บทที่ 3

ระเบียบวิธีวิจัย

ในบทนี้จะกล่าวถึงระเบียบวิธีวิจัยที่เป็นแนวทางในการตอบวัตถุประสงค์ของการวิจัย แผนแบบการทดลอง (Experimental Design) การทดสอบสมมติฐาน การทำงานของเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบต่าง ๆ ที่งานวิจัยกำหนดในส่วนของภาพรวมของเครื่องมือทดสอบและในส่วนของรายละเอียดของเครื่องมือทดสอบ รวมทั้งแสดงขั้นตอนวิธีการพัฒนาเครื่องมือทดสอบและการทดสอบประสิทธิภาพของการค้นคืนเอกสาร ประเด็นของความเชื่อถือได้ (Reliability) ความถูกต้อง (Validity) และกรอบการวิเคราะห์ข้อมูล (Data Analysis Framework) ดังรายละเอียดต่อไปนี้

3.1 แผนแบบการทดลอง

งานวิจัยนี้มีวัตถุประสงค์ในการทดลองเพื่อศึกษาประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) ในการกำหนดกรอบค่าความคล้ายคลึงของเอกสารที่เป็นผลลัพธ์ในการค้นคืนเอกสารสำหรับเทคนิคปริภูมิเวกเตอร์ที่วัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน เพื่อเพิ่มประสิทธิภาพของระบบค้นคืนเอกสาร (Document Retrieval) ให้สามารถค้นคืนเอกสารที่เกี่ยวข้องกับข้อสอบถามและตรงตามความต้องการของผู้ใช้ได้ โดยศึกษาเปรียบเทียบประสิทธิภาพการค้นคืนเอกสาร (Document Retrieval) กับเทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม

จากวัตถุประสงค์งานวิจัยที่กล่าวมาข้างต้น งานวิจัยนี้เป็นงานวิจัยเชิงประจักษ์ (Empirical Research) ผู้วิจัยต้องการศึกษาตัวแปรต้นที่ประยุกต์ใช้ทฤษฎีมาอธิบายค่าที่เกิดขึ้นของตัวแปรตามจากการเลือกทดลองกับการใช้ตัวแปรต้นที่แตกต่างกัน เพื่อที่ต้องการทราบค่าประสิทธิภาพของระบบค้นคืนเอกสารว่ามีค่าแตกต่างกันอย่างไร โดยกำหนดให้มีตัวแปรในการทดลองเปรียบเทียบการค้นคืนเอกสารในรูปแบบที่ต้องการทดสอบ ดังต่อไปนี้

3.1.1. ตัวแปรต้น

งานวิจัยนี้สนใจว่าการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการกำหนดกรอบค่าความคล้ายคลึงในการค้นคืนที่ประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) กำหนดกรอบเอกสารที่จะถูกค้นคืนสามารถให้ประสิทธิภาพของการค้นคืนเอกสารให้ตรงตามความต้องการของผู้ใช้ได้ดีกว่าการค้นคืนเอกสารที่

ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมหรือไม่ และซึ่งทั้ง 2 วิธีการเป็นตัวแปรที่ผู้วิจัยต้องการศึกษาและเป็นตัวแปรสำคัญต่องานวิจัยนี้ ในงานวิจัยสนใจวิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถามสองวิธีการด้วยกัน ดังนั้นตัวแปรต้นของการศึกษาในครั้งนี้จะเป็นการเปรียบเทียบการค้นคืนเอกสารด้วยวิธีการวัดความคล้ายคลึง 2 รูปแบบดังนี้

1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle)

จากการค้นคืนเอกสารทั้ง 2 รูปแบบ ผู้วิจัยจะเรียกการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ว่า"การค้นคืนเอกสารรูปแบบที่ 1" และเรียกการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมว่า"การค้นคืนเอกสารรูปแบบที่ 2"

3.1.2 ตัวแปรตาม

เนื่องจากงานวิจัยนี้สนใจเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารโดยการใช้เทคนิคดังที่กล่าวในหัวข้อตัวแปรต้น ดังนั้นการเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสารจะพิจารณาความถูกต้องและความครอบคลุมในการค้นคืนเอกสารตามความต้องการของผู้ใช้ โดยในงานวิจัยนี้จะวัดประสิทธิภาพของระบบค้นคืนเอกสารจากค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ซึ่งรายละเอียดและวิธีการคำนวณค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค ได้กล่าวไว้ในบทที่ 2

3.1.3 ตัวแปรควบคุม

ในการวิจัยเพื่อได้ผลการทดลองที่สะท้อนประสิทธิภาพของเทคนิคทั้งสองรูปแบบอย่างแท้จริง ผู้วิจัยต้องควบคุมปัจจัยอื่นๆที่อาจส่งผลกับการทดลองให้มีความคงที่หรือเหมือนกันมากที่สุด ตัวแปรที่ผู้วิจัยจะควบคุมในการสร้างเครื่องมือทดสอบการค้นคืนเอกสารทั้งสองรูปแบบ คือ

1) เอกสาร

เอกสารเป็นชุดเอกสารที่นำมาทดสอบกับระบบการค้นคืนเอกสารทั้งสองรูปแบบ ซึ่งผู้วิจัยหวังว่าจะทดลองระบบกับเอกสารภาษาอังกฤษทุกเอกสาร แต่ในทางปฏิบัตินั้นไม่สามารถนำเอกสารประเภทนั้นมาทั้งหมดได้ เนื่องจากผู้วิจัยต้องการเลือกใช้ชุดเอกสารที่มีความใกล้เคียงกับเอกสารเพื่อการตัดสินใจทางธุรกิจ ดังที่ผู้วิจัยได้กล่าวไว้ในบทที่ 1 แล้วว่าผู้บริหารในองค์กรธุรกิจมี

ความจำเป็นต้องใช้เอกสารหลากหลายหัวข้อประกอบการตัดสินใจดำเนินงานทางด้านธุรกิจ เช่น เอกสารที่มีข้อมูลทางด้านการเมือง เศรษฐกิจ เทคโนโลยี รวมถึงสังคมและวัฒนธรรม ผู้วิจัยจึงเห็นว่าแหล่งหนึ่งของเอกสารดังกล่าวคือนิตยสารข่าว ดังนั้นผู้วิจัยจึงเลือกใช้ฐานข้อมูลนิตยสารไทม์ (TIME Collection) เป็นหน่วยตัวอย่างในการพัฒนาระบบค้นคืนเอกสารในการศึกษาครั้งนี้

ฐานข้อมูลนิตยสารไทม์เป็นฐานข้อมูลมาตรฐานที่สร้างโดยมหาวิทยาลัยคอแนล (Cornell University) ในปี 1963 ประกอบไปด้วยเอกสารจำนวน 425 เอกสาร ที่มีความยาวเฉลี่ย 546 คำ หรือ 53 บรรทัด ถูกสร้างขึ้นเพื่อใช้เป็นหน่วยทดสอบระบบค้นคืนเอกสารในงานวิจัยการค้นคืนเอกสาร (Smart Collection 1963) และเป็นฐานข้อมูลเอกสารที่ใช้ในการทดสอบงานวิจัยทางด้านการค้นคืนเอกสารมากมาย (Salton 1971; Chowdhury 2004)

2) **ข้อสอบถาม**

ข้อสอบถามจะใช้ในการทดสอบการค้นคืนเอกสารเพื่อให้ระบบค้นคืนเอกสารที่เกี่ยวเนื่องกับข้อสอบถามออกมาแสดง ซึ่งผู้วิจัยหวังว่าจะทดลองระบบค้นคืนเอกสารกับข้อสอบถามที่ได้จากคำทุกคำที่มีอยู่ในระบบ แต่ในความเป็นจริงจำนวนคำที่มีอยู่ในระบบมีจำนวนมาก ไม่สามารถที่จะระบุคำทุกคำที่มีในระบบกับข้อสอบถามได้ และเนื่องจากฐานข้อมูลนิตยสารไทม์ (TIME Collection) ได้กำหนดข้อสอบถามสำหรับทดสอบระบบค้นคืนเอกสารจำนวน 83 ข้อสอบถาม (Smart Collection, 1963) ดังนั้นผู้วิจัยจึงเลือกข้อสอบถามดังกล่าวเป็นหน่วยตัวอย่างในการทดสอบระบบค้นคืนเอกสารที่พัฒนาขึ้นทั้งสองรูปแบบ

ข้อสอบถามทั้ง 83 ข้อสอบถามมีจำนวนคำในแต่ละข้อสอบถามแตกต่างกันไป โดยจำนวนคำที่น้อยที่สุดคือ 4 คำ จำนวนคำที่มากที่สุดคือ 38 คำ และมีจำนวนคำเฉลี่ย 15 คำ ในการศึกษาครั้งนี้ ถึงแม้ว่าจำนวนคำที่แตกต่างกันในแต่ละข้อสอบถามจะส่งผลให้น้ำหนักของคำในข้อสอบถามแต่ละอันที่คำนวณได้จากสมการ 2.6 มีความแตกต่างกัน แต่ผู้วิจัยก็มีความเห็นว่าน้ำหนักของคำในข้อสอบถามแต่ละอันที่แตกต่างกันนั้นได้สะท้อนความสำคัญของคำเหล่านั้นในข้อสอบถามแล้ว กล่าวคือ ในข้อสอบถามที่มีจำนวนคำน้อย คำในข้อสอบถามจะได้รับน้ำหนักมากกว่าคำในข้อสอบถามที่มีจำนวนคำมาก ซึ่งในข้อสอบถามที่ใช้คำน้อย ได้สะท้อนว่าคำเหล่านั้นมีความเกี่ยวข้องกับเอกสารที่ต้องการมากกว่าคำที่ปรากฏในข้อสอบถามที่ใช้คำมาก

