสัมพันธ์ของคำจะนำกฎความสัมพันธ์ของคำนั้นเก็บลงฐานข้อมูลไว้เพื่อเป็นข้อมูลในการขยายคำในข้อสอบถามในการค้นคืนเอกสาร เมื่อเตรียมข้อมูลเวกเตอร์เอกสาร เวกเตอร์ข้อสอบถามและกฎความสัมพันธ์ของคำแล้วจะนำข้อมูลเหล่านี้มาทดสอบกับเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบตามที่กำหนดไว้ โดยการค้นคืนเอกสารรูปแบบที่ 1 จะใช้ข้อมูลเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม การค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 3 จะใช้ข้อมูลเวกเตอร์เอกสาร เวกเตอร์ข้อสอบถามและกฎความสัมพันธ์ของคำ ซึ่งรายละเอียดของขั้นตอนและเทคนิคที่ใช้ในส่วนที่ 1 และส่วนที่ 2 จะกล่าวต่อไป



รูปที่ 3.1 รูปแสดงภาพรวมของเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบ

3.6 องค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

รายละเอียดส่วนนี้จะแสดงหลักการที่ใช้และวิธีการของการสร้างเครื่องมือทดสอบการค้นคืนเอกสาร โดยจะแบ่งเป็นส่วนของการทำงานหลัก 2 ส่วน ดังที่กล่าวในหัวข้อภาพรวมของระบบข้างต้น สามารถแสดงรายละเอียดได้ดังรูปที่ 3.2 โดยในขั้นตอนในส่วนที่ 1 ขั้นตอนการตัดคำในเอกสาร การดึงคำยกเว้นออก การลดรูปคำ การสร้างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม ทั้งหมดเหล่านี้เป็นการทำงานของโปรแกรมทีเอ็มจี (TMG) โดยจะแสดงเทคนิคต่าง ๆ ที่โปรแกรมทีเอ็มจี (TMG) ทำงานและโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) จะทำหน้าที่ในการค้นหากฎความสัมพันธ์ของคำออกมาเพื่อเก็บลงในฐานข้อมูลที่กำหนดให้เป็นคลังคำศัพท์ และในส่วนที่ 2 เป็นส่วนที่แสดงเทคนิคที่ใช้ในการค้นคืนเอกสารและการให้ผลสะท้อนกลับ



รูปที่ 3.2 รูปแสดงภาพรวมองค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารโดยรวม

3.7 เครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

จากองค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารแสดงเทคนิคต่าง ๆ ที่นำมาใช้ในแต่ละขั้นตอน โดยแสดงรายละเอียดดังต่อไปนี้

3.7.1 ส่วนการเตรียมข้อมูลเอกสาร ข้อสอบถามและคำที่มีความสัมพันธ์กัน แบ่งส่วนการเตรียมข้อมูลนี้ตามขั้นตอน ดังนี้

1) ตัดคำในประโยค
2) ดึงคำที่เป็นคำยกเว้นออก (Eliminate Stop word)
3) ลดรูปคำ (Stemming)
4) จัดทำเวกเตอร์เอกสารหรือดรรชนีและเวกเตอร์ข้อสอบถาม
5) ค้นหากฎความสัมพันธ์ (Discover Association Rule)

จากขั้นตอนดังกล่าวจะกล่าวถึงเทคนิคที่ใช้และการทำงานดังต่อไปนี้

- **เทคนิคที่ใช้ในส่วนการเตรียมข้อมูลเอกสาร ข้อสอบถามและคำที่มีความสัมพันธ์กัน**

ขั้นตอนที่ 1 ถึงขั้นตอนที่ 4 นั้นเป็นส่วนการทำงานในโปรแกรมทีเอ็มจี (TMG) และขั้นตอนที่ 5 เป็นส่วนการทำงานของโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) โดยรายละเอียดเทคนิคที่ใช้ในแต่ละขั้นตอนสามารถแสดงได้ดังต่อไปนี้

1) ตัดคำในประโยค

เมื่อได้เอกสารหรือข้อสอบถามมาแล้ว จะนำเอกสารหรือข้อสอบถามเหล่านั้นมาตัดคำในแต่ละเอกสารให้เป็นคำเดี่ยว โดยตัดคำจากการพิจารณาจากช่องว่างระหว่างคำ คำที่แยกด้วยช่องว่างจะถูกตัดออกเป็นคำ 1 คำแล้วเก็บลงฐานข้อมูลไว้

2) ดึงคำที่เป็นคำยกเว้นออก (Eliminate Stop word)

จากคำที่ได้จากการตัดคำในประโยคนำมาเทียบกับตารางคำยกเว้น (แสดงในภาคผนวก ข) ถ้าคำที่ตัดมาจากเอกสารหรือข้อสอบถามคำใดเหมือนกับคำที่อยู่ในตารางคำยกเว้นจะตัดคำนั้นทิ้งไปไม่นำมาพิจารณา เนื่องจากเป็นคำที่ไม่สื่อความหมายของเอกสารหรือข้อสอบถามนั้น ๆ

3) ลดรูปคำ (Stemming)

ขั้นตอนนี้จะใช้ขั้นตอนวิธีของพอร์ทเตอร์ (Porter's Algorithm) (รายละเอียดแสดงในภาคผนวก ค) เมื่อได้คำในเอกสารหรือข้อสอบถามที่ตัดคำยกเว้นออกแล้ว ต่อจากนั้นจะมาเข้าสู่ขั้นตอนวิธีการลดรูปคำ (Stemming) ด้วยขั้นตอนวิธีของพอร์ทเตอร์ ซึ่งผู้วิจัยจะใช้ซอร์สโค้ด (Source code) ของขั้นตอนพอร์ทเตอร์ที่ดาวด์โหลด (Download) ได้จากเว็บไซด์ของพอร์ทเตอร์ (Porter) ที่สร้างไว้ที่ http://www.tartarus.org/~martin/PorterStemmer/index.html (Porter,

1980) โดยคำที่ผ่านการลดรูปคำ (Stemming) แล้วจะถูกนำเข้าสู่ขั้นตอนการจัดทำเวกเตอร์เอกสารหรือดรรชนีและเวกเตอร์ข้อสอบถามต่อไป

4) จัดทำเวกเตอร์เอกสารหรือดรรชนีและเวกเตอร์ข้อสอบถาม

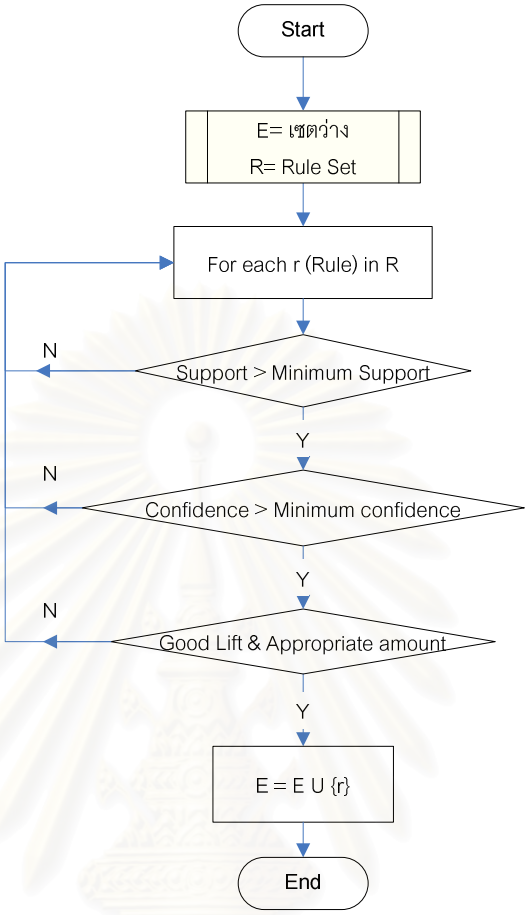
วิธีนี้เป็นการเก็บคำศัพท์ที่ได้จากเอกสารหรือข้อสอบถามที่ตัดคำที่เป็นคำยกเว้นออกและลดรูปคำแล้ว สำหรับเอกสารจะเก็บดรรชนีโดย ใช้หลักการดรรชนีแบบผกผัน (Inverted index) เนื่องจากวิธี Inverted index เป็นวิธีที่ง่าย รวดเร็วในการค้นหาและไม่เสียพื้นที่มากจนเกินไป (Baeza-Yates and Ribeiro-Neto, 1999) ซึ่งทำงานร่วมกับวิธีเก็บคำศัพท์ในคลังคำศัพท์ดังที่กล่าวไว้ในบทที่ 2 ส่วนสำหรับข้อสอบถามนั้นจะจัดเก็บเป็นเวกเตอร์ข้อสอบถาม ซึ่งอยู่ในรูปแบบของดรรชนีแบบผกผัน (Inverted index) เช่นเดียวกัน

เนื่องจากผู้วิจัยเลือกวิธีแบบจำลองปริภูมิเวกเตอร์มาใช้ในการทดสอบเทคนิคการค้นคืนเอกสารนี้ ดังนั้นจึงมีการจัดเอกสารหรือข้อสอบถามให้อยู่ในรูปแบบเวกเตอร์ โดยการจัดทำเวกเตอร์เอกสารหรือข้อสอบถามนั้นจะแปลงเอกสารหรือข้อสอบถามในฐานข้อมูลให้อยู่ในรูปแบบเวกเตอร์ โดยที่ค่าน้ำหนักของแต่ละมิติของเวกเตอร์นั้นจะแสดงถึงความสำคัญของคำนั้นในเอกสารหรือข้อสอบถามสามารถคำนวณค่าน้ำหนักของคำโดยใช้ความถี่ของคำ (Term Frequency: tf) และความถี่ของเอกสารแบบผกผัน (Inverse Document Frequent : idf) ดังที่กล่าวรายละเอียดในบทที่ 2

5) ค้นหากฎความสัมพันธ์ (Discover Association Rule)

การหาคำที่มีความสัมพันธ์กันนั้นสามารถใช้เทคนิคของการทำเหมืองข้อมูลที่เรียกว่าเทคนิคการค้นหากฎความสัมพันธ์ของคำ (Association Rule Discovery) โดยการดึงกฎความสัมพันธ์ที่มีนัยสำคัญจะพิจารณาจากค่าสนับสนุน (Support) ค่าความเชื่อมั่น (confidence) นอกจากนั้นยังพิจารณาร่วมกับค่าลิฟท์ (Lift) ดังที่กล่าวไว้ในบทที่ 2 (Ye, 2001)

ดังนั้นงานวิจัยนี้จะพิจารณาค่าทั้งสิ้น 3 ค่า คือ ค่าสนับสนุน ค่าความเชื่อมั่น ค่าลิฟท์ (Lift) เพื่อหาความสัมพันธ์ที่มีคุณภาพออกมา โดยมีขั้นตอนพิจารณาคัดเลือกกฎความสัมพันธ์ดังรูปที่ 3.3 โดยกำหนดให้ $E = \phi$ และ R = {r | r คือ กฎความสัมพันธ์}



รูปที่ 3.3 รูปแสดงขั้นตอนการคัดกรองกฎความสัมพันธ์ที่มีคุณภาพมาใช้ในระบบ

- **การทำงานของส่วนการเตรียมข้อมูลเอกสาร ข้อสอบถามและคำที่มีความสัมพันธ์กัน**

เทคนิคที่ใช้ในขั้นตอนทั้ง 5 ขั้นตอนดังที่กล่าวมา ในส่วนขั้นตอนที่ 1 ถึงขั้นตอนที่ 4 นั้นเป็นการทำงานของโปรแกรมทีเอ็มจี (TMG) ที่ใช้เทคนิคต่างๆ ดังที่กล่าวมา โดยการใช้โปรแกรมทีเอ็มจี (TMG) เริ่มแรกผู้วิจัยจะโหลดเอกสารและคำหยุดเข้าไปในโปรแกรมทีเอ็มจี (TMG) และเลือกเทคนิคในการให้น้ำหนักคำนั่นคือ ความถี่ของคำ (Term Frequency: tf) และความถี่ของเอกสารแบบผกผัน (Inverse Document Frequent: idf) เมื่อกำหนดเทคนิคต่าง ๆ ทั้งหมดแล้วระบบจะสร้างเวกเตอร์เอกสารออกมาให้

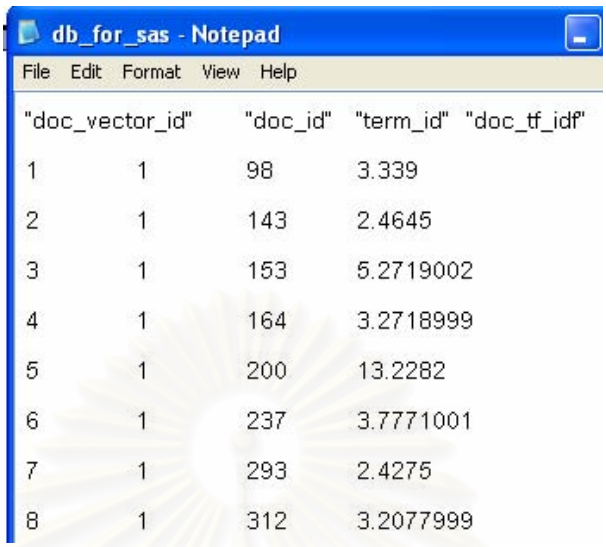
การสร้างเวกเตอร์ข้อสอบถามจะทำหลังจากสร้างเวกเตอร์ข้อสอบถามเสร็จเรียบร้อยแล้ว โดยขั้นตอนแรกจะโหลดข้อสอบถามเข้าไปยังโปรแกรมทีเอ็มจี (TMG) นี้เช่นเดียวกับการสร้าง

เวกเตอร์เอกสาร สามารถดูขั้นตอนและภาพโปรแกรมประกอบของการสร้างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามได้ในภาคผนวก จ

ส่วนในขั้นตอนที่ 5 เป็นขั้นตอนที่นำเวกเตอร์เอกสารที่ได้จากโปรแกรมทีเอ็มจี (TMG) มาหาคำที่มีความสัมพันธ์กันโดยโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) จากนั้นจะเก็บคำที่มีความสัมพันธ์กันออกมาเหล่านั้นลงในคลังคำศัพท์ โดยขั้นตอนการทำงานแสดงดังภาคผนวก ฉ

ขั้นตอนเริ่มแรกผู้วิจัยจะทำการทดลองกำหนดค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ขั้นต่ำในการคัดเลือกกฎก่อน ต่อจากนั้นโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) คัดเลือกกฎความสัมพันธ์ที่มีค่าสนับสนุนและค่าความเชื่อมั่นเกินกว่าค่าที่ตั้งไว้ออกมา จากนั้นผู้วิจัยจะพิจารณาดูค่า Lift โดยผู้วิจัยจะสนใจกฎความสัมพันธ์ที่มีค่า Lift สูง ควบคู่กับการดูจำนวนกฎความสัมพันธ์ที่ได้ออกมา ซึ่งจำนวนกฎความสัมพันธ์จะต้องไม่มากเกินไปและน้อยเกินไปโดยขึ้นอยู่กับจำนวนคำทั้งหมดที่ทำเป็นดรรชนีด้วย ผลของกฎความสัมพันธ์ที่ได้นั้นจะช่วยเพิ่มประสิทธิภาพในการขยายคำในคลังคำศัพท์ให้มีประสิทธิภาพนั่นคือมีการกำหนดคำที่มีความสัมพันธ์กันในตารางคลังคำศัพท์ให้มีมากขึ้น

ในขั้นตอนต่อมาเป็นส่วนของการเตรียมข้อมูลของคำที่มีความสัมพันธ์ โดยนำเวกเตอร์เอกสารมาหาความสัมพันธ์ของคำที่ปรากฏในเวกเตอร์เอกสาร ซึ่งผู้วิจัยเลือกใช้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ในการหาความสัมพันธ์ที่เกิดขึ้น นั่นคือผู้วิจัยจะพัฒนาโปรแกรมโดยส่งข้อมูลตารางเวกเตอร์เอกสารในฐานข้อมูลเอสคิวแอลเซิฟท์เวอร์ 2000 (SQL Server 2000) ออกมาเขียนลงแฟ้มข้อมูลนามสกุล .txt ในรูปแบบที่โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) กำหนด ซึ่งผู้วิจัยเลือกรูปแบบที่แต่ละคอลัมน์ของตารางจะคั่นด้วยแท็บ (Tab) และแต่ละแถวเป็นแต่ละบรรทัดในแฟ้มข้อมูลที่สร้างขึ้น โดยคอลัมน์แรกคือ doc_vector_id จะเป็นคีย์หลัก (Primary key) ของตาราง ซึ่งเป็นรหัสคีย์หลักของตารางเวกเตอร์เอกสาร คอลัมน์ต่อมาเป็น doc_id เป็นรหัสเอกสาร ถัดมาคือ term_id เป็นรหัสคำและ doc_tf_idf เป็นค่าน้ำหนักของคำคำนั้นกับเอกสารรหัสนั้นนั่นคือเป็นรหัสเอกสารนั้นในแต่ละแถว นั่นคือแต่ละแถวจะแสดงคำในแต่ละมิติของเอกสาร ดังตัวอย่างรูปที่ 3.4 จากนั้นจะโหลดแฟ้มข้อมูลตารางที่ได้นี้เข้าไปโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) เพื่อค้นหาคำที่มีความสัมพันธ์กัน โดยขั้นตอนการค้นหากฎความสัมพันธ์ด้วยโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) แสดงดังภาคผนวก ฉ



db_for_sas - Notepad

File Edit Format View Help

"doc_vector_id"	"doc_id"	"term_id"	"doc_tf_idf"
1	1	98	3.339
2	1	143	2.4645
3	1	153	5.2719002
4	1	164	3.2718999
5	1	200	13.2282
6	1	237	3.7771001
7	1	293	2.4275
8	1	312	3.2077999

รูปที่ 3.4 รูปแสดงตัวอย่างตารางก่อนนำเข้าโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1)

ในการค้นหาคำที่มีความสัมพันธ์กัน เนื่องจากผู้วิจัยกำหนดใช้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) เป็นเครื่องมือในการค้นหากฎความสัมพันธ์ของคำ ดังนั้นในการกำหนดค่าสนับสนุนที่ต่ำที่สุด (Minimum Support) และค่าความเชื่อมั่นต่ำที่สุด (Minimum Confidence) จึงพิจารณาจากค่าลิฟท์ (Lift) และจำนวนกฎความสัมพันธ์ที่ค้นหาออกมาได้ โดยผู้วิจัยกำหนดค่าสนับสนุนที่ต่ำที่สุด (Minimum Support) เท่ากับ 1.6470 และค่าความเชื่อมั่นต่ำที่สุด (Minimum Confidence) เท่ากับ 70 เปอร์เซนต์ สำหรับการค้นหากฎความสัมพันธ์ในงานวิจัยนี้ ซึ่งเมื่อดำเนินงาน (Run) โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) โดยการกำหนดค่าสนับสนุนที่ต่ำที่สุด (Minimum Support) และค่าความเชื่อมั่นต่ำที่สุด (Minimum Confidence) ไว้ที่ค่าดังกล่าวข้างต้น ผลลัพธ์กฎความสัมพันธ์ที่ออกมามีจำนวนกฎความสัมพันธ์ 1000 กฎความสัมพันธ์ ที่มีค่าลิฟท์ (Lift) มากที่สุดเท่ากับ 60.71 และค่าลิฟท์ (Lift) น้อยที่สุดเท่ากับ 4.83

จากผลลัพธ์กฎความสัมพันธ์ที่ได้ ในการคัดเลือกกฎความสัมพันธ์ผู้วิจัยจะเลือกกฎความสัมพันธ์ที่มีค่าลิฟท์ (Lift) สูง ๆ และจำนวนกฎความสัมพันธ์ที่ได้ ดังนั้นผู้วิจัยจะเลือกกฎความสัมพันธ์จากผลลัพธ์กฎความสัมพันธ์ที่ได้โดยเลือกกฎความสัมพันธ์ที่มีค่าลิฟท์ (Lift) มากกว่า 40 30 20 และ 10 ตามลำดับเป็น 4 กรณี และจากการเลือกกฎความสัมพันธ์ที่มากกว่าค่าดังกล่าวจะได้จำนวนกฎความสัมพันธ์ในแต่ละกรณีออกมา ดังตารางที่ 3.1

ตารางที่ 3.1 ตารางสรุปการพิจารณาค่าลิฟท์ (Lift) เฉลี่ยและจำนวนกฎความสัมพันธ์ต่าง ๆ โดยที่ค่าสนับสนุนต่ำที่สุด (Minimum support) มีค่าเท่ากับ 1.6471 และค่าความเชื่อมั่นต่ำที่สุด (Minimum confidence) มีค่าเท่ากับ 70

กรณี	ค่าลิฟท์ (Lift) น้อยที่สุด	ค่าลิฟท์ (Lift) เฉลี่ย	จำนวนกฎความสัมพันธ์
1	40	46.8007	74
2	30	41.2096	136
3	20	32.4301	291
4	10	23.3695	571

จากสมการการคำนวณค่าลิฟท์ (Lift) ในบทที่ 2 ผู้วิจัยได้คำนวณค่า Lift จากการกำหนดค่าสนับสนุนต่ำที่สุด (Minimum support) มีค่าเท่ากับ 1.6471 และค่าความเชื่อมั่นต่ำที่สุด (Minimum confidence) มีค่าเท่ากับ 70 ดังนั้นผู้วิจัยจะได้ค่าลิฟท์ออกมาเท่ากับ 25.71 ดังนั้นผู้วิจัยจึงเห็นว่าไม่ควรคัดเลือกกฎความสัมพันธ์ที่มีค่าลิฟท์ (Lift) ที่มีค่าต่ำกว่า 25.71 ออกมา เนื่องจากหากคัดเลือกกฎความสัมพันธ์ที่มีค่าลิฟท์ต่ำกว่าค่านี้ อาจทำให้กฎความสัมพันธ์ที่คัดเลือกออกมานั้นมีจำนวนกฎความสัมพันธ์ที่ไม่มีนัยสำคัญมากเกินไป ดังนั้นจากตาราง 3.1 ผู้วิจัยจึงเลือกกรณีที่ 2 ที่มีค่าลิฟท์ (Lift) น้อยที่สุดเท่ากับ 30 และมีจำนวนกฎความสัมพันธ์เท่ากับ 136 ซึ่งมีค่าลิฟท์ (Lift) ที่ไม่สูงเกินไปและต่ำจนเกินไปเมื่อเทียบกับกรณีอื่น ๆ อีกทั้งจำนวนกฎความสัมพันธ์ไม่ได้มีจำนวนมากเกินไปหรือน้อยเกินไปเช่นกัน

สรุปผลลัพธ์กฎความสัมพันธ์ที่ค้นคืนออกมาได้จากโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) มีทั้งสิ้น 136 กฎความสัมพันธ์ กฎความสัมพันธ์ทั้งหมดแสดงได้ดังภาคผนวก ช

3.7.2 ส่วนการค้นคืนเอกสาร

เป็นส่วนที่ผู้วิจัยพัฒนาเครื่องมือทดสอบขึ้นเอง โดยจะเป็นส่วนเครื่องมือทดสอบที่แตกต่างกันทั้ง 3 รูปแบบ

- **เทคนิคที่ใช้ในส่วนการค้นคืนเอกสาร**

จากขั้นตอนการทำงานของเครื่องมือทดสอบระบบค้นคืนเอกสารทั้ง 3 รูปแบบดังกล่าวจะใช้เทคนิคแตกต่างกันไปตามรูปแบบของการค้นคืนเอกสาร โดยเทคนิคทั้งหมดที่ใช้ในการค้นคืนเอกสารทั้ง 3 รูปแบบแสดงได้ดังนี้

1) การกำหนดเวกเตอร์ข้อสอบถาม

เมื่อผู้วิจัยเลือกข้อสอบถามที่ต้องการทดสอบเข้ามา เครื่องมือจะดึงเวกเตอร์เอกสารที่สร้างไว้แล้วจากโปรแกรมทีเอ็มจี (TMG) ที่เก็บไว้ในฐานข้อมูลออกมา

2) การขยายคำในเวกเตอร์ข้อสอบถาม

เมื่อผู้วิจัยเลือกข้อสอบถามที่ต้องการทดสอบเข้ามายังระบบแล้ว ระบบจะนำคำนั้นไปเทียบกับกฎความสัมพันธ์ที่สร้างไว้จากโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ในฐานข้อมูลก่อนหน้านี้ ถ้ามีกฎความสัมพันธ์ของคำคำนั้นอยู่ จะนำคำที่มีความสัมพันธ์กันกับคำในข้อสอบถามนั้นเข้ามาพิจารณาเพื่อค้นคืนเอกสารด้วย

คำที่ปรากฏในข้อสอบถามและคำที่สัมพันธ์กันตามกฎความสัมพันธ์นั้นตรงกับตำแหน่งใดมิติของเวกเตอร์ตำแหน่งนั้นก็จะถูกให้ค่าน้ำหนักตามสมการที่ 2.6 ที่กล่าวในบทที่ 2 โดยคำที่มีความสัมพันธ์จะมีค่าน้ำหนักเป็นสัดส่วนตามค่าความเชื่อมั่นของกฎนั้น ๆ ที่คำนวณมาได้ เช่น กำหนดให้ t มีค่าเท่ากับ 6 ถ้าเวกเตอร์คำในเอกสารทั้งหมดเป็นดังนี้ T= (a, b, c, d, e, f) เมื่อผู้วิจัยกรอกข้อมูลคำว่า “c” เข้ามายังระบบ ระบบค้นพบกฎความสัมพันธ์ “ $c \rightarrow e, f$ ” ที่มีค่าความเชื่อมั่นเท่ากับ 90% ดังนั้นคำที่จะนำมาเพื่อค้นคืนเอกสารจึงมีคำ “c” “e” และ “f” ด้วย โดยที่ถ้าสมมุติระบบหาค่าน้ำหนักของ c ในข้อสอบถามออกมาได้เท่ากับ 0.8 จะทำให้ค่าน้ำหนักในตำแหน่งที่ e และ f มีค่าน้ำหนักเท่ากับ 0.8 * (90 / 100) = 0.72 ทำให้เวกเตอร์ข้อสอบถามจะมีค่าน้ำหนักของคำในแต่ละตำแหน่งดังนี้ (0, 0, 0.8, 0, 0.72, 0.72)

3) เปรียบเทียบความเหมือนระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถาม

งานวิจัยนี้ของเสนอแบบจำลองปริภูมิเวกเตอร์จะคำนวณระดับความเหมือนของเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถาม โดยหาความสัมพันธ์ (Correlation) ระหว่างเวกเตอร์ข้อสอบถามและเวกเตอร์เอกสารที่สามารถคำนวณจากสมการที่ 2.9 ด้วยวิธีคำนวณค่าความเหมือนโคไซน์ (Cosine coefficient) (Baeza-Yates and Ribeiro-Neto, 1999; Chowdhury, 2004)

4) แสดงเอกสารแก่ผู้วิจัย

ในการค้นคืนเอกสารจะต้องนำเวกเตอร์ข้อสอบถามมาคำนวณค่าความเหมือนกับเวกเตอร์เอกสาร โดยจะตั้งค่าความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามต่ำสุดไว้ ดังนั้นหลังจากที่คำนวณระดับความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามแล้วเครื่องมือทดสอบจะค้นคืนเอกสารที่มีความเหมือนมากกว่าค่าความเหมือนต่ำสุดที่ตั้งไว้ออกมาแสดงต่อผู้วิจัยทางหน้าจอคอมพิวเตอร์

จากงานวิจัยของ Udomchaiporn Akadej ดังที่กล่าวมาแล้วในบทที่ 2 ได้เสนอไว้ว่าการตั้งค่าความเหมือนสามารถตั้งได้ตามความเหมาะสมกับระบบค้นคืนเอกสารนั้น ๆ โดยจะสามารถ

ตั้งไว้ที่ค่าเฉลี่ย (Mean) หรือค่าเฉลี่ยบวกกับค่าเบี่ยงเบนมาตรฐาน (Mean + Standard Deviation) หรือมากกว่านี้ได้ตามความเหมาะสม ดังนั้นผู้วิจัยจึงคำนวณหาค่าเฉลี่ย (Mean) และค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าความเหมือนของทุกข้อสอบถามกับทุกเอกสาร ได้ผลออกมาดังนี้

ค่าเฉลี่ย (Mean) = 0.0120

ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) = 0.0319

จากผลการทดลองที่ได้นั้น ค่าความเหมือนแต่ละข้อสอบถามและเอกสารนั้นมีค่าเฉลี่ย (Mean) เท่ากับ 0.0120 และมีค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) เท่ากับ 0.0319 เพราะฉะนั้นถ้าผู้วิจัยตั้งค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) หรือค่าเฉลี่ย (Mean) ลบกับค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ค่าที่ได้เป็นค่าติดลบ นั่นคือถ้าผู้วิจัยตั้งค่านี้เป็นค่าความเหมือนต่ำสุด จะทำให้ระบบค้นคืนเอกสารออกมาทุกเอกสาร เนื่องจะค่าการคำนวณความเหมือนโคไซน์ (Cosine coefficient) มี่ค่าความเหมือนตั้งแต่ค่า 0 ถึง 1

และจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) มีการกำหนดเอกสารที่เกี่ยวเนื่องกับข้อสอบถาม ดังนั้นจึงสามารถหาเปอร์เซนต์ของเอกสารที่ไม่เกี่ยวเนื่องถูกดึงออกมาแสดงได้ ซึ่งถ้าหากตั้งค่าความเหมือนต่ำสุดเท่ากับค่าเฉลี่ย (Mean) จะทำให้การค้นคืนเอกสารมีเอกสารที่ไม่เกี่ยวเนื่องถูกค้นคืนออกมาของข้อสอบถามทั้ง 83 ข้อสอบถามคิดเป็น 24.1644 เปอร์เซนต์ ในขณะที่ถ้าตั้งค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) จะทำให้การค้นคืนเอกสารมีเอกสารที่ไม่เกี่ยวเนื่องถูกค้นคืนออกมาของข้อสอบถามทั้ง 83 ข้อสอบถาม 4.4026 เปอร์เซนต์ แสดงว่าเมื่อตั้งค่าความเหมือนต่ำสุดเท่ากับค่าเฉลี่ย (Mean) จะค้นคืนเอกสารที่ไม่เกี่ยวเนื่องกับทั้ง 83 ข้อสอบถามออกมามากกว่าการตั้งค่าความเหมือนต่ำสุดไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) 5 เท่า ดังนั้นเอกสารที่ไม่เกี่ยวเนื่องถูกค้นคืนออกมามากจนเกินไป ดังนั้นผู้วิจัยจึงกำหนดค่าความเหมือนไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) นั่นคือมีค่าเท่ากับ 0.0439

5) รับผลสะท้อนกลับจากผู้วิจัย

เอกสารที่แสดงออกมาเป็นผลลัพธ์ทางหน้าจอ จะมีช่องเลือก (Checkbox) ให้ผู้วิจัยกำหนดเอกสารที่ไม่เกี่ยวเนื่องดังที่ได้กำหนดมาจากข้อมูลตัวอย่างที่ใช้ โดยผู้วิจัยจะพิจารณาเทียบจากกลุ่มเอกสารที่เป็นคำตอบที่ถูกต้องของข้อสอบถามนั้น ๆ ซึ่งได้มีการกำหนดไว้แล้วมาจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) เพื่อส่งผลสะท้อนกลับไปยังเครื่องมือทดสอบ

เพื่อให้เครื่องมือทดสอบปรับปรุงข้อสอบถามให้ค้นคืนเอกสารออกมาให้เกี่ยวเนื่องกับข้อสอบถามที่ผู้วิจัยกรอกเข้าไปยังระบบมากขึ้นในขั้นตอนต่อไป

6) ปรับปรุงข้อสอบถามจากผลสะท้อนกลับของผู้วิจัย

ผู้วิจัยจะปรับปรุงข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องกับความต้องการของผู้วิจัยมากขึ้น นั่นคือเครื่องมือทดสอบสามารถค้นคืนเอกสารที่มีคำในข้อสอบถามปรากฏอยู่ โดยผู้วิจัยหวังว่าจะใช้เทคนิคการปรับปรุงข้อสอบถามของรอคชิโอ (Rochio) (Baeza-Yates and Ribeiro-Neto, 1999) นั่นคือผู้วิจัยต้องพิจารณาข้อสอบถามเดิมด้วย โดยพิจารณาร่วมกับการให้ค่าน้ำหนักกับกลุ่มเอกสารที่เกี่ยวเนื่องมากกว่ากลุ่มเอกสารที่ไม่เกี่ยวเนื่องตามสูตรที่ 2.11 ที่แสดงในบทที่ 2 ซึ่งจะต้องให้ค่าน้ำหนักในเวกเตอร์ข้อสอบถามเดิมหรือค่า α ค่าน้ำหนักกับกลุ่มเวกเตอร์เอกสารที่เกี่ยวเนื่องหรือค่า β และค่าน้ำหนักกับกลุ่มเวกเตอร์เอกสารที่ไม่เกี่ยวเนื่องหรือค่า γ ดังนั้นผู้วิจัยจึงกำหนดค่าคงที่หรือค่าน้ำหนักที่เหมาะสมเมื่อได้รับผลสะท้อนกลับจากผู้วิจัยคือค่า α, β, γ เท่ากับ 8, 16, 4 ตามที่กำหนดในงานวิจัยที่กล่าวข้างต้น เนื่องจากผลการทดลองของงานวิจัยความถูกต้องของการให้ผลสะท้อนกลับและการจัดกลุ่มเอกสารของ Iwayama (Iwayama, 2000) และงานวิจัยเรื่องผลกระทบเมื่อให้ผลสะท้อนกลับของ Buckley และคณะ (Buckley et al., 1994) ในงานวิจัยอ้างอิง (ในบทที่ 2) นั้นสามารถช่วยเพิ่มประสิทธิภาพให้กับระบบค้นคืนเอกสารที่ใช้เทคนิคการให้ผลสะท้อนกลับได้ เมื่อตั้งค่า α, β, γ ไว้ที่ 8, 16, และ 4 ตามลำดับ

7) คำนวณค่าประสิทธิภาพของการค้นคืนเอกสาร

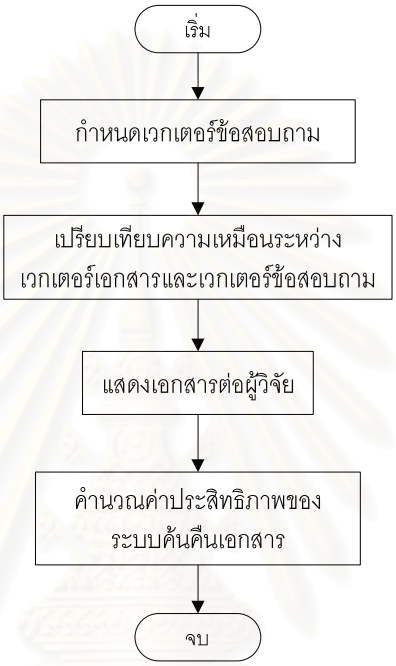
งานวิจัยนี้ใช้วิธีการคำนวณค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) เนื่องจากเป็นค่าที่คำนวณจากค่าเรียกคืนและค่าความถูกต้องมาเฉลี่ย การคำนวณค่าเฉลี่ยฮาร์โมนิคนั้นจะต้องรู้เอกสารที่เกี่ยวเนื่องกับข้อสอบถามแต่ละข้อสอบถาม ข้อมูลดังกล่าวระบุไว้แล้วในฐานข้อมูลนิตยสารไทม์ (TIME Collection) ที่ใช้ในการทดลอง

- **การทำงานของส่วนการค้นคืนเอกสาร**

การทำงานของเครื่องมือทดสอบการค้นคืนเอกสารนั้นจะต้องมีการเตรียมข้อมูลเพื่อทดสอบการค้นคืนเอกสารเรียบร้อยแล้ว นั่นคือข้อมูลเวกเตอร์เอกสาร เวกเตอร์ข้อสอบถาม และกฎความสัมพันธ์ของคำที่ได้จากการทำงานส่วนที่ 1 โดยเก็บลงฐานข้อมูลไว้ การพัฒนาเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารในรูปแบบต่าง ๆ ทั้ง 3 รูปแบบดังที่กำหนดไว้ข้างต้นมีรายละเอียดการทำงานดังต่อไปนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้

ขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 ซึ่งเป็นรูปแบบการใช้เทคนิคปริภูมิเวกเตอร์เท่านั้น มีขั้นตอนการทำงานดังรูปที่ 3.5

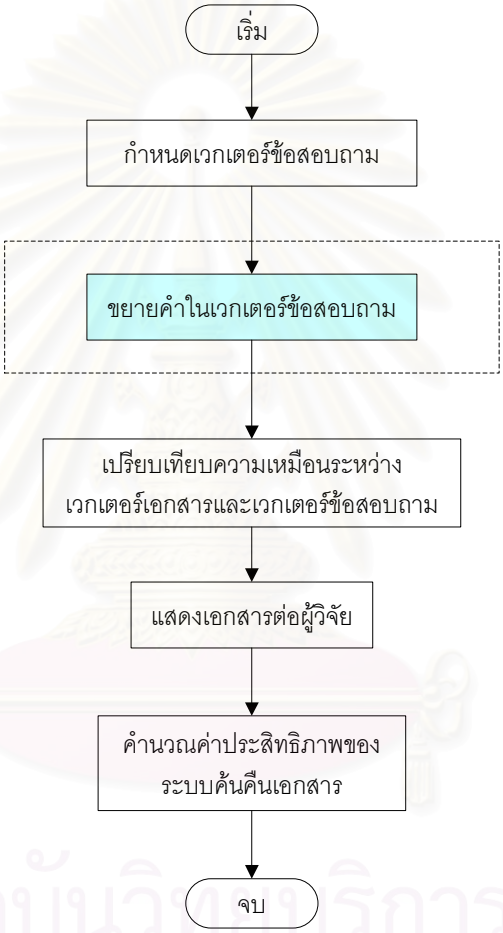


รูปที่ 3.5 รูปแสดงขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1

การทำงานจากเริ่มผู้วิจัยจะกำหนดข้อสอบถามที่ต้องการทดสอบเข้ามายังเครื่องมือทดสอบ ต่อมาเครื่องมือจะเทียบความเหมือนระหว่างเอกสารและข้อสอบถาม โดยคำนวณหาค่าความเหมือนระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถามนั้น ๆ ถ้าค่าความเหมือนของเวกเตอร์เอกสารกับข้อสอบถามใดมีค่ามากกว่าค่าความเหมือนต่ำที่สุดที่ตั้งไว้จะแสดงเอกสารนั้นออกทางหน้าจอ จากนั้นนำเอกสารที่เป็นผลลัพธ์มาเปรียบเทียบหาจำนวนเอกสารที่ถูกต้อง เพื่อคำนวณหาค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) แล้วเก็บลงฐานข้อมูล โดยเอกสารที่ถูกต้องสำหรับข้อสอบถามแต่ละข้อได้ถูกกำหนดไว้ในฐานข้อมูลนิตยสารไทม์ (TIME Collection) เครื่องมือจะแสดงค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องทางหน้าจอแก่ผู้วิจัยด้วย เป็นอันเสร็จสิ้นระบบค้นคืนเอกสารของข้อสอบถามหนึ่ง ๆ

