การค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น

นางสาวปรารถนา จันพลโท

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2550

WEB PAGE RETRIEVAL FROM USER RELEVANCE FEEDBACK USING WEB PAGE SEGMENTATION AND PROBABILISTIC MODEL

Miss Prattana Chanpolto

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2007

หัวข้อวิทยานิพนธ์	การค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น
โดย	นางสาวปรารถนา จันพลโท
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ นครทิพย์ พร้อมพูล

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร. บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร. วันชัย ริ้วไพบูลย์)

..... อาจารย์ที่ปรึกษา
(ผู้ช่วยศาสตราจารย์ นครทิพย์ พร้อมพูล)

..... กรรมการ
(รองศาสตราจารย์ ดร. พรศิริ หมื่นไชยศรี)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. ดาริชา สุธีวงศ์)

ปรารถนา จันพลโท : การค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น. (WEB PAGE RETRIEVAL FROM USER RELEVANCE FEEDBACK USING WEB PAGE SEGMENTATION AND PROBABILISTIC MODEL) อ. ที่ปรึกษา : ผศ. นครทิพย์ พร้อมพูล, 104 หน้า.

เว็บเพจเป็นหนึ่งในสารสนเทศที่ได้รับการพัฒนาขึ้นจากความก้าวหน้าของเทคโนโลยีสารสนเทศในปัจจุบัน เว็บเพจประกอบด้วยเนื้อหาหรือสารสนเทศที่มีรูปแบบหลากหลายชนิด เว็บเบราว์เซอร์เป็นส่วนสำคัญของเทคโนโลยีอินเทอร์เน็ตมีกลไกเพื่อรองรับการสืบค้นข้อมูลสารสนเทศจากหลายแหล่ง กระบวนการค้นคืนสารสนเทศแบบเว็บเพื่อให้ได้เว็บเพจที่ตรงกับความต้องการของผู้ใช้นั้นจะค้นคืนด้วยการคำนวณความคล้ายกันจากข้อคำถามเปรียบเทียบกับคำที่ปรากฏในเว็บเพจ ซึ่งผลการค้นคืนที่ได้อาจจะไม่ตรงตามความต้องการของผู้ใช้เท่าที่ควร จึงจำเป็นต้องทำการค้นคืนใหม่โดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ช่วยให้ผลการค้นคืนที่ได้ตรงตามความต้องการของผู้ใช้ได้มากขึ้น

งานวิทยานิพนธ์นี้จึงมีวัตถุประสงค์เพื่อวิเคราะห์และออกแบบระบบสำหรับการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม และเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ที่ใช้ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองความน่าจะเป็น โดยวีไอพีเอสอัลกอริทึมนั้นเป็นเทคนิคที่ใช้ในการแบ่งเว็บเพจเป็นส่วนย่อยหรือบล็อก ซึ่งผู้ใช้จะให้ผลป้อนกลับด้วยการเลือกบล็อกที่เห็นว่าตรงประเด็นกับที่ต้องการ และคำที่ปรากฏในบล็อกที่เลือก จะนำมากำหนดข้อคำถามใหม่ พร้อมทั้งเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น งานวิจัยนี้ได้พัฒนาเครื่องมือเพื่อทดสอบแนวคิดที่นำเสนอ และสามารถประเมินประสิทธิผลของระบบการค้นคืนเว็บเพจด้วยค่าเรียกคืนและค่าความแม่นยำ

ผลการทดลองที่ได้จากงานวิทยานิพนธ์นี้แสดงให้เห็นว่า ที่ระดับนัยสำคัญ 0.05 การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ให้ผลค่าความแม่นยำมากกว่าวิธีการค้นคืนแบบไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ 69.69 เปอร์เซ็นต์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นให้ผลค่าความแม่นยำมากกว่าวิธีการค้นคืนแบบไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ 26.59 เปอร์เซ็นต์ เนื่องจากแบบจำลองปริภูมิเวคเตอร์สามารถเปลี่ยนแปลงได้ทั้งคำและค่าน้ำหนักของคำในข้อคำถาม ในขณะที่แบบจำลองความน่าจะเป็นเปลี่ยนแปลงเฉพาะค่าน้ำหนักของคำในข้อคำถามเท่านั้น

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....ปรารถนา.....จันพลโท.....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่ออาจารย์ที่ปรึกษา.....นครทิพย์ พร้อมพูล.....

ปีการศึกษา 2550

4770346421 : MAJOR COMPUTER SCIENCE
KEY WORD : RELEVANCE FEEDBACK / WEB PAGE SEGMENTATION / PROBABILISTIC MODEL / QUERY REFORMULATION / QUERY EXPANSION / TERM REWEIGHTING

PRATTANA CHANPOLTO : WEB PAGE RETRIEVAL FROM USER RELEVANCE FEEDBACK USING WEB PAGE SEGMENTATION AND PROBABILISTIC MODEL. THESIS ADVISOR : ASST.PROF. NAKORNTHIP PROMPOON, 104 pp.

Web page is one of various kinds of information developed in the current information technology advancement. A web page contains difference types and forms of content or information. Web browser, an important part of internet technology, provides a search mechanism to retrieve information from various sources. In web information retrieval process, retrieving a web page to meet user requirements is based on the similarity computation between terms in a query and terms in a web page. The query results may not meet user requirements so it is necessary to retrieve again with user relevance feedback which can improve the result of web information retrieval.

The objective of this thesis is to analyze and design a system an approach for web page retrieval from user relevance feedback using web page segmentation with Vision based Page Segmentation (VIPS) Algorithm and term reweighting in a new query with probabilistic model. The VIPS algorithm is a technique to segment a web page into several blocks. User may select relevant blocks meet his/her need. Terms contain in the selected blocks are used to produce a new query and to recompute the weight of terms in a new query using probabilistic model. As a result, this research develops a tool to test the proposed approach. The two widely used metrics named recall and precision are use to evaluate web page retrieval results.

The results of our experiment indicate that, for a level of significance 0.05, the precision value of our approach, using VIPS algorithm and vector space model (first model) is 69.69 percent greater than web page retrieval without user relevance feedback, and the precision value using VIPS algorithm and the probabilistic model (second model) is 26.59 percent greater than web page retrieval without user relevance feedback. The first model can adjusts both terms and weight of terms in the query whereas the second model can adjusts only weight of terms in that query.

Department : Computer Engineering Student's Signature Prattana Chanpolto

Field of Study : Computer Science Advisor's Signature Nakornthip Prompoon

Academic Year : 2007

กิตติกรรมประกาศ

ขอขอบพระคุณอาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ นครทิพย์ พร้อมพูล ที่เสียสละเวลาช่วยเหลือให้คำปรึกษา คำแนะนำและข้อคิดเห็นที่มีประโยชน์ต่องานวิจัยนี้ ความรู้และประสบการณ์ในงานวิชาการอื่นๆ รวมถึงคำสอน ข้อคิดด้านคุณธรรม จริยธรรม และการขลัดเกลาความรู้ความสามารถให้ดียิ่งขึ้น ทำให้งานวิทยานิพนธ์นี้ ประสบความสำเร็จด้วยดี

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ทุกท่าน รองศาสตราจารย์ ดร.วันชัย ริ้วไพบูลย์ รองศาสตราจารย์ ดร.พรศิริ หมื่นไชยศรี และ ผู้ช่วยศาสตราจารย์ ดร.ดาริชา สุธีวงศ์ ที่กรุณาสละเวลาในการให้คำแนะนำ ข้อคิดเห็นต่างๆที่เป็นประโยชน์อย่างยิ่งต่องานวิจัย รวมทั้งการตรวจสอบความถูกต้องสมบูรณ์ของวิทยานิพนธ์ฉบับนี้

ขอขอบพระคุณคณาจารย์ทุกท่านในภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ให้ความรู้ คำแนะนำในการเรียน และการทำวิจัย

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ ห้องปฏิบัติการวิศวกรรมซอฟต์แวร์ทุกคน และเพื่อนวิทยาศาสตร์คอมพิวเตอร์ทุกท่าน ที่คอยช่วยเหลือ ให้กำลังใจ และคำแนะนำที่ดีเสมอมา

สุดท้ายนี้ ขอกราบขอบพระคุณ บิดา มารดา และสมาชิกในครอบครัวทุกท่าน ที่คอยให้กำลังใจ ผลักดัน และสนับสนุนในทุกๆด้าน แก่ผู้วิจัยตลอดมา จนสำเร็จการศึกษา

สารบัญ

หน้า

หน้า

หน้า

สารบัญตาราง

หน้า

หน้า

สารบัญภาพ

หน้า

หน้า

บทที่ 1
บทนำ

ในบทนี้ จะกล่าวถึงแนวคิดหลักของงานวิจัย อันประกอบไปด้วย ที่มาและความสำคัญ วัตถุประสงค์ ขอบเขต ประโยชน์ ขั้นตอนและโครงสร้างของเนื้อหางานวิจัย และบทความทางวิชาการที่ได้รับการตีพิมพ์ ซึ่งมีเนื้อหาดังต่อไปนี้

1.1 ที่มาและความสำคัญของงานวิจัย

ความก้าวหน้าทางด้านเทคโนโลยีสารสนเทศในปัจจุบันทำให้เกิดรูปแบบของสารสนเทศขึ้นมากมาย ซึ่งระบบสารสนเทศนั้นจัดได้ว่ามีความจำเป็นในการดำรงชีวิตของมนุษย์ เช่น เพื่อการศึกษาหาความรู้ เพื่อติดตามข่าวสารข้อมูลในปัจจุบัน และที่ได้รับความนิยมเป็นอย่างมากคือ ข้อมูลทางอินเทอร์เน็ต (Internet) ผ่านสื่อประเภทเว็บเพจ (Web Page) เพราะสามารถเข้าถึงกลุ่มผู้สนใจได้ทั่วโลก และผู้ใช้จากทุกแห่งหนที่สามารถติดต่อเข้าสู่ระบบอินเทอร์เน็ตได้ ก็สามารถเรียกดูข้อมูลได้ตลอดเวลา ดังนั้นข้อมูลในระบบอินเทอร์เน็ตจึงสามารถเผยแพร่ได้รวดเร็ว และกว้างไกลตลอดจนข้อมูลสามารถนำเสนอข้อมูลได้หลากหลายรูปแบบเช่น ตัวอักษร ข้อความ หรือรูปภาพ

เว็บเพจนั้นจะมีโครงสร้างของหน้าเว็บเพจที่แตกต่างกัน และข้อมูลที่ปรากฏในหน้าเว็บเพจสามารถแบ่งออกได้เป็นส่วนๆ โดยที่แต่ละส่วนจะแสดงข้อมูลที่แตกต่างกัน และเป็นอิสระจากกัน ตามโครงสร้างที่ได้กำหนดไว้ในแต่ละหน้าเว็บเพจ ซึ่งส่วนใหญ่จะแบ่งหน้าเว็บเพจออกเป็น 3 ส่วน คือ ส่วนหัว ส่วนเนื้อหา และส่วนท้าย โดยส่วนหัวเป็นส่วนที่ประกอบด้วยชื่อเว็บหรือชื่อเรื่องของเว็บเพจนั้นๆ ส่วนเนื้อหาเป็นส่วนที่ประกอบด้วยเนื้อหาซึ่งจะมีใจความสำคัญที่ตรงประเด็นกับส่วนชื่อเรื่องของแต่ละเว็บเพจ และส่วนท้ายของหน้าเว็บเพจเป็นส่วนของบริเวณที่จะให้ข้อมูลเพิ่มเติมเกี่ยวกับเนื้อหาและเว็บเพจ

เว็บเพจจัดได้ว่าเป็นสารสนเทศที่ประกอบด้วยเนื้อหาหรือข้อมูลที่มีความหลากหลาย โดยการจัดเก็บและค้นคืนเว็บเพจจะมีหลักการเดียวกันกับการจัดเก็บและค้นคืนสารสนเทศ (Information Storage and Retrieval System) คือ การกำหนดลักษณะที่เหมาะสมสำหรับใช้เป็นตัวแทนของเอกสาร (Document Representation) การกำหนดข้อคำถาม (Query) การกำหนดฟังก์ชันในการคำนวณหาค่าความคล้ายระหว่างเอกสารแบบเว็บเพจกับข้อคำถาม (Similarity Function) ตลอดจนการจัดอันดับผลของเอกสารที่ค้นคืนได้จากการทดสอบความคล้ายกันนั้น โดยใช้ฟังก์ชันในการจัดอันดับ (Ranking Function) ที่เหมาะสม

ในการค้นคืนเว็บเพจจะทำการเปรียบเทียบความคล้ายระหว่างคำในข้อคำถาม กับคำที่ปรากฏในเอกสารเว็บเพจ แล้วแสดงผลการค้นคืนที่ได้เป็นรายการเอกสารเว็บเพจที่ตรงประเด็นกับข้อคำถามนั้นๆ โดยทั่วไปการค้นคืนเว็บเพจจะทำการเปรียบเทียบคำในข้อคำถามกับเอกสารแบบเว็บเพจที่ปรากฏคำๆ นั้นทั้งหน้าเว็บเพจ เช่น ถ้าต้องการค้นคืนเอกสารที่มีคำว่า

Information Retrieval ก็จะค้นหาคำว่า Information Retrieval ที่ปรากฏในทุกส่วนของหน้าเว็บเพจ โดยผลการค้นคืนที่ได้ อาจจะไม่ตรงตามความต้องการของผู้ใช้ ถ้าข้อคำถามที่ใช้ในการค้นคืนนั้นตรงกับคำที่ปรากฏในส่วนที่ไม่สำคัญของเว็บเพจ เช่น โฆษณา แต่เอกสารนั้นก็ถูกนำมาแสดงในผลลัพธ์รายการเอกสารที่ค้นคืนได้ ซึ่งเป็นปัญหาสำคัญ ดังนั้นเพื่อให้ได้ผลการค้นคืนตามความต้องการของผู้ใช้ จึงต้องทำการให้ผลป้อนกลับที่ตรงประเด็น (Relevance Feedback) [1] โดยการกำหนดข้อคำถามใหม่ (Query Reformulation) เพราะการให้ผลป้อนกลับที่ตรงประเด็น จะใช้ข้อมูลที่ได้กลับมาจากผู้ใช้เพื่อนำมากำหนดข้อคำถามใหม่ที่เหมาะสมกว่าข้อคำถามเดิม ซึ่งจะช่วยให้ระบบรู้ได้ว่าเอกสารใดบ้างที่ตรงประเด็นและตรงตามความต้องการของผู้ใช้งาน เพื่อที่ระบบจะได้นำข้อมูลเหล่านั้นไปใช้ในการค้นคืนครั้งถัดไป ส่งผลให้ได้มาซึ่งเอกสารที่ตรงตามความต้องการ และสามารถทำการจัดอันดับได้อย่างเหมาะสมมากยิ่งขึ้น

แม้ว่าจะทำการกำหนดข้อคำถามใหม่เพื่อใช้ในการให้ผลป้อนกลับที่ตรงประเด็นแล้ว แต่ยังคงมีปัญหาเกิดขึ้นคือ การกำหนดข้อคำถามใหม่นั้นจะมีการปรับเปลี่ยนคำหรือการเพิ่มคำ (Term) เข้าไปในข้อคำถามเดิม ซึ่งคำเหล่านั้นอาจจะเป็นคำที่ปรากฏในส่วนที่ไม่สำคัญของหน้าเว็บเพจ เพื่อแก้ปัญหาดังกล่าวจึงนำวีไอพีเอส อัลกอริทึม [2] (Vision-based Pages Segmentation :VIPS algorithm) มาใช้ในการแบ่งเว็บเพจที่ได้จากการค้นคืนครั้งแรกเป็นบล็อก (Block) เพื่อแสดงให้เห็นว่าข้อคำถามนั้นปรากฏในส่วนใดของเอกสารเว็บเพจที่ค้นคืนมาได้ แล้วผู้ใช้จะให้ผลป้อนกลับ (Feedback) ด้วยการเลือกบล็อก ที่ตรงประเด็นกับข้อคำถามที่ต้องการ ซึ่งจะช่วยในการกำหนดข้อคำถามใหม่ ในการให้ผลป้อนกลับที่ตรงประเด็นจากบล็อกที่ผู้ใช้เลือก แทนที่การเลือกคำจากทุกส่วนของหน้าเว็บเพจ

จากงานวิจัย [3] พบว่าการนำวีไอพีเอสอัลกอริทึม มาใช้ในการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียม (Pseudo Relevance Feedback) ซึ่งเป็นการให้ผลป้อนกลับที่ตรงประเด็นแบบหนึ่งที่ไม่สนใจผลตอบกลับจากผู้ใช้ แต่ทำการเลือกคำที่จะมากำหนดข้อคำถามใหม่จากรายการเอกสารที่ถูกจัดอันดับแล้วว่าตรงประเด็นกับข้อคำถามใน 10 อันดับแรก ผลปรากฏว่าสามารถเพิ่มประสิทธิผลของการค้นคืนได้เพิ่มขึ้น 27 เปอร์เซ็นต์ เมื่อเปรียบเทียบกับการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมโดยไม่ได้ใช้ วีไอพีเอสอัลกอริทึม และจากงานวิจัย [4] นั้นพบว่าการให้ค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น (Probabilistic Model) สามารถช่วยเพิ่มประสิทธิผลของการให้ผลป้อนกลับที่ตรงประเด็นได้ โดยเฉพาะในส่วนของการเปลี่ยนแปลงค่าน้ำหนักของคำ

ดังนั้นงานวิจัยนี้จึงเสนอวิธีการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น มาประยุกต์เข้าด้วยกัน โดยส่วนของวีไอพีเอสอัลกอริทึมนั้น จะใช้ในการแบ่งเว็บเพจออกเป็นบล็อก ซึ่งผู้ใช้จะเป็นผู้ตัดสินใจว่าบล็อกที่ได้จากการแบ่งโดยวีไอพีเอสอัลกอริทึมนั้นบล็อกใดบ้างที่ตรงประเด็นกับที่ต้องการ เพื่อช่วยในการเลือกคำที่จะนำมากำหนดข้อคำถามใหม่จากบล็อกเหล่านั้น แล้วทำการเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น ซึ่งข้อคำถามใหม่ที่ได้นำไปใช้ใน

การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยทำการเปรียบเทียบค่าความคล้ายระหว่างข้อคำถามใหม่ที่ได้ กับเอกสารเว็บเพจในฐานข้อมูล แล้วแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้อีกครั้ง สำหรับผลลัพธ์ที่ได้ จากวิธีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นที่ผู้วิจัยได้นำเสนอนี้ จะใช้ค่าความแม่นยำและค่าเรียกคืน ในการประเมินประสิทธิผลของระบบ

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อศึกษาและออกแบบวิธีการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อย และใช้แบบจำลองความน่าจะเป็น

2) เพื่อพัฒนาเครื่องมือในการสนับสนุนการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อย และใช้แบบจำลองความน่าจะเป็น

1.3 ขอบเขตของงานวิจัย

1) สามารถทำการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ได้เฉพาะเอกสารแบบเว็บเพจที่เป็นภาษาอังกฤษเท่านั้น

2) งานวิจัยนี้จะเปรียบเทียบผลการค้นคืนเว็บเพจที่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น กับผลการค้นคืนเว็บเพจที่ไม่ได้ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

3) งานวิจัยนี้จะใช้ค่าของความแม่นยำในการวัดประสิทธิผลของการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

4) ในการค้นคืนเว็บเพจนั้นไม่ได้มีการพิจารณาถึงเรื่องความกำกวมของคำ

5) ทดสอบการทำงานเครื่องมือด้วยข้อคำถามอย่างน้อย 50 ข้อคำถาม

1.4 ประโยชน์ของงานวิจัย

สามารถทำการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็นได้ และทำให้ได้เครื่องมือในการสนับสนุนการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อย และใช้แบบจำลองความน่าจะเป็น

1.5 ขั้นตอนและวิธีการวิจัย

1) ศึกษาระบบการจัดเก็บและค้นคืนสารสนเทศที่เหมาะสมกับการค้นคืนเว็บเพจ

2) วิเคราะห์ปัญหาของการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้และแนวทางในการแก้ไข

3) วิเคราะห์และออกแบบโครงสร้างของข้อมูล และโครงสร้างแฟ้มข้อมูลที่ใช้ตามแนววิธีนี้

4) วิเคราะห์และออกแบบขั้นตอนในการจัดเก็บ การค้นคืน และประเมินผลการค้นคืน

5) วิเคราะห์และออกแบบขั้นตอนในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้และประเมินผลการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยการนำวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมาใช้ พร้อมทั้งสร้างเครื่องมือเพื่อรองรับการทำงานทั้งหมด

6) พัฒนาเครื่องมือจากแนวคิดและวิธีการที่นำเสนอ เพื่อใช้ในการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และประเมินผลการทดสอบการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

7) จัดทำเอกสารวิทยานิพนธ์และข้อเสนอแนะ

1.6 โครงสร้างของเนื้อหางานวิจัย

สำหรับโครงสร้างเนื้อหาทั้งหมดของวิทยานิพนธ์ฉบับนี้แบ่งออกเป็น 6 บทดังนี้ คือ บทที่ 1 เป็นบทนำกล่าวถึงที่มาและความสำคัญของปัญหา วัตถุประสงค์ของงานวิจัย เป็นต้น บทที่ 2 จะกล่าวถึงทฤษฎีและงานวิจัยต่างๆ ที่เกี่ยวข้อง เช่น ระบบจัดเก็บและการค้นคืนสารสนเทศ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ แบบจำลองความน่าจะเป็น เป็นต้น บทที่ 3 กล่าวถึงขั้นตอนวิธีการในการค้นคืนเว็บเพจ และวิธีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ รวมถึงวิธีการประเมินประสิทธิผลของวิธีการที่ทำการวิจัย บทที่ 4 จะกล่าวถึงการออกแบบและพัฒนาเครื่องมือที่สนับสนุนตามขั้นตอนวิธีที่นำเสนอ บทที่ 5 กล่าวถึงการทดลองและผลการทดลองของขั้นตอนวิธีที่นำเสนอ และบทที่ 6 ซึ่งเป็นบทสุดท้ายจะเป็นบทสรุปของการวิจัย รวมทั้งข้อเสนอแนะต่างๆ เพื่อให้การพัฒนาระบบมีความสามารถเพิ่มขึ้นในอนาคต

1.7 บทความทางวิชาการ

ในงานวิจัยนี้ ผู้วิจัยมีผลงานวิชาการร่วมกับคณะผู้วิจัย เป็นบทความวิชาการระดับชาติ 1 บทความ ดังแสดงใน ภาคผนวก จ. ได้แก่

บทความวิชาการเรื่อง “การค้นคืนย้อนกลับจากผู้ใช้ในสารสนเทศแบบเว็บโดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น (User Relevance Feedback for Web Information Retrieval Using Web Page Segmentation and Probabilistic Model)” ซึ่งได้รับการคัดเลือกเพื่อนำเสนอและตีพิมพ์ในงาน “การประชุมวิชาการร่วมสาขาวิทยาการคอมพิวเตอร์และวิศวกรรมซอฟต์แวร์ ครั้งที่ 3 (The 3rd Joint Conference on Computer Science and Software Engineering: JCSSE 2006)” ระหว่างวันที่ 29 – 30 มิถุนายน 2549 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า วิทยาเขตพระนครเหนือ กรุงเทพฯ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ จะกล่าวถึงทฤษฎีที่สำคัญ ที่นำมาประยุกต์ใช้ สนับสนุน และอ้างอิง รวมถึงงานวิจัยต่างๆ ที่เกี่ยวข้อง โดยมีเนื้อหาดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

งานวิทยานิพนธ์นี้เกี่ยวข้องกับการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการนำวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น มาประยุกต์เข้าด้วยกัน ซึ่งทฤษฎีที่ตรงประเด็นมีดังนี้

2.1.1 ระบบการจัดเก็บและค้นคืนสารสนเทศ (Information Storage and Retrieval System)

ระบบการจัดเก็บและค้นคืนสารสนเทศ [5] เป็นระบบที่มีการจัดเก็บสารสนเทศเพื่อใช้ในการประมวลผล การค้นคืนสารสนเทศ รวมทั้งนำเสนอในรูปแบบที่เหมาะสม ซึ่งระบบค้นคืนสารสนเทศจะช่วยให้ผู้ใช้สามารถค้นหาข้อมูลที่ตรงตามความต้องการได้สะดวกรวดเร็วขึ้น โดยผู้ใช้จะค้นหาข้อมูลที่ตนเองต้องการด้วยการใช้ข้อคำถามซึ่งประกอบด้วยคำ สำหรับใช้ค้นหาสารสนเทศที่ต้องการ ระบบจะทำการค้นคืนเอกสาร ด้วยการเปรียบเทียบความคล้ายระหว่างตัวแทนเอกสารในคอลเลคชันกับข้อคำถาม ซึ่งผลลัพธ์ที่ได้จะเป็นรายการเอกสารที่มีค่าความคล้ายกับข้อคำถามมากกว่าหรือเท่ากับค่าขีดแบ่งเริ่มต้นความคล้ายที่กำหนดไว้ แต่ผลลัพธ์ที่ได้จะตรงกับความต้องการผู้ใช้มากน้อยเพียงใดก็ขึ้นอยู่กับ การออกแบบและพัฒนาระบบนั้นๆว่ามีความสามารถเพียงใด กระบวนการในการค้นคืนสารสนเทศนั้นแสดงได้ดังรูปที่ 2.1 ซึ่งประกอบด้วยส่วนต่างๆ ดังนี้

1) ส่วนต่อประสานกับผู้ใช้ (User Interface) ส่วนนี้เป็นส่วนเชื่อมต่อการดำเนินงานต่างๆ ระหว่างผู้ใช้ กับระบบสารสนเทศ ซึ่งผู้ใช้จะป้อนข้อคำถามเข้าสู่ระบบเพื่อให้ระบบทำการค้นคืนเอกสารที่ตรงประเด็นกับข้อคำถาม เป็นส่วนแสดงผลตอบกลับที่ได้จากการค้นคืนสารสนเทศแก่ผู้ใช้งาน และเป็นส่วนที่ผู้ดูแลระบบใช้ในการติดต่อกับระบบในการนำเข้าสารสนเทศต่างๆ เพื่อจัดเก็บเข้าสู่ฐานข้อมูล

2) ส่วนจัดการฐานข้อมูล (Database Management Module) ทำหน้าที่ในการจัดการงานต่างๆ ที่เกี่ยวกับฐานข้อมูล

3) ฐานข้อมูล (Database) ทำหน้าที่จัดเก็บข้อมูล เอกสารต่างๆ ที่ผู้ใช้ป้อนเข้าสู่ระบบ

4) การดำเนินการข้อความ เป็นการจัดการกับข้อความ ซึ่งเป็นการเตรียมเอกสารก่อนนำไปทำดรรชนี เช่นกระบวนการในการตัดคำ (Word Segmentation) การกำจัดคำยกเว้น (Elimination of Stop Words) การลดรูปคำ (Stemming)

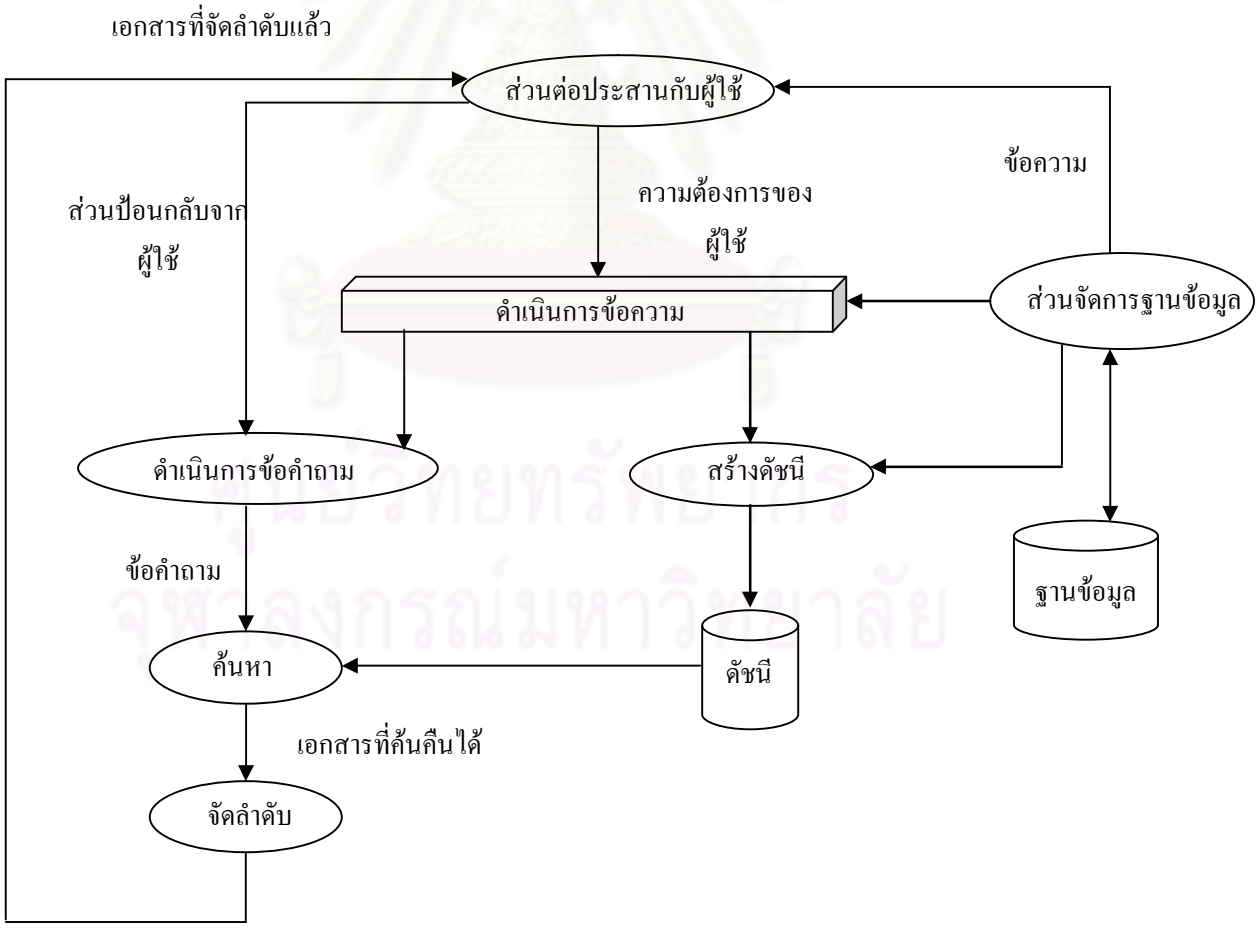
5) การดำเนินการข้อคำถาม ในส่วนนี้จะเป็นการเปลี่ยนความต้องการของผู้ใช้ ให้เป็นข้อคำถาม ซึ่งมีขั้นตอนคล้ายกับการดำเนินการข้อความ เพื่อให้ได้คำที่จะใช้ในการค้นคืนสารสนเทศ และยังทำหน้าที่ในการเปลี่ยนแปลงรูปแบบข้อคำถามที่รับมาจากผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ เพื่อทำการให้ผลป้อนกลับที่ตรงประเด็น

6) การทำดรรชนี เป็นส่วนของการสร้างดรรชนี เพื่อให้ได้คำในการแสดงถึงตัวแทนของเอกสารต่างๆ ที่มีในฐานข้อมูล โดยคำที่ได้นั้นผ่านขั้นตอนการดำเนินการข้อความเรียบร้อยแล้ว และจัดเก็บตามโครงสร้างที่ได้ออกแบบไว้

7) แฟ้มดรรชนี ส่วนนี้ทำหน้าที่ในการจัดเก็บดรรชนี ที่ได้มาจากขั้นตอนการสร้างดรรชนี เพื่อใช้เป็นตัวแทนของเอกสาร

8) การค้นหา เป็นส่วนที่ทำหน้าที่ในการค้นคืนเอกสารโดยทำการเปรียบเทียบความคล้ายระหว่างข้อคำถามที่ได้จากผู้ใช้ กับดรรชนีซึ่งเป็นตัวแทนของแต่ละเอกสาร แล้วแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้

9) การจัดลำดับ เป็นส่วนที่ใช้ในการจัดลำดับผลที่ได้การค้นคืนเอกสารตามค่าความคล้ายกลับสู่ผู้ใช้งานผ่านทางส่วนต่อประสานกับผู้ใช้

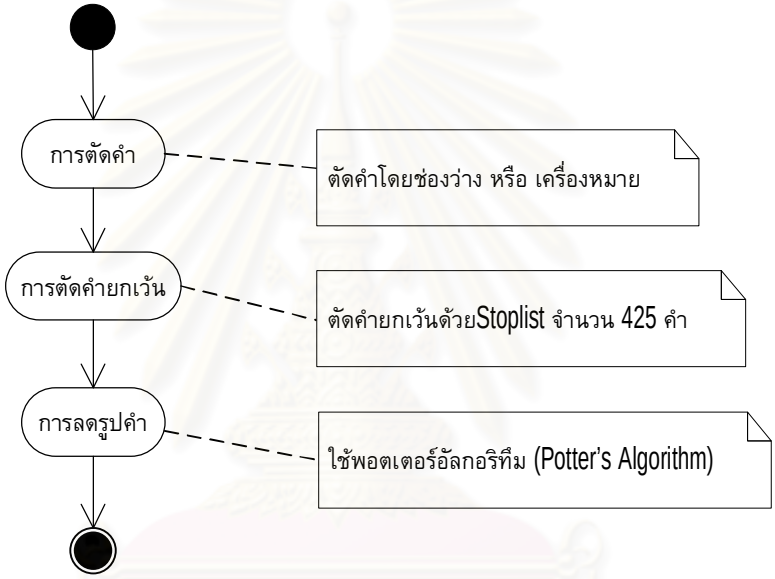


รูปที่ 2.1 กระบวนการในระบบการจัดเก็บและค้นคืนสารสนเทศ

ในกระบวนการจัดเก็บและค้นคืนสารสนเทศของงานวิจัยนี้แบ่งเป็น 3 ส่วนหลักๆ คือ การจัดเก็บสารสนเทศ การค้นคืนสารสนเทศ และการประเมินผลการค้นคืนสารสนเทศ ซึ่งมีดังนี้

2.1.1.1 **การจัดเก็บสารสนเทศ** (Information Storage)

ในการจัดเก็บสารสนเทศนั้นจะต้องอาศัยโครงสร้างพื้นฐานของการทำดรรชนี การให้ค่าน้ำหนักคำ การกำหนดโครงสร้างของการจัดเก็บข้อมูล เพื่อใช้ในการดำเนินการข้อความ หรือเอกสารที่ต้องการจัดเก็บเข้าสู่ฐานข้อมูล โดยในงานวิจัยนี้จะใช้การทำดรรชนีอัตโนมัติ (Automatic Indexing) ซึ่งมีการตัดคำ การตัดคำยกเว้น และการลดรูปคำ โดยใช้พอตเตอร์อัลกอริทึม (Porter's Algorithm) ก็จะได้ชุดของดรรชนี เพื่อใช้ในการอ้างอิงในการค้นคืนเอกสารต่อไป โดยมีขั้นตอนการทำงานดังรูปที่ 2.2



รูปที่ 2.2 แผนภาพกิจกรรมการทำดรรชนีอัตโนมัติ

1) การตัดคำ (Word Segmentation) พิจารณาจากเครื่องหมายวรรคตอน (Punctuation Marks) หรือช่องว่างระหว่างคำ

2) การตัดคำยกเว้น (Elimination of Stop words) โดยคำยกเว้นที่ถูกตัดออกไปในขั้นตอนนี้มีทั้งหมด 425 คำ [5] เนื่องจากเป็นคำที่ไม่สามารถเป็นตัวแทนของเอกสารได้ และยังช่วยลดขนาดของแฟ้มดรรชนีได้

3) การลดรูปคำ (Stemming) เป็นการตัดคำนำหน้า (Prefix) และคำต่อท้าย (Suffix) ออกจากคำให้เหลือเพียงรากคำศัพท์ (Root Word) ของคำนั้นๆ โดยใช้พอตเตอร์อัลกอริทึม (Porter's Algorithm) [6]

หลังจากทำดรรชนีอัตโนมัติแล้วก็จะมีการให้ค่าน้ำหนักของคำที่ปรากฏอยู่ในแฟ้มดรรชนี เพื่อเป็นการบอกความสำคัญของคำนั้นๆ ในการเป็นตัวแทนของเอกสาร โดยมีวิธีการ

คำนวณในการหาค่าน้ำหนักของคำจากหนังสือ [7] จำนวน 4 วิธี คือใช้ความถี่ของคำมาตรฐาน (Term Frequency: TF) ใช้ค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน (Inverse Document Frequency: IDF) ใช้ค่าอัตราส่วนของสัญญาณรบกวน (The Signal-Noise Ratio) และใช้ค่าความถี่ของคำและค่าการแบ่งแยกคำ (The Term Discrimination Value)

โดยการให้ค่าน้ำหนักของคำนั้นงานวิจัยนี้ใช้แบบค่าความถี่ของคำและความถี่ของเอกสารแบบผกผันหรือ IDF (Inverse Document Frequency) ซึ่งการให้น้ำหนักคำเป็นวิธีการประมาณค่าความสำคัญของคำ ซึ่งมีสมมุติฐานคือ ความสำคัญของคำ (Term Importance) เป็นสัดส่วนตามกับความถี่ของคำหรือเทอมที่ k ที่เกิดขึ้นใน เอกสารที่ i ($Freq_{ik}$) และเป็นสัดส่วนผกผันกับจำนวนเอกสารทั้งหมดที่มีคำหรือเทอมที่ k ปรากฏอยู่ ($TotFreq_k$) มีสูตรการคำนวณดังสมการที่ (1)

$$Weight_{ik} = \frac{Freq_{ik}}{TotFreq_k} \qquad (1)$$

โดยที่ $Weight_{ik}$ คือ ค่าน้ำหนักของคำ k ในเอกสาร i

$Freq_{ik}$ คือ ค่าความถี่ของคำ k ในเอกสาร i

$TotFreq_k$ คือ ความถี่ของคำ k ที่ปรากฏในฐานข้อมูล

2.1.1.2 **การค้นคืนสารสนเทศ** (Information Retrieval)