3) **ผลเฉลย**

ผลเฉลยเป็นชุดผลเฉลยของแต่ละข้อสอบถามที่ใช้ในการทดสอบการค้นคืนเอกสาร ซึ่งเป็นการระบุถึงกลุ่มเอกสารที่เป็นคำตอบของแต่ละข้อสอบถาม ผลเฉลยจะมีการระบุถึงข้อสอบถามและเอกสารทั้งหมดที่เกี่ยวข้องกับข้อสอบถามนั้น ในฐานข้อมูลนิตยสารไทม์ (TIME Collection) ได้กำหนดผลเฉลยไว้สำหรับทดสอบระบบค้นคืนเอกสาร (Smart Collection, 1963)

จากข้อสอบถามทั้ง 83 ข้อสอบถาม ซึ่งผลเฉลยนี้เกิดจากการกำหนดโดยกลุ่มคนกลุ่มหนึ่งซึ่งเป็นผู้เชี่ยวชาญกับบทความนั้น ดังนั้นผู้วิจัยจะทราบจำนวนเอกสารที่ถูกต้องในการค้นคืนเอกสารแต่ละครั้ง ทำให้สามารถวัดค่าประสิทธิภาพของระบบค้นคืนเอกสารด้วยการคำนวณหาค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคได้

5) เครื่องมือสร้างระบบค้นคืนเอกสาร

งานวิจัยนี้ต้องใช้เครื่องมือในการพัฒนาเครื่องมือทดสอบการค้นคืนเอกสาร ดังต่อไปนี้

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

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

โปรแกรมทีเอ็มจี (TMG) ทำงานบนโปรแกรมแมทแล็บเวอร์ชั่น 6.5 (MATLAB version 6.5) โปรแกรมทีเอ็มจี (TMG) จะสร้างเวกเตอร์ให้กับเอกสารและข้อสอบถาม โดยที่แต่ละมิติของเวกเตอร์จะเป็นตำแหน่งของคำต่าง ๆ (ศิรัตน์ ศิรนานนท์ 2549) ซึ่งโปรแกรมทีเอ็มจี (TMG) ผู้ใช้สามารถเลือกรูปแบบการสร้างเวกเตอร์ด้วยเทคนิคต่าง ๆ และสามารถกำหนดวิธีการคำนวณค่าน้ำหนักคำในแต่ละมิติของเวกเตอร์ได้ ในงานวิจัยนี้ผู้วิจัยได้กำหนดเทคนิคต่างๆ สำหรับการสร้างเวกเตอร์เอกสารและข้อสอบถามนั่นคือ การตัดคำยกเว้น (Stop word) การลดรูปคำ (Stemming) การให้ค่าน้ำหนักคำด้วยค่าความถี่และค่าความถี่ของเอกสารแบบผกผัน (tf-idf)

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

โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ออกแบบมาเพื่อวิเคราะห์ข้อมูลและการทำเหมืองข้อมูล (Data Mining) ซึ่งโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) ผู้ใช้สามารถเลือกส่วนของการวิเคราะห์ข้อมูลด้วยเทคนิคต่างๆ ของการทำเหมืองข้อมูลได้ เช่น เทคนิคการค้นหากฎความสัมพันธ์ (Association Discovery) การจัดกลุ่มข้อมูล (Clustering) หรือการแบ่งประเภทข้อมูล (Classification) เป็นต้น

ในงานวิจัยนี้ผู้วิจัยได้เลือกใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) สำหรับทดสอบการค้นคืนเอกสาร ผลจากการใช้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 สามารถค้นหาข้อมูลต่าง ๆ จากการจัดกลุ่มข้อมูล ได้แก่ กราฟแสดงผลการจัดกลุ่มของข้อมูล, ตารางแสดงรายละเอียดผลการจัดกลุ่มของแต่ละข้อมูล, ตารางแสดงรายละเอียดค่าสถิติต่าง ๆ ของการจัดกลุ่ม

- Apache 2.2.4

Apache คือ HTTP หรือ Web Server ซึ่งสามารถรันได้ทั้งบนแพลตฟอร์มยูนิกซ์และไมโครซอฟต์วินโดวส์ มีหน้าที่ในการจัดเก็บ Homepage และส่ง Homepage ไปยัง Browser ที่มีการเรียกเข้ายัง Web server ที่เก็บ HomePage นั้นอยู่ ซึ่งสามารถทำให้ระบบที่ผู้วิจัยพัฒนาจากภาษาพีเอชพี (PHP) สามารถทำงานได้

- **พีเอชพี (PHP)**

ภาษาพีเอชพี (PHP) คือภาษาที่ทำให้ข้อมูลถูกเปลี่ยนแปลงโดยอัตโนมัติตามเงื่อนไขต่าง ๆ ที่ผู้เขียนกำหนด (Dynamic Language) และเป็นภาษาประเภทสคริปต์ (Script) ที่สามารถติดต่อกับผู้ใช้ได้ (กิตติ ภักดีวัฒนะกุล และคณะ 2545) ซึ่งงานวิจัยนี้ใช้เครื่องมือ EditPlus ในการพัฒนาระบบค้นคืนเอกสารโดยใช้ภาษาพีเอชพี (PHP)

- SQL Server 2000

เป็นโปรแกรมฐานข้อมูลที่ใช้เก็บข้อมูลภายในองค์กรต่าง ๆ ซึ่งนิยมใช้กันทั่วไป โดยเป็นฐานข้อมูลเชิงสัมพันธ์ (Relational Database) ของบริษัทไมโครซอฟท์ที่เป็นรุ่นถัดมาของเอสคิวแอลเซิร์ฟเวอร์ (SQL Server) โดยจะสนับสนุนภาษาเอสคิวแอล (SQL) ที่สามารถสอบถาม (Query) วิเคราะห์ (Analyze) ตลอดจนจัดการข้อมูลผ่านเว็บ ด้วยการสนับสนุนภาษาเอกซ์เอ็มแอล (XML) ช่วยในการจัดการข้อมูลทั้งแบบโอแอลทีพี (OLTP: Online Transaction Processing) และโอแอลเอพี (OLAP: Online Analytical Processing) เป็นไปได้อย่างง่ายดาย มีประสิทธิภาพสูงสุดในการจัดเก็บข้อมูลและวิเคราะห์ข้อมูล (สมพร จิวรสกุล 2545) อีกทั้งยังจัดการฐานข้อมูลเชิงสัมพันธ์ที่สนับสนุนการทำ "Two phased Commit" (Tight Consistency) เพื่อรักษาเสถียรภาพของข้อมูลระหว่างเซิร์ฟเวอร์ (Server) หลาย ๆ ตัว

- **ภาษาซี (C Programming Language)**

ภาษาซี (C) เป็นภาษาโปรแกรมเชิงโครงสร้างระดับสูงที่ได้รับการพัฒนาขึ้นในช่วงทศวรรษ 1970 โดย เคน ธอมป์สัน (Ken Thompson) และ เดนนิส ริทชี่ (Dennis Ritchie) ขณะทำงานอยู่ที่ เบลล์เทเลโฟน เลบอราทอรี่ เป็นภาษาโปรแกรมหนึ่งที่ใช้กันแพร่หลายมากที่สุด ภาษาซีมีจุดเด่นที่ประสิทธิภาพในการทำงาน เนื่องจากมีความสามารถใกล้เคียงกับภาษาระดับต่ำแต่เขียนแบบภาษาระดับสูง โปรแกรมคอมพิวเตอร์ที่เขียนด้วยภาษาซีจึงทำงานได้รวดเร็ว ภาษาซีเป็นภาษาโปรแกรมที่นิยมใช้กันมากสำหรับพัฒนาระบบปฏิบัติการ (วิกิพีเดีย สารานุกรมเสรี 2551) ซึ่งงานวิจัยนี้ใช้เครื่องมือ Borland C++ Builder 5

(http://www.borland.com/bcppbuilder/) ในการพัฒนาการจัดกลุ่มเอกสารด้วยภาษาซี (C programming language)

3.2 สมมติฐานงานวิจัย

สำหรับงานวิจัยนี้ผู้วิจัยต้องการศึกษาวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการกำหนดค่าความคล้ายคลึงในการค้นคืนแบบกรอบความคล้ายคลึง สามารถเพิ่มประสิทธิภาพค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคของการค้นคืนเอกสารได้หรือไม่ ดังนั้นงานวิจัยนี้จึงต้องการศึกษาประสิทธิภาพของระบบค้นคืนเอกสารทั้งสองรูปแบบตามที่กำหนดไว้แล้วในหัวข้อตัวแปรต้น โดยจะตั้งสมมติฐานไว้ดังนี้

กำหนดให้

μ_1 คือ ค่าเฉลี่ยค่าความแม่นยำของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_2 คือ ค่าเฉลี่ยค่าความแม่นยำของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_3 คือ ค่าเฉลี่ยค่าความระลึกของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_4 คือ ค่าเฉลี่ยค่าความระลึกของการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมหรือการค้นคืนเอกสารรูปแบบที่ 2