2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

ขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ซึ่งเป็นรูปแบบการใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ มีขั้นตอนการทำงานดังรูปที่ 3.6

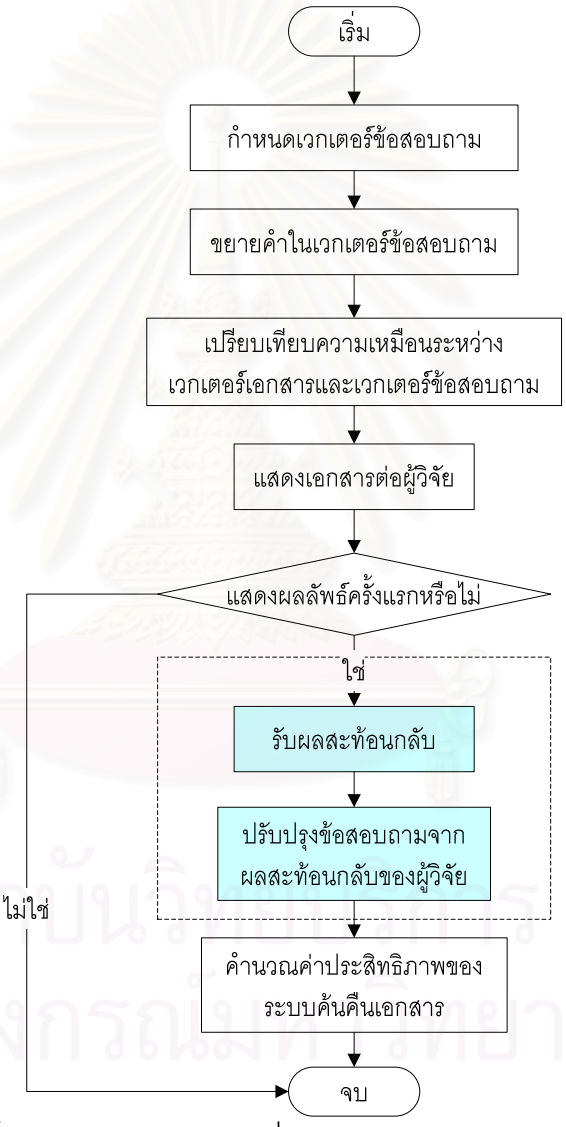


รูปที่ 3.6 รูปแสดงขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2

การทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 นั้นจะเพิ่มการทำงานจากระบบค้นคืนเอกสารรูปแบบที่ 1 ในส่วนที่อยู่ในกรอบสี่เหลี่ยมเส้นประดังรูปที่ 3.6 โดยจะเป็นส่วนการทำงานของการหาคำที่มีความสัมพันธ์กับคำที่ปรากฏในข้อสอบถามที่ผู้วิจัยเลือกเข้ามา

3) การค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

ขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ซึ่งเป็นรูปแบบการใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ มีขั้นตอนการทำงานดังรูปที่ 3.7



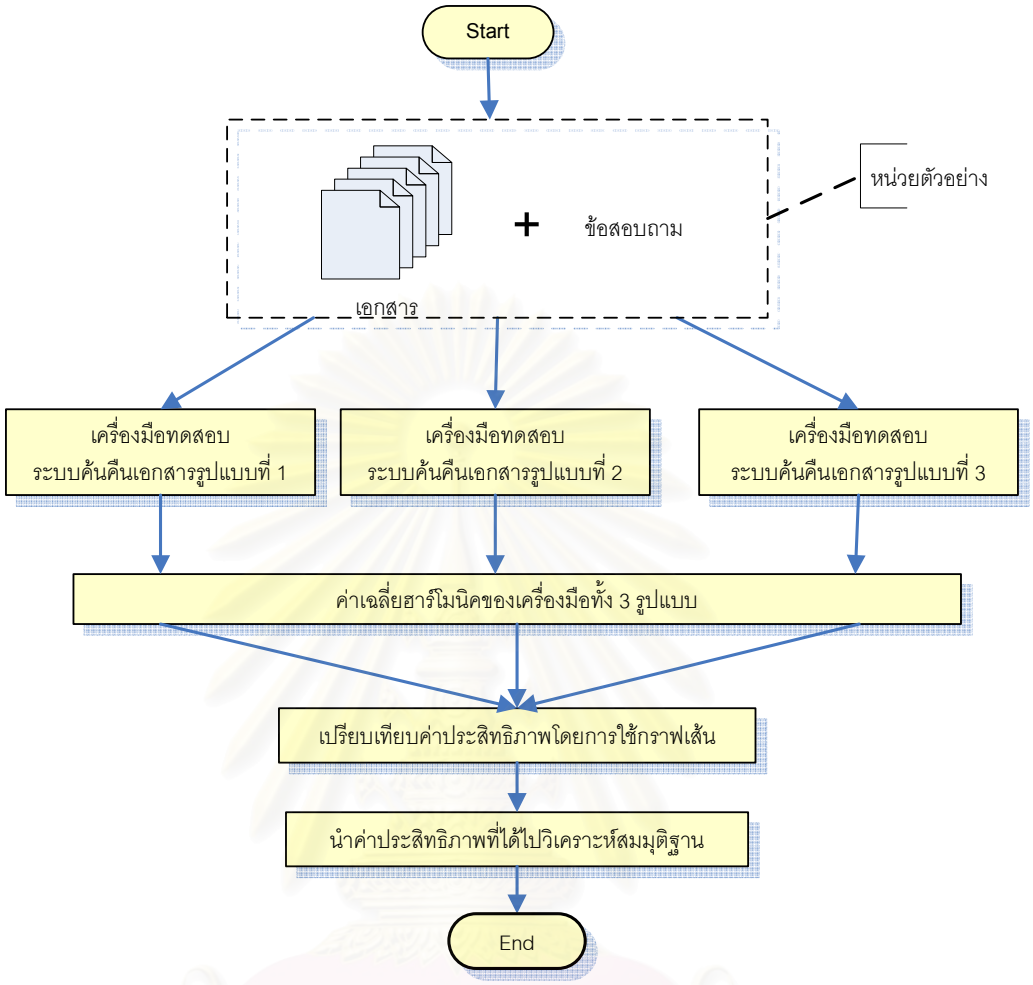
รูปที่ 3.7 รูปแสดงขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3

การทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 นั้นจะมีการทำงานเช่นเดียวกับระบบค้นคืนเอกสารรูปแบบที่ 2 แต่เพิ่มเติมส่วนที่อยู่ในกรอบสี่เหลี่ยมเส้นประดังรูปที่ 3.7 ซึ่งเป็นส่วนของการให้ผลสะท้อนกลับจากผู้วิจัย โดยผู้วิจัยจะเลือกเอกสารที่เกี่ยวเนื่องกับข้อสอบถามตามที่ฐานข้อมูลนิตยสารไทม์ (TIME Collection) กำหนดมาเข้ามายังระบบเพื่อให้

ระบบคำนวณค่าน้ำหนักคำในแต่ละมิติของเวกเตอร์ข้อสอบถามใหม่อีกครั้งตามสูตรของรอคชิโอ (Rochio) ซึ่งขั้นตอนการปรับน้ำหนักคำแต่ละมิติจะไม่พิจารณาถึงความสัมพันธ์ของคำแต่จะพิจารณาเพียงคำที่อยู่ในเอกสารที่เกี่ยวเนื่องและคำที่อยู่ในเอกสารที่ไม่เกี่ยวเนื่องตามที่ผู้วิจัยกำหนดเข้ามาเท่านั้น เช่น ถ้าข้อสอบถามที่ผู้วิจัยเลือกเข้ามามีคำ “a” แล้วคำ “a” มีความสัมพันธ์กับคำ “b” และ ”c” ดังนั้นเมื่อใช้เทคนิคกฎความสัมพันธ์ของคำแล้วข้อสอบถามจะมีคำ “a” ”b” และ ”c” ไปค้นคืนเอกสารออกมา เมื่อผู้วิจัยให้ผลสะท้อนกลับการปรับน้ำหนักคำ “a” ”b” และ ”c” จะเป็นอิสระต่อกันความสัมพันธ์ของทั้ง 3 คำไม่มีผลกระทบต่อกันนั่นคือ ถ้าผลสะท้อนกลับเอกสารที่เกี่ยวเนื่องมีเพียงคำ “a” เท่านั้น ระบบจะปรับน้ำหนักคำ “a” จะไม่ปรับน้ำหนักคำ “b” และ “c” ตามความสัมพันธ์ที่มีการกำหนดไว้ เมื่อปรับค่าน้ำหนักคำในเวกเตอร์ข้อสอบถามแล้วจากนั้นนำเวกเตอร์ข้อสอบถามใหม่ที่ได้นี้ไปเลือกเอกสารที่เกี่ยวเนื่องใหม่อีกครั้ง แล้วคำนวณหาค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ต่อไป โดยการให้ผลสะท้อนกลับนี้ ผู้วิจัยได้กำหนดให้ผู้วิจัยให้ผลสะท้อนกลับเพียงครั้งเดียวเท่านั้น

3.8 **ขั้นตอนในการทดสอบประสิทธิภาพเทคนิคการค้นคืนเอกสาร** (Baeza-Yates and Ribeiro-Neto, 1999)

เมื่อสร้างระบบค้นคืนเอกสารทั้ง 3 รูปแบบเสร็จสิ้นแล้ว ต่อจากนั้นจะวัดประสิทธิภาพการค้นคืนเอกสาร โดยจุดมุ่งหมายหลักของการวัดประสิทธิภาพการค้นคืนเอกสารคือ เอกสารที่เกี่ยวเนื่องกับข้อสอบถามที่ผู้วิจัยเลือกเข้าไปทดสอบถูกค้นคืนออกมา ซึ่งงานวิจัยนี้จะใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) มาเป็นค่าแสดงประสิทธิภาพของระบบค้นคืนเอกสาร เมื่อนำชุดเอกสารของฐานข้อมูลนิตยสารไทม์ (TIME Collection) มาทดสอบกับเครื่องมือที่สร้างขึ้น เครื่องมือจะค้นคืนเอกสารทั้งหมดที่ตรงกับเงื่อนไขที่ผู้วิจัยกำหนดไว้ในหัวข้อข้างต้น ดังนั้นจะได้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องทั้ง 83 ข้อสอบถามจากการค้นคืนเอกสารทั้ง 3 รูปแบบออกมา จากนั้นพิจารณาประสิทธิภาพการค้นคืนเอกสารตามค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องที่คำนวณได้มาเปรียบเทียบโดยการใช้กราฟในรูปแบบที่เหมาะสม เพื่อง่ายต่อการพิจารณาเปรียบเทียบค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของระบบค้นคืนเอกสารทั้ง 3 รูปแบบ ซึ่งขั้นตอนโดยสรุปของการทดสอบแสดงได้ดังรูปที่ 3.8



รูปที่ 3.8 รูปแสดงขั้นตอนการทดสอบระบบ

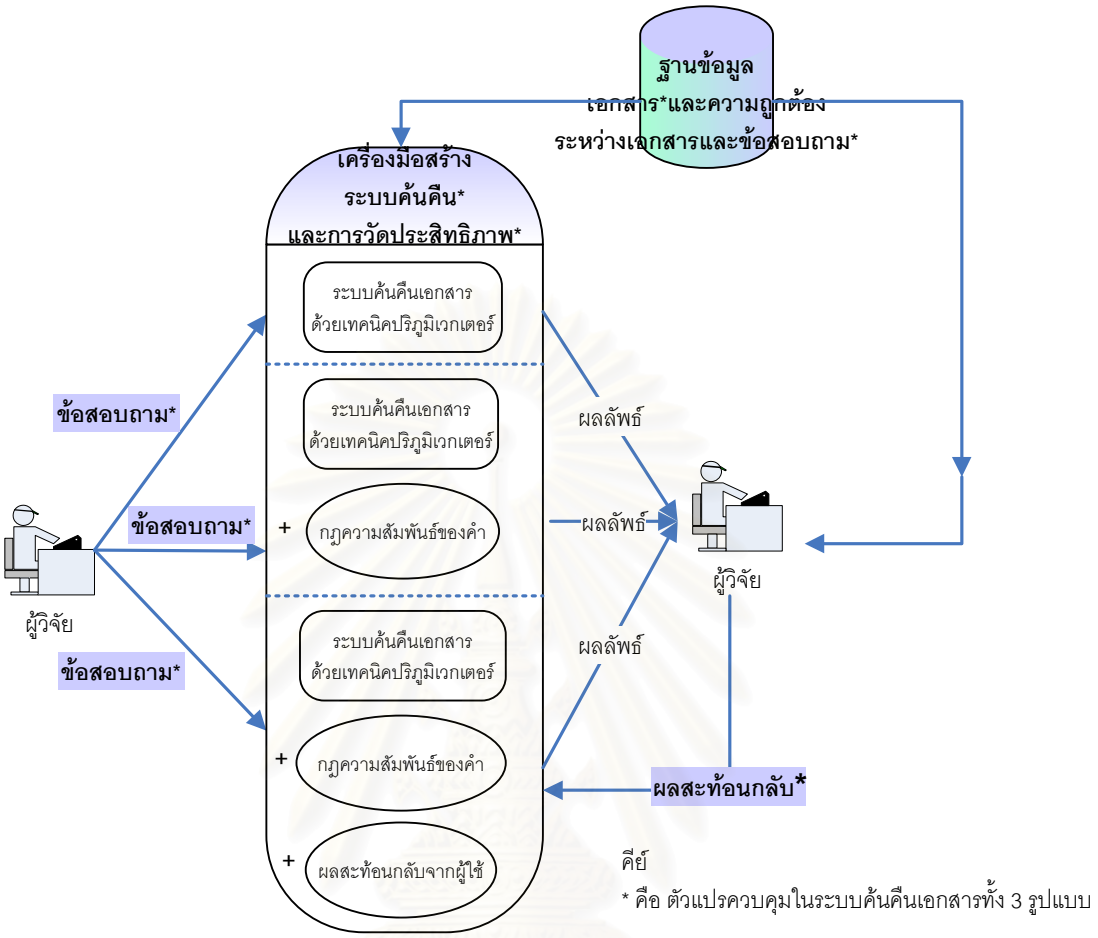
3.9 ความถูกต้อง (Validity) และความน่าเชื่อถือ (Reliability) ของข้อมูลที่เก็บ

การตอบวัตถุประสงค์ของข้อมูลงานวิจัยให้เชื่อถือได้ (Reliability) และถูกต้อง (Validity) จำเป็นต้องควบคุมปัจจัยที่เกี่ยวข้องอันได้แก่ การเลือกเอกสาร การเลือกข้อสอบถามและทดสอบประสิทธิภาพของระบบ

เนื่องจากงานวิจัยนี้ต้องการที่จะศึกษาถึงผลกระทบจากตัวแปรอิสระ (ตัวแปรต้น) คือ ระบบค้นคืนเอกสารที่ใช้เทคนิคต่าง ๆ กัน ซึ่งตัวแปรอิสระนี้เป็นปัจจัยที่ต้องเปลี่ยนค่าไปตามแบบแผนการทดลองเพื่อดูความแตกต่างอันเกิดขึ้นจากการทดลอง นอกจากนั้นยังต้องสามารถควบคุมปัจจัยในด้านต่าง ๆ ให้มีความเหมือนกันหรือมีความคงที่ภายใต้สภาวะเดียวกัน เพื่อผลการทดลองที่สะท้อนเป็นค่าของตัวแปรอิสระที่ได้รับการทรีตเมนต์ (Treatment) นั่นคือ เทคนิคการใช้

กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้วิจัยเท่านั้น โดยในการทดลองระบบมีปัจจัยที่ต้องควบคุมดังรูปที่ 3.9 โดยมีรายละเอียดดังนี้

1) การเลือกเอกสารและข้อสอบถามที่นำมาใช้ทดลองในระบบค้นคืนเอกสารทั้ง 3 รูปแบบ ผู้วิจัยกำหนดให้เอกสารและข้อสอบถามที่จะทดสอบคือกลุ่มข้อมูลเดียวกัน เพื่อให้ประสิทธิภาพที่วัดเกิดจากเทคนิคที่แตกต่างกันอย่างแท้จริง
2) เอกสารและข้อสอบถามที่ผู้วิจัยเลือกมาเป็นหน่วยตัวอย่างนั้นเป็นข้อมูลที่เผยแพร่ต่อสาธารณะและได้ถูกนำไปศึกษาในงานวิจัยอื่น ๆ อย่างแพร่หลาย (Willet, 1988)
3) เอกสารที่นำมาทดสอบในงานวิจัยนี้เป็นบทความของนิตยสารไทม์ (TIME) ในปี 1963 ซึ่งเป็นเรื่องราวข่าวสารทั่วไป ทำให้สามารถเป็นตัวแทนของเอกสารที่มีความหลากหลายประเภทและสามารถนำไปประยุกต์ใช้กับเอกสารด้านอื่น ๆ รวมทั้งงานเอกสารด้านธุรกิจได้
4) การกำหนดเอกสารที่เกี่ยวเนื่องกับข้อสอบถามแต่ละข้อสอบถามทั้ง 83 ข้อสอบถาม จะถูกกำหนดมาในชุดเอกสารที่นำมาทดสอบนั่นคือฐานข้อมูลนิตยสารไทม์ (TIME Collection)
5) เนื่องจากฐานข้อมูลเอกสารและข้อสอบถามที่นำมาทดสอบระบบ ได้มีการกำหนดข้อสอบถามและรายการเอกสารที่เป็นคำตอบของข้อสอบถามไว้ชัดเจนแล้ว โดยผู้สร้างกลุ่มทดสอบนี้ อีกทั้งฐานข้อมูลนี้ยังได้รับการนำไปใช้ทดสอบในงานวิจัยมากมาย (Dumais, 1991; Lee et al., 1997; Rauber and Merkl, 1999; Rauber and Merkl, 2000) ซึ่งทำให้กระบวนการพิจารณาผลลัพธ์ที่ระบบแสดงออกมาในขั้นตอนการทดสอบระบบค้นคืนเอกสารในงานวิจัยนี้จะสามารถเชื่อถือความถูกต้องของรายการเอกสารที่เป็นผลลัพธ์ของข้อสอบถาม
6) การทดสอบประสิทธิภาพของระบบซึ่งจะใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ที่เป็นการคำนวณจากการคำนวณร่วมกันระหว่างค่าของค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) เป็นมาตรวัดที่นิยมใช้ในการทดลองทดสอบระบบการค้นคืนสารสนเทศ (Baeza-Yates and Ribeiro-Neto, 1999) โดยจะเป็นการวัดว่าระบบสามารถค้นคืนเอกสารออกมาได้ถูกต้องตรงกับความต้องการของผู้วิจัยหรือไม่
7) เครื่องมือที่ใช้ในการพัฒนาระบบค้นคืนเอกสารทั้ง 3 รูปแบบในงานวิจัยเป็นเครื่องมือประเภทเดียวกัน



รูปที่ 3.9 รูปแสดงตัวแปรที่ควบคุมในการสร้างระบบค้นคืนเอกสารทั้ง 3 รูปแบบ

3.10 **กรอบการวิเคราะห์ข้อมูล** (Data Analysis Framework)

งานวิจัยนี้ใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องเป็นค่าที่แสดงถึงประสิทธิภาพของระบบค้นคืนเอกสาร เมื่อทดสอบประสิทธิภาพของระบบเสร็จสิ้นแล้ว ทำให้ได้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง 83 ค่า เนื่องจากในการทดสอบประสิทธิภาพของระบบในงานวิจัยใช้ชุดข้อสอบถามจำนวน 83 ข้อสอบถาม จากนั้นในขั้นตอนแรกจะตรวจสอบการแจกแจกของค่าประสิทธิภาพที่ได้มาว่า มีการแจกแจงปกติหรือไม่ เพื่อเลือกทางเลือกในการทดสอบสมมติฐานได้ว่าจะให้การทดสอบสมมติฐานแบบใช้พารามิเตอร์ (Parametric Test) หรือแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ถ้าผลการทดสอบพบว่าประชากรมีการแจกแจงแบบปกติ จึงใช้การวิเคราะห์ความแปรปรวน (ANOVA: Analysis of Variant) อันเป็นวิธีการทดสอบสมมติฐานที่ใช้กับการทดลองที่มีปัจจัย 2 ปัจจัยขึ้นไป (Parametric Test) โดยอาศัยตัวทดสอบสถิติทดสอบแบบ F โดยแสดงค่าลงในตารางวิเคราะห์ความแปรปรวน

แบบจำแนก 2 ทาง (Two-Factor Analysis of Variant) เนื่องจากมีตัวแปรอิสระสองตัวที่มีอิทธิพลต่อค่าประสิทธิภาพนั่นคือ เทคนิคที่ใช้ในระบบค้นคืนเอกสารและข้อสอบถาม แต่ถ้าผลทดสอบการแจกแจงประชากรพบว่ามีการแจกแจงไม่ปกติในแต่ละระบบค้นคืนเอกสารทั้ง 3 รูปแบบ ต้องใช้วิธีการทดสอบสมมติฐานแบบไม่ใช้พารามิเตอร์ (Non Parametric Test) ต่อไป

เมื่อพิสูจน์สมมติฐานแรกแล้วผลการทดสอบสมมติฐานออกมามีค่าปฏิเสธ H_0 จะทดสอบสมมติฐานแต่ละคู่ของระบบเอกสารทั้ง 3 รูปแบบว่ามีความแตกต่างกันอย่างไรในแต่ละคู่ ดังที่กล่าวไว้ในหัวข้อสมมติฐานของงานวิจัยมาแล้ว โดยใช้เทคนิคการเปรียบเทียบเชิงซ้อน (Multiple Comparison) ของ Turkey เพื่อเปรียบเทียบผลต่างระหว่างค่าเฉลี่ยครั้งละคู่ ซึ่งจะสามารถพิจารณาค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของระบบค้นคืนเอกสารแต่ละคู่ว่าเท่ากันหรือไม่เท่ากันอย่างมีนัยสำคัญหรือไม่

บทที่ 4

ผลการวิเคราะห์ข้อมูล

4.1 บทนำ

ในบทนี้จะแสดงผลและวิเคราะห์เปรียบเทียบข้อมูลจากการทดสอบเทคนิคการค้นคืนเอกสารทั้ง 3 รูปแบบ เพื่อนำมาตอบวัตถุประสงค์ของงานวิจัยที่กล่าวมาข้างต้น ซึ่งได้แก่การเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ที่ไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้ การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วมและการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ และการศึกษาและวิเคราะห์ข้อมูลเพิ่มเติม

4.2 ผลการทดลอง

การศึกษานี้มีจุดประสงค์เพื่อวัดประสิทธิภาพของเทคนิคการค้นคืนเอกสาร 3 รูปแบบที่กำหนดในบทที่ 3 โดยสร้างเครื่องมือทดสอบตามการค้นคืนเอกสารที่ต้องการทดสอบ ดังนี้

เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 ใช้เทคนิคค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ที่ไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้

เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ใช้เทคนิคการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำ

เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 3 ใช้เทคนิคการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ที่ได้จากการทดลองแสดงในตารางที่ 4.1

ตารางที่ 4.1 ตารางแสดงผลการทดลองค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องการค้นคืนเอกสารทั้ง 3 รูปแบบ

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสาร		
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*
1	0.3636	0.3636	0.3636
2*	0.1111	0.1143	0.1143
3	0.1132	0.1132	0.1071
4	0.2222	0.2222	0.2222
5	0.2326	0.2326	0.2326
6	0.3913	0.3913	0.3913
7	0.2000	0.2000	0.2000
8*	0.2353	0.2857	0.2857
9*	0.6364	0.7000	0.7368
10*	0.5217	0.5714	0.5714
11	0.2353	0.2353	0.2222
12*	0.5600	0.5600	0.5600
13	0.3158	0.3158	0.3158
14	0.3333	0.3333	0.2857
15	0.5714	0.5714	0.5714
16	0.2727	0.2727	0.2609
17*	0.2222	0.2667	0.2500
18	0.3333	0.3333	0.2857
19	0.4167	0.4167	0.4167
20	0.0000	0.0000	0.1538
21	0.1250	0.1250	0.1290

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสาร		
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*
22	0.2000	0.2000	0.1905
23	0.4000	0.4000	0.2000
24	0.0870	0.0870	0.0952
25	0.1000	0.1000	0.1000
26	0.0000	0.0000	0.2500
27	0.2143	0.2143	0.1818
28	0.2424	0.2424	0.2500
29*	0.2105	0.2353	0.2353
30	0.2941	0.2941	0.2857
31	0.3478	0.3478	0.4516
32	0.0400	0.0400	0.0333
33	1.0000	1.0000	1.0000
34	0.2857	0.2857	0.2857
35	0.1818	0.1818	0.1818
36	0.5000	0.5000	0.5000
37	0.2222	0.2222	0.2500
38	0.0000	0.0000	0.0800
39	0.4865	0.4865	0.4737
40	0.3830	0.3830	0.3830
41	0.2703	0.2703	0.2500
42	0.0526	0.0526	0.0500

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสาร		
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*
43	0.1739	0.1739	0.1429
44	0.2667	0.2667	0.2222
45	0.2353	0.2353	0.2857
46	0.5667	0.5667	0.5484
47	0.3448	0.3448	0.2941
48	0.1538	0.1538	0.1053
49	0.5161	0.5161	0.5000
50	0.0833	0.0833	0.0800
51	0.1579	0.1579	0.1395
52	0.0500	0.0500	0.0455
53	0.1905	0.1905	0.1379
54	0.0755	0.0755	0.0702
55	0.4490	0.4490	0.4151
56	0.0909	0.0909	0.0769
57	0.2222	0.2222	0.2222
58	0.5185	0.5185	0.5000
59	0.1600	0.1600	0.1538
60	0.1818	0.1818	0.1818
61	0.6842	0.6842	0.5532
62	0.2222	0.2222	0.2500
63	0.4400	0.4400	0.5000

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสาร		
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*
64	0.1905	0.1905	0.2000
65	0.0870	0.0870	0.0870
66	0.1429	0.1429	0.1290
67	0.3750	0.3750	0.3750
68	0.3429	0.3429	0.3243
69*	0.7647	0.7879	0.7879
70	0.1333	0.1333	0.1250
71	0.5714	0.5714	0.5714
72	0.0645	0.0645	0.0571
73*	0.1538	0.1429	0.1538
74*	0.1905	0.2353	0.2353
75*	0.0833	0.1053	0.1053
76	0.3030	0.3030	0.3030
77	0.0833	0.0833	0.0800
78	0.1538	0.1538	0.1538
79*	0.0952	0.1000	0.1176
80	0.4571	0.4571	0.4324
81	0.0364	0.0364	0.0667
82	0.3704	0.3704	0.3030
83	0.1176	0.1176	0.1176

หมายเหตุ เครื่องหมาย * หมายถึงข้อสอบถามที่ถูกขยายคำในเวกเตอร์ด้วยกฎความสัมพันธ์

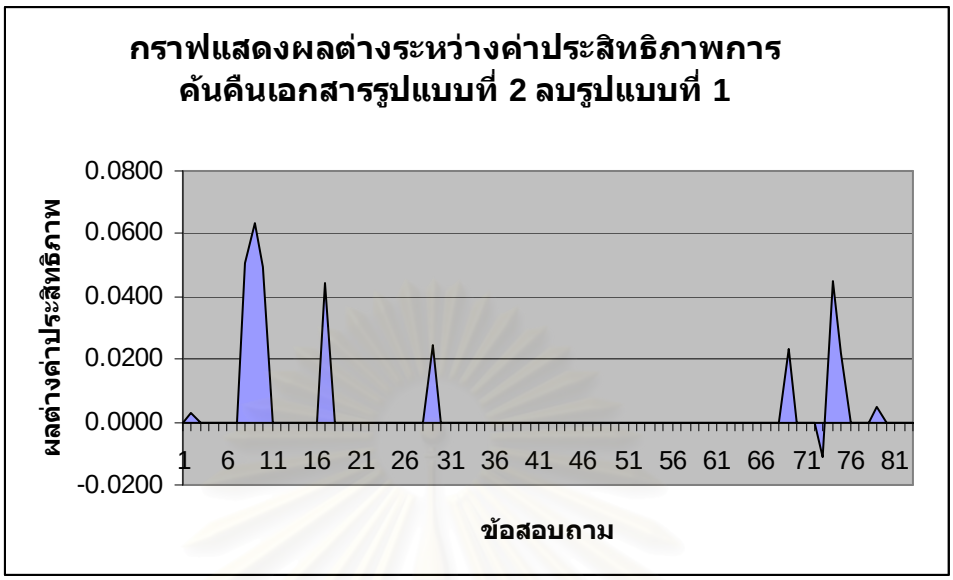
จากการทดลองการค้นคืนเอกสารที่ใช้เทคนิคกฎความสัมพันธ์ของคำนั้น ข้อสอบถามที่ขยายด้วยกฎความสัมพันธ์มีจำนวน 12 ข้อสอบถามจาก 83 ข้อสอบถาม โดยการเปรียบเทียบค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ของข้อสอบถาม 12 ข้อสอบถามระหว่างการค้นคืนเอกสารรูปแบบที่ 1 และการค้นคืนเอกสารรูปแบบที่ 2 สรุปได้ดังนี้

- ข้อสอบถามที่สามารถเพิ่มประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 ให้มีค่ามากกว่าการค้นคืนเอกสารรูปแบบที่ 1 มีจำนวน 10 ข้อสอบถาม
- จำนวนข้อสอบถามที่ทำให้ประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มีค่าเท่ากับการค้นคืนเอกสารรูปแบบที่ 1 มี 1 ข้อสอบถาม
- ข้อสอบถามที่ทำให้การค้นคืนเอกสารรูปแบบที่ 2 มีประสิทธิภาพลดลงเมื่อเทียบกับการค้นคืนเอกสารรูปแบบที่ 1 มีจำนวน 1 ข้อสอบถาม

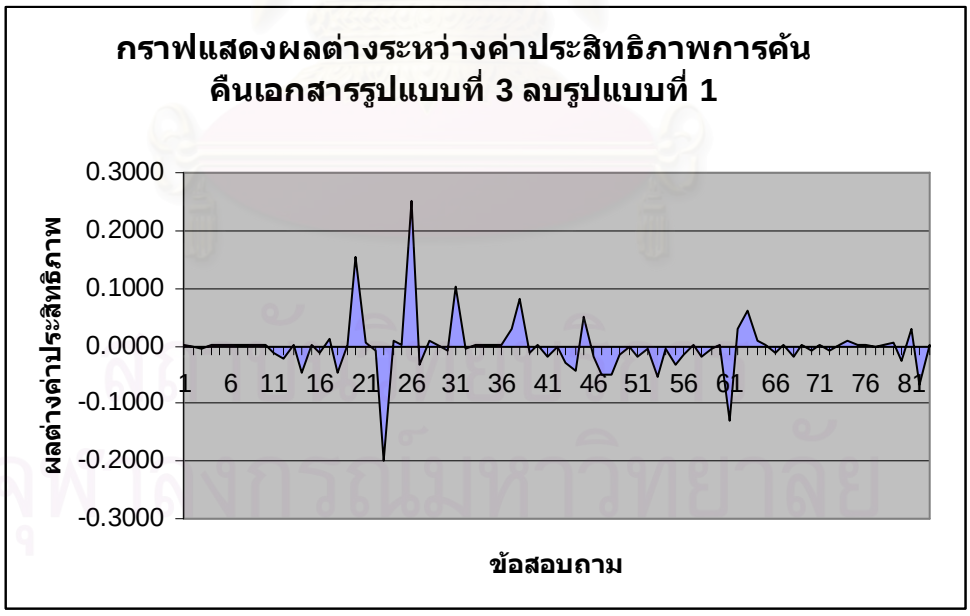
การค้นคืนเอกสารรูปแบบที่ 3 มีข้อสอบถามที่ได้รับทั้งสองเทคนิคคือเทคนิคกฎความสัมพันธ์และเทคนิคผลสะท้อนกลับจากผู้ใช้จำนวน 12 ข้อสอบถามเช่นกัน ซึ่งเป็นข้อสอบถามเดียวกันกับข้อสอบถามที่ได้รับการขยายคำด้วยกฎความสัมพันธ์ การเปรียบเทียบค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ระหว่างการค้นคืนเอกสารรูปแบบที่ 2 และการค้นคืนเอกสารรูปแบบที่ 3 สามารถสรุปได้ดังนี้

- ข้อสอบถามของการค้นคืนเอกสารรูปแบบที่ 3 ที่มีประสิทธิภาพเพิ่มขึ้นเมื่อเทียบกับการค้นคืนเอกสารรูปแบบที่ 2 มีจำนวน 3 ข้อสอบถาม
- ข้อสอบถาม 1 ข้อสอบถามของการค้นคืนเอกสารรูปแบบที่ 3 มีประสิทธิภาพน้อยลงเมื่อเทียบกับการค้นคืนเอกสารรูปแบบที่ 1
- ข้อสอบถามอีก 8 ข้อสอบถามมีค่าเท่ากันระหว่างการค้นคืนเอกสารรูปแบบที่ 2 และการค้นคืนเอกสารรูปแบบที่ 3

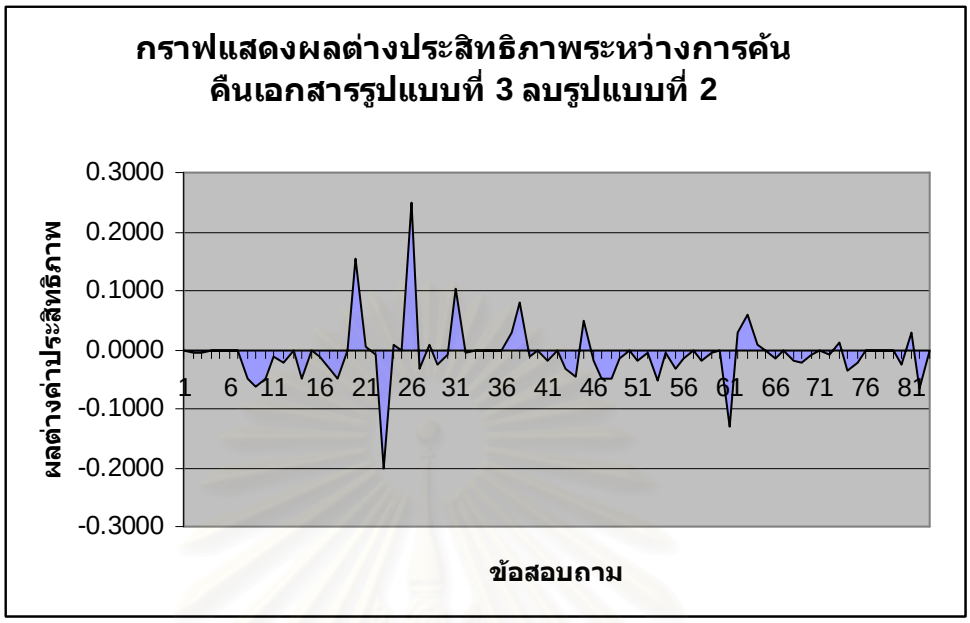
จากผลการทดลองค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ตารางที่ 4.1 ผู้วิจัยจะแสดงกราฟเพื่อเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบโดยใช้กราฟแสดงพื้นที่ (Area Chart) ที่แสดงผลต่างค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารทั้ง 3 รูปแบบดังรูปที่ 4.1 รูปที่ 4.2 และรูปที่ 4.3 และสรุปผลทดลองดังตารางที่ 4.4



รูปที่ 4.1 รูปแสดงกราฟแสดงค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 2 ลบกับรูปแบบที่ 1



รูปที่ 4.2 รูปแสดงกราฟแสดงค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 3 ลบกับรูปแบบที่ 1



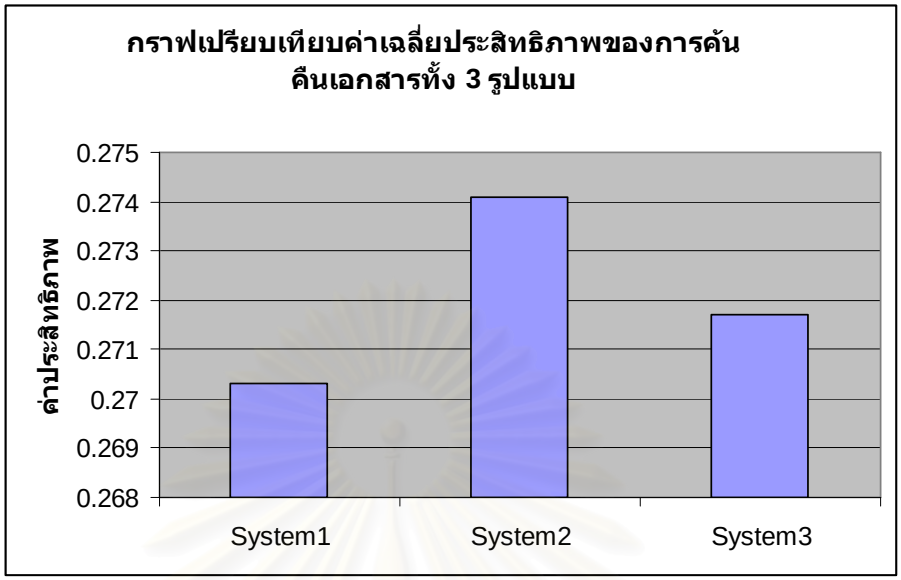
รูปที่ 4.3 รูปแสดงกราฟแสดงค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 3 ลบกับรูปแบบที่ 2