เมื่อผู้ใช้มีการติดต่อกับระบบผ่านส่วนต่อประสานงานกับผู้ใช้ เพื่อค้นคืนเอกสารโดยการป้อนข้อคำถามเข้าสู่ระบบ จากนั้นระบบจะเริ่มต้นดำเนินการของการค้นคืน คือทำการเปรียบเทียบความคล้ายของคำในข้อคำถามกับชุดของดรรชนีที่มีในฐานข้อมูล ซึ่งเปรียบเสมือนตัวแทนของเอกสาร ซึ่งงานวิจัยนี้จะใช้ค่าสัมประสิทธิ์โคซายน์(Cosine Coefficient) [1] ดังสมการที่ 2

$$Similarity\,(DOC_i, Query_j) = \frac{\sum_{k=1}^{t}(Term_{ik} \cdot QTerm_{jk})}{\sqrt{\sum_{k=1}^{t}(Term_{ik})^2 \cdot \sum_{k=1}^{t}(QTerm_{jk})^2}} \qquad (2)$$

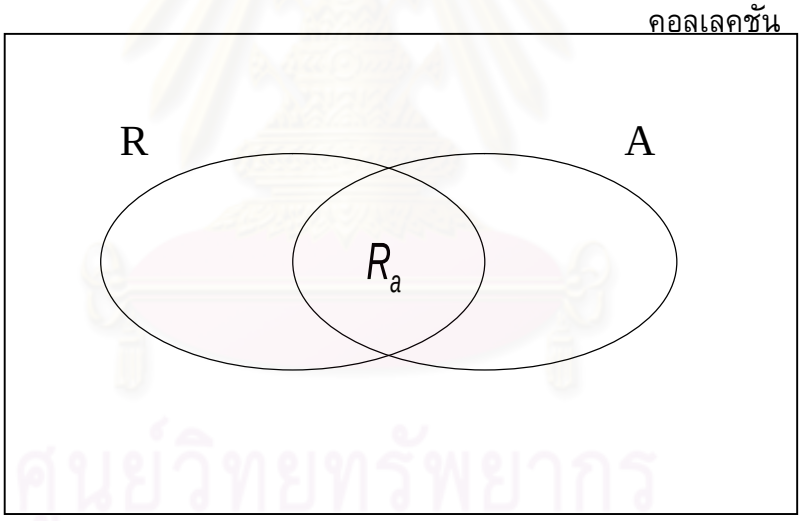
โดยที่ $Similarity(DOC_i, Query_j)$ คือ ความคล้ายระหว่างเอกสารที่ i กับข้อคำถามที่ j

$Term_{ik}$ คือ ค่าน้ำหนักของคำ k ในเอกสารที่ i

$QTerm_{jk}$ คือ ค่าน้ำหนักของคำ k ในข้อคำถามที่ j

2.1.1.3 **การประเมินผลการค้นคืนสารสนเทศ** (Evaluation of Information Retrieval)

หลังจากทำการค้นคืนเอกสารเรียบร้อยแล้วระบบก็จะแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้งานโดยเรียงลำดับเอกสารที่ได้ตามค่าความคล้ายจากมากไปน้อย ซึ่งสามารถประเมินประสิทธิผล ของการค้นค้นคืน [5] ได้ด้วย ค่าเรียกคืน (Recall) และค่าความแม่นยำ (Precision) ซึ่งค่าเรียกคืน เป็นปริมาณที่แสดงความครอบคลุม เช่น ถ้าฐานข้อมูลมีเอกสารที่ตรงตามต้องการทั้งสิ้น R ฉบับ และการค้นคืนสามารถดึงเอกสารที่ตรงตามต้องการได้ Ra ฉบับ ค่าเรียกคืนเป็น Ra/R ส่วนค่าความแม่นยำเป็นปริมาณที่แสดงว่าค้นคืนเอกสารได้ตรงตามต้องการเพียงใด เช่น ถ้าค้นคืนเอกสารออกมาได้ A ฉบับ และมีเอกสารอยู่ Ra ฉบับที่ตรงตามต้องการ ดังนั้นค่าความแม่นยำมีค่าเป็น Ra/A หรือเป็นโอกาสของเอกสารที่ค้นคืนออกมาตรงตามต้องการ ทั้งค่าเรียกคืนและค่าความแม่นยำมีค่าอยู่ระหว่าง 0 ถึง 1 ค่าเรียกคืนและค่าความแม่นยำมีความสัมพันธ์แบบปฏิภาคผกผันคือ หากค่าเรียกคืนสูงค่าความแม่นยำก็จะต่ำ และในทางตรงกันข้าม หากค่าความแม่นยำสูง ค่าเรียกคืนก็จะต่ำ [5] แสดงได้ดังรูปที่ 2.3 และสมการที่ 3 และ สมการที่ 4



รูปที่ 2.3 ค่าความแม่นยำและค่าเรียกคืน

$$\text{ค่าเรียกคืน (Recall)} = \frac{|R_a|}{|R|} \quad (3)$$

$$\text{ค่าความแม่นยำ(Precision)} = \frac{|R_a|}{|A|} \quad (4)$$

โดยที่ R_a คือ จำนวนเอกสารตรงประเด็นที่ค้นคืนออกได้

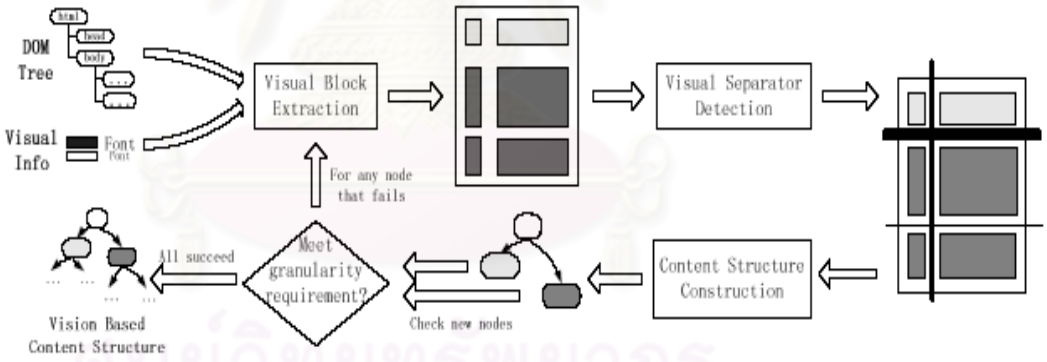
A คือ จำนวนเอกสารทั้งหมดที่ค้นคืนออกมา

R คือ จำนวนเอกสารที่ตรงประเด็นทั้งสิ้นในฐานข้อมูล

2.1.2 วีไอพีเอสอัลกอริทึม (Vision-based Pages Segmentation: VIPS algorithm)

ในการแยกแยะความแตกต่างของเว็บเพจนั้น จำเป็นที่จะต้องทำการแบ่งเว็บเพจออกเป็นบล็อก ซึ่งก็มีวิธีการหลายวิธีในการแบ่งเว็บเพจ แต่ที่ได้รับความนิยมคือ แบบดอม (DOM-based segmentation) [8] แบบตำแหน่ง (location-based segmentation) [9] และแบบวีไอพีเอสอัลกอริทึม ซึ่งจากงานวิจัย [2] พบว่า วีไอพีเอสอัลกอริทึม นั้นสามารถเพิ่มประสิทธิผลในการค้นคืนสารสนเทศได้ดีที่สุด

วีไอพีเอสอัลกอริทึม เป็นอัลกอริทึมที่ใช้ในการแบ่งหน้าเว็บเพจออกเป็นบล็อก เพื่อให้ง่ายต่อการค้นคืนสารสนเทศ ซึ่งมาจากการรวมโครงสร้างแบบดอม (DOM Structure) และโครงสร้างเนื้อหา (Vision-based Content Structure) จุดประสงค์หลักของวีไอพีเอสอัลกอริทึม คือ เพื่อทำการแบ่งเว็บเพจให้มีลักษณะเป็นโครงสร้างของเนื้อหาสำคัญ โดยที่โครงสร้างของเนื้อหาที่สำคัญนั้นจะอยู่บนพื้นฐานของการแบ่งเนื้อหาที่สำคัญตามความหมายซึ่งสามารถช่วยให้ส่วนที่ไม่สำคัญในเว็บเพจนั้นถูกตัดออกไป โดยกระบวนการในการแบ่งเว็บเพจออกเป็นบล็อกนั้น สามารถแสดงได้ดังรูปที่ 2.4



รูปที่ 2.4 ขั้นตอนการทำงานของวีไอพีเอสอัลกอริทึม

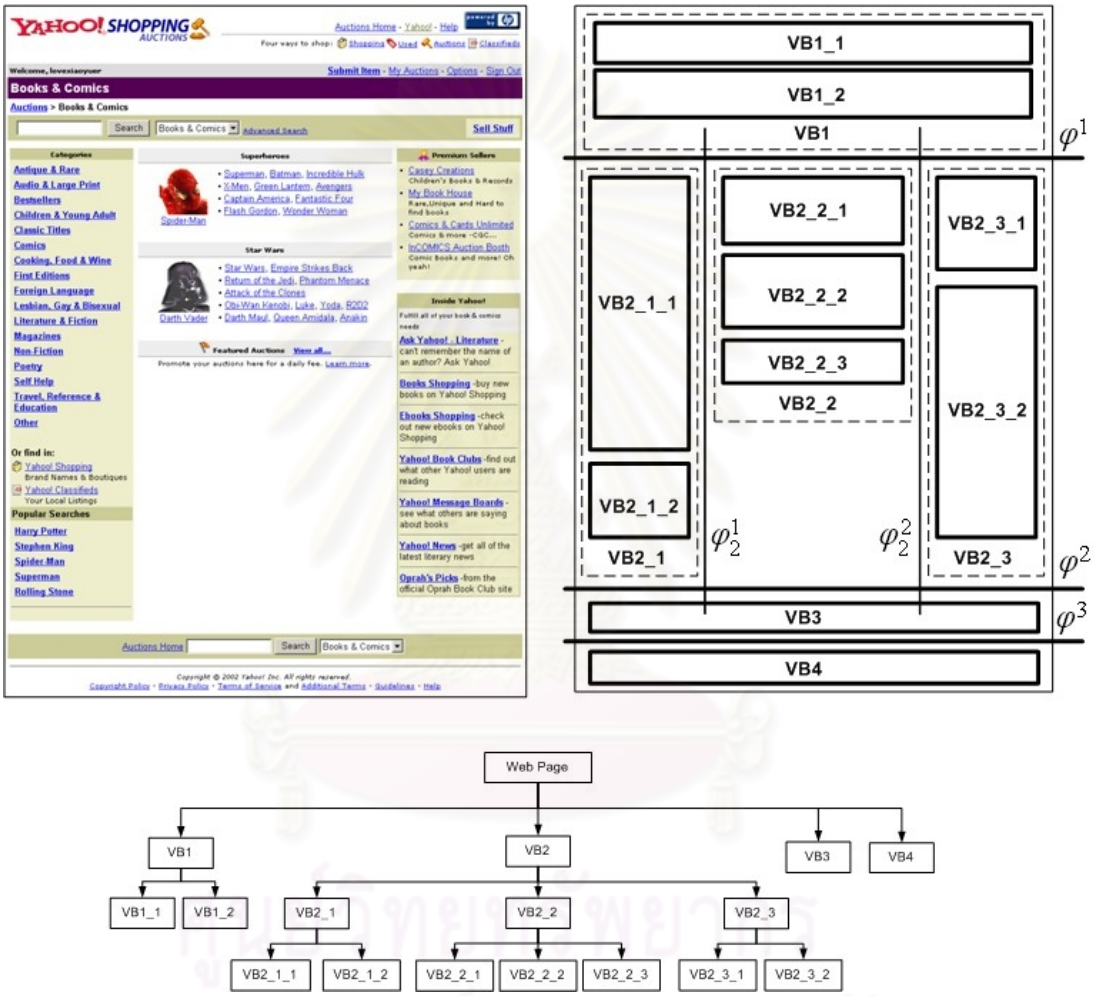
ขั้นตอนการทำงานของวีไอพีเอสอัลกอริทึม

1) การตัดออกเป็นบล็อก (Visual Block Extraction) ขั้นตอนนี้เป็นขั้นตอนแรกของวีไอพีเอสอัลกอริทึม ซึ่งจะทำการตัดหน้าเว็บเพจออกเป็นส่วนคือ เป็นบล็อกโดยที่แต่ละบล็อกมีเนื้อหาคล้ายกันหรือทำนองเดียวกัน

2) การตรวจหาตัวแบ่งแยก (Visual Separator Detection) ขั้นตอนนี้จะเป็นการกำหนดส่วนที่ถูกแบ่งออกเป็นบล็อกให้แยกออกจากกันโดยตัวแบ่งแยก ซึ่งอาจจะแบ่งตามแนวตั้งหรือแนวนอนก็ได้ แล้วให้ค่าน้ำหนักของแต่ละส่วนตามความสำคัญ

3) การสร้างโครงสร้างของเนื้อหา (Content Structure Construction) ในขั้นตอนนี้จะทำการสร้างส่วนของเนื้อหาที่สำคัญให้มีลักษณะเป็นโครงสร้างที่ชัดเจน

4) ทำกระบวนการข้างต้นซ้ำอีก (Iterating the Above Step) ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการทำงานของวีไอพีเอสอัลกอริทึม ซึ่งจะทำตามข้อ 1-3 ที่กล่าวมาข้างต้นซ้ำจนกว่าจะสามารถพบส่วนตรงกับความต้องการของผู้ใช้ โดยโครงสร้างการแบ่งหน้าเว็บเพจของวีไอพีเอสอัลกอริทึมแสดงได้ดังรูปที่ 2.5



รูปที่ 2.5 โครงสร้างการแบ่งหน้าเว็บเพจของวีไอพีเอสอัลกอริทึม

โดยในงานวิจัยนี้จะนำเทคนิคด้านวีไอพีเอสอัลกอริทึมที่มีอยู่แล้วมาใช้ให้เกิดประโยชน์เพิ่มขึ้น ด้วยการนำมาประยุกต์ร่วมกับแบบจำลองความน่าจะเป็น โดยวีไอพีเอสอัลกอริทึมช่วยกำหนดข้อคำถามใหม่ในส่วนของการขยายคำในข้อคำถามเดิม ซึ่งได้มาจากการเลือกคำที่ปรากฏในบล็อกที่ผ่านการแบ่งโดยวีไอพีเอสอัลกอริทึม แล้วเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ในการให้ผลป้อนกลับที่ตรงประเด็นด้วยแบบจำลองความน่าจะเป็น

2.1.3 แบบจำลองความน่าจะเป็น

เป็นแบบจำลองแบบหนึ่งในการค้นคืนสารสนเทศ ซึ่งนำเสนอครั้งแรกโดย Maron & Kuhns ต่อมามีการขยายความโดย Spark Jone ซึ่งภายหลังรู้จักในชื่อ Binary Independence Retrieval (BIR) Model โดยในแนวคิดของการค้นคืนนั้นพยายามที่จะตอบคำถามที่ว่าอะไรคือความน่าจะเป็นที่เอกสารชิ้นนั้นตรงประเด็นกับข้อคำถามที่ใช้ในการค้นคืนสารสนเทศ โดยการใช้แบบจำลองความน่าจะเป็นนั้นจะค้นคืนเอกสารโดยเรียงลำดับผลการค้นคืนที่ได้ ตามลำดับความน่าจะเป็นที่เอกสารนั้นตรงประเด็นกับข้อคำถามจากมากไปน้อย และพยายามที่จะแก้ปัญหาเกี่ยวกับการค้นคืนสารสนเทศ เพื่อใช้ในการให้ผลป้อนกลับที่ตรงประเด็นให้มีประสิทธิผลมากขึ้น

ข้อดีของการเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็นคือ [5]

1) กระบวนการในการให้ผลป้อนกลับที่ตรงประเด็นนั้นจะขึ้นอยู่กับการให้ค่าน้ำหนักของคำใหม่ในข้อคำถามโดยตรง เพราะเมื่อเปลี่ยนแปลงค่าน้ำหนักของคำใหม่แล้ว จะทำให้เอกสารที่ค้นคืนมาได้อยู่ในลำดับการจัดที่สูงขึ้น

2) การเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็นนั้นเป็นการเปลี่ยนแปลงค่าน้ำหนักของคำได้เหมาะสม เพราะคำที่ตรงประเด็นกับข้อคำถามซึ่งใช้ในการค้นคืนเอกสารนั้นจะมีค่าน้ำหนักเพิ่มขึ้น ทำให้เอกสารที่ได้จากการค้นคืนเว็บเพจโดยการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้จะตรงตามความต้องการของผู้ใช้มากขึ้น เพราะค่าน้ำหนักของคำในเอกสารที่ไม่ตรงประเด็นนั้นจะลดลง

ข้อเสียของการเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็นคือ [5]

1) ค่าน้ำหนักของคำในเอกสารที่กำหนดไว้ตั้งแต่เริ่มต้นไม่ได้นำมาใช้ในการพิจารณาเมื่อทำการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่

2) การเปลี่ยนแปลงค่าน้ำหนักของคำใหม่จะไม่สนใจค่าน้ำหนักของคำในการเปลี่ยนแปลงข้อคำถามครั้งก่อน

3) ไม่มีการปรับเปลี่ยนคำที่ใช้ในข้อคำถาม คือ เปลี่ยนแปลงแค่ค่าน้ำหนักคำในข้อคำถามเดิมเท่านั้น ซึ่งปัญหาข้อนี้สามารถแก้ไขได้ด้วยการใช้วิไอพีเอสอัลกอริทึมในการขยายคำที่ใช้ในข้อคำถาม

ซึ่งจากงานวิจัยที่ผ่านมา [4] จะพบว่าการให้ค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น นั้นสามารถช่วยเพิ่มประสิทธิผลของการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นได้ดี โดยเฉพาะในส่วนของการเปลี่ยนแปลงค่าน้ำหนักของคำ ดังนั้นงานวิจัยนี้ได้นำส่วนของแบบจำลองความน่าจะเป็นมาใช้ในการเปลี่ยนแปลงค่าน้ำหนักของคำ โดยการหาค่าความคล้ายกันของข้อคำถามกับเอกสารแบบเว็บเพจ

2.1.4 การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ (User Relevance feedback)

ในการกำหนดข้อคำถามใหม่ด้วยการขยายคำในข้อคำถาม และการเปลี่ยนแปลงค่าน้ำหนักคำในข้อคำถามสามารถทำได้โดยผ่านทางการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

(User Relevance Feedback) [4] โดยการใช้แบบจำลองปริภูมิเวกเตอร์หรือแบบจำลองความน่าจะเป็น ซึ่งในงานวิจัยนี้จะใช้แบบจำลองทั้งสองแบบ โดยในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น จะเริ่มด้วยการค้นคืนเว็บเพจที่ต้องการด้วยข้อคำถามหลังจากที่ระบบแสดงผลการค้นคืนเว็บเพจที่ตรงประเด็นแล้ว ผู้ใช้จึงทำการตรวจสอบเอกสารที่ค้นคืนออกมาได้ ซึ่งผู้ใช้จะพิจารณาเลือกเอกสารที่เห็นว่าตรงประเด็นโดยการทำเครื่องหมายไว้ จากนั้นระบบจะทำการเลือกคำที่จะใช้ในการกำหนดข้อคำถามใหม่จากเอกสารเหล่านั้นเพื่อให้ได้ข้อคำถามใหม่ที่ดีขึ้น แล้วผู้ใช้ก็จะป้อนข้อคำถามใหม่ที่ได้กลับสู่ระบบอีกครั้งเพื่อทำการค้นคืนใหม่ ซึ่งจุดประสงค์สำคัญของการขยายคำในข้อคำถามคือ ข้อคำถามใหม่นั้นควรมีความคล้าย (Similarity) กับเอกสารที่ตรงประเด็นมากขึ้น และมีความคล้ายกับเอกสารที่ไม่ตรงประเด็นน้อยลงเมื่อเปรียบเทียบกับข้อคำถามเดิม ซึ่งกระบวนการในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นแสดงได้ดังรูปที่ 2.3

รูปที่ 2.6 กระบวนการในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

โดยในงานวิจัยนี้จะใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ 2 วิธีการ คือ

1) การขยายคำและค่าน้ำหนักคำในข้อคำถาม ด้วยแบบจำลองปริภูมิเวคเตอร์ (Query Expansion and Term Reweighting for the Vector Space Model) [10] เป็นการเปลี่ยนแปลงข้อคำถามเดิมที่ใช้ในการค้นคืนเว็บเพจ ด้วยการเพิ่มคำเข้าไปในข้อคำถามใหม่ พร้อมทั้งเปลี่ยนแปลงค่าน้ำหนักของคำ โดยคำที่จะเพิ่มเข้าไปในข้อคำถามใหม่นั้นเลือกมาจากเอกสารที่ตรงประเด็นที่ถูกค้นคืนมาได้ในครั้งแรก และมีการเปลี่ยนแปลงค่าน้ำหนักของคำ โดยมีสูตรในการคำนวณเพื่อกำหนดข้อคำถามใหม่ดังสมการที่ 5

$$Q' = aQ + b\left(\frac{1}{R'}\sum_{i\in D_{R'}} DOC_i\right) - g\left(\frac{1}{N'}\sum_{i\in D_{N'}} DOC_i\right) \qquad (5)$$

โดยที่ Q' คือ ข้อคำถามใหม่
Q คือ ข้อคำถามเดิม
R' คือ จำนวนเอกสารที่ตรงประเด็น
N' คือ จำนวนเอกสารที่ไม่ตรงประเด็น
DOC_i คือ เวคเตอร์ของเอกสารที่ i
$D_{R'}$ คือ จำนวนเอกสารที่ตรงประเด็นจากเอกสารที่ค้นคืนได้ทั้งหมด
$D_{N'}$ คือ จำนวนของเอกสารที่ไม่ตรงประเด็นจากเอกสารที่ค้นคืนได้ทั้งหมด

ส่วนค่า a, b, g คือ ค่าคงที่สำหรับการปรับค่า ซึ่งในงานวิจัยนี้ จะกำหนด a = 1 และกำหนด b = 0.5 ส่วนค่า g = 0 เพื่อแสดงให้เห็นว่า การกำหนด a = 1 เพื่อให้ค่าน้ำหนักของข้อคำถามเดิมนั้น ไม่มีการเปลี่ยนแปลงใดๆทั้งสิ้น ส่วนการกำหนด b = 0.5 เพื่อแสดงให้เห็นว่ามีการนำคำและน้ำหนักของคำ ในเอกสารที่ผู้ใช้ระบุว่าตรงประเด็นมาทำการเปลี่ยนแปลงเพื่อใช้ในการกำหนดข้อคำถามใหม่ โดยให้ค่าน้ำหนักเพิ่มขึ้นไม่มากหรือน้อยเกินไป และการกำหนดค่า g = 0 เพราะไม่สนใจคำและน้ำหนักของคำในเอกสารที่ผู้ใช้ไม่ได้ระบุว่าตรงประเด็นกับที่ต้องการมาใช้ในการเปลี่ยนแปลงข้อคำถามใหม่

2) การเปลี่ยนแปลงค่าน้ำหนักคำในข้อคำถามด้วยแบบจำลองความน่าจะเป็น (Term Reweighting for Probabilistic Model) [10] เป็นการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ ที่จะใช้ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยคำที่ถูกเลือกมาเป็นข้อคำถามใหม่นั้นได้มาจากการเลือกส่วนของเอกสาร ที่ผู้ใช้เห็นว่าตรงประเด็นกับที่ต้องการ แล้วเพิ่มคำเหล่านั้นเข้าไปในข้อคำถามเดิม โดยมีสูตรการคำนวณดังสมการที่ (6)

$$W_{i,j} = \log \frac{\frac{r}{R-r}}{\frac{n-r}{(N-n)-(R-r)}} \qquad (6)$$

โดยที่ $W_{i,j}$ คือ ค่าน้ำหนักของคำ i ในข้อคำถาม j
r คือ จำนวนของเอกสารที่ตรงประเด็นกับข้อคำถาม j ที่มีคำ i ปรากฏอยู่
R คือ จำนวนของเอกสารทั้งหมดที่ตรงประเด็นกับข้อคำถาม j
n คือ จำนวนของเอกสารในคอลเลคชันที่มีคำ i ปรากฏอยู่
N คือ จำนวนของเอกสารทั้งหมดในคอลเลคชัน

โดยข้อดีของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ [4] นั้นสามารถช่วยลดรายละเอียดของกระบวนการกำหนดข้อคำถามใหม่ให้น้อยลง เพราะผู้ใช้งานเป็นผู้ตัดสินใจเลือกเอกสารที่เห็นว่าตรงประเด็นที่ได้มาจากการค้นคืนครั้งแรก เพื่อนำไปใช้ในส่วนของการขยายคำในข้อคำถาม การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นมีลำดับขั้นตอนในการทำงาน

ที่ชัดเจน ไม่ซับซ้อน คือ ผู้ใช้พิจารณาเอกสารที่เห็นว่าตรงตามที่ต้องการจากผลการค้นคืนที่ได้แล้วทำเครื่องหมายไว้ หลังจากนั้นระบบจะเลือกคำที่มีในเอกสารเหล่านั้นมาเพิ่มเข้าไปในข้อคำถามใหม่ แล้วจึงทำการค้นคืนซ้ำอีกครั้งด้วยข้อคำถามใหม่ที่ได้ ทำให้ไม่เกิดความยุ่งยากในการดำเนินการและเนื่องจากคำที่ใช้ในการกำหนดข้อคำถามใหม่ถูกเลือกมาจากเอกสารที่ผู้ใช้ระบุไว้ว่าตรงประเด็นทำให้คำที่ตรงประเด็น (Relevance Term) ถูกเลือกมาใช้ในการกำหนดข้อคำถามใหม่ ส่วนคำที่ไม่ตรงประเด็น (Non-Relevance Term) ซึ่งปรากฏในเอกสารที่ไม่ตรงประเด็นก็จะไม่ถูกเลือกมาใช้ในการกำหนดข้อคำถามใหม่ ทำให้ผลลัพธ์ที่ได้ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยข้อคำถามใหม่นั้นตรงตามความต้องการของผู้ใช้มากขึ้น

2.1.5 โครงสร้างข้อมูลที่ใช้ในระบบการจัดเก็บและค้นคืนเว็บเพจ

สำหรับโครงสร้างข้อมูลที่ใช้ในระบบการจัดเก็บและค้นคืนเว็บเพจ จำเป็นต้องมีการสร้างดรรชนี (Index) โดยการสร้างดรรชนีจะสร้างจากเอกสารที่ถูกประมวลผลแล้วจากเอกสารต้นฉบับเก็บเป็นแฟ้มข้อมูลที่เรียกว่า แฟ้มข้อมูลผกผัน (Inverted File) ซึ่งเป็นโครงสร้างข้อมูลที่ใช้ตารางเป็นเครื่องมือในการเข้าถึงข้อมูล ตารางดังกล่าวประกอบด้วยคำและตัวที่ชี้ไปยังข้อมูลในแฟ้มข้อมูลตามคำนั้นๆ ซึ่งเป็นโครงสร้างที่มีรูปแบบง่าย ไม่ซับซ้อน ทำให้สะดวกต่อการเขียนโปรแกรม แต่ข้อเสียคือใช้เนื้อที่ในการจัดเก็บค่อนข้างมาก เนื่องจากต้องนำส่วนของข้อมูลที่เป็นดรรชนีไปเก็บไว้ในแฟ้มข้อมูลผกผันด้วย

ลักษณะของดรรชนีผกผันคือ ตารางของคำที่ถูกสร้างขึ้น เพื่อเก็บคำที่มีทั้งหมด พร้อมทั้งรายการของที่อยู่ของคำเหล่านั้นว่าอยู่ในเอกสารใด เมื่อมีการค้นหาเกิดขึ้นก็ไม่จำเป็นต้องไปค้นหาคำจากทีละเอกสาร แต่สิ่งที่โปรแกรมจะทำก็คือ ไปดูที่ตารางคำว่าคำที่ค้นหาอยู่ที่เอกสารใดบ้าง ซึ่งก็จะทำให้การค้นหาเป็นไปได้อย่างรวดเร็ว การจัดเก็บดรรชนีผกผันแสดงตัวอย่างในตารางที่ 2.1

ตารางที่ 2.1 รูปแบบการจัดเก็บดรรชนีผกผัน

Term	WebPageID	Visual BlockID	Weight
Alex	25	3	0.083
Captain	105	4	0.142
Fulham	119	6	0.200
Lehman	44	5	0.092
Davenport	51	1	0.065
Everton	95	8	0.018
Woosnam	213	7	0.013

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 VIPS: A Vision based Page Segmentation Algorithm [2]

นำเสนอเกี่ยวกับอัลกอริทึมชื่อวีไอพีเอสอัลกอริทึม ในการแบ่งหน้าเว็บเพจออกเป็นบล็อก เพื่อลดส่วนของเอกสารแบบเว็บเพจที่ใช้ในการค้นคืนสารสนเทศเหลือเพียงเฉพาะส่วนที่ตรงประเด็นเท่านั้น และสามารถเพิ่มประสิทธิผลในการค้นคืนสารสนเทศได้ โดยทำการเปรียบเทียบประสิทธิผลของการค้นคืนสารสนเทศกับวิธีการแบบดอม(DOM-based Page Segmentation) [8] และแบบค้นคืนทั้งเอกสาร (Full Document) ซึ่งผลที่ได้คือ ค่าความแม่นยำจากการค้นคืนสารสนเทศโดยใช้ วีไอพีเอสอัลกอริทึมนั้น ให้ค่าที่มากที่สุด

จะเห็นได้ว่าวีไอพีเอสอัลกอริทึมที่ได้จากงานวิจัยนี้นั้นมีข้อดีคือ ผลที่ได้จากการค้นคืนเอกสารนั้นมีค่าความแม่นยำที่มากที่สุด ดังนั้นงานวิจัยนี้จึงนำข้อดีของวีไอพีเอสอัลกอริทึมนี้มาประยุกต์ใช้กับการค้นคืนเว็บเพจ

2.2.2 Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation [3]

งานวิจัยนี้ ได้ทดลองนำวีไอพีเอสอัลกอริทึม มาใช้ในการเพิ่มประสิทธิผลของการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียม (Pseudo-Relevance Feedback) ในการค้นคืนเว็บเพจโดยใช้การแบ่งเว็บเพจเป็นบล็อก เพื่อช่วยเลือกคำที่จะนำมากำหนดข้อคำถามใหม่ในการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมส่งผลให้ประสิทธิผลของการค้นคืนเพิ่มขึ้นได้ถึง 27 เปอร์เซ็นต์ โดยมีวิธีการนำวีไอพีเอสอัลกอริทึมมาประยุกต์ใช้กับการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมดังนี้

1) เริ่มต้นการค้นคืนเว็บเพจด้วยวิธีการธรรมดา ที่มีทั่วไปไม่ได้ระบุว่าใช้วิธีใด
2) แบ่งเว็บเพจเป็นบล็อกด้วยวีไอพีเอสอัลกอริทึม
3) เลือกบล็อกที่ถูกแบ่งแล้วมาเรียงลำดับตามความตรงประเด็นที่มีกับข้อคำถามจากมากไปน้อย
4) เลือกคำที่จะนำมาปรับเปลี่ยนคำในข้อคำถามจากบล็อกที่เรียงลำดับไว้แล้วจำนวน 10 คำ ทำให้ส่วนที่ไม่ตรงประเด็นไม่ถูกเลือกมาปรับเปลี่ยนคำในข้อคำถาม ซึ่งถ้าเป็นวิธีการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมที่เคยใช้นั้นจะเลือกส่วนที่จะนำมาปรับเปลี่ยนคำในข้อคำถามจากทั้งหมดของหน้าเว็บเพจ
5) เพิ่มค่าน้ำหนักให้กับคำเดิม และคำใหม่ ในข้อคำถามใหม่

ในการทดลองได้นำวีไอพีเอสอัลกอริทึมมาช่วยในการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียม และเปรียบเทียบประสิทธิผลในส่วนของค่าความแม่นยำ (Precision) และค่าเรียกคืน (Recall) กับวิธีการแบบดอมและแบบเอกสารทั้งหมด ซึ่งผลการทดลองที่ได้ปรากฏว่าวีไอพีเอสอัลกอริทึมนั้นสามารถช่วยเพิ่มประสิทธิผลการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมได้ดีที่สุด

จากงานวิจัยนี้ผู้วิจัยคิดว่าการการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมนั้น อาศัยการเรียงลำดับค่าคำที่มีความสำคัญ โดยเลือกคำที่มีความสำคัญลำดับต้นๆ ไม่เกิน 10 อันดับมา

เพิ่มเข้าไปในข้อคำถามเดิมได้เป็นข้อคำถามใหม่ โดยผู้ใช้งานไม่ได้มีส่วนในการให้ผลป้อนกลับที่ตรงประเด็น ซึ่งผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ถือได้ว่าเป็นเป็นส่วนสำคัญยิ่งที่จะทำให้ผลการค้นคืนที่ได้ตรงตามความต้องการของผู้ใช้งาน ดังนั้นผู้วิจัยเห็นว่าน่าจะประยุกต์การค้นคืนเว็บเพจโดยใช้วีไอพีเอสอัลกอริทึม เข้ากับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ เพื่อให้ผู้ใช้มีส่วนร่วม กับการกำหนดข้อคำถามใหม่ แทนการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียม และเพิ่มในส่วนของการเปลี่ยนแปลงค่าน้ำหนักของคำในการให้ผลป้อนกลับที่ตรงประเด็นด้วยแบบจำลองความน่าจะเป็น ซึ่งยังไม่มีงานวิจัยใดนำ วีไอพีเอสอัลกอริทึมมาประยุกต์กับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้และแบบจำลองความน่าจะเป็น เพื่อพิสูจน์ว่าผลการทดลองที่ได้จะเป็นเช่นไร

2.2.3 Pseudo-Relevance Feedback in Web Information Retrieval Using Segments' Subjective Importance Values [11]

นำเสนองานวิจัยเกี่ยวกับการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมในการค้นคืนเว็บเพจ ที่มีการแบ่งเว็บเพจเป็นบล็อกโดยใช้วีไอพีเอสอัลกอริทึมจากงานวิจัย [2] และการลดความกำกวมของคำมาประยุกต์รวมกัน ซึ่งผลที่ได้จากการให้ผลป้อนกลับที่ตรงประเด็นแบบเทียมนั้น พบว่าค่าความแม่นยำของการค้นคืนนั้นยังมีค่าน้อย เมื่อเทียบกับค่าเรียกคืนที่ได้ค่ามากกว่า ซึ่งยังคงเป็นปัญหาสำคัญที่ต้องแก้ไข

จากงานวิจัยนี้ผู้วิจัยเห็นว่า ผลลัพธ์ที่ได้จากการค้นคืนเว็บเพจด้วยการใช้วีไอพีเอสอัลกอริทึมและการลดความกำกวมของคำไม่ได้ช่วยให้ประสิทธิผลของระบบดีขึ้น ดังนั้นระบบในการค้นคืนเว็บเพจที่ผู้วิจัยพัฒนาขึ้นจึงไม่พิจารณาเรื่องความกำกวมของคำ

บทที่ 3
วิธีการวิจัย

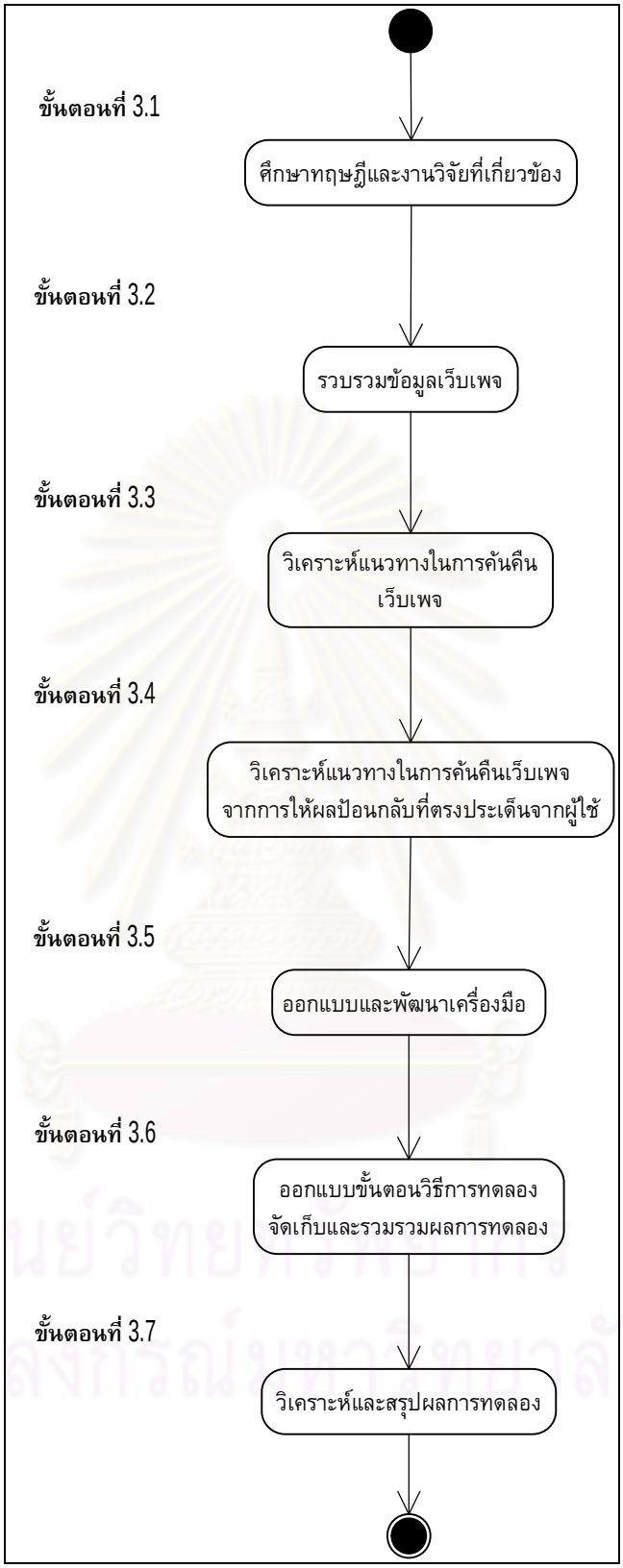
ในบทนี้จะกล่าวถึงขั้นตอนวิธีการวิจัยของการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น ซึ่งประกอบด้วยขั้นตอนหลักๆ ทั้งหมด 7 ขั้นตอน ดังต่อไปนี้

1) ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2) รวบรวมข้อมูลเว็บเพจ
3) วิเคราะห์แนวทางในการค้นคืนเว็บเพจ
4) วิเคราะห์แนวทางในการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้
5) ออกแบบและพัฒนาเครื่องมือเพื่อสนับสนุนแนวทางที่นำเสนอไว้
6) ออกแบบขั้นตอนวิธีการทดลองรวมถึงจัดเก็บและรวบรวมผลการทดลองที่ได้
7) วิเคราะห์และสรุปผลการทดลอง

สำหรับขั้นตอนที่ 5-7 จะกล่าวถึงใน บทที่ 4 5 และ 6 ตามลำดับ โดยในบทนี้จะกล่าวถึงขั้นตอนที่ 1-4 ดังต่อไปนี้

3.1 ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ขั้นตอนนี้เป็นขั้นตอนในการศึกษาทฤษฎีที่เกี่ยวข้องกับการจัดเก็บและค้นคืนสารสนเทศ การทำดรรชนีอัตโนมัติ การกำหนดค่าน้ำหนักคำด้วยค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน การกำหนดโครงสร้างการจัดเก็บข้อมูลสารสนเทศ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ การหาค่าความคล้ายระหว่างเอกสารกับข้อคำถามด้วยค่าสัมประสิทธิ์โคซายน์ แบบจำลองความน่าจะเป็น การเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น การขยายคำและค่าน้ำหนักคำในข้อคำถามด้วยแบบจำลองปริภูมิเวคเตอร์ การประเมินประสิทธิผลของระบบการจัดเก็บและค้นคืนสารสนเทศ วีไอพีเอสอัลกอริทึม งานวิจัยต่างๆ ที่เกี่ยวข้อง



รูปที่ 3.1 ขั้นตอนวิธีการวิจัย

3.2 รวบรวมข้อมูลเว็บเพจ

ในขั้นตอนนี้ผู้วิจัยได้ทำการเก็บรวบรวมเว็บเพจ จากเว็บไซต์ 8 เว็บไซต์ ได้เว็บเพจทั้งหมดจำนวน 300 เว็บเพจ ซึ่งเก็บติดต่อกันเป็นเวลา 5 วัน เพื่อให้ได้เนื้อหาข่าวที่มีลักษณะต่อเนื่องและใกล้เคียงกัน โดยทุกเว็บเพจเป็นภาษาอังกฤษทั้งหมด และมีเนื้อหาข่าวเกี่ยวกับกีฬา ซึ่งผู้วิจัยมีความสนใจและมีความเข้าใจในด้านนี้ นอกจากนี้ก็เพื่อให้เนื้อหาของข้อมูลหรือเอกสารที่ใช้ในระบบมีเชิงลึกและเฉพาะเจาะจงมากขึ้น สามารถแบ่งเว็บเพจตามประเภทของข่าวกีฬาได้ทั้งหมด 7 ประเภท เช่น ข่าวกีฬาประเภทฟุตบอล ข่าวกีฬาประเภทเทนนิส เป็นต้น จากนั้นนำเว็บเพจทั้งหมดที่จัดเก็บไว้ไปแบ่งเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึมเพื่อใช้ในการค้นคืนเว็บเพจต่อไป

3.3 วิเคราะห์แนวทางในการค้นคืนเว็บเพจ

การค้นคืนเว็บเพจ ใช้ข้อคำถามในการค้นคืนทั้งหมดจำนวน 50 ข้อคำถาม ซึ่งรายละเอียดของข้อคำถามแสดงไว้ในหัวข้อที่ 5.2.3 เริ่มจากการป้อนข้อคำถามเข้าสู่ระบบการค้นคืนเว็บเพจ ทำการเปรียบเทียบค่าความคล้ายระหว่างข้อคำถามกับเอกสารเว็บเพจ โดยใช้สูตรดังสมการที่ 2 ของบทที่ 2 ซึ่งจะต้องทำการเปรียบเทียบข้อคำถามกับเอกสารเว็บเพจที่มีทั้งหมดในฐานข้อมูล เพื่อค้นคืนเอกสารเว็บเพจที่มีค่าความคล้ายกับข้อคำถามมากที่สุด แล้วแสดงผลการค้นคืนที่ได้ให้แก่ผู้ใช้งานเพื่อใช้ในการพิจารณาในขั้นตอนของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ต่อไป

3.4 วิเคราะห์แนวทางในการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

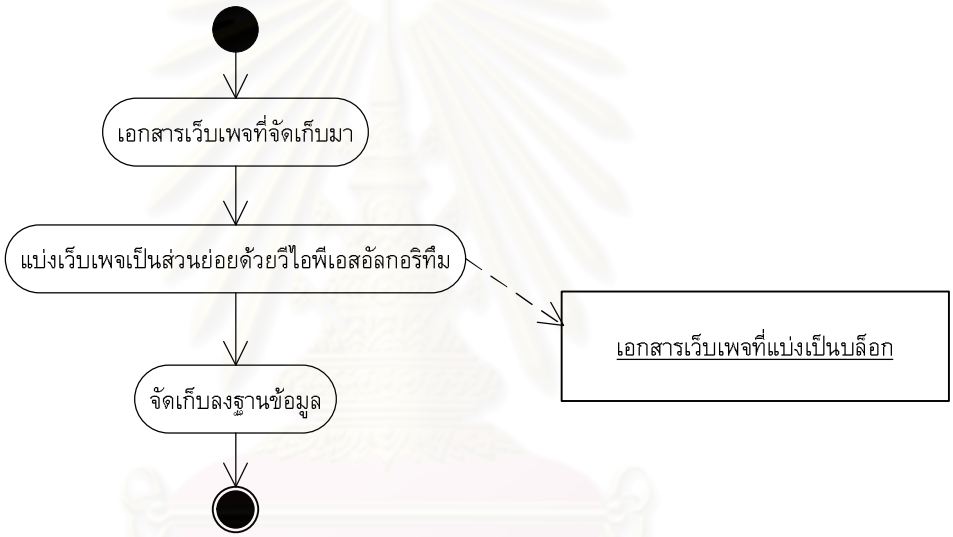
ในส่วนของการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ จะดำเนินการหลังจากที่ได้ทำการค้นคืนครั้งแรกเรียบร้อยแล้ว โดยผู้ใช้จะทำหน้าที่ในการให้ผลป้อนกลับแก่ระบบ ซึ่งจะพิจารณารายการเอกสารที่เห็นว่าตรงประเด็นกับที่ต้องการ จากรายการเอกสารที่ค้นคืนมาได้ทั้งหมดในครั้งแรก ผ่านทางส่วนต่อประสานกับผู้ใช้ แล้วเข้าสู่ขั้นตอนของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยระบบจะนำผลป้อนกลับที่ได้จากผู้ใช้ ไปคำนวณในการเปลี่ยนแปลงข้อคำถามใหม่ตามแนวทางที่นำเสนอไว้ทั้ง 2 วิธีการ คือ การขยายคำและค่าน้ำหนักคำในข้อคำถามด้วยแบบจำลองปริภูมิเวคเตอร์ และการเปลี่ยนแปลงค่าน้ำหนักคำในข้อคำถามด้วยแบบจำลองความน่าจะเป็น โดยการเปลี่ยนแปลงข้อคำถามใหม่ทั้ง 2 วิธีการที่นำเสนอนั้น จะมีการนำวีไอพีเอสอัลกอริทึม มาช่วยในการเลือกคำที่จะมากำหนดข้อคำถามใหม่ด้วย

3.5 ภาพรวมการทำงานของแนวทางที่นำเสนอ

ในขั้นตอนนี้จะนำเสนอภาพรวมการทำงานทั้งหมดของแนวทางที่นำเสนอไว้ ประกอบไปด้วยการพัฒนาการทำงานของเครื่องมือการค้นคืนเว็บเพจ โดยมีการพัฒนาฟังก์ชันในส่วนของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ตามแนวทาง หลักการ และการวิเคราะห์ต่างๆ ที่นำเสนอไว้ข้างต้น เพื่อให้ได้ระบบและเครื่องมือเพื่อสนับสนุนแนวความคิดและวิธีการที่งานวิจัยนี้ได้นำเสนอไว้ ซึ่งภาพรวมการทำงานของระบบนั้นประกอบด้วยขั้นตอนหลักๆ 4 ขั้นตอน ดังรูปที่ 3.3

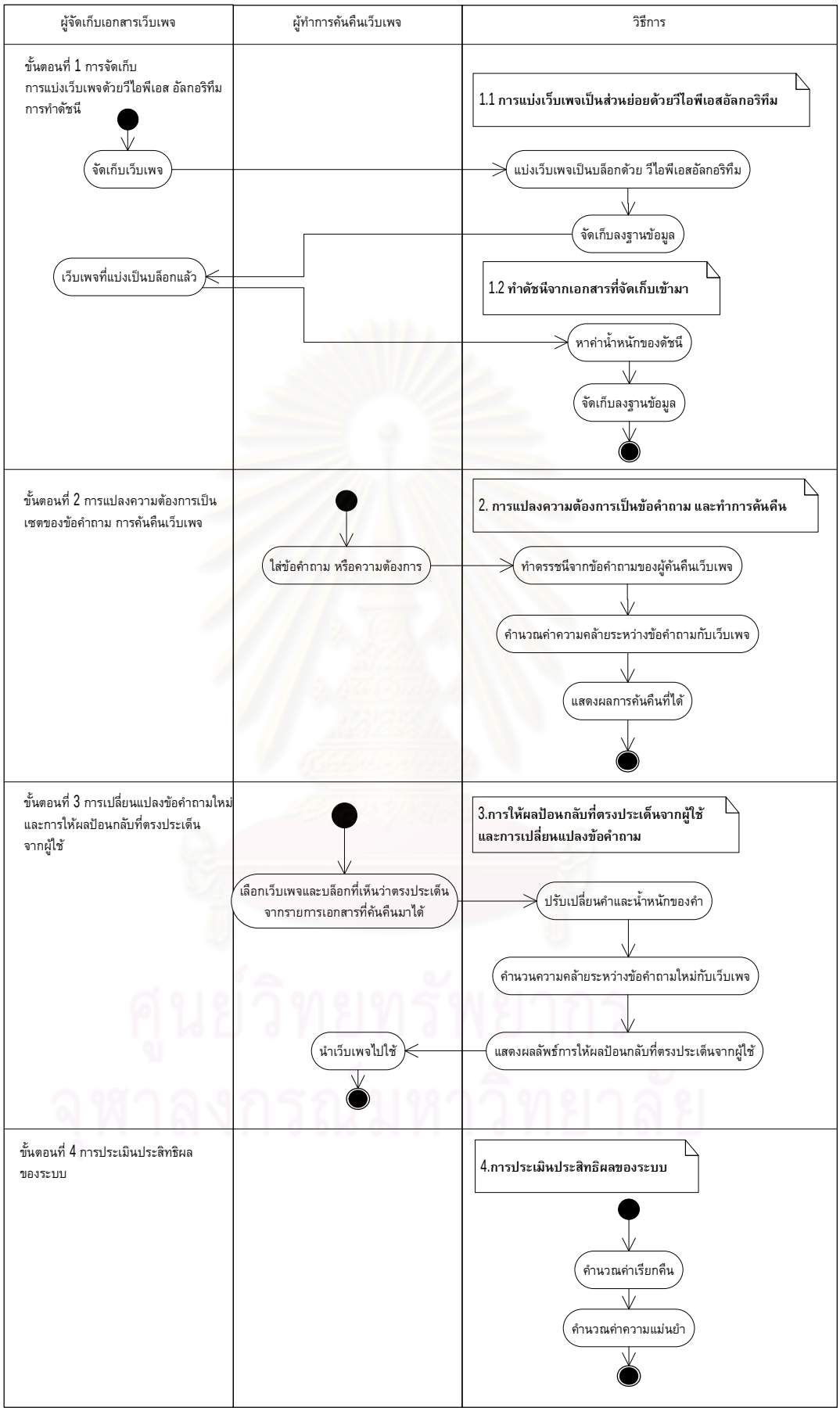
3.5.1 ขั้นตอนการจัดเก็บเว็บเพจ

ในขั้นตอนของการจัดเก็บเว็บเพจนี้ ประกอบด้วยขั้นตอนย่อย 2 ขั้นตอน คือ ขั้นตอนการแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม และขั้นตอนการทำดรรชนีอัตโนมัติ ซึ่งสามารถแสดงได้ดังรูปที่ 3.2 และ 3.4

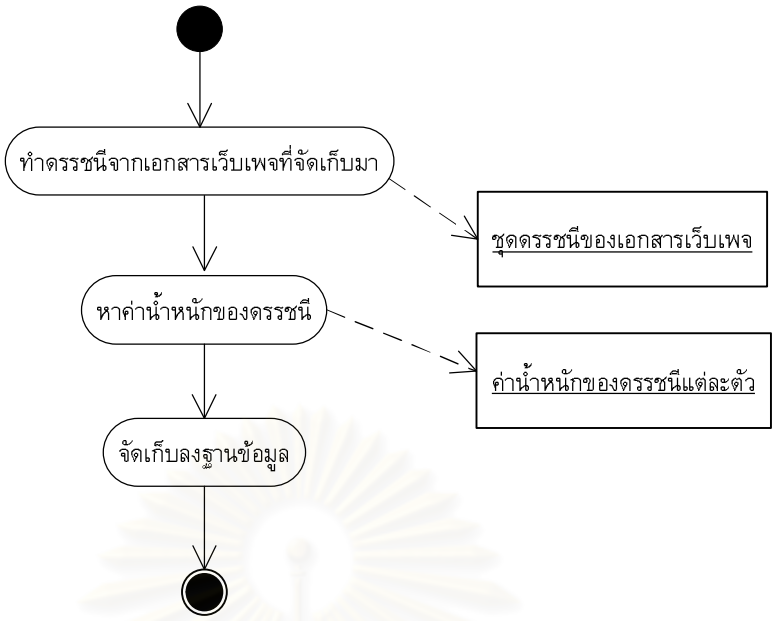


รูปที่ 3.2 แผนภาพกิจกรรมขั้นตอนการแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม

หลังจากจัดเก็บเว็บเพจ จากเว็บไซต์เรียบร้อยแล้ว ทำการแบ่งเว็บเพจที่จัดเก็บมาทั้งหมดด้วยวีไอพีเอสอัลกอริทึมออกเป็นบล็อกก่อนที่จะเข้าสู่ขั้นตอนในการทำดรรชนีอัตโนมัติ เพื่อรองรับขั้นตอนในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ซึ่งผู้ใช้จะเลือกบล็อกที่ตรงประเด็นกับที่ต้องการ หลังจากได้ผลของการค้นคืนครั้งแรก แล้วจึงเข้าสู่ขั้นตอนการทำดรรชนีอัตโนมัติ



รูปที่ 3.3 ภาพรวมการทำงานโดยรวมของแนวทางที่นำเสนอ

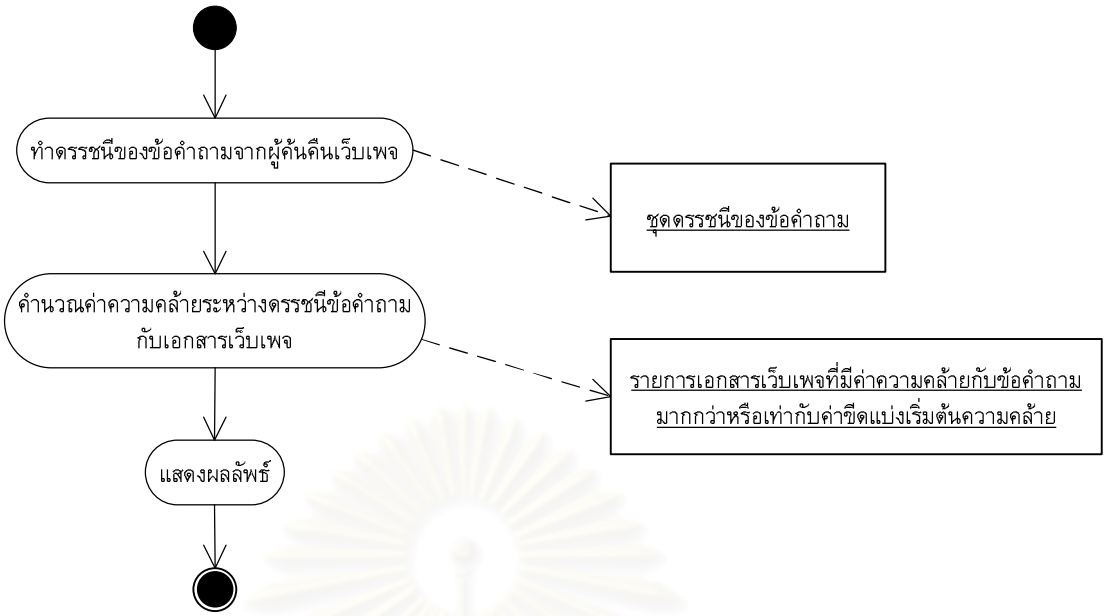


รูปที่ 3.4 แผนภาพกิจกรรมการจัดเก็บเอกสารเว็บเพจ และขั้นตอนการทำดรรชนีอัตโนมัติ

ในส่วนของขั้นตอนนี้เป็นการสร้างดรรชนีอัตโนมัติ ซึ่งระบบจะใช้ฟังก์ชันในการสร้างดรรชนีให้กับเอกสารเว็บเพจแต่ละบล็อกแล้วหาค่าน้ำหนักของคำที่เป็นดรรชนีเหล่านั้นแบบอัตโนมัติ โดยใช้ค่าน้ำหนักแบบค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน ซึ่งขั้นตอนในการทำดรรชนีอัตโนมัตินั้น อธิบายไว้ในหัวข้อที่ 2.1.1.1 ของบทที่ 2 ดังนั้นจะได้เอกสารเว็บเพจที่จัดเก็บเป็นบล็อกโดยมีดรรชนีและค่าน้ำหนัก ในการเชื่อมโยงไปยังเอกสารเหล่านั้น จากนั้นจัดเก็บเข้าสู่ฐานข้อมูล เพื่อใช้ในการค้นคืนเว็บเพจต่อไป

3.5.2 ขั้นตอนการค้นคืนเว็บเพจ

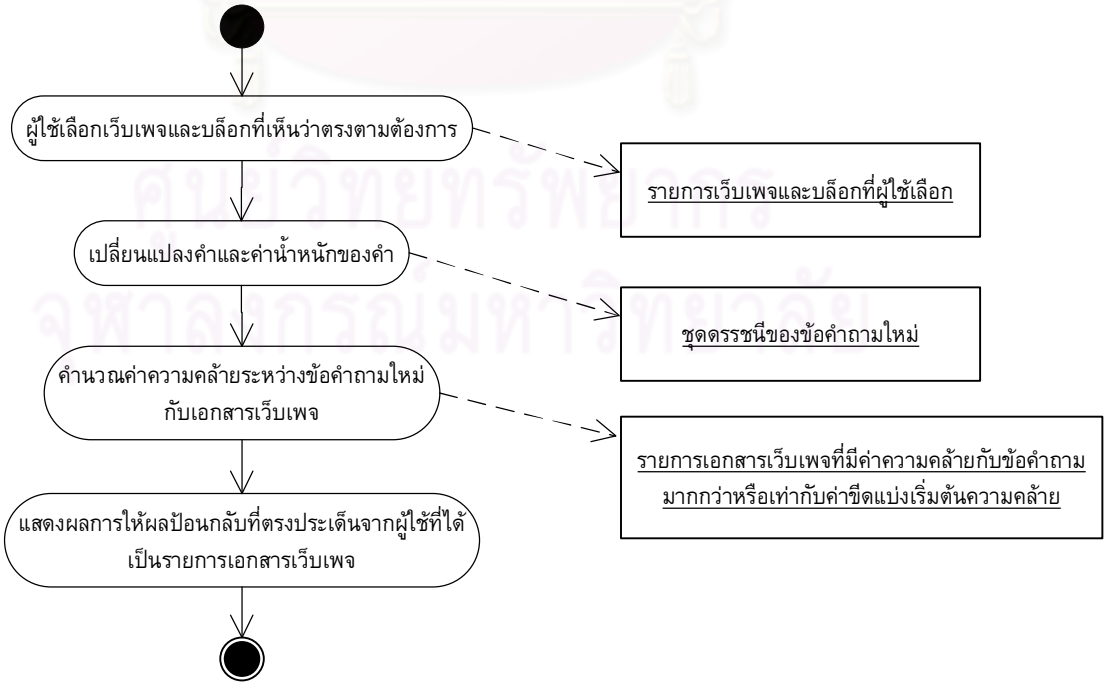
ในการค้นคืนเว็บเพจนั้น ผู้ใช้จะป้อนข้อคำถามเข้าสู่ระบบ ระบบจะแปลงข้อคำถามเหล่านั้นให้เป็นชุดของดรรชนี เพราะตัวแทนของเอกสารเว็บเพจที่จัดเก็บไว้ในฐานข้อมูลจัดเก็บเป็นชุดดรรชนี เปรียบเทียบค่าความคล้ายระหว่างชุดดรรชนีข้อคำถามกับชุดดรรชนีตัวแทนของเอกสารเว็บเพจที่มีในฐานข้อมูล โดยมีการกำหนดค่าขีดแบ่งเริ่มต้นความคล้าย เพื่อให้ระบบแสดงผลการค้นคืนกลับสู่ผู้ใช้งาน ถ้าเอกสารที่ค้นคืนมาได้เหล่านั้นมีค่าความคล้ายมากกว่าหรือเท่ากับค่าขีดแบ่งเริ่มต้นความคล้ายที่กำหนดไว้ ซึ่งรายละเอียดของการกำหนดค่าขีดแบ่งเริ่มต้นความคล้ายจะขอกล่าวถึงในบทที่ 5 สำหรับกิจกรรมการค้นคืนเว็บเพจในขั้นตอนนี้ แสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 แผนภาพกิจกรรมของการค้นคืนเว็บเพจ

3.5.3 ขั้นตอนการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นจะดำเนินการหลังจากระบบแสดงผลการค้นคืนครั้งแรกกลับสู่ผู้ใช้งานผ่านทางส่วนต่อประสานกับผู้ใช้เป็นรายการเว็บเพจ จากนั้นผู้ใช้ทำการตรวจสอบรายการเว็บเพจที่เห็นว่าตรงประเด็นกับที่ต้องการ แล้วทำการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยการป้อนรายการเอกสารที่ตรงประเด็นนั้นกลับสู่ระบบ แล้วระบบจะนำผลป้อนกลับจากผู้ใช้นั้นไปทำการคำนวณใหม่อีกครั้ง โดยกิจกรรมการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น แสดงได้ดังรูปที่ 3.6



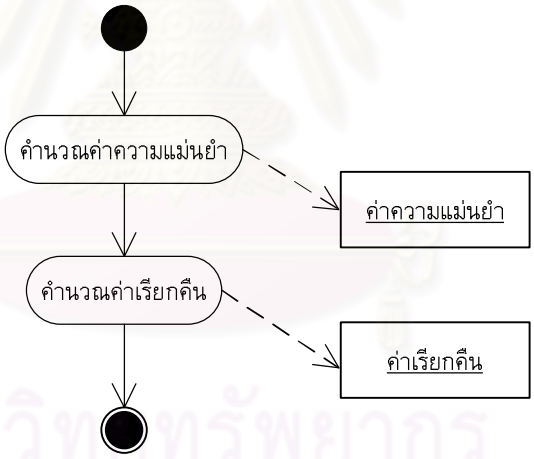
รูปที่ 3.6 แผนภาพกิจกรรมของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

เมื่อระบบรับรายการเว็บเพจที่ตรงประเด็นมาจากผู้ใช้แล้ว ระบบจะทำการกำหนดข้อคำถามใหม่ โดยแบ่งออกเป็น 2 แบบ คือ ในแบบจำลองปริภูมิเวคเตอร์จะทำการปรับเปลี่ยนคำและค่าน้ำหนักของคำ ซึ่งสามารถคำนวณได้จากสูตรในสมการที่ 5 ของบทที่ 2 แต่ในแบบจำลองความน่าจะเป็นจะทำการปรับเปลี่ยนค่าน้ำหนักของคำเท่านั้น ซึ่งสามารถคำนวณได้จากสูตรในสมการที่ 6 ของบทที่ 2 จากนั้นระบบคำนวณหาค่าความคล้ายระหว่างข้อคำถามใหม่กับเอกสารเว็บเพจ ท้ายที่สุดระบบจะแสดงผลการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ เป็นรายการเอกสารที่ค้นคืนมาได้เรียงลำดับตามค่าความคล้ายระหว่างข้อคำถามใหม่ กับเว็บเพจจากมากไปน้อย กลับสู่ผู้ใช้ผ่านทางส่วนต่อประสานกับผู้ใช้งาน

3.5.4 การประเมินประสิทธิผลของระบบ

การประเมินประสิทธิผลของระบบค้นคืนเว็บเพจ จะทำการประเมินจากเว็บเพจที่ค้นคืนออกมาได้นั้น ว่ามีค่าความแม่นยำตรงตามความต้องการของผู้ค้นคืนหรือไม่ ดังนั้นงานวิทยานิพนธ์นี้จึงใช้การคำนวณหา ค่าความแม่นยำ และค่าเรียกคืน ในการประเมินประสิทธิผลของการค้นคืนเว็บเพจ

ซึ่งสามารถคำนวณได้จากสูตรในสมการที่ 3 และ 4 ของบทที่ 2 ตามลำดับ สำหรับกิจกรรมในขั้นตอนนี้แสดงเป็นแผนภาพกิจกรรมได้ดังรูปที่ 3.7



รูปที่ 3.7 แผนภาพกิจกรรมการประเมินประสิทธิผลของระบบ

บทที่ 4
การพัฒนาเครื่องมือ

ในบทนี้จะกล่าวถึงรายละเอียดของการพัฒนาเครื่องมือสำหรับการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ แบบจำลองความน่าจะเป็น และมีการนำวีไอพีเอสอัลกอริทึมมาประยุกต์ใช้ร่วมด้วย โดยจะกล่าวถึงสภาพแวดล้อมที่ใช้ในการพัฒนาเครื่องมือ สถาปัตยกรรมในการพัฒนาเครื่องมือ โครงสร้างของเครื่องมือ และแบบจำลองข้อมูล ซึ่งมีรายละเอียดดังต่อไปนี้

4.1 สภาพแวดล้อมที่ใช้ในการพัฒนาเครื่องมือ

ฮาร์ดแวร์ (Hardware) และซอฟต์แวร์ (Software) ที่ใช้ในการพัฒนาระบบมีดังนี้

4.1.1 ฮาร์ดแวร์

1) เครื่องคอมพิวเตอร์ส่วนบุคคล หน่วยประมวลผล อินเทล เซเลรอน 1.4 กิกะเฮิร์ท โปรเซสเซอร์ 600 (Intel Pentium M 1.6 GHz Processor 600)

2) หน่วยความจำ (Memory) 256 เมกะไบต์

3) จานบันทึกแบบแข็ง (Hard disk) ความจุ 40 กิกะไบต์

4) จอภาพ 14 นิ้ว

4.1.2 ซอฟต์แวร์

1) ระบบปฏิบัติการ

(1) ระบบปฏิบัติการไมโครซอฟท์วินโดวส์ เอ็กซ์พี โปรเฟสชันแนล (Microsoft Window XP Professional)

2) ซอฟต์แวร์ที่ใช้ในจัดทำเอกสาร

(1) ไมโครซอฟท์ออฟฟิส 2003 (Microsoft Office 2003)

(2) ไมโครซอฟท์ออฟฟิสวิซิโอ โปรเฟสชันแนล 2003 (Microsoft Office Visio Professional 2003)

3) ซอฟต์แวร์ที่ใช้ในการพัฒนาส่วนต่อประสานกับผู้ใช้

(1) เครื่องมือช่วยจัดการส่วนต่อประสานเว็บเพจ ได้แก่ มาโครมีเดียร์ ดรีมวีเวอร์ เอ็มเอ็กซ์ 2002 รุ่นที่ 6 (Macromedia Dreamweaver: MX 2002 version 6)

(2) อินเทอร์เน็ต เอกซ์โพลเรอร์ รุ่นที่ 6 (Internet Explorer 6.0)

4) ซอฟต์แวร์ที่ใช้ในการเขียนโปรแกรมเพื่อพัฒนาเครื่องมือ

(1) อิดีสพลัส รุ่นที่ 2(EditPlus version 2)

(2) จาวาสแตนดาร์ดเอดีชัน รุ่น 1.4.2.08 สำหรับวินโดว์ส (J2SDK 1.4.2.08 for Windows)

5) ซอฟต์แวร์ที่ใช้ในส่วนให้บริการและส่วนสนับสนุน

(1) อาปาเช จากาตา ทอมแคท รุ่นที่ 5.0.30 (Apache Jakarta Tomcat 5.0.30)

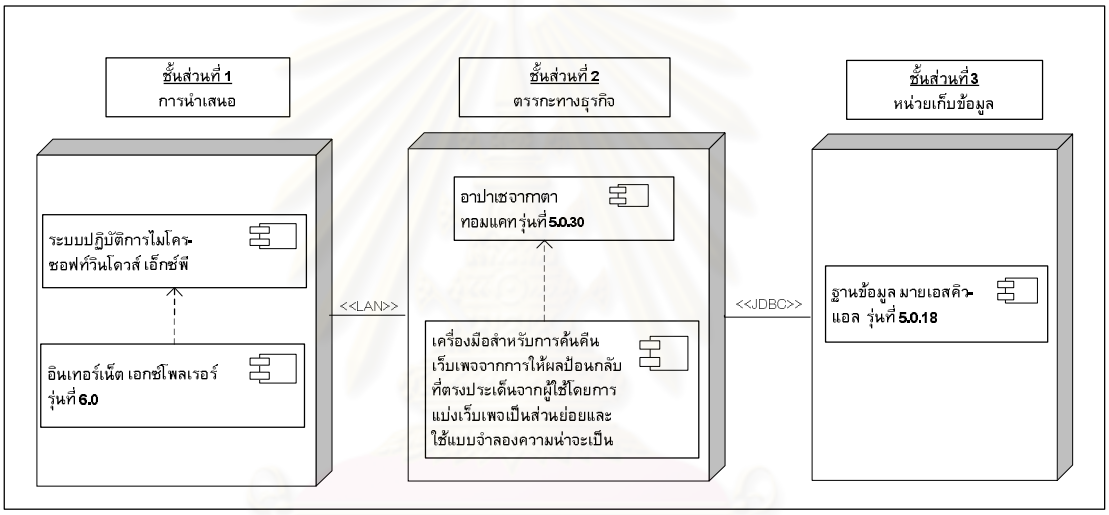
(2) ฐานข้อมูล มายเอสคิวแอล รุ่นที่ 5.0.18 (MySQL 5.0.18)

(3) โปรแกรมจัดการฐานข้อมูลพรีเมียมซอฟท์ นาวิแคท 2006 (PremiumSoft Navicat 2006)

(4) ตัวเชื่อมต่อมายเอสคิวแอลรุ่นที่ J 3.0.15 (MySQL Connector/J 3.0.15)

4.2 สถาปัตยกรรมในการพัฒนาเครื่องมือ

เป็นส่วนที่แสดงให้เห็นถึงส่วนประกอบต่าง ๆ ของสถาปัตยกรรมที่ใช้ในการพัฒนาเครื่องมือเพื่อแสดงโครงสร้างเทคโนโลยีของเครื่องมือที่สร้างขึ้น ทั้งส่วนที่ใช้ในการให้บริการและส่วนระบบที่พัฒนา มีการออกแบบสถาปัตยกรรมแบบหลายส่วนชั้น (Multi-Tiers) ดังรูปที่ 4.1 ซึ่งสามารถแบ่งออกเป็น 3 ส่วน ได้แก่



รูปที่ 4.1 แผนภาพส่วนประกอบของสถาปัตยกรรมในการพัฒนาเครื่องมือ

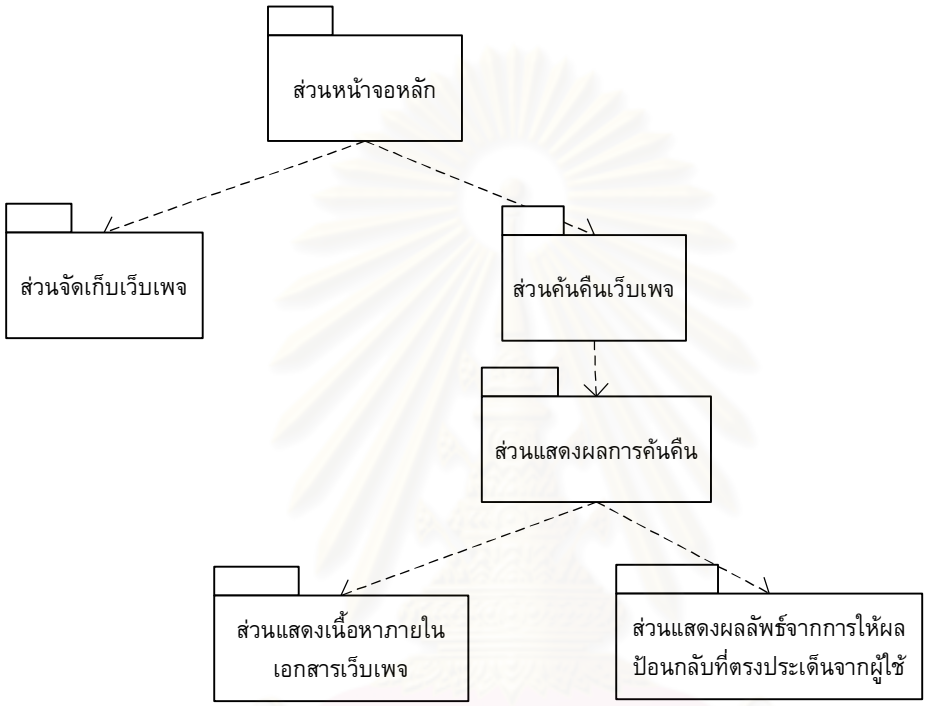
1) ชั้นส่วนการนำเสนอ (Presentation Tier) เป็นชั้นส่วนลูกข่ายซึ่งทำหน้าที่เป็นส่วนต่อประสานกับผู้ใช้ระบบโดยใช้โปรแกรมประยุกต์แบบเว็บเบส เพื่อแก้ปัญหาข้อจำกัดเรื่องสถานที่ทำงาน

2) ชั้นส่วนตรรกะทางธุรกิจ (Business Logic Tier) เป็นส่วนกลางทำหน้าที่ให้บริการข้อมูลและประมวลผลการทำงานให้แก่เครื่องลูกข่าย

3) ชั้นส่วนหน่วยเก็บข้อมูลของระบบ (Data Storage Tier) เป็นส่วนที่ทำหน้าที่จัดเก็บข้อมูลของระบบ โดยใช้ฐานข้อมูลมายเอสคิวแอล (MY SQL) ในการจัดเก็บข้อมูลของระบบ

4.3 โครงสร้างของเครื่องมือ

โครงสร้างของเครื่องมือที่พัฒนา ประกอบด้วยการทำงาน 2 ส่วนสำคัญคือ ส่วนของการจัดเก็บเว็บเพจ และ ส่วนของการค้นคืนเว็บเพจ ซึ่งสามารถอธิบายด้วยแผนภาพส่วนประกอบ ซึ่งแสดงความสัมพันธ์ระหว่างส่วนประกอบต่าง ๆ ในระบบ แผนภาพแสดงส่วนประกอบโครงสร้างของเครื่องมือ แสดงดังรูปที่ 4.2



รูปที่ 4.2 แผนภาพส่วนประกอบโครงสร้างของเครื่องมือ

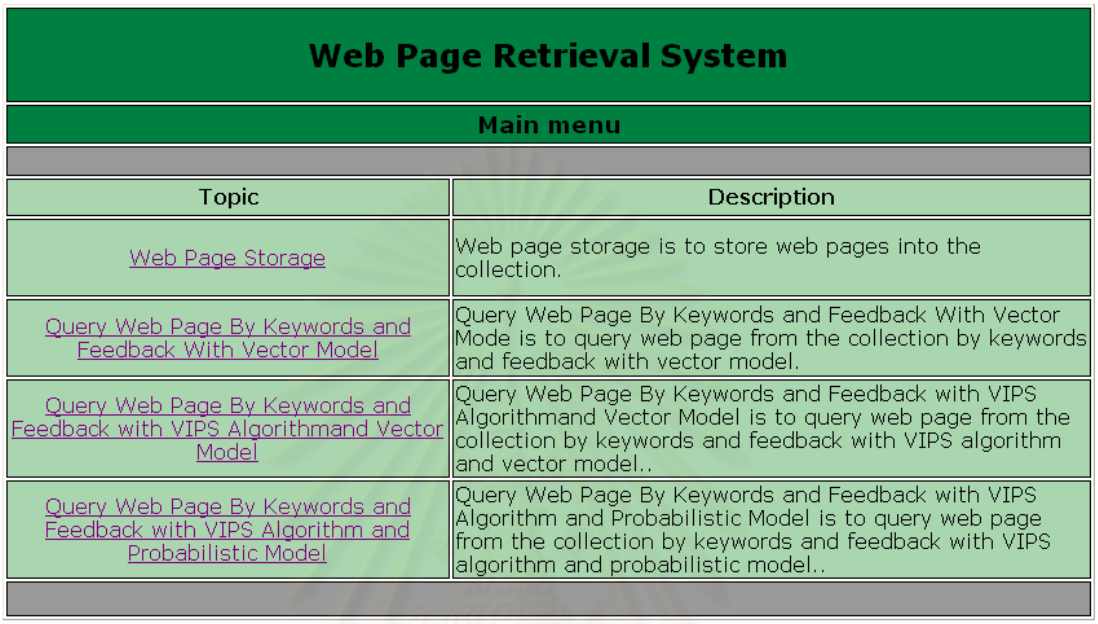
หน้าจอแสดงรายการของโครงสร้างหลักของเครื่องมือ ในระบบการจัดเก็บและค้นคืนเว็บเพจ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ แสดงได้ดังรูปที่ 4.3 หน้าจอแสดงรายการนี้มีรายการให้เลือกได้ 4 รายการ ได้แก่