μ_5 คือ ค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคในการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) หรือการค้นคืนเอกสารรูปแบบที่ 1

μ_6 คือ ค่าเฉลี่ยของค่าเฉลี่ยฮาร์โมนิคในการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมหรือการค้นคืนเอกสารรูปแบบที่ 2

1) วิเคราะห์เปรียบเทียบประสิทธิภาพค่าความแม่นยำของการค้นคืนเอกสารทั้ง 2 รูปแบบว่ามีความแตกต่างกันหรือไม่

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

2) วิเคราะห์เปรียบเทียบประสิทธิภาพค่าความระลึกของการค้นคืนเอกสารทั้ง 2 รูปแบบว่ามีความแตกต่างกันหรือไม่

$H_0 : \mu_3 - \mu_4 = 0$

$H_1 : \mu_3 - \mu_4 \neq 0$

3) วิเคราะห์เปรียบเทียบประสิทธิภาพค่าเฉลี่ยฮาร์โมนิคของการค้นคืนเอกสารทั้ง 2 รูปแบบว่ามีความแตกต่างกันหรือไม่

$H_0 : \mu_5 - \mu_6 = 0$

$H_1 : \mu_5 - \mu_6 \neq 0$

หากผลการทดสอบสมมติฐานพบว่าปฏิเสธ H_0 แสดงว่ามีเทคนิคการค้นคืนเอกสารทั้ง 2 รูปแบบที่มีค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคแตกต่างกัน

3.3 แนวทางการทำวิจัย

งานวิจัยนี้เป็นงานวิจัยเชิงประจักษ์ (Empirical Research) เนื่องจากเป็นการศึกษาทดสอบระบบการค้นคืนเอกสารที่ประยุกต์ใช้ทฤษฎี เพื่ออธิบายประสิทธิภาพของระบบค้นคืนเอกสารที่เกิดขึ้น โดยในงานวิจัยนี้สนใจการค้นคืนเอกสารวิธีการวัดความคล้ายคลึงเชิงมุม (Cosine Angle) และการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ภายในการกำหนดกรอบค่าความคล้ายคลึงด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering) ซึ่งผู้วิจัยศึกษาว่าการประยุกต์ใช้ทฤษฎีเทคนิคการจัดกลุ่มข้อมูล (Clustering) บนระยะห่างเชิงยูคลิเดียน สามารถกำหนดกรอบความคล้ายคลึงของเอกสารที่จะถูกค้นคืนได้ตรงตามความต้องการ และสามารถช่วยเพิ่มประสิทธิภาพให้กับระบบค้นคืนเอกสารหรือไม่ ซึ่งในงานวิจัยจะควบคุมตัวแปรอื่น ๆ ให้เหมือนกันเพื่อให้ตัวแปรควบคุมต่าง ๆ ที่กำหนดมีผลกระทบกับตัวแปรตามน้อยที่สุดและผลของงานวิจัยจะได้เป็นผลที่เกิดขึ้นจากการเปลี่ยนเฉพาะตัวแปรต้น

อย่างแท้จริง ตัวแปรที่ต้องควบคุมให้เหมือนกันในงานวิจัยนี้คือ เอกสาร, ข้อสอบถาม, และเครื่องมือทดสอบ นั่นคืองานวิจัยจะทดลองว่าการค้นคืนเอกสารจะมีประสิทธิภาพเปลี่ยนแปลงไปอย่างไรเมื่อใช้วิธีการวัดความคล้ายคลึงที่แตกต่างกัน โดยทดลองสร้างเครื่องมือเพื่อทดสอบประสิทธิภาพของการค้นคืนเอกสารเป็น 2 รูปแบบ ดังนี้

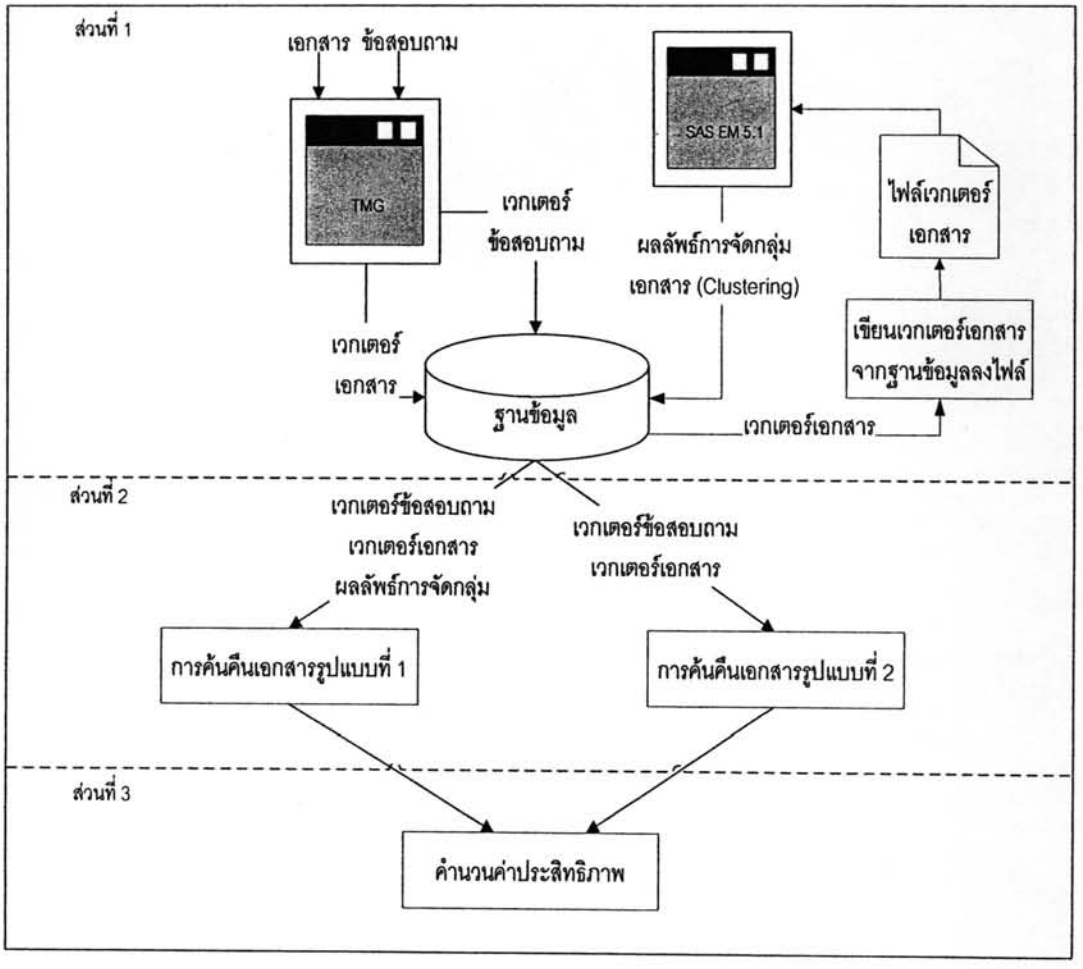
1) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียน (Euclidean Distance) ด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)
2) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม

เนื่องจากผู้วิจัยสนใจว่าวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูลนั้น สามารถเพิ่มประสิทธิภาพของระบบค้นคืนเอกสารได้หรือไม่ ดังนั้นในการสร้างระบบค้นคืนเอกสารจึงต้องสร้างระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุมหรือการค้นคืนเอกสารรูปแบบที่ 2 เป็นกลุ่มควบคุม เพื่อเป็นกลุ่มเปรียบเทียบกับระบบค้นคืนเอกสารที่ได้รับการกระตุ้น นั่นคือการค้นคืนเอกสารรูปแบบที่ 1 เป็นกลุ่มทดลอง

3.4 ภาพรวมการทำงานของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

จากแนวทางการทำวิจัยตามที่ได้กล่าวมาในหัวข้อที่แล้วนั้น ผู้วิจัยวางแผนที่จะพัฒนาเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร 2 รูปแบบ โดยระบบที่พัฒนานั้นพัฒนาด้วยภาษา PHP (Professional Home Page Language) และระบบจัดการฐานข้อมูล (Database Management System) ของ MSSQL Server 2000 การออกแบบการทำงานของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารทั้ง 2 รูปแบบนั้นมีภาพรวมดังรูปที่ 3.1 ซึ่งระบบค้นคืนเอกสารนี้จะแบ่งการสร้างเครื่องมือทดสอบเป็น 3 ส่วน โดยส่วนที่ 1 เป็นส่วนของการแปลงเอกสารและข้อสอบถามให้เป็นเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม นำจัดเก็บลงฐานข้อมูล เพื่อเตรียมข้อมูลไว้ก่อนจะนำข้อมูลเหล่านี้ไปทดสอบต่อในส่วนที่ 2 ซึ่งเครื่องมือทดสอบการค้นคืนเอกสารที่พัฒนาในส่วนที่ 2 เป็นส่วนการค้นคืนเอกสารที่ผู้วิจัยพัฒนาเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบต่าง ๆ นั่นคือ เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม และเทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนด้วยการกำหนดกรอบค่าความคล้ายคลึงการค้นคืนโดยการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) และส่วนสุดท้ายของเครื่องมือทดสอบระบบค้นคืนเอกสารคือ ส่วนที่ 3 เป็นส่วนของการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบระบบค้นคืนเอกสารทั้ง 2 รูปแบบ เพื่อวัดประสิทธิภาพของ