ตารางที่ 4.2 ตารางสรุปผลการทดลองค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสาร

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision)		
	การค้นคืนเอกสาร รูปแบบที่ 1	**การค้นคืนเอกสาร รูปแบบที่ 2**	**การค้นคืนเอกสาร รูปแบบที่ 3**
ค่าเฉลี่ย (Mean)	0.2703	0.2741	0.2717
ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)	0.1906	0.1932	0.1861

จากตารางสรุปผลการทดลองแสดงให้เห็นว่า

- ค่าเฉลี่ยของการค้นคืนเอกสารรูปแบบที่ 2 มีค่ามากที่สุด ต่อมาเป็นรูปแบบที่ 1 และรูปแบบที่ 3 ตามลำดับ ดังรูปที่ 4.4
- การค้นคืนเอกสารรูปแบบที่ 2 มีค่าเบี่ยงเบนมาตรฐานของค่าประสิทธิภาพมากที่สุด ตามด้วยรูปแบบที่ 1 และรูปแบบที่ 3 ตามลำดับ



รูปที่ 4.4 รูปแสดงกราฟเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision)
ระหว่างการค้นคืนเอกสารทั้ง 3 รูปแบบ

เนื่องจากค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบมีค่าแตกต่างกันดังรูปที่ 4.4 แต่ไม่สามารถสรุปได้ว่าประสิทธิภาพแตกต่างกันอย่างมีนัยสำคัญ จึงวิเคราะห์ผลการทดลองต่อไปว่าประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบแตกต่างกันอย่างมีนัยสำคัญหรือไม่ ดังรายละเอียดดังต่อไปนี้

4.3 ผลการวิเคราะห์ข้อมูล

การตรวจสอบเงื่อนไขพื้นฐานขั้นแรก ผู้วิจัยต้องตรวจสอบการแจกแจกของประชากรว่าการแจกแจงปกติหรือไม่ เพื่อเลือกทางเลือกในการทดสอบสมมติฐานว่าจะใช้วิธีการทดสอบสมมติฐานแบบอิงพารามิเตอร์ (Parametric Test) หรือแบบไม่อิงกับพารามิเตอร์ (Non Parametric Test) ถ้าผลการทดสอบพบว่าประชากรมีการแจกแจงแบบปกติ จึงจะสามารถใช้การวิเคราะห์ความแปรปรวน (ANOVA: Analysis of variance) อันเป็นวิธีทดสอบสมมติฐานที่ใช้กับการทดลองที่มีปัจจัย 2 ปัจจัยขึ้นไปในแบบอิงพารามิเตอร์ (Parametric Test) ได้ที่กล่าวไว้ในบทที่ 3 แต่ถ้าผลทดสอบพบว่าการแจกแจงประชากรไม่เป็นแบบปกติ ผู้วิจัยต้องใช้วิธีการทดสอบสมมติฐานแบบไม่ใช้พารามิเตอร์ (Non Parametric Test) (กัลยา วานิชย์บัญชา, 2543)

4.3.1 การวิเคราะห์การแจกแจงข้อมูล

ในงานวิจัยนี้ผู้วิจัยสนใจตัวแปร คือประสิทธิภาพของเทคนิคการค้นคืนเอกสาร ดังนั้นจึงตรวจสอบการแจกแจงของข้อมูลที่ได้จากหน่วยทดลอง นั่นคือประสิทธิภาพของการค้นคืนเอกสาร ซึ่งได้มาจากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ที่คำนวณจากการค้นคืนเอกสารที่ทดสอบโดยใช้ข้อสอบถามจำนวน 83 ข้อ สอบถามไปดึงเอกสารในฐานข้อมูลซึ่งมีจำนวน 423 เอกสารออกมา ดังนั้นผู้วิจัยจะตรวจสอบว่าค่าประสิทธิภาพของการค้นคืนเอกสารมีการแจกแจงแบบปกติหรือไม่จากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) 83 ค่าที่ได้ในการค้นคืนเอกสารแต่ละรูปแบบ โดยตั้งสมมติฐานของการทดสอบสำหรับทดสอบค่าตัวแปรประสิทธิภาพของการค้นคืนเอกสารแต่ละกลุ่มมีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1

H_0: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบไม่ปกติ

2) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 2

H_0: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบไม่ปกติ

3) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3

H_0: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบไม่ปกติ

ตัวสถิติทดสอบคือ Kolmogorov-Sminov เมื่อขนาดตัวอย่างมากกว่า 50 หน่วยและของ Shapiro-Wilk เมื่อขนาดตัวอย่างอย่างน้อยกว่า 50 หน่วย (กัลยา วานิชย์บัญชา,2548) เนื่องจากในงานวิจัยนี้ตัวอย่างในแต่ละกลุ่มมีขนาดมากกว่า 50 จึงใช้วิธีตรวจสอบการแจกแจงโดยใช้

เทคนิคของ Kolmogorov-Sminov โดยจะยอมรับสมมติฐาน H_0 ถ้ามีค่า Sig.มีค่ามากกว่าค่านัยสำคัญ α ซึ่งกำหนดให้เท่ากับ 0.05 ดังตารางต่อไปนี้

ตารางที่ 4.3 ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง

	รูปแบบการค้นคืนเอกสาร	Kolmogorov-Sminov		
		Statistic	df	Sig.
ประสิทธิภาพของการค้นคืนเอกสาร	1	0.139	83	0.000
	2	0.122	83	0.004
	3	0.145	83	0.000

ผลการทดสอบในตารางที่ 4.3 ข้างต้นพบว่าค่าสถิติค่า Sig.ของตัวแปรของการค้นคืนเอกสารทั้ง 3 รูปแบบเป็นดังนี้

1) การค้นคืนเอกสารรูปแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

2) การค้นคืนเอกสารรูปแบบที่ 2 มีค่า Sig. เท่ากับ 0.004 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

3) การค้นคืนเอกสารรูปแบบที่ 3 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นไม่เป็นแบบปกติ

4.3.2 การวิเคราะห์ความแตกต่างประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบ

จากการวิเคราะห์การแจกแจงของข้อมูลในหัวข้อ 4.3.1 ได้ผลออกมาว่าการแจกแจงของตัวแปรประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นไม่เป็นแบบปกติ ดังนั้นทำให้ต้องใช้วิธีวิเคราะห์ความแปรปรวนแบบไม่อิงพารามิเตอร์ (Nonparametric Test) โดยในงานวิจัยนี้ผู้วิจัยเลือกใช้เทคนิคการวิเคราะห์ความแปรปรวนแบบไม่อิงพารามิเตอร์วิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) เนื่องจากงานวิจัยนี้วิเคราะห์เปรียบเทียบลักษณะของข้อมูลที่มากกว่า 2 กลุ่มว่ามีกลุ่มใดแตกต่างกันหรือไม่ ซึ่งจะกำจัดอิทธิพลอื่น ๆ ออกไป นั่นคือ

ข้อสอบถามแต่ละข้อสอบถาม เพื่อที่จะวัดประสิทธิภาพของการค้นคืนเอกสารเพียงอย่างเดียว โดยจะตั้งสมมติฐานดังต่อไปนี้

วิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 3

H_0 : $\mu_1 = \mu_2 = \mu_3$

H_1 : ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารอย่างน้อย 1 คู่มีค่าไม่เท่ากัน

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) จะได้ผลการวิเคราะห์ดังตารางที่ 4.4

ตารางที่ 4.4 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) ของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง

	ประสิทธิภาพของการค้นคืนเอกสาร
N	83
Chi-Square	11.322
df	2
ค่า Asymp.Sig. (2-tailed)	0.003

จากตาราง 4.4 ค่าสถิติทดสอบมีการแจกแจงแบบไคว์สแควร์และมีค่าเท่ากับ 11.322 ที่องศาความเป็นอิสระ (degree of freedom: df) เท่ากับ 2 โดยมีค่า Sig. เท่ากับ 0.003 ซึ่งมีค่า

น้อยกว่าค่า α =0.05 ดังนั้น จึงปฏิเสธสมมติฐาน H_0 นั่นหมายความว่าการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นมีอย่างน้อย 1 คู่แตกต่างกัน

จากผลการวิเคราะห์ความแตกต่างของประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบ ว่ามีความแตกต่างกันอย่างน้อย 2 รูปแบบการค้นคืนเอกสาร ดังนั้นขั้นตอนต่อไปจะต้องตรวจสอบว่าการค้นคืนเอกสารคู่ใดมีประสิทธิภาพแตกต่างกัน โดยผู้วิจัยกำหนดใช้เทคนิคเครื่องหมายลำดับที่ของวิลคอกซัน สำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) เนื่องจากในส่วนนี้ผู้วิจัยต้องการวิเคราะห์เปรียบเทียบค่าข้อมูล 2 กลุ่มที่มีความสัมพันธ์กันว่าแตกต่างกันหรือไม่ ซึ่งจะกำจัดอิทธิพลอื่น ๆ ออกไป และวิธีนี้จะพิจารณาเครื่องหมายและปริมาณผลต่างของค่าที่ต้องการทดสอบว่ามากน้อยเพียงใด (กัลยา วานิชย์บัญชา, 2548) ดังนั้นจะวิเคราะห์เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารทั้ง 3 รูปแบบทีละคู่ ดังต่อไปนี้

- **เปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_2 \leq \mu_1$

H_1: $\mu_2 > \mu_1$

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์ดังตารางที่ 4.5

ตารางที่ 4.5 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ระหว่างการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2

	ประสิทธิภาพของ การค้นคืนเอกสารรูปแบบที่ 2 – รูปแบบที่1[a]
Z	-2.667
ค่า Asymp.Sig. (2-tailed)	0.008

a Based on negative ranks

จากตาราง 4.5 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.667 ซึ่งน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.008 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.004 ซึ่งน้อยกว่าค่า α =0.05 และผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 2 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือยืนยันได้ว่าค่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มากกว่ารูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

- **เปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_1$

H_1: $\mu_3 > \mu_1$

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์ดังตารางที่ 4.6

ตารางที่ 4.6 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ระหว่างการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 3

	ประสิทธิภาพของ การค้นคืนเอกสารรูปแบบที่ 3 – รูปแบบที่ 1[b]
Z	-0.449
ค่า Asymp.Sig. (2-tailed)	0.653

b Base on positive ranks

จากตาราง 4.6 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -0.449 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.653 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.327 ซึ่งมีค่ามากกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้ง

บนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 1 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

- **เปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และการใช้กฎความสัมพันธ์กับการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_2$

H_1: $\mu_3 > \mu_2$

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์ดังตารางที่ 4.7

ตารางที่ 4.7 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ ระหว่างการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 3

	ประสิทธิภาพของ การค้นคืนเอกสารรูปแบบที่ 3 - รูปแบบที่ 2[b]
Z	-1.806
ค่า Asymp.Sig. (2-tailed)	0.071

b Base on positive rank

จากตาราง 4.7 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -1.806 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.071 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.036 ซึ่งมีค่าน้อยกว่าค่า α =0.05 แต่เนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 2 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 น้อยกว่าหรือเท่ากับรูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองจะแสดงว่าค่าประสิทธิภาพของการค้นคืนเอกสาร

รูปแบบที่ 3 น้อยกว่ารูปแบบที่ 2 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานจะแสดงว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 น้อยกว่ารูปแบบที่ 2

4.4 สรุปผลการวิเคราะห์ข้อมูล

จากการวิเคราะห์ผลการทดลองการวัดประสิทธิภาพการค้นคืนเอกสาร โดยใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ผู้วิจัยสามารถสรุปได้ว่าการค้นคืนเอกสารรูปแบบที่ 1 มีประสิทธิภาพน้อยที่สุด รองลงมาคือการค้นคืนเอกสารรูปแบบที่ 3 และประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มีประสิทธิภาพมากที่สุด

จากผลสรุปดังกล่าวแสดงว่า การค้นคืนเอกสารที่ใช้เทคนิคการใช้กฎความสัมพันธ์เข้าร่วมนั้นสามารถเพิ่มประสิทธิภาพให้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ได้ แต่เทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ร่วมกับการใช้กฎความสัมพันธ์นั้นไม่สามารถเพิ่มประสิทธิภาพให้กับการค้นคืนเอกสารได้

4.5 ผลการศึกษาเพิ่มเติม

จากการทดลองการค้นคืนเอกสารเบื้องต้นทำให้ผู้วิจัยมีความต้องการวิเคราะห์ผลการทดลองเพิ่มเติมว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้นั้นมีประสิทธิภาพการค้นคืนเอกสารมากกว่าการไม่ใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้หรือไม่ โดยเรียกการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ว่าการค้นคืนเอกสารรูปแบบที่ 4 และศึกษาเพิ่มเติมเรื่องหากไม่ใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) เป็นค่าที่แสดงถึงประสิทธิภาพการค้นคืนเอกสารในภาพรวมแล้ว แต่ใช้ค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ผลการทดลองที่ได้จะเป็นเช่นเดียวกันกับการใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) หรือไม่

4.5.1 การวัดประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

จากผลการทดลองการสรุปผลการทดลอง ผู้วิจัยต้องการทดสอบเพิ่มเติมในส่วนการเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารทั้ง 3 รูปแบบในงานวิจัยข้างต้น

1) ผลการทดลอง

จากการทดลองการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ ที่ทดสอบกับเอกสารไทม์ 425 เอกสารและข้อสอบถาม 83 ข้อสอบถาม ในฐานข้อมูลนิตยสารไทม์ (Time Collection) แสดงได้ดังตารางที่ 4.8

ตารางที่ 4.8 ตารางแสดงผลการทดลองของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4		**ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4**		**ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4**
1	0.3636	18	0.2857	35	0.1818
2	0.1081	19	0.4167	36	0.5000
3	0.1071	20	0.1538	37	0.2500
4	0.2222	21	0.1290	38	0.0800
5	0.2326	22	0.1905	39	0.4737
6	0.3913	23	0.2000	40	0.3830
7	0.2000	24	0.0952	41	0.2500
8	0.2353	25	0.1000	42	0.0500
9	0.6364	26	0.2500	43	0.1429
10	0.5217	27	0.1818	44	0.2222
11	0.2222	28	0.2500	45	0.2857
12	0.5385	29	0.2105	46	0.5484
13	0.3158	30	0.2857	47	0.2941
14	0.2857	31	0.4516	48	0.1053
15	0.5714	32	0.0333	49	0.5000
16	0.2609	33	1.0000	50	0.0800
17	0.2353	34	0.2857	51	0.1395

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4
52	0.0455
53	0.1379
54	0.0702
55	0.4151
56	0.0769
57	0.2222
58	0.5000
59	0.1538
60	0.1818
61	0.5532
62	0.2500

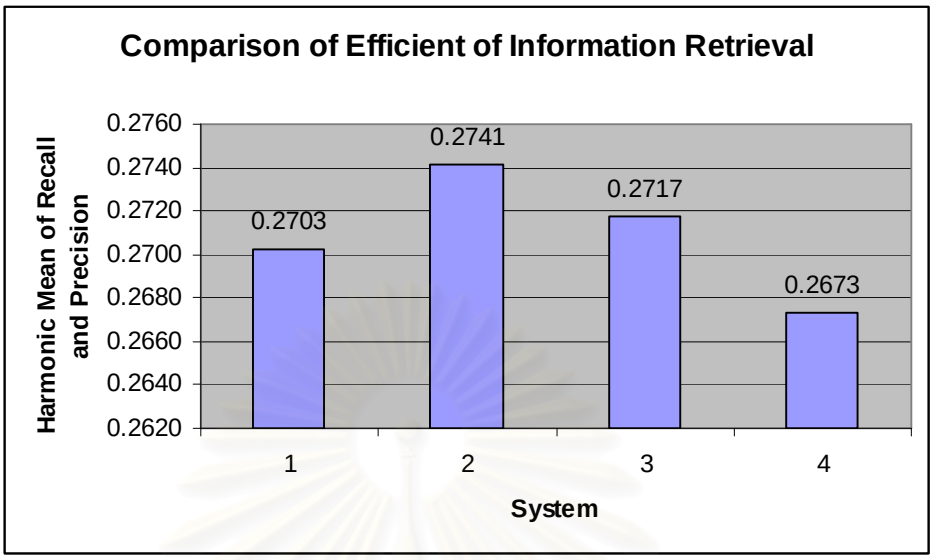
	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4
63	0.5000
64	0.2000
65	0.0870
66	0.1290
67	0.3750
68	0.3243
69	0.7647
70	0.1250
71	0.5714
72	0.0571
73	0.1538

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4
74	0.2000
75	0.0833
76	0.3030
77	0.0800
78	0.1538
79	0.1000
80	0.4324
81	0.0667
82	0.3030
83	0.1176

จากผลการทดลองค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ตารางที่ 4.8 และสรุปการวิเคราะห์ผลทดลองของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ของการค้นคืนเอกสารรูปแบบที่ 4 ได้ดังตารางที่ 4.9 และแสดงค่าเฉลี่ยของประสิทธิภาพการค้นคืนเอกสารทั้ง 4 รูปแบบดังรูป 4.5

ตารางที่ 4.9 ตารางสรุปผลการทดลองของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องการค้นคืนเอกสารรูปแบบที่ 4
ค่าเฉลี่ย (Mean)	0.2717
ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)	0.1861



รูปที่ 4.5 รูปแสดงกราฟเปรียบเทียบค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องระหว่างการค้นคืนเอกสารทั้ง 4 รูปแบบ

2) **การวิเคราะห์ข้อมูล**

- **การวิเคราะห์การแจกแจงข้อมูล**

ทดสอบการแจกแจงของข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4

H_0: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบไม่ปกติ

ตัวสถิติทดสอบคือ Kolmogorov-Sminov เมื่อขนาดตัวอย่างมากกว่า 50 หน่วยและของ Shapiro-Wilk เมื่อขนาดตัวอย่างอย่างน้อยกว่า 50 หน่วย (กัลยา วานิชย์บัญชา,2548) เนื่องจากในงานวิจัยนี้ตัวอย่างในแต่ละกลุ่มมีขนาดมากกว่า 50 จึงใช้วิธีตรวจสอบการแจกแจงโดยใช้เทคนิคของ Kolmogorov-Sminov โดยจะยอมรับสมมติฐาน H_0 ถ้ามีค่า Sig.มีค่ามากกว่าค่านัยสำคัญ α ซึ่งกำหนดให้เท่ากับ 0.05 ดังตาราง 4.10 ต่อไปนี้

ตารางที่ 4.10 ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง

	Kolmogorov-Sminov		
	Statistic	df	Sig.
ประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4	0.140	83	0.000

ผลการทดสอบในตารางที่ ข้างต้นพบว่าค่าสถิติค่า Sig.เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0 ดังนั้นสรุปได้ว่าการแจกแจงของตัวแปรประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 นั้นไม่เป็นแบบปกติ ทำให้การวิเคราะห์ข้อมูลของการค้นคืนเอกสารแต่ละรูปแบบต้องใช้วิธีวิเคราะห์ความแปรปรวนแบบไม่อิงพารามิเตอร์ (Nonparametric Test) โดยกำหนดให้

μ_1 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 3

μ_4 คือ ค่าเฉลี่ยประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 4

เนื่องจากได้มีการวิเคราะห์เปรียบเทียบลักษณะของข้อมูลที่มากกว่า 2 กลุ่มว่ามีกลุ่มใดแตกต่างกันหรือไม่แล้วด้วยเทคนิคการวิเคราะห์ความแปรปรวนแบบไม่อิงพารามิเตอร์วิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) ของการค้นคืนเอกสาร 3 รูปแบบ ซึ่งผลแสดงว่าค่าประสิทธิภาพของระบบค้นคืนเอกสารมีอย่างน้อย 1 คู่แตกต่างกัน ดังนั้นแม้ว่าจะนำการค้นคืนเอกสารรูปแบบที่ 4 ไปวิเคราะห์ร่วมด้วยผลสรุปการทดสอบสมมติฐานก็ไม่แตกต่างจากข้างต้น ดังนั้นขั้นตอนจึงวิเคราะห์เปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 4 กับอีก 3 รูปแบบทีละคู่ได้ดังนี้

- **การวิเคราะห์การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_1$

H_1: $\mu_4 > \mu_1$

- **การวิเคราะห์การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำกับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_2$

H_1: $\mu_4 > \mu_2$

- **การวิเคราะห์การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำและการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_4$

H_1: $\mu_3 > \mu_4$

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์เปรียบเทียบทั้ง 3 คู่ดังกล่าวข้างต้นแสดงดังตารางที่ 4.11

ตารางที่ 4.11 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่ ระหว่างการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 4

	ประสิทธิภาพการค้นคืนเอกสาร		
	รูปแบบที่ 4 – รูปแบบที่ 1[b]	รูปแบบที่ 4 – รูปแบบที่ 2[b]	รูปแบบที่ 4 – รูปแบบที่ 3[b]
	กรณีที่1	**กรณีที่2**	**กรณีที่3**
Z	-2.101	-3.048	-2.934
ค่า Asymp.Sig. (2-tailed)	0.036	0.002	0.003

b Base on positive ranks

จากตาราง 4.11 ผลการวิเคราะห์ผลการทดลองแสดงดังนี้

กรณีที่1 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.101 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.036 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.018 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 1 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่าหรือเท่ากับรูปแบบที่ 4 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองแสดงให้เห็นว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 4 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานสามารถสรุปได้ว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 4

กรณีที่ 2 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -3.048 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.002 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.001 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 2 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มากกว่าหรือเท่ากับรูปแบบที่ 4 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองแสดงให้เห็นว่าประสิทธิภาพการค้นคืนเอกสาร

รูปแบบที่ 2 มากกว่ารูปแบบที่ 4 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานสามารถสรุปได้ว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มากกว่ารูปแบบที่ 4

กรณีที่ 3 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.934 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.003 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.015 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 3 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 4 ที่ระดับนัยสำคัญ 0.05

3) สรุปผลการศึกษาเพิ่มเติมของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

จากการวิเคราะห์ข้อมูลเพิ่มเติมในส่วนการวัดประสิทธิภาพของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ เมื่อทดสอบและวิเคราะห์ข้อมูลผลการทดลองแล้วสรุปได้ว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 มีประสิทธิภาพน้อยกว่าการค้นคืนเอกสารรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 นั่นคือ การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้มีประสิทธิภาพด้อยกว่าการค้นคืนเอกสารรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 ดังนี้ การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้ การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ และการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

4.5.2 การวัดประสิทธิภาพของการค้นคืนเอกสารด้วยค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision)

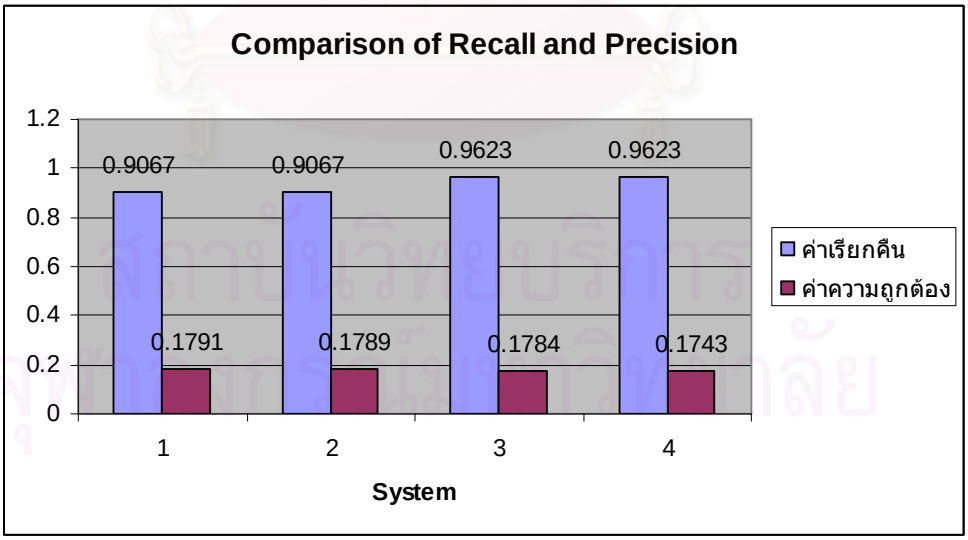
การพิจารณาประสิทธิภาพการค้นคืนเอกสาร นอกจากจะพิจารณาจากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) แล้ว ยังสามารถพิจารณาจากค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ดังนั้นผู้วิจัยจึงใช้ค่าทั้ง 2 นี้มาทดสอบตามสมมติฐานที่ตั้งไว้ในการเปรียบเทียบประสิทธิภาพการทดลองการค้นคืนเอกสารทั้ง 4 รูปแบบเพิ่มเติมจากค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) โดยผลการทดลองค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารทั้ง 4 รูปแบบแสดงในภาคผนวก ฌ

1) **ผลการทดลอง**

ตารางสรุปผลการทดลองค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารทั้ง 4 รูปแบบแสดงดังตารางที่ 4.12 และเปรียบเทียบค่าเฉลี่ยค่าประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบดังรูป 4.6

ตารางที่ 4.12 ตารางสรุปผลการทดลองของค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision)

		การค้นคืนเอกสาร			
		รูปแบบที่ 1	**รูปแบบที่ 2**	**รูปแบบที่ 3**	**รูปแบบที่ 4**
ค่าเรียกคืน	ค่าเฉลี่ย (Mean)	0.9067	0.9067	0.9623	0.9623
	ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)	0.1675	0.2198	0.1110	0.1110
	ค่ามัธยฐาน (Median)	0.1333	1.0000	1.0000	1.0000
ค่าความถูกต้อง	ค่าเฉลี่ย (Mean)	0.1791	0.1789	0.1784	0.1743
	ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation)	0.1649	0.1675	0.1633	0.1577
	ค่ามัธยฐาน (Median)	0.1250	0.1333	0.1333	0.1250



รูปที่ 4.6 รูปแสดงกราฟเปรียบเทียบค่าเรียกคืนและค่าความถูกต้องระหว่างการค้นคืนเอกสารทั้ง 4 รูปแบบ

เนื่องจากค่าเฉลี่ยประสิทธิภาพของค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ทั้ง 4 รูปแบบมีค่าแตกต่างกันดังรูปที่ 4.4 แต่ไม่สามารถสรุปได้ว่าประสิทธิภาพแตกต่างกันอย่างมี

นัยสำคัญ จึงวิเคราะห์ผลการทดลองต่อไปว่าประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบแตกต่างกันอย่างมีนัยสำคัญหรือไม่ โดยจะแบ่งการวิเคราะห์ออกเป็นค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ดังต่อไปนี้

2) การวิเคราะห์ข้อมูลประสิทธิภาพค่าเรียกคืน (Recall)

- **การวิเคราะห์การแจกแจงข้อมูล**

ผู้วิจัยจะตรวจสอบว่าค่าประสิทธิภาพค่าเรียกคืน (Recall) ของการค้นคืนเอกสารมีการแจกแจงแบบปกติหรือไม่จากค่าเรียกคืน (Recall) 83 ค่าที่ได้ในการค้นคืนเอกสารแต่ละรูปแบบ โดยตั้งสมมติฐานของการทดสอบสำหรับทดสอบค่าตัวแปรประสิทธิภาพค่าเรียกคืน (Recall) ของการค้นคืนเอกสารแต่ละกลุ่มมีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 1

H_0: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบไม่ปกติ

2) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 2

H_0: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบไม่ปกติ

3) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 3

H_0: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบไม่ปกติ

4) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 4

H_0: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบไม่ปกติ

ตัวสถิติทดสอบคือ Kolmogorov-Sminov เมื่อขนาดตัวอย่างมากกว่า 50 หน่วยและของ Shapiro-Wilk เมื่อขนาดตัวอย่างอย่างน้อยกว่า 50 หน่วย (กัลยา วานิชย์บัญชา, 2548) เนื่องจากในงานวิจัยนี้ตัวอย่างในแต่ละกลุ่มมีขนาดมากกว่า 50 จึงใช้วิธีตรวจสอบการแจกแจงโดยใช้เทคนิคของ Kolmogorov-Sminov โดยจะยอมรับสมมติฐาน H_0 ถ้ามีค่า Sig.มีค่ามากกว่าค่านัยสำคัญ α ซึ่งกำหนดให้เท่ากับ 0.05 ดังตารางต่อไปนี้

ตารางที่ 4.13 ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของประสิทธิภาพค่าเรียกคืน

	รูปแบบการค้นคืนเอกสาร	**Kolmogorov-Sminov**		
		Statistic	**df**	**Sig.**
ประสิทธิภาพค่าเรียกคืน (Recall)	1	0.411	83	0.000
	2	0.411	83	0.000
	3	0.488	83	0.000
	4	0.488	83	0.000

ผลการทดสอบในตารางที่ 4.13 ข้างต้นพบว่าค่าสถิติค่า Sig.ของตัวแปรของการค้นคืนเอกสารทั้ง 4 รูปแบบเป็นดังนี้

1) การค้นคืนเอกสารรูปแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

2) การค้นคืนเอกสารรูปแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

3) การค้นคืนเอกสารรูปแบบที่ 3 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

4) การค้นคืนเอกสารรูปแบบที่ 4 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของค่าเรียกคืน (Recall) ของการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นไม่เป็นแบบปกติ

- **การวิเคราะห์ความแตกต่างประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบ**

วิเคราะห์เปรียบเทียบค่าเรียกคืน (Recall) ของการค้นคืนเอกสารทั้ง 3 รูปแบบว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1 คือ ค่าเฉลี่ยค่าเรียกคืนของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยค่าเรียกคืนของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยค่าเรียกคืนของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 3

μ_4 คือ ค่าเฉลี่ยค่าเรียกคืนของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 4

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : ค่าเฉลี่ยค่าเรียกคืนของการค้นคืนเอกสารอย่างน้อย 1 คู่มีค่าไม่เท่ากัน

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) จะได้ผลการวิเคราะห์ดังตารางที่ 4.14

ตารางที่ 4.14 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) ของค่าเรียกคืนในการวิเคราะห์ความแตกต่างประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบ

	ประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสาร
N	83
Chi-Square	30.000
df	3
ค่า Asymp.Sig. (2-tailed)	0.000

จากตาราง 4.14 ค่าสถิติทดสอบมีการแจกแจงแบบไคว์สแควร์และมีค่าเท่ากับ 30.000 ที่องศาความเป็นอิสระ (degree of freedom: df) เท่ากับ 3 โดยมีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α =0.05 ดังนั้น จึงปฏิเสธสมมติฐาน H_0 นั่นหมายความว่าการค้นคืนเอกสารทั้ง 4 รูปแบบนั้นมีอย่างน้อย 1 คู่แตกต่างกัน

จากผลการวิเคราะห์ความแตกต่างของประสิทธิภาพค่าเรียกคืน (Recall) ของการค้นคืนเอกสารทั้ง 4 รูปแบบว่ามีความแตกต่างกันอย่างน้อย 2 รูปแบบการค้นคืนเอกสาร ดังนั้นขั้นตอนต่อไปจะต้องตรวจสอบว่าการค้นคืนเอกสารคู่ใดมีประสิทธิภาพค่าเรียกคืน (Recall) แตกต่างกัน

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_2 \leq \mu_1$

H_1: $\mu_2 > \mu_1$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และการใช้กฎความสัมพันธ์กับการค้นคืนเอกสารโดยใช้

เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_2$

H_1: $\mu_3 > \mu_2$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำและการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_4$

H_1: $\mu_3 > \mu_4$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_1$

H_1: $\mu_3 > \mu_1$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_1$

H_1: $\mu_4 > \mu_1$

o การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 4

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าเรียกคืน (Recall) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำกับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_2$

H_1: $\mu_4 > \mu_2$

จากสมมติฐานข้างต้นและผลการทดลองค่าเรียกคืน (Recall) ที่ได้ เมื่อวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซัน สำหรับการทดสอบแบบจับคู่ (The Wilcoxon Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์ข้อมูลโดยเปรียบเทียบรูปแบบการค้นคืนเอกสารทีละคู่ดังตารางที่ 4.15

ตารางที่ 4.15 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าเรียกคืนด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซันสำหรับการทดสอบแบบจับคู่การค้นคืนเอกสารแต่ละรูปแบบ

ประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสาร	กรณีที่	Z	ค่า Asymp.Sig. (2-tailed)
รูปแบบที่ 2 – รูปแบบที่1[c]	1	0.000	1.000
รูปแบบที่ 3 – รูปแบบที่2[a]	2	-2.814	0.005
รูปแบบที่ 4 – รูปแบบที่3[c]	3	0.000	1.000
รูปแบบที่ 3 – รูปแบบที1[a]	4	-2.814	0.005
รูปแบบที่ 1 – รูปแบบที4[a]	5	-2.814	0.005
รูปแบบที่ 2 – รูปแบบที4[a]	6	-2.814	0.005

a Base on negative ranks

c The sum of negative ranks equal the sum of positive ranks

จากตาราง 4.15 วิเคราะห์ผลการทดลองได้ดังนี้

กรณีที่1 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ 0.000 ซึ่งเท่ากับศูนย์และมีค่า Sig. เท่ากับ 1.000 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.5 ซึ่งมากกว่าค่า α =0.05 ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 และผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานเท่ากัน (The sum of negative ranks equal the sum of positive ranks) นั่นคือยืนยันได้ว่าค่าประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 2 เท่ากับรูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 2 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -1.806 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.005 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.003 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 2 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 3 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ 0.000 ซึ่งมีค่าเท่ากับศูนย์และมีค่า Sig. เท่ากับ 1.000 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.5 ซึ่งมีค่ามากกว่าค่า α =0.05 ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานเท่ากัน (The sum of negative ranks equal the sum of positive ranks) นั่นคือประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 3 เท่ากับรูปแบบที่ 4 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 4 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.814 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.005 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.003 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 3 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 5 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.814 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.005 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสอง

ได้ค่าเท่ากับ 0.003 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 4 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 6 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.814 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.005 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสอง ได้ค่าเท่ากับ 0.003 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มากกว่ารูปแบบที่ 4 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05

3) สรุปการวิเคราะห์ข้อมูลประสิทธิภาพของค่าเรียกคืน (Recall)

ประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 1 เท่ากับรูปแบบที่ 2 และประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 3 มีค่าเท่ากับรูปแบบที่ 4 และประสิทธิภาพค่าเรียกคืน (Recall) การค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบ 2 น้อยกว่าการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบ 4 นั่นคือ ประสิทธิภาพค่าเรียกคืน (Recall) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้มีค่าเท่ากันกับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ ซึ่งทั้ง 2 รูปแบบนี้มีประสิทธิภาพค่าเรียกคืน (Recall) น้อยกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ที่มีค่าประสิทธิภาพเรียกคืน (Recall) เท่ากับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

4) การวิเคราะห์ข้อมูลประสิทธิภาพของค่าความถูกต้อง (Precision)

- **การวิเคราะห์การแจกแจงข้อมูล**

ผู้วิจัยจะตรวจสอบว่าค่าประสิทธิภาพค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารมีการแจกแจงแบบปกติหรือไม่จากค่าความถูกต้อง (Precision) 83 ค่าที่ได้ในการค้นคืนเอกสารแต่ละรูปแบบ โดยตั้งสมมติฐานของการทดสอบสำหรับทดสอบค่าตัวแปรประสิทธิภาพค่าความ

ถูกต้อง (Precision) ของการค้นคืนเอกสารแต่ละกลุ่มมีการแจกแจงแบบปกติหรือไม่ภายใต้สมมติฐานทางสถิติ ดังนี้

1) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 1

H_0: ข้อมูลค่าประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบปกติ

H_1: ข้อมูลค่าประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 1 มีการแจกแจงแบบไม่ปกติ

2) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 2

H_0: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 2 มีการแจกแจงแบบไม่ปกติ

3) ทดสอบการแจกแจงของข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 3

H_0: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 3 มีการแจกแจงแบบไม่ปกติ

4) ทดสอบการแจกแจงของข้อมูลค่าประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4

H_0: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบปกติ

H_1: ข้อมูลประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารรูปแบบที่ 4 มีการแจกแจงแบบไม่ปกติ

ตัวสถิติทดสอบคือ Kolmogorov-Sminov เมื่อขนาดตัวอย่างมากกว่า 50 หน่วยและของ Shapiro-Wilk เมื่อขนาดตัวอย่างอย่างน้อยกว่า 50 หน่วย (กัลยา วานิชย์บัญชา,2548) เนื่องจากในงานวิจัยนี้ตัวอย่างในแต่ละกลุ่มมีขนาดมากกว่า 50 จึงใช้วิธีตรวจสอบการแจกแจงโดยใช้เทคนิคของ Kolmogorov-Sminov โดยจะยอมรับสมมติฐาน H_0 ถ้ามีค่า Sig.มีค่ามากกว่าค่านัยสำคัญ α ซึ่งกำหนดให้เท่ากับ 0.05 ดังตารางต่อไปนี้

ตารางที่ 4.16 ตารางแสดงค่าสถิติทดสอบการแจกแจงปกติ (Normality Test) ของประสิทธิภาพค่าความถูกต้อง