1) Web Page Storage เป็นส่วนที่ใช้ในการจัดเก็บเว็บเพจ โดยรายละเอียดของการทำงานในส่วนนี้จะอธิบายไว้ในข้อ 4.3.1

2) Query Web Page by Keywords and Feedback With Vector Space Model เป็นส่วนที่ใช้ในการค้นคืนเว็บเพจและมีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ โดยรายละเอียดของการทำงานในส่วนนี้จะอธิบายไว้ในข้อ 4.3.2

3) Query Web Page by Keywords and Feedback With VIPS Algorithm and Vector Space Model เป็นส่วนที่ใช้ในการค้นคืนเว็บเพจและมีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ โดยรายละเอียดของการทำงานในส่วนนี้จะอธิบายไว้ในข้อ 4.3.2

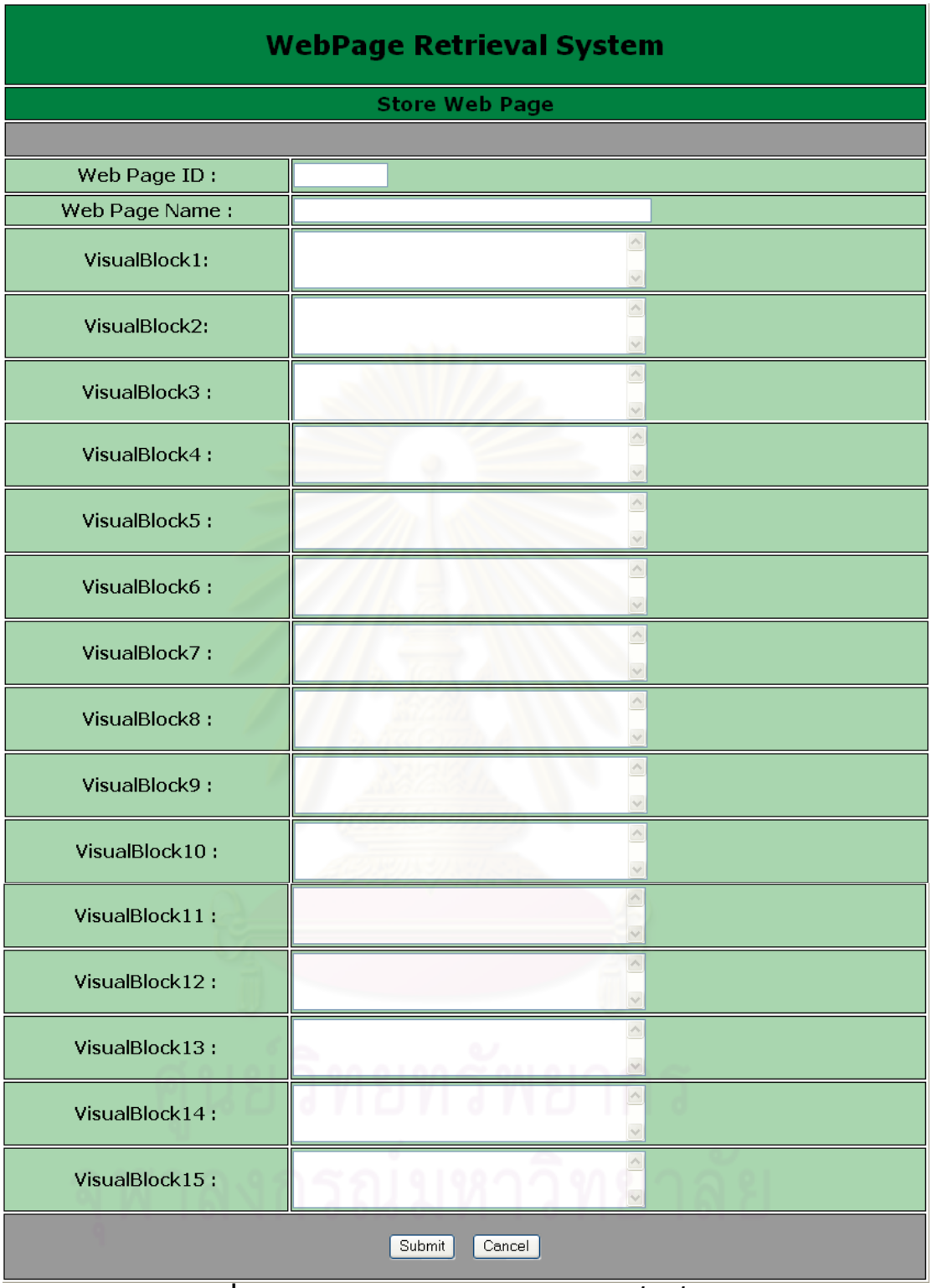
4) Query Web Page by Keywords and Feedback With VIPS Algorithm and Probabilistic Model เป็นส่วนที่ใช้ในการค้นคืนเว็บเพจและมีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น โดยรายละเอียดของการทำงานในส่วนนี้จะอธิบายไว้ในข้อ 4.3.2



รูปที่ 4.3 หน้าจอแสดงรายการโครงสร้างหลักของเครื่องมือ

4.3.1 การจัดเก็บเว็บเพจ

การจัดเก็บเว็บเพจ ในส่วนนี้จะดำเนินการหลังจากทำการจัดเก็บเว็บเพจจากเว็บไซต์จำนวน 300 เว็บเพจ โดยแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม ก่อนที่จะป้อนข้อมูลเว็บเพจเหล่านั้นผ่านทางส่วนต่อประสานกับผู้ใช้ หน้าจอสำหรับจัดเก็บเว็บเพจนี้ จึงออกแบบให้เก็บเนื้อหาของเว็บเพจเป็นบล็อก เพื่อช่วยสนับสนุนในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ เพราะในส่วนของการให้ผลป้อนกลับที่ตรงประเด็น ผู้ใช้จะต้องเลือกบล็อกที่เห็นว่าตรงประเด็นกับที่ต้องการ โดยแบ่งการจัดเก็บชื่อของเว็บเพจเป็นส่วนแรก แล้วตามด้วยส่วนเนื้อหาทั้งหมดของเว็บเพจที่แบ่งเป็นบล็อกด้วยวีไอพีเอสอัลกอริทึมแล้ว หลังจากผู้ใช้งานใส่เนื้อหาทั้งหมดของเว็บเพจเรียบร้อย และกดปุ่ม Submit แล้ว ข้อมูลในเอกสารนี้จะถูกแปลงเป็นชุดของดรรชนีโดยผ่านการทำดรรชนีอัตโนมัติ (Automatic Indexing) และจัดเก็บไว้ในฐานข้อมูลเพื่อใช้ในการค้นคืนต่อไป ซึ่งการทำดรรชนีอัตโนมัตินั้น ได้กล่าวไว้ใน หัวข้อที่ 2.1.1.1 ของบทที่ 2 ในส่วนการจัดเก็บสารสนเทศ หน้าจอในการจัดเก็บเว็บเพจแสดงได้ดังรูปที่ 4.4

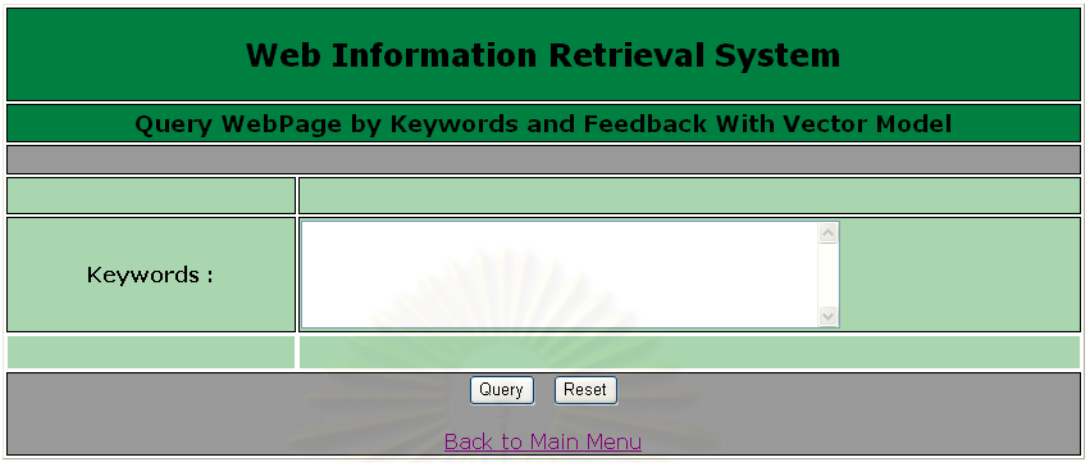


รูปที่ 4.4 หน้าจอแสดงส่วนสำหรับการจัดเก็บเว็บเพจ

4.3.2 การค้นคืนเว็บเพจ

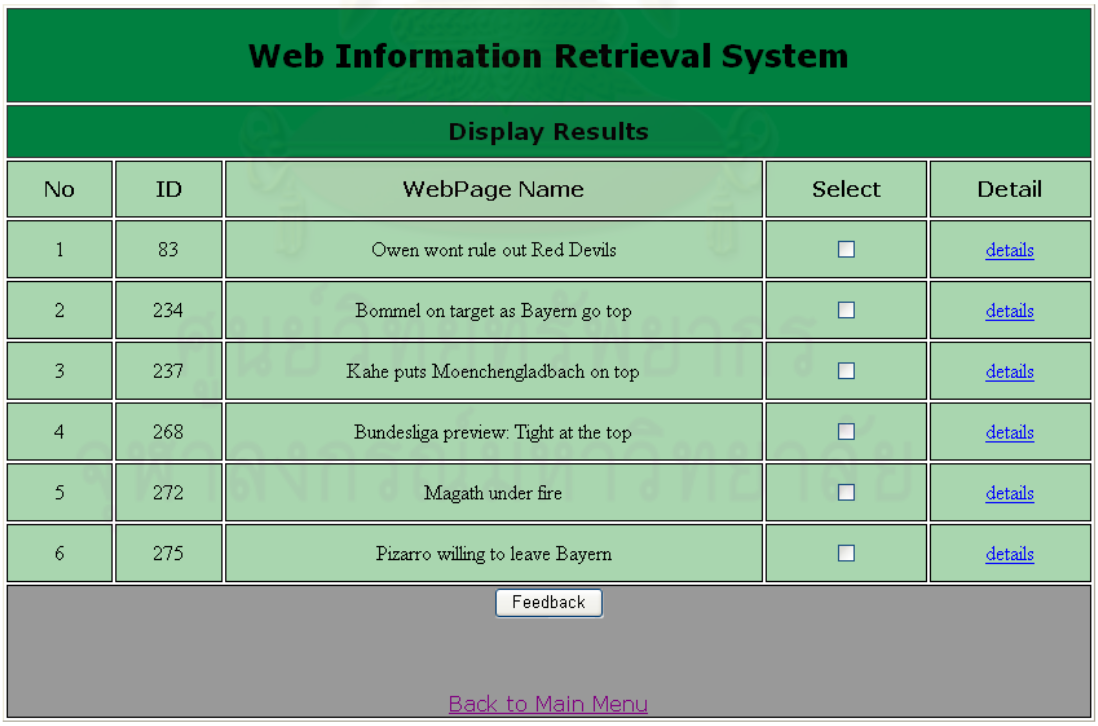
ในการค้นคืนเว็บเพจ มีการออกแบบส่วนต่อประสานกับผู้ใช้ให้สามารถค้นหาเว็บเพจโดยการใช้คำ ซึ่งผู้ใช้จะทำการป้อนข้อคำถามที่เป็นภาษาอังกฤษ ลงในช่องข้อคำถาม จากนั้นกดปุ่ม Query แล้วระบบจะทำการแปลงคำข้อคำถามเหล่านั้น ให้กลายเป็นชุดของดรรชนี เพื่อ

นำไปหาค่าความคล้ายระหว่างชุดตรรชนีของข้อคำถาม กับชุดตรรชนีของเว็บเพจที่มีในฐานข้อมูล หน้าจอสำหรับการค้นคืนเว็บเพจ แสดงดังรูปที่ 4.5

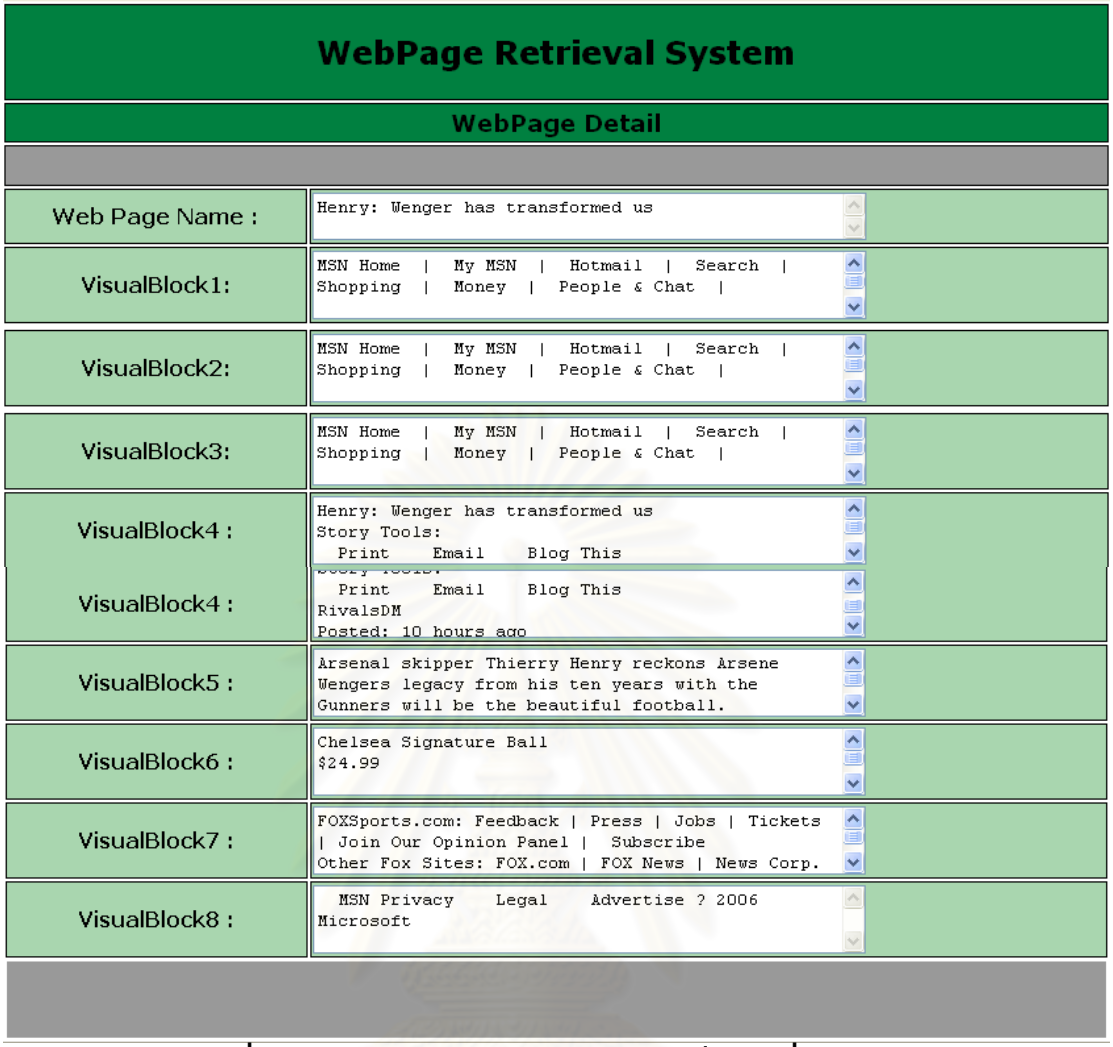


รูปที่ 4.5 หน้าจอแสดงส่วนการค้นคืนเว็บเพจ

ผลลัพธ์ที่ได้จากการค้นคืนเว็บเพจนั้น ประกอบด้วยส่วนหลักๆ 2 ส่วน คือ ส่วนของผลที่ได้จากการค้นคืนเว็บเพจ ซึ่งจะแสดงชื่อของเว็บเพจ ดังรูปที่ 4.6 และส่วนที่สองแสดงรายละเอียดของแต่ละเว็บเพจ ดังรูปที่ 4.7



รูปที่ 4.6 หน้าจอแสดงผลลัพธ์ของเว็บเพจที่ค้นคืนมาได้โดยไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้



รูปที่ 4.7 หน้าจอแสดงรายละเอียดของเว็บเพจที่ผู้ใช้เลือก

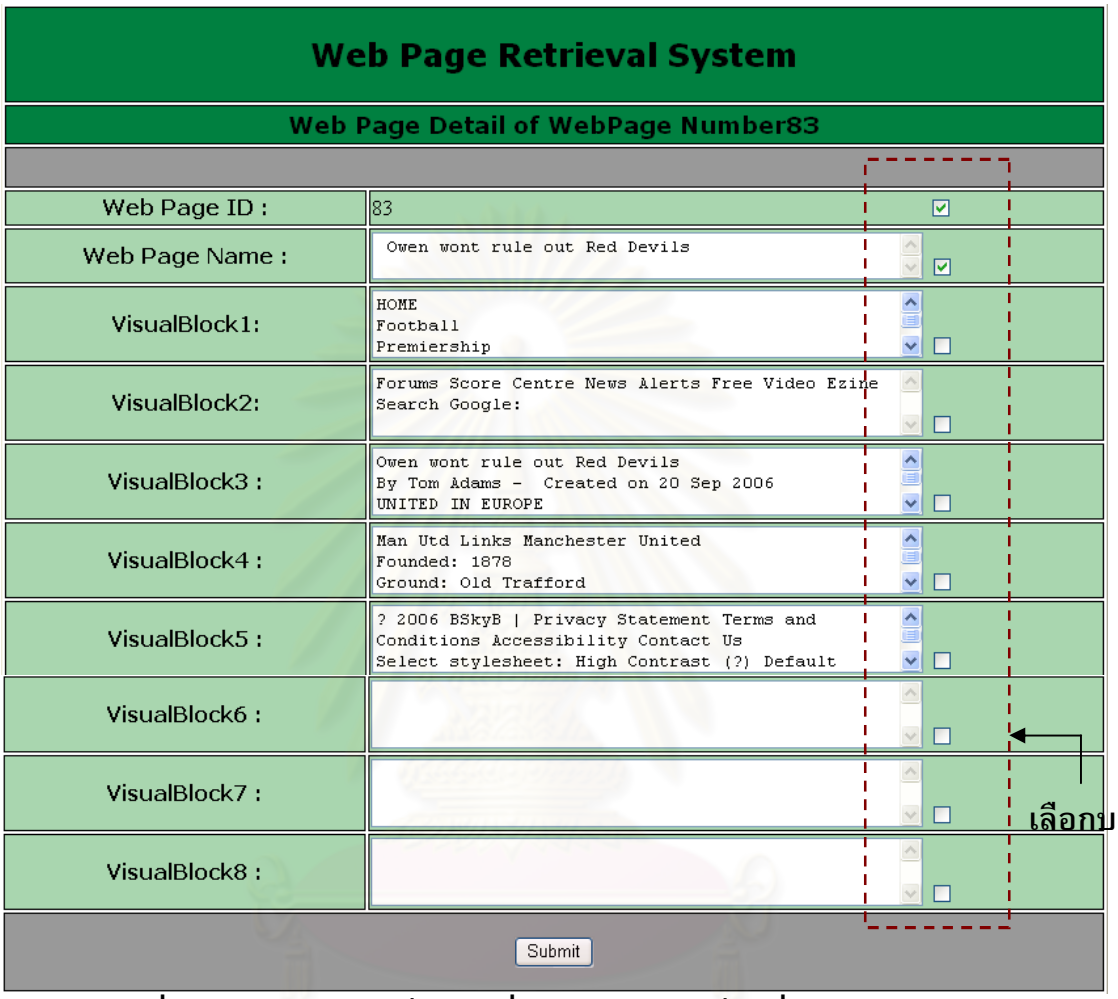
สำหรับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น มีทั้งหมด 3 วิธีการ ดังนี้

1) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยใช้แบบจำลองปริภูมิเวคเตอร์ วิธีการนี้ผู้ใช้จะทำการเลือกเว็บเพจที่เห็นว่าตรงประเด็นกับที่ต้องการ เพื่อนำคำในเว็บเพจที่ผู้ใช้เลือกไปทำการกำหนดข้อคำถามใหม่ โดยใช้แบบจำลองปริภูมิเวคเตอร์ในการขยายคำและค่าน้ำหนักของคำ

2) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยใช้วีไอพีเอสอัลกอริทึม และแบบจำลองปริภูมิเวคเตอร์ วิธีการนี้ผู้ใช้จะทำการเลือกเว็บเพจที่เห็นว่าตรงประเด็นกับที่ต้องการ และเลือกบล็อกของเว็บเพจที่เห็นว่าตรงตามต้องการ เพื่อนำคำในบล็อกที่ผู้ใช้เลือกนั้น ไปช่วยในการกำหนดข้อคำถามใหม่ ซึ่งใช้แบบจำลองปริภูมิเวคเตอร์ในการขยายคำและค่าน้ำหนักของคำ

3) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยใช้วีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็น วิธีการนี้ผู้ใช้จะทำการเลือกเว็บเพจที่เห็นว่าตรงประเด็นกับที่ต้องการ และเลือกบล็อกของเว็บเพจที่เห็นว่าตรงตามต้องการ เพื่อนำคำในบล็อกที่ผู้ใช้เลือกนั้น ไปกำหนดข้อคำถามใหม่ ซึ่งใช้แบบจำลองความน่าจะเป็นในการเปลี่ยนแปลงค่าน้ำหนักของคำ

โดยหน้าจอของการเลือกบล็อกที่ผู้ใช้เห็นว่าตรงประเด็น จากเว็บเพจที่ผู้ใช้เลือกแล้วว่าตรงประเด็นกับที่ต้องการ แสดงได้ดังรูปที่ 4.8 และหลังจากผู้ใช้ให้ผลป้อนกลับแล้ว ระบบจะแสดงผลการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ที่ได้กลับสู่ผู้ใช้งาน แสดงได้ดังรูปที่ 4.9



รูปที่ 4.8 หน้าจอแสดงเว็บเพจเพื่อให้ผู้ใช้เลือกบล็อกที่ตรงตามต้องการ

4.4 แบบจำลองข้อมูล

ในการออกแบบแบบจำลองข้อมูลของระบบการจัดเก็บและค้นคืนเว็บเพจ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น เป็นการออกแบบโครงสร้างข้อมูลและความสัมพันธ์ระหว่างข้อมูลภายในระบบ โดยใช้ฐานข้อมูลเชิงสัมพันธ์ (Relational Database) ซึ่งแบบจำลองข้อมูลทั้งหมดสามารถแสดงเป็นแผนภาพความสัมพันธ์ระหว่างข้อมูล ดังรูปที่ 4.10

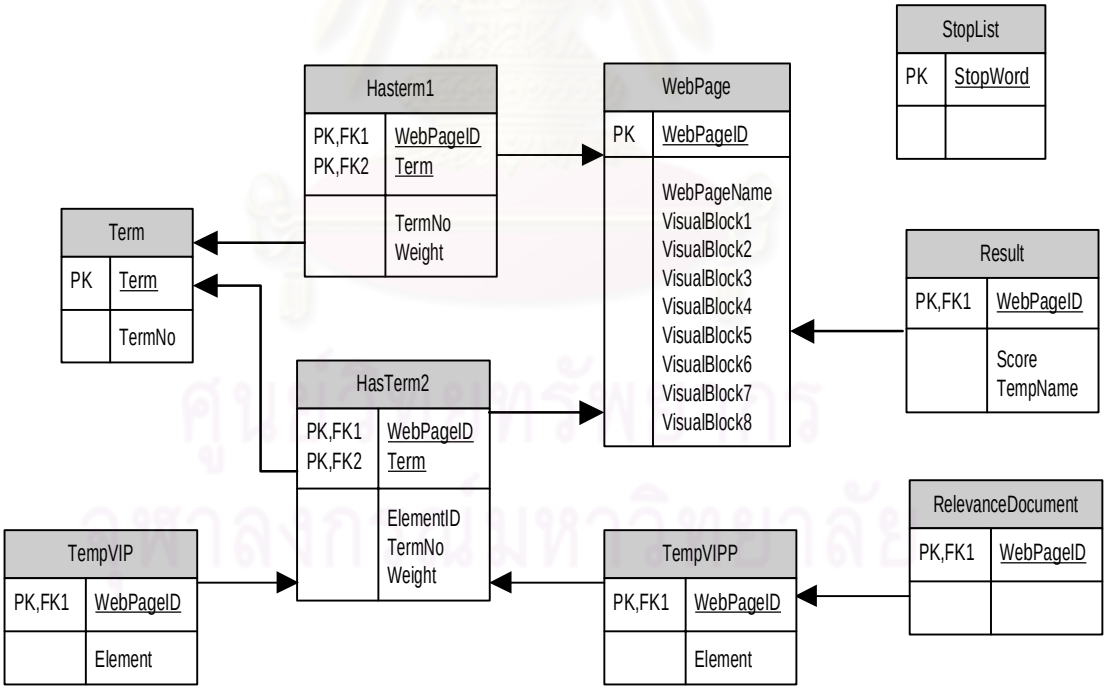
Web Information Retrieval System

Display Results

No	WebPage ID	WebPage Name
1	83	Owen wont rule out Red Devils
2	234	Bommel on target as Bayern go top
3	237	Kahe puts Moenchengladbach on top
4	268	Bundesliga preview: Tight at the top
5	272	Magath under fire
6	275	Pizarro willing to leave Bayern

Back to Main Menu

รูปที่ 4.9 หน้าจอแสดงผลลัพธ์รายการเว็บเพจที่ได้จากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้



รูปที่ 4.10 แผนภาพแสดงความสัมพันธ์ระหว่างข้อมูลของระบบ

จากแผนภาพแสดงความสัมพันธ์ระหว่างข้อมูลของระบบการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้โดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็นสามารถแปลงเป็นตารางได้ทั้งหมด 9 ตาราง ดังรายละเอียดในตารางที่ 4.1

ตารางที่ 4.1 อธิบายตารางข้อมูลของระบบ

หมายเลข	ชื่อตาราง	คำอธิบาย
1	WebPage	ตารางจัดเก็บเอกสารเว็บเพจ
2	Term	ตารางจัดเก็บคำและจำนวนของคำทั้งหมดในชุดข้อมูล
3	HasTerm1	ตารางจัดเก็บคำทั้งหมด และจำนวนคำแต่ละคำในแต่ละเอกสารเว็บเพจ รวมทั้งค่าน้ำหนักของคำแต่ละคำ
4	Hasterm2	ตารางจัดเก็บคำทั้งหมด จำนวนคำแต่ละคำ และส่วนประกอบ ในแต่ละเอกสารเว็บเพจ รวมทั้งค่าน้ำหนักของคำแต่ละคำ
5	TempVIP	ตารางจัดเก็บรหัสเว็บเพจ และส่วนประกอบเว็บเพจที่ได้จากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองปริภูมิเวคเตอร์
6	TempVIPP	ตารางจัดเก็บรหัสเว็บเพจ และส่วนประกอบเว็บเพจที่ได้จากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองความน่าจะเป็น
7	RelevanceDocument	ตารางจัดเก็บรหัสเว็บเพจเพื่อบอกจำนวนเว็บเพจที่ได้จากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้
8	Result	ตารางจัดเก็บผลการค้นคืนทั้งหมด
9	Stoplist	ตารางจัดเก็บคำ Stop words ทั้งหมด

จากตารางอธิบายตารางข้อมูลของระบบ ผู้วิจัยได้สร้างโครงสร้างของแต่ละตารางข้อมูลดังต่อไปนี้

ตารางที่ 4.2 โครงสร้างตารางข้อมูล WebPage

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WPID	รหัสเว็บเพจ	varchar	4	PK
Name	ชื่อเว็บเพจ	text		
VisualBlock1	เนื้อหาเว็บเพจในส่วนประกอบที่ 1	mediumtext		
VisualBlock2	เนื้อหาเว็บเพจในส่วนประกอบที่ 2	mediumtext		
VisualBlock3	เนื้อหาเว็บเพจในส่วนประกอบที่ 3	mediumtext		
VisualBlock4	เนื้อหาเว็บเพจในส่วนประกอบที่ 4	mediumtext		
VisualBlock5	เนื้อหาเว็บเพจในส่วนประกอบที่ 5	mediumtext		
VisualBlock6	เนื้อหาเว็บเพจในส่วนประกอบที่ 6	mediumtext		
VisualBlock7	เนื้อหาเว็บเพจในส่วนประกอบที่ 7	mediumtext		
VisualBlock8	เนื้อหาเว็บเพจในส่วนประกอบที่ 8	mediumtext		

ตารางที่ 4.3 โครงสร้างตารางข้อมูล Term

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
Term	คำ	char	30	PK
TermNo	จำนวนคำ	int	10	

ตารางที่ 4.4 โครงสร้างตารางข้อมูล HasTerm1

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	varchar	4	PK,FK1
Term	คำ	char	10	PK,FK2
TermNo	จำนวนคำ	int	10	
Weight	น้ำหนักคำ	float		

ตารางที่ 4.5 โครงสร้างตารางข้อมูล Hasterm2

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	varchar	4	PK,FK1
Term	คำ	char	10	PK,FK2
ElementID	รหัสส่วนประกอบเว็บเพจ	char	10	
TermNo	จำนวนคำ	int	10	
Weight	น้ำหนักคำ	float		

ตารางที่ 4.6 โครงสร้างตารางข้อมูล TempVIP

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	varchar	4	PK,FK1
Element	ส่วนประกอบเว็บเพจ	int	10	

ตารางที่ 4.7 โครงสร้างตารางข้อมูล TempVIPP

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	varchar	4	PK,FK1
Element	ส่วนประกอบเว็บเพจ	int	10	

ตารางที่ 4.8 โครงสร้างตารางข้อมูล RelevanceDocument

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	char	30	PK,FK1

ตารางที่ 4.9 โครงสร้างตารางข้อมูล Result

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
WebPageID	รหัสเว็บเพจ	varchar	4	PK,FK1
Score	ค่าความคล้าย	float		
TempName	ชื่อเรื่อง	text		

ตารางที่ 4.10 โครงสร้างตารางข้อมูล Stoplist

คุณลักษณะ	คำอธิบาย	ชนิด	ขนาด	คีย์
StopWord	คำ Stop word	char	20	PK

บทที่ 5
การทดลอง

สำหรับบทนี้จะกล่าวถึงรายละเอียดของการทดลอง เครื่องมือที่ผู้วิจัยพัฒนาขึ้นเพื่อทดสอบแนวคิดที่ได้นำเสนอไว้ รวมถึงวัตถุประสงค์ของการทดลอง วิธีการทดลอง ขั้นตอนการทดลอง สภาพแวดล้อมการทดลอง ผลการทดลอง และการวิเคราะห์ผลการทดลอง ซึ่งมีรายละเอียดดังต่อไปนี้

5.1 วัตถุประสงค์ของการทดลอง

ในการทดลองครั้งนี้ มีวัตถุประสงค์เพื่อทดสอบสมมุติฐานของงานวิจัย นั่นคือ การค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ มีประสิทธิผลมากกว่าการค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้หรือไม่ โดยเปรียบเทียบประสิทธิผลทั้งในแบบจำลองปริภูมิเวคเตอร์ และแบบจำลองความน่าจะเป็น ซึ่งมีการใช้วีไอพีเอสอัลกอริทึมด้วยทั้ง 2 แบบ และแบบที่ไม่ได้ใช้วีไอพีเอสอัลกอริทึม ซึ่งสามารถสรุปได้เป็น 4 วิธีการด้วยกัน โดยจะกล่าวถึงในหัวข้อที่ 5.2.2

โดยในการทดลองนั้น จะมีการตั้งเงื่อนไขและปัจจัยควบคุม เพื่อให้การทดลองเป็นไปอย่างเหมาะสม และทำการเปรียบเทียบผลการค้นคืนเว็บเพจและการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองต่างๆ โดยผลลัพธ์ที่ได้จากแต่ละวิธีการค้นคืนเว็บเพจและการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นสามารถประเมินประสิทธิผลได้ด้วย ค่าความแม่นยำ และค่าเรียกคืน

5.2 วิธีการทดลอง

ในการออกแบบวิธีการทดลองนั้น จะต้องมีปัจจัยควบคุมที่เหมาะสม เพื่อลดความโน้มเอียงหรืออคติที่จะเกิดขึ้น และส่งผลกระทบต่อผลการทดลองได้ สำหรับการทดลองในวิทยานิพนธ์นี้ มีปัจจัยที่ต้องควบคุม ดังต่อไปนี้

5.2.1 เว็บเพจ (Web Pages)

เว็บเพจที่ใช้ในการทดลองครั้งนี้ มีทั้งสิ้น 300 เว็บเพจ ซึ่งมีเนื้อหาเกี่ยวข้องกับข่าวกีฬา เนื่องจากผู้วิจัยมีความสนใจ และมีความเข้าใจในเนื้อหาด้านนี้ นอกจากนี้เพื่อให้เนื้อหาของข้อมูลหรือเอกสารที่ใช้ในระบบมีเชิงลึกและเฉพาะเจาะจงมากขึ้น โดยทำการเก็บรวบรวมต่อเนื่องกันเป็นเวลา 5 วัน จากวันที่ 19 กันยายน 2549 ถึงวันที่ 23 กันยายน 2549 เพื่อให้เนื้อหาข่าวมีความต่อเนื่องและเป็นไปในแนวทางเดียวกัน ซึ่งเป็นภาษาอังกฤษทั้งหมดจากเว็บไซต์ทั้งหมด 8 เว็บไซต์ โดยรายละเอียดของเว็บเพจที่จัดเก็บมาทั้งหมดสามารถแสดงได้ใน ภาคผนวก ก

5.2.2 วิธีการค้นคืนเว็บเพจ (Web Pages Retrieval Methods)

สำหรับวิธีการค้นคืนเว็บเพจนั้น จะทำการค้นคืนแบบที่ไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และการค้นคืนแบบใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ จึงมีการพัฒนาเครื่องมือให้สามารถรองรับได้บนแบบจำลองทั้ง 2 แบบคือแบบจำลองปริภูมิเวคเตอร์และแบบจำลองความน่าจะเป็น เพื่อให้สอดคล้องกับการทดลอง โดยมีฟังก์ชันการค้นคืนเว็บเพจที่รองรับการทดลองทั้ง 4 วิธีการ ดังต่อไปนี้

1) การค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้
2) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์
3) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์
4) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

ในงานวิจัยนี้จะไม่มีการใช้หน่วยตัวอย่างในการทดลอง เนื่องจากขอบเขตเนื้อหาของเว็บเพจที่ทำการจัดเก็บนั้นมีลักษณะเฉพาะเจาะจง และผู้วิจัยมีการเรียนรู้และทำความเข้าใจในเนื้อหาของเว็บเพจที่ทำการจัดเก็บ ดังนั้นผู้วิจัยจึงเป็นผู้ที่ทำการทดลองในการค้นคืนเว็บเพจแต่เพียงผู้เดียว ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น กำหนดให้สามารถเลือกรายการเอกสารเว็บเพจที่เห็นว่าตรงประเด็นกับที่ต้องการได้ 2 เว็บเพจ สำหรับกรณีที่มีการใช้วีไอพีเอสอัลกอริทึม สามารถเลือกบล็อกได้เว็บเพจละ 1 บล็อก และเนื้อหาของบล็อกนั้นเป็นเนื้อหาเดียวกันกับหัวข้อข่าวของเว็บเพจนั้นๆ เนื่องจากว่าการใช้ข้อคำถามที่ประกอบด้วยคำมากเกินไปในการค้นคืนไม่ก่อให้เกิดประโยชน์ และการดำเนินการของข้อคำถามกับระบบเป็นการดำเนินการแบบ ออร์ (Or Operation) จึงจำเป็นต้องจำกัดจำนวนของเว็บเพจและจำนวนของบล็อกในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

5.2.3 ข้อคำถาม (Queries)

เนื่องจากผู้วิจัยเป็นผู้ที่ทำการทดลองในการค้นคืนเว็บเพจ ดังนั้นข้อคำถามที่ใช้ในการค้นคืนเว็บเพจจะได้มาจากการพิจารณาของผู้วิจัย ซึ่งมีหลักการสำคัญในการสร้างข้อคำถามคือข้อคำถามที่สร้างขึ้นเกิดจากการเรียนรู้ ความเข้าใจและความคุ้นเคย ในเนื้อหาของเว็บเพจที่ทำการจัดเก็บโดยพยายามให้ข้อคำถามปรากฏในข่าวกีฬาทุกประเภท เพื่อให้เกิดความครอบคลุมในเนื้อหาของข่าวกีฬาทั้งหมดที่จัดเก็บในฐานข้อมูล เป็นคำที่มีความหมายและมีรายการเว็บเพจที่ตรงประเด็นมากกว่า 1 รายการ ซึ่งมีทั้งหมด 50 ข้อคำถาม ดังนี้