เครื่องมือทดสอบการค้นคืนเอกสารที่พัฒนานั้นสามารถค้นคืนเอกสารที่มีความถูกต้องและครอบคลุมกับความต้องการหรือไม่



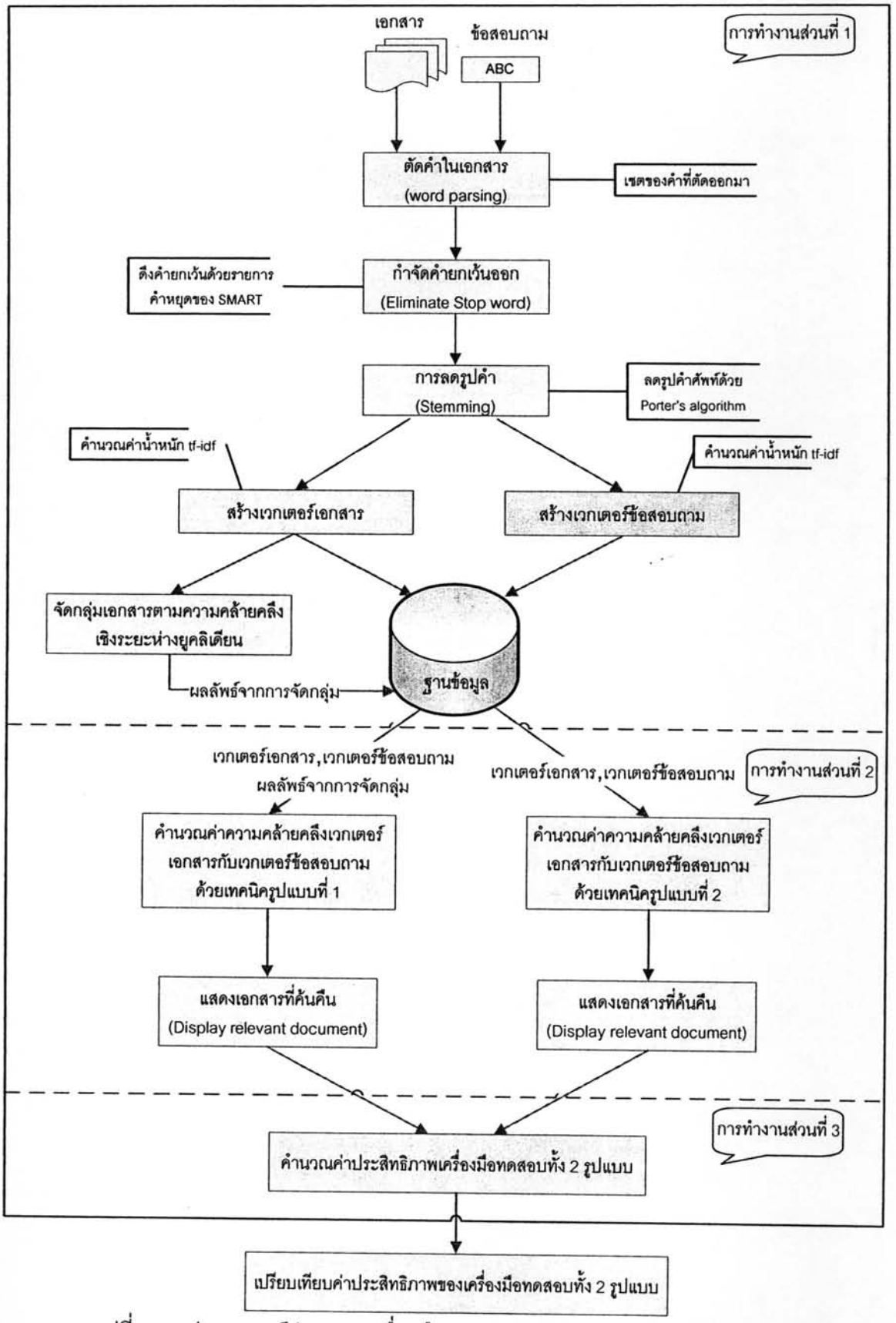
รูปที่ 3.1 รูปแสดงภาพรวมของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารทั้ง 2 รูปแบบ

จากภาพรวมการทำงานของเครื่องมือทดสอบดังรูปที่ 3.1 ในส่วนที่ 1 ขั้นตอนแรกจะนำเอกสารเข้าโปรแกรมทีเอ็มจี (TMG) เพื่อสร้างเวกเตอร์เอกสารเก็บลงฐานข้อมูล จากนั้นจะนำข้อสอบถามเข้าโปรแกรมทีเอ็มจี (TMG) เพื่อสร้างเวกเตอร์ข้อสอบถามเก็บลงฐานข้อมูล และในส่วนนี้จะนำข้อมูลเวกเตอร์เอกสารในฐานข้อมูล เข้าโปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) เพื่อจัดกลุ่มให้กับเอกสาร โดยโปรแกรมจะนำข้อมูลเวกเตอร์เอกสารในฐานข้อมูลออกมาเขียนลงแฟ้มข้อมูล ก่อนนำแฟ้มข้อมูลเวกเตอร์เอกสารนั้นเป็นข้อมูลนำเข้าให้โปรแกรมแซสเอนเตอร์ไพส์ไมน์เนอร์ 5.1 (SAS Enterprise Miner 5.1) เมื่อได้ผลการจัดกลุ่มชุดเอกสารแล้วจะนำค่าสถิติการจัดกลุ่มลงฐานข้อมูลไว้เพื่อนำไปใช้ในการค้นคืนเอกสารส่วนที่ 2

ต่อไป เมื่อเตรียมข้อมูลเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามแล้วจะนำข้อมูลเหล่านี้มาทดสอบในส่วนที่ 2 กับเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 2 รูปแบบตามที่กำหนดไว้ และนำเครื่องมือทดสอบการค้นคืนเอกสารที่พัฒนาขึ้นทั้ง 2 รูปแบบไปคำนวณประสิทธิภาพการค้นคืนเอกสารในส่วนที่ 3 ซึ่งรายละเอียดของขั้นตอนและเทคนิคที่ใช้ในส่วนที่ 1 ส่วนที่ 2 และส่วนที่ 3 จะกล่าวในหัวข้อต่อไป

3.5 องค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

เครื่องมือทดสอบเทคนิคการค้นคืนเอกสารในงานวิจัยนี้ จะแบ่งเป็นส่วนของการทำงานหลัก 3 ส่วน (ภายในกรอบสี่เหลี่ยม) ดังที่กล่าวในหัวข้อภาพรวมของระบบข้างต้น สามารถแสดงรายละเอียดได้ดังรูปที่ 3.2 ซึ่งจะประกอบด้วยรายละเอียดแสดงหลักการที่ใช้และวิธีการของการสร้างเครื่องมือทดสอบการค้นคืนเอกสาร โดยในขั้นตอนในส่วนที่ 1 ขั้นตอนการตัดคำในเอกสาร การกำจัดคำยกเว้นออก การลดรูปคำ การสร้างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม และจัดเก็บลงฐานข้อมูล ในส่วนที่ 2 เป็นส่วนที่แสดงการค้นคืนเอกสารด้วยเทคนิคที่ใช้ในการค้นคืนเอกสาร โดยขั้นตอนในส่วนนี้จะประกอบด้วยขั้นตอนการเปรียบเทียบความเหมือนระหว่างเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารทั้ง 2 รูปแบบ และการแสดงเอกสารแก่ผู้วิจัย และในส่วนที่ 3 เป็นส่วนของการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสารทั้ง 2 รูปแบบ



รูปที่ 3.2 รูปแสดงองค์ประกอบเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

เครื่องมือทดสอบเทคนิคการค้นคืนเอกสารในหัวข้อข้างต้น แบ่งการทำงานของเครื่องมือทดสอบเป็น 3 ส่วน ซึ่งผู้วิจัยจะอธิบายเทคนิคต่าง ๆ ที่นำมาใช้ในแต่ละขั้นตอนทั้ง 3 ส่วน โดยแสดงรายละเอียดดังต่อไปนี้

3.5.1. ส่วนที่ 1 การเตรียมข้อมูลเบื้องต้น

ส่วนการทำงานของเครื่องมือทดสอบในส่วนที่ 1 เป็นส่วนของการเตรียมข้อมูลเอกสาร และข้อสอบถามให้อยู่ในรูปแบบที่สามารถนำไปใช้ในขั้นตอนการค้นคืนเอกสารต่อไปได้ ซึ่งเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 2 รูปแบบมีกระบวนการทำงานในส่วนการเตรียมข้อมูลเบื้องต้นที่เหมือนกัน จะแบ่งส่วนการเตรียมข้อมูลเบื้องต้นนี้ตามขั้นตอนเทคนิคที่ใช้และการทำงานร่วมกัน ดังต่อไปนี้

- **เทคนิคที่ใช้ในส่วนการเตรียมข้อมูลเบื้องต้น**

สามารถแสดงขั้นตอนและเทคนิคที่ใช้ในแต่ละขั้นตอน ดังนี้

1) **ตัดคำในเอกสาร** (word parsing)