	รูปแบบการค้นคืนเอกสาร	Kolmogorov-Sminov		
		Statistic	df	Sig.
ประสิทธิภาพค่าความถูกต้อง (Precision)	1	0.411	83	0.000
	2	0.411	83	0.000
	3	0.488	83	0.000
	4	0.488	83	0.000

ผลการทดสอบในตารางที่ 4.16 ข้างต้นพบว่าค่าสถิติค่า Sig.ของตัวแปรของการค้นคืนเอกสารทั้ง 4 รูปแบบเป็นดังนี้

1) การค้นคืนเอกสารรูปแบบที่ 1 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

2) การค้นคืนเอกสารรูปแบบที่ 2 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

3) การค้นคืนเอกสารรูปแบบที่ 3 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

4) การค้นคืนเอกสารรูปแบบที่ 4 มีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α = 0.05 ดังนั้นจึงปฏิเสธสมมติฐาน H_0

ดังนั้นสรุปได้ว่าการแจกแจงของค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นไม่เป็นแบบปกติ

- **การวิเคราะห์ความแตกต่างประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบ**

วิเคราะห์เปรียบเทียบค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารทั้ง 3 รูปแบบว่ามีความแตกต่างกันหรือไม่

กำหนดให้ μ_1 คือ ค่าเฉลี่ยค่าความถูกต้องของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยค่าความถูกต้องของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยค่าความถูกต้องของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 3

μ_4 คือ ค่าเฉลี่ยค่าความถูกต้องของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือการค้นคืนเอกสารรูปแบบที่ 4

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : ค่าเฉลี่ยค่าความถูกต้องของการค้นคืนเอกสารอย่างน้อย 1 คู่มีค่าไม่เท่ากัน

จากผลการทดลองที่ได้เมื่อนำมาวิเคราะห์ข้อมูลด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) จะได้ผลการวิเคราะห์ดังตารางที่ 4.17

ตารางที่ 4.17 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพด้วยวิธีฟรีดแมน (The Friedman F_r Test for a Randomized Block Design) ของค่าความถูกต้องในการวิเคราะห์ความแตกต่างประสิทธิภาพของการค้นคืนเอกสารทั้ง 4 รูปแบบ

	ประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสาร
N	83
Chi-Square	19.598
df	3
ค่า Asymp.Sig. (2-tailed)	0.000

จากตาราง 4.17 ค่าสถิติทดสอบมีการแจกแจงแบบไคว์สแควร์และมีค่าเท่ากับ 19.598 ที่องศาความเป็นอิสระ (degree of freedom: df) เท่ากับ 3 โดยมีค่า Sig. เท่ากับ 0.000 ซึ่งมีค่าน้อยกว่าค่า α =0.05 ดังนั้น จึงปฏิเสธสมมติฐาน H_0 นั่นหมายความว่าการค้นคืนเอกสารทั้ง 4 รูปแบบนั้นมีอย่างน้อย 1 คู่แตกต่างกัน

จากผลการวิเคราะห์ความแตกต่างของประสิทธิภาพค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารทั้ง 4 รูปแบบว่ามีความแตกต่างกันอย่างน้อย 2 รูปแบบการค้นคืนเอกสาร ดังนั้นขั้นตอนต่อไปจะต้องตรวจสอบว่าการค้นคืนเอกสารคู่ใดมีประสิทธิภาพค่าความถูกต้อง (Precision) แตกต่างกัน

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 2**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_2 \leq \mu_1$

H_1: $\mu_2 > \mu_1$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และการใช้กฎความสัมพันธ์กับการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_2$

H_1: $\mu_3 > \mu_2$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 3 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำและการใช้เทคนิคผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_4$

H_1: $\mu_3 > \mu_4$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 3**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_3 \leq \mu_1$

H_1: $\mu_3 > \mu_1$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 1 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_1$

H_1: $\mu_4 > \mu_1$

- **การเปรียบเทียบการค้นคืนเอกสารรูปแบบที่ 2 และรูปแบบที่ 4**

การเปรียบเทียบนี้เป็นการเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Precision) ระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้กฎความสัมพันธ์ของคำกับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ สามารถตั้งสมมติฐานได้ดังนี้

H_0: $\mu_4 \leq \mu_2$

H_1: $\mu_4 > \mu_2$

จากสมมติฐานข้างต้นและผลการทดลองค่าความถูกต้อง (Precision) ที่ได้ เมื่อวิเคราะห์ข้อมูลด้วยวิธีเครื่องหมายลำดับที่ของวิลคอกซัน สำหรับการทดสอบแบบจับคู่ (The Wilcoxon

Signed Rank Sum Test for the Matched Paired Difference) แล้ว จะได้ผลการวิเคราะห์ข้อมูลโดยเปรียบเทียบรูปแบบการค้นคืนเอกสารทีละคู่ดังตารางที่ 4.18

ตารางที่ 4.18 ตารางแสดงค่าสถิติทดสอบประสิทธิภาพค่าความถูกต้องด้วยวิธีเครื่องหมายลำดับที่ของ วิลคอกซันสำหรับการทดสอบแบบจับคู่การค้นคืนเอกสารแต่ละคู่

ประสิทธิภาพค่าเรียกคืนของการค้นคืนเอกสาร	กรณีที่	Z	ค่า Asymp.Sig. (2-tailed)
รูปแบบที่ 2 – รูปแบบที่1[a]	1	-0.279	0.780
รูปแบบที่ 3 – รูปแบบที่2[a]	2	-0.821	0.412
รูปแบบที่ 4 – รูปแบบที่3[b]	3	-2.934	0.003
รูปแบบที่ 3 – รูปแบบที่1[a]	4	-0.906	0.365
รูปแบบที่ 4 – รูปแบบที่1[b]	5	-2.518	0.012
รูปแบบที่ 4 – รูปแบบที่2[a]	6	-2.083	0.037

a Base on negative ranks

b Base on positive ranks

c The sum of negative ranks equal the sum of positive ranks

กรณีที่ 1 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -0.279 ซึ่งเท่ากับศูนย์และมีค่า Sig. เท่ากับ 0.780 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.390 ซึ่งมากกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางลบ (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 2 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 1 มากกว่าหรือเท่ากับรูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองจะแสดงว่าค่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 2 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานจะแสดงว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 1 มากกว่ารูปแบบที่ 2

กรณีที่ 2 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -0.821 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.421 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสอง

ได้ค่าเท่ากับ 0.211 ซึ่งมีค่ามากกว่าค่า α =0.05 และผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on negative ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 2 มากกว่ารูปแบบที่ 3 (ตัวลบมากกว่าตัวตั้ง) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าความถูกต้อง (Precision) การค้นคืนเอกสารรูปแบบที่ 3 น้อยกว่าหรือเท่ากับรูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองจะแสดงว่าค่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 มีค่าน้อยกว่าจนเกือบเท่ากับรูปแบบที่ 2 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานจะแสดงว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 เท่ากับรูปแบบที่ 2

กรณีที่ 3 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.934 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.003 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.002 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 3 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าความถูกต้อง (Precision) การค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 4 ที่ระดับนัยสำคัญ 0.05

กรณีที่ 4 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -0.906ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.365 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.183 ซึ่งมีค่ามากกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 3 มากกว่ารูปแบบที่ 1 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าความถูกต้อง (Precision) การค้นคืนเอกสารรูปแบบที่ 3 น้อยกว่าหรือเท่ากับรูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองจะแสดงว่าค่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 มีค่าน้อยกว่าจนเกือบเท่ากับรูปแบบที่ 1 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานจะแสดงว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 3 เท่ากับรูปแบบที่ 1

กรณีที่ 5 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.518 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.012 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.006 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้น

คืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 1 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าความถูกต้อง (Precision) การค้นคืนเอกสารรูปแบบที่ 4 น้อยกว่าหรือเท่ากับรูปแบบที่ 1 ที่ระดับนัยสำคัญ 0.05 และจากผลการทดลองจะแสดงว่าค่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 น้อยกว่ารูปแบบที่ 1 ดังนั้นจากผลการทดลองและการสรุปสมมติฐานจะแสดงว่าประสิทธิภาพของการค้นคืนเอกสารรูปแบบที่ 4 น้อยกว่ารูปแบบที่ 1

กรณีที่ 6 จากสถิติค่าทดสอบค่า Z มีค่าเท่ากับ -2.083 ซึ่งมีค่าน้อยกว่าศูนย์และมีค่า Sig. เท่ากับ 0.037 จากการตั้งสมมติฐานในงานวิจัยเป็นแบบทางเดียวจึงต้องนำค่า Sig.หารสองได้ค่าเท่ากับ 0.019 ซึ่งมีค่าน้อยกว่าค่า α =0.05 และเนื่องจากผลการวิเคราะห์ที่ออกมาตั้งบนพื้นฐานทางบวก (Based on positive ranks) นั่นคือวิเคราะห์ในเชิงที่เทียบว่าประสิทธิภาพการค้นคืนเอกสารรูปแบบที่ 4 มากกว่ารูปแบบที่ 2 (ตัวตั้งมากกว่าตัวลบ) ดังนั้นจึงไม่สามารถปฏิเสธสมมติฐาน H_0 นั่นคือประสิทธิภาพค่าความถูกต้อง (Precision) การค้นคืนเอกสารรูปแบบที่ 4 น้อยกว่าหรือเท่ากับรูปแบบที่ 2 ที่ระดับนัยสำคัญ 0.05

5) สรุปการวิเคราะห์ข้อมูลประสิทธิภาพของค่าความถูกต้อง (Precision)

ประสิทธิภาพค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารรูปแบบที่ 1 รูปแบบที่ 2 และรูปแบบที่ 3 มีค่าเท่ากันและมีค่ามากกว่ารูปแบบที่ 4 นั่นคือ ประสิทธิภาพค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้มีค่าเท่ากับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ ซึ่งมีค่าน้อยกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

บทที่ 5

สรุปผลการวิจัย

5.1 บทนำ

บทนี้จะแสดงการสรุปผลของงานวิจัยและปัญหาที่เกิดขึ้นในการทดลอง สุดท้ายเป็นข้อเสนอแนะของงานวิจัย เพื่อปรับเปลี่ยนรูปแบบของงานวิจัยหรือพัฒนาการทดลองให้มีประสิทธิภาพยิ่งขึ้น

5.2 การทดลองและลักษณะของข้อมูลที่ใช้ทดสอบการค้นคืนเอกสาร

งานวิจัยนี้เป็นการวิจัยเชิงทดลอง (Experimental Research) โดยใช้เอกสารและข้อสอบถามของฐานข้อมูลนิตยสารไทม์ (TIME Collection) ปี 1963 มาทดสอบเทคนิคการค้นคืนเอกสารที่พัฒนาขึ้น ซึ่งมีเอกสารจำนวน 425 เอกสารและข้อสอบถามจำนวน 83 ข้อสอบถาม (สามารถดูตัวอย่างได้ในภาคผนวก ง) โดยเป็นเอกสารที่เก็บรวบรวมโดยมหาวิทยาลัยคอแนล (Cornell University) เพื่อนำมาทดลองกับระบบค้นคืนเอกสารสมาร์ท (SMART Information Retrieval System) ที่พัฒนาด้วยเทคนิคแบบจำลองปริภูมิเวกเตอร์ที่ใช้ในงานวิจัยระบบค้นคืนเอกสาร (Williamson and Lesk, 1971)

5.3 สรุปผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อทดสอบประสิทธิภาพของการค้นคืนเอกสารโดยการใช้เทคนิคปริภูมิเวกเตอร์นั้นเป็นเทคนิคที่เปลี่ยนเอกสารและข้อสอบถามให้อยู่ในรูปแบบเวกเตอร์ที่แต่ละตำแหน่งมิติเป็นคำที่อยู่ในระบบ โดยใช้ทฤษฎีของการค้นคืนสารสนเทศ (Information Retrieval) คือ การกำหนดดรรชนี (indexing) การให้น้ำหนักคำ (Term Weighting) และการคำนวณค่าความเหมือนระหว่างเอกสารและข้อสอบถาม และสามารถค้นคืนเอกสารได้โดยเปรียบเทียบความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม แล้วดึงเอกสารที่มีค่าความเหมือนกันมากกว่าค่าที่กำหนดออกมาแสดง นอกจากเทคนิคการค้นคืนเอกสารนี้ ผู้วิจัยสนใจที่จะเพิ่มประสิทธิภาพการค้นคืนเอกสารด้วยเทคนิคของการทำเหมืองข้อมูลคือการค้นหากฎความสัมพันธ์ (Association Rule Discovery) และเทคนิคของการค้นคืนสารสนเทศ (Information Retrieval) คือ การให้ผลสะท้อนกลับจากผู้ใช้ (Relevant Feedback) เข้ามาร่วมกับเทคนิคปริภูมิเวกเตอร์ด้วย โดยทดสอบว่าถ้าใช้เทคนิคการใช้กฎความสัมพันธ์และการใช้เทคนิคการให้ผลสะท้อนกลับ

จากผู้ใช้ หรือใช้เทคนิคการใช้กฎความสัมพันธ์อย่างเดียวจะสามารถเพิ่มประสิทธิภาพของการค้นคืนเอกสารได้หรือไม่ ผู้วิจัยจึงกำหนดให้ทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ
3) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

การทดลองการค้นคืนเอกสารผู้วิจัยกำหนดใช้ฐานข้อมูลนิตยสารไทม์ (Time Collection) กับเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบโดยประสิทธิภาพของการค้นคืนเอกสารในงานวิจัยนี้จะวัดด้วยค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic of recall and precision) เมื่อนำผลประสิทธิภาพของการค้นคืนเอกสารทั้ง 3 รูปแบบมาวิเคราะห์เปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร ผลการทดลองค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic of recall and precision) ที่ได้จากการค้นคืนทั้ง 3 รูปแบบได้ผลดังนี้

5.3.1 เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำและไม่ใช้เทคนิคการใช้เทคนิคกฎความสัมพันธ์ของคำ

การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำมีประสิทธิภาพมากกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เพียงอย่างเดียว โดยผลการทดลองผู้วิจัยพบว่าค่าเฉลี่ยทั้ง 83 ข้อสอบถามของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำมีประสิทธิภาพมากกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เท่ากับ 0.0028 สรุปได้ว่าการใช้เทคนิคกฎความสัมพันธ์ของคำร่วมด้วยสามารถค้นคืนเอกสารที่เกี่ยวเนื่องกับความต้องการออกมาได้มากกว่าไม่ใช้เทคนิคกฎความสัมพันธ์ของคำ และผู้วิจัยพบว่าข้อสอบถามที่ได้รับการขยายคำโดยกฎความสัมพันธ์ของคำมีเพียง 12 ข้อสอบถามจาก 83 ข้อสอบถามเท่านั้น ดังนั้นถ้าจำนวนข้อสอบถามถูกขยายคำมีมากกว่านี้น่าจะทำให้ประสิทธิภาพของการค้นคืนเอกสารโดยรวมมีค่ามากยิ่งขึ้น

5.3.2 เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคกฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เพียงอย่างเดียว

การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้มีประสิทธิภาพมากกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เท่านั้น เนื่องจากการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้มีการใช้เทคนิคกฎความสัมพันธ์ของคำที่สามารถทำให้ประสิทธิภาพการค้นคืนเอกสารเพิ่มขึ้น (จากผลในข้อ 1)

5.3.3 เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารระหว่างการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคกฎความสัมพันธ์ของคำ

ผลการทดลองเปรียบเทียบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้มีประสิทธิภาพน้อยกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำ จากผลการทดลองที่ว่าการใช้เทคนิคการใช้กฎความสัมพันธ์ของคำสามารถช่วยเพิ่มประสิทธิภาพให้กับการค้นคืนเอกสารแบบใช้เทคนิคปริภูมิเวกเตอร์อย่างมีนัยสำคัญ แต่เมื่อการค้นคืนเอกสารใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้นั้นทำให้ประสิทธิภาพลดลง

5.3.4 เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้กับการค้นคืนเอกสารทั้ง 3 รูปแบบข้างต้น

ผลการวิเคราะห์ประสิทธิภาพการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ ผู้วิจัยตั้งข้อสังเกตว่าการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้ไม่ควรใช้ร่วมกับกฎความสัมพันธ์ของคำหรือไม่ จึงทำการทดสอบต่อไปว่าการค้นคืนเอกสารที่ใช้เทคนิคการให้ผลสะท้อนกลับมีประสิทธิภาพอย่างไรเมื่อเทียบกับการค้นคืนเอกสารทั้ง 3 รูปแบบที่กล่าวมาข้างต้น ซึ่งเมื่อวิเคราะห์ผลการทดลองพบว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้มีประสิทธิภาพน้อยกว่าการค้นคืนเอกสารทั้ง 3 รูปแบบข้างต้น

ดังนั้นผู้วิจัยจึงสรุปว่าการใช้เทคนิคการให้ผลสะท้อนกลับในการศึกษาครั้งนี้ไม่สามารถเพิ่มประสิทธิภาพการค้นคืนเอกสารได้ ซึ่งผู้วิจัยเห็นว่าการที่ประสิทธิภาพของการค้นคืนเอกสารที่

ใช้เทคนิคการให้ผลสะท้อนกลับมีประสิทธิภาพลดลงกว่าการไม่ใช้ อาจเกิดจากค่า α, β, γ ในสมการที่ 2.11 โดยกำหนดไม่เหมาะสมกับฐานข้อมูลนิตยสารไทม์ (Time Collection) เนื่องจากงานวิจัยความถูกต้องของการให้ผลสะท้อนกลับและการจัดกลุ่มเอกสารของ Iwayama (Iwayama, 2000) และงานวิจัยผลกระทบเมื่อให้ผลสะท้อนกลับของ Buckley และคณะ (Buckley et al., 1994) ที่ผู้วิจัยอ้างอิงค่า α, β, γ เท่ากับ 8, 16, 4 ตามลำดับ ได้แสดงให้เห็นว่าการค้นคืนเอกสารโดยใช้เทคนิคผลสะท้อนกลับจากผู้ใช้นั้นมีประสิทธิภาพที่ดีกับฐานข้อมูลทีอาร์อีซี (TREC Collection) จึงเป็นการตั้งค่า α, β, γ เท่ากับค่าที่ตั้งไว้ในงานวิจัยดังกล่าว อาจให้ผลที่แตกต่างจากงานวิจัยนี้ ซึ่งให้ผลการทดลองของการค้นคืนเอกสารมีประสิทธิภาพน้อยลงเมื่อใช้เทคนิคการให้ผลสะท้อนกลับ นอกจากนี้ผู้วิจัยเห็นว่าอาจเป็นเพราะเอกสารนิตยสารไทม์ (TIME) จากฐานข้อมูลนิตยสารไทม์ (TIME Collection) เป็นเรื่องราวข่าวสารทั่วไป จึงมีเนื้อหาสาระหลากหลายไม่เป็นหมวดหมู่ เอกสารที่เกี่ยวเนื่องกับข้อสอบถามที่ทดสอบอาจไม่อยู่เป็นกลุ่ม ทำให้การปรับเวกเตอร์ข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องนั้นเป็นไปได้ยาก จึงทำให้ผลการทดลองไม่ได้ผลที่ดีขึ้นเช่นเดียวกับผลการทดลองในการให้ผลสะท้อนกลับในงานวิจัยที่ผ่านมา

5.3.5 เปรียบเทียบประสิทธิภาพการค้นคืนเอกสารทั้ง 4 รูปแบบเมื่อใช้ค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision)

จากการวิเคราะห์ข้างต้นเป็นการวัดประสิทธิภาพการค้นคืนเอกสารโดยใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic of recall and precision) ซึ่งเป็นค่าที่หาค่าเฉลี่ยระหว่างค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) เมื่อผู้วิจัยจะวิเคราะห์ค่าประสิทธิภาพโดยใช้ค่าเรียกคืนจะให้ผลลัพธ์ว่าค่าเรียกคืน (Recall) ผลการวิเคราะห์สรุปว่าค่าเรียกคืน (Recall) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้มีค่าเท่ากันกับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ ซึ่งทั้ง 2 รูปแบบมีค่าเรียกคืน (Recall) น้อยกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ที่มีค่าเรียกคืน (Recall) เท่ากับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ ผู้วิจัยเห็นว่าการค้นคืนเอกสารที่ใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้จะสามารถให้ค่าประสิทธิภาพค่าเรียกคืนมากกว่าการค้นคืนเอกสารที่ใช้เทคนิคกฎความสัมพันธ์ นั่นคือสามารถดึงเอกสารที่เกี่ยวเนื่องออกมาได้มากขึ้น และเมื่อวัดประสิทธิภาพของการค้นคืนเอกสารโดยใช้ค่าความถูกต้อง (Precision) จะให้ผลลัพธ์ว่าค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้มีค่า

เท่ากับการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ และการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ ซึ่งมีค่าน้อยกว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

จากผลสรุปการวิเคราะห์ผลการทดลองผู้วิจัยเห็นว่าประสิทธิภาพค่าความถูกต้องของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้มีค่าความถูกต้องน้อยที่สุด แต่มีค่าเรียกคืนมากอาจเป็นเพราะเอกสารถูกค้นคืนออกมามากทำให้เจอเอกสารที่เกี่ยวเนื่องมากขึ้น

5.4 การนำงานวิจัยไปประยุกต์ใช้

ในงานวิจัยนี้สามารถใช้เป็นแนวทางในการศึกษาต่อไปหรือนำไปประยุกต์ใช้ในการค้นคืนเอกสารด้านต่าง ๆ โดยผู้วิจัยสามารถแบ่งข้อเสนอได้ดังต่อไปนี้

5.4.1 การนำงานวิจัยไปใช้ในเชิงทฤษฎี

ในอดีตมีงานวิจัยทดสอบประสิทธิภาพการค้นคืนเอกสารโดยการใช้เทคนิคการใช้กฎความสัมพันธ์ที่ให้ผลประสิทธิภาพดี ดังนั้นผู้วิจัยจึงสนใจเพิ่มประสิทธิภาพการค้นคืนเอกสารมากขึ้นจากการใช้เทคนิคกฎความสัมพันธ์โดยนำเทคนิคการให้ผลสะท้อนกลับเข้ามาด้วย เนื่องจากเป็นเทคนิคที่ทำให้คอมพิวเตอร์สามารถโต้ตอบกับผู้ใช้ได้ โดยไม่ได้เป็นการสื่อสารทางเดียวคือเมื่อผู้ใช้กรอกข้อสอบถามแล้วระบบจะแสดงเอกสารที่เป็นผลลัพธ์ออกมา และจากการที่มนุษย์จะรู้และเข้าใจความหมาย (Meaning) ของเอกสารที่เป็นผลลัพท์ แต่คอมพิวเตอร์ไม่รู้ ทำให้การทำงานของระบบจะค้นคืนเอกสารจากการที่มนุษย์กำหนดให้ทำงานเท่านั้น ดังนั้นการที่ผู้ใช้สามารถให้ผลสะท้อนกลับที่ถูกต้องกลับเข้ามายังระบบ จะทำให้ระบบสามารถค้นคืนเอกสารที่เกี่ยวเนื่องออกมามากขึ้น ดังนั้นการใช้เทคนิคการให้ผลสะท้อนกลับสามารถเป็นแนวทางให้กับผู้ที่สนใจในด้านการพัฒนาการค้นคืนเอกสารได้นำข้อมูลเหล่านี้ไปใช้ในการศึกษาต่อไป

5.4.2 การนำงานวิจัยไปใช้ในเชิงประยุกต์

1. เทคนิคการค้นหากฎความสัมพันธ์ (Association Discovery) สามารถนำไปใช้ร่วมกับการค้นคืนเอกสารในส่วนของการกำหนดคำที่มีความสัมพันธ์กัน โดยเฉพาะในกลุ่มของเอกสารที่กำหนดคำที่มีความสัมพันธ์กันได้ยาก เช่น เอกสารในเชิงธุรกิจ เนื่องจากคำที่ใช้ในบทความเชิงธุรกิจเป็นคำทั่วไปที่มีความหลากหลาย การกำหนดคำที่มีความสัมพันธ์กันนั้นอาจกำหนดได้ยาก ไม่เหมือนกับเอกสารด้านวิทยาศาสตร์ที่จะมีการกำหนดใช้คำที่เป็นคำศัพท์เฉพาะ

(Technical term) ที่มีความสัมพันธ์กันไว้อย่างเป็นสากล ดังนั้นจึงสามารถใช้เทคนิคการใช้กฎความสัมพันธ์มาเพื่อช่วยในการหาคำที่มีความสัมพันธ์กันที่ได้

2. จากผลการทดลองประสิทธิภาพโดยใช้ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ซึ่งเป็นค่าที่วัดประสิทธิภาพโดยรวมของระบบค้นคืนเอกสาร แสดงให้เห็นว่าผู้ใช้สามารถนำกฎความสัมพันธ์ของคำไปใช้ร่วมกับการค้นคืนเอกสาร เพื่อเพิ่มประสิทธิภาพให้กับการค้นคืนเอกสารให้ดียิ่งขึ้นได้ แต่หากผู้ใช้ต้องการวัดประสิทธิภาพโดยใช้ค่าเรียกคืน (Recall) ผลการศึกษาเพิ่มเติมแสดงให้เห็นว่า การใช้เทคนิคกฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้หรือเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้ไปใช้ร่วมกับการค้นคืนเอกสารให้ค่าประสิทธิภาพดี ในกรณีที่ผู้ใช้ต้องการวัดประสิทธิภาพโดยใช้ค่าความถูกต้อง (Precision) จากผลการทดลองเพิ่มเติมแสดงให้เห็นว่าเทคนิคที่ให้ค่าความถูกต้องที่ดีคือ การใช้เทคนิคกฎความสัมพันธ์ของคำหรือเทคนิคกฎความสัมพันธ์ของคำร่วมกับเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้หรือไม่ใช้เทคนิคทั้งสองเลย

5.5 ข้อจำกัดของงานวิจัย

จากการทดลองการค้นคืนเอกสารในงานวิจัยนี้ มีข้อจำกัดบางประการดังนี้

1) ผลการทดลองของการค้นคืนเอกสารในงานวิจัยนี้ เป็นผลจากการทดสอบกับชุดเอกสารและข้อสอบถามของฐานข้อมูลนิตยสารไทม์ (TIME Collection) เท่านั้น
2) การทดลองของการค้นคืนเอกสารในงานวิจัยนี้ ผลการทดลองเป็นผลที่ได้จากการที่ผู้วิจัยกำหนดค่าต่างๆ ในเครื่องมือทดสอบการค้นคืนเอกสารดังนี้
   - ค่าความเหมือนต่ำสุดในการค้นคืนเอกสารไว้เท่ากับ 0.0439
   - ค่าสนับสนุนต่ำที่สุด (Minimum support) และค่าความเชื่อมั่นต่ำสุด (Minimum confidence) ในงานวิจัยนี้ ผู้วิจัยกำหนดให้มีค่าเท่ากับ 1.6471 และ 70 เปอร์เซนต์ ตามลำดับ
   - การให้ค่าน้ำหนักในสมการของการให้ผลสะท้อนกลับของรอคชิโอ (Rochio) (Baeza-Yates and Ribeiro-Neto, 1999) ได้กำหนดค่า α, β, γ ให้มีค่าเท่ากับ 8, 16, 4 ตามลำดับ

5.6 แนวทางการศึกษาต่อเนื่อง

จากข้อจำกัดของงานวิจัย ผู้ที่สนใจศึกษาต่อเนื่องอาจเป็นแนวทางในการศึกษาดังนี้

1) ผู้วิจัยสามารถทดสอบกับชุดเอกสารและข้อสอบถามที่นอกเหนือจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) เพื่อให้ผลการทดลองครอบคลุมกับเอกสารหลากหลายประเภทมากยิ่งขึ้น
2) ในการทดสอบการค้นคืนเอกสารสามารถตั้งค่าต่าง ๆ ได้ตามความเหมาะสมของเครื่องมือทดสอบการค้นคืนเอกสารที่พัฒนา ได้แก่ค่าดังต่อไปนี้
    - การตั้งค่าความเหมือนต่ำสุดในการค้นคืนเอกสาร
    - ค่าสนับสนุนต่ำสุด (Minimum Support) และค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) ผู้ที่จะนำไปศึกษาต่อสามารถกำหนดค่าทั้งสองแตกต่างออกไปจากงานวิจัยนี้ ไม่ว่าฐานข้อมูลที่ต้องการศึกษาจะเป็นฐานข้อมูลนิตยสารไทม์ (TIME Collection) ในหรือฐานข้อมูลเอกสารอื่น ๆ ผู้ศึกษาสามารถเปลี่ยนแปลงได้ตามความเหมาะสม
    - ค่าที่ใช้ในการปรับค่าน้ำหนักในเวกเตอร์ข้อสอบถาม เพื่อไปดึงเอกสารอีกครั้ง นั่นคือค่า α, β, γ .โดยการตั้งค่า α, β, γ จึงตั้งได้ตามความเหมาะสมของระบบค้นคืนเอกสารที่พัฒนา ผู้วิจัยได้ศึกษาต่อในการปรับค่าความเหมือนค่า α ให้มีค่ามากขึ้นและปรับค่า β ให้มีค่าน้อยลง พบว่าผลประสิทธิภาพดีขึ้นกว่าการตั้งค่า α, β, γ เท่ากับ 8,16,4 ตามลำดับ ผู้วิจัยปรับค่าดังที่กล่าวมาเนื่องจากค่า α เป็นค่าที่พิจารณาร่วมกับกลุ่มเอกสารที่เกี่ยวเนื่อง และค่า β เป็นค่าที่พิจารณาร่วมกับกลุ่มเอกสารที่ไม่เกี่ยวเนื่อง โดยจะเป็นการเพิ่มน้ำหนักคำในเอกสารที่เกี่ยวเนื่องมากขึ้นและลดน้ำหนักคำในเอกสารที่ไม่เกี่ยวเนื่อง

รายการอ้างอิง

ภาษาไทย:

กฤษณี อริยชาญศิลป์. (2545). "ระบบค้นคืนสารสนเทศภาษาไทย-อังกฤษ สำหรับคำทับศัพท์และแสดงผลลัพธ์ด้วยวิธีจัดกลุ่มข้อมูล". วิทยานิพนธ์ปริญญาดุษฎีบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

กิตติ ภักดีวัฒนะกุล อังศุมาลิน เวชนารายณ์ กิตติพงษ์ ธีรวัฒน์เสถียร. (2545). "PHP ฉบับโปรแกรมเมอร์". KTP COMP & CONSULT.

กัลยา วานิชย์บัญชา. (2546). "การวิเคราะห์สถิติ: สถิติสำหรับการบริหารและวิจัย". ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย.

กัลยา วานิชย์บัญชา. (2548). "การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล". ศูนย์หนังสือแห่งจุฬาลงกรณ์มหาวิทยาลัย.

บัณฑิต จามรภูติ. (2541). "การใช้งานฐานข้อมูลเชิงสัมพันธ์ Microsoft SQL Server". บริษัท ว. เพ็ชรสกุล จำกัด.

สมประสงค์ ธิตินิลนิธิ. (2545). "เรียนลัด PHP 4 ครอบคลุมเวอร์ชั่น 4.2". บริษัท โปรวิชั่น จำกัด.

สมพร จิวรสกุล. (2545). "คู่มือการติดตั้งและใช้งาน Microsoft SQL Server 2000 ฉบับสมบูรณ์". สำนักพิมพ์ อินโฟเพรส.

ภาษาอังกฤษ:

Antonie, M., and Zaiane, O. (2002). Text Document Categorization by Term Association. IEEE International Conference on Data Mining (ICDM'02) p.19.

Baeza-Yates, R., and Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press Book.

Bass, L., Clements, P., and Kazman, R. (1998). Software Architecture in Practice. Addison Wesley.

Buckly, C., Salton, G., and Allan, J. (1994) The Effect of Adding Relevant Information in a Relevant Feedback Environment. ACM SIGIR.

Chakrabarti, S. (2003). Mining The Web Discovering Knowledge from Hypertext Data. Morgan Kaufmann Publishers.

Cherfi, H., Napoli, A., and Toussaint, Y. (2006). Towards A Text Mining Methodology Using Association Rule Extraction. Springer-Verlag Berlin Heidelberg. Issue: Volume 10, Number 5, March. 2006.

Chowdhury, G. (2004). Introduction to Modern Information Retrieval. second edition. London : Library Association Publishing.

Cognitive Science Laboratory. (2005). "Wordnet". [Online] Available: http://www.wordnet.princiton.edu. Princeton University.

Delgado, M., Martin-Bautists, M., Sanchez, D., Serrano, J, and Vila, M. (2002). Association Rule Extraction for Text Mining. Springer-Verlag Berlin Heidelberg. 5th International Conference. FQAS 2002, Copenhagen, Denmark, October 27-29. 2002.

Dimitrios, Z. and Gallopoulos, E. (2005) TMG : A MATLAB Toolbox for generating term-document matrices from text collections. http://scgroup.hpclab.ceid.upatras.gr/scgroup/Projects/TMG/.

Dumais, S. (1991). Improving the Retrieval of Information from External Source. Behavior Research Methods, Instruments, & Computers 1991, 23 (2), 229-236.

Haddad, M., Chevallet, J., Bruandet, M. (2000). Relation between Terms Discovery by Association Rules. in 4th European conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, Lyon France, september 12, 2000.

Iwayama, M. (2000) Relevant Feedback with a Small Number of Relevant Judgements: Incremental Relevant Feedback vs. Document Clustering. ACM SIGIR 2000 7/00 Athens, Greece.

Kou,H., and Gardarin, G. (2002). Similarity Model and Association For Document Categorization. IEEE Proceeding of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02).

Lee, D., Chuang, H., and Seamons, K. (1997) Document Ranking and the Vector-Space Model. Parallel and Distributed Systems, IEEE Transactions, on Volume 15, Issue 1, Jan. 2004 Page(s):18 – 27.

Matsumura, N., Ohsawa, Y., and Ishizuka, M. (2002). PAI - Automatic Indexing for Extracting Asserted Keywords from a Document. American Association for Artificial Intelligence (www.aaai.org).

Matsuo, Y., and Ishizuka, M. (2003). Keyword Extraction from a Single Document using Word Co-occurance Statistical Information, American Association for Artificial Intelligence (www.aaai.org).

Meadow, T.C., Boyce, R.B., and Kraft, H.D. (2000). Text Information Retrieval System. second edition. Academic Press.

Porter, M. (1980). "The Porter Stemming Algorithm". [Online] Available: http://www.tartarus.org/~martin/PorterStemmer/index.html.

Qin, Z., Liu, L., and Zhang, S. (2004). Mining Term Association Rules for Heuristic Query Construction. Springer-Verlag Berlin Heidelberg. PAKDD 2004, LNAI 3056, pp. 145-154.

Rauber, A., and Merkl, D. (1999). Mining Text Archives: Creating Readable Maps to Structure and Describe Document Collections. Springer-Verlag Berlin Heidelberg. PKDD'99, LNAI 1704, pp. 524-529.

Rauber, A. and Merkl, D. (2000). Providing Topically Sorted Access to Subsequently Released Newspaper Edition or: How to Built Your Private Digital Library. Springer-Verlag Berlin Heidelberg DEXA 2000. LNCS 1873, pp. 499-508.

Robert, R. (1997). Information Storage and Retrieval. Wiley Computer Publishing.

SAS and all other SAS Institute Inc. (2005). "SAS Enterprise Miner 5.2". Available: http://www.sas.com/technologies/analytics/datamining/miner/factsheet.pdf.

Silva, I., Souza, J., and Santos, S. (2004). Dependence Among Terms in Vector Space Model. IEEE Proceedings of the International Database Engineering and Applications Symposium (IDEAS'04).

Smart Collection (1963) "Time Collection". [Online]. Available: ftp://ftp.cs.cornell.edu/pub/smart/time.

Smart System. (2005). Stop word list. http://www.unine.ch/info/clef/englishST.txt.

Song, M., Song, I. Y., Hu, X., and Allen, R. (2005). Semantic Query Expansion Combining Association Rules with Ontologies and Information Retrieval Techniques. Springer-Verlag Berlin Heidelberg. 7th International Conference, DaWaK , Copenhagen, Denmark, August 22-26.

Sullivan, D. (2001). Document Warehouse Text Mining. Wiley Computer Publishing.

Udomchaiporn, Akadej. (2005). Use Case Retrieval using Terms and Use Case Structure Similarity Computation. A Thesis Submitted in Partial Fulfillment of the Requirements for the Degreeof Master of Science Program in Software Engineering Department of Computer Engineering. Faculty of Engineering. Chulalongkorn University.

Weiss, S., Indurkhya, N., Zhang, T., and Dameran, F. (2005). Text Mining Predictive Methods for Analyzing Unstructured Information. Springer.

Williamson, D., and Lesk M. (1971). The Cornell Implementation of the SMART System. In The SMART Retrieval System. edited by G. Salton. Prentice-Hall, Englewood Cliffs, N.J.

Yahoo. (2005). “Lemur Search”. [Online] Available: http://www.rollyo.com/shaper/lemur.

Ye, N. (2001). The Handbook of Data Mining. Lawrence Erlbaum Association. Arizona State University.

Zaki, M (2004). Mining Non-Redundant Association Rule. Kluwer Academic Publishers. Manufactured in The Netherlands. Data Mining and Knowledge Discovery, 9 2004. Page(s):223–248.

Zhuang, L., and Dai, H. (2004). A Maximal Frequent Itemset Approach For Web Document Clustering. IEEE Proceeding of the Fourth International Conference on Computer and international Technology (CIT’04).