1) ข้อคำถามที่ประกอบด้วยคำ 1 คำ จำนวน 10 ข้อคำถาม
2) ข้อคำถามที่ประกอบด้วยคำ 2 คำ จำนวน 10 ข้อคำถาม
3) ข้อคำถามที่ประกอบด้วยคำ 3 คำ จำนวน 10 ข้อคำถาม
4) ข้อคำถามที่ประกอบด้วยคำ 4 คำ จำนวน 10 ข้อคำถาม

5) ข้อคำถามที่ประกอบด้วยคำ 5 คำ จำนวน 10 ข้อคำถาม

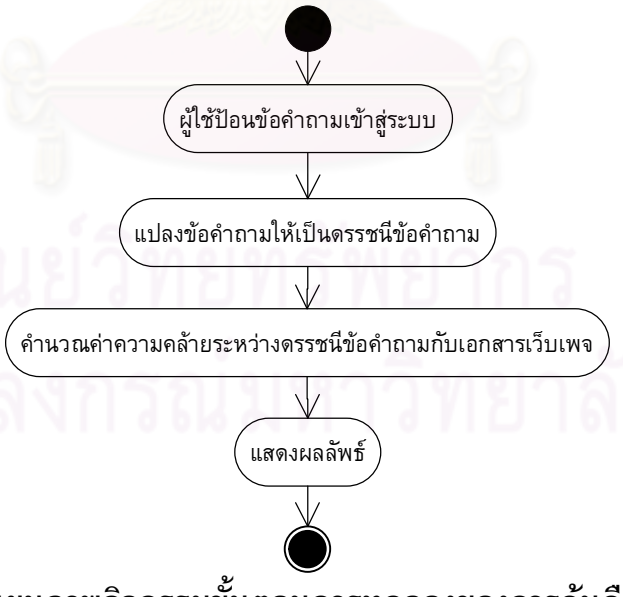
ข้อคำถามทั้งหมดนี้ จะใช้ในการค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้และที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยใช้แบบจำลองปริภูมิเวคเตอร์ และแบบจำลองความน่าจะเป็น ซึ่งการแบ่งข้อคำถามให้ประกอบด้วยคำที่มีขนาดตั้งแต่ 1 คำไปจนถึง 5 คำ เพื่อให้เกิดความหลากหลายของข้อคำถามและจากประสบการณ์โดยทั่วไปผู้ที่ทำการค้นคืนระบบสารสนเทศส่วนใหญ่จะใช้ข้อคำถามที่ประกอบด้วยคำจำนวนน้อยเช่น 1 หรือ 2 คำ แต่เพื่อให้เกิดความครอบคลุมเนื้อหาในคอลเลคชัน และเพื่อเพิ่มความมั่นใจในการค้นคืนเว็บเพจที่ต้องการลักษณะเฉพาะการใช้คำเพียง 1 หรือ 2 คำ อาจจะไม่เพียงพอ จึงได้กำหนดให้ข้อคำถามที่ขนาดคำได้ถึง 5 คำ นอกจากนี้เพื่อต้องการตรวจสอบขนาดของข้อคำถามว่ามีผลต่อการทดลองในเรื่องของค่าความแม่นยำ และค่าเรียกคืนหรือไม่

5.3 ขั้นตอนการทดลอง

สำหรับขั้นตอนในการทดลองนี้แบ่งได้เป็น 4 วิธีการ ซึ่งมีลำดับขั้นตอนและการดำเนินการของแต่ละวิธีการดังต่อไปนี้

5.3.1 การค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

ขั้นตอนการทดลองในการค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้เริ่มจากผู้ใช้ป้อนข้อคำถามเข้าสู่ระบบ แล้วระบบเปลี่ยนข้อคำถามเหล่านั้นเป็นชุดดรรชนีข้อคำถาม เพื่อใช้ในการคำนวณหาค่าความคล้ายระหว่างชุดดรรชนีของข้อคำถาม กับเอกสารเว็บเพจ จากนั้นแสดงผลลัพธ์ที่ได้จากการค้นคืนกลับสู่ผู้ใช้งาน แสดงได้ดังรูปที่ 5.1



รูปที่ 5.1 แผนภาพกิจกรรมขั้นตอนการทดลองของการค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

5.3.2 การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์

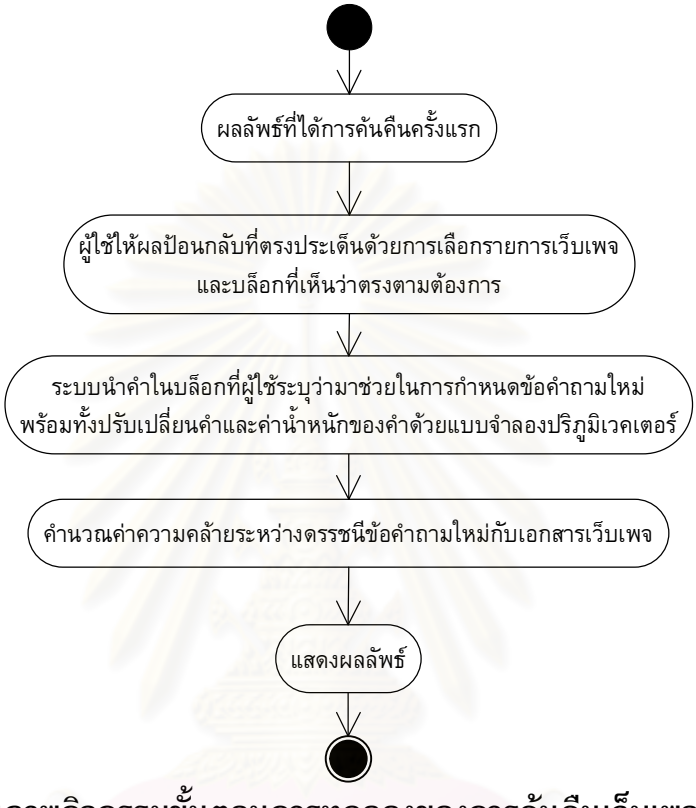
ขั้นตอนการทดลองในการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ เริ่มจากผู้ใช้ให้ผลป้อนกลับสู่ระบบด้วยการเลือกรายการเอกสารที่ค้นคืนมาได้ครั้งแรกและเห็นว่าตรงประเด็น จากนั้นระบบจะเลือกคำที่มีค่าความถี่สูงสุดจากรายการเอกสารที่ผู้ใช้เลือกแล้ว ซึ่งจะใช้การเลือกแบบสุ่มถ้าในกรณีมีคำที่มีค่าความถี่สูงสุดมากกว่า 1 คำ นำคำที่ได้ไปช่วยในการกำหนดข้อคำถามใหม่ พร้อมทั้งเปลี่ยนแปลงคำและค่าน้ำหนักของคำด้วยแบบจำลองปริภูมิเวคเตอร์ เพื่อใช้ในการค้นคืนอีกครั้งด้วยข้อคำถามใหม่ โดยเปรียบเทียบค่าความคล้ายของข้อคำถามใหม่กับเอกสารเว็บเพจ แล้วแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้งาน แสดงได้ดังรูปที่ 5.2

รูปที่ 5.2 แผนภาพกิจกรรมขั้นตอนการทดลองของการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์

5.3.3 การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์

ขั้นตอนการทดลองในการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ เริ่มจากผู้ใช้ให้ผลป้อนกลับสู่ระบบด้วยการเลือกรายการเอกสารเว็บเพจ และบล็อกที่ได้จากการแบ่งเป็นส่วนย่อยด้วยวีไอพีเอส

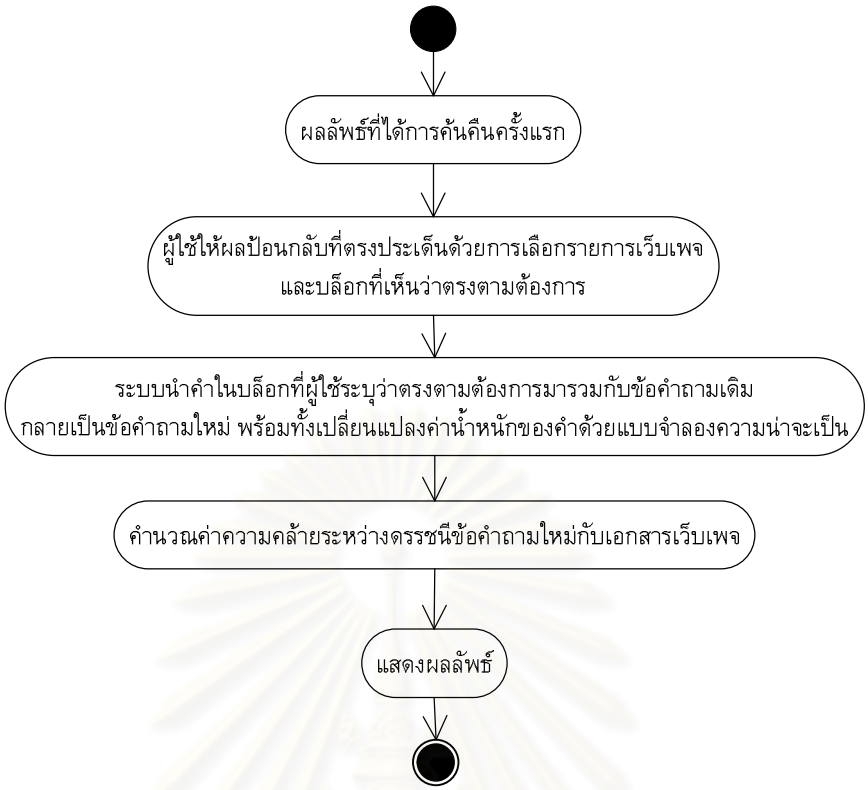
อัลกอริทึม ที่ค้นคืนมาได้ครั้งแรกและเห็นว่าตรงตามประเด็น จากนั้นระบบนำคำที่ได้จากบล็อกที่ผู้ใช้เลือกแล้วไปช่วยในการกำหนดข้อคำถามใหม่ พร้อมทั้งเปลี่ยนแปลงคำและค่าน้ำหนักของคำด้วยแบบจำลองปริภูมิเวคเตอร์ เพื่อใช้ในการค้นคืนอีกครั้งด้วยข้อคำถามใหม่ โดยเปรียบเทียบค่าความคล้ายของข้อคำถามใหม่กับเอกสารเว็บเพจ แล้วแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้งาน แสดงได้ดังรูปที่ 5.3



รูปที่ 5.3 แผนภาพกิจกรรมขั้นตอนการทดลองของการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์

5.3.4 การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

ขั้นตอนการทดลองในการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น เริ่มจากผู้ใช้ให้ผลป้อนกลับสู่ระบบด้วยการเลือกรายการเอกสาร และบล็อกที่ได้จากการแบ่งเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม ที่ค้นคืนมาได้ครั้งแรกและเห็นว่าตรงประเด็น จากนั้นระบบนำคำที่ได้จากบล็อกที่ผู้ใช้เลือกแล้วไปรวมกับข้อคำถามเดิมกลายเป็นข้อคำถามใหม่ พร้อมทั้งเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น เพื่อใช้ในการค้นคืนอีกครั้งด้วยข้อคำถามใหม่ โดยเปรียบเทียบค่าความคล้ายของข้อคำถามใหม่กับเอกสารเว็บเพจ แล้วแสดงผลการค้นคืนที่ได้กลับสู่ผู้ใช้งาน แสดงได้ดังรูปที่ 5.4



รูปที่ 5.4 แผนภาพกิจกรรมขั้นตอนการทดลองของการค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

5.4 การกำหนดค่าขีดแบ่งเริ่มต้นความคล้าย

ในการค้นคืนระบบสารสนเทศนั้นจะต้องมีการกำหนด ค่าขีดแบ่งเริ่มต้นความคล้าย (Similarity Threshold) ซึ่งเป็นค่าตัวเลขค่าหนึ่ง ที่ใช้เพื่อกำหนดให้ระบบค้นคืนนำเอกสารหนึ่ง ๆ ออกมาแสดงแก่ผู้ค้นคืน ถ้าเอกสารเหล่านั้นมีค่าความคล้ายกับข้อคำถามที่ใช้ในการค้นคืน มากกว่าหรือเท่ากับค่าขีดแบ่งเริ่มต้นความคล้ายที่กำหนดไว้ โดยผู้วิจัยใช้ชุดข้อคำถามจำนวน 50 ข้อ ที่ได้กำหนดไว้ มาทำการทดสอบกับเครื่องมือที่สร้างขึ้น โดยมีวิธีการค้นคืนดังกล่าวไว้ในหัวข้อที่ 5.2.2 และทำการค้นคืน จากเว็บเพจ จำนวน 300 เว็บเพจ ซึ่งค่าขีดแบ่งเริ่มต้นความคล้าย หาได้จากค่าเฉลี่ยความคล้ายที่ได้จากวิธีการค้นคืนของงานวิจัยนี้ ดังสมการที่ 7

ค่าขีดแบ่งเริ่มต้นความคล้าย = ค่าเฉลี่ยความคล้าย (7)

ซึ่งค่าขีดแบ่งเริ่มต้นความคล้ายนั้นอาจจะใช้ค่าเฉลี่ยหักออกด้วยค่าเบี่ยงเบนมาตรฐาน ถ้าในกรณีที่ต้องการให้เอกสารที่ค้นคืนมาได้มีจำนวนมากขึ้น หรืออาจจะใช้ค่าเฉลี่ยรวมกับค่าเบี่ยงเบนมาตรฐาน ในกรณีที่ต้องการให้เอกสารที่ค้นคืนมาได้มีจำนวนน้อยลง แต่งานวิจัยนี้จะ

ใช้ค่าขีดแบ่งเริ่มต้นความคล้ายเท่ากับค่าเฉลี่ย เพราะเอกสารที่จัดเก็บทั้งหมดมีขนาดเล็กและไม่ต้องการให้รายการเอกสารที่ไม่ตรงประเด็นถูกนำมาแสดงในผลการค้นคืนที่ได้มากเกินไป ดังนั้นค่าขีดแบ่งเริ่มต้นความคล้ายของการค้นคืนเว็บเพจในงานวิจัยนี้ มีค่าเท่ากับ 0.000703163

5.5 ผลการทดลอง

ผลลัพธ์ที่ได้จากการค้นคืนเว็บเพจทั้งหมดจากการทดลอง ทั้ง 4 วิธีการ แสดงไว้ในภาคผนวก ค ซึ่งสามารถสรุปได้เป็นค่าเฉลี่ยของค่าเรียกคืน ค่าเฉลี่ยของค่าความแม่นยำ แบ่งตามขนาดของข้อคำถามซึ่งประกอบด้วยคำตั้งแต่ 1 คำ ถึง 5 คำ โดยทำการเปรียบเทียบระหว่างการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบต่างๆ ดังตารางที่ 5.1

ตารางที่ 5.1 แสดงค่าเฉลี่ยค่าเรียกคืน ค่าความแม่นยำ ของการค้นคืนแบบไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองต่าง ๆ โดยแบ่งตามขนาดของข้อคำถามที่ใช้ในการค้นคืน

ขนาดของข้อคำถาม (คำ)	ไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้		ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้											
			แบบจำลองปริภูมิเวคเตอร์				วีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์				วีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น			
	R	P	R	P	ร้อยละ เพิ่มขึ้น (+) / ลดลง (-)		R	P	ร้อยละ เพิ่มขึ้น (+) / ลดลง (-)		R	P	ร้อยละ เพิ่มขึ้น (+) / ลดลง (-)	
					R	P			R	P			R	P
1	0.844	0.423	0.766	0.647	-6.58	+52.95	0.823	0.427	-2.48	+0.95	0.755	0.306	-10.54	-27.66
2	0.984	0.295	0.978	0.415	-0.61	+40.67	0.987	0.289	+3.04	-2.03	1.000	0.255	+1.63	-1.54
3	0.762	0.350	0.784	0.368	+2.89	+5.14	0.851	0.351	+11.24	+0.29	0.926	0.353	+21.52	+0.86
4	0.914	0.207	0.943	0.236	+3.17	+14.01	0.941	0.198	+2.95	-0.34	0.965	0.206	+5.58	-0.48
5	0.906	0.222	0.961	0.177	+6.07	-20.27	1.000	0.178	+10.37	-19.81	0.996	0.167	+9.93	-24.77
ค่าเฉลี่ย	0.882	0.299	0.886	0.369	+0.99	+23.41	0.920	0.289	+4.31	-3.34	0.928	0.258	+5.22	-13.71

จากตารางที่ 5.1 จะเห็นได้ว่าค่าเฉลี่ยของค่าเรียกคืนที่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ทุกกรณี มีค่ามากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้เพียงเล็กน้อย คือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าเรียกคืนเพิ่มขึ้นร้อยละ 0.99 การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์มีค่าเรียกคืนเพิ่มขึ้นร้อยละ 4.31 และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าเรียกคืนเพิ่มขึ้นร้อยละ 5.22 แสดงว่าหลังจากทำการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้แล้วระบบจะทำการค้นคืนรายการเอกสารที่ตรงตามความต้องการของผู้ใช้ได้มากขึ้น

ส่วนค่าเฉลี่ยค่าความแม่นยำของการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองปริภูมิเวคเตอร์ให้ค่ามากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ถึงร้อยละ 23.41 แสดงว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ในแบบจำลองปริภูมิเวคเตอร์สามารถลดรายการเอกสารที่ไม่ตรงประเด็นกับความต้องการของผู้ใช้ลดน้อยลง ทั้งนี้เนื่องมาจากคำที่ใช้เพิ่มเข้าไปในข้อคำถามใหม่ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นได้มาจากคำที่มีค่าน้ำหนักสูงสุดของเว็บเพจที่ผู้ใช้เลือกเพียงคำเดียว ทำให้ข้อคำถามใหม่ที่ได้ประกอบด้วยคำที่เพิ่มขึ้นเพียง 1 คำ จึงไม่ส่งผลกระทบต่อค่าความแม่นยำ แต่กลับทำให้มีค่าความแม่นยำเพิ่มขึ้นด้วยเพราะมีการปรับเปลี่ยนคำและค่าน้ำหนักของคำด้วยแบบจำลองปริภูมิเวคเตอร์

แต่การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่า การไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ เพียงเล็กน้อย แสดงว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ทำให้ระบบการค้นคืนแสดงรายการเอกสารที่ไม่ตรงประเด็นเพิ่มขึ้นด้วย ซึ่งมีค่าเหมือนกับกรณีของการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นเช่นกัน แต่ทั้งนี้ทั้งนั้นก็เนื่องมาจากขนาดของข้อคำถาม เพราะจะเห็นได้ชัดว่า เมื่อขนาดของคำในข้อคำถามมีขนาดเพิ่มขึ้น คือขนาดของข้อคำถามที่ประกอบด้วยคำ 1 คำ ไปจนถึง 5 คำ ค่าความแม่นยำของทุกกรณีจะมีค่าลดลง นั่นแสดงให้เห็นถึงขนาดของข้อคำถามมีผลต่อค่าความแม่นยำ

นอกจากนี้ ในกรณีของการใช้วีไอพีเอสอัลกอริทึมสำหรับแบบจำลองทั้ง 2 แบบในการเลือกบล็อกของเว็บเพจที่ผู้ใช้เห็นว่าตรงประเด็นกับที่ต้องการเพื่อนำคำในบล็อกที่ผู้ใช้เลือกแล้วไปช่วยในการกำหนดข้อคำถามใหม่ ซึ่งบล็อกเหล่านั้นประกอบด้วยคำมากกว่า 5 คำขึ้นไป ทำให้ขนาดของข้อคำถามใหม่ที่ใช้ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีคำเพิ่มมากขึ้นด้วย และในการค้นคืนนั้นใช้การดำเนินการแบบ ออร์ (OR Operation) ทำให้ผลลัพธ์ที่ได้มีรายการเอกสารที่ไม่ตรงประเด็นเพิ่มขึ้นไปด้วย ส่งผลให้ค่าความแม่นยำของทั้ง 2 แบบจำลองที่ใช้วีไอพีเอสอัลกอริทึมมีค่าลดลง มากกว่าการไม่ใช้วีไอพีเอสอัลกอริทึมในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

ในทางกลับกันถ้าไม่พิจารณาถึงขนาดของข้อคำถามที่ใช้ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ แต่พิจารณาจากผลการทดลองด้วยการนำค่าความแม่นยำมาเฉลี่ยตามค่าเรียกคืน 11 ค่า จาก 0 ถึง 1 แยกตามวิธีการค้นคืนทั้ง 4 วิธีการจะให้ผลที่แตกต่างออกไปดังตารางที่ 5.2

ตารางที่ 5.2 สรุปผลค่าความแม่นยำเฉลี่ยจากข้อคำถามทั้งหมด 50 ข้อคำถามตามค่าเรียกคืนทั้ง 11 ค่า จาก 0 ถึง 1 ในแต่ละวิธีการค้นคืน

ค่าเรียกคืน	ค่าความแม่นยำ			
	วิธีที่ 1	วิธีที่ 2	วิธีที่ 3	วิธีที่ 4
0.0	0.061	0.333	0.500	0.061
0.1	0.061	0.333	0.500	0.061
0.2	0.500	0.766	0.577	0.400
0.3	0.500	0.766	0.577	0.400
0.4	0.500	0.766	0.577	0.400
0.5	0.102	0.573	0.577	0.400
0.6	0.161	0.555	0.577	0.455
0.7	0.161	0.157	0.591	0.659
0.8	0.571	0.087	0.409	0.733
0.9	0.403	0.724	0.436	0.340
1.0	0.247	0.323	0.228	0.223
ค่าเฉลี่ย	0.297	0.489	0.504	0.376

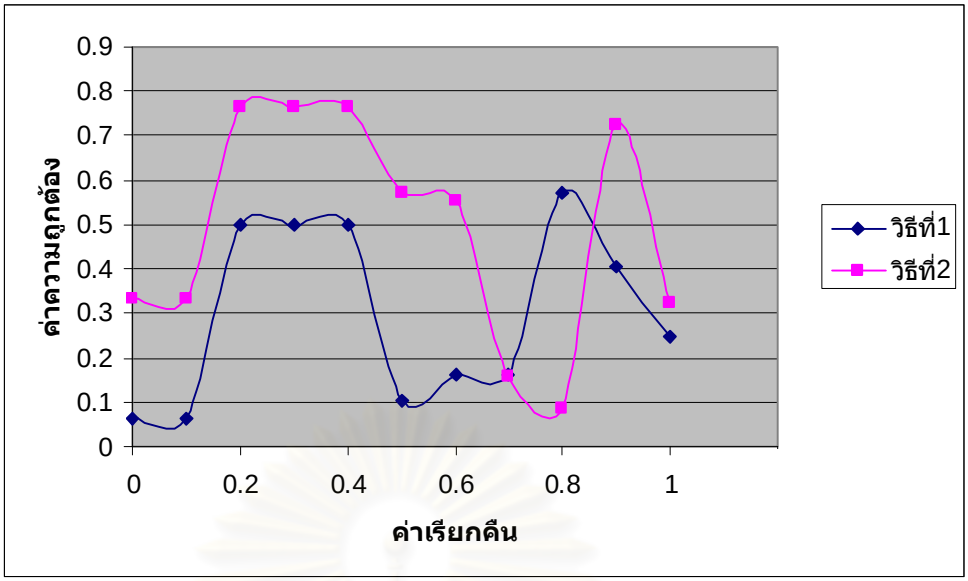
จากตารางที่ 5.2 พบว่า วิธีการค้นคืนวิธีที่ 2 3 และ 4 ให้ค่าความแม่นยำเฉลี่ยมากกว่าวิธีการที่ 1 นั่นแสดงว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ให้ค่าความแม่นยำมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ให้ค่าความแม่นยำมากที่สุดอย่างมีนัยสำคัญ เพราะค่าเฉลี่ยค่าความแม่นยำไม่เพียงพอที่จะแบ่งแยกความแตกต่างได้ ควรทำการทดสอบด้วยการทดสอบสมมุติฐานซึ่งจะกล่าวถึงในหัวข้อที่ 5.6

จากตารางข้างต้นผู้วิจัยนำข้อมูลค่าความแม่นยำเฉลี่ยตามค่าเรียกคืนทั้ง 11 ระดับ มาแสดงในตารางค่าร้อยละของค่าความแม่นยำที่เพิ่มขึ้นหรือลดลงของแต่ละวิธีการ รวมทั้งแสดงผลในรูปแบบกราฟ เปรียบเทียบค่าความแม่นยำ ระหว่างวิธีการใช้และไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ รวมทั้งเปรียบเทียบวิธีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ซึ่งได้แก่

1) คู่ของวิธีการที่ 1 กับ 2 ดังตารางที่ 5.3 และ รูปที่ 5.5
2) คู่ของวิธีการที่ 1 กับ 3 ดังตารางที่ 5.4 และ รูปที่ 5.6
3) คู่ของวิธีการที่ 1 กับ 4 ดังตารางที่ 5.5 และ รูปที่ 5.7
4) คู่ของวิธีการที่ 2 กับ 3 ดังตารางที่ 5.6 และ รูปที่ 5.8
5) คู่ของวิธีการที่ 2 กับ 4 ดังตารางที่ 5.7 และ รูปที่ 5.9
6) คู่ของวิธีการที่ 3 กับ 4 ดังตารางที่ 5.8 และ รูปที่ 5.10

ตารางที่ 5.3 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 1 และวิธีการที่ 2 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 1	วิธีที่ 2	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.061	0.333	+445.90
0.1	0.061	0.333	+445.90
0.2	0.500	0.766	+53.20
0.3	0.500	0.766	+53.20
0.4	0.500	0.766	+53.20
0.5	0.102	0.573	+461.76
0.6	0.161	0.555	+244.72
0.7	0.161	0.157	-2.48
0.8	0.571	0.087	-84.76
0.9	0.403	0.724	+79.65
1.0	0.247	0.323	+30.76
ค่าเฉลี่ย	0.297	0.489	+64.65

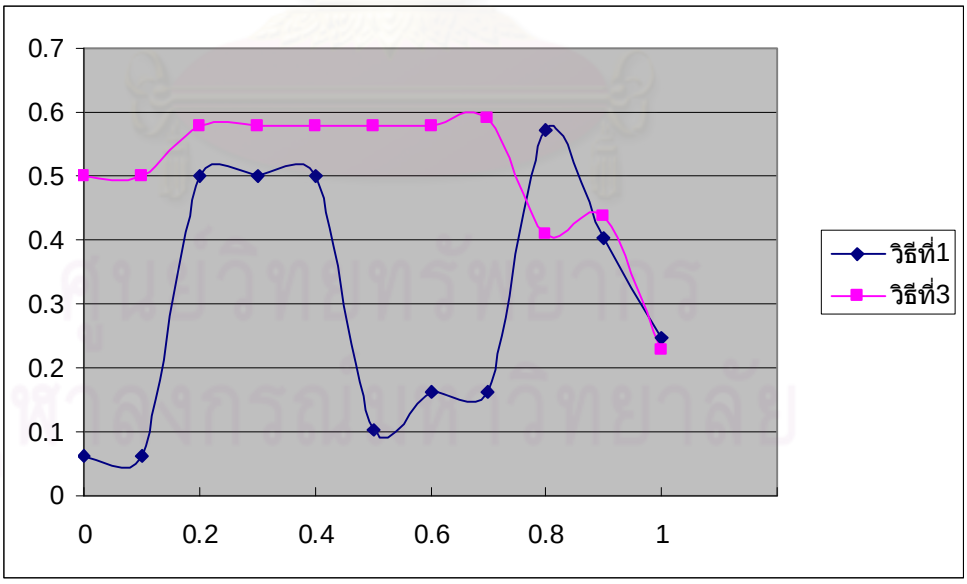


รูปที่ 5.5 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ (วิธีที่ 1) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 2)

จากตารางที่ 5.3 และกราฟในรูปที่ 5.5 พบว่าการค้นคืนวิธีที่ 2 มีค่าความแม่นยำโดยเฉลี่ยมากกว่าการค้นคืนวิธีที่ 1 ร้อยละ 64.65 ซึ่งจะเห็นได้ว่าค่าความแม่นยำจะมีค่าเพิ่มขึ้นทุกๆ ค่าเรียกคืน ยกเว้นค่าเรียกคืนที่ 0.7 และ 0.8 เท่านั้นที่มีค่าความแม่นยำลดลง นั่นแสดงว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

ตารางที่ 5.4 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 1 และวิธีการที่ 3 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 1	วิธีที่ 3	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.061	0.500	+719.67
0.1	0.061	0.500	+719.67
0.2	0.500	0.577	+15.40
0.3	0.500	0.577	+15.40
0.4	0.500	0.577	+15.40
0.5	0.102	0.577	+465.68
0.6	0.161	0.577	+258.38
0.7	0.161	0.591	+267.08
0.8	0.571	0.409	-28.37
0.9	0.403	0.436	+8.18
1.0	0.247	0.228	-7.69
ค่าเฉลี่ย	0.297	0.504	+69.69

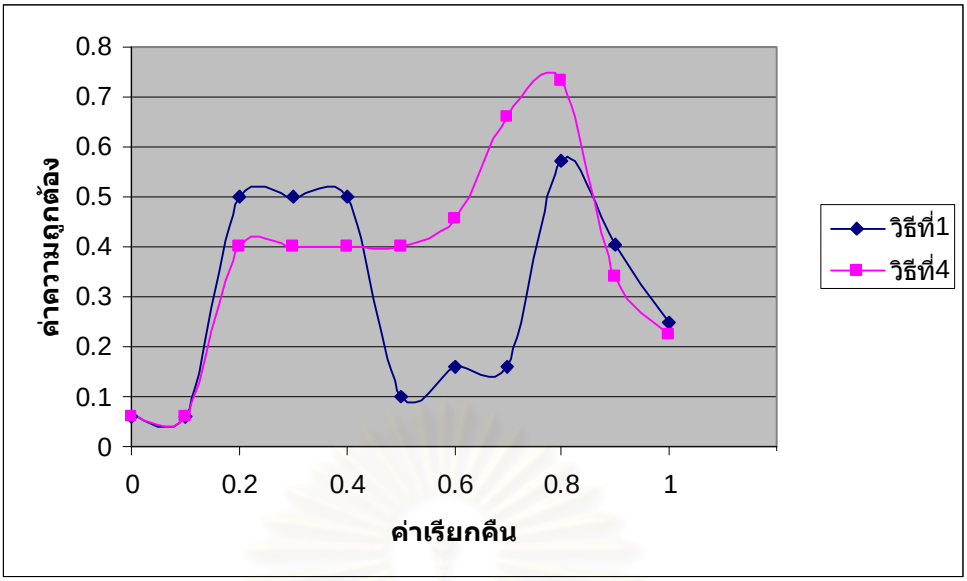


รูปที่ 5.6 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ (วิธีที่ 1) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 3)

จากตารางที่ 5.4 และกราฟในรูปที่ 5.6 จะเห็นได้อย่างชัดเจนว่าการค้นคืนด้วยวิธีการที่ 3 มีค่าเฉลี่ยของค่าความแม่นยำมากกว่าการค้นคืนด้วยวิธีการที่ 1 ในทุกๆ ค่าเรียกคืน ยกเว้นค่าเรียกคืนที่ 0.8 และ 1 เมื่อพิจารณาโดยรวมแล้ววิธีการที่ 3 มีค่าความแม่นยำเพิ่มขึ้นมากกว่าวิธีการที่ 1 ถึงร้อยละ 69.69 แสดงให้เห็นว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองปริภูมิเวคเตอร์ มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

ตารางที่ 5.5 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 1 และวิธีการที่ 4 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 1	วิธีที่ 4	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.061	0.061	0.00
0.1	0.061	0.061	0.00
0.2	0.500	0.400	-20.00
0.3	0.500	0.400	-20.00
0.4	0.500	0.400	-20.00
0.5	0.102	0.400	+292.15
0.6	0.161	0.455	+182.61
0.7	0.161	0.659	+309.32
0.8	0.571	0.733	+28.37
0.9	0.403	0.340	-15.63
1.0	0.247	0.223	-9.72
ค่าเฉลี่ย	0.297	0.376	+26.59

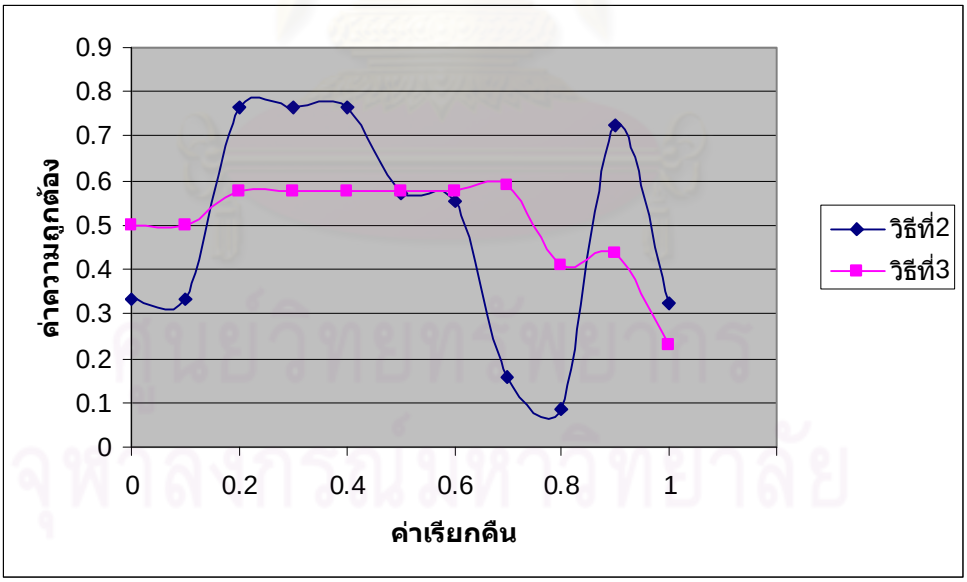


รูปที่ 5.7 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ (วิธีที่ 1) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น (วิธีที่ 4)

จากตารางที่ 5.5 และกราฟในรูปที่ 5.7 พบว่าการค้นคืนด้วยวิธีการที่ 3 และวิธีการค้นคืนที่ 4 มีค่าความแม่นยำเฉลี่ยใกล้เคียงกัน แต่จะแตกต่างในส่วนของค่าเรียกคืนที่ 0.5 ถึง 0.8 อย่างเห็นได้ชัดเจนว่า วิธีการ 4 นั้นมีค่าความแม่นยำเฉลี่ยมากกว่า วิธีการที่ 1 เมื่อพิจารณาโดยรวมแล้วค่าความแม่นยำของวิธีการที่ 4 มีค่ามากกว่า วิธีการที่ 1 ร้อยละ 26.59 แสดงว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองความน่าจะเป็นมีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

ตารางที่ 5.6 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 2 และวิธีการที่ 3 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 2	วิธีที่ 3	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.333	0.500	+50.15
0.1	0.333	0.500	+50.15
0.2	0.766	0.577	-24.67
0.3	0.766	0.577	-24.67
0.4	0.766	0.577	-24.67
0.5	0.573	0.577	+0.69
0.6	0.555	0.577	+3.96
0.7	0.157	0.591	+276.43
0.8	0.087	0.409	+370.11
0.9	0.724	0.436	-39.78
1.0	0.323	0.228	-29.41
ค่าเฉลี่ย	0.489	0.504	+3.07

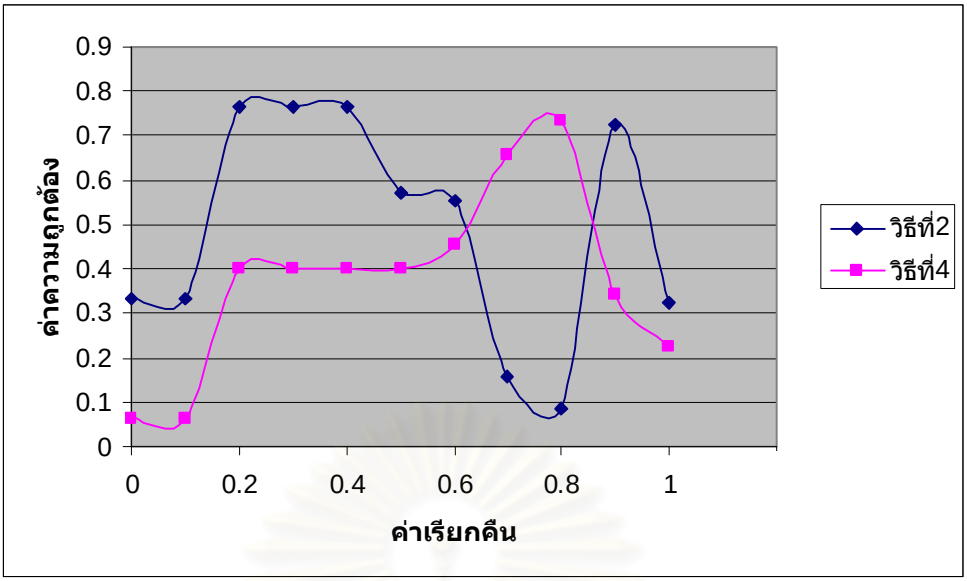


รูปที่ 5.8 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 2) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 3)

จากตารางที่ 5.6 และกราฟในรูปที่ 5.8 พบว่าวิธีการค้นคืนที่ 2 มีค่าเฉลี่ยค่าความแม่นยำมากกว่าวิธีการค้นคืนที่ 3 ในช่วง ค่าเรียกคืนที่ 0.2 ถึง 0.4 และ ช่วงที่ 0.9 ถึง 1.0 แต่มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าวิธีการที่ 3 ในช่วงค่าเรียกคืนที่ 0.0 ถึง 0.1 และ 0.6 ถึง 0.8 เมื่อพิจารณาโดยรวมแล้วค่าความแม่นยำของวิธีการที่ 3 มีค่ามากกว่าวิธีการที่ 2 ร้อยละ 3.07 แสดงให้เห็นว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองปริภูมิเวคเตอร์มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว

ตารางที่ 5.7 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 2 และวิธีการที่ 4 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 2	วิธีที่ 4	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.333	0.061	-81.68
0.1	0.333	0.061	-81.68
0.2	0.766	0.400	-47.78
0.3	0.766	0.400	-47.78
0.4	0.766	0.400	-47.78
0.5	0.573	0.400	-30.19
0.6	0.555	0.455	-18.02
0.7	0.157	0.659	+319.74
0.8	0.087	0.733	+742.53
0.9	0.724	0.340	-53.04
1.0	0.323	0.223	-30.96
ค่าเฉลี่ย	0.489	0.376	-23.11

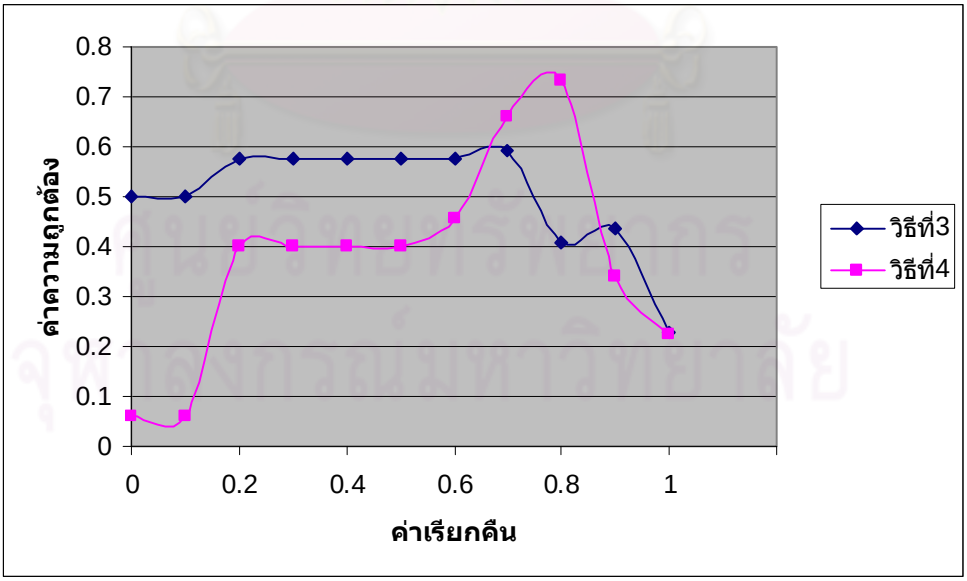


รูปที่ 5.9 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 2) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น (วิธีที่ 4)

จากตารางที่ 5.7 และกราฟในรูปที่ 5.9 จะเห็นได้ว่า วิธีการที่ 4 ให้ค่าความแม่นยำเพิ่มสูงมากในค่าเรียกคืนที่ 0.7 และ 0.8 ซึ่งมากกว่าวิธีการที่ 2 ถึง 3 และ 7 เท่า แต่จะขัดแย้งกับค่าเฉลี่ยค่าความแม่นยำโดยรวมของทั้ง 2 วิธีการ ซึ่งจะเห็นได้ว่า วิธีการที่ 2 ให้ค่าความแม่นยำมากกว่าวิธีการที่ 4 และเมื่อพิจารณาจากค่าความแม่นยำเฉลี่ยในทุกๆ ค่าเรียกคืนจะเห็นได้ชัดเจนว่า วิธีการที่ 4 ทำให้ค่าความแม่นยำเฉลี่ยลดลงร้อยละ 23.11 เมื่อเปรียบเทียบกับวิธีการค้นคืนที่ 2 ดังนั้นการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองความน่าจะเป็นมีผลให้การค้นคืนมีความถูกต้องเฉลี่ยลดลงมากกว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์

ตารางที่ 5.8 เปรียบเทียบค่าความแม่นยำเฉลี่ยระหว่างวิธีการที่ 3 และวิธีการที่ 4 ตามค่าเรียกคืนทั้ง 11 ค่าจาก 0 ถึง 1

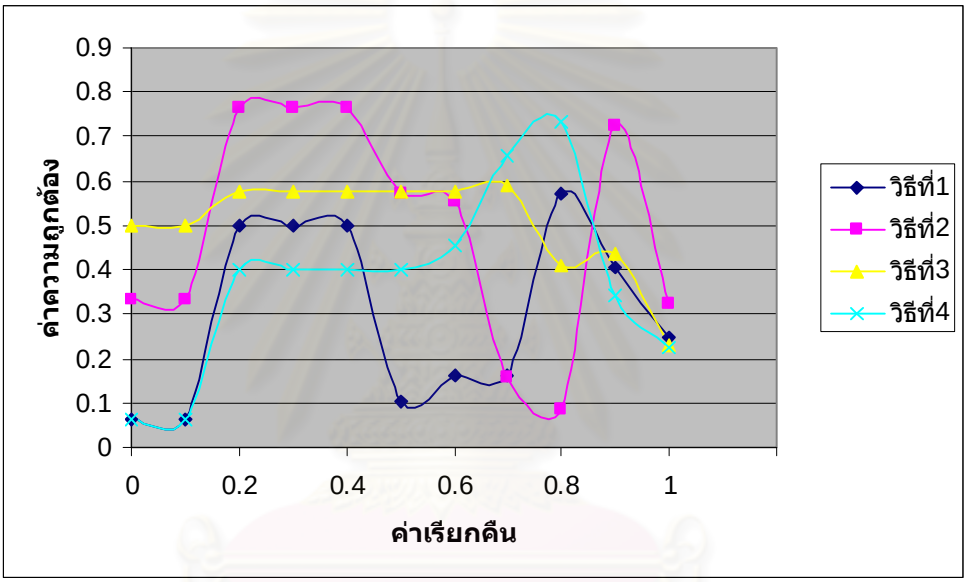
ค่าเรียกคืน	ค่าความแม่นยำ		
	วิธีที่ 3	วิธีที่ 4	ร้อยละ เพิ่มขึ้น(+) / ลดลง(-)
0.0	0.500	0.061	-87.80
0.1	0.500	0.061	-87.80
0.2	0.577	0.400	-30.67
0.3	0.577	0.400	-30.67
0.4	0.577	0.400	-30.67
0.5	0.577	0.400	-30.67
0.6	0.577	0.455	-21.14
0.7	0.591	0.659	+11.51
0.8	0.409	0.733	+79.22
0.9	0.436	0.340	-30.27
1.0	0.228	0.223	-2.19
ค่าเฉลี่ย	0.504	0.376	-25.39



รูปที่ 5.10 กราฟค่าเรียกคืนและค่าความแม่นยำ ระหว่างการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ (วิธีที่ 3) และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น (วิธีที่ 4)

และจากตารางที่ 5.8 และกราฟในรูปที่ 5.10 พบว่าวิธีการค้นคืนที่ 4 มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าวิธีการค้นคืนที่ 3 ในช่วงค่าเรียกคืนที่ 0.0 ถึง 0.6 และ 0.9 ถึง 1.0 ซึ่งแสดงให้เห็นว่าวิธีการที่ 4 มีค่าเฉลี่ยค่าความแม่นยำลดลง ร้อยละ 25.39 เมื่อเปรียบเทียบกับวิธีการที่ 3 และเมื่อพิจารณาโดยภาพรวมแล้วแสดงให้เห็นว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น มีผลให้การค้นคืนมีความถูกต้องลดลงมากกว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองปริภูมิเวคเตอร์

ผลการค้นคืนที่ได้ทั้งหมดของทุกวิธีการที่นำค่าความแม่นยำมาเฉลี่ยตามค่าเรียกคืนทั้ง 11 ค่า สามารถนำมาแสดงในรูปแบบกราฟได้ดังรูปที่ 5.11



รูปที่ 5.11 กราฟค่าเรียกคืนและค่าความแม่นยำของการค้นคืนทั้ง 4 วิธีการ

จากตารางที่ 5.2 และรูปที่ 5.11 ในการนำผลของค่าความแม่นยำที่ได้จากการค้นคืนเว็บเพจด้วยข้อคำถามจำนวน 50 ข้อคำถาม มาทำการเฉลี่ยเพื่อหาค่าความแม่นยำในระดับค่าเรียกคืนทั้งหมด 11 จุด คือ ค่าเรียกคืนที่ 0.0 ถึง 1.0 ของแต่ละวิธีการค้นคืน ซึ่งเป็นการพิจารณาผลการทดลองโดยไม่คำนึงถึงขนาดของข้อคำถามที่ใช้ในการทดลอง คือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ทั้งในแบบจำลองปริภูมิเวคเตอร์ และแบบจำลองความน่าจะเป็น รวมถึงการใช้และไม่ใช้วีไอพีเอสอัลกอริทึม สามารถทำให้ค่าความแม่นยำ มีค่าเพิ่มขึ้นมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ทุกวิธีการ เพราะในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ผู้ใช้มีส่วนร่วมในการให้ผลป้อนกลับจากรายการเอกสารที่ค้นคืนมาได้ครั้งแรก รวมทั้งมีการปรับเปลี่ยนคำและค่าน้ำหนักของคำในข้อคำถามใหม่ทำให้ผลการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ที่ได้มีค่าเพิ่มขึ้น และเมื่อเปรียบเทียบในส่วนของการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้เพียงอย่างเดียวพบว่า การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอ

พีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ให้ค่าความแม่นยำมากที่สุด รองลงมาคือ การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นตามลำดับ นั่นก็เพราะในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์จะทำการปรับเปลี่ยนคำและค่าน้ำหนักของคำในการกำหนดข้อคำถามใหม่ แต่ในแบบจำลองความน่าจะเป็นนั้น จะเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่เพียงอย่างเดียว จึงทำให้ผลการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ที่ได้ในการใช้วีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าความแม่นยำน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์

5.6 การวิเคราะห์ผลการทดลองโดยทางสถิติ

จากผลการทดลองทั้งหมดที่ได้นำเสนอไปข้างต้นนั้น ยังไม่สามารถที่จะสรุปได้ว่าวิธีการค้นคืนแบบใดที่เหมาะสมหรือมีประสิทธิผลในการค้นคืนเว็บเพจได้ดีที่สุด โดยงานวิจัยนี้จะวิเคราะห์จากจากค่าความแม่นยำเป็นหลัก เพราะเป็นมาตรวัดที่ใช้ในการคำนวณว่ารายการเอกสารที่ค้นคืนได้นั้นมีรายการเอกสารที่ตรงตามความต้องการของผู้ใช้เพียงใด

ดังนั้นจึงนำการทางวิเคราะห์สถิติ มาใช้ในการสนับสนุนความน่าเชื่อถือของงานวิจัยในอีกแง่มุมหนึ่งนอกเหนือจากผลการทดลอง และต้องมีการตั้งสมมุติฐาน การทดสอบสมมุติฐานโดยประการแรกผู้วิจัยจะต้องทำการทดสอบการแจกแจงของประชากรหรือข้อมูล จากตารางที่ 5.2 ว่ามีการแจกแจงแบบปกติหรือไม่ เพื่อเลือกใช้วิธีการทางสถิติในการทดสอบสมมุติฐานที่เหมาะสมได้

สำหรับงานวิจัยนี้ใช้การตรวจสอบการแจกแจงของประชากรด้วย Kolmogorov-Smirnov Test (K-S Test) ซึ่งเป็นเทคนิคที่ไม่ใช้พารามิเตอร์และขนาดตัวอย่างเล็กในการทดสอบการแจกแจงของประชากร ซึ่งสามารถทดสอบได้ว่าประชากรมีการแจกแจงแบบปกติหรือแบบอื่นๆ ได้ [12] ในการทดสอบว่าค่าความแม่นยำเฉลี่ยตามค่าเรียกคืน 11 ค่าของวิธีการที่ 1 2 3 และ 4 นั้นมีการแจกแจงแบบปกติหรือไม่ สามารถตั้งสมมุติฐานเพื่อทดสอบได้ดังนี้

$H_{0,1}$: ค่าความแม่นยำเฉลี่ยของวิธีการที่ 1 ตามค่าเรียกคืน11 ค่า มีการแจกแจงแบบปกติ

$H_{0,2}$: ค่าความแม่นยำเฉลี่ยของวิธีการที่ 2 ตามค่าเรียกคืน 11 ค่า มีการแจกแจงแบบปกติ

$H_{0,3}$: ค่าความแม่นยำเฉลี่ยของวิธีการที่ 3 ตามค่าเรียกคืน 11 ค่า มีการแจกแจงแบบปกติ

$H_{0,4}$: ค่าความแม่นยำเฉลี่ยของวิธีการที่ 4 ตามค่าเรียกคืน 11 ค่า มีการแจกแจงแบบปกติ

H_1 : ค่าความแม่นยำเฉลี่ยของทุกวิธีการ ตามค่าเรียกคืน 11 ค่า มีการแจกแจงแบบไม่ปกติ

ทำการทดสอบสมมุติฐานด้วยโปรแกรม SPSS โดยพิจารณาจากค่านัยสำคัญที่ 0.05 แสดงผลได้ดังตารางที่ 5.3

ตารางที่ 5.9 แสดงสถิติทดสอบการแจกแจงของประชากรที่ได้จาก 50 ข้อคำถามในแต่ละวิธีการค้นคืน

วิธีการค้นคืนที่	Kolmogorov-Smirnov Test (K-S Test)		H_0
	Kolmogorov-Smirnov Z	Sig.	
1	0.693	0.723	**ยอมรับ**
2	0.617	0.842	**ยอมรับ**
3	0.955	0.321	**ยอมรับ**
4	0.606	0.857	**ยอมรับ**

จากตารางที่ 5.3 แสดงให้เห็นว่า ค่า Sig มีค่ามากกว่าระดับนัยสำคัญที่กำหนดไว้ คือ 0.05 จึงไม่สามารถปฏิเสธสมมุติฐาน $H_{0,1}$ $H_{0,2}$ $H_{0,3}$ และ $H_{0,4}$ ได้ ดังนั้นค่าความแม่นยำเฉลี่ยตามค่าเรียกคืนที่แตกต่างกัน ตั้งแต่ 0 ถึง 1 ของวิธีการที่ 1 2 3 และ 4 มีการแจกแจงแบบปกติ จากนั้นผู้วิจัยใช้การทดสอบสมมุติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ย 2 ประชากรแบบจับคู่ (Paired t-test) ในการทดสอบสมมุติฐาน [13] ของการเปรียบเทียบค่าความแม่นยำเฉลี่ยตามค่าเรียกคืนทั้ง 11 ค่าของแต่ละวิธีการแบบจับคู่

สำหรับการตั้งสมมุติฐานในการเปรียบเทียบวิธีการค้นคืนเว็บเพจในแต่ละวิธีนั้น จะกำหนดระดับนัยสำคัญ a เท่ากับ 0.05 โดยในการทดสอบสมมุติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ย 2 ประชากรแบบจับคู่จะใช้กับประชากรที่มีการแจกแจงปกติ ในงานวิจัยนี้จะทดสอบสมมุติฐานแบบด้านเดียวคือ ด้านซ้ายลบ

ในการทดสอบแบบด้านเดียวเขตการปฏิเสธสมมุติฐาน H_0 คือ

1) Sig. (2-tailed) / 2 < a
2) t < 0 หรือ มีค่าลบ

ในการทดสอบสมมุติฐานนั้นจะปฏิเสธ H_0 ได้ก็ต่อเมื่อเงื่อนไขทั้ง 2 ข้อเป็นจริง ถ้าเงื่อนไขข้อใดข้อหนึ่งไม่เป็นจริงจะไม่สามารถปฏิเสธสมมุติฐาน H_0 ได้

กำหนดให้

m_1 คือ ค่าความแม่นยำเฉลี่ย จากการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

m_2 คือ ค่าความแม่นยำเฉลี่ย จากการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์

m_3 คือ ค่าความแม่นยำเฉลี่ย จากการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์

m_4 คือ ค่าความแม่นยำเฉลี่ย จากการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

ประชากรที่ใช้ในการทดสอบสมมุติฐาน คือ ค่าเฉลี่ยความแม่นยำตามค่าเรียกคืนทั้ง 11 ค่าของแต่ละวิธีการในตารางที่ 5.2 ซึ่งมีการทดสอบแล้วว่ามีการแจกแจงแบบปกติและใช้การทดสอบสมมุติฐานแบบด้านเดียวโดยมีการตั้งสมมุติฐานดังต่อไปนี้

5.6.1 การค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ สมมุติฐานคือ

$$\begin{aligned} H_0 &: \mu_1 \geq \mu_2 \\ H_1 &: \mu_1 < \mu_2 \end{aligned} \quad \ldots\ldots (1)$$

5.6.2 การค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ สมมุติฐานคือ

$$\begin{aligned} H_0 &: \mu_1 \geq \mu_3 \\ H_1 &: \mu_1 < \mu_3 \end{aligned} \quad \ldots\ldots (2)$$

5.6.3 การค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น สมมุติฐานคือ

$$\begin{aligned} H_0 &: \mu_1 \geq \mu_4 \\ H_1 &: \mu_1 < \mu_4 \end{aligned} \quad \ldots\ldots (3)$$

5.6.4 การค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ สมมุติฐานคือ

$$\begin{aligned} H_0 &: \mu_2 \geq \mu_3 \\ H_1 &: \mu_2 < \mu_3 \end{aligned} \quad \ldots\ldots (4)$$

5.6.5 การค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น สมมุติฐานคือ

$$H_0 : \mu_2 \geq \mu_4$$
$$H_1 : \mu_2 < \mu_4 \quad \ldots\ldots (5)$$

5.6.6 การค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น สมมุติฐานคือ

$$H_0 : \mu_3 \geq \mu_4$$
$$H_1 : \mu_3 < \mu_4 \quad \ldots\ldots (6)$$

จากนั้นใช้โปรแกรม SPSS ในการทดสอบสมมุติฐานทั้ง 6 ข้อ ด้วยการทดสอบสมมุติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ย 2 ประชากรแบบจับคู่ ผลการทดสอบสมมุติฐาน ได้ค่าสถิติดังตารางที่ 5.8

ตารางที่ 5.10 แสดงค่าสถิติทดสอบสมมุติฐานเกี่ยวกับผลต่างระหว่างค่าเฉลี่ย 2 ประชากรแบบจับคู่ ของการทดสอบสมมุติฐานที่ 1 ถึง 6

การทดสอบสมมุติฐานที่	t	Sig. (2-tailed)/2	H_0
1) วิธีค้นคืนที่ 1 เปรียบเทียบกับวิธีค้นคืนที่ 2	-2.459	0.017	**ปฏิเสธ**
2) วิธีค้นคืนที่ 1 เปรียบเทียบกับวิธีค้นคืนที่ 3	-2.958	0.007	**ปฏิเสธ**
3) วิธีค้นคืนที่ 1 เปรียบเทียบกับวิธีค้นคืนที่ 4	-1.932	0.041	**ปฏิเสธ**
4) วิธีค้นคืนที่ 2 เปรียบเทียบกับวิธีค้นคืนที่ 3	-1.905	0.043	**ปฏิเสธ**
5) วิธีค้นคืนที่ 2 เปรียบเทียบกับวิธีค้นคืนที่ 4	1.057	0.158	**ยอมรับ**
6) วิธีค้นคืนที่ 3 เปรียบเทียบกับวิธีค้นคืนที่ 4	1.988	0.037	**ยอมรับ**

5.7 สรุปผลการทดลอง

ผลที่ได้จากการวิเคราะห์ผลการทดลองซึ่งเป็นการทดสอบสมมุติฐานต่าง ๆ ในหัวข้อที่ 5.6 นำมาสรุปได้ดังต่อไปนี้

5.7.1 การทดสอบสมมุติฐานที่ 1 ผลจากค่าสถิติทดสอบ ทำให้ปฏิเสธสมมุติฐาน H_0 (ยอมรับ H_1) แสดงว่าการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ ในอีกความหมายหนึ่งคือการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ ทำให้มีค่าความแม่นยำเพิ่มขึ้น

5.7.2 การทดสอบสมมุติฐานที่ 2 ผลจากค่าสถิติทดสอบ ทำให้ปฏิเสธสมมุติฐาน H_0 (ยอมรับ H_1) แสดงว่าการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ในอีกความหมายหนึ่งคือการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ทำให้มีค่าความแม่นยำเพิ่มมากขึ้น

5.7.3 การทดสอบสมมุติฐานที่ 3 ผลจากค่าสถิติทดสอบ ทำให้ปฏิเสธสมมุติฐาน H_0 (ยอมรับ H_1) แสดงว่าการค้นคืนเว็บเพจโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในอีกความหมายหนึ่งคือการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ทำให้มีค่าความแม่นยำเพิ่มมากขึ้น

5.7.4 การทดสอบสมมุติฐานที่ 4 ผลจากค่าสถิติทดสอบ ทำให้ปฏิเสธสมมุติฐาน H_0 (ยอมรับ H_1) แสดงว่าการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าเฉลี่ยค่าความแม่นยำน้อยกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ในอีกความหมายหนึ่งคือการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ทำให้มีค่าความแม่นยำมากกว่าการไม่ใช้วีไอพีเอสอัลกอริทึม

5.7.5 การทดสอบสมมุติฐานที่ 5 ผลจากค่าสถิติทดสอบ ทำให้ยอมรับสมมุติฐาน H_0 แสดงว่าการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในอีกความหมายหนึ่งคือการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีผลทำให้ค่าเฉลี่ยค่าความแม่นยำมีค่าน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว

5.7.6 การทดสอบสมมุติฐานที่ 6 ผลจากค่าสถิติทดสอบ ทำให้ยอมรับสมมุติฐาน H_0 แสดงว่าการค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ มีค่าเฉลี่ยค่าความแม่นยำมากกว่าหรือเท่ากับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในอีกความหมายหนึ่งคือการใช้แบบจำลองความน่าจะเป็นมีผลให้ค่าเฉลี่ยค่าความแม่นยำมีค่าน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์ ในกรณีที่มีการใช้วีไอพีเอสอัลกอริทึมร่วมด้วย

5.8 ข้ออภิปราย

จากการทดลอง ผลการทดลอง การวิเคราะห์ผลการทดลองด้วยหลักสถิติและสรุปผลทดลอง ในหัวข้อดังกล่าวข้างต้น สามารถนำมาอภิปรายได้ดังต่อไปนี้

1) จากตารางที่ 5.1 เป็นการทดลองเพื่อให้เห็นถึงขนาดของข้อคำถามว่ามีผลต่อค่าความแม่นยำในการค้นคืนหรือไม่ ซึ่งพบว่าในการค้นคืนเว็บเพจนั้น ต้องคำนึงถึงขนาดของข้อคำถามที่ใช้ในการค้นคืน คือ ข้อคำถามนั้นจะต้องประกอบด้วยคำที่ไม่มากจนเกินไป เพราะจะเห็นได้ว่าขนาดของข้อคำถามที่ประกอบด้วยคำเพิ่มขึ้น มีผลให้ค่าเฉลี่ยค่าความแม่นยำที่ได้จากการค้นคืนมีค่าลดลงในทุกกรณีที่ใช้ทดสอบ ทั้งในการค้นคืนโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และการค้นคืนที่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ แต่ขนาดของข้อคำถามนั้นที่ใช้ในการค้นคืน ไม่มีผลต่อค่าเรียกคืน คือค่าเรียกคืนไม่ได้มีค่าลดลงแต่มีค่าเพิ่มขึ้นเพียงเล็กน้อย

2) จากตารางที่ 5.2 และรูปที่ 5.5 5.6 และ 5.7 พบว่าเมื่อนำค่าความแม่นยำมาเฉลี่ยตามค่าเรียกคืน 11 ค่า จาก 0 ถึง 1 แยกตามวิธีการค้นคืนทั้ง 4 วิธีการ โดยไม่พิจารณาถึงขนาดของข้อคำถาม จะเห็นได้ว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ทั้ง 3 วิธีการ คือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น มีค่าเฉลี่ยค่าความแม่นยำมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ซึ่งสอดคล้องกับการทดสอบสมมุติฐานที่ 1 2 และ 3 ที่แสดงให้เห็นว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ให้ค่าความแม่นยำมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

3) จากตารางที่ 5.2 และรูปที่ 5.11 พบว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ให้ค่าความแม่นยำเฉลี่ยมากที่สุด รองลงมาคือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นตามลำดับ ซึ่งถ้าไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้นจะให้ค่าความแม่นยำเฉลี่ยน้อยที่สุด

4) จากตารางที่ 5.6 และรูปที่ 5.8 พบว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองปริภูมิเวคเตอร์มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว ซึ่งสอดคล้องกับการทดสอบสมมุติฐานที่ 4 คือ การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ทำให้มีค่าความแม่นยำมากกว่าการไม่ใช้วีไอพีเอสอัลกอริทึม เนื่องจากการใช้วีไอพีเอสอัลกอริทึมผู้ใช้สามารถเลือกบล็อกที่ตรงประเด็นกับที่ต้องการทำให้คำที่ใช้ในการกำหนดข้อคำถามใหม่นั้นเหมาะสมมากกว่าการใช้การเลือกคำที่มีค่าความถี่สูงสุดจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ที่ใช้แบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว

5) จากตารางที่ 5.7 และรูปที่ 5.9 พบว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นซึ่งสอดคล้องกับการทดสอบสมมุติฐานที่ 5 คือ การค้นคืนเว็บเพจโดยใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีผลทำให้ค่าเฉลี่ยค่าความแม่นยำมีค่าน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว เนื่องจากในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นผู้ใช้มีการให้ผลป้อนกลับด้วยการเลือกรายการเอกสารและบล็อกที่เห็นว่าตรงประเด็นกับที่ต้องการโดยนำคำในบล็อกเหล่านั้นไปช่วยในการกำหนดข้อคำถามใหม่ ซึ่งทำให้ขนาดของข้อคำถามใหม่มีคำเพิ่มมากขึ้นและมีการเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลองความน่าจะเป็น ส่งผลให้เอกสารที่ค้นคืนมาได้มีเอกสารที่ไม่ตรงประเด็นจำนวนเพิ่มขึ้น ค่าความแม่นยำจึงมีค่าน้อยกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ แต่ในขณะที่การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียวนั้นคำที่นำเข้ามาเพิ่มในการกำหนดข้อคำถามใหม่ มีเพียง 1 คำ คือคำที่มีค่าความถี่สูงสุดจากเว็บเพจที่ผู้ใช้เลือกแล้วว่าตรงประเด็นกับที่ต้องการ และมีการขยายคำและค่าน้ำหนักของคำด้วยแบบจำลองปริภูมิเวคเตอร์ ส่งผลให้เอกสารที่ค้นคืนมาได้มีรายการเอกสารที่ไม่ตรงประเด็นน้อยลงทำให้ค่าความแม่นยำที่ได้จากการค้นคืนมีค่ามากกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

6) จากตารางที่ 5.8 และรูปที่ 5.10 พบว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ มีผลให้การค้นคืนมีความถูกต้องเพิ่มขึ้นมากกว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ซึ่งสอดคล้องกับสมมุติฐานที่ 6 คือ การใช้แบบจำลองความน่าจะเป็นมีผลให้ค่าเฉลี่ยค่าความแม่นยำมีค่าน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์ ในกรณีที่มีการใช้วีไอพีเอสอัลกอริทึมร่วมด้วย เนื่องจากว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์นั้น มีการปรับเปลี่ยนคำและค่าน้ำหนักของ

คำที่ใช้ในการกำหนดข้อคำถามใหม่ แต่การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นนั้นมีเพียงการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่เพียงอย่างเดียว จึงทำให้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมที่มีการใช้แบบจำลองปริภูมิเวคเตอร์ มีค่าความแม่นยำมากกว่าการใช้แบบจำลองความน่าจะเป็น

7) จากการทดลองและการทดสอบสมมุติฐานทั้งหมดพบว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ ให้ค่าความแม่นยำในการค้นคืนได้ดีที่สุดเมื่อเปรียบเทียบกับ การไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ และการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ที่ระดับนัยสำคัญ 0.05

บทที่ 6
สรุปผลงานวิจัย

ในบทนี้กล่าวถึงส่วนสุดท้ายที่ได้จากผลงานวิจัยนั่นคือ บทสรุปของผลงานวิจัย รวมทั้งงานวิจัยในอนาคต ซึ่งมีรายละเอียดดังต่อไปนี้

6.1 สรุปผลงานวิจัย

งานวิจัยนี้นำเสนอวิธีการค้นคืนเว็บเพจโดยใช้วีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ร่วมกับการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ ซึ่งใช้และไม่ใช้วีไอพีเอสอัลกอริทึมด้วย รวมถึงการค้นคืนโดยไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ซึ่งเว็บเพจที่นำมาจัดเก็บทั้งหมด มี 300 เว็บเพจ รวบรวมมาจากเว็บไซต์ข่าวกีฬาจำนวน 8 เว็บไซต์ เป็นเวลาติดต่อกัน 5 วัน เพื่อให้ได้เนื้อหาของเว็บเพจ ในแนวทางเดียวกัน หรือใกล้เคียงกัน และเป็นภาษาอังกฤษทั้งหมด โดยนำหลักการพื้นฐานจากทฤษฎีการจัดเก็บและค้นคืนสารสนเทศ อันได้แก่ การทำดรรชนีอัตโนมัติ การหาน้ำหนักของคำ การใช้สูตรคำนวณหาค่าความคล้าย การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ แบบจำลองความน่าจะเป็น แบบจำลองปริภูมิเวคเตอร์ เป็นต้น มาประยุกต์ใช้

โดยในงานวิจัยนี้จึงประกอบด้วย 5 ขั้นตอนที่สำคัญ คือ

1) การจัดเก็บเว็บเพจ

สำหรับการจัดเก็บเว็บเพจ จะดำเนินการ 2 ขั้นตอนคือ การแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม และการทำดรรชนีอัตโนมัติ โดยทำการแบ่งเว็บเพจที่จัดเก็บมาทั้งหมดด้วยวีไอพีเอสอัลกอริทึมออกเป็นเป็นบล็อกก่อนที่จะเข้าสู่ขั้นตอนในการทำดรรชนีอัตโนมัติ เพื่อรองรับขั้นตอนการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ที่จะต้องเลือกบล็อกของรายการเอกสารที่ค้นคืนมาได้ในครั้งแรกที่ตรงประเด็นกับที่ต้องการ แล้วจึงเข้าสู่ขั้นตอนการทำดรรชนีอัตโนมัติ ซึ่งใช้ค่าน้ำหนักแบบค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน

2) การค้นคืนเว็บเพจ

ในการค้นคืนเว็บเพจนั้นจะพิจารณาจากข้อคำถามเพียงอย่างเดียว โดยที่จะมีขนาดของข้อคำถามที่ประกอบด้วยคำตั้งแต่ 1 คำ ถึง 5 คำ เพื่อให้เกิดความหลากหลาย และโดยส่วนใหญ่จะใช้คำในการค้นคืนไม่มากนัก นอกจากนี้ยังมีจุดประสงค์เพื่อทดสอบขนาดของคำว่ามีผลต่อประสิทธิผลในการค้นคืนหรือไม่ ซึ่งจะใช้ข้อคำถามทั้งหมด 50 ข้อคำถามในการค้นคืนเว็บเพจ

3) การค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

สำหรับการค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น จะแบ่งออกเป็น 3 วิธีการ คือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิ

เวคเตอร์ และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น โดยในแบบจำลองปริภูมิเวคเตอร์จะเปลี่ยนแปลงคำในข้อคำถามเดิมด้วยการเพิ่มคำที่มีค่าความถี่สูงสุดจากรายการเอกสารที่ผู้ใช้เลือกว่าตรงประเด็น แล้วเปลี่ยนแปลงค่าน้ำหนักของคำได้เป็นข้อคำถามใหม่ ส่วนในวีไอพีเอสอัลกอริทึมของแบบจำลองทั้ง 2 แบบ จะใช้คำที่ได้มาจากบล็อกที่ผู้ใช้เลือกแล้วว่าตรงประเด็น มาช่วยในการกำหนดข้อคำถามใหม่ ซึ่งจะแตกต่างในส่วนของการเปลี่ยนแปลงค่าน้ำหนักของคำตามแบบจำลองแต่ละแบบ โดยในแบบจำลองปริภูมิเวคเตอร์จะมีการปรับเปลี่ยนคำและค่าน้ำหนักของคำ ส่วนในแบบจำลองปริภูมิเวคเตอร์จะเปลี่ยนแปลงค่าน้ำหนักของคำเพียงอย่างเดียว

4) การออกแบบและพัฒนาเครื่องมือ

หลังจากได้แนวคิดในงานวิจัยเรียบร้อยแล้ว จึงทำการออกแบบและพัฒนาเครื่องมือที่เหมาะสมเพื่อสนับสนุนแนวคิดดังกล่าว โดยเครื่องมือที่พัฒนาขึ้น แบ่งออกเป็น 2 ส่วน ดังนี้

(1) ส่วนการจัดเก็บเว็บเพจ หลังจากแบ่งเว็บเพจเป็นส่วนย่อยด้วย วีไอพีเอสอัลกอริทึมแล้วแล้วจึงดำเนินการกับข้อความ หาค่าดรรชนีและค่าน้ำหนักของคำในแต่ละเอกสาร แล้วจัดเก็บลงสู่ฐานข้อมูล

(2) ส่วนการค้นคืนเว็บเพจ ข้อคำถามที่ผู้ใช้ป้อนเข้าสู่ระบบ จะถูกนำมาแปลงเป็นชุดดรรชนีของข้อคำถามก่อนแล้วจึงทำการเปรียบเทียบความคล้ายกับดรรชนีชุดเอกสารในฐานข้อมูล หากข้อคำถามดังกล่าวมีค่าความคล้ายกับเอกสารพอเพียงกับค่าขีดแบ่งเริ่มต้นความคล้ายที่ได้กำหนดไว้ ระบบจึงแสดงผลกลับสู่ผู้ใช้งาน โดยในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้นั้น ผู้ใช้จะมีส่วนร่วมในการเลือกรายการเอกสารที่เห็นว่าตรงประเด็นมาใช้ในการกำหนดข้อคำถามใหม่ในแบบจำลองปริภูมิเวคเตอร์ แต่ถ้ามีการใช้วีไอพีเอสอัลกอริทึม ผู้ใช้จะเลือกบล็อกที่เห็นว่าตรงประเด็นเพื่อนำคำเหล่านั้นไปช่วยกำหนดข้อคำถามใหม่ ร่วมกับการขยายคำและค่าน้ำหนักของคำในข้อคำถามเดิมด้วยแบบจำลองปริภูมิเวคเตอร์ ส่วนในแบบจำลองความน่าจะเป็นจะมีเพียงการเปลี่ยนแปลงค่าน้ำหนักของคำในบล็อกที่ผู้ใช้เลือกแล้วเท่านั้น

5) การทดลองและประเมินผลการทดลอง

เมื่อพัฒนาเครื่องมือเรียบร้อยแล้ว ผู้วิจัยได้ออกแบบการทดลองและทำการทดลองเพื่อประเมินประสิทธิผลความถูกต้องของวิธีการที่นำเสนอ ซึ่งเครื่องมือดังกล่าวถูกออกแบบให้สามารถทำการค้นคืนเว็บเพจ และค้นคืนเว็บเพจจากการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ทั้งที่ใช้และไม่ใช้วีไอพีเอสอัลกอริทึมในการเลือกบล็อก และใช้แบบจำลองปริภูมิเวคเตอร์ หรือแบบจำลองความน่าจะเป็นร่วมด้วย การทดลองจะทำการเปรียบเทียบประสิทธิผลระหว่างการค้นคืนเว็บเพจโดยใช้และไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และเปรียบเทียบการใช้และไม่ใช้วีไอพีเอสอัลกอริทึมในแบบจำลองปริภูมิเวคเตอร์ และการใช้วีไอพีเอสอัลกอริทึมในแบบจำลองทั้ง 2 แบบ ใช้ค่าเรียกคืน ค่าความแม่นยำ เป็นมาตรวัดในการประเมินประสิทธิผล

ของการค้นคืนเว็บเพจ จากข้อคำถามทั้งสิ้น 50 ข้อคำถาม ในเว็บเพจ จำนวน 300 เว็บเพจ มีขนาดของข้อคำถามตั้งแต่ 1 คำ ถึง 5 คำ อย่างละ 10 ข้อที่มีความหลากหลายต่างกันไป

6) สรุปผลการทดลอง

ผลลัพธ์ที่ได้จากการทดลอง แบ่งออกเป็น 2 กรณีด้วยกัน กรณีแรกคือ พิจารณาจากขนาดของข้อคำถาม เพื่อทดสอบว่า ขนาดของข้อคำถามที่ประกอบด้วยคำตั้งแต่ 1 คำ ถึง 5 คำ ว่าจะมีผลต่อประสิทธิผลในการค้นคืนหรือไม่ ซึ่งจากการทดลองพบว่า มีผลต่อประสิทธิผลต่อการค้นคืนในค่าความแม่นยำ เพราะจากการทดสอบที่มีการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ และไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ พบว่าเมื่อคำที่ใช้เป็นข้อคำถามมีคำเพิ่มมากขึ้น ค่าความแม่นยำที่ได้จากการค้นคืนและการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้มีค่าลดลง แต่ไม่มีผลต่อค่าเรียกคืน

อีกกรณีหนึ่ง คือ พิจารณาจากค่าเรียกคืน และค่าความแม่นยำ โดยไม่คำนึงถึงขนาดของข้อคำถามที่แตกต่างกัน ซึ่งนำค่าความแม่นยำจากทุกวิธีการค้นคืนมาเฉลี่ยตามจุดของค่าเรียกคืนทั้ง 11 จุด ดังตารางที่ 5.2 และ กราฟรูปที่ 5.11 พบว่าการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์มีค่าความแม่นยำโดยเฉลี่ยเป็น 0.504 ซึ่งมีค่ามากที่สุด รองลงมาคือ การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์มีค่าความแม่นยำโดยเฉลี่ยเป็น 0.489 และการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าความแม่นยำ โดยเฉลี่ยเป็น 0.376 ซึ่งมีค่าความแม่นยำเฉลี่ยมากกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ซึ่งมีค่าความแม่นยำเฉลี่ยเพียง 0.297

สำหรับการพิจารณาจากค่าร้อยละที่เพิ่มขึ้นหรือลดลงของค่าความแม่นยำเฉลี่ยตามค่าเรียกคืนในแต่ละคู่วิธีการที่ทำการเปรียบเทียบ พบว่า

(1) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ มีค่าความแม่นยำเฉลี่ยเพิ่มขึ้นร้อยละ 64.65 เมื่อเปรียบเทียบกับการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

(2) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ มีค่าความแม่นยำเฉลี่ยเพิ่มขึ้นร้อยละ 69.69 เมื่อเปรียบเทียบกับการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

(3) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าความแม่นยำเฉลี่ยเพิ่มขึ้นร้อยละ 26.59 เมื่อเปรียบเทียบกับการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

(4) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์ มีค่าความแม่นยำเฉลี่ยเพิ่มขึ้นร้อยละ 3.07 เมื่อเปรียบเทียบกับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว

(5) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าความแม่นยำเฉลี่ยลดลงร้อยละ 23.11 เมื่อเปรียบเทียบกับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์

(6) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นมีค่าความแม่นยำเฉลี่ยลดลงร้อยละ 25.39 เมื่อเปรียบเทียบกับการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์

จากนั้นได้ใช้หลักการทางสถิติในการวิเคราะห์ผลการทดลอง โดยได้ตั้งสมมุติฐานต่างๆ และทำการทดสอบสมมุติฐานพบว่า

(1) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้สามารถทำให้ระบบการค้นคืนมีค่าความแม่นยำมากขึ้นกว่าการไม่ใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

(2) การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสและแบบจำลองปริภูมิเวคเตอร์มีผลให้การค้นคืนมีความแม่นยำเพิ่มขึ้นมากกว่าการใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์เพียงอย่างเดียว

(3) การใช้การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยแบบจำลองปริภูมิเวคเตอร์ มีผลให้การค้นคืนมีความแม่นยำเพิ่มขึ้นมากกว่า การให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

(4) การใช้แบบจำลองความน่าจะเป็นมีผลให้ค่าเฉลี่ยค่าความแม่นยำมีค่าน้อยกว่าการใช้แบบจำลองปริภูมิเวคเตอร์ ในกรณีที่มีการใช้วีไอพีเอสอัลกอริทึมร่วมด้วย

จากผลการทดลองและการทดสอบสมมุติฐานต่างๆ สรุปได้ว่า ในการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์จะให้ประสิทธิผลในการค้นคืนดีที่สุดในเรื่องของค่าความแม่นยำ

6.2 งานวิจัยในอนาคต

1) การจัดเก็บและค้นคืนเอกสารก็ยังคงเป็นงานวิจัยหนึ่งที่มีผลงานการวิจัยออกมาเผยแพร่อย่างสม่ำเสมอ เนื่องจากเอกสารในยุคสมัยนี้มีรูปแบบหลากหลาย และต้องการการจัดเก็บและค้นคืนที่ให้ทั้งประสิทธิภาพและประสิทธิผล ซึ่งสามารถนำมาประยุกต์ใช้ได้ในเอกสารหลากหลาย เช่น เอกสารแบบยูสเคส

2) สำหรับการใช้วีไอพีเอสอัลกอริทึม ในการเลือกบล็อกของเว็บเพจที่ผู้ใช้เห็นว่าตรงตามต้องการ อาจจะทำการปรับให้ระบบเลือกคำที่มีค่าความถี่สูงสุดจากบล็อกเหล่านั้น แทนการใช้คำที่ปรากฏในบล็อกทั้งหมดมากำหนดข้อคำถามใหม่ เพื่อช่วยลดขนาดของคำที่มากำหนดข้อคำถามใหม่ ให้มีขนาดลดลง เพราะขนาดของคำที่ใช้ในการค้นคืนมีผลต่อค่าความแม่นยำ

รายการอ้างอิง

[1] G. Salton, C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, (1990): 288-297.