เมื่อได้เอกสารหรือข้อสอบถามมาแล้ว จะนำเอกสารหรือข้อสอบถามเหล่านั้นมาตัดแยกคำเป็นคำเดี่ยว ๆ ออกมาจากเอกสาร เนื่องจากในงานวิจัยนี้ใช้ฐานข้อมูลทดสอบที่เป็นเอกสารและข้อสอบถามภาษาอังกฤษ ดังนั้นการตัดคำจะพิจารณาจากช่องว่างระหว่างคำในแต่ละประโยค คำที่แยกด้วยช่องว่างจะถูกตัดออกเป็นคำ 1 คำ แล้วเก็บลงฐานข้อมูลไว้

2) **กำจัดคำที่เป็นคำยกเว้นออก** (Eliminate Stop word)

เมื่อคำเดี่ยว ๆ ที่ได้จากขั้นตอนการตัดคำในประโยคแล้ว จะพิจารณาคำที่มีความหมายที่ไม่เป็นประโยชน์ต่อการสืบค้นออกไป โดยนำมาเทียบกับตารางคำยกเว้น งานวิจัยนี้กำหนดใช้รายการคำยกเว้นจากสมาร์ท (SMART) ซึ่งประกอบด้วยรายการของคำที่เป็นคำยกเว้น 571 คำ (ftp://ftp.cs.cornell.edu/pub/smart/english.stop) ดังแสดงในภาคผนวก ข ถ้าคำที่ตัดมาจากเอกสารหรือข้อสอบถามคำใดเหมือนกับคำที่อยู่ในตารางคำยกเว้นจะตัดคำนั้นทิ้งไปไม่นำมาพิจารณา เพื่อลดคำศัพท์ที่เก็บในระบบให้มีเฉพาะคำที่สำคัญ

3) **ลดรูปคำ** (Stemming)

เมื่อได้คำในเอกสารหรือข้อสอบถามที่กำจัดคำยกเว้นออกแล้ว จากนั้นจะมาเข้าสู่ขั้นตอนวิธีการลดรูปคำ (Stemming) ซึ่งเป็นการวิเคราะห์หารากศัพท์ของคำและลดรูปให้อยู่ในรากศัพท์เดียวกัน เพราะคำที่ปรากฏในเอกสารและข้อสอบถามอยู่ในรูปแบบที่หลากหลาย จึงต้องทำให้คำเหล่านี้อยู่ในรูปแบบเดียวกัน ด้วยวิธีการลดรูปคำ (Stemming) วิธีการลดรูปคำที่ผู้วิจัยนำมาใช้คือขั้นตอนวิธีของพอร์ทเตอร์ (Porter's Algorithm) ซึ่งผู้วิจัยจะใช้ซอร์สโค้ด (Source code) ของ

ขั้นตอนพอร์ทเตอร์ที่สามารถดาวน์โหลด (Download) ได้จากเว็บไซต์ของพอร์ทเตอร์ (Porter) ที่สร้างไว้ที่ http://www.tartarus.org/~martin/PorterStemmer/index.html (Porter, 1980)

4) จัดทำเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถาม

การทำเวกเตอร์เอกสารหรือดรรชนีเป็นวิธีการจัดทำดรรชนีของคำสำคัญที่พบภายในเอกสาร เป็นวิธีการเก็บคำศัพท์ที่ได้จากเอกสารหรือข้อสอบถามที่ผ่านขั้นตอนการตัดคำที่เป็นคำยกเว้นออกและลดรูปคำแล้ว ดรรชนีของคำสำคัญที่ได้จากเอกสารจะถูกเก็บรวบรวมไว้เป็นฐานข้อมูลขนาดใหญ่เพื่อจัดเตรียมไว้สำหรับการสืบค้น เอกสารจะเก็บดรรชนีโดยใช้หลักการดรรชนีแบบผกผัน (Inverted index) ซึ่งเป็นวิธีที่ง่าย และรวดเร็วในการค้นหารายละเอียดกล่าวไว้ในบทที่ 2

เนื่องจากในงานวิจัยนี้ได้สนใจเทคนิคการค้นคืนเอกสารด้วยแบบจำลองปริภูมิเวกเตอร์ที่มีวิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถามแตกต่างกัน ดังนั้นจึงต้องมีการจัดเอกสารและข้อสอบถามให้อยู่ในรูปแบบเวกเตอร์ของคำ ซึ่งนำคำที่ได้จากขั้นตอนการตัดคำยกเว้นและลดรูปคำแล้วมากำหนดเวกเตอร์เอกสารหรือข้อสอบถาม โดยการจัดทำเวกเตอร์เอกสารหรือข้อสอบถามนั้นจะแปลงเอกสารหรือข้อสอบถามในฐานข้อมูลให้อยู่ในรูปแบบเวกเตอร์ โดยค่าแต่ละมิติของเวกเตอร์จะแสดงถึงค่าน้ำหนักของความสำคัญของคำนั้นในเอกสารหรือข้อสอบถาม ซึ่งการคำนวณค่าน้ำหนักของคำในงานวิจัยนี้ ผู้วิจัยได้เลือกวิธีโดยใช้ความถี่ของคำ (Term Frequency: tf) และความถี่ของเอกสารแบบผกผัน (Inverse Document Frequent : idf) ดังที่ได้กล่าวรายละเอียดในบทที่ 2

5) จัดกลุ่มเอกสารตามความคล้ายคลึงเชิงระยะห่างยูคลิเดียน

ในการทำงานของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 1 จะมีการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูลเพื่อจัดกลุ่มเอกสารตามความคล้ายคลึงเชิงระยะห่างยูคลิเดียน แล้วใช้ผลลัพธ์ที่ได้ เช่น ขนาดหรือรัศมีของกลุ่มเอกสาร ในการกำหนดกรอบค่าความคล้ายคลึงของข้อสอบถามแต่ละข้อ ดังนั้นหลังจากการแปลงเอกสารให้อยู่ในรูปแบบเวกเตอร์ ผู้วิจัยจะนำเวกเตอร์ของเอกสารทั้งหมดที่อยู่ในฐานข้อมูลมาทำการจัดกลุ่มของเอกสารด้วยการใช้เทคนิคการจัดกลุ่มข้อมูลแบบ K-means Clustering เนื่องจากการจัดกลุ่มข้อมูลแบบ K-means เป็นเทคนิคการจัดกลุ่มข้อมูลที่วัดระยะห่างระหว่างเอกสารและจุดศูนย์กลางของกลุ่มโดยการใช้ระยะห่างแบบยูคลิเดียน และเป็นเทคนิคที่เหมาะสำหรับจัดกลุ่มข้อมูลที่มีจำนวนมากกว่า 200 ข้อมูลเป็นต้นไป ซึ่งเหมาะสมกับวัตถุประสงค์และจำนวนเอกสารที่จะใช้ในการศึกษาครั้งนี้ โดยผู้วิจัยจะทำการทดลองจัดกลุ่มเอกสารโดยการกำหนดจำนวนกลุ่มของเอกสารหรือค่า K หลาย ๆ ค่า และเลือกค่า K ที่ทำให้ผล

ลัพธ์เป็นกลุ่มเอกสารที่มีคุณภาพดีที่สุด และบันทึกค่าสถิติของกลุ่มเอกสารที่ได้เพื่อนำไปใช้ในขั้นตอนต่อไป

- **การทำงานของส่วนการเตรียมข้อมูลเบื้องต้น**

จากขั้นตอนวิธีและเทคนิคที่ใช้ส่วนของการเตรียมข้อมูลเบื้องต้นทั้ง 5 ขั้นตอนข้างต้น ดังนั้นการทำงานร่วมกันของส่วนการเตรียมข้อมูลเบื้องต้นนั้นสามารถที่จะอธิบายการทำงานได้ดังนี้คือ