ภาคผนวก

ภาคผนวก ก

นิยามคำศัพท์

ศัพท์ (ไทย)	ศัพท์ (อังกฤษ)	นิยาม
คำ	Term	ตัวอักษรที่มาเรียงต่อกันแล้วเกิดเป็นความหมาย
เซตข้อมูล	Item set	เซตของของข้อมูลตั้งแต่หนึ่งข้อมูลขึ้นไป
ความสัมพันธ์ของคำ	Correlation	ความเกี่ยวข้องในเชิงความหมายของคำต่าง ๆ เหล่านั้น
พจนานุกรม	Dictionary	ที่เก็บคำศัพท์ทั้งหมดที่มีอยู่ในระบบ เพื่อให้ระบบรู้จักคำศัพท์เหล่านี้ โดยจะนิยามคำศัพท์ ระบุคำที่เหมือนกันและคำที่เกี่ยวข้องกัน (Meadow et al., 2000)
เอกสาร	Document	ข้อมูลชนิดข้อความที่ประกอบไปด้วยตัวอักษรที่นำมาเรียงต่อกันเป็นคำ ข้อความ จนถึงเป็นบทความ
ข้อสอบถาม	Query	คำที่ผู้ใช้กรอกเข้ามายังระบบเพื่อต้องการให้ระบบค้นคืนเอกสารที่ต้องการออกมา
เทคนิคการค้นคืนสารสนเทศ	Information Retrieval	กระบวนการในการรวบรวมสารสนเทศและทำรายการให้กับสารสนเทศที่รวบรวมไว้ ทั้งนี้เพื่อให้ทราบที่อยู่ของสารสนเทศและสามารถแสดงผลการค้นออกมาตามรูปแบบที่ต้องการ (Baeza-Yates and Ribeiro-Neto, 1999)
คำเดี่ยว	Single word	คำที่เป็นตัวอักษรมาเรียงต่อกันแล้วเกิดความหมายที่ระหว่างการเชื่อมตัวอักษรไม่มีเว้นวรรคคั่น และจะต้องเป็นอิสระกับคำอื่น ๆ คือ ไม่มีความหมายร่วมกับคำเดี่ยวอื่น ๆ ด้วย (Weiss et al., 2005)
ดรรชนี	Index	เป็นเครื่องชี้นำที่จัดทำขึ้นอย่างเป็นระบบเพื่อชี้ไปยังตำแหน่งของคำหรือแนวคิดที่สำคัญในเอกสาร ประกอบด้วยรายการดรรชนีที่จัดเรียงอย่างเป็นระบบ ซึ่งจะกำหนดจากสาระสำคัญของเอกสาร ชื่อบุคคล ชื่อสถานที่ ชื่อเฉพาะต่าง ๆ ที่เป็นแกนเรื่องของเอกสาร (Baeza-Yates and Ribeiro-

ศัพท์ (ไทย)	ศัพท์ (อังกฤษ)	นิยาม
		Neto, 1999)
การให้ผลสะท้อนกลับ	Relevant feedback	เป็นวิธีการปรับปรุงข้อสอบถามที่นิยม โดยจะเป็นการเลือกคำหรือสำนวนที่มีความสำคัญที่อยู่ในเอกสารที่ผู้ใช้ระบุมายังระบบค้นคืนว่าเกี่ยวเนื่องกับความต้องการของผู้ใช้ เพื่อนำไปปรับเปลี่ยนข้อสอบถามให้มีประสิทธิภาพมากยิ่งขึ้น (Baeza-Yates and Ribeiro-Neto, 1999)
ประสิทธิภาพของข้อสอบถาม		ข้อสอบถามนั้นสามารถนำไปค้นคืนเอกสารจากระบบออกมาแสดงแก่ผู้ใช้ได้อย่างถูกต้อง ตรงกับความต้องการของผู้ใช้
กฎความสัมพันธ์	Association Rule	กฎที่แสดงความสัมพันธ์ของข้อมูล
การทำเหมืองข้อมูลข้อความ	Text Mining	เป็นวิธีทางวิทยาศาสตร์ที่ดึงสารสนเทศและองค์ความรู้ออกมาจากเอกสาร (Sullivan, 2001)
คำยกเว้น	Stop word	คำที่เป็นคำนำหน้านาม คำบุพบทและคำเชื่อมในภาษาอังกฤษซึ่งส่วนใหญ่จะเป็นคำที่มีความถี่เกิน 80 เปอร์เซ็นต์ในเอกสาร ไม่มีความสามารถในการแยกแยะเอกสาร (Baeza-Yates and Ribeiro-Neto, 1999)
การค้นหาแบบแฮช	Hashing	เป็นการค้นหาที่มุ่งเน้นที่ประสิทธิภาพการค้นหา โดยที่มีค่าเวลาที่ใช้ในการค้นหาเป็นค่าคงที่คือ O(1) (บิ๊กโอ 1)ซึ่งหมายความว่าค้นหาเพียงครั้งเดียวก็พบแล้ว การค้นหาแบบนี้จะประกอบด้วย 3 ส่วนหลัก ๆ คือ คีย์ ฟังก์ชั่นแฮชและตารางแฮช
คีย์	Key	เลขที่ใช้จัดเก็บแถวลำดับ (Array) ข้อมูล
ฟังก์ชั่นแฮช	Hash Function	วิธีการหาว่าคีย์นี้อยู่ที่ตำแหน่งใดของแถวลำดับ (Array)
ตารางแฮช	Hash Table	แถวลำดับ (Array) ที่ใช้เก็บข้อมูล
ตัวชี้	Pointer	เป็นตัวแปรชนิดหนึ่งซึ่งมีเนื้อที่อยู่ในหน่วยความจำเหมือนตัวแปรอื่น ๆ ทั่วไป โดยจะมีหน้าที่หลักคือ การชี้ไปยัง address ใด ๆ เพื่อไปจัดการกับข้อมูลที่เก็บอยู่ใน address นั้น ๆ

ภาคผนวก ข

รายการคำยกเว้น (Stop words list)

รายการคำที่เป็นคำยกเว้นที่ใช้ในงานวิจัยมีดังต่อไปนี้ โดยนำมาจากระบบค้นคืนเอกสารสมาร์ท (SMART) ที่พัฒนาโดยมหาวิทยาลัยคอแนล (Cornell University) (SMART System, 2005)

a	everywhere	N	Thanx
a's	Ex	Name	That
able	Exactly	Namely	that's
about	Example	Nd	thats
above	Except	near	the
according	F	nearly	their
accordingly	Far	necessary	theirs
across	Few	need	them
actually	Fifth	needs	themselves
after	First	neither	then
afterwards	Five	never	thence
again	Followed	nevertheless	there
against	Following	New	there's
ain't	Follows	Next	thereafter
all	For	Nine	thereby
allow	Former	No	therefore
allows	Formerly	nobody	therein
almost	Forth	Non	theres
alone	Four	none	thereupon
along	From	noone	these
already	Further	Nor	they

also	Furthermore	normally	they'd
although	G	Not	they'll
always	Get	nothing	they're
am	Gets	novel	they've
among	Getting	Now	think
amongst	Given	nowhere	third
an	Gives	O	this
and	Go	obviously	thorough
another	Goes	Of	thoroughly
any	Going	Off	those
anybody	Gone	often	though
anyhow	Got	Oh	three
anyone	Gotten	Ok	through
anything	Greetings	okay	throughout
anyway	H	Old	thru
anyways	Had	On	thus
anywhere	hadn't	once	to
apart	Happens	One	together
appear	Hardly	ones	too
appreciate	Has	Only	took
appropriate	hasn't	Onto	toward
are	Have	Or	towards
aren't	haven't	other	tried
around	Having	others	tries
as	He	otherwise	truly
aside	he's	ought	try
ask	Hello	Our	trying
asking	Help	Ours	twice
associated	Hence	ourselves	two

at	Her	Out	u
available	Here	outside	un
away	here's	over	under
awfully	Hereafter	overall	unfortunately
b	Hereby	Own	unless
be	Herein	P	unlikely
became	Hereupon	particular	until
because	Hers	particularly	unto
become	Herself	Per	up
becomes	Hi	perhaps	upon
becoming	Him	placed	us
been	Himself	please	use
before	His	Plus	used
beforehand	Hither	possible	useful
behind	Hopefully	presumably	uses
being	How	probably	using
believe	Howbeit	provides	usually
below	However	Q	uucp
beside	I	Que	v
besides	i'd	quite	value
best	i'll	Qv	various
better	i'm	R	very
between	i've	rather	via
beyond	Ie	Rd	viz
both	If	Re	vs
brief	Ignored	really	w
but	Immediate	reasonably	want
by	In	regarding	wants
c	Inasmuch	regardless	was

c'mon
c's
came
can
can't
cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
couldn't
course
currently
Inc
Indeed
Indicate
Indicated
Indicates
Inner
Insofar
Instead
Into
Inward
Is
isn't
It
it'd
it'll
it's
Its
Itself
J
Just
K
Keep
Keeps
Kept
Know
Knows
Known
L
Last
regards
relatively
respectively
right
S
Said
same
Saw
Say
saying
says
second
secondly
See
seeing
seem
seemed
seeming
seems
seen
Self
selves
sensible
Sent
serious
seriously
seven
several
shall
wasn't
way
we
we'd
we'll
we're
we've
welcome
well
went
were
weren't
what
what's
whatever
when
whence
whenever
where
where's
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while

d	Lately	She	whither
definitely	Later	should	who
described	Latter	shouldn't	who's
despite	Latterly	since	whoever
did	Least	Six	whole
didn't	Less	So	whom
different	Lest	some	whose
do	Let	somebody	why
does	let's	somehow	will
doesn't	Like	someone	willing
doing	Liked	something	wish
don't	Likely	sometime	with
done	Little	sometimes	within
down	Look	somewhat	without
downwards	Looking	somewhere	won't
during	Looks	soon	wonder
e	Ltd	sorry	would
each	M	specified	wouldn
edu	Mainly	specify	wouldn't
eg	Many	specifying	x
eight	May	Still	y
either	Maybe	Sub	yes
else	Me	such	yet
elsewhere	Mean	Sup	you
enough	Meanwhile	Sure	you'd
entirely	Merely	T	you'll
especially	Might	t's	you're
et	More	Take	you've
etc	Moreover	taken	your

even	Most	Tell	yours
ever	Mostly	tends	yourself
every	Much	Th	yourselves
everybody	Must	Than	z
everyone	My	thank	zero
everything	Myself	thanks	

ภาคผนวก ค

ขั้นตอนวิธีของพอร์ทเตอร์ (Porter's Algorithm)

กฎ (rule) ของขั้นตอนวิธีของพอร์ทเตอร์กำหนดข้อตกลงเริ่มแรกดังนี้ (Baeza-Yates and Ribeiro-Neto, 1999)

- ตัวแปรกลมกลืนถูกแสดงโดย C แต่ไม่ใช่ตัวอักษร a,e,i,o,u และไม่ใช่ตัวอักษร y ที่ตามหลังตัวแปรกลมกลืน
- ตัวแปรที่เป็นเสียงสระถูกแสดงโดย V แสดงถึงทุกตัวอักษรที่ไม่ใช่ตัวแปรกลมกลืน
- ตัวอักษรโดยทั่วไปถูกแสดงโดย L
- สัญลักษณ์ ϕ ถูกใช้เพื่อแสดงถึงสายอักขระว่าง
- การรวมกันของ C V และ L จะเรียกว่า pattern
- สัญลักษณ์ * ถูกใช้เพื่อแสดงถึงการเกิดซ้ำของ pattern ตั้งแต่ศูนย์หรือมากกว่าของ
- สัญลักษณ์ + ถูกใช้เพื่อแสดงถึงการเกิดซ้ำของ pattern ตั้งแต่หนึ่งหรือมากกว่าของ
- วงเล็บจะถูกใช้ในส่วนขยายของลำดับตัวแปรที่ใช้ตัวดำเนินการ * และ +
- Pattern ธรรมดา ถูกรวมตัวโดยสัญลักษณ์ วงเล็บและตัวดำเนินการ * และ +
- กฎการแทนที่จะถูกปฏิบัติเช่นเดียวกับคำสั่งที่แยกโดย ;
- กฎการแทนที่ถูกประยุกต์ไปเป็นส่วนต่อท้ายคำ (suffix) ในคำปัจจุบัน
- เงื่อนไขคำสั่ง if จะแสดงเป็น "if (pattern) rule" และ rule นั้นจะถูกดำเนินการเมื่อ pattern นั้นอยู่ในเงื่อนไขที่ตรงกับ current word
- บรรทัดที่เริ่มจาก "%" จะแสดงถึงข้อคิดเห็น (comment)
- "{ }" จะใช้เพื่อสร้างคำสั่งที่ประกอบกัน
- จะเลือกกฎเดียวที่จะดำเนินการจากทุกกฎในคำสั่งที่ประกอบกัน โดยที่การเลือกกฎนั้นจะต้องเหมือนกับส่วนต่อท้ายคำ (suffix)ที่ยาวที่สุด

ตัวอย่างคำสั่ง

$$if\ (*V*L)\ then\ \ ed \rightarrow \phi$$

แสดงว่าจะมี การแทนที่ suffix "ed" ที่เป็นว่าง ถ้าคำในปัจจุบันนั้นมีเสียงสระและตัวอักษรอย่างน้อยหนึ่งตัว

ขั้นตอนวิธีของพอร์ทเตอร์จะประยุกต์ใช้กับคำทุกคำในเอกสาร ซึ่งจะมีกระบวนคำสั่งดังต่อไปนี้

% Phase 1:Plural and past participles.

Select rule with longest suffix {

sses $\rightarrow$ ss;

ies $\rightarrow$ I;

ss $\rightarrow$ ss;

s $\rightarrow$ ϕ;

}

select rule with longest suffix{

If ((C)*((V)+(C)+)+(V)*eed) then eed $\rightarrow$ ee;

If (*V*ed or *V*ing) then{

select rule with longest suffix {

ed $\rightarrow$ ϕ;

ing $\rightarrow$ ϕ; }

select rule with longest suffix {

at $\rightarrow$ ate;

bl $\rightarrow$ ble;

iz $\rightarrow$ ize;

if ((*C1C2) and (C1 = C2) and (C1 $\notin$ {l,s,z}))

then C1C2 $\rightarrow$ C1;

if (((C)*((V)+(C)+)C1V1C2) and (C2 $\notin$ {w,x,y}))

then C1V1C2 $\rightarrow$ C1V1C2e;}

}

}

If (*V*y) then y $\rightarrow$ I;

If(C)*((V)+(C)+)+V) then

select rule with longest suffix {

ational $\rightarrow$ ate;

tional → tion;

enci → ence;

anci → ance;

izer → ize;

abli → able;

alli → al;

entli → ent;

eli → e;

ousli → ous;

ization → ize;

ation → ate;

ator → ate;

alism → al;

iveness → ive;

fulness → ful;

ousness → ous;

aliti → al;

iviti → ive;

biliti → ble; }

if((C)*((V)+(C)+)+(V)*) then

select rule with longest suffix {

icate → ic;

ative → ϕ;

alize → al;

iciti → ic;

ical → ic;

ful → ϕ;

ness → ϕ;

if((C)*((V)+(C)+)((V)+(C)+)+(V)*) then

select rule with longest suffix {

al $\rightarrow \phi$;

ance $\rightarrow \phi$;

ence $\rightarrow \phi$;

er $\rightarrow \phi$

ic $\rightarrow \phi$;

able $\rightarrow \phi$;

ible $\rightarrow \phi$;

ant $\rightarrow \phi$;

ement $\rightarrow \phi$;

ment $\rightarrow \phi$;

ent $\rightarrow \phi$;

ou $\rightarrow \phi$;

ism $\rightarrow \phi$;

ate $\rightarrow \phi$;

iti $\rightarrow \phi$;

ous $\rightarrow \phi$;

ive $\rightarrow \phi$;

ize $\rightarrow \phi$;

if(*s or *t) then ion $\rightarrow \phi$;)

select rule with longest suffix {

if ((C)*((V)+(C)+)((V)+(C)+)+(V)*) then e $\rightarrow \phi$;

if (((C)*((V)+(C)+)(V)*) and not ((*C1V1C2) and (C2 $\notin$ {w,x,y})))

then e $\rightarrow$ nil; }

if((C)*((V)+(C)+)((V)+(C)+)+V*ll) then LL $\rightarrow$ l

ภาคผนวก ง

ตัวอย่างเอกสารและข้อสอบถาม

งานวิจัยนี้กำหนดใช้เอกสารและข้อสอบถามของฐานข้อมูลนิตยสารไทม์ (TIME Collection) ปี 1963 มาทดสอบระบบค้นคืนเอกสารที่สร้างขึ้น ซึ่งมีเอกสารจำนวน 425 เอกสาร และข้อสอบถามจำนวน 83 ข้อสอบถาม (สามารถดูตัวอย่างได้ในภาคผนวก ง) โดยเป็นเอกสารที่เก็บรวบรวมโดย Cornell University (Smart Collection, 1963)

- **ตัวอย่างเอกสาร**

*TEXT 017 01/04/63 PAGE 020

THE ALLIES AFTER NASSAU IN DECEMBER 1960, THE U.S . FIRST

PROPOSED TO HELP NATO DEVELOP ITS OWN NUCLEAR STRIKE FORCE . BUT EUROPE MADE NO ATTEMPT TO DEVISE A PLAN . LAST WEEK, AS THEY STUDIED THE NASSAU ACCORD BETWEEN PRESIDENT KENNEDY AND PRIME MINISTER MACMILLAN, EUROPEANS SAW EMERGING THE FIRST OUTLINES OF THE NUCLEAR NATO THAT THE U.S . WANTS AND WILL SUPPORT . IT ALL SPRANG FROM THE ANGLO-U.S. CRISIS OVER CANCELLATION OF THE BUG-RIDDEN SKYBOLT MISSILE, AND THE U.S . OFFER TO SUPPLY BRITAIN AND FRANCE WITH THE PROVED POLARIS (TIME,DEC . 28) . THE ONE ALLIED LEADER WHO UNRESERVEDLY WELCOMED THE POLARIS OFFER WAS HAROLD MACMILLAN, WHO BY THUS KEEPING A SEPARATE NUCLEAR DETERRENT FOR BRITAIN HAD SAVED HIS OWN NECK . BACK FROM NASSAU, THE PRIME MINISTER BEAMED THAT BRITAIN NOW HAD A WEAPON THAT " WILL LAST A GENERATION . THE TERMS ARE VERY GOOD . " MANY OTHER BRITONS WERE NOT SO SURE . THOUGH THE GOVERNMENT WILL SHOULDER NONE OF THE $800 MILLION DEVELOPMENT COST OF POLARIS, IT HAS ALREADY POURED $28 MILLION INTO SKYBOLT AND WILL HAVE TO SPEND PERHAPS $1 BILLION MORE FOR A FLEET OF MISSILE-PACKING SUBMARINES . AT BEST, THE BRITISH WILL NOT BE ABLE TO DESIGN, BUILD AND PROVE ITS NUCLEAR FLEET BEFORE 1970, THREE YEARS AFTER BRITAIN'S BOMBER FORCE HAS PRESUMABLY BECOME OBSOLETE . THEN WHAT? TORY BACKBENCHERS ARE LOUDLY SKEPTICAL OF WHAT THEY CALL " THE SMALL TYPE " IN THE NASSAU PACT, WHICH STIPULATES THAT BRITAIN'S POLARIS SUBMARINE FLEET, EXCEPT WHEN " SUPREME NATIONAL INTERESTS "INTERVENE, MUST BE COMMITTED TO A TRULY MULTILATERAL NATO FORCE . DOES THAT MEAN THAT BRITAIN WILL EVENTUALLY HAVE NO STRIKE FORCE OF ITS OWN? WHO WILL DECIDE WHEN OR WHETHER NATIONAL INTERESTS JUSTIFY WITHDRAWAL OF SUBMARINES FROM NATO, PARTICULARLY IF THOSE NATIONAL INTERESTS CONFLICT WITH U.S . POLICY ? THE BIGGEST QUESTION OF ALL IS WHETHER FRANCE'S INCLUSION IN THE OFFER WAS A DELIBERATE PLOY BY JACK KENNEDY TO END OR AT LEAST DOWNGRADE BRITAIN'S PRIZED " SPECIAL RELATIONSHIP " WITH THE U.S . THE CARTOONISTS WENT EVEN FARTHER . THEY NOT ONLY SHOWED SUPERMAC JUMPING TO SUPERJACK'S COMMANDS, BUT DE GAULLE AND ADENAUER AS WELL . AS EDITH SAID . THE FRENCH, WHO GOT NO HELP FROM THE U.S . IN DEVELOPING THEIR FORCE DE FRAPPE, WERE QUICK TO CROW THAT BRITAIN'S VAUNTED TIES WITH THE U.S . HAD BROUGHT IT NOTHING BUT HUMILIATION . BY CONTRAST, BRAGGED FRENCH OFFICIALS, THE SKYBOLT FIASCO

ONLY VINDICATED FRANCE'S DECISION TO DEVELOP ITS OWN BOMBS AND DELIVERY SYSTEMS . THUS, THOUGH CHARLES DE GAULLE PROMISED TO " REFLECT " ON THE POLARIS OFFER, THERE WAS LITTLE LIKELIHOOD THAT HE WOULD ACCEPT ANY OFFER THAT WOULD SUBJECT A FRENCH FORCE TO ALLIED CONTROL . IT IS DE GAULLE'S UNSWERVING CONVICTION THAT IF THE RUSSIANS WERE ACTUALLY TO INVADE WESTERN EUROPE, NO NATION THAT WAS NOT DIRECTLY ATTACKED MEANING THE U.S . WOULD INVITE NUCLEAR DEVASTATION BY HELPING ITS ALLIES . THUS UNLIKE BRITAIN'S BOMBER FORCE, WHICH ALL ALONG HAS BEEN PLEDGED TO "THE WESTERN STRATEGIC DETERRENT, " FRANCE'S FORCE DE FRAPPE WILL BE RESPONSIBLE ONLY FOR FRANCE'S DEFENSE . AT THE SAME TIME, DE GAULLE HAS LONG ARGUED THAT THE ATLANTIC ALLIANCE COULD BE RUN MOST

EFFICIENTLY BY A TRIUMVIRATE THAT WOULD INCLUDE FRANCE AS AN EQUAL OF THE U.S . AND BRITAIN . THIS IS ONE OF HIS MAJOR, IF UNSPOKEN, CONDITIONS FOR BRITISH MEMBERSHIP IN THE COMMON MARKET ; AND DE GAULLE SUGGESTED POINTEDLY TO MACMILLAN THAT IT WOULD HELP IF BRITAIN WERE TO SHARE ITS ADVANCED MISSILE TECHNOLOGY WITH FRANCE . WHEN MACMILLAN REPLIED NONCOMMITTALLY
THAT HE WOULD HAVE TO DISCUSS THIS WITH KENNEDY, DE GAULLE TOLD HIS
GUEST WITH HAUTEUR THAT FRANCE IN THAT CASE COULD DO NOTHING TO EASE
BRITAIN'S ENTRY INTO EUROPE . GO-IT-ALONE GRANDEUR . KONRAD ADENAUER,
ON THE OTHER HAND, IS FEARFUL THAT DE GAULLE WILL SNAP UP THE POLARIS
OFFER AND IN THIS WAY ACHIEVE HIS GOAL OF A THREE-NATION NATO
DIRECTORATE . THOUGH HIS GOVERNMENT VOWED IN 1954 NOT TO MANUFACTURE
NUCLEAR WEAPONS, ADENAUER HAS BECOME INCREASINGLY APPREHENSIVE THAT
WITHOUT THEM, AND WITH NO SAY IN THEIR USE, WEST GERMANY WILL BE
RELEGATED TO SECOND-CLASS CITIZENSHIP IN THE ALLIANCE . LAST WEEK AN
OFFICIAL BULLETIN EVEN REVIVED THE OLD, BITTER CRY THAT U.S . PLEAS FOR GREATER RELIANCE ON CONVENTIONAL FORCES ARE AIMED AT RAISING GERMAN "CANNON FODDER " FOR U.S . " ATOMIC KNIGHTS . " A FROSTY LETTER FROM THE CHANCELLOR TO PRESIDENT KENNEDY SUGGESTED THAT GERMANY, WHICH ALREADY SUPPLIES ALMOST 50 PER CENT OF NATO GROUND STRENGTH, DOES NOT INTEND TO RAISE ANY MORE DIVISIONS FOR CONVENTIONAL WARFARE . YET U.S . STRATEGIC PLANNERS REASON THAT THE ONLY CREDIBLE DETERRENT TO SOVIET ATTACK IS A STRONG ARMY ON THE GROUND, BACKED BY THE VAST U.S . NUCLEAR ARSENAL . FACT IS, THE BRITISH AND FRENCH NUCLEAR WEAPONS COULD NEVER BE USED INDEPENDENTLY OF THE U.S . AGAINST RUSSIA WITHOUT INVITING DEVASTATING SOVIET RETALIATION . AFTER ALL THEIR EFFORTS, THE BRITISH AND FRENCH WILL HAVE MANAGED TO CREATE A NUCLEAR CAPACITY THAT REPRESENTS ONLY 4
PER CENT OF U.S . NUCLEAR POWER . " IT IS JUST A DAMNED NUISANCE, "
SAID A STATE DEPARTMENT OFFICIAL LAST WEEK . " IT MEANS NOTHING
MILITARILY EXCEPT THAT WE WILL BE EXPECTED TO BAIL OUT THE FIRST
COUNTRY THAT THROWS THE FIRST PEA AT THE RUSSIANS OR ANYONE ELSE . "
CHARLES DE GAULLE COULD HARDLY BE EXPECTED TO AGREE, AT LEAST UNTIL HIS FORCE DE FRAPPE BECOMES OBSOLETE . FOR BRITAIN AND GERMANY, THE
MULTILATERAL DETERRENT MAKES IMMEDIATE SENSE . EVENTUALLY, FRANCE, TOO, MAY WELL FIND A NATO-CONTROLLED POLARIS FLEET, OR ITS POSSIBLE
SUCCESSOR, A EUROPEAN MINUTEMAN ARSENAL, THE ONLY ANSWER TO THE
SPIRALING COST AND DIMINISHING VALUE OF GO-IT-ALONE GRANDEUR .

*TEXT 018 01/04/63 PAGE 021

RUSSIA WHO'S IN CHARGE HERE ? IT WAS IN 1954 THAT NIKITA
KHRUSHCHEV LAUNCHED HIS GRANDIOSE " VIRGIN LANDS " GAMBLE . PART OF THE PLAN WAS TO PLOW UP 32 MILLION ACRES OF MARGINAL LAND IN KAZAKHSTAN, AND SETTLE IT WITH COMMUNIST " PIONEERS, " WHO WERE TO PLANT AND PRODUCE HUGE QUANTITIES OF DESPERATELY NEEDED GRAIN WITHIN TWO YEARS. NIKITA'S SCHEME FLOPPED . THERE WAS NOT ENOUGH RAINFALL, AND THE PIONEERS DID NOT TAKE TO TRACTOR LIFE ON THE BLEAK FRONTIER . EXCEPT FOR 1958, EACH HARVEST HAS BEEN LOWER THAN THE PREVIOUS YEAR'S. WORST YEAR OF ALL WAS 1962, WHEN THE VIRGIN LANDS DELIVERED ONLY HALF THEIR QUOTAS . NATURALLY, KHRUSHCHEV TAKES NONE OF THE BLAME FOR THE FIASCO . THREE YEARS AGO HE FOUND A SCAPEGOAT IN KAZAKHSTAN PARTY BOSS NIKOLAI BELYAEV, FIRED HIM FOR HIS " ERRORS . " LAST WEEK BELYAEV'S SUCCESSOR, DINMUKHAMED KUNAEV, WAS SIMILARLY BOUNCED FOR " LAPSES " IN HIS WORK . FOR GOOD MEASURE, MOSCOW ALSO PURGED THE FORMER PREMIER OF THE TERRITORY FROM THE LOCAL PARTY'S CENTRAL COMMITTEE . IT WAS PERHAPS NO COINCIDENCE THAT NIKOLAI IGNATOV, 61, A ONETIME KHRUSHCHEV CRONY, LAST WEEK ABRUPTLY LEFT HIS POST AS A SOVIET DEPUTY PREMIER AFTER ONLY NINE MONTHS ON THE JOB . FARM EXPERT IGNATOV HAD THE MISFORTUNE TO BE BOSS OF A SPECIAL COMMITTEE TO BOOST FOOD PRODUCTION .

*TEXT 019 01/04/63 PAGE 021

BERLIN ONE LAST RUN HANS WEIDNER HAD BEEN HOPING FOR MONTHS TO

ESCAPE DRAB EAST GERMANY AND MAKE HIS WAY TO THE WEST . THE ODDS WERE
AGAINST HIM, FOR WEIDNER, 40, WAS A CRIPPLE ON CRUTCHES WHO LIVED IN
THE VILLAGE OF NEUGERSDORF, 115 MILES SOUTHEAST OF THE FRONTIER OF
FREEDOM BUT HANS WEIDNER DID HAVE ONE MAJOR ASSET, THE BUS THAT
HE OPERATED FOR THE LOCAL COMMUNIST REGIME . IT WAS AN UGLY THING, AND ANCIENT . ITS CHASSIS CREAKED, AND THE ENGINE COUGHED ; A
CREAM-COLORED COAT OF PAINT COULD NOT DISGUISE THE WELTS AND BRUISES OF TWO DECADES OF CHUGGING SERVICE . IN FACT, THE BUS WAS READY FOR THE JUNK PILE WHEN WEIDNER DECIDED TO PRESS IT INTO SERVICE FOR ONE LAST RUN . SHARP BLADES . THE HAZARDS WOULD BE GREAT ON THE JOURNEY TO THE BORDER ; SO WEIDNER SIGNED UP A FELLOW VILLAGER, JURGEN WAGNER, 22, TO TAKE THE WHEEL . EIGHT DAYS BEFORE CHRISTMAS, THE PAIR BEGAN THE FEVERISH PREPARATIONS IN WEIDNER'S GARAGE . FIRST WEIDNER AND WAGNER ATTACHED A HEAVY SNOWPLOW TO THE FRONT OF THE BUS, NOT TO PLOW SNOW, BUT TO SCOOP AWAY THE HEAVY OBSTACLES THEY KNEW AWAITED THEM AT ROADBLOCKS AHEAD . TO ALL SIX LUGS ON EACH FRONT WHEEL THEY BOLTED SHARP BLADES OF THE TOUGHEST STEEL, AFFIXED SO THAT THE WHIRLING EDGES WOULD CHOP BARBED WIRE TO BITS . THEN THEY WEDGED ONE-QUARTER-INCH SECTIONS OF STEEL PLATE INSIDE THE BUS TO STOP BULLETS . AT LAST ALL WAS READY . ON CHRISTMAS EVE, WEIDNER AND WAGNER PILED THEIR WIVES AND FOUR CHILDREN ABOARD, NOT FORGETTING THREE TONS OF HOUSEHOLD BELONGINGS. FOR ADDED PROTECTION THE PLOTTERS SHOVELED A TON OF COAL AND POTATOES INTO THE BACK OF THE BUS . THEN THEY CHUGGED OFF NORTH TOWARD BERLIN ALONG BACK ROADS TO ESCAPE COMMUNIST PATROLS. JUST BEFORE THEY REACHED THE WALL, THEY PLANNED TO SWING WEST IN ORDER TO ENTER THE EAST-WEST AUTOBAHN LEADING TO THE U.S . SECTOR OF THE CITY . EN ROUTE, THE RADIATOR FROZE IN THE SUBZERO WEATHER . THAT FIXED, THEY WERE ONLY A FEW MILES FARTHER WHEN A TIRE BLEW OUT . THE KIDS WERE CRYING AND THE WIVES SHIVERING WITH COLD AND PANIC WHEN, AT LAST, THEY ARRIVED AT DREWITZ, THE MOST HEAVILY GUARDED CHECKPOINT ON THE ENTIRE AUTOBAHN TO BERLIN . IT WAS NO TIME TO STOP AND RECONSIDER. FLYING POTATOES . "WAH-AH, WAH-AH, " SHRIEKED THE POLICE-TYPE KLAXONS THAT WEIDNER HAD THOUGHTFULLY INSTALLED IN ADVANCE . THE COMMUNIST GUARDS OBEDIENTLY RAISED THE FIRST OF THREE BARRIERS . BUT WHAT WAS A BUS DOING ON EMERGENCY DUTY ? SUDDENLY THE SHOOTING BEGAN TOO LATE . WAGNER, AT 40 M.P.H., WAS ALREADY CRASHING THROUGH THE SECOND BARRIER 100 YARDS AHEAD, THEN THE THIRD, ONLY 20 YARDS AWAY . ITS WINDSHIELD SMASHED, ITS PASSENGERS SHAKEN, ITS CARGO OF COAL AND POTATOES IN EVERY CORNER OF THE CAB, THE OLD BUS FINALLY LURCHED TO A STOP A FEW MILES DOWN THE ROAD WHERE THE COMMUNISTS NO LONGER MATTERED AT THE U.S . CHECKPOINT, A FOOT OR TWO INSIDE WEST BERLIN .
*TEXT 020 01/04/63 PAGE 021
THE ROAD TO JAIL IS PAVED WITH NONOBJECTIVE ART SINCE THE
KREMLIN'S SHARPEST BARBS THESE DAYS ARE AIMED AT MODERN ART AND "
WESTERN ESPIONAGE, " IT WAS JUST A MATTER OF TIME BEFORE THE KGB'S COPS WOULD TURN UP A VICTIM WHOSE WRONGDOINGS COMBINED BOTH EVILS . HE TURNED OUT TO BE A LENINGRAD PHYSICS TEACHER WHOSE TASTE FOR ABSTRACT PAINTING ALLEGEDLY LED HIM TO JOIN THE U.S . SPY SERVICE . POLICE SAID THEY FIRST SPOTTED THE TEACHER, ONE RUDOLF FRIEDMAN, AS HE MUTTERED UNCOMPLIMENTARY REMARKS ABOUT SOCIALIST REALISM WHILE STROLLING THROUGH LENINGRAD'S RUSSIAN MUSEUM . A WELL-DRESSED U.S . TOURIST APPROACHED HIM, ENTHUSIASTICALLY SHARED HIS SENTIMENTS, AND PROMISED TO SEND FRIEDMAN REPRODUCTIONS OF AVANT-GARDE PAINTINGS FROM AMERICA . THE PICTURE FRIEDMAN LIKED BEST, SAID THE COPS INDIGNANTLY, WAS A " CHAOS OF BLACK, RED AND BLUE SPLOTCHES CAPTIONED I NEED YOU TONIGHT . " SOON, THEY SAID, THE TEACHER WAS GETTING MESSAGES FROM THE U.S . WRITTEN IN INVISIBLE INK . JUST AS FRIEDMAN PREPARED TO DELIVER INFORMATION " VERY REMOTE FROM THEORETICAL ARGUMENTS ABOUT ABSTRACT ART, " POLICEMOVED IN AND HUSTLED HIM OFF TO JAIL .

ตัวอย่างข้อสอบถาม

*FIND 1
KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP
SUPPRESSING THE BUDDHISTS .
*FIND 2

EFFORTS OF AMBASSADOR HENRY CABOT LODGE TO GET VIET NAM'S
PRESIDENT DIEM TO CHANGE HIS POLICIES OF POLITICAL REPRESSION .
*FIND 3
NUMBER OF TROOPS THE UNITED STATES HAS STATIONED IN SOUTH
VIET NAM AS COMPARED WITH THE NUMBER OF TROOPS IT HAS STATIONED
IN WEST GERMANY .
*FIND 4
U.S . POLICY TOWARD THE NEW REGIME IN SOUTH VIET NAM WHICH OVERTHREW
PRESIDENT DIEM .
*FIND 5
PERSONS INVOLVED IN THE VIET NAM COUP .
*FIND 6
CEREMONIAL SUICIDES COMMITTED BY SOME BUDDHIST MONKS IN SOUTH VIET NAM
AND WHAT THEY ARE SEEKING TO GAIN BY SUCH ACTS .
*FIND 7
REJECTION BY PRINCE NORODOM SIHANOUK, AN ASIAN NEUTRALIST LEADER,
OF ALL FURTHER U.S . AID TO HIS NATION .
*FIND 8
U.N . TEAM SURVEY OF PUBLIC OPINION IN NORTH BORNEO AND SARAWAK ON
THE QUESTION OF JOINING THE FEDERATION OF MALAYSIA .
*FIND 9
OPPOSITION OF INDONESIA TO THE NEWLY-CREATED MALAYSIA .
*FIND 10
GROWING CONTROVERSY IN SOUTHEAST ASIA OVER THE PROPOSED
CREATION OF A FEDERATION OF MALAYSIA .
*FIND 11
ARRANGEMENTS FOR INDONESIA TO TAKE OVER THE ADMINISTRATION
OF WEST IRIAN, WHICH HAS BEEN UNDER UNITED NATIONS ADMINISTRATION .
*FIND 12
CONTROVERSY BETWEEN INDONESIA AND MALAYA ON THE PROPOSED
FEDERATION OF MALAYSIA, WHICH WOULD UNITE FIVE TERRITORIES .