[2] Ji-Rong Wen, Shipeng Yu, Deng Cai and Wei-Ying Ma VIPS: a vision-based page segmentation algorithm, Microsoft Technical Report, MSR-TR-2003-79, 2003.

[3] Ji-Rong Wen, Shipeng Yu, Deng Cai and Wei-Ying Ma , Improving pseudo-relevance feedback in web information retrieval using web page segmentation. Proceedings of the 12th International World Wide Web Conference, WWW2003. (2003): 11–18.

[4] D. Harman. Relevance feedback revisited. 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval., 1992.

[5] Recardo, B. Yates and Berthier, R. Neto. Retrieval Evaluation. Modern Information Retrieval, Addison-Wesley, (1999): 73-82.

[6] Porter, M. F. An algorithm for suffix stripping: Program. Automated library and information system 14 (1980): 130-137.

[7] Salton, G. and Buckley, C. Introduction to Modern Information Retrieval, McGraw Hill, (1983): 141-143

[8] J. Chen, B. Zhou, J. Shi, H. Zhang, and F. Qiu. Function-Based Object Model Towards Website Adaptation. Proceedings of the 10th World Wide Web conference (WWW10)., May 2001.

[9] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic. Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification. Proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02)., 2002.

[10] Sparck Jones Search Term Relevance Weighting Given Little Relevance Information. Journal of Documentation 35 (1979): 30-48.

[11] Seung Yeol and Achim Hoffman . Pseudo-Relevance Feedback in Web information Retrieval Using Segments'Subjective Importance Value Proceedings of International Conference on Web Intelligence., 2005.

[12] กัลยา วานิชย์บัญชา. การวิเคราะห์สถิติ: สถิติสำหรับการบริหารและวิจัย. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2546.

13. กัลยา วานิชย์บัญชา. การใช้ SPSS for Windows ในการวิเคราะห์ข้อมูล. กรุงเทพมหานคร: โรงพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2549.

ภาคผนวก

ภาคผนวก ก
เอกสารเว็บเพจทั้งหมดที่จัดเก็บ

ในภาคผนวก ก จะอธิบายถึงรายละเอียดของเอกสารเว็บเพจทั้งหมดที่นำมาใช้ในงานวิจัย ซึ่งประกอบด้วยเว็บเพจทั้งหมด 300 เว็บเพจ และเป็นข่าวกีฬาภาษาอังกฤษทั้งหมด แสดงรายชื่อเว็บไซต์ได้ดังตารางที่ ก.1 และรายชื่อเว็บเพจได้ดังตารางที่ ก.2

ตารางที่ ก.1 รายชื่อเว็บไซต์ที่นำเว็บเพจมาใช้กับระบบ

ลำดับที่	ชื่อเว็บไซต์	จำนวนเว็บเพจ	จำนวนบล็อก
1	http://news.bbc.co.uk/sport	33	8
2	http://edition.cnn.com/	40	7
3	http://msn.foxsports.com/	52	8
4	http://cbs.sportsline.com/	35	8
5	http://www.espnstar.com/	36	4
6	http://eurosport.yahoo.com/	35	6
7	http://www.skysports.com/	43	5
8	http://www.usatoday.com/	26	7

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด

รหัสเว็บเพจ	ชื่อเว็บเพจ
1	Acasuso in Iraq jibe at Hewitt
2	Clarke relishing Ryder Cup return
3	No grudge with Monty Olazabal
4	Kanu rejected Europe for Redknapp
5	San Marino race gets F1 lifeline
6	Russia feels the heat against U.S. in Davis Cup semis
7	Agassi 2006 win-loss
8	Davis Cup: Argentina faces visiting Australia in semifinals
9	Russia feeling pressure against U.S. in Davis Cup semis
10	Henin-Hardenne to miss 3 weeks because of knee injury
11	Henin-Hardenne temporarily sidelined with knee injury
12	Knee injury to sideline Henin-Hardenne, perhaps until November

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด (ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
13	Record $5m F1 fine hits Turkish GP
14	Henin-Hardenne faces 3-week rest
15	Mets first to claim play-off berth
16	Gregan puts World Cup before tour
17	Jankovic returns with emphatic win
18	Woods partners Furyk at Ryder Cup
19	Relaxed Lehman knows his pairings
20	Pound puts cycling top of hit list
21	Hingis hurries to win in India
22	Ryder captains agree to fight hard
23	Schalke lift suspension on Asamoah
24	Grieving Clarke says he is ready for Ryder Cup
25	Wenger fires title warning after Arsenal win
26	Parnevik warns US rookies could see Boston-like Ryder Cup shock
27	Hingis makes sparkling debut in India
28	Robson leaves West Brom
29	Jankovic powers into Beijing second round
30	Hingis works quickly to advance in Calcutta
31	Fergie calms Giggs injury fears
32	Chelsea rule out Ballack appeal
33	Drogba beaten boo-boys - Cech
34	Adebayor draws confidence from goal
35	Veterans must lead this team to win
36	Italy wins Fed Cup title, beats Belgium 3-2
37	Beckham: England, EPL not in my future
38	Rio: I feared for United future
39	Baghdatis wins China Open for first title
40	Wenger: Players had WC hangover
41	Becks to prove Macca wrong for England snub
42	Jose wants more from Sheva
43	Acasuso Blasts Hewitt Circus

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
44	Woosnam wants frank assessment
45	Ferdinand feared United exit
46	Rafa refusing to panic
47	Henin-Hardenne Sidelined
48	K Club built to strike fear
49	Roundup: Big names tumble in Calcutta; Li rolls at China Open
50	Davenport struggles to China win
51	Lerner completes takeover of English soccer club Aston Villa
52	Southgate takes blame for defeat
53	Wycombe delight at Fulham scalp
54	Woods angry at fake nude photos
55	Manager Blackwell sacked by Leeds
56	Peng, Davenport win at China Open
57	Hingis to face Tanasugarn in Calcutta quarterfinals
58	Davis Cup: Fitzgerald says Aussies will upset Argentina
59	Alonso finds his range as Reds win
60	Tiger bares teeth over porn link
61	Arsenal restructure $493m of debt
62	Crespo gives Inter victory in Rome
63	Hingis survives Obziler fightback
64	English FA launches bung inquiry
65	Pauleta gives PSG some rare cheer
66	Pearce under pressure as City lose
67	Real deal for Beckham within weeks
68	Struggling Leeds, QPR ring changes
69	Rooney frustrated by slump
70	Hurricane Gordon causes Ryder Cup chaos
71	English FA to investigate claims Allardyce took bribes
72	Hurricane causes Ryder Cup
73	Injured Henin-Hardenne out for three weeks
74	Safin picked ahead of Davydenko for singles play

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
75	Arsenal debt increases by £100m
76	Real president: Beckham to agree new deal
77	Furyk next in line for Woods
78	Ryder Cup revived Montgomeries career
79	Swiss star into the last eight
80	Davenport Struggles past Chakvetadze
81	Benitez never worried
82	Bolton may ask for more details
83	Owen wont rule out Red Devils
84	Rooney below usual standard
85	European roundup: AC Milan pulls into positive territory in Serie A
86	Davenport advances at China Open
87	English soccer officials open inquiry into alleged bribery
88	Rusedski is closer to retiring
89	Carver takes Leeds caretaker role
90	Nalbandian stokes up Hewitt feud
91	Mickelson dismisses form concerns
92	Weather fears for Ryder ceremony
93	Ryder Cup opening ceremony begins
94	Pompey boss hits out at Panorama
95	Pearce shrugs off pressure talk
96	Woosnam switches pairings again
97	Davis Cup: Safin, Youzhny to play singles, Davydenko dropped
98	Second-ranked Nadal to face Seppi in World Group playoffs
99	Safin leapfrogs Davydenko, will play Roddick in Day 1
100	Davis Cup: Nadal play for Spain
101	Federer to play for Switzerland against Serbia in World Group playoffs
102	Russia drops Davydenko for Davis Cup against U.S.
103	Mauresmo tears past Sun, moves into China Open quarters
104	Federer set to play three matches in World Group playoffs
105	Clarke and Westwood paired again

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
106	Davydenko axed for opening singles
107	Morientes happy to return to Spain
108	Kewell will undergo two operations
109	Mauresmo breezes into last eight
120	US hopes in hands of Tiger-trained rookies
121	Alonsos wonder-goal sinks Newcastle
122	Benitez defends Bellamy over alleged tunnel bust-up
123	Davenport struggles in China Open first round
124	Hingis makes Kolkata Open quarter-final
125	Monty on the trail of place in history
126	Kewell to go under the knife again
127	Peng knocks out title holder in Beijing
128	Nadal will try to keep Spain in World Group
129	Chelseas so-called slow start picking up speed
130	Essien discipline pleases Mourinho
131	Ryder Cup practice delayed by high winds
132	Woods agent considers possible lawsuit
133	Nadal to play for Spain in World Group playoffs
134	Davis Cup: Safin, Youzhny to play singles, Davydenko dropped
135	Federer to play for Swiss in World Group playoffs
136	Hingis stuggles a bit but advances in Calcutta
137	Mauresmo, Petrova cruise into quarterfinals
138	Cool Murray confident of Davis cup win
139	Mauresmo shine at suns expense
140	Mauresmo motors on in China
141	Alonso out to repeat trick
142	Europe set for final practice
143	Ljungberg eyes Gunners glory
144	Mandaric stands down
145	Mickelson rested and ready
146	Mourinho praise for Essien

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
147	Ugly Americans get silly at Ryder Cup
148	USA, in 11-year Davis Cup title drought, eager for win in Russia
149	Woods and Furyk start first for U.S. in Ryder Cup matches
150	Mauresmo beats Davenport hoodoo
151	Cink & Henry stage great comeback
152	Benitez defends Bellamy after row
153	Clarke savours emotional victory
154	Spanish duo seal comfortable win
155	Battling Brits boost survival bid
156	No. 1 Mauresmo defeats Davenport for first time since 2000
157	Safin gets Russia started with straight-set victory over Roddick
158	Vieira is given a three-match ban
159	Wildcards win to put Europe ahead
160	Rose and Cejka lead at Texas Open
161	Platini outlines aims as president
162	Newcastles Babayaro charged by FA
163	Russians hold 2-0 Davis advantage
164	Clarke birdie helps Europe to lead
165	Local favourite Mirza regales crowd to enter last eight
166	Babayaro charged with violent conduct
167	We cant afford to slip again says Uniteds Carrick
168	Davis Cup: Safin opens against Roddick in Moscow
169	Heavy artillery to get Ryder Cup off to explosive start
170	Chelsea and Fulham fight for more than bragging rights
171	Europe into early lead at Ryder Cup
172	Alonso banks on home comforts
173	Mauresmo, Davenport set up quarter-final clash
174	Davis Cup: Safin opens against Roddick in Moscow
175	Russia goes with Safin, Youzhny vs. USA
176	Alonso vows to repeat party piece
177	Carrick wary of chasing pack

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
178	Henry: Wenger has transformed us
179	Lamps: New boys learning Jose way
180	Redknapp leaps to Big Sams defence
181	Chris aims to give Chelsea the Blues
182	Spurs without influential trio
183	Clarke brings ray of sunshine to Ryder Cup start
184	Clarke shines, Tiger gets wet in Ryder Cup start
185	Europe has Cup and wants to keep it
186	Europe has the cup and wants to keep it
187	Woods gets another crack at Monty-Harrington
188	Klepac enjoys big win before home fans
189	Jankovic Through in Beijing smog
190	Rose and Cejka set pace in texas
191	Double success for GB Davis cup team
192	Safin and Youznhy set Russia Rolling
193	Calderon voices Becks hope
194	Benitez defends Bellamy
195	Wenger: Probe like a witch hunt
196	Given set to leave hospital
197	Hughes was never worried
198	Real make plans for Reyes
199	Johnson happy with Harry
200	Rio wary of Royals threat
201	Lamps: The Jose way
202	Hughes: Todd in plans
203	United urge patience on Rooney
204	Woods all wet to start Ryder Cup, but Furyk bails him out
205	Jang breaks course record, takes 3-stroke LPGA lead
206	Woods-Furyk start off with win, but Europeans finish fourball with lead
207	Russia takes 2-0 lead over U.S. in Davis Cup
208	Wet forecast muddies views on lift-and-clean rules at Ryder Cup

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
209	Roundup: Mauresmo beats Davenport to reach China Open semifinals
210	Americans need to match Europeans attitude to win Ryder Cup
211	Garcia and Olazabal turn on style
212	Woods misery as Europeans thrive
213	Johnson fires US to vital victory
214	Murray & Delgado crash in doubles
215	Acasuso wraps up Hewitt victory
216	Corretja calls time on his career
217	Doubles win puts US back in match
218	Hingis storms into Kolkata final
219	Argentina leads Australia 2-0 in Davis Cup
220	Corretja, battling eye injury, ends career after 16 years
221	Argentina leads Australia 2-0 in Davis Cup semifinal
222	Austria close to ticket for world group after 2-0 lead over Mexico
223	Davis Cup: Doubles match postponed due to rain
224	Bryan brothers win, cut Russias advantage in half
225	Hingis set up semifinal at Sunfeast
226	Federer plays doubles, gives Swiss 2-1 lead over Serbs
227	Federer gives Switzerland 1-0 lead in Davis Cup
228	Nadal helps Spain take 2-1 lead over Italy
229	world No. 2 Corretja
230	Safin beats Roddick in opening Davis Cup match
231	twins keep U.S. alive in Davis Cup
232	Top seeds Mauresmo, Kuznetsova advance to China Open final
233	Loits injury sends 15-year-old Paszek into final
234	Bommel on target as Bayern go top
235	Old Firm victory delights Strachan
236	Lehman sticks to struggling stars
237	Kahe puts Moenchengladbach on top
238	Bryan brothers maintain U.S. hopes
239	Liverpool extend winning home run

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
240	Mauresmo and Kuznetsova make final
241	Carrick says we can not afford to slip again
242	Brazil v Sweden Davis Cup play-off hit by rain delay
243	Davis Cup: Rain stops play with Hewitt on the ropes
244	Garcia just loving it as he claims Tigers scalp
245	Hingis, Mirza to clash in Kolkata Open semi-final
246	Mauresmo ends six-year draught with win over Davenport
247	Super Sergio stuns Americans in Ryder Cup
248	Carlos expects battle with Royals
249	Decade in charge not special - AW
250	English Premier League Roundup, Sep. 23
251	Harry and Sam united in adversity
252	Hughes: Southgate must prove himself
253	Kuyt: Jenas miss was key to victory
254	Moyes wary of Magpies new boys
255	Henry, Johnson pull out halves for U.S.
256	Close, but in the end still Europe
257	Federer whips Tipsarevic in straight sets
258	Safin records Davis Cup win over Roddick
259	British pair crushed in doubles
260	Gerrard eyes further glory
261	Henry hails Gunners new boy
262	Sheva vows to start firing
263	Rafa hails battling Sissoko
264	Tennis player Alex Corretja retires
265	Mauresmo, Kuznetsova to decide China Open
266	Beckham to sign new deal soon
267	Buffon confident of promotion
268	Bundesliga preview: Tight at the top
269	Lampard double downs Fulham
270	Gunners tracking Ledley

ตารางที่ ก.2 รายชื่อเว็บเพจทั้งหมด(ต่อ)

รหัสเว็บเพจ	ชื่อเว็บเพจ
271	Cisse a month away from fitness
272	Magath under fire
273	Liverpool blunt Spurs
274	Morientes critical of Premiership
275	Pizarro willing to leave Bayern
276	Mourinhos Shevchenko challenge
277	Ronaldo happy at United
278	Valeron could be back in December
279	Captains unconcerned by weather
280	DiMarco wants revenge
281	Europe take slender lead
282	Plans in place as rain continues
283	US team praise crowds
284	Woods singing in the rain...and wind
285	Woosnam goes for nationalistic feel
286	Woosnam expected US pairs
287	Give others credit, says Shikha
288	Coach leaves distracted Coria
289	Hewitt the unlikely hero
290	Hewitt feud continues
291	Nadal levels things up
292	Murray wants Davis whitewash
293	Nadal wants improvement
294	Nadal: We must show respect
295	Nalbandian predicts victory
296	Nalbandian gives Argentina perfect start
297	Philippoussis ready for dogfight
298	Roddick consults Sampras
299	Nalbandian stokes fires
300	Rusedski ready to retire

ภาคผนวก ข
ข้อคำถาม

สำหรับข้อคำถามทั้งหมดที่ใช้ในงานวิจัยนี้มีจำนวน 50 ข้อคำถาม ซึ่งสามารถแบ่งได้ดังต่อไปนี้

1) ข้อคำถามที่ประกอบด้วยคำ 1 คำ จำนวน 10 ข้อคำถาม
2) ข้อคำถามที่ประกอบด้วยคำ 2 คำ จำนวน 10 ข้อคำถาม
3) ข้อคำถามที่ประกอบด้วยคำ 3 คำ จำนวน 10 ข้อคำถาม
4) ข้อคำถามที่ประกอบด้วยคำ 4 คำ จำนวน 10 ข้อคำถาม
5) ข้อคำถามที่ประกอบด้วยคำ 5 คำ จำนวน 10 ข้อคำถาม

โดยข้อคำถามทั้งหมดสามารถแสดงได้ดังตารางที่ ข.1

ตารางที่ ข.1 แสดงข้อคำถามทั้งหมดจำนวน 50 ข้อคำถาม

ข้อคำถามที่	ข้อคำถาม
1	Retire
2	Riddick
3	Injure
4	Davenport
5	Chelsea
6	Acasuso
7	Gunner
8	Philippoussis
9	Beckham
10	Australia
11	Martina Hingis
12	Roger Federer
13	Tiger Wood
14	Rafael Nadal
15	Alex Corretja
16	China Open
17	Arsenal Player
18	Rafael Benitez
19	Lleyton Hewitt

ตารางที่ ข.1 แสดงข้อคำถามทั้งหมดจำนวน 50 ข้อคำถาม (ต่อ)

ข้อคำถามที่	ข้อคำถาม
20	Ian Woosnam
21	Tennis China Open
22	Tennis Kolkata Open
23	Tennis Davis Cup
24	Golf Ryder Cup
25	Football Laliga Spain
26	Football Premier League
27	Football Bundesliga Germany
28	Preview Football Premiership
29	Ryder Cup Delay
30	Manchester United Team
31	Kolkata Open Quarter Final
32	China Open Semi Final
33	Football English Premier League
34	Davis Cup Semi Final
35	Tennis World Group Playoff
36	China Open First Round
37	Chelsea Manager Jose Mourinho
38	Liverpool Manager Rafael Benitez
39	Europe Captain Ian Woosnam
40	Ryder Cup Opening Ceremony
41	Golf Ryder Cup Europe Team
42	Tennis Davis Cup Argentina Team
43	Manchester United Team Football Player
44	Spain Team World Group Playoff
45	Australia Player Tennis Davis Cup
46	United State Team Davis Cup
47	Chelsea Football Club Premier League
48	Liverpool Football Club English Premiership
49	Arsenal Football Club Premier League
50	Golf Ryder Cup USA Team

ภาคผนวก ค
ค่าเรียกคืน และค่าความแม่นยำที่ได้จากการทดลอง

ค่าเรียกคืน (R) ค่าความแม่นยำ (P) ทั้งหมดที่ได้จากการทดลอง แยกตามวิธีการที่ใช้ทั้งหมด 4 วิธีการในการทดลองได้แก่

1) การค้นคืนเว็บเพจที่ไม่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้

2) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ ด้วยแบบจำลองปริภูมิเวคเตอร์

3) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองปริภูมิเวคเตอร์

4) การค้นคืนเว็บเพจที่มีการให้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ด้วยวีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น

โดยแยกตามข้อคำถามที่ใช้ทั้งหมด 50 ข้อคำถามในแต่ละวิธีการค้นคืน ตามลำดับข้างต้นซึ่งมีรายละเอียดดังต่อไปนี้

ตารางที่ ค.1 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 1

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
1	0.184	1.000
2	0.320	1.000
3	0.061	0.182
4	0.333	1.000
5	0.462	0.800
6	0.313	1.000
7	0.385	0.909
8	0.250	1.000
9	0.333	1.000
10	0.773	0.944
11	0.706	1.000
12	0.200	1.000
13	0.359	0.933
14	0.226	1.000
15	0.121	1.000
16	0.432	0.905
17	0.143	1.000
18	0.133	1.000
19	0.455	1.000
20	0.174	1.000
21	0.350	1.000
22	0.167	0.700
23	0.811	0.741
24	0.797	0.797
25	0.167	0.500
26	0.507	0.384
27	0.182	1.000
28	0.036	0.500
29	0.152	1.000
30	0.326	1.000

ตารางที่ ค.1 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 1 (ต่อ)

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
31	0.058	1.000
32	0.067	1.000
33	0.492	0.372
34	0.525	0.837
35	0.368	0.933
36	0.095	1.000
37	0.149	1.000
38	0.095	1.000
39	0.163	1.000
40	0.056	1.000
41	0.373	0.846
42	0.189	1.000
43	0.231	1.000
44	0.194	1.000
45	0.193	0.923
46	0.214	0.800
47	0.154	0.666
48	0.211	1.000
49	0.157	1.000
50	0.302	0.826

ตารางที่ ค.2 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 2

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
1	0.555	0.555
2	0.727	1.000
3	0.333	0.182
4	0.750	1.000
5	0.800	0.533
6	0.454	1.000
7	0.545	1.000
8	0.307	1.000
9	1.000	1.000
10	1.000	0.389
11	0.706	1.000
12	0.286	1.000
13	0.424	0.933
14	0.241	1.000
15	0.571	1.000
16	0.545	0.851
17	0.267	1.000
18	0.462	1.000
19	0.454	1.000
20	0.195	1.000
21	0.273	1.000
22	0.179	0.700
23	0.693	0.896
24	0.934	0.891
25	0.333	0.500
26	0.532	0.384
27	0.231	1.000
28	0.060	0.750
29	0.135	0.714
30	0.311	1.000

ตารางที่ ค.2 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 2 (ต่อ)

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
31	0.065	1.000
32	0.059	1.000
33	0.587	0.546
34	0.479	0.946
35	0.500	0.933
36	0.106	1.000
37	0.179	1.000
38	0.146	1.000
39	0.163	1.000
40	0.077	1.000
41	0.316	0.923
42	0.129	1.000
43	0.152	1.000
44	0.149	1.000
45	0.153	1.000
46	0.189	0.933
47	0.113	0.800
48	0.184	1.000
49	0.103	1.000
50	0.278	0.956

ตารางที่ ค.3 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 3

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
1	0.429	0.778
2	0.348	1.000
3	0.500	0.182
4	0.429	1.000
5	0.458	0.733
6	0.333	1.000
7	0.600	0.818
8	0.267	1.000
9	0.316	1.000
10	0.591	0.722
11	0.387	1.000
12	0.240	1.000
13	0.419	0.867
14	0.219	1.000
15	0.222	1.000
16	0.344	1.000
17	0.129	1.000
18	0.273	1.000
19	0.435	1.000
20	0.222	1.000
21	0.323	1.000
22	0.239	1.000
23	0.719	0.793
24	0.725	0.906
25	0.200	0.750
26	0.577	0.523
27	0.214	1.000
28	0.050	0.750
29	0.162	0.857
30	0.302	0.928

ตารางที่ ค.3 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 3 (ต่อ)

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
31	0.053	1.000
32	0.057	1.000
33	0.578	0.558
34	0.430	0.919
35	0.292	0.933
36	0.094	1.000
37	0.179	1.000
38	0.122	1.000
39	0.136	1.000
40	0.040	1.000
41	0.321	1.000
42	0.262	1.000
43	0.121	1.000
44	0.104	1.000
45	0.159	1.000
46	0.153	1.000
47	0.134	1.000
48	0.174	1.000
49	0.103	1.000
50	0.250	1.000

ตารางที่ ค.4 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 4

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
1	0.229	1.000
2	0.242	1.000
3	0.061	0.182
4	0.300	1.000
5	0.400	0.267
6	0.455	1.000
7	0.400	0.545
8	0.267	1.000
9	0.250	1.000
10	0.455	0.556
11	0.706	1.000
12	0.115	1.000
13	0.385	1.000
14	0.250	1.000
15	0.182	1.000
16	0.339	1.000
17	0.069	1.000
18	0.136	1.000
19	0.227	1.000
20	0.143	1.000
21	0.420	1.000
22	0.188	0.900
23	0.706	0.828
24	0.792	0.953
25	0.170	1.000
26	0.659	0.721
27	0.200	1.000
28	0.034	1.000
29	0.128	0.857
30	0.237	1.000

ตารางที่ ค.4 ค่าเรียกคืน และค่าความแม่นยำ ในการค้นคืนด้วยข้อคำถาม 50 ข้อ ของวิธีการที่ 4 (ต่อ)

ข้อคำถาม	ค่าเรียกคืน	ค่าความแม่นยำ
31	0.065	1.000
32	0.068	1.000
33	0.733	0.767
34	0.353	0.946
35	0.378	0.933
36	0.086	1.000
37	0.129	1.000
38	0.087	1.000
39	0.127	1.000
40	0.037	1.000
41	0.312	0.962
42	0.244	1.000
43	0.154	1.000
44	0.125	1.000
45	0.131	1.000
46	0.119	1.000
47	0.109	1.000
48	0.145	1.000
49	0.095	1.000
50	0.237	1.000

ภาคผนวก ง
สรุปสูตรที่ใช้ในงานวิจัย

ตารางที่ ง.1 สรุปสูตรทั้งหมดที่ใช้ในงานวิจัย

ที่	สูตร
1.	$Weight_{ik} = \frac{Freq_{ik}}{TotFreq_k}$
	เป็นสูตรในการคำนวณหาค่าน้ำหนักของคำแต่ละคำในแต่ละเอกสารเว็บเพจ โดยใช้ค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน
2.	$Similarity\,(DOC_i, Query_j) = \frac{\sum_{k=1}^{t}(Term_{ik} \cdot QTerm_{jk})}{\sqrt{\sum_{k=1}^{t}(Term_{ik})^2 \cdot \sum_{k=1}^{t}(QTerm_{jk})^2}}$
	เป็นสูตรในการคำนวณหาค่าความคล้ายระหว่างข้อคำถามกับเอกสารเว็บเพจ
3.	Recall (R) = $\frac{\lvert R_a \rvert}{\lvert R \rvert}$
	เป็นสูตรในการคำนวณหาค่าเรียกคืน เพื่อประเมินประสิทธิผลของระบบ
4.	Precision (P) = $\frac{\lvert R_a \rvert}{\lvert A \rvert}$
	เป็นสูตรในการคำนวณหาค่าความแม่นยำ เพื่อประเมินประสิทธิผลของระบบ
5.	$Q' = \alpha Q + \beta\left(\frac{1}{R'}\sum_{i \in D_{R'}} DOC_i\right) - \gamma\left(\frac{1}{N'}\sum_{i \in D_{N'}} DOC_i\right)$
	เป็นสูตรในการขยายคำและค่าน้ำหนักของคำจากข้อคำถามเดิมเป็นข้อคำถามใหม่ในแบบจำลองปริภูมิเวคเตอร์
7.	$W_{i,j} = \log \frac{\frac{r}{R-r}}{\frac{n-r}{(N-n)-(R-r)}}$
	เป็นสูตรในการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ด้วยแบบจำลองความน่าจะเป็น

ภาคผนวก จ
บทความวิชาการที่ตีพิมพ์

ในการวิจัยนี้ ผู้วิจัยมีผลงานวิชาการร่วมกับคณะผู้วิจัย เป็นบทความวิชาการระดับชาติ 1 บทความ ได้แก่

จ.1 บทความวิชาการเรื่อง “การค้นคืนย้อนกลับจากผู้ใช้ในสารสนเทศแบบเว็บโดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น (User Relevance Feedback for Web Information Retrieval Using Web Page Segmentation and Probabilistic Model)” ซึ่งได้รับการคัดเลือกเพื่อนำเสนอและตีพิมพ์ในงาน “การประชุมวิชาการร่วมสาขาวิทยาการคอมพิวเตอร์และวิศวกรรมซอฟต์แวร์ ครั้งที่ 3 (The 3rd Joint Conference on Computer Science and Software Engineering: JCSSE 2006)” ระหว่างวันที่ 29 – 30 มิถุนายน 2549 ณ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้า วิทยาเขตพระนครเหนือ กรุงเทพฯ

การค้นคืนย้อนกลับจากผู้ใช้ในสารสนเทศแบบเว็บโดยการแบ่งเว็บเพจเป็นส่วนย่อย และใช้แบบจำลองความน่าจะเป็น

User Relevance Feedback for Web Information Retrieval Using Web Page Segmentation and Probabilistic Model

ปรารถนา จันพลโท และ นครทิพย์ พร้อมพูล
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
อีเมล : prattana.c@chula.ac.th และ nakornthip.s@chula.ac.th

บทคัดย่อ

การค้นคืนสารสนเทศแบบเว็บนั้นจะค้นคืนหาคำที่ใช้ในข้อคำถามเปรียบเทียบกับสารสนเทศแบบเว็บที่ปรากฏคำๆนั้นทั้งหน้าเว็บเพจ ซึ่งผลการค้นคืนที่ได้อาจจะไม่ตรงตามความต้องการของผู้ใช้เท่าที่ควร จึงจำเป็นต้องทำการค้นคืนย้อนกลับด้วยการกำหนดข้อคำถามใหม่จากผลตอบกลับซึ่งอาศัยผลการประเมินจากผู้ใช้เพื่อช่วยให้ผลการค้นคืนที่ได้ตรงตามความต้องการของผู้ใช้ได้มากขึ้น งานวิจัยนี้เสนอการค้นคืนย้อนกลับสารสนเทศแบบเว็บโดยการแบ่งเว็บเพจเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม และเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ที่ใช้ในการค้นคืนย้อนกลับด้วยแบบจำลองความน่าจะเป็น รวมทั้งแสดงตัวอย่างของการดำเนินการและวิธีการประเมินผลลัพธ์ของการค้นคืนย้อนกลับตามวิธีการที่นำเสนอมาด้วยค่าความถูกต้อง

คำสำคัญ: การค้นคืนย้อนกลับจากผู้ใช้ การแบ่งเว็บเพจเป็นส่วนย่อย แบบจำลองความน่าจะเป็น การกำหนดข้อคำถามใหม่ การเปลี่ยนแปลงคำในข้อคำถาม การเปลี่ยนแปลงค่าน้ำหนักคำ

Abstract

Web information retrieval process is concerned with the similarity between term in query and index term of web information. The query results often do not meet user requirements so it is necessary to retrieve again with user relevance feedback which can improve the result of web information retrieval. This research purposes a user relevance feedback for web information retrieval using web page segmentation with VIPS algorithm and reweighting term with probabilistic model. The example of this process and evaluation using precision are also presented.