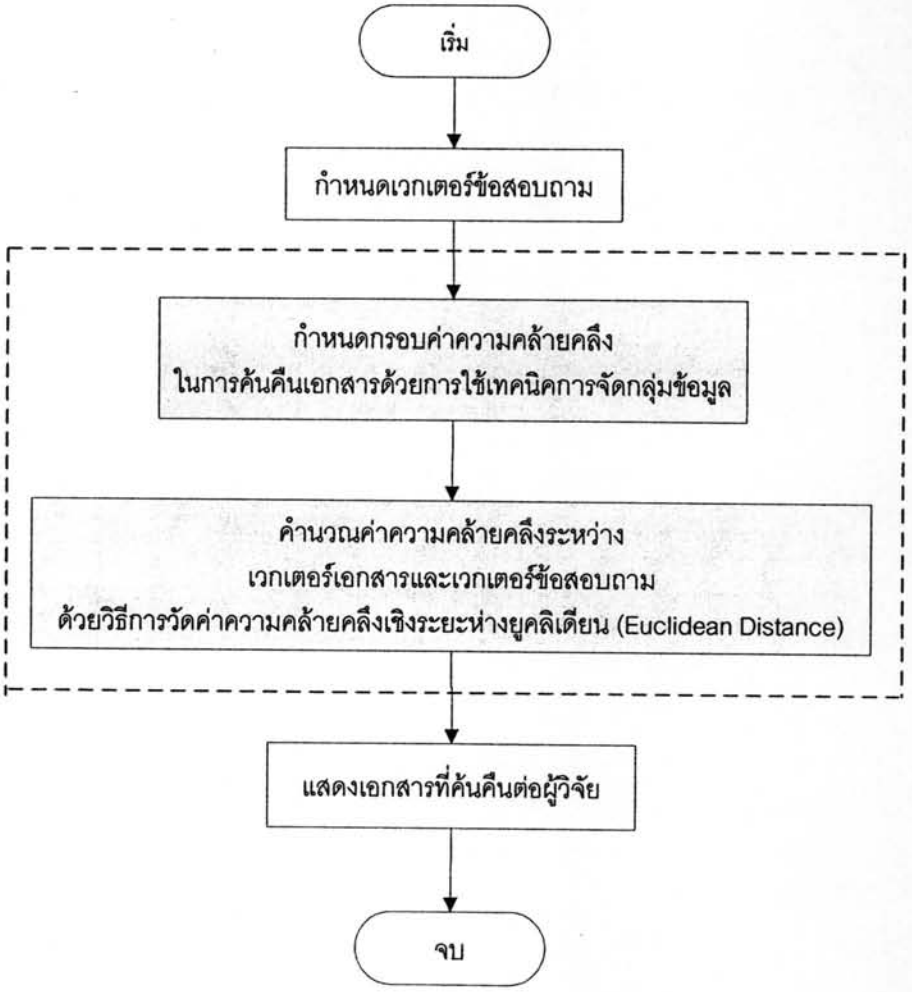
เริ่มแรกเมื่อผู้วิจัยโหลดเอกสารและข้อสอบถามจากฐานข้อมูลทดสอบ เครื่องมือทดสอบจะนำเอกสารนั้นไปผ่านขั้นตอนการตัดคำ การกำจัดคำที่เป็นคำยกเว้นออก และการลดรูปคำศัพท์ที่มีรากศัพท์เดียวกัน ซึ่งผลลัพธ์ที่ได้คือคำสำคัญทั้งหมดที่แสดงในเอกสาร หลังจากนั้นเครื่องมือทดสอบจะนำเอกสารที่ผ่านขั้นตอนทั้ง 3 ไปสร้างเป็นเวกเตอร์เอกสาร โดยที่ค่าน้ำหนักของแต่ละมิติของเวกเตอร์นั้นจะแสดงถึงความสำคัญของคำนั้นในเอกสาร เมื่อได้เวกเตอร์เอกสารแล้ว เครื่องมือทดสอบจะจัดเก็บเวกเตอร์เอกสารลงในฐานข้อมูลพร้อมทั้งนำเวกเตอร์เอกสารที่ได้ไปจัดกลุ่มด้วยเทคนิค K-means Clustering ดังที่ได้กล่าวมาแล้ว ในส่วนของการสร้างเวกเตอร์ของข้อสอบถามจะมีขั้นตอนการทำงานที่เหมือนกับการสร้างเวกเตอร์เอกสาร ซึ่งเมื่อได้สร้างเวกเตอร์ข้อสอบถามแล้ว เครื่องมือทดสอบจะจัดเก็บเวกเตอร์ข้อสอบถามในฐานข้อมูลเช่นเดียวกับเวกเตอร์เอกสาร การสร้างเวกเตอร์ข้อสอบถามจะทำหลังจากการสร้างเวกเตอร์เอกสารเสร็จเรียบร้อยแล้ว นั่นคือหลังจากการกำหนดข้อสอบถามที่ต้องการทดสอบเข้ามายังระบบ

3.5.2. ส่วนที่ 2 การค้นคืนเอกสาร

ส่วนการค้นคืนเอกสาร คือ ส่วนการทำงานของเครื่องมือทดสอบในส่วนที่ 2 ซึ่งเป็นส่วนที่แสดงขั้นตอนการค้นคืนด้วยเทคนิคที่ใช้ในการค้นคืนเอกสารที่แตกต่างกันทั้ง 2 รูปแบบ เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 และ 2 จะแตกต่างกันในส่วนของวิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถาม และในเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 ได้มีการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) เข้ามาช่วยในการกำหนดกรอบของเอกสารที่จะถูกนำมาแสดงให้กับผู้วิจัย ดังนั้นสามารถแบ่งส่วนการค้นคืนเอกสารนี้ตามขั้นตอน เทคนิคที่ใช้ และการทำงานของเครื่องมือทดสอบในแต่ละรูปแบบ ดังรายละเอียดต่อไปนี้

1.) การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนภายในกรอบค่าความคล้ายคลึงที่กำหนดด้วยผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มข้อมูล (Clustering)

ผู้วิจัยได้กำหนดให้การค้นคืนเอกสารวิธีการนี้เป็นเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 ซึ่งมีรูปแบบการใช้วิธีการวัดความคล้ายคลึงเชิงระยะห่างยูคลิเดียนและค้นคืนเอกสารต่อผู้ใช้ภายในกรอบค่าความคล้ายคลึงที่กำหนดโดยการประยุกต์ใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) เข้ามากำหนดกรอบของเอกสารที่จะถูกค้นคืนให้กับผู้วิจัย การทำงานของเครื่องมือทดสอบรูปแบบที่ 1 แสดงได้ดังรูปที่ 3.3



รูปที่ 3.3 รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 1

จากรูปข้างต้นแสดงการทำงานส่วนการค้นคืนเอกสารของเครื่องมือทดสอบรูปแบบที่ 1 ซึ่งส่วนขั้นตอนการทำงานภายในกรอบเส้นประสี่เหลี่ยม เป็นส่วนที่แตกต่างออกไปจากส่วนของเครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 ผู้วิจัยต้องการทดสอบว่าการนำเทคนิคการจัดกลุ่มข้อมูล (Clustering) มากำหนดกรอบค่าความคล้ายคลึงบนระยะห่างเชิงยูคลิเดียน (Euclidean distance) เครื่องมือทดสอบสามารถค้นคืนเอกสารได้ตรงตามความต้องการได้หรือไม่

- **การกำหนดเวกเตอร์ข้อสอบถาม**

การทำงานของเครื่องมือทดสอบรูปแบบที่ 1 เริ่มเมื่อผู้วิจัยเลือกข้อสอบถามที่ต้องการทดสอบ เครื่องมือทดสอบการค้นคืนจะนำข้อสอบถามที่ได้ไปแปลงให้อยู่ในรูปแบบเวกเตอร์ เมื่อได้เวกเตอร์ข้อสอบถามแล้ว เครื่องมือทดสอบการค้นคืนจะนำเวกเตอร์ข้อสอบถามไปกำหนดกรอบค่าความคล้ายคลึงของเอกสาร เพื่อใช้ในการกำหนดกรอบของผลลัพธ์ที่จะค้นคืนต่อไป

- **กำหนดกรอบค่าความคล้ายคลึงในการค้นคืนเอกสาร**

ในขั้นตอนนี้เครื่องมือทดสอบการค้นคืนรูปแบบที่ 1 จะนำเวกเตอร์ข้อสอบถามที่ได้ไปทำการเปรียบเทียบว่า มีความใกล้เคียงกับกลุ่มเอกสารใดในกลุ่มทั้งหมดที่ได้จัดไว้แล้ว ด้วยการพิจารณาหาเอกสารที่อยู่ใกล้กับข้อสอบถามมากที่สุด โดยตั้งสมมติฐานว่า หากข้อสอบถามมีความใกล้เคียงกับเอกสารใด เอกสารนั้นและเอกสารในบริเวณใกล้เคียงก็ควรจะเป็นผลลัพธ์ของการค้นคืน อย่างไรก็ตาม เนื่องจากเอกสารในแต่ละบริเวณ หรือแต่ละกลุ่มมีลักษณะการกระจายตัวที่แตกต่างกัน ดังนั้นจึงนำค่าสถิติของกลุ่มเอกสารนั้น ๆ เช่น ขนาดของกลุ่ม หรือ ค่ารัศมีของกลุ่ม มาใช้กำหนดกรอบค่าความคล้ายคลึงในการค้นคืนเอกสารสำหรับข้อสอบถามนั้น ๆ

- **คำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถาม**

หลังจากกำหนดกรอบค่าความคล้ายคลึงในการค้นคืนเอกสารแล้ว ขั้นตอนการทำงานต่อไปคือ การคำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถาม เพื่อนำไปพิจารณาค้นหาเอกสารที่มีค่าความคล้ายคลึงภายในกรอบที่กำหนด ซึ่งจะหาค่าความคล้ายคลึงด้วยการใช้สูตรการคำนวณระยะห่างยูคลิเดียนจากสมการที่ 2.8

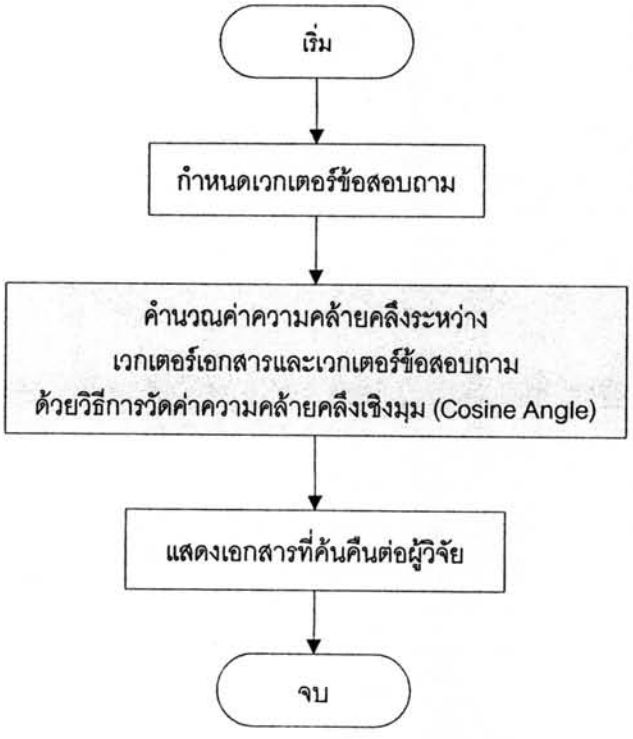
- **แสดงเอกสารแก่ผู้วิจัย**

ขั้นตอนการทำงานนี้เป็นขั้นตอนที่เครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 1 จะแสดงเอกสารที่เป็นผลลัพธ์ของการค้นคืนต่อผู้วิจัย ดังนั้นหลังจากเครื่องมือทดสอบการค้นคืนได้กำหนดกรอบค่าความคล้ายคลึงในการค้นคืนเอกสาร และคำนวณค่าความคล้ายคลึงระหว่างเอกสารและข้อสอบถามนั้น ๆ แล้ว เครื่องมือทดสอบจะต้องค้นหาเอกสารที่มีค่าความคล้ายคลึงกับข้อสอบถามภายใต้กรอบค่าความคล้ายคลึงที่ได้กำหนดไว้เพื่อแสดงต่อผู้วิจัยทางหน้าจอคอมพิวเตอร์