ภาคผนวก จ

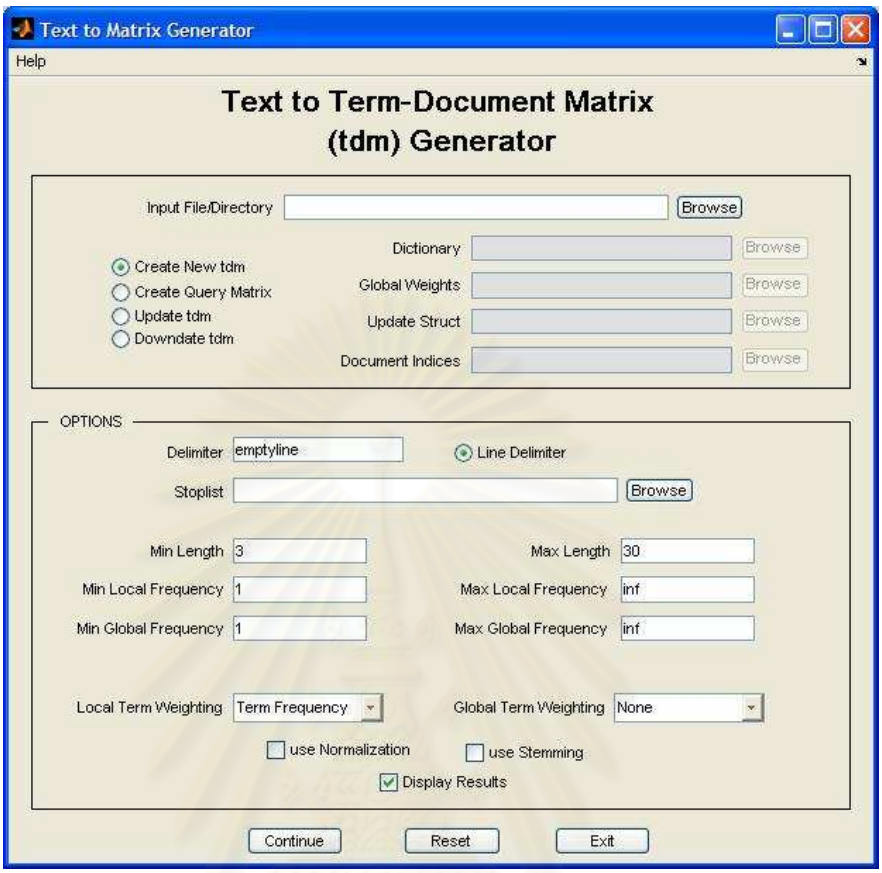
โปรแกรมทีเอ็มจี TMG (A MATLAB Toolbox for generating term-document matrices from text collections)

โปรแกรมทีเอ็มจี (TMG) เวอร์ชั่น 2.0R3.0 (Dimitrios and Gallopoulos, 2005) คือ โปรแกรมสร้างเวกเตอร์ของเอกสารและข้อสอบถาม ซึ่งโปรแกรมทีเอ็มจี (TMG) ทำงานบนโปรแกรมแม็ทแลบ เวอร์ชั่น 6.5 (MATLAB version 6.5) แล้วดำเนินงาน (Run) แฟ้มข้อมูล tmg_gui ที่เป็นแฟ้มข้อมูลชนิดเอ็ม (M-file)

งานวิจัยนี้จะให้โปรแกรมทีเอ็มจี (TMG) ในการสร้างเวกเตอร์ให้กับบทความของนิตยสารไทม์ (TIME Magazine) และข้อสอบถามที่มีการกำหนดให้เป็นเอกสารที่ใช้ทดสอบระบบการค้นคืนเอกสาร ซึ่งขั้นตอนการสร้างเวกเตอร์ให้กับเอกสารและข้อสอบถามดังกล่าวจะต้องสร้างเวกเตอร์ของเอกสารก่อนแล้วจึงทำการสร้างเวกเตอร์ข้อสอบถามดังขั้นตอนต่อไปนี้

ขั้นตอนการสร้างเวกเตอร์ให้กับเอกสาร

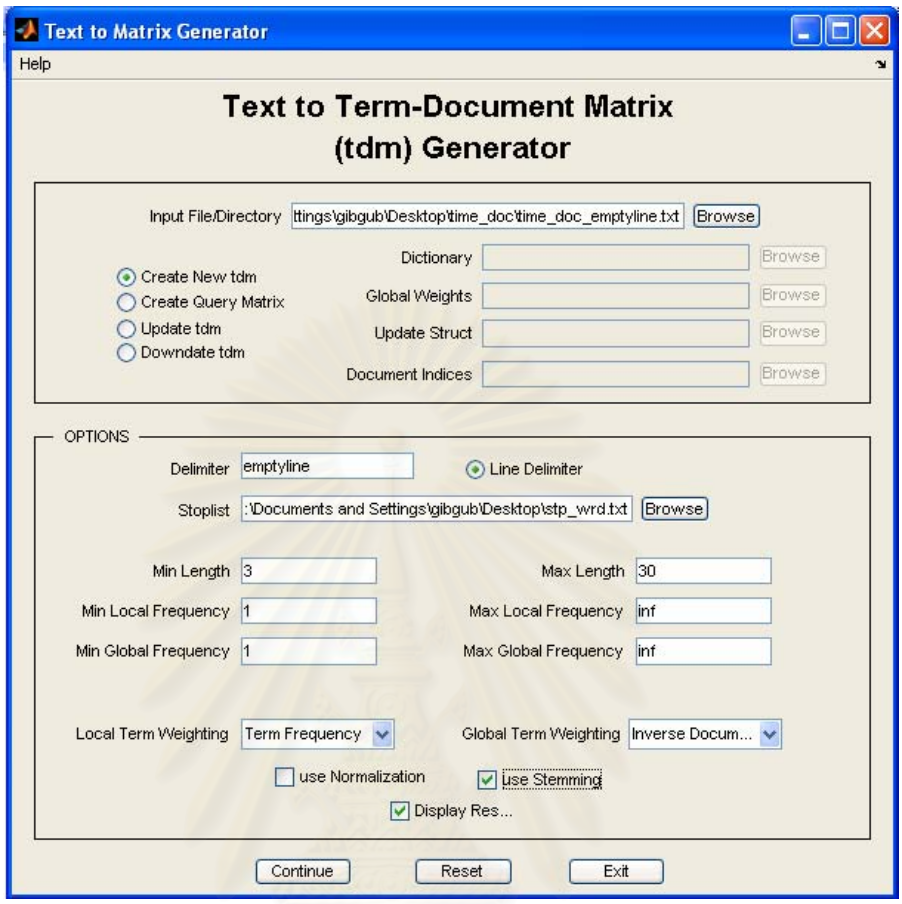
ขั้นตอนที่ 1 เลือกการทำงานที่ต้องการจากปุ่มตัวเลือก (Radio Button) ซึ่งจะมี 4 ลักษณะการทำงาน ดังรูป จ.1 ซึ่งจะเลือกการสร้างเวกเตอร์เอกสารขึ้นใหม่ โดยจะเลือกปุ่มตัวเลือก (Radio Button) สร้างเวกเตอร์เอกสาร (Create New tdm)



รูปที่ จ.1 รูปแสดงหน้าจอแรกของโปรแกรมทีเอ็มจี (TMG) เพื่อให้ผู้ใช้เลือกการทำงานที่ต้องการ

ขั้นตอนที่ 2 เลือกแฟ้มข้อมูลเอกสารที่ต้องการเปลี่ยนให้อยู่ในรูปแบบของเวกเตอร์โดยคลิกปุ่มค้นดู (Browse) เพื่อเลือกแฟ้มข้อมูลในช่องของรับเข้าปฟ้มข้อมูลหรือสารบบ (Input File/Directory)

ขั้นตอนที่ 3 เลือกแฟ้มข้อมูลคำยกเว้นที่ไม่พิจารณาโดยคลิกปุ่มค้นดู (Browse) เพื่อเลือกแฟ้มข้อมูลในช่องของรายการคำหยุด (Stoplist) ดังรูปที่ จ.2



รูปที่ จ.2 รูปแสดงหน้าจอกำหนดคุณสมบัติในการสร้างเวกเตอร์เอกสาร

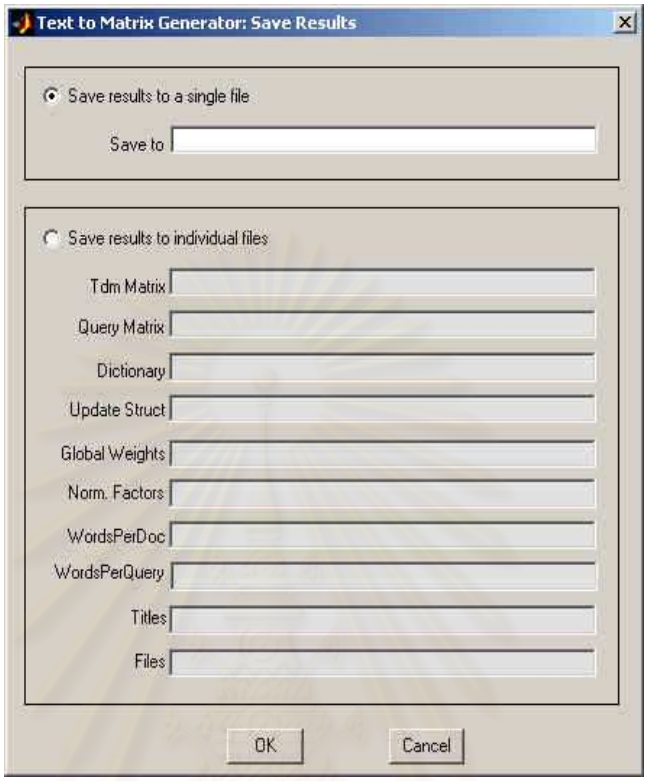
ขั้นตอนที่ 4 กำหนดคุณสมบัติต่างๆ โดยจากรูปจะมีการกำหนดดังนี้

- แยกแต่ละบทความโดยใช้บรรทัดว่างเป็นตัวแบ่ง
- กำหนดค่าน้ำหนักโดยใช้ค่าความถี่ของคำและค่าความถี่ของเอกสารแบบผกผัน (tf-idf)
- ทำการลดรูปคำ (Stemming)

ขั้นตอนที่ 5 ดำเนินการให้ระบบออกผลลัพธ์ โดยคลิกปุ่มดำเนินต่อไป (Continue) จะปรากฏหน้าจอดังรูปที่ จ.3

รูปที่ จ.3 รูปแสดงหน้าจอยืนยันการบันทึกผลลัพธ์ลงแฟ้มข้อมูล

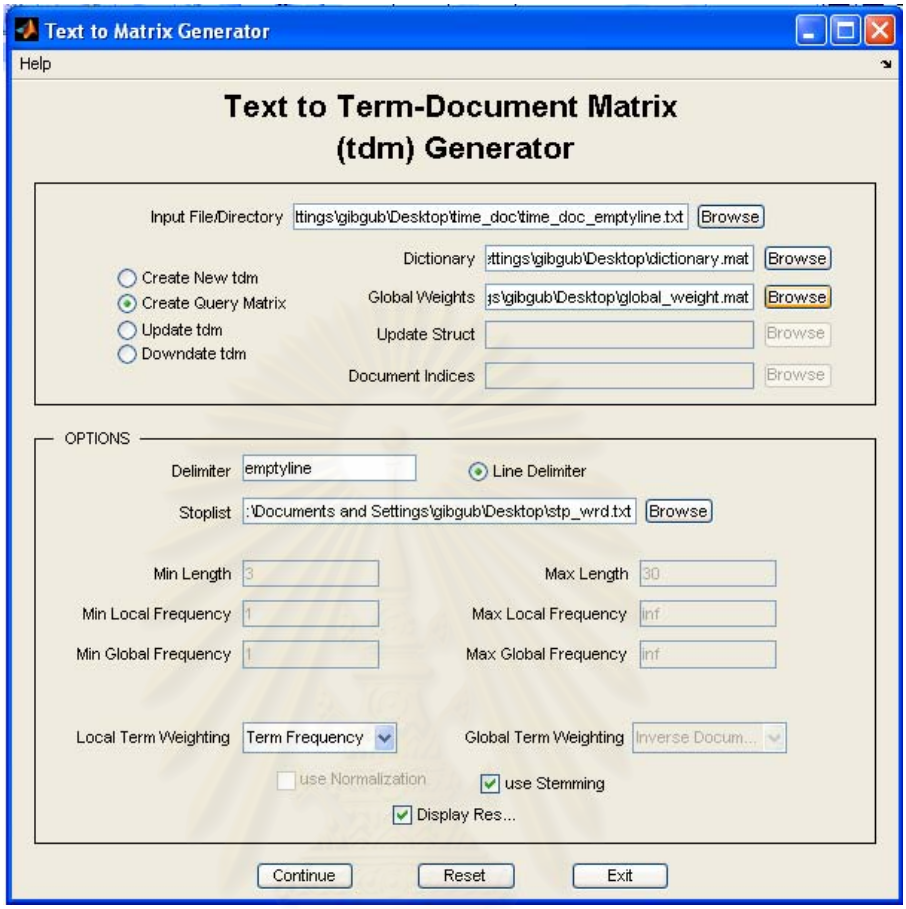
ขั้นตอนที่ 6 บันทึกผลลัพธ์ไปยังตำแหน่งที่ต้องการ



รูปที่ จ.4 รูปแสดงหน้าจอในการบันทึกผลลัพธ์ลงในแฟ้มข้อมูลที่กำหนด

ขั้นตอนการสร้างเวกเตอร์ให้กับข้อสอบถาม

ขั้นตอนที่ 1 เลือกการทำงานที่ต้องการจากปุ่มตัวเลือก (Radio Button) ซึ่งจะมี 4 ลักษณะการทำงาน ดังรูปที่ จ.5 ซึ่งจะเลือกการสร้างเวกเตอร์ข้อสอบถาม โดยจะเลือกปุ่มตัวเลือกสร้างเวกเตอร์เอกสาร (Radio Button Create Query Metrix)



รูปที่ จ.5 รูปแสดงหน้าจอเลือกกำหนดคุณสมบัติต่างๆในการสร้างเวกเตอร์ข้อสอบถาม

ขั้นตอนที่ 2 เลือกแฟ้มข้อมูลดิกชันนารี (Dictionary) และแฟ้มข้อมูลค่าน้ำหนักครอบคลุม (Global Weight) จากการสร้างเวกเตอร์เอกสารดังที่กล่าวมาแล้ว

ขั้นตอนที่ 3 เลือกแฟ้มข้อมูลคำยกเว้นที่ไม่พิจารณาโดยคลิกปุ่มค้นดู (Browse) เพื่อเลือกแฟ้มข้อมูลในช่องของรายการคำหยุด (Stoplist)

ขั้นตอนที่ 4 กำหนดคุณสมบัติต่างๆ โดยจากรูปจะมีการกำหนดดังนี้

- แยกแต่ละบทความโดยใช้บรรทัดว่างเป็นตัวแบ่ง
- กำหนดค่าน้ำหนักโดยใช้ค่าความถี่ของคำและค่าความถี่ของเอกสารแบบผกผัน (tf-idf)
- ทำการลดรูปคำ (Stemming)

ขั้นตอนที่ 5 ดำเนินการให้ระบบออกผลลัพธ์ โดยคลิกปุ่มดำเนินต่อไป (Continue)

ขั้นตอนที่ 6 บันทึกผลลัพธ์ไปยังตำแหน่งที่ต้องการดังเช่นขั้นตอนการสร้างเวกเตอร์ให้กับเอกสาร

ภาคผนวก ฉ

โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1)

โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) คือ โปรแกรมที่ช่วยการทำเหมืองข้อมูล ซึ่งงานวิจัยนี้จะใช้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) นี้ช่วยในการหาความสัมพันธ์ของคำในเวกเตอร์เอกสารนิตยสารไทม์ (TIME Magazine) ที่ใช้ทดสอบระบบการค้นคืนเอกสารที่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วม โดยมีขั้นตอนดังต่อไปนี้

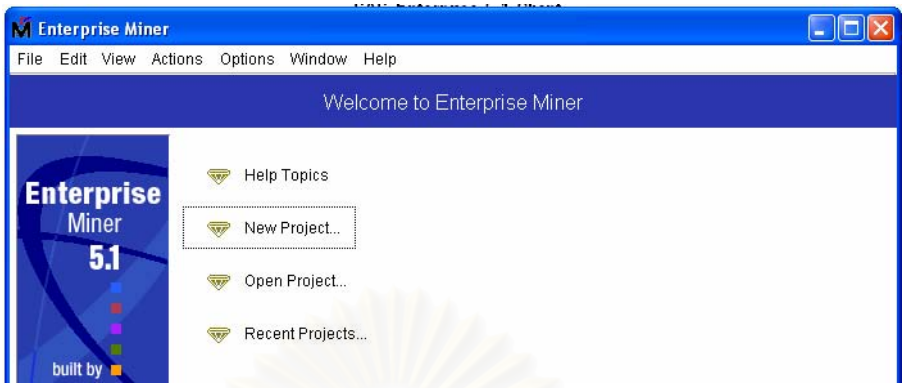
ขั้นตอนที่ 1 เปิดโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ที่โปรแกรม ดังรูปที่ ฉ.1

รูปที่ ฉ.1 รูปแสดงลักษณะโปรแกรมของแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise 5.1 Client)

ขั้นตอนที่ 2 เมื่อคลิกเปิดโปรแกรมแล้วจะปรากฏหน้าจอให้เลือกประเภทผู้ใช้ ดังรูปที่ ฉ. 2 แล้วคลิกปุ่มเริ่ม (Start) ในที่นี้ผู้วิจัยเลือกสถานีงานส่วนบุคคล (Personal Workstation)

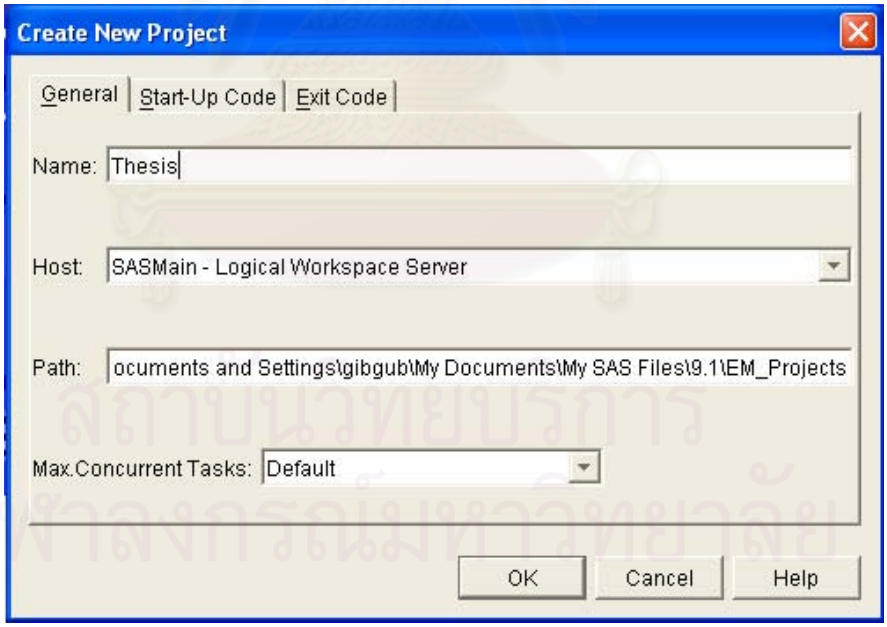
รูปที่ ฉ.2 รูปแสดงหน้าจอเลือกประเภทผู้ใช้ของโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1)

ขั้นตอนที่ 3 เมื่อเลือกประเภทผู้ใช้แล้วจะปรากฏหน้าจอการทำงานดังรูป ฉ.3



รูปที่ ฉ.3 รูปแสดงหน้าจอการทำงานแรกของโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1)

ขั้นตอนที่ 4 จากนั้นเลือกสร้างโครงการใหม่ (New Project) เพื่อสร้างโครงการขึ้นมาใหม่ ซึ่งจะปรากฏหน้าจอ ดังรูป ฉ.4 จากนั้นพิมพ์ชื่อโครงการที่ต้องการสร้างขึ้นมาใหม่และกำหนดรายละเอียดโครงการอื่นๆ



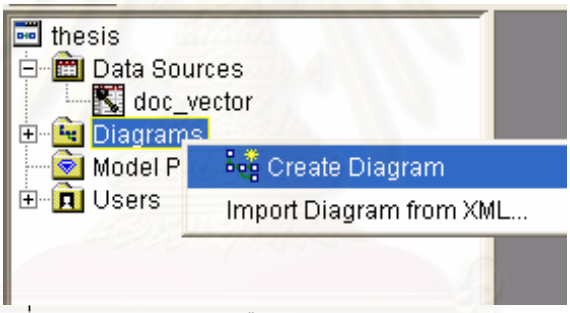
รูปที่ ฉ.4 รูปแสดงหน้าจอกำหนดรายละเอียดโครงการ

ขั้นตอนที่ 5 เมื่อสร้างโครงการเรียบร้อยแล้ว จากนั้นสร้างอุปกรณ์ส่งข้อมูล (Data Source) โดยการคลิกขวาที่โฟลเดอร์อุปกรณ์ส่งข้อมูล (Folder Data Sources) แล้วเลือกสร้างอุปกรณ์ส่งข้อมูล (Create Data Source) ดังรูป ฉ.5

รูปที่ ฉ.5 รูปแสดงการเลือกสร้างอุปกรณ์ส่งข้อมูล (Data Source)

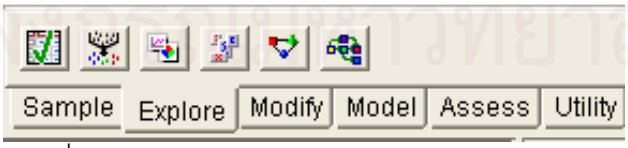
ขั้นตอนที่ 6 เมื่อเลือกสร้างอุปกรณ์ส่งข้อมูล (Create Data Source) แล้ว ต่อจากนั้น กำหนดรายละเอียดของอุปกรณ์ส่งข้อมูล (Data source) นั้น ๆ โดยการเลือกตาราง เมื่อเสร็จแล้ว จะปรากฏอุปกรณ์ส่งข้อมูล (Data Source) ในโฟลเดอร์อุปกรณ์ส่งข้อมูล (Folder Data Sources)

ขั้นตอนที่ 7 จากนั้นสร้างแผนภาพ (Diagram) โดยคลิกขวาที่แผนภาพ (Diagram) แล้ว เลือกสร้างแผนภาพ (Create Diagram) ดังรูป ฉ.6



รูปที่ ฉ.6 รูปแสดงการเลือกสร้างแผนภาพ (Diagram)

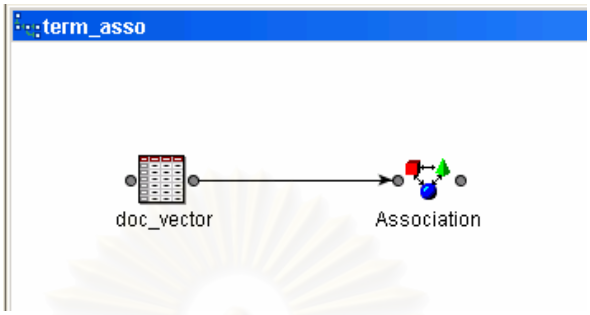
ขั้นตอนที่ 8 เมื่อสร้างแผนภาพ (Diagram) เสร็จแล้วจะปรากฏหน้าต่างพื้นที่ว่าง ๆ จากนั้นลากอุปกรณ์ส่งข้อมูล (Data Source) ที่สร้างไว้มาวางที่พื้นที่นั้น เนื่องจากงานวิจัยนี้ ต้องการหากฎความสัมพันธ์ของคำ ดังนั้นเลือกแท็บค้นหา (Tab Explore) ดังรูปที่ ฉ.7



รูปที่ ฉ.7 รูปแสดงหน้าจอเลือกแบบจำลอง (Model)

ขั้นตอนที่ 9 จากนั้นลากแบบจำลองกฎความสัมพันธ์ (Model Association) มาวางไว้ บนพื้นที่ใกล้ ๆ กับอุปกรณ์ส่งข้อมูล (Data Source) ที่ลากมาวางก่อนหน้านี้ แล้วลากเส้นเชื่อม

ระหว่างอุปกรณ์ส่งข้อมูล (Data Source) แบบจำลองกฎความสัมพันธ์ (Model Association) ดังรูปที่ ฉ.8



รูปที่ ฉ.8 รูปแสดงหน้าจอการสร้างแผนภาพ (Create Diagram)

ขั้นตอนที่ 10 เมื่อสร้าง Diagram แล้ว จากนั้นให้ดำเนินงาน (Run) โปรแกรม โดยคลิกที่ปุ่มดำเนินงาน (Run)  และเมื่อดำเนินงานแผนภาพ (Run Diagram) เสร็จเรียบร้อย เมื่อต้องการดูผลลัพธ์จากการดำเนินงานแผนภาพ (Run Diagram) ให้แสดงผลลัพธ์ที่ได้ โดยการคลิกที่ปุ่มแสดงผลลัพธ์ (Result)  จะปรากฏหน้าต่างแสดงผลลัพธ์ขึ้นมา

ภาคผนวก ช

กฎความสัมพันธ์ของคำ

จากการตั้งค่าสนับสนุนต่ำที่สุด (Minimum Support) ไว้ที่ค่า 1.6471 และค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) ไว้ที่ค่า 70 เปอร์เซนต์ โปรแกรม SAS Exterprise Miner 5.1 จะค้นหากฎความสัมพันธ์ที่ค้นหาออกมาได้ในงานวิจัยนี้ โดยผู้วิจัยจะนำมาเฉพาะกฎความสัมพันธ์ที่มีค่าลิฟท์ (Lift) สูง ๆ ออกมา จำนวน 136 กฎความสัมพันธ์ โดยสามารถแสดงผลลัพธ์กฎความสัมพันธ์ในตารางต่อไปนี้ โดยที่จะเรียงลำดับกฎความสัมพันธ์จากค่าลิฟท์ (Lift) มากไปน้อย

ตารางที่ ช.1 ตารางแสดงกฎความสัมพันธ์ที่คัดเลือกออกมาได้

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
100.0000	1.6471	60.7143	phouma	->	souvanna
100.0000	1.6471	60.7143	souvanna	->	phouma
100.0000	1.6471	53.1250	imam	->	royalist
100.0000	1.8824	53.1250	pietro	->	sinistra
100.0000	1.8824	53.1250	apertura	->	sinistra
87.5000	1.6471	53.1250	pathet	->	souvanna
87.5000	1.6471	53.1250	rahman	->	Tunku
100.0000	1.8824	53.1250	amintor	->	fanfani
100.0000	1.8824	53.1250	fanfani	->	amintor
87.5000	1.6471	53.1250	royalist	->	Imam
100.0000	1.8824	53.1250	sinistra	->	apertura
100.0000	1.8824	53.1250	pietro	->	apertura
100.0000	1.6471	53.1250	souvanna	->	pathet
100.0000	1.6471	53.1250	phouma	->	pathet

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
87.5000	1.6471	53.1250	pathet	->	phouma
100.0000	1.8824	53.1250	sinistra	->	pietro
100.0000	1.8824	53.1250	apertura	->	pietro
100.0000	1.6471	53.1250	tunku	->	rahman
100.0000	2.1176	47.2222	borneo	->	sarawak
100.0000	1.8824	47.2222	rahman	->	sarawak
100.0000	1.6471	47.2222	tunku	->	sarawak
77.7778	1.6471	47.2222	laotian	->	souvanna
77.7778	1.6471	47.2222	sarawak	->	Tunku
77.7778	1.6471	47.2222	borneo	->	Tunku
100.0000	2.1176	47.2222	sarawak	->	borneo
100.0000	1.8824	47.2222	rahman	->	borneo
100.0000	1.6471	47.2222	tunku	->	borneo
88.8889	1.8824	47.2222	nkrumah	->	kwame
100.0000	1.6471	47.2222	souvanna	->	laotian
100.0000	1.6471	47.2222	phouma	->	laotian
100.0000	1.8824	47.2222	kwame	->	nkrumah
77.7778	1.6471	47.2222	laotian	->	phouma
88.8889	1.8824	47.2222	sarawak	->	rahman
88.8889	1.8824	47.2222	borneo	->	rahman
87.5000	1.6471	46.4844	fanfani	->	sinistra
87.5000	1.6471	46.4844	amintor	->	sinistra
87.5000	1.6471	46.4844	sinistra	->	fanfani
87.5000	1.6471	46.4844	pietro	->	fanfani

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
87.5000	1.6471	46.4844	apertura	->	fanfani
87.5000	1.6471	46.4844	sinistra	->	amintor
87.5000	1.6471	46.4844	pietro	->	amintor
87.5000	1.6471	46.4844	apertura	->	amintor
87.5000	1.6471	46.4844	fanfani	->	apertura
87.5000	1.6471	46.4844	amintor	->	apertura
87.5000	1.6471	46.4844	fanfani	->	pietro
87.5000	1.6471	46.4844	amintor	->	pietro
70.0000	1.6471	42.5000	abdullah	->	Sallal
80.0000	1.8824	42.5000	nenni	->	sinistra
70.0000	1.6471	42.5000	ahm	->	Bella
70.0000	1.6471	42.5000	malaysia	->	Tunku
100.0000	1.6471	42.5000	bella	->	ahm
80.0000	1.8824	42.5000	nenni	->	apertura
100.0000	1.6471	42.5000	kassem	->	iraqi
70.0000	1.6471	42.5000	iraqi	->	kassem
100.0000	1.8824	42.5000	rahman	->	malaysia
100.0000	1.6471	42.5000	tunku	->	malaysia
100.0000	1.8824	42.5000	sinistra	->	nenni
100.0000	1.8824	42.5000	pietro	->	nenni
100.0000	1.8824	42.5000	apertura	->	nenni
100.0000	1.6471	42.5000	sallal	->	abdullah
80.0000	1.8824	42.5000	nenni	->	pietro
80.0000	1.8824	42.5000	malaysia	->	rahman

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
88.8889	1.8824	41.9753	brunei	->	sarawak
88.8889	1.8824	41.9753	brunei	->	borneo
88.8889	1.8824	41.9753	sarawak	->	brunei
88.8889	1.8824	41.9753	borneo	->	brunei
87.5000	1.6471	41.3194	rahman	->	brunei
77.7778	1.6471	41.3194	pagoda	->	hue
87.5000	1.6471	41.3194	pathet	->	laotian
87.5000	1.6471	41.3194	rahman	->	malaya
87.5000	1.6471	41.3194	hue	->	pagoda
77.7778	1.6471	41.3194	laotian	->	pathet
77.7778	1.6471	41.3194	malaya	->	rahman
77.7778	1.6471	41.3194	brunei	->	rahman
100.0000	2.5882	38.6364	cyril	->	adoula
100.0000	2.5882	38.6364	adoula	->	cyril
80.0000	1.8824	37.7778	malaysia	->	sarawak
80.0000	1.8824	37.7778	malaysia	->	borneo
80.0000	1.8824	37.7778	malaysia	->	brunei
88.8889	1.8824	37.7778	persian	->	gulf
88.8889	1.8824	37.7778	sarawak	->	malaysia
88.8889	1.8824	37.7778	brunei	->	malaysia
88.8889	1.8824	37.7778	borneo	->	malaysia
80.0000	1.8824	37.7778	gulf	->	persian
87.5000	1.6471	37.1875	indonesian	->	sukarno
87.5000	1.6471	37.1875	churchil	->	winston

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
70.0000	1.6471	37.1875	winston	->	churchil
70.0000	1.6471	37.1875	nenni	->	fanfani
70.0000	1.6471	37.1875	nenni	->	amintor
70.0000	1.6471	37.1875	sukarno	->	indonesian
87.5000	1.6471	37.1875	fanfani	->	nenni
87.5000	1.6471	37.1875	amintor	->	nenni
77.7778	1.6471	36.7284	malaya	->	sarawak
77.7778	1.6471	36.7284	malaya	->	borneo
77.7778	1.6471	36.7284	malaya	->	brunei
77.7778	1.6471	36.7284	sarawak	->	malaya
77.7778	1.6471	36.7284	brunei	->	malaya
77.7778	1.6471	36.7284	borneo	->	malaya
100.0000	1.8824	35.4167	rusk	->	dean
100.0000	1.8824	35.4167	gaitskel	->	hugh
83.3333	2.3529	35.4167	baghdad	->	iraqi
100.0000	2.3529	35.4167	iraqi	->	baghdad
100.0000	1.6471	35.4167	kassem	->	baghdad
72.7273	1.8824	34.3434	singapor	->	brunei
72.7273	1.8824	34.3434	singapor	->	malaya
88.8889	1.8824	34.3434	malaya	->	singapor
88.8889	1.8824	34.3434	brunei	->	singapor
80.0000	1.8824	34.0000	elisabethvil	->	secessionist
80.0000	1.8824	34.0000	secessionist	->	elisabethvil
87.5000	1.6471	33.8068	rahman	->	singapor

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
87.5000	1.6471	33.8068	cabot	->	lodg
70.0000	1.6471	33.0556	malaysia	->	malaya
77.7778	1.6471	33.0556	malaya	->	malaysia
100.0000	3.0588	32.6923	mois	->	tshomb
100.0000	2.3529	32.6923	secessionist	->	tshomb
100.0000	2.3529	32.6923	elisabethvil	->	tshomb
100.0000	3.0588	32.6923	tshomb	->	mois
100.0000	2.3529	32.6923	secessionist	->	mois
100.0000	2.3529	32.6923	elisabethvil	->	mois
100.0000	2.1176	32.6923	jawaharl	->	nehru
76.9231	2.3529	32.6923	tshomb	->	secessionist
76.9231	2.3529	32.6923	mois	->	secessionist
76.9231	2.3529	32.6923	tshomb	->	elisabethvil
76.9231	2.3529	32.6923	mois	->	elisabethvil
83.3333	2.3529	32.1970	wire	->	barb
90.9091	2.3529	32.1970	barb	->	wire
72.7273	1.8824	30.9091	singapor	->	malaysia
80.0000	1.8824	30.9091	malaysia	->	singapor
100.0000	1.6471	30.3571	yemeni	->	yemen
100.0000	1.6471	30.3571	sallal	->	yemen
100.0000	1.6471	30.3571	imam	->	yemen
100.0000	1.6471	30.3571	kenyatta	->	kenya
92.3077	2.8235	30.1775	baathist	->	baath
92.3077	2.8235	30.1775	baath	->	baathist

Confidence (%)	Support (%)	Lift	กฎความสัมพันธ์		
			Antecedent		Consequent
77.7778	1.6471	30.0505	sarawak	->	singapor
77.7778	1.6471	30.0505	borneo	->	singapor

ภาคผนวก ซ

การออกแบบการทำงานของเครื่องมือทดสอบ

ในการออกแบบการทำงานเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 3 รูปแบบ ดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์โดยไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้

2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

3) การค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

โดยจะออกแบบในส่วนสถาปัตยกรรมของเครื่องมือทดสอบ (Software Architecture) แผนภาพการไหลข้อมูล (Data Flow Diagram) การออกแบบการทำงานของเครื่องมือ (System Domain Design) การออกแบบฐานข้อมูล (Database Design) การออกแบบหน้าจอ (Interface Design) และการออกแบบการทดสอบ (Test Design) โดยมีรายละเอียดดังต่อไปนี้

ซ.1 สถาปัตยกรรมของเครื่องมือทดสอบ (Software Architecture)

เครื่องมือที่จะพัฒนาขึ้นได้ถูกออกแบบให้มีโครงสร้างในรูปแบบเลเยอร์ (Layer) 3 เลเยอร์ ประกอบด้วย

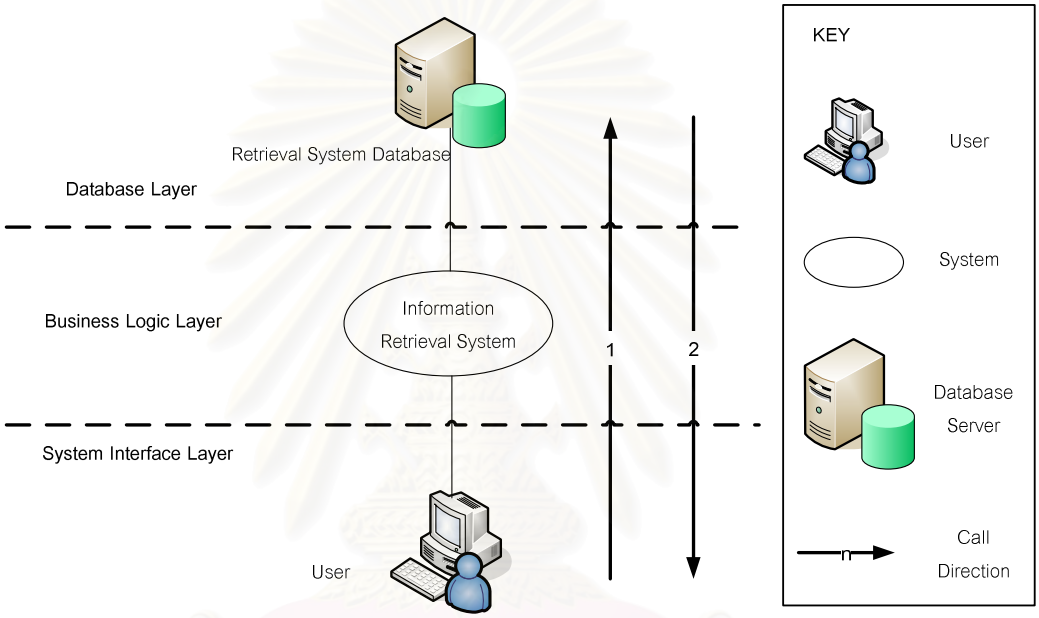
1) System Interface Layer เป็นส่วนของหน้าจอที่ติดต่อกับผู้ใช้เครื่องมือ
2) Business Logic Layer ซึ่งทำหน้าที่ค้นหาและประมวลผลข้อมูลหรือเงื่อนไขที่รับมาจาก System Interface Layer และ Database Layer
3) Database Layer เป็นส่วนจัดการระบบฐานข้อมูลเอกสาร

ผู้วิจัยเลือกใช้รูปแบบสถาปัตยกรรมนี้ เนื่องจากสถาปัตยกรรมเลเยอร์มีความเหมาะสมกับเครื่องมือที่จะพัฒนาและมีคุณสมบัติที่ดีหลายด้าน โดยระบบจะแบ่งหน้าที่การทำงานกันอย่างชัดเจนในแต่ละเลเยอร์ (Layer) ทำให้สามารถ (Bass et al., 1998)

1) ปรับปรุงแก้ไขระบบ (Modifiability) ได้ง่าย
2) นำแต่ละส่วนไปใช้ใหม่กับระบบอื่น ๆ ได้ (Reusability)

3) สามารถขยายขีดความสามารถได้ โดยไม่กระทบกับการทำงานส่วนอื่น (Scalability) เมื่อเพิ่มฟังก์ชันการทำงานเข้าไป