Keywords: User Relevance Feedback, Web Page Segmentation, Probabilistic Model, Query Reformulation, Query Expansion, Term Reweighting

1. บทนำ

ในการค้นคืนสารสนเทศแบบเว็บจะเกี่ยวข้องกับการค้นหาหน้าเว็บเพจที่ตรงตามความต้องการของผู้ใช้กับข้อคำถาม (Query) โดยจะค้นคืนหาคำที่ใช้ในข้อคำถามเปรียบเทียบกับสารสนเทศแบบเว็บเพจในฐานข้อมูลแล้วแสดงผลการค้นคืนที่ได้เป็นรายการเหล่านั้น ซึ่งโดยทั่วไปจะทำการเปรียบเทียบคำในข้อคำถามกับเอกสารแบบเว็บเพจที่ปรากฏคำๆนั้นทั้งหน้าเว็บเพจ เช่น ถ้าต้องการค้นคืนเอกสารที่มีคำว่า Information Retrieval ก็จะค้นหาคำว่า Information Retrieval ที่ปรากฏในทุกส่วนของหน้าเว็บเพจ โดยผลการค้นคืนที่ได้นั้น อาจจะไม่ตรงตามความต้องการของผู้ใช้เพราะ อาจจะพบคำในส่วนอื่นที่ไม่ใช่ส่วนสำคัญ เช่น โฆษณา แต่เอกสารนั้นก็ถูกนำมาแสดงในรายการค้น

คืน ซึ่งเป็นปัญหาสำคัญ ดังนั้นเพื่อให้ได้ผลการค้นคืนตรงตามความต้องการของผู้ใช้ จึงต้องทำการค้นคืนย้อนกลับ (Relevance Feedback) [2] โดยการกำหนดข้อคำถามใหม่ (Query Reformulation) ด้วยการเพิ่มคำ (Term) เข้าไปในข้อคำถามเดิม แล้วทำการค้นคืนซ้ำอีกครั้ง ซึ่งสามารถช่วยให้ผลการค้นคืนที่ได้ตรงตามความต้องการของผู้ใช้เพิ่มขึ้นบ้าง แต่ก็ยังคงมีปัญหาเกิดขึ้นคือ คำที่เพิ่มเข้าไปในข้อคำถามเดิมนั้น ถูกเลือกมาจากทั้งหน้าเว็บเพจ ดังนั้นเพื่อแก้ปัญหาดังกล่าวจึงนำวีไอพีเอสอัลกอริทึม [5] (Vision-based Pages Segmentation :VIPS algorithm) มาใช้ในการแบ่งสารสนเทศแบบเว็บเพจที่ได้จากการค้นคืนครั้งแรกเป็นส่วนย่อย เพื่อช่วยเลือกคำที่จะมากำหนดข้อคำถามใหม่จากส่วนที่สำคัญ และแบบจำลองความน่าจะเป็นนำมาใช้ในการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามใหม่ที่ใช้ในการค้นคืนย้อนกลับ

โดยงานวิจัยที่เกี่ยวข้องจะกล่าวถึงในหัวข้อที่ 2 เรื่องของระบบการจัดเก็บและค้นคืนสารสนเทศ การค้นคืนย้อนกลับจากผู้ใช้ การประเมินผลลัพธ์ที่ได้ จะกล่าวถึงใน หัวข้อที่ 3, 4, และ 5 ตามลำดับ วีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็นนั้น จะอยู่ในหัวข้อที่ 6 และ 7 โดยวิธีการในการค้นคืนย้อนกลับของงานวิจัยนี้ และกรณีตัวอย่าง จะกล่าวถึงในหัวข้อที่ 8 และ 9 ซึ่งส่วนสุดท้าย คือสรุปผลงานวิจัยและการดำเนินการในอนาคต

2. งานวิจัยที่เกี่ยวข้อง

Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma [5] นำเสนอเกี่ยวกับวีไอพีเอสอัลกอริทึม ซึ่งใช้ในการแบ่งหน้าเว็บเพจออกเป็นบล็อกทำให้สามารถเพิ่มประสิทธิภาพในการค้นคืนสารสนเทศได้ โดยทำการเปรียบเทียบประสิทธิภาพของการค้นคืนสารสนเทศกับวิธีการแบบดอม(DOM-based Page Segmentation) [4] และแบบค้นคืนทั้งเอกสาร (Full Document) ซึ่งผลที่ได้คือ ค่าความถูกต้องจากการค้นคืนสารสนเทศโดยใช้ วีไอพีเอสอัลกอริทึมนั้น ให้ค่าที่มากที่สุด โดยผลการทดลองที่ได้จากงานวิจัยนี้ทำให้เกิดงานวิจัยอีกงานหนึ่ง คือ Shipeng Yu, Deng Cai, Ji-Rong Wen, Wei-Ying Ma [6] งานวิจัยนี้นำวีไอพีเอสอัลกอริทึมมาใช้ในการเพิ่มประสิทธิภาพของการค้นคืนย้อนกลับแบบเทียม (Pseudo-Relevance Feedback) ในการค้นคืนสารสนเทศแบบเว็บ โดยใช้การแบ่งเว็บเพจเป็นบล็อก เพื่อช่วยเลือกคำที่จะนำมากำหนดข้อคำถามใหม่ในการค้นคืนย้อนกลับแบบเทียมส่งผลให้ประสิทธิภาพของการค้นคืนเพิ่มขึ้นได้ถึง 27 เปอร์เซนต์ และยังมีงานวิจัยของ Seung Yeol และ Achim Hoffman [9] ที่นำเสนอเกี่ยวกับการค้นคืนย้อนกลับแบบเทียมในการค้นคืนสารสนเทศแบบเว็บ ที่มีการแบ่งเว็บเพจเป็นบล็อกโดยใช้วีไอพีเอสอัลกอริทึมจากงานวิจัย [5] และการลดความกำกวมของคำมาประยุกต์รวมกัน ซึ่งผลที่ได้จากการค้นคืนย้อนกลับแบบเทียมนั้นพบว่าค่าความถูกต้องของการค้นคืนยังมีค่าน้อย เมื่อเทียบกับค่าเรียกคืนที่ได้ค่ามากกว่า ซึ่งยังคงเป็นปัญหาสำคัญที่ต้องแก้ไข

จากงานวิจัยที่กล่าวมาทั้งหมดแสดงให้เห็นว่า วีไอพีเอสอัลกอริทึม สามารถช่วยเพิ่มประสิทธิภาพในการค้นคืนย้อนกลับแบบเทียมได้ แต่ไม่ได้สนใจในส่วนของผลตอบกลับจากผู้ใช้แต่จะทำการเลือกคำที่จะมากำหนดข้อคำถามใหม่จากรายการเอกสารที่ถูกจัดอันดับแล้วว่าเกี่ยวข้องกับข้อคำถามใน 10 อันดับแรก หรือมากกว่าแล้วแต่จะกำหนด ดังนั้นงานวิจัยนี้ขอเสนอการค้นคืนย้อนกลับสารสนเทศแบบเว็บ โดยการแบ่งเว็บเพจเป็นส่วยย่อย และใช้แบบจำลองความน่าจะเป็นโดยที่จะสนใจผลตอบกลับจากผู้ใช้ในการเลือกคำจากส่วนของเว็บเพจที่แบ่งเป็นส่วนย่อยซึ่งนำมาใช้ในการกำหนดข้อคำถามใหม่ เพราะการค้นคืนย้อนกลับจากผู้ใช้นั้นถือว่ามีส่วนสำคัญที่จะทำให้ผลการค้นคืนย้อนกลับที่ได้ตรงตามความต้องการของผู้ใช้ และแบบจำลองความน่าจะเป็นจะใช้ในการเปลี่ยนแปลงค่าน้ำหนักของคำที่ใช้ในการกำหนดข้อคำถามใหม่ เพราะจากงานวิจัยที่ผ่านมาของ Harman [1] พบว่าการเปลี่ยนแปลงค่าน้ำหนักของคำด้วยแบบจำลอง

ความน่าจะเป็นช่วยเพิ่มประสิทธิภาพของการค้นคืนย้อนกลับได้

3. ระบบการจัดเก็บและค้นคืนสารสนเทศ

เป็นระบบที่มีการจัดเก็บสารสนเทศเพื่อใช้ในการประมวลผล การค้นคืนสารสนเทศ รวมทั้งนำเสนอในรูปแบบที่เหมาะสม ซึ่งระบบค้นคืนสารสนเทศจะช่วยให้ผู้ใช้สามารถค้นหาข้อมูลที่ตรงตามความต้องการได้สะดวกรวดเร็วขึ้น โดยผู้ใช้จะค้นหาข้อมูลที่ตนเองต้องการด้วยการใช้ข้อคำถามซึ่งประกอบด้วยคำสำคัญ (Keyword) สำหรับใช้ค้นหาสารสนเทศที่ต้องการ จากนั้นระบบจะทำการค้นคืนเอกสารที่ตรงกับสิ่งที่ผู้ใช้ต้องการ กระบวนการในการค้นคืนสารสนเทศ [8] แสดงได้ดังรูปที่ 1

รูปที่ 1. กระบวนการจัดเก็บและค้นคืนสารสนเทศ

ในกระบวนการจัดเก็บและค้นคืนสารสนเทศของงานวิจัยนี้แบ่งเป็น 2 ส่วนหลักๆ คือ ส่วนของการจัดเก็บสารสนเทศแบบเว็บ และส่วนของการค้นคืนสารสนเทศแบบเว็บ ซึ่งมีดังนี้

1) ในการค้นคืนสารสนเทศแบบเว็บนั้นเริ่มต้นด้วยการจัดเก็บสารสนเทศแบบเว็บไว้ในฐานข้อมูลโดยใช้โครงสร้างการจัดเก็บแบบแฟ้มข้อมูลผกผัน (Inverted File)

2) ในส่วนของการสร้างดัชนีซึ่งเป็นคำสำคัญที่ปรากฏในสารสนเทศแบบเว็บจะใช้พอตเตอร์อัลกอริทึม (Potter Algorithm) ในการดำเนินการต่างๆของขั้นตอนการสร้างดัชนี และกำหนดค่าน้ำหนักของคำสำคัญ ด้วยค่าความถี่ของคำและความถี่ของเอกสารแบบผกผัน (Inverse Document Frequency: IDF)

3) ขั้นตอนต่อไปคือการดำเนินการข้อคำถามที่จะใช้ในการค้นคืนสารสนเทศแบบเว็บที่มีในฐานข้อมูล โดยในงานวิจัยนี้จะทำการสร้างข้อคำถามที่ประกอบด้วยคำสำคัญ(Keyword) ขึ้นมาจากการศึกษาข้อมูลสารสนเทศแบบเว็บที่มีในปัจจุบันอย่างน้อย 50 ข้อคำถาม

4) เมื่อได้ข้อคำถามที่จะนำมาค้นคืนสารสนเทศแบบเว็บแล้ว ก็จะเข้าสู่กระบวนการในการค้นคืน คือทำการเปรียบเทียบความคล้ายของคำสำคัญในข้อคำถามกับสารสนเทศแบบเว็บที่มีในฐานข้อมูล ซึ่งจะใช้ค่าความคล้ายแบบโคซายน์ (Cosine Similarity) [3] ดังสมการที่ (1)

$$COSINE(DOC_i, QUERY_j) = \frac{\sum_{k=1}^{t}(TERM_{ik} \cdot QTERM_{jk})}{\sqrt{\sum_{k=1}^{t}(TERM_{ik})^2 \cdot \sum_{k=1}^{t}(QTERM_{jk})^2}} \quad (1)$$

$TERM_{ik}$ = ค่าน้ำหนักของคำ k ในเอกสาร i

$QTERM_{jk}$ = ค่าน้ำหนักของคำ k ในข้อคำถาม j

5) เรียงลำดับผลการค้นคืนที่ได้ตามค่าความคล้ายจากมากไปน้อย แล้วแสดงผลลัพธ์ที่ได้กลับสู่ผู้ใช้งาน

4. การประเมินผลลัพธ์ที่ได้

ค่าความถูกต้อง (Precision) หมายถึง สัดส่วนของเอกสารที่ค้นคืนมาได้และตรงกับความต้องการ ใช้ในการวัดประสิทธิภาพของระบบการค้นคืนสารสนเทศ โดยมีสูตรการคำนวณ ดังนี้คือ [3]

$$Precision = Ra/A$$

Ra = จำนวนเอกสารตรงตามต้องการที่ค้นคืนได้

A = จำนวนเอกสารทั้งหมดที่ค้นคืนออกมา

ค่าความถูกต้องเป็นปริมาณที่แสดงว่าการค้นคืนเอกสารได้ตรงตามต้องการเพียงใด เช่น ถ้าค้นคืนเอกสารออกมาได้ A ฉบับ และมีเอกสารอยู่ Ra ฉบับที่ตรงตามต้องการ ดังนั้นค่าความถูกต้องมีค่าเป็น Ra/A หรือเป็นโอกาสของเอกสารที่ค้นคืนออกมาตรงตามต้องการ ค่าความถูกต้องมีค่าอยู่ระหว่าง 0 ถึง 1 และ

งานวิจัยนี้จะใช้ค่าความถูกต้องในการวัดประสิทธิภาพของการค้นคืนย้อนกลับที่ได้จากการทดสอบกับกลุ่มตัวอย่างทั้ง 2 กลุ่ม ซึ่งจะกล่าวถึงในส่วนของการค้นคืนย้อนกลับสารสนเทศแบบเว็บในหัวข้อที่ 8

5. การค้นคืนย้อนกลับจากผู้ใช้

การค้นคืนย้อนกลับเป็นวิธีการที่ช่วยในการกำหนดข้อคำถามใหม่ ให้สามารถค้นคืนเอกสารได้ตรงตามความต้องการของผู้ใช้มากขึ้นกว่าข้อคำถามเดิม โดยรูปแบบการกำหนดข้อคำถามใหม่มี 2 แบบ คือ

1) การเปลี่ยนแปลงคำในข้อคำถาม (Query Expansion) เป็นการเปลี่ยนแปลงข้อคำถามเดิมที่ใช้ในการค้นคืนสารสนเทศด้วยการเพิ่มคำเข้าไปในข้อคำถามใหม่ โดยคำที่จะเพิ่มเข้าไปในข้อคำถามใหม่นั้นเลือกมาจากเอกสารที่เกี่ยวข้องที่ถูกค้นคืนมาได้ในครั้งแรก

2) การเปลี่ยนแปลงค่าน้ำหนักคำในข้อคำถาม(Term Reweighting) เป็นการเปลี่ยนแปลงค่าน้ำหนักของคำในข้อคำถามที่ตรงกับเอกสารที่เกี่ยวข้องให้มีค่าเพิ่มขึ้นและลดค่าน้ำหนักของคำในข้อคำถามในเอกสารที่ไม่เกี่ยวข้องให้มีค่าลดลง

รูปแบบการกำหนดข้อคำถามใหม่ทั้ง 2 แบบข้างต้น สามารถทำได้โดยผ่านทางการค้นคืนย้อนกลับจากผู้ใช้ (User Relevance Feedback) ซึ่งกระบวนการในการค้นคืนย้อนกลับจากผู้ใช้นั้นแสดงได้ดังรูปที่ 2

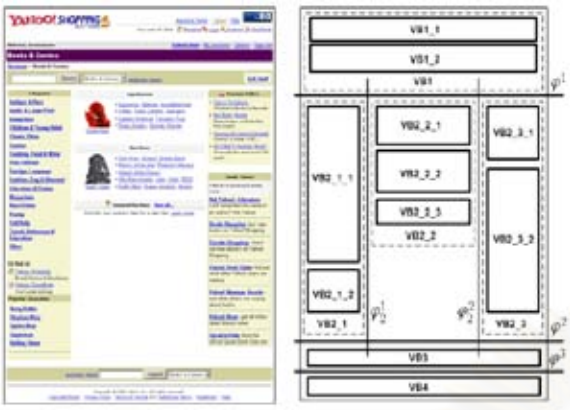
รูปที่ 2. กระบวนการค้นคืนย้อนกลับจากผู้ใช้

การค้นคืนย้อนกลับจากผู้ใช้จะทำการค้นคืนสารสนเทศโดยการเปรียบเทียบความคล้ายระหว่างข้อคำถามกับเอกสารที่มีในฐานข้อมูลแล้วระบบจะแสดงผลการค้นคืนสารสนเทศที่เกี่ยวข้องแก่ผู้ใช้งาน แล้วผู้ใช้จึงทำการตรวจสอบเอกสารที่ค้นคืนออกมาได้ ให้ผู้ใช้พิจารณาเลือกเอกสารที่เห็นว่าตรงตามที่ต้องการ โดยการทำเครื่องหมายไว้จากนั้นระบบจะทำการเลือกคำที่จะใช้ในการเปลี่ยนแปลงข้อคำถามใหม่จากเอกสารเหล่านั้นเพื่อให้ได้ข้อคำถามใหม่ที่ดีขึ้น แล้วผู้ใช้ก็จะป้อนข้อคำถามใหม่ที่ได้กลับสู่ระบบอีกครั้ง เพื่อทำการค้นคืนใหม่ ซึ่งจุดประสงค์สำคัญของการเปลี่ยนแปลงคำในข้อคำถามคือ ข้อคำถามใหม่นั้นควรมีความคล้าย (Similarity) กับเอกสารที่เกี่ยวข้องมากขึ้น และมีความคล้ายกับเอกสารที่ไม่เกี่ยวข้องน้อยลง เมื่อเปรียบเทียบกับข้อคำถามเดิม

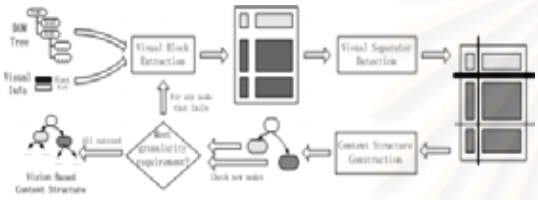
6. วีไอพีเอสอัลกอริทึม

ในการแยกแยะความแตกต่างของสารสนเทศแบบเว็บเพจนั้น จำเป็นที่จะต้องทำการแบ่งเว็บเพจออกเป็นบล็อก ซึ่งก็มีวิธีการหลายวิธีในการแบ่งเว็บเพจ แต่ที่ได้รับความนิยมคือ แบบดอม (DOM-based segmentation) [4] แบบตำแหน่ง (Location-based segmentation) [7] และแบบวีไอพีเอส (Vision-based Pages Segmentation: VIPS) ซึ่งจากงานวิจัย [5] พบว่า วีไอพีเอสอัลกอริทึมนั้นสามารถเพิ่มประสิทธิภาพในการค้นคืนสารสนเทศได้ดีที่สุด

วีไอพีเอสอัลกอริทึม เป็นอัลกอริทึมที่ใช้ในการแบ่งหน้าเว็บเพจออกเป็นบล็อก เพื่อให้ง่ายต่อการค้นคืนสารสนเทศ ซึ่งมาจากการรวมโครงสร้างแบบดอม (DOM Structure) และโครงสร้างเนื้อหา (Vision-based Content Structure) จุดประสงค์หลักของวีไอพีเอสอัลกอริทึมคือ เพื่อแบ่งหน้าเว็บเพจให้มีลักษณะเป็นโครงสร้างของเนื้อหาสำคัญ โดยอยู่บนพื้นฐานของการแบ่งเนื้อหาที่สำคัญตามความหมายซึ่งสามารถช่วยให้ส่วนที่ไม่สำคัญในหน้าเว็บเพจนั้นถูกตัดออกไป โครงสร้างของวีไอพีเอสอัลกอริทึม [5] แสดงได้ดังรูปที่ 3 และกระบวนการในการแบ่งเว็บเพจออกเป็นบล็อก [5] นั้น สามารถแสดงได้ดังรูปที่ 4



รูปที่ 3. ตัวอย่างโครงสร้างของเว็บเพจที่แบ่งเป็นส่วนย่อยด้วยวิไอพีเอสอัลกอริทึม



รูปที่ 4. ขั้นตอนการทำงานของวิไอพีเอสอัลกอริทึม

ขั้นตอนการทำงานของวิไอพีเอสอัลกอริทึม

1) การตัดออกเป็นบล็อก (Visual Block Extraction) ขั้นตอนนี้จะทำการตัดหน้าเว็บเพจออกเป็นส่วน คือเป็นบล็อก โดยที่แต่ละบล็อกมีเนื้อหาคล้ายกันหรือทำนองเดียวกัน

2) การตรวจหาตัวแบ่งแยก (Visual Separator Detection) ขั้นตอนนี้จะเป็นการกำหนดส่วนที่ถูกแบ่งออกเป็นบล็อกให้แยกออกจากกันโดยตัวแบ่งแยกซึ่งอาจจะแบ่งตามแนวตั้งหรือแนวนอนก็ได้ แล้วให้ค่าน้ำหนักของแต่ละส่วนตามความสำคัญ

3) การสร้างโครงสร้างของเนื้อหา (Content Structure Construction) ในขั้นตอนนี้จะทำการสร้างส่วนของเนื้อหาที่สำคัญให้มีลักษณะเป็นโครงสร้างที่ชัดเจน

4) ทำกระบวนการข้างต้นซ้ำอีก (Iterating the Above Step) ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการทำงานของวิไอพีเอสอัลกอริทึม ซึ่งจะทำตามข้อ 1-3 ที่กล่าวมาข้างต้นซ้ำจนกว่าจะสามารถพบส่วนตรงกับความต้องการของผู้ใช้

7. แบบจำลองความน่าจะเป็น

เป็นแบบจำลองแบบหนึ่งในการค้นคืนสารสนเทศ ซึ่งนำเสนอครั้งแรกโดย Maron & Kuhns ต่อมามีการขยายความโดย Spark Jone ภายหลังรู้จักในชื่อ Binary Independence Retrieval (BIR) Model โดยแนวคิดของการค้นคืนนั้นพยายามที่จะตอบคำถามที่ว่าอะไรคือความน่าจะเป็นที่เอกสารชิ้นนั้นเกี่ยวข้องกับข้อคำถามที่ให้ในการค้นคืนสารสนเทศ ซึ่งการใช้แบบจำลองความน่าจะเป็นนั้น จะค้นคืนเอกสารโดยเรียงลำดับผลการค้นคืนที่ได้ ตามลำดับความน่าจะเป็นที่เอกสารนั้นเกี่ยวข้องกับข้อคำถามจากมากไปน้อย และพยายามที่จะแก้ปัญหาเกี่ยวกับการค้นคืนสารสนเทศ เพื่อใช้ในการค้นคืนย้อนกลับให้มีประสิทธิภาพมากขึ้น สามารถหา ค่าน้ำหนักของคำที่เกี่ยวข้องกับเอกสารได้ดังสูตรที่ (2) ของ Robertson and Spark Jones ซึ่งค่าน้ำหนักส่วนนี้จะใช้ในการกำหนดค่าน้ำหนักของคำสำหรับเอกสารที่ค้นคืนมาได้ในครั้งแรกแล้วเรียงลำดับค่าน้ำหนักของคำเหล่านั้นเพื่อนำไปใช้ในการปรับเปลี่ยนคำในข้อคำถาม

$$W_i = \log \frac{\frac{r}{R-r}}{\frac{n-r}{(N-n)-(R-r)}} \qquad (2)$$

- W_i = ค่าน้ำหนักของคำ i
- r = จำนวนของเอกสารที่เกี่ยวข้องที่มีคำ i ปรากฏอยู่
- R = จำนวนของเอกสารทั้งหมดที่เกี่ยวข้อง
- n = จำนวนของเอกสารในคอลเลคชันที่มีคำ i ปรากฏอยู่
- N = จำนวนของเอกสารทั้งหมดในคอลเลคชัน

8. วิธีการค้นคืนย้อนกลับสารสนเทศแบบเว็บ

ในการค้นคืนย้อนกลับสารสนเทศแบบเว็บด้วยการแบ่งเว็บเพจเป็นส่วนย่อย และใช้แบบจำลองความน่าจะเป็นนั้น มีขั้นตอนแสดงได้ดังรูปที่ 5

รูปที่ 5. แนวคิดการทำงานของโมเดลในการค้นคืนย้อนกลับ

1) เริ่มต้นด้วยการค้นคืนสารสนเทศครั้งแรกแบบธรรมดาตามแนวทางที่กล่าวไว้ในหัวข้อที่ 3 ผลลัพธ์ที่ได้คือ รายการเอกสารที่เกี่ยวข้องซึ่งเรียงลำดับตามความคล้ายจากมากไปน้อย

2) หลังจากได้เอกสารที่เกี่ยวข้องจากการค้นคืนครั้งแรกแล้วจึงทำการแบ่งเอกสารแบบเว็บเพจเหล่านั้นออกเป็นส่วนย่อยโดยใช้วิธีไอพีเอสอัลกอริทึม ซึ่งจะได้หน้าเว็บเพจที่มีลักษณะเป็นบล็อก จากนั้นผู้ใช้งานเลือกเฉพาะบล็อกที่เห็นว่าเกี่ยวข้องเพื่อใช้ในการหาคำที่จะมากำหนดข้อคำถามใหม่ ซึ่งทำให้ส่วนที่ไม่เกี่ยวข้องถูกตัดออกไป

3) ทำการหาคำที่จะใช้ในการกำหนดข้อคำถามใหม่จากบล็อกที่ผู้ใช้ระบุว่าเกี่ยวข้องแล้วเปลี่ยนแปลงค่าน้ำหนักของคำ ตามสมการที่ (2)

4) เมื่อได้คำที่จะใช้ในการกำหนดข้อคำถามใหม่แล้วจึงทำการค้นคืนย้อนกลับ

5) แสดงผลการค้นคืนที่ได้เป็นรายการอกสารที่เกี่ยวข้องกลับสู่ผู้ใช้งาน

ทำการทดสอบการทำงานของโมเดลที่สร้างขึ้นโดยที่ส่งชุดของข้อคำถามที่จะใช้ในการทดสอบเข้าไปในโมเดล แล้วให้โมเดลทำการค้นคืนเอกสารมาให้โดยแบ่งการทดสอบออกเป็น 2 กลุ่มคือ

กลุ่มที่ 1 จะใช้วิธีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในการค้นคืนแบบย้อนกลับซึ่งหลังจากที่ได้เอกสารที่ค้นคืนมาแล้วผู้ใช้งานเลือกบล็อก ที่เห็นว่าเกี่ยวข้องกับที่ต้องการ แล้วโมเดลจะทำการปรับเปลี่ยนค่าน้ำหนักของคำจากบล็อกที่ผู้ใช้เลือกแล้ว ด้วยแบบจำลองความน่าจะเป็นจึงนำคำที่ได้ไปรวมกับข้อคำถามเดิมได้เป็นข้อคำถามใหม่ ใช้ในการค้นคืนอีกครั้งแล้วแสดงผลที่ได้กลับสู่ผู้ใช้งาน

กลุ่มที่ 2 จะไม่ใช้วิธีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นในการค้นคืนแบบย้อนกลับโดยการหาคำที่จะนำมาปรับเปลี่ยนในข้อคำถามจากคำที่ปรากฏในเอกสารที่ค้นคืนมาได้ทั้งหมดของหน้าเว็บเพจแล้วเพิ่มคำที่ได้เข้าไปในข้อคำถามใหม่จากนั้นทำการค้นคืนอีกครั้งแสดงผลที่ได้กลับสู่ผู้ใช้งาน

9. กรณีตัวอย่าง

เพื่อความเข้าใจในวิธีการค้นคืนย้อนกลับสารสนเทศแบบเว็บตามที่ได้กล่าวมาข้างต้น ในหัวข้อนี้จะนำเสนอกรณีตัวอย่างในการค้นคืนสารสนเทศแบบเว็บซึ่งแบ่งเป็น 2 กรณี คือ

กรณีที่ 1 ใช้วิธีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็นในการค้นคืนย้อนกลับ

กรณีที่ 2 ไม่ใช้วิธีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็นในการค้นคืนย้อนกลับ

9.1) กรณีที่ 1 ใช้วิธีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในการค้นคืนแบบย้อนกลับ

- สมมุติให้เอกสารแบบเว็บเพจทั้งหมดในฐานข้อมูลมี 30 เว็บเพจ

- ทำการค้นคืนสารสนเทศครั้งแรกแบบธรรมดาซึ่งเวคเตอร์ของข้อคำถามแรกและผลการค้นคืนที่ได้แสดงในตารางที่ 1

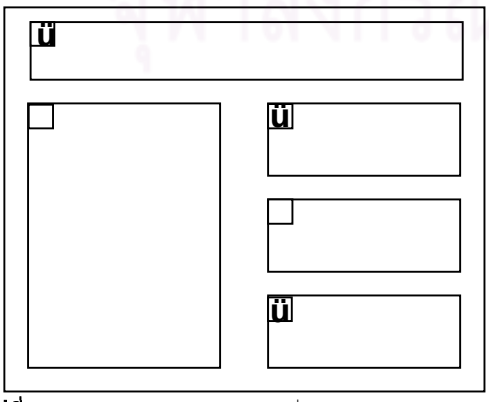
- สมมุติให้เอกสารแบบเว็บเพจที่ค้นคืนมาได้ในครั้งแรกมี 20 เว็บเพจ แต่ที่ผู้ใช้ระบุว่าตรงตามต้องการมี 6 เว็บเพจ คือ Doc 1, Doc 2, Doc 3,Doc 11,Doc 13,Doc 15

ตารางที่ 1. แสดงเวคเตอร์ข้อคำถามแรกและผลการค้นคืนที่ได้

	Term1	Term2	Term3	Term4	Term5
Q	2	1	0	0	0

Doc 1	3	2	2	1	1
Doc 2	2	2	1	0	0
Doc 3	2	2	3	2	0
Doc 4	0	2	2	3	1
Doc 5	1	3	0	0	0
Doc 6	2	0	3	1	1
Doc 7	1	0	1	2	1
Doc 8	0	1	3	2	0
Doc 9	2	0	2	2	0
Doc10	1	0	3	1	1
Doc 11	1	1	0	1	1
Doc 12	1	1	2	0	3
Doc 13	3	2	1	1	0
Doc 14	0	1	0	2	2
Doc 15	1	1	3	0	1
Doc 16	3	0	1	0	3
Doc 17	0	2	2	1	1
Doc 18	2	0	3	0	0
Doc 19	0	0	3	2	1
Doc 20	0	2	1	3	0

จากนั้นทำการแบ่งเอกสารเหล่านี้ออกเป็นส่วนย่อยด้วย วีไอพีเอสอัลกอริทึม เพื่อให้ผู้ใช้เลือกเฉพาะบล็อกที่เห็นว่าเกี่ยวข้องกับที่ต้องการ ซึ่งตัวอย่างของหน้าเว็บเพจที่ถูกแบ่งเป็นบล็อกแล้วนั้นสามารถแสดงได้ดังรูปที่ 6



รูปที่ 6. แสดงตัวอย่างหน้าเว็บเพจที่ถูกแบ่งเป็นบล็อก

เลือกคำที่จะมากำหนดข้อคำถามใหม่จากบล็อกที่ผู้ใช้เลือกแล้วว่าเกี่ยวข้องกับที่ต้องการ โดยเปลี่ยนแปลงค่าน้ำหนักของคำใช้สูตรหาค่าน้ำหนักของคำตามสมการที่ (2) แล้วนำ มารวมกับข้อคำถามเดิมก็จะได้ข้อคำถามใหม่ซึ่งใช้ในการค้นคืนย้อนกลับ สมมุติให้คำที่เลือกมาจากแต่ละบล็อกของทั้ง3 เว็บเพจมีค่าน้ำหนักดังตารางที่ 2

ตารางที่ 2. แสดงค่าน้ำหนักของคำที่ได้มาจากบล็อก

Term	Weight
T1	0.301
T2	0.477
T3	0.602
T4	0.698
T5	0.698
T6	0.301
T7	0.477
T8	0.602

เลือกคำที่มีค่าน้ำหนักมากกว่า 0.65 มาใช้ในการกำหนดข้อคำถามใหม่ ดังนั้นก็จะได้คำที่จะมากำหนดข้อคำถามใหม่ 2 คำ คือ T4 และ T5 แล้วทำการค้นคืนย้อนกลับด้วยข้อคำถามใหม่พร้อมทั้งเปลี่ยนแปลงค่าน้ำหนักของคำด้วยสมการที่ (3) ซึ่งค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการกับข้อคำถามเดิมแสดงในตารางที่ 3 และค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการ กับข้อคำถามใหม่แสดงในตารางที่ 4 แล้วเปรียบเทียบผลการค้นคืนที่ได้ด้วยค่าความถูกต้อง

- สมมุติให้ ค่าน้ำหนักของคำในข้อคำถามเดิมเป็นดังนี้ Term 1 = 0.698, Term 2 = 0.778

- สมมุติให้ ค้นคืนย้อนกลับเอกสารได้ 15 เว็บเพจ ซึ่งผู้ใช้ระบุว่าตรงตามต้องการ 8 เว็บเพจ

ตารางที่ 3. แสดงค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการกับข้อคำถามเดิม

Similarity	Doc1	Doc2	Doc3	Doc11	Doc13	Doc15
Q	0.99	0.94	0.94	0.95	0.99	0.95

ตารางที่ 4. แสดงค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการกับข้อคำถามใหม่

Similarity	Doc1	Doc2	Doc3	Doc4
Q'	2.87	2.17	2.87	2.17
	Doc11	Doc12	Doc13	Doc15
	2.17	2.17	2.87	2.17

ตารางที่ 5. แสดงค่าความถูกต้อง

Query	Precision	% Improvement
Q	6/20 = 0.3	76.66%
Q'	8/15 = 0.53	

จากตารางที่ 5 แสดงให้เห็นว่า ค่าความถูกต้องของการค้นคืนย้อนกลับโดยใช้วีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็นเพิ่มขึ้น จากเดิมที่ไม่ได้ทำการค้นคืนย้อนกลับมีค่า 0.3 เป็น 0.53 ซึ่งมีประสิทธิภาพเพิ่มขึ้นถึง 76.66%

9.2) **กรณีที่ 2 ไม่ใช้วีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็น ในการค้นคืนแบบย้อนกลับ**

- สมมุติให้ข้อคำถามที่ใช้ในการค้นคืนและผลการค้นคืนครั้งแรกที่ได้นั้นเหมือนกับกลุ่มที่ 1

- เลือกคำที่จะมากำหนดข้อคำถามใหม่จากเอกสารแบบเว็บเพจทั้ง 6 เว็บเพจที่ผู้ใช้ระบุว่าเกี่ยวข้องจากทุกส่วนของหน้าเว็บเพจโดยสมมุติให้คำที่จะใช้เพิ่มเข้าไปในข้อคำถามคือ T3 และ T5 ซึ่งมีค่าน้ำหนักของคำเท่ากับ 3

แสดงข้อคำถามใหม่ที่ได้ ในตารางที่ 6 แล้วทำการค้นคืนย้อนกลับโดยไม่มีการเปลี่ยนแปลงค่าน้ำหนักของคำใดๆทั้งสิ้น แสดงค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการกับข้อคำถามใหม่ ได้ในตารางที่ 7 และค่าความถูกต้องในตารางที่ 8

- สมมุติให้ ค้นคืนย้อนกลับเอกสารได้ 20 เว็บเพจ ซึ่งผู้ใช้ระบุว่าตรงตามต้องการ 8 เว็บเพจ

ตารางที่ 6. เวคเตอร์ของข้อคำถามใหม่

	Term1	Term2	Term3	Term4	Term5
Q'	2	1	3	0	3

ตารางที่ 7. แสดงค่าความคล้ายระหว่างเอกสารที่ผู้ใช้ระบุว่าตรงตามต้องการกับข้อคำถามใหม่

Similarity	Doc 1	Doc 3	Doc 6	Doc 10
Q'	0.85	0.75	0.94	0.93
	Doc12	Doc15	Doc16	Doc18
	0.97	0.90	0.86	0.75

ตารางที่ 8. แสดงค่าความถูกต้อง

Query	Precision	% Improvement
Q	6/20 = 0.3	33.33%
Q'	8/20 = 0.4	

จากตารางที่ 8 แสดงให้เห็นว่า ค่าความถูกต้องของการค้นคืนย้อนกลับโดยไม่ใช้วีไอพีเอสอัลกอริทึม และแบบจำลองความน่าจะเป็นนั้นพิ่มขึ้น จากเดิมที่ไม่ได้ทำการ ค้นคืนย้อนกลับมีค่า 0.3 เป็น 0.4 ซึ่งมีประสิทธิภาพเพิ่มขึ้นเพียง 33.33%

9.3) **เปรียบเทียบผลการค้นคืนย้อนกลับที่ได้จากทั้ง 2 กรณีตัวอย่าง**

ตารางที่ 9. แสดงค่าความถูกต้องของ 2 กรณีตัวอย่าง

Group	Precision	% Improvement
1	8/15 = 0.53	76.66%
2	8/20 = 0.4	33.33%

จากตารางที่ 9 แสดงให้เห็นว่าค่าความถูกต้องที่ได้จากการค้นคืนย้อนกลับ ด้วยการใช้วีไอพีเอสอัลกอริทึมและแบบจำลองความน่าจะเป็นนั้น สามารถเพิ่มประสิทธิภาพในการค้นคืนได้มากกว่าการค้นคืนย้อนกลับแบบที่ไม่ได้ใช้ถึง 43.33%

10. สรุปผลการวิจัยและการดำเนินการในอนาคต

งานวิจัยนี้นำเสนอวิธีการในการค้นคืนย้อนกลับสารสนเทศแบบเว็บโดยการแบ่งเว็บเพจเป็นส่วนย่อยและใช้แบบจำลองความน่าจะเป็น โดยในส่วนของการค้นคืนย้อนกลับนั้นผู้ใช้จะมีส่วนร่วมในการเลือกบล็อกของเว็บเพจที่ถูกแบ่งออกเป็นส่วนย่อยด้วยวีไอพีเอสอัลกอริทึม เพื่อใช้ในการเลือกคำที่จะนำมาปรับเปลี่ยนในการกำหนด

ข้อคำถามใหม่ และในส่วนของแบบจำลองความน่าจะเป็นนั้นจะใช้ในการเปลี่ยนแปลงค่าน้ำหนักของคำก่อนทำการค้นคืนย้อนกลับ ข้อดีของงานวิจัยนี้คือ ส่วนของเว็บเพจที่ไม่เกี่ยวข้องกับที่ผู้ใช้ต้องการไม่ถูกเลือกมากำหนดข้อคำถามใหม่เพราะผู้ใช้ระบุไว้แล้วว่าส่วนไหนบ้างของหน้าเว็บเพจที่เกี่ยวข้องกับที่ต้องการ และทำให้ผลการค้นคืนย้อนกลับที่ได้ตรงตามความต้องการของผู้ใช้มากขึ้นอีกด้วย แต่อย่างไรก็ตาม ในส่วนของการทดลอง และผลการทดลองซึ่งถือว่าเป็นส่วนสำคัญที่สุดนั้นงานวิจัยนี้จะดำเนินการต่อไปในอนาคต

11. บรรณานุกรม

[1] D. Harman, “Relevance feedback revisited”, Paper presented at 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Chicago, IL, 1992.

[2] G. Salton, C. Buckley, “Improving retrieval performance by relevance feedback”, Journal of the American Society for Information Science, 1990, 288-297.

[3] G. Salton, C. Buckley, “Introduction to Modern Information Retrieval”, New York: McGraw-Hill, 1983.

[4] J. Chen, B. Zhou, J. Shi, H. Zhang, and F. Qiu, “Function-Based Object Model towards Website Adaptation”, in the proceedings of the 10th World Wide Web conference (WWW10), Budapest, Hungary, May 2001.

[5] Ji-Rong Wen Shipeng Yu, Deng Cai and Wei-Ying Ma, “VIPS: a vision-based page segmentation algorithm”, Microsoft Technical Report, MSR-TR-2003-79, 2003.

[6] Ji-Rong Wen Shipeng Yu, Deng Cai and Wei-Ying Ma, “Improving pseudo-relevance feedback in web information retrieval using web page segmentation”, In Proceedings of the 12th International World Wide Web Conference, WWW2003, 2003, pages 11–18.

[7] M. Kovacevic, M. Diligenti, M. Gori, and V. Milutinovic, “Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification”, in the proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December, 2002.

[8] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, “Modern Information Retrieval”, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1999.

[9] Seung Yeol and Achim Hoffman , “Pseudo-Relevance Feedback in Web information Retrieval Using Segments Subjective Importance Value”, In Proceedings of International Conference on Web Intelligence, 2005.

ประวัติผู้เขียนวิทยานิพนธ์

นางสาวปรารถนา จันพลโท เกิดวันที่ 23 ธันวาคม พ.ศ. 2524 ที่จังหวัดภูเก็ต สำเร็จการศึกษาระดับปริญญาวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตภูเก็ต ในปีการศึกษา 2546 และเข้าศึกษาต่อในหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2547