และในการแสดงเอกสารต่อผู้วิจัยนั้นเครื่องมือทดสอบจะกำหนดลำดับการแสดงเอกสารด้วยค่าความคล้ายคลึงกับข้อสอบถามจากมากไปหาน้อย เพื่อให้เอกสารที่มีความคล้ายคลึงกับข้อสอบถามนั้นถูกแสดงในลำดับต้น ๆ

2.) **การค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ด้วยวิธีการวัดความคล้ายคลึงเชิงมุม**

ผู้วิจัยได้กำหนดให้การค้นคืนเอกสารวิธีการนี้เป็นเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ซึ่งเป็นเครื่องมือทดสอบที่ใช้วิธีการวัดความคล้ายคลึงเชิงมุมระหว่างเอกสารและข้อสอบถาม การทำงานของเครื่องมือทดสอบรูปแบบที่ 2 แสดงได้ดังรูปที่ 3.4



รูปที่ 3.4 รูปแสดงการทำงานของเครื่องมือทดสอบรูปแบบที่ 2

จากรูปข้างต้นแสดงการทำงานส่วนการค้นคืนเอกสารของเครื่องมือทดสอบรูปแบบที่ 2 ซึ่งสามารถแสดงขั้นตอนและเทคนิคทั้งหมดที่ใช้ในแต่ละขั้นตอนการทำงาน แสดงได้ดังนี้

- **การกำหนดเวกเตอร์ข้อสอบถาม**

เริ่มเมื่อผู้วิจัยเลือกข้อสอบถามที่ต้องการทดสอบเข้ามา เครื่องมือทดสอบการค้นคืนจะนำข้อสอบถามที่ได้ไปแปลงให้อยู่ในรูปแบบเวกเตอร์ และเครื่องมือทดสอบการค้นคืนจะดึงเวกเตอร์เอกสารทั้งหมดที่เก็บไว้ในฐานข้อมูลออกมา เพื่อจะดำเนินการในส่วนของขั้นตอนต่อไป

- **คำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์เอกสารกับเวกเตอร์ข้อสอบถาม**

เมื่อได้เตรียมข้อมูลของเอกสารและข้อสอบถามในรูปแบบเวกเตอร์แล้ว ขั้นตอนการทำงานต่อไปคือ การคำนวณค่าความคล้ายคลึงระหว่างเอกสารกับข้อสอบถาม โดยเครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 จะคำนวณหาค่าความคล้ายคลึงโดยหาความสัมพันธ์ (Correlation) ระหว่างทุกเวกเตอร์เอกสารและเวกเตอร์ข้อสอบถามนั้น ๆ ด้วยวิธีคำนวณค่าความเหมือนโคไซน์ (Cosine coefficient) จากสมการที่ 2.7 ค่าความคล้ายคลึงที่ได้ จะนำไปพิจารณาหาเอกสารที่จะเป็นผลลัพธ์ของการค้นคืนในขั้นตอนต่อไป

- **แสดงเอกสารแก่ผู้วิจัย**

หลังจากคำนวณค่าความคล้ายคลึงแล้ว เครื่องมือทดสอบจะค้นคืนเอกสารที่มีค่าความคล้ายคลึงตามที่กำหนดไว้ออกมาแสดงต่อผู้วิจัยทางหน้าจอคอมพิวเตอร์ ในงานวิจัยของ Udomchaiporn (2005) ได้เสนอว่าการกำหนดค่าความคล้ายคลึงสามารถกำหนดได้ตามความเหมาะสมซึ่งแตกต่างไปตามชุดเอกสาร ซึ่งในงานวิจัยของศิรัตน์ ศิรนานนท์ (2549) ได้ใช้ฐานข้อมูลนิตยสารไทม์ (TIME Collection) ในการทดสอบ ซึ่งเป็นฐานข้อมูลเดียวกับที่ผู้วิจัยนำมาพัฒนาเครื่องมือทดสอบการค้นคืนในงานวิจัยนี้ นอกจากนี้งานวิจัยของศิรัตน์ ศิรนานนท์ได้ใช้วิธีการคำนวณค่าความคล้ายคลึงเชิงมุมระหว่างเอกสารและข้อสอบถาม เช่นเดียวกับเครื่องมือทดสอบการค้นคืนเอกสารรูปแบบที่ 2 ดังนั้นผู้วิจัยจึงกำหนดค่าความคล้ายคลึงไว้ที่ค่าเฉลี่ย (Mean) บวกค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) ของค่าความคล้ายคลึงทุกข้อสอบถามกับทุกเอกสารเช่นเดียวกับที่กำหนดในงานวิจัยของศิรัตน์ ศิรนานนท์

เครื่องมือทดสอบการค้นคืนรูปแบบที่ 2 จะค้นคืนเอกสารที่มีค่าความคล้ายคลึงระหว่างเอกสารและข้อสอบถามนั้น ๆ มากกว่าค่าความคล้ายคลึงที่กำหนดไว้ ออกมาแสดงต่อผู้วิจัยทางหน้าจอคอมพิวเตอร์

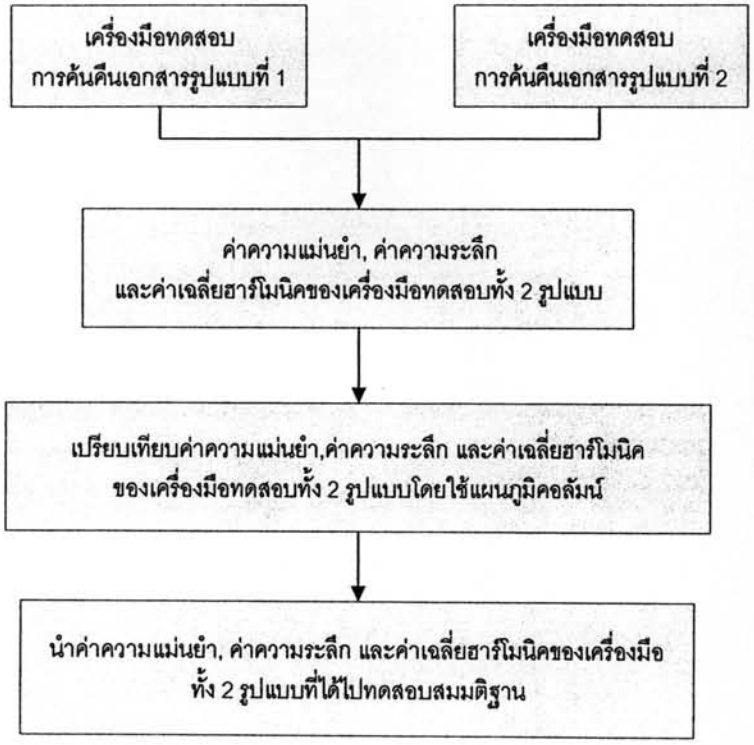
3.5.3. ส่วนที่ 3 การคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

เป็นส่วนการทำงานของเครื่องมือทดสอบการค้นคืนทั้ง 2 รูปแบบในส่วนสุดท้าย ซึ่งเป็นส่วนของการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบที่ถูกพัฒนาในงานวิจัยนี้ทั้ง 2 รูปแบบ ในเครื่องมือทดสอบเมื่อมีการค้นคืนเอกสารออกมาแสดงให้กับผู้วิจัยแล้วก็จะมาทำการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบ เพื่อพิจารณาว่าเครื่องมือทดสอบนั้นสามารถค้นคืนเอกสารมีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด งานวิจัยนี้ใช้วิธีการคำนวณค่าประสิทธิภาพของเครื่องมือทดสอบทั้ง 2 รูปแบบด้วยค่าประสิทธิภาพการค้นคืน 3 ค่า คือ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) ซึ่งรายละเอียดการคำนวณในแต่ละค่าประสิทธิภาพได้กล่าวไว้ในบทที่ 2 ผู้วิจัยคำนวณค่าประสิทธิภาพการค้นคืน 3 ค่าด้วยกัน เนื่องจากเป็นวิธีการคำนวณค่าประสิทธิภาพที่นำมาใช้ในระบบการค้นคืนเอกสารส่วนใหญ่ (Baeza-Yates and Ribeiro-Neto 1999) ในการคำนวณค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคนั้นจะต้องรู้จำนวนเอกสารที่เกี่ยวเนื่องกับข้อสอบถามแต่ละข้อสอบถาม โดยเอกสารที่ถูกต้องสำหรับข้อสอบถามแต่ละข้อได้ถูกกำหนดไว้ในฐานข้อมูลทดสอบของนิตยสารไทม์ (TIME Collection) แล้ว จากนั้นเครื่องมือทดสอบนำค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคที่คำนวณได้เก็บลงฐานข้อมูล และแสดงค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคนั้นทางหน้าจอแก่ผู้วิจัยด้วย เป็นอันเสร็จสิ้นขั้นตอนการทำงานของเครื่องมือทดสอบการค้นคืนเอกสารที่พัฒนาขึ้น