ผู้วิจัยได้ออกแบบเครื่องมือดังรูปที่ 3.1 โดยลำดับการทำงานของเครื่องมือนี้จะเริ่มจากที่ผู้ใช้ป้อนข้อสอบถามไปยังเครื่องมือเพื่อค้นคืนเอกสารจากฐานข้อมูล จากนั้นเครื่องมือจะค้นคืนเอกสารในฐานข้อมูลออกมาแสดงแก่ผู้ใช้ทางหน้าจอ

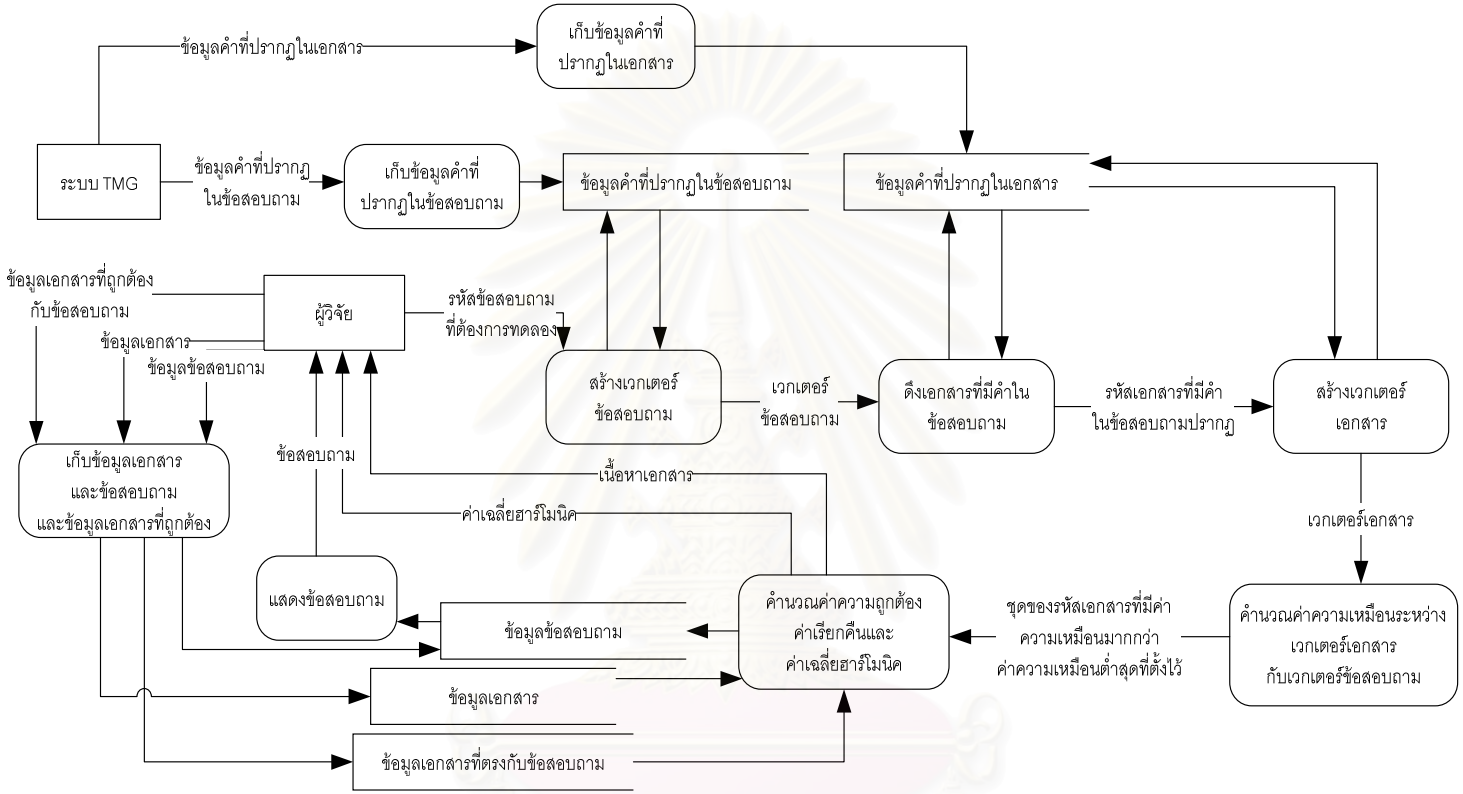


รูปที่ ซ.1 สถาปัตยกรรมของระบบแบบ 3 เลเยอร์

ซ.2 แผนภาพการไหลข้อมูล (Data Flow Diagram)

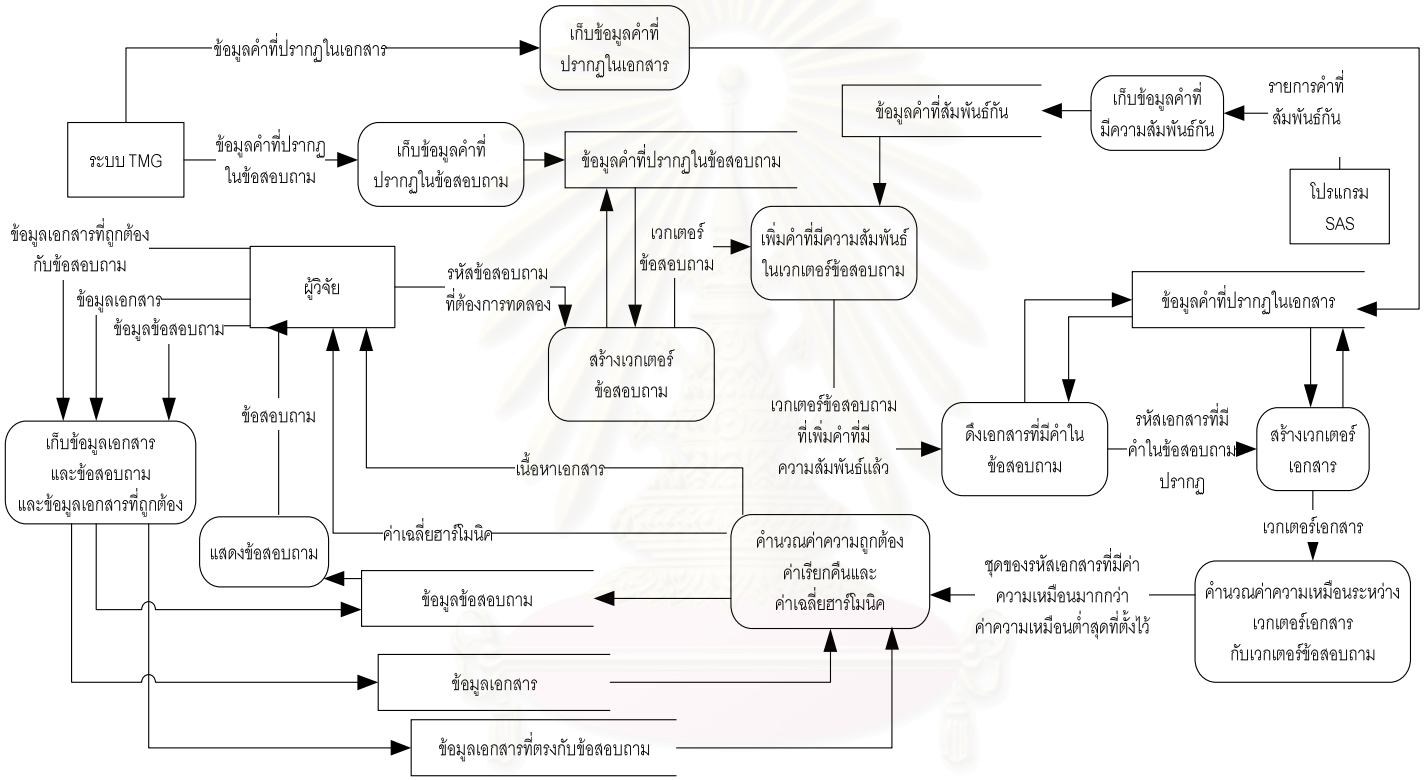
เป็นการออกแบบในส่วนของการไหลของข้อมูลในระบบทั้ง 3 รูปแบบ ซึ่งสามารถแสดงได้ดังรูปที่ ซ.2 ถึง ซ.7

รูปที่ ซ.2 รูปแสดงแผนภาพการไหลของข้อมูลบริบท (Context Diagram) ของการค้นคืนเอกสารระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ที่ไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้

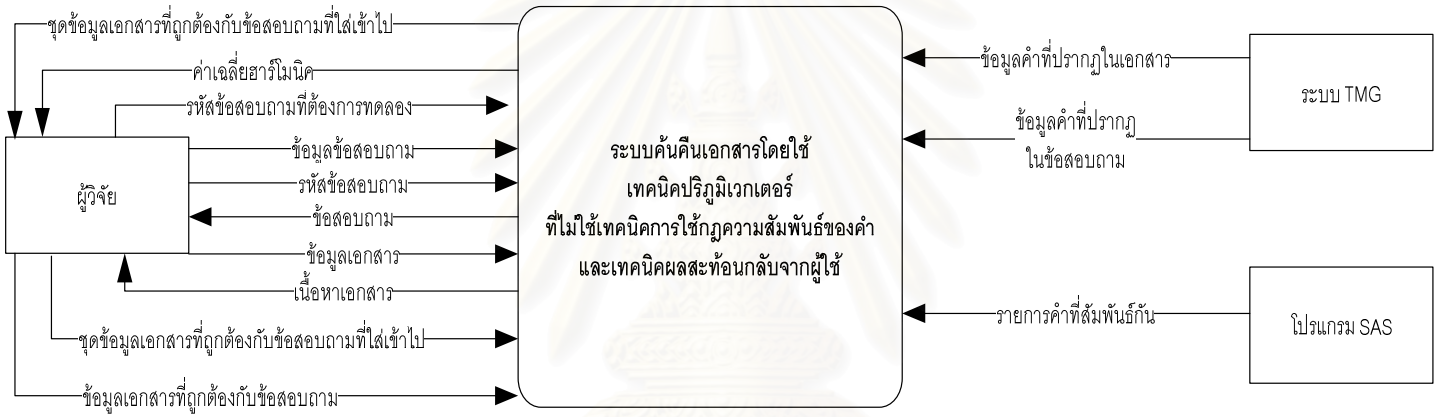


รูปที่ ซ.3 รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้

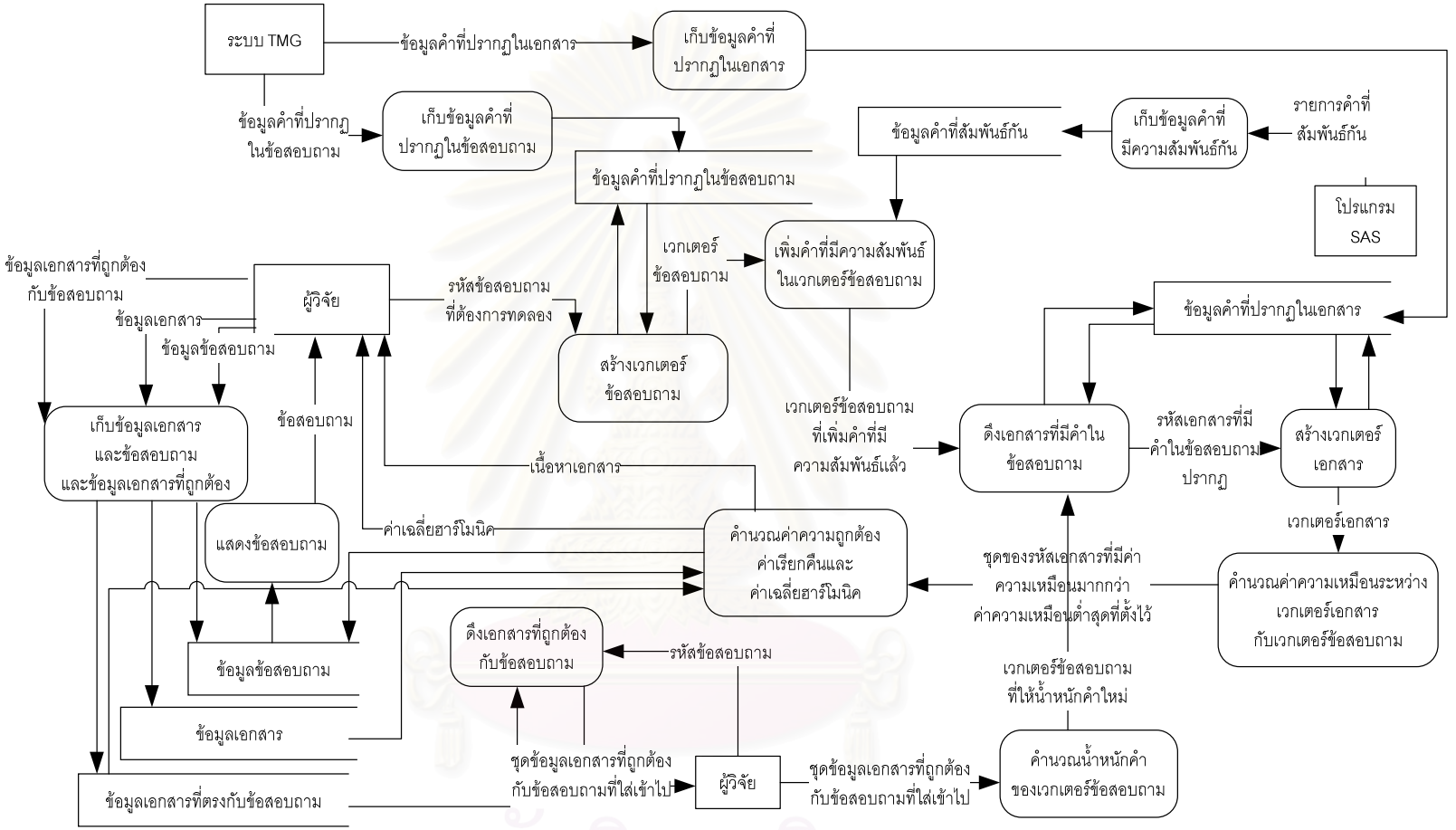
รูปที่ ซ.4 รูปแสดงแผนภาพการไหลของข้อมูลบริบท (Context Diagram) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ



รูปที่ ซ.5 รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ



รูปที่ ซ.6 รูปแสดงแผนภาพการไหลของข้อมูลบริบท (Context Diagram) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้



รูปที่ ซ.7 รูปแสดงแผนภาพการไหลของข้อมูลระดับที่ 2 (Data Flow Diagram Level 2) ของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

ซ.3 การออกแบบการทำงานของระบบ (System Domain Design)

เป็นส่วนการออกแบบการทำงานหลักของระบบค้นคืนเอกสารแบบต่าง ๆ โดยมีขั้นตอนการทำงานตั้งแต่เริ่มจนสิ้นสุดโปรแกรม ดังต่อไปนี้

1) ระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้

ขั้นตอนการทำงานของการค้นคืนเอกสารรูปแบบที่ 1 ซึ่งเป็นรูปแบบการใช้เทคนิคปริภูมิเวกเตอร์เท่านั้น สามารถแสดงได้ดังรูปที่ ซ.8 โดยเริ่มการทำงานจากเครื่องมือจะดึงข้อสอบถามจากฐานข้อมูลออกมาแสดงทางหน้าจอ จากนั้นผู้วิจัยจะกำหนดข้อสอบถามที่ต้องการทดสอบเข้ามายังระบบ ต่อมาเครื่องมือจะสร้างเวกเตอร์โดยดึงคำที่ปรากฏในข้อสอบถามนั้น ๆ จากฐานข้อมูลเพื่อไปดึงเอกสารที่มีคำในข้อสอบถามนั้นปรากฏอยู่แล้วเก็บลงแถวลำดับ (Array) เก็บเอกสารไว้ก่อน จากนั้นทำการสร้างเวกเตอร์เอกสารที่เก็บในแถวลำดับ (Array) เอกสารนั้น แล้วคำนวณหาค่าความเหมือนระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถามนั้น ๆ จนกว่าจะเอกสารในแถวลำดับ (Array) เก็บเอกสารจะหมด ถ้าค่าความเหมือนของเวกเตอร์เอกสารกับข้อสอบถามใดมีค่ามากกว่าค่าความเหมือนต่ำที่สุดที่ตั้งไว้จะเก็บเอกสารนั้นลงแถวลำดับ (Array) ที่เก็บเอกสารผลลัพธ์ เมื่อเครื่องมือได้เอกสารที่เป็นผลลัพธ์ทั้งหมดแล้ว เครื่องมือจะดึงผลลัพธ์ที่ถูกต้องของเอกสารนั้นตามที่กำหนดไว้แล้วโดยฐานข้อมูลนิตยสารไทม์ (TIME Collection) ในฐานข้อมูล จากนั้นนำรายการเอกสารที่เป็นผลลัพธ์มาเปรียบเทียบหาจำนวนเอกสารที่ถูกต้อง เพื่อคำนวณหาค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) แล้วเก็บลงฐานข้อมูล ส่วนค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้องจะทำการแสดงผลทางหน้าจอแก่ผู้วิจัยด้วย เป็นอันเสร็จสิ้นการค้นคืนเอกสารของข้อสอบถามหนึ่ง ๆ

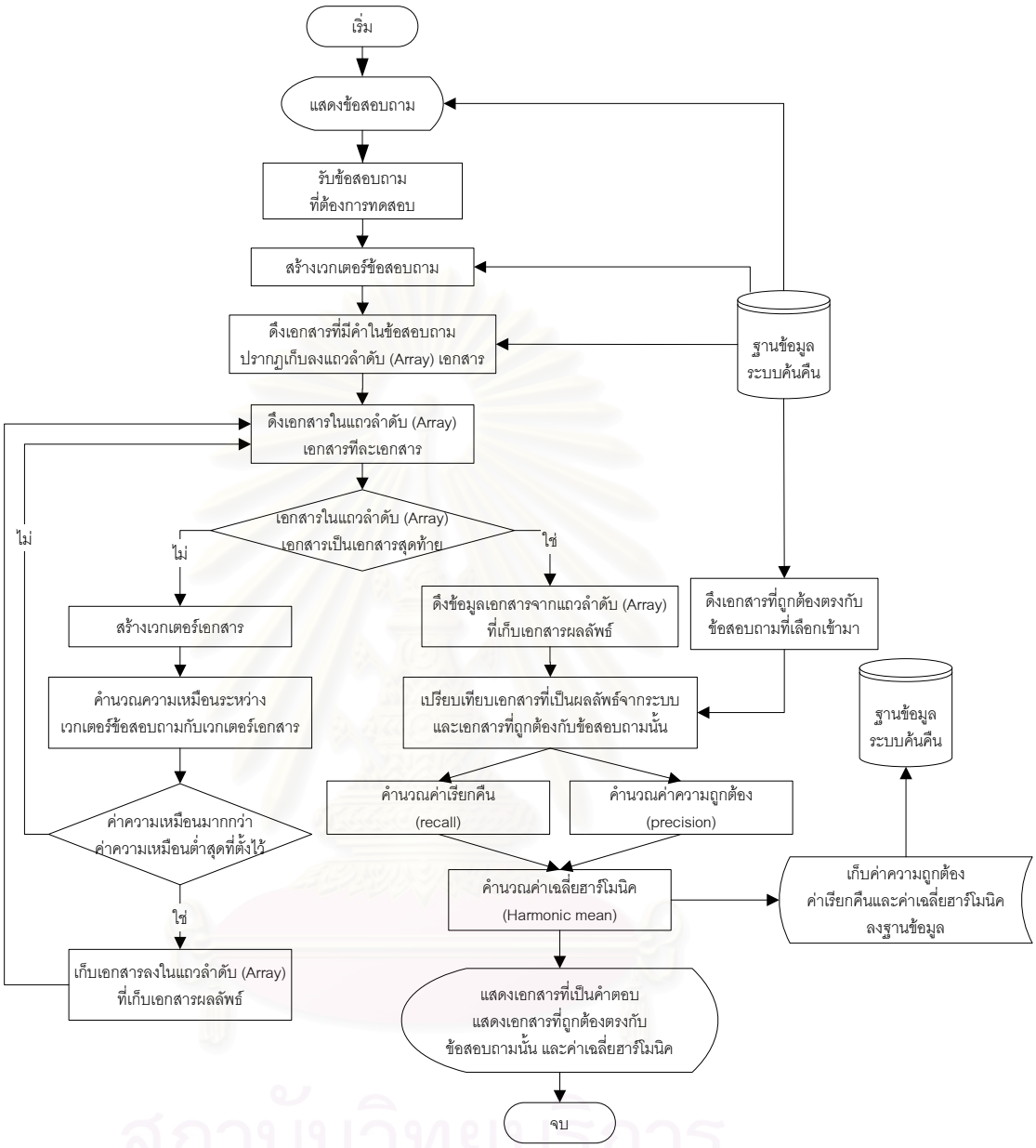
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

ขั้นตอนการทำงานของการค้นคืนเอกสารรูปแบบที่ 2 นี้ จะเพิ่มการทำงานจากการค้นคืนเอกสารรูปแบบที่ 1 ในส่วนที่อยู่ในกรอบสี่เหลี่ยมเส้นประดังรูปที่ ซ.9 โดยจะเป็นส่วนการทำงานของการหาคำที่มีความสัมพันธ์กับคำที่ปรากฏในข้อสอบถามที่ผู้วิจัยเลือกเข้ามา ก่อนที่จะนำคำเหล่านั้นไปสร้างเวกเตอร์ข้อสอบถาม โดยจะมีการคำนวณหาค่าน้ำหนักของคำที่มีความสัมพันธ์กับคำในเวกเตอร์ข้อสอบถามนั้นด้วย เนื่องจากคำที่มีความสัมพันธ์กับคำในข้อสอบถามนั้นยังไม่มีการกำหนดค่าน้ำหนักมาล่วงหน้าสำหรับเวกเตอร์ข้อสอบถามนั้น โดยจะคำนวณจากค่าน้ำหนักของคำในข้อสอบถามที่ผู้วิจัยเลือกเข้ามาคูณกับค่าความเชื่อมั่น

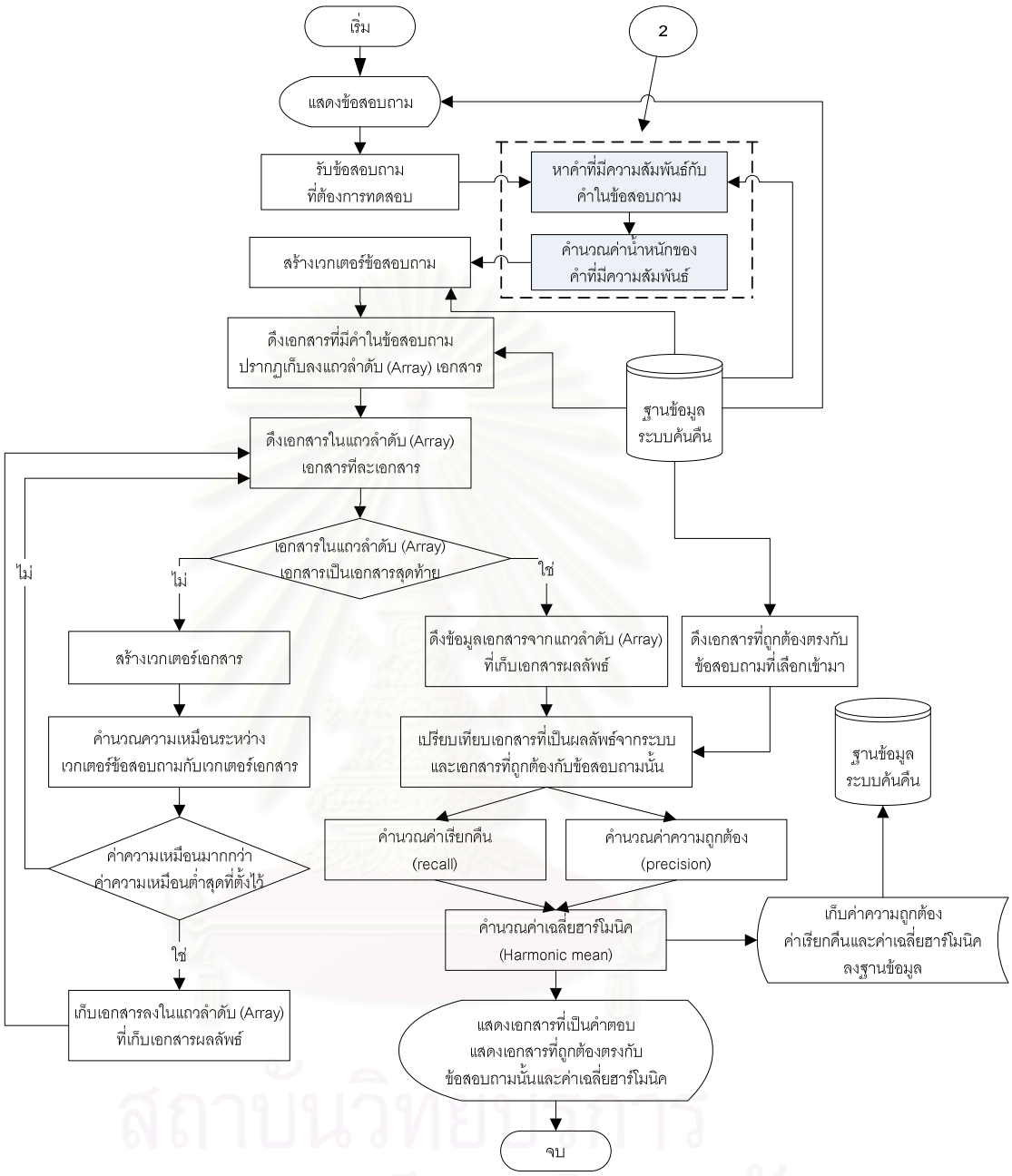
(Confidence) ของกฎความสัมพันธ์ของคำหารด้วยหนึ่งร้อยจะได้ค่าน้ำหนักของคำที่สัมพันธ์กับคำในข้อสอบถามที่ผู้วิจัยเลือกเข้ามาทดสอบ

3) การค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

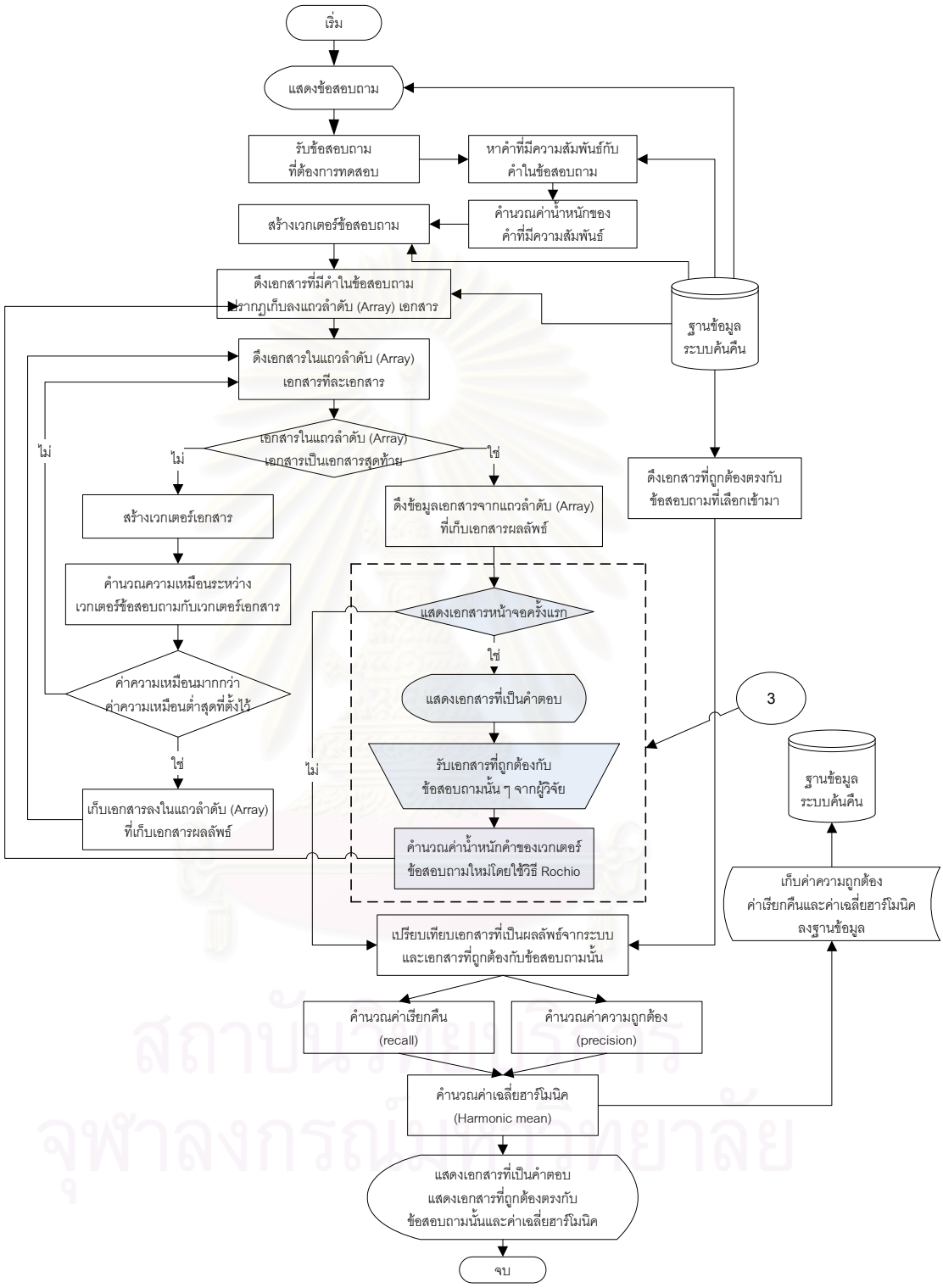
การค้นคืนเอกสารรูปแบบที่ 3 นี้ เช่นเดียวกับการค้นคืนเอกสารรูปแบบที่ 2 แต่เพิ่มเติมส่วนที่อยู่ในกรอบสี่เหลี่ยมเส้นประดังรูปที่ ซ.9 ซึ่งเป็นส่วนของการให้ผลสะท้อนกลับจากผู้วิจัย โดยผู้วิจัยจะเลือกเอกสารที่เกี่ยวเนื่องกับข้อสอบถามตามที่ฐานข้อมูลนิตยสารไทม์ (TIME Collection) กำหนดมาเข้ามายังเครื่องมือเพื่อให้เครื่องมือคำนวณค่าน้ำหนักคำในแต่ละมิติของเวกเตอร์ข้อสอบถามใหม่อีกครั้งตามสูตรของรอคชิโอ (Rochio) ซึ่งขั้นตอนการปรับน้ำหนักคำแต่ละมิติจะไม่พิจารณาถึงความสัมพันธ์ของคำแต่จะพิจารณาเพียงคำที่อยู่ในเอกสารที่เกี่ยวเนื่องและคำที่อยู่ในเอกสารที่ไม่เกี่ยวเนื่องตามที่ผู้วิจัยกำหนดเข้ามาเท่านั้น เช่น ถ้าข้อสอบถามที่ผู้วิจัยเลือกเข้ามามีคำ “a” แล้วคำ “a” มีความสัมพันธ์กับคำ “b” และ ”c” ดังนั้นเมื่อใช้เทคนิคกฎความสัมพันธ์ของคำแล้วข้อสอบถามจะมีคำ “a” ”b” และ ”c” ไปค้นคืนเอกสารออกมา เมื่อผู้วิจัยให้ผลสะท้อนกลับการปรับน้ำหนักคำ “a” ”b” และ ”c” จะเป็นอิสระต่อกันความสัมพันธ์ของทั้ง 3 คำไม่มีผลกระทบต่อกันนั่นคือ ถ้าผลสะท้อนกลับเอกสารที่เกี่ยวเนื่องมีเพียงคำ “a” เท่านั้น เครื่องมือจะปรับน้ำหนักคำ “a” จะไม่ปรับน้ำหนักคำ “b” และ “c” ตามความสัมพันธ์ที่มีการกำหนดไว้ เมื่อปรับค่าน้ำหนักคำในเวกเตอร์ข้อสอบถามแล้ว จากนั้นนำเวกเตอร์ข้อสอบถามใหม่ที่ได้นี้ไปเลือกเอกสารที่เกี่ยวเนื่องใหม่อีกครั้ง แล้วคำนวณหาค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ต่อไป โดยการให้ผลสะท้อนกลับนี้ ผู้วิจัยได้กำหนดให้ผู้วิจัยให้ผลสะท้อนกลับเพียงครั้งเดียวเท่านั้น



รูปที่ ซ.8 รูปแสดงขั้นตอนการทำงานของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคผลสะท้อนกลับจากผู้ใช้



รูปที่ ซ.9 รูปแสดงขั้นตอนการทำงานของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

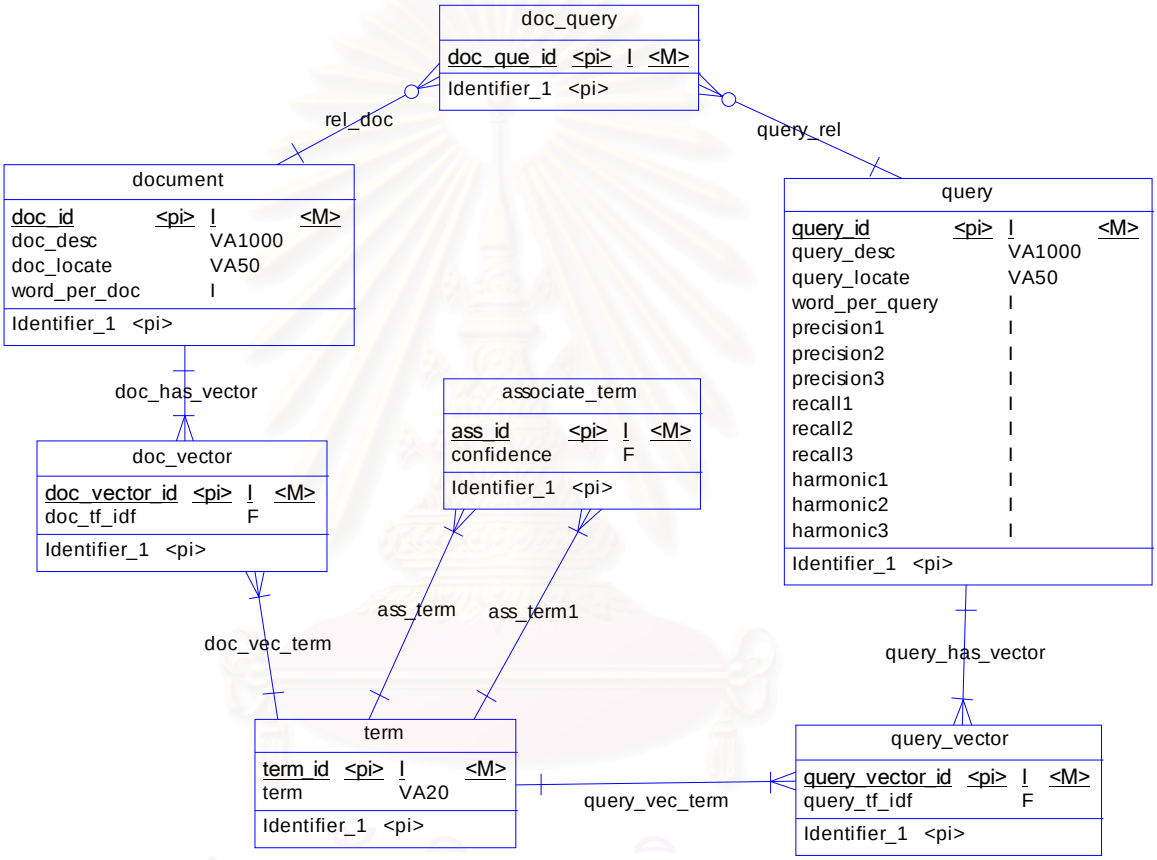


รูปที่ ซ.10 รูปแสดงขั้นตอนการทำงานของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคกฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้

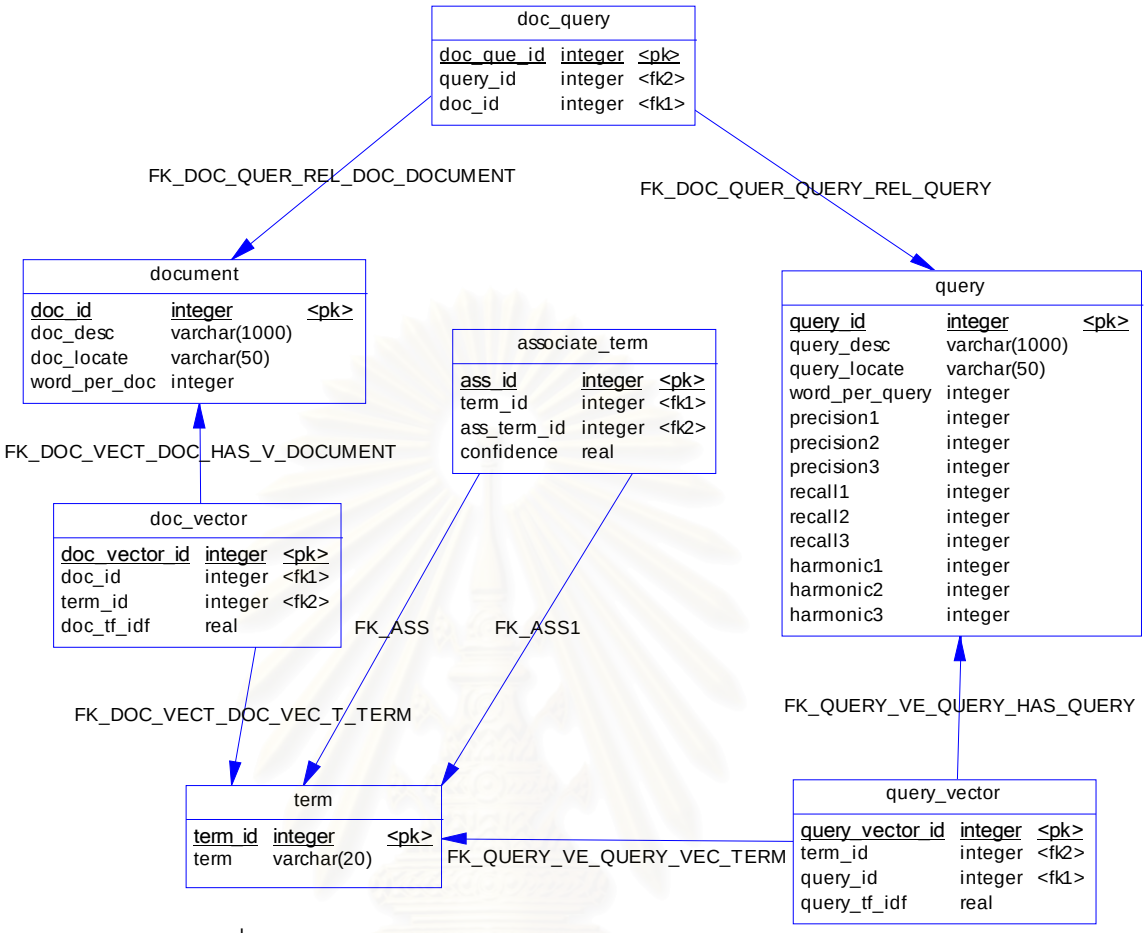
ซ.4 การออกแบบฐานข้อมูล (Database Design)

การออกแบบในส่วนของฐานข้อมูลของการค้นคืนเอกสาร เพื่อเก็บข้อมูลของการทดสอบค้นคืนเอกสารทั้งหมดในการทดลอง โดยจะแสดง ER Diagram เป็นส่วนการออกแบบความสัมพันธ์ของตารางทั้งหมดในฐานข้อมูลของการค้นคืนเอกสาร

1) ER Diagram



รูปที่ ซ.11 รูปแสดงแผนภาพเชิงแนวคิด (Conceptual Diagram)



รูปที่ ซ.12 รูปแสดงแผนภาพเชิงกายภาพ (Physical Diagram)

2) **พจนานุกรมข้อมูล** (Data Dictionary)

ชื่อตาราง :	document			
คำอธิบาย :	ตารางเก็บข้อมูลเอกสาร			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
doc_id	PK	integer	4	รหัสเอกสาร ซึ่งใช้เป็นคีย์หลักตารางเอกสาร
doc_desc		varchar	1000	เอกสารที่ใช้ในการทดสอบการค้นคืนเอกสาร โดยจะเก็บบทความเพียงส่วนหนึ่งของบทความทั้งหมด
doc_locate		varchar	50	สถานที่เก็บแฟ้มข้อมูลของบทความทั้งหมด
word_per_doc		integer	4	จำนวนคำของบทความนั้น ๆ

ชื่อตาราง :	query			
คำอธิบาย :	ตารางเก็บข้อมูลข้อสอบถาม			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
query_id	PK	integer	4	รหัสข้อสอบถาม ซึ่งใช้เป็นคีย์หลักตารางเอกสาร
query_desc		varchar	1000	ข้อสอบถามที่ใช้ในการทดลอง ซึ่งมีรูปแบบเป็นข้อความ
word_per_query		integer	4	จำนวนคำของข้อสอบถามนั้น ๆ
precision1		integer	4	ค่าความถูกต้อง (precision) ของการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์
precision2		integer	4	ค่าความถูกต้อง (precision) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วม
precision3		integer	4	ค่าความถูกต้อง (precision) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้
recall1		integer	4	ค่าเรียกคืน (recall) ของการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์
recall2		integer	4	ค่าเรียกคืน (recall) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วม
recall3		integer	4	ค่าเรียกคืน (recall) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

ชื่อคอลัมน์	คีย์	ชนิด	ความยาว	คำอธิบาย
harmonic1		integer	4	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ของการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์
harmonic2		integer	4	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และเทคนิคการใช้กฎความสัมพันธ์ของคำร่วม
harmonic3		integer	4	ค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ของการการค้นคืนเอกสารโดยใช้เทคนิคปริภูมิเวกเตอร์และใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้

ชื่อตาราง :	doc_query			
คำอธิบาย :	ตารางเก็บข้อมูลเอกสารที่ตรงกับข้อสอบถามแต่ละข้อสอบถาม			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
doc_query_id	PK	integer	4	คีย์หลักของตารางเก็บข้อมูลเอกสารที่ตรงกับข้อสอบถามแต่ละข้อสอบถาม
doc_id	FK	integer	4	รหัสเอกสารซึ่งเป็นคีย์อ้างอิงมาจากตารางเอกสาร โดยจะจับคู่กับข้อสอบถามที่เกี่ยวเนื่องกัน
query_id	FK	integer	4	รหัสข้อสอบถามซึ่งเป็นคีย์อ้างอิงมาจากตารางข้อสอบถาม โดยจะจับคู่กับเอกสารที่เกี่ยวเนื่องกัน

ชื่อตาราง :	term			
คำอธิบาย :	ตารางเก็บคำทั้งหมดที่ปรากฏในเอกสารและข้อสอบถามที่นำมาทดสอบการค้นคืนเอกสาร			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
term_id	PK	integer	4	คีย์หลักของคำ
term		varchar	50	คำต่าง ๆ

ชื่อตาราง :	doc_vector			
คำอธิบาย :	ตารางเก็บเวกเตอร์เอกสารโดยจะเก็บรหัสเอกสารและรหัสคำที่อยู่ในเอกสารนั้น ๆ			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
doc_vector_id	PK	integer	4	คีย์หลักของเวกเตอร์เอกสาร
doc_id	FK	integer	4	รหัสเอกสารซึ่งเป็นคีย์อ้างอิงมาจากตารางเอกสาร โดยจะจับคู่กับรหัสคำที่ปรากฏในเอกสารนั้น ๆ
term_id	FK	integer	4	รหัสคำซึ่งเป็นคีย์อ้างอิงมาจากตารางคำ โดยจะจับคู่กับเอกสารที่มีคำนั้นๆปรากฏ
doc_tf_idf		real	4	ค่าน้ำหนักของคำนั้น ๆ ในเอกสารหนึ่ง ๆ

ชื่อตาราง :	query_vector			
คำอธิบาย :	ตารางเก็บเวกเตอร์ข้อสอบถามโดยจะเก็บรหัสข้อสอบถามและรหัสคำที่อยู่ในข้อสอบถามนั้น ๆ			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
query_vector_id	PK	integer	4	คีย์หลักของเวกเตอร์ข้อสอบถาม
query_id	FK	integer	4	รหัสเอกสารซึ่งเป็นคีย์อ้างอิงมาจากตารางข้อสอบถาม โดยจะจับคู่กับรหัสคำที่ปรากฏในข้อสอบถามนั้น ๆ

ชื่อคอลัมน์	คีย์	ชนิด	ความยาว	คำอธิบาย
term_id	FK	integer	4	รหัสคำซึ่งเป็นคีย์อ้างอิงมาจากตารางคำ โดยจะจับคู่กับข้อสอบถามที่มีคำนั้นๆ ปรากฏ
query _tf_idf		real	4	ค่าน้ำหนักของคำนั้น ๆ ในข้อสอบถามหนึ่ง ๆ

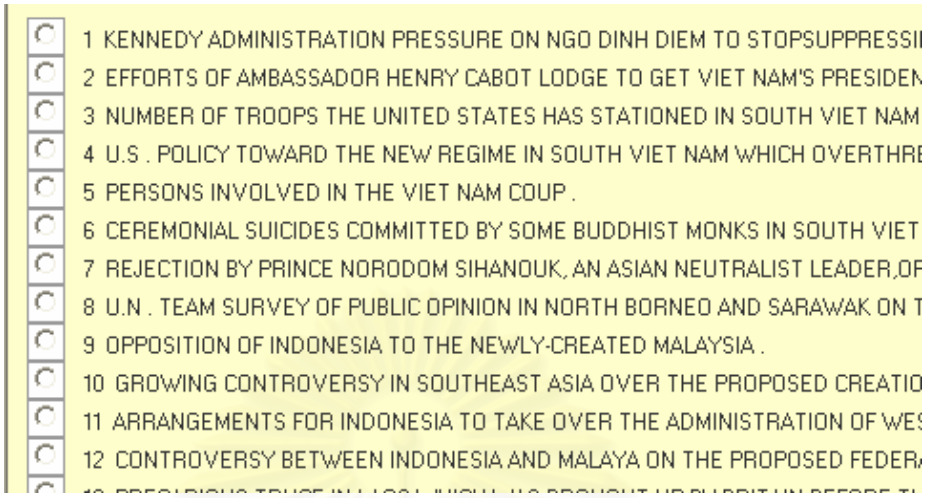
ชื่อตาราง :	associate_term			
คำอธิบาย :	ตารางเก็บคำที่มีความสัมพันธ์กัน			
ชื่อคอลัมน์	**คีย์**	**ชนิด**	**ความยาว**	**คำอธิบาย**
ass_id	PK	integer	4	คีย์หลักของเวกเตอร์ข้อสอบถาม
term_id	FK	integer	4	รหัสคำซึ่งเป็นคีย์อ้างอิงมาจากตารางคำ โดยจะจับคู่กับรหัสคำที่มีความสัมพันธ์กัน
ass_term_id	FK	integer	4	รหัสคำซึ่งเป็นคีย์อ้างอิงมาจากตารางคำ โดยจะจับคู่กับรหัสคำที่มีความสัมพันธ์กัน
confidence		real	4	ค่าความเชื่อมั่นของคำที่มีความสัมพันธ์กัน

ซ.5 การออกแบบหน้าจอ (Interface Design)

การออกแบบหน้าจอของการค้นคืนเอกสารทั้ง 3 รูปแบบนั้นจะมีทั้งหมด 3 หน้าจอ คือ

หน้าจอ 1 : หน้าจอเลือกข้อสอบถาม

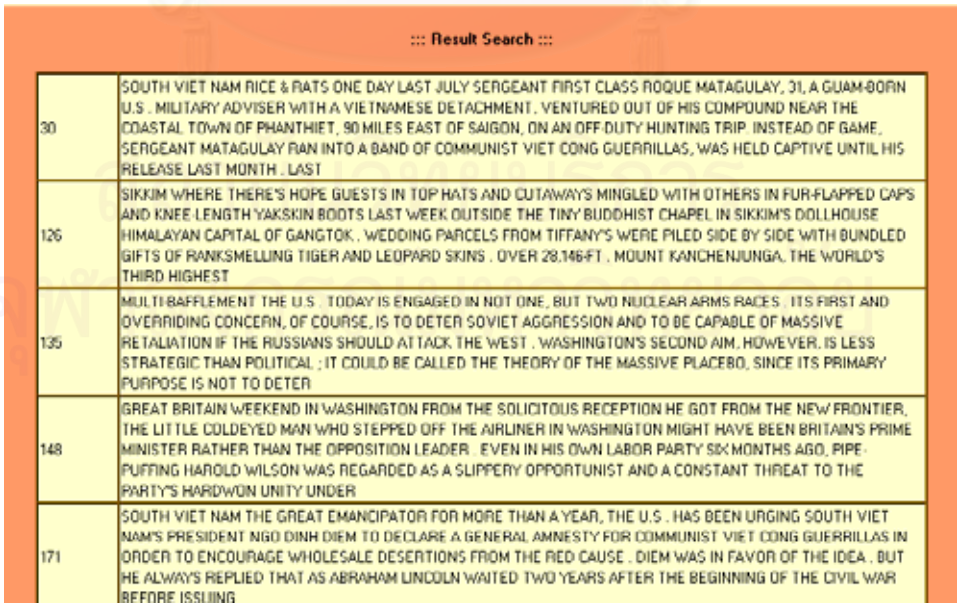
หน้าจอแสดง Radio Button ของข้อสอบถามทั้งหมดที่ต้องการจะทดสอบ เพื่อให้ผู้ใช้เลือกข้อสอบถามที่ต้องการจะทดสอบ เมื่อผู้ใช้เลือกข้อสอบถามที่ต้องการแล้วจากนั้นกดปุ่ม Submit ซึ่งหน้าจอนี้จะปรากฏอยู่ในการค้นคืนเอกสารทั้ง 3 รูปแบบดังรูปที่ ซ.13



รูปที่ ซ.13 รูปแสดงหน้าจอเลือกข้อสอบถามที่ต้องการทดลอง

หน้าจอ 2 : หน้าจอแสดงผลการค้นคืนเอกสาร

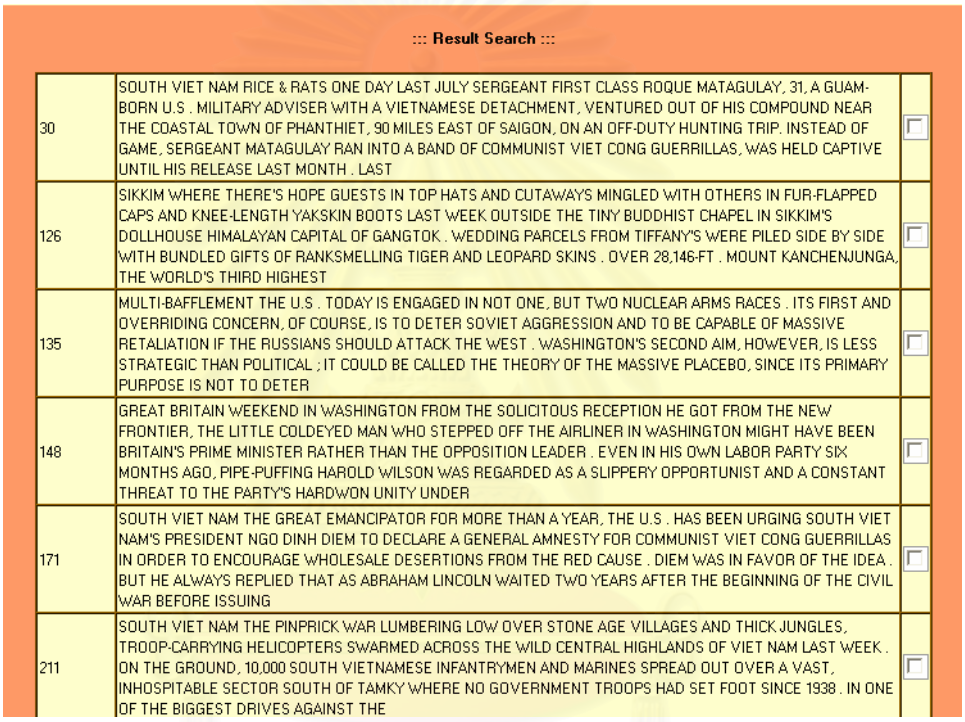
หน้าจอแสดงผลลัพธ์หลังจากที่ผ่านกระบวนการที่ออกแบบไว้ในแต่ละการค้นคืนเอกสารทั้ง 3 รูปแบบ โดยหน้าจอนี้จะแสดงเอกสารที่ค้นคืนออกมา รายการรหัสเอกสารที่ถูกต้องตรงกับข้อสอบถามที่ผู้ใช้เลือกเข้ามาและค่าประสิทธิภาพทั้ง 3 ค่า คือ ค่าความถูกต้อง (Precision) ค่าเรียกคืน (Recall) และค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision) ซึ่งหน้าจอนี้จะปรากฏอยู่ในการค้นคืนเอกสารทั้ง 3 รูปแบบดังรูปที่ ซ.14



รูปที่ ซ.14 รูปแสดงหน้าจอแสดงเอกสารและผลการค้นคืนเอกสาร

หน้าจอ 3 : หน้าจอให้ผลสะท้อนกลับ

หน้าจอที่แสดงรายการเอกสารที่การค้นคืนเอกสารค้นคืนออกมาโดยจะแสดงเป็น Checkbox เพื่อให้ผู้ใช้เลือกเอกสารที่ถูกต้องตรงกับข้อสอบถามนั้น ๆ โดยเป็นการให้ผลสะท้อนกลับมายังการค้นคืนเอกสาร เพื่อให้การค้นคืนเอกสารออกมาอีกครั้ง ดังรูปที่ ซ.15 ซึ่งหน้าจอนี้จะปรากฏอยู่ในการค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์เท่านั้นและที่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำร่วมกับเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้



::: Result Search :::

30	SOUTH VIET NAM RICE & RATS ONE DAY LAST JULY SERGEANT FIRST CLASS ROQUE MATAGULAY, 31, A GUAM-BORN U.S . MILITARY ADVISER WITH A VIETNAMESE DETACHMENT, VENTURED OUT OF HIS COMPOUND NEAR THE COASTAL TOWN OF PHANTHIET, 90 MILES EAST OF SAIGON, ON AN OFF-DUTY HUNTING TRIP. INSTEAD OF GAME, SERGEANT MATAGULAY RAN INTO A BAND OF COMMUNIST VIET CONG GUERRILLAS, WAS HELD CAPTIVE UNTIL HIS RELEASE LAST MONTH . LAST	☐
126	SIKKIM WHERE THERE'S HOPE GUESTS IN TOP HATS AND CUTAWAYS MINGLED WITH OTHERS IN FUR-FLAPPED CAPS AND KNEE-LENGTH YAKSKIN BOOTS LAST WEEK OUTSIDE THE TINY BUDDHIST CHAPEL IN SIKKIM'S DOLLHOUSE HIMALAYAN CAPITAL OF GANGTOK . WEDDING PARCELS FROM TIFFANY'S WERE PILED SIDE BY SIDE WITH BUNDLED GIFTS OF RANKSMELLING TIGER AND LEOPARD SKINS . OVER 28,146-FT . MOUNT KANCHENJUNGA, THE WORLD'S THIRD HIGHEST	☐
135	MULTI-BAFFLEMENT THE U.S . TODAY IS ENGAGED IN NOT ONE, BUT TWO NUCLEAR ARMS RACES . ITS FIRST AND OVERRIDING CONCERN, OF COURSE, IS TO DETER SOVIET AGGRESSION AND TO BE CAPABLE OF MASSIVE RETALIATION IF THE RUSSIANS SHOULD ATTACK THE WEST . WASHINGTON'S SECOND AIM, HOWEVER, IS LESS STRATEGIC THAN POLITICAL ; IT COULD BE CALLED THE THEORY OF THE MASSIVE PLACEBO, SINCE ITS PRIMARY PURPOSE IS NOT TO DETER	☐
148	GREAT BRITAIN WEEKEND IN WASHINGTON FROM THE SOLICITOUS RECEPTION HE GOT FROM THE NEW FRONTIER, THE LITTLE COLDEYED MAN WHO STEPPED OFF THE AIRLINER IN WASHINGTON MIGHT HAVE BEEN BRITAIN'S PRIME MINISTER RATHER THAN THE OPPOSITION LEADER . EVEN IN HIS OWN LABOR PARTY SIX MONTHS AGO, PIPE-PUFFING HAROLD WILSON WAS REGARDED AS A SLIPPERY OPPORTUNIST AND A CONSTANT THREAT TO THE PARTY'S HARDWON UNITY UNDER	☐
171	SOUTH VIET NAM THE GREAT EMANCIPATOR FOR MORE THAN A YEAR, THE U.S . HAS BEEN URGING SOUTH VIET NAM'S PRESIDENT NGO DINH DIEM TO DECLARE A GENERAL AMNESTY FOR COMMUNIST VIET CONG GUERRILLAS IN ORDER TO ENCOURAGE WHOLESALE DESERTIONS FROM THE RED CAUSE . DIEM WAS IN FAVOR OF THE IDEA . BUT HE ALWAYS REPLIED THAT AS ABRAHAM LINCOLN WAITED TWO YEARS AFTER THE BEGINNING OF THE CIVIL WAR BEFORE ISSUING	☐
211	SOUTH VIET NAM THE PINPRICK WAR LUMBERING LOW OVER STONE AGE VILLAGES AND THICK JUNGLES, TROOP-CARRYING HELICOPTERS SWARMED ACROSS THE WILD CENTRAL HIGHLANDS OF VIET NAM LAST WEEK . ON THE GROUND, 10,000 SOUTH VIETNAMESE INFANTRYMEN AND MARINES SPREAD OUT OVER A VAST, INHOSPITABLE SECTOR SOUTH OF TAMKY WHERE NO GOVERNMENT TROOPS HAD SET FOOT SINCE 1938 . IN ONE OF THE BIGGEST DRIVES AGAINST THE	☐

รูปที่ ซ.15 รูปแสดงหน้าจอเลือกเอกสารเพื่อให้ผลสะท้อนกลับเอกสารที่ตรงกับข้อสอบถามนั้น ๆ

การค้นคืนเอกสารทั้ง 3 รูปแบบนี้ แต่ละรูปแบบจะประกอบด้วยหน้าจอดังที่กล่าวมาดังตารางที่ ซ.1 และมีลำดับการแสดงหน้าจอของแต่ละการค้นคืนเอกสารตารางที่ ซ.2 โดยการค้นคืนเอกสารรูปแบบที่ 1 และการค้นคืนเอกสารรูปแบบที่ 2 จะประกอบด้วยหน้าจอ 1 และหน้าจอ 2 โดยจะแสดงหน้าจอ 1 ก่อนแล้วจึงแสดงหน้าจอ 2 ส่วนการค้นคืนเอกสารรูปแบบที่ 3 จะประกอบด้วยหน้าจอทั้งหน้าจอ 1 หน้าจอ 2 และหน้าจอ 3 โดยจะแสดงหน้าจอ1 หน้าจอ 2 และหน้าจอ 3 ตามลำดับ

ตารางที่ ซ.1 ตารางแสดงหน้าจอที่ปรากฏในการค้นคืนเอกสารแต่ละรูปแบบ

การค้นคืนเอกสาร	หน้าจอเลือกข้อสอบถาม (หน้าจอ 1)	หน้าจอแสดงผลการค้นคืนเอกสาร (หน้าจอ 2)	หน้าจอให้ผลสะท้อนกลับ (หน้าจอ 3)
1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เท่านั้น	✓	✓	✗
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ	✓	✓	✗
3) การค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้	✓	✓	✓

ตารางที่ ซ.2 ตารางแสดงลำดับการแสดงหน้าจอของแต่ละการค้นคืนเอกสาร

การค้นคืนเอกสาร	ลำดับหน้าจอ
1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์เท่านั้น	หน้าจอ 1 → หน้าจอ 2
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ	หน้าจอ 1 → หน้าจอ 2
3) การค้นคืนเอกสารที่ใช้โดยใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการใช้ผลสะท้อนกลับจากผู้ใช้	หน้าจอ 1 → หน้าจอ 2 → หน้าจอ 3

ซ.6 การออกแบบการทดสอบ (Test Design)

การค้นคืนเอกสารทั้ง 3 รูปแบบจะประกอบด้วยฟังก์ชันการทำงานหลัก ๆ ทั้งหมดดังตารางที่ ซ.3 ซึ่งงานวิจัยนี้จะใช้ Functional Testing เป็นเทคนิคของ Black Box Testing โดยใช้วิธีการของ Equivalence Partitioning Steps ซึ่งเป็นการทดสอบ Features, Function และการไหล (Flow) ต่าง ๆ ที่มีอยู่ในเครื่องมือว่าถูกต้องและครบถ้วน มีข้อผิดพลาดใด ๆ หรือไม่ โดยจะออกแบบกรณีทดสอบ (Test Case) ในการทำงานต่าง ๆ ดังตารางที่ ซ.4

ตารางที่ ซ.3 ตารางแสดงกรณีทดสอบ (Test Case) ของแต่ละฟังก์ชันการทำงาน

Test Case	Function	Test Case	Expect Result
TC-1	คำนวณค่าน้ำหนักคำให้กับคำที่มีความสัมพันธ์กับคำในเวกเตอร์ข้อสอบถาม	- ค่าน้ำหนักของคำในข้อสอบถามเดิม - ค่าความเชื่อมั่น (Confidence)	ค่าน้ำหนักคำที่มีความสัมพันธ์กับคำในข้อสอบถามเดิม = ค่าความเชื่อมั่น (Confidence) * ค่าน้ำหนักของคำในข้อสอบถามเดิม
TC-2	การคำนวณค่าความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม	- ค่าน้ำหนักแต่ละมิติของเวกเตอร์ข้อสอบถาม q ($w_{i,q}$) - ค่าน้ำหนักแต่ละมิติของเวกเตอร์เอกสารที่ j ($w_{i,j}$)	ค่าความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม = $\frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^{2}} \times \sqrt{\sum_{i=1}^{t} w_{i.q}^{2}}}$ โดยที่ t คือจำนวนมิติของเวกเตอร์ทั้งสองเวกเตอร์
TC-3	การคำนวณค่าความถูกต้อง (Precision)	- จำนวนเอกสารเกี่ยวเนื่องตามความต้องการที่ค้นคืนออกมาได้ (|Ra|) - จำนวนเอกสารทั้งหมดที่ค้นคืนออกมา (|A|)	ค่าความถูกต้อง (Precision) = $\frac{\lvert Ra \rvert}{\lvert A \rvert}$

Test Case	Function	Test Case	Expect Result
TC-4	การคำนวณค่าเรียกคืน (Recall)	- จำนวนเอกสารเกี่ยวเนื่องตามความต้องการที่ค้นคืนออกมาได้ (|Ra|) - จำนวนเอกสารเกี่ยวเนื่องกับความต้องการที่อยู่ในฐานข้อมูลทั้งหมด (|R|)	ค่าเรียกคืน (Recall) = $\frac{\lvert Ra \rvert}{\lvert R \rvert}$
TC-5	การคำนวณค่าเฉลี่ยฮาร์โมนิคของค่าเรียกคืนและค่าความถูกต้อง (Harmonic mean of recall and precision)	- ค่าความถูกต้อง (P) - ค่าเรียกคืน (R)	$F - measure = \frac{2}{\frac{1}{R} + \frac{1}{P}}$
TC-6	คำนวณค่าน้ำหนักคำของเวกเตอร์ข้อสอบถามใหม่โดยใช้วิธี Rochio	- เวกเตอร์ข้อสอบถามที่กำหนดขึ้นใหม่ ($\vec{q}_m$) - เวกเตอร์ข้อสอบถามเริ่มต้น ($\vec{q}$) - จำนวนเอกสารที่เกี่ยวเนื่องกับตามต้องการ ($\lvert D_r \rvert$) - จำนวนเอกสารที่ไม่เกี่ยวเนื่องกับตามต้องการ ($\lvert D_n \rvert$)	$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{\lvert D_r \rvert}\left(\sum_{\forall \vec{d}_j \in D_r'} \vec{d}_j\right) - \frac{\gamma}{\lvert D_n \rvert}\left(\sum_{\forall \vec{d}_j \in D_n'} \vec{d}_j\right)$

Test Case	Function	Test Case	Expect Result
		- เวกเตอร์ของเอกสารที่ j ($\vec{d}_j$) - เซตของเอกสารที่เกี่ยวเนื่องในจำนวนเอกสารที่ค้นคืนได้ทั้งหมด (D_r ') - เซตของเอกสารที่ไม่เกี่ยวเนื่องในจำนวนเอกสารที่ค้นคืนได้ทั้งหมด (D_n ') - α เท่ากับ 8 - β เท่ากับ 16 - γ เท่ากับ 4	

ภาคผนวก ฌ

ผลการทดลองค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision)

การวิเคราะห์ผลการทดลองเพิ่มเติมในส่วนของการเปรียบเทียบประสิทธิภาพของระบบค้นคืนเอกสารโดยใช้ค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ของการค้นคืนเอกสารที่ใช้เทคนิคต่าง ๆ ดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ที่ไม่ใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคผลสะท้อนกลับจากผู้ใช้
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ
3) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับใช้เทคนิคการใช้กฎความสัมพันธ์ของคำและเทคนิคการให้ผลสะท้อนกลับจากผู้ใช้
4) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับการใช้เทคนิคการให้ผลสะท้อนกลับจากผู้ใช้

จากการค้นคืนเอกสารทั้ง 4 รูปแบบให้ผลการทดลองค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision) ดังตารางที่ ฌ.1

ตารางที่ ฌ.1 ตารางแสดงผลการทดลองค่าเรียกคืน (Recall) และค่าความถูกต้อง (Precision)

	ค่าความถูกต้องของการค้นคืนเอกสาร				**ค่าเรียกคืนของการค้นคืนเอกสาร**			
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*	*รูปแบบที่ 1*	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*
1	0.2222	0.2400	0.2222	0.2222	1.0000	1.0000	1.0000	1.0000
2	0.0588	0.0606	0.0606	0.0571	1.0000	1.0000	1.0000	1.0000
3	0.0612	0.0625	0.0577	0.0577	0.7500	0.7500	0.7500	0.7500
4	0.1250	0.1316	0.1250	0.1250	1.0000	1.0000	1.0000	1.0000
5	0.1316	0.1250	0.1316	0.1316	1.0000	1.0000	1.0000	1.0000
6	0.2432	0.2500	0.2432	0.2432	1.0000	1.0000	1.0000	1.0000

	ค่าความถูกต้องของการค้นคืนเอกสาร				ค่าเรียกคืนของการค้นคืนเอกสาร			
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*	*รูปแบบที่ 1*	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*
7	0.1111	0.1111	0.1111	0.1111	1.0000	1.0000	1.0000	1.0000
8	0.1333	0.0800	0.1667	0.1333	1.0000	1.0000	1.0000	1.0000
9	0.5000	0.5833	0.6364	0.5000	0.8750	0.8750	0.8750	0.8750
10	0.3529	0.4000	0.4000	0.3529	1.0000	1.0000	1.0000	1.0000
11	0.1333	0.1333	0.1250	0.1250	1.0000	1.0000	1.0000	1.0000
12	0.3889	0.3889	0.3889	0.3684	1.0000	1.0000	1.0000	1.0000
13	0.1875	0.1875	0.1875	0.1875	1.0000	1.0000	1.0000	1.0000
14	0.2000	0.2000	0.1667	0.1667	1.0000	1.0000	1.0000	1.0000
15	0.4444	0.4444	0.4444	0.4444	0.8000	0.8000	0.8000	0.8000
16	0.1579	0.1579	0.1500	0.1500	1.0000	1.0000	1.0000	1.0000
17	0.1250	0.1538	0.1429	0.1333	1.0000	1.0000	1.0000	1.0000
18	0.2000	0.2000	0.1667	0.1667	1.0000	1.0000	1.0000	1.0000
19	0.2632	0.1724	0.2632	0.2632	1.0000	1.0000	1.0000	1.0000
20	0.0000	0.0000	0.0833	0.0833	0.0000	0.0000	1.0000	1.0000
21	0.0667	0.0667	0.0690	0.0690	1.0000	1.0000	1.0000	1.0000
22	0.1111	0.1111	0.1053	0.1053	1.0000	1.0000	1.0000	1.0000
23	0.2500	0.2500	0.1111	0.1111	1.0000	1.0000	1.0000	1.0000
24	0.0455	0.0455	0.0500	0.0500	1.0000	1.0000	1.0000	1.0000
25	0.0526	0.0526	0.0526	0.0526	1.0000	1.0000	1.0000	1.0000
26	0.0000	0.0000	0.1667	0.1667	0.0000	0.0000	0.5000	0.5000
27	0.1200	0.1200	0.1000	0.1000	1.0000	1.0000	1.0000	1.0000
28	0.1429	0.1429	0.1429	0.1429	0.8000	0.8000	1.0000	1.0000
29	0.1176	0.1333	0.1333	0.1176	1.0000	1.0000	1.0000	1.0000
30	0.1724	0.1852	0.1667	0.1667	1.0000	1.0000	1.0000	1.0000

	ค่าความถูกต้องของการค้นคืนเอกสาร				ค่าเรียกคืนของการค้นคืนเอกสาร			
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*	*รูปแบบที่ 1*	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*
31	0.2500	0.2500	0.2917	0.2917	0.5714	0.5714	1.0000	1.0000
32	0.0204	0.0204	0.0169	0.0169	1.0000	1.0000	1.0000	1.0000
33	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
34	0.1667	0.1667	0.1667	0.1667	1.0000	1.0000	1.0000	1.0000
35	0.1000	0.1000	0.1000	0.1000	1.0000	1.0000	1.0000	1.0000
36	0.3333	0.3333	0.3333	0.3333	1.0000	1.0000	1.0000	1.0000
37	0.1429	0.1429	0.1429	0.1429	0.5000	0.5000	1.0000	1.0000
38	0.0000	0.0000	0.0417	0.0417	0.0000	0.0000	1.0000	1.0000
39	0.3214	0.3214	0.3103	0.3103	1.0000	1.0000	1.0000	1.0000
40	0.2368	0.2368	0.2368	0.2368	1.0000	1.0000	1.0000	1.0000
41	0.1613	0.1613	0.1429	0.1429	0.8333	0.8333	1.0000	1.0000
42	0.0270	0.0270	0.0256	0.0256	1.0000	1.0000	1.0000	1.0000
43	0.0952	0.0952	0.0769	0.0769	1.0000	1.0000	1.0000	1.0000
44	0.1538	0.1538	0.1250	0.1250	1.0000	1.0000	1.0000	1.0000
45	0.1379	0.1379	0.1667	0.1667	0.8000	0.8000	1.0000	1.0000
46	0.4048	0.4048	0.3864	0.3864	0.9444	0.9444	0.9444	0.9444
47	0.2174	0.2174	0.1786	0.1786	0.8333	0.8333	0.8333	0.8333
48	0.0833	0.0833	0.0556	0.0556	1.0000	1.0000	1.0000	1.0000
49	0.3478	0.3077	0.3333	0.3333	1.0000	1.0000	1.0000	1.0000
50	0.0435	0.0435	0.0417	0.0417	1.0000	1.0000	1.0000	1.0000
51	0.0857	0.0857	0.0750	0.0750	1.0000	1.0000	1.0000	1.0000
52	0.0263	0.0263	0.0238	0.0238	0.5000	0.5000	0.5000	0.5000
53	0.1053	0.1053	0.0741	0.0741	1.0000	1.0000	1.0000	1.0000
54	0.0392	0.0392	0.0364	0.0364	1.0000	1.0000	1.0000	1.0000

	ค่าความถูกต้องของการค้นคืนเอกสาร				ค่าเรียกคืนของการค้นคืนเอกสาร			
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*	*รูปแบบที่ 1*	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*
55	0.2973	0.2973	0.2683	0.2683	0.9167	0.9167	0.9167	0.9167
56	0.0476	0.0476	0.0400	0.0400	1.0000	1.0000	1.0000	1.0000
57	0.1250	0.1250	0.1250	0.1250	1.0000	1.0000	1.0000	1.0000
58	0.3684	0.3500	0.3333	0.3333	0.8750	0.8750	1.0000	1.0000
59	0.0870	0.0870	0.0833	0.0833	1.0000	1.0000	1.0000	1.0000
60	0.1000	0.1000	0.1000	0.1000	1.0000	1.0000	1.0000	1.0000
61	0.5652	0.5652	0.4063	0.4063	0.8667	0.8667	0.8667	0.8667
62	0.1250	0.1250	0.1429	0.1429	1.0000	1.0000	1.0000	1.0000
63	0.2821	0.3056	0.3333	0.3333	1.0000	1.0000	1.0000	1.0000
64	0.1053	0.0952	0.1111	0.1111	1.0000	1.0000	1.0000	1.0000
65	0.0455	0.0370	0.0455	0.0455	1.0000	1.0000	1.0000	1.0000
66	0.0769	0.0606	0.0690	0.0690	1.0000	1.0000	1.0000	1.0000
67	0.2308	0.2308	0.2308	0.2308	1.0000	1.0000	1.0000	1.0000
68	0.2222	0.2222	0.2069	0.2069	0.7500	0.7500	0.7500	0.7500
69	0.6190	0.6500	0.6500	0.6190	1.0000	1.0000	1.0000	1.0000
70	0.0714	0.0526	0.0667	0.0667	1.0000	1.0000	1.0000	1.0000
71	0.5000	0.5000	0.5000	0.5000	0.6667	0.6667	0.6667	0.6667
72	0.0333	0.0333	0.0294	0.0294	1.0000	1.0000	1.0000	1.0000
73	0.0833	0.0769	0.0833	0.0833	1.0000	1.0000	1.0000	1.0000
74	0.1053	0.1333	0.1333	0.1111	1.0000	1.0000	1.0000	1.0000
75	0.0435	0.0556	0.0556	0.0435	1.0000	1.0000	1.0000	1.0000
76	0.1786	0.1786	0.1786	0.1786	1.0000	1.0000	1.0000	1.0000
77	0.0435	0.0435	0.0417	0.0417	1.0000	1.0000	1.0000	1.0000
78	0.0833	0.0833	0.0833	0.0833	1.0000	1.0000	1.0000	1.0000

	ค่าความถูกต้องของการค้นคืนเอกสาร				ค่าเรียกคืนของการค้นคืนเอกสาร			
	รูปแบบที่ 1	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*	*รูปแบบที่ 1*	*รูปแบบที่ 2*	*รูปแบบที่ 3*	*รูปแบบที่ 4*
79	0.0500	0.0526	0.0625	0.0526	1.0000	1.0000	1.0000	1.0000
80	0.4444	0.3810	0.4000	0.4000	0.4706	0.4706	0.4706	0.4706
81	0.0189	0.0189	0.0345	0.0345	0.5000	0.5000	1.0000	1.0000
82	0.2273	0.2273	0.1786	0.1786	1.0000	1.0000	1.0000	1.0000
83	0.0625	0.0625	0.0625	0.0625	1.0000	1.0000	1.0000	1.0000

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวศิรัตน์ ศิรนานนท์ เกิดเมื่อวันที่ 24 กันยายน พ.ศ. 2525 ที่จังหวัดชลบุรี สำเร็จการศึกษาปริญญาตรีวิทยาศาสตรบัณฑิต จากภาควิชาคณิตศาสตร์ สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2546 และเข้าศึกษาต่อในหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาการพัฒนาซอฟต์แวร์ด้านธุรกิจ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยในปีการศึกษา 2547