3.6 การทดสอบประสิทธิภาพของเครื่องมือทดสอบเทคนิคการค้นคืนเอกสาร

เครื่องมือทดสอบการค้นคืนเอกสารทั้ง 2 รูปแบบ พัฒนาขึ้นเพื่อทดสอบประสิทธิภาพการค้นคืนเอกสารว่ามีความแตกต่างกันหรือไม่ ถ้าแตกต่างเครื่องมือทดสอบการค้นคืนรูปแบบใดจะให้ประสิทธิภาพการค้นคืนที่ดีกว่า โดยจุดมุ่งหมายหลักของการวัดประสิทธิภาพการค้นคืนเอกสารคือ เอกสารที่เกี่ยวเนื่องกับข้อสอบถามจะถูกค้นคืนออกมาแสดงต่อผู้วิจัย (Baeza-Yates and Ribeiro-Neto, 1999) ดังนั้นเมื่อสร้างระบบค้นคืนเอกสารทั้ง 2 รูปแบบเสร็จสิ้นแล้ว ต่อจากนั้นจะวัดประสิทธิภาพการค้นคืนเอกสารด้วยค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) (ตามที่ได้กล่าวรายละเอียดการคำนวณไว้ในบทที่ 2) มาเป็นค่าแสดงประสิทธิภาพของเครื่องมือทดสอบ การทดสอบประสิทธิภาพจะแสดงด้วยการเปรียบเทียบค่าความแม่นยำ, ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคของเครื่องมือทดสอบทั้ง 2 รูปแบบด้วยการใช้กราฟเส้นแสดงการเปรียบเทียบ เพื่อง่ายต่อการพิจารณาประสิทธิภาพ ซึ่ง

ขั้นตอนโดยสรุปของการทดสอบประสิทธิภาพเครื่องมือทดสอบการค้นคืนทั้ง 2 รูปแบบ แสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 รูปแสดงขั้นตอนการทดสอบประสิทธิภาพของเครื่องมือทดสอบทั้ง 2 รูปแบบ

3.7 ความถูกต้อง (Validity) และความน่าเชื่อถือ (Reliability)

งานวิจัยนี้เป็นการทดลองศึกษาระบบค้นคืนเอกสารที่ใช้เทคนิคการค้นคืนรูปแบบที่ต่างกัน ซึ่งเทคนิคการค้นคืนรูปแบบที่ต่างกันนี้ จัดเป็นตัวแปรต้นที่เป็นปัจจัยที่ต้องเปลี่ยนค่าไปตามแผนแบบการทดลองเพื่อดูความแตกต่างอันเกิดขึ้นจากการทดลอง เพื่อให้ผลวิจัยมีความเชื่อถือได้ (Reliability) และถูกต้อง (Validity) จำเป็นต้องควบคุมปัจจัยอื่นที่เกี่ยวข้องซึ่งอาจส่งผลต่อตัวแปรตามอันได้แก่ การเลือกชุดเอกสารทดสอบ การเลือกข้อสอบถาม และการวัดประสิทธิภาพของระบบ โดยมีรายละเอียดดังนี้

1) การเลือกชุดเอกสารทดสอบและข้อสอบถาม

- ชุดเอกสารและข้อสอบถามที่ผู้วิจัยนำมาใช้ในการทดลองระบบค้นคืนเอกสารทั้ง 2 รูปแบบนั้น เป็นข้อมูลที่ได้นำมาจากฐานข้อมูลที่ถูกสร้างขึ้นเพื่อให้นักวิจัยได้นำมาใช้ทดสอบ

ระบบค้นคืนเอกสารที่พัฒนาขึ้น (Smart Collection 1963) และเป็นฐานข้อมูลที่ใช้ในการทดสอบงานวิจัยทางด้านการค้นคืนเอกสารมากมาย (Salton 1971; Chowdhury 2004)

- ชุดเอกสารที่นำมาทดสอบในงานวิจัยนี้เป็นเอกสารที่เกี่ยวกับข่าวสารทั่วไป ทำให้สามารถเป็นตัวแทนของเอกสารที่มีความหลากหลายประเภทและสามารถนำไปประยุกต์ใช้กับเอกสารเพื่อการตัดสินใจทางธุรกิจได้

- ชุดเอกสารและข้อสอบถามที่นำมาใช้ทดลองในระบบค้นคืนเอกสารทั้ง 2 รูปแบบ เป็นชุดเอกสารและข้อสอบถามกลุ่มเดียวกัน เพื่อให้ประสิทธิภาพการทดลองเกิดจากเทคนิคที่แตกต่างกันอย่างแท้จริง

- ฐานข้อมูลชุดเอกสารและข้อสอบถามที่นำมาทดสอบ ได้มีการกำหนดข้อสอบถามและผลเฉลยซึ่งเป็นรายการเอกสารที่เป็นคำตอบของแต่ละข้อสอบถามไว้ชัดเจนแล้ว โดยกลุ่มคนที่มีความเชี่ยวชาญในชุดเอกสารและข้อสอบถามนั้น (Salton 1971) ซึ่งทำให้กระบวนการพิจารณาผลลัพธ์ที่ระบบแสดงออกมาในขั้นตอนการทดสอบระบบค้นคืนเอกสารในงานวิจัยนี้จะสามารถเชื่อถือความถูกต้องของผลเฉลยของรายการเอกสารที่เป็นผลลัพธ์ของข้อสอบถาม

2) การวัดประสิทธิภาพของระบบ

- การวัดประสิทธิภาพระบบค้นคืนเอกสารทั้ง 2 รูปแบบ จะวัดด้วยค่าความแม่นยำ (Precision), ค่าความระลึก (Recall) และค่าเฉลี่ยฮาร์โมนิค (Harmonic mean) มาเป็นค่าแสดงประสิทธิภาพของระบบค้นคืนเอกสารทดสอบ (ศิรัตน์ ศิรนานนท์ 2549; Cyril and Eric 2005; Greenwood 2002) ซึ่งเป็นค่าแสดงประสิทธิภาพที่นิยมใช้ในการทดสอบระบบการค้นคืนสารสนเทศ (Baeza-Yates and Ribeiro-Neto 1999) โดยจะเป็นค่าที่แสดงการวัดว่าระบบสามารถค้นคืนเอกสารออกมาได้ถูกต้องและครอบคลุมตรงกับความต้องการของผู้วิจัยหรือไม่

3.8 กรอบการวิเคราะห์ข้อมูล (Data Analysis Framework)

งานวิจัยนี้ต้องการเปรียบเทียบประสิทธิภาพของการค้นคืนเอกสาร (Document Retrieval) ด้วยวิธีการวัดความคล้ายคลึงระหว่างเอกสารและข้อสอบถามที่ต่างกัน ซึ่งทดสอบประสิทธิภาพของระบบการค้นคืนเอกสารที่พัฒนาด้วยค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิค เมื่อทดสอบประสิทธิภาพของระบบค้นคืนเอกสารที่พัฒนาทั้ง 2 รูปแบบเสร็จสิ้นแล้ว ทำให้ได้ค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคออกมาเท่ากับจำนวนของข้อสอบถามที่ทดสอบ คืออย่างละ 83 ค่า ซึ่งข้อมูลค่าประสิทธิภาพนี้จะถูกนำไปวิเคราะห์เพื่อตอบวัตถุประสงค์ และทดสอบสมมติฐานของงานวิจัย ดังนั้นจะต้องตรวจสอบการแจกแจงของค่าประสิทธิภาพที่ได้มาว่า มีการแจกแจงปกติ (Normal Distribution) หรือไม่ เพื่อเลือกทางเลือกใน

การทดสอบสมมติฐาน หากการแจกแจงค่าประสิทธิภาพเป็นปกติ จะทดสอบสมมติฐานด้วยสถิติทดสอบแบบที (t-test) ถ้าหากการแจกแจงค่าประสิทธิภาพไม่ปกติจะต้องมีการใช้การทดสอบสมมติฐานแบบไม่อิงกับพารามิเตอร์ (Nonparametric Test) (กัลยา วาณิชย์บัญชา, 2546) ดังนั้น งานวิจัยนี้เป็นการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของค่าประสิทธิภาพในการค้นคืนของเครื่องมือทดสอบการค้นคืนเอกสารทั้ง 2 รูปแบบ และไม่ทราบค่าความแปรปรวนของประชากร จึงใช้สถิติทดสอบแบบที (t-test) เพื่อทดสอบสมมติฐานที่ตั้งไว้ ซึ่งจะสามารถพิจารณาเปรียบเทียบค่าความแม่นยำ ค่าความระลึก และค่าเฉลี่ยฮาร์โมนิคของระบบค้นคืนเอกสารทั้ง 2 รูปแบบได้