รายงานโครงการวิจัยประจำปีงบประมาณ 2551

เรื่อง

การศึกษาเปรียบเทียบประสิทธิภาพของเทคนิควิธีการขยายคำสืบค้นและการวิเคราะห์เชื่อมโยง

Comparative Study Effectiveness of Query Expansion and Link Analysis

โดย

พรฤดี เนติโสภากุล

และ

วิริศ ลีลาภัทร

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520

พ.ศ. 2552

รายงานโครงการวิจัยประจำปีงบประมาณ 2551

เรื่อง

การศึกษาเปรียบเทียบประสิทธิภาพของเทคนิคการขยายคำสืบค้นและการวิเคราะห์เชื่อมโยง

Comparative Study Effectiveness of Query Expansion and Link Analysis

โดย

พรฤดี เนติโสภากุล

และ

วิริศ ลีลาภัทร

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ 10520

พ.ศ. 2552

การศึกษาเปรียบเทียบประสิทธิภาพของเทคนิควิธีการขยายคำสืบค้นและการวิเคราะห์เชื่อมโยง

Comparative Study Effectiveness of Query Expansion and Link Analysis

พรฤดี เนติโสภากุล และ วิริศ ลีลาภัทร

บทคัดย่อ

เทคนิคดั้งเดิมที่ใช้ในการปรับปรุงประสิทธิภาพของโปรแกรมค้นหาบนอินเตอร์เน็ต คือ เทคนิคการขยายคำสืบค้น แนวคิดหลักของเทคนิคในกลุ่มนี้ คือ การเพิ่มจำนวนหน้าเว็บที่เกี่ยวข้องโดยการเพิ่มเติมหรือผสมคำสืบค้นใหม่ กับคำสืบค้นเริ่มต้นของผู้ใช้ ต่อมา มีการนำเสนอเทคนิคการวิเคราะห์เชื่อมโยง เพื่อใช้ในงานจัดกลุ่มหน้าเว็บที่คล้ายคลึงกัน แนวคิดหลักของการวิเคราะห์เชื่อมโยง คือการดึงหน้าเว็บที่เกี่ยวข้องจากลิงค์ที่ฝังตัวอยู่ในหน้านั้นๆ ส่วนงานวิจัยนี้ ได้นำเสนอวิธีการผสมผสานเทคนิควิเคราะห์เชื่อมโยง กับ เทคนิคการขยายคำสืบค้นแบบดั้งเดิม เพื่อศึกษาประสิทธิภาพของการสืบค้น โดยมีสองส่วนงานหลักคือ ส่วนแรกเป็นการออกแบบและจัดสร้างโปรแกรมสืบค้นที่รวมเอาเทคนิคทั้งสองเข้าด้วยกัน และส่วนที่สองเป็นการทดลองเพื่อเปรียบเทียบประสิทธิภาพการสืบค้นด้วยเทคนิคต่างๆ

Abstract

Traditional techniques to improve effectiveness of internet search engine are query expansion techniques. The main idea of this class of techniques is to retrieve more relevant webpages by adding or combining new search terms to origin search terms. Recently, a link analysis technique has been proposed to classify webpages into closely related groups. The main idea of a link analysis is to retrieve more relevant webpages by following links embedded in the webpages. This research proposed to combine a link analysis technique to relevant feedback query expansion techniques in order to study the impact on their search efficiencies. There are two main tasks: the first one is to design and implement a new search engine which incorporated link analysis and query expansion, and the second one is to conduct an experiment to compare the effectiveness of various modes of proposed techniques used to improve the search engine.

สารบัญ

หน้า

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

โปรแกรมสืบค้นทางอินเตอร์เน็ตเป็นเครื่องมือที่สำคัญในชีวิตประจำวันไปแล้ว อย่างไรก็ตาม ผลลัพธ์ของการสืบค้นบนเครือข่ายอินเตอร์เน็ตมีปริมาณที่มาก และอาจไม่ตรงกับความต้องการที่แท้จริงของผู้สืบค้นที่ไม่เชี่ยวชาญในการสืบค้น การใช้โปรแกรมสืบค้นในเบื้องต้นนั้นจะต้องอาศัยการป้อนคำสืบค้นเป็นหลัก ซึ่งมีข้อจำกัดหลายประการด้วยกัน คือ คำที่ป้อนเพื่อสืบค้นอาจไม่ปรากฏในหน้าเว็บที่เกี่ยวข้อง ทำให้ได้ผลลัพธ์ของการค้นคืนที่ไม่ครบถ้วน หรือ อาจไปปรากฏในหน้าเว็บที่ไม่เกี่ยวข้องก็ได้ ทำให้ผลลัพธ์ที่ได้มีทั้งเว็บเพจที่เกี่ยวข้อง และเว็บเพจที่ไม่เกี่ยวข้องทำให้การค้นคืน ดังนั้น การสืบค้นโดยใช้คำสืบค้น ยังให้ผลลัพธ์ไม่ดีเท่าที่ควร ทั้งนี้สาเหตุหลักที่สำคัญประการหนึ่งคือความคลุมเครือของความหมายของคำที่เกิดขึ้นในภาษาธรรมชาตินั่นเอง

เทคนิคดั้งเดิมที่ใช้ในปรับปรุงประสิทธิภาพการสืบค้น มักเกี่ยวข้องกับงานวิจัยด้านการขยายคำสืบค้น (Query Expansion Techniques) ซึ่งเป็นเทคนิคที่ใช้ในการปรับปรุงคำสืบค้นของผู้ใช้ โดยการเพิ่มคำใหม่ที่มีความหมายใกล้เคียงกันหรือคล้ายคลึงกัน ลงในคำสืบค้นเดิมที่มีอยู่เพื่อเพิ่มโอกาสในการค้นคืนเอกสารที่เกี่ยวข้องเพิ่มมากยิ่งขึ้น โดยคำใหม่ที่เพิ่มลงไปนั้นอาจนำมาจากสองแหล่งด้วยกันคือ แหล่งที่หนึ่ง ใช้พจนานุกรมคำศัพท์ที่มีความหมายใกล้เคียงกัน (Thesaurus) และแหล่งที่สอง นำมาจากคำที่ปรากฏร่วมในเอกสารที่ถูกค้นคืนมาในครั้งแรกที่ปรากฏอยู่ในอันดับต้นๆ หรือถูกระบุโดยผู้ใช้งานว่า มีความเกี่ยวข้องกับสิ่งที่ต้องการ หรือที่เรียกว่า Relevance Feedback Techniques อย่างไรก็ตาม เทคนิคเหล่านี้ ยังไม่สามารถแก้ปัญหาความคลุมเครือของคำสืบค้นที่มีหลายความหมายได้ นอกจากนี้แล้ว เทคนิคการขยายคำสืบค้น อาจสามารถใช้ได้ดีพอควรกับระบบเอกสารแบบดั้งเดิมที่มีจำนวนเอกสารจำกัด แต่เมื่อนำมาประยุกต์ใช้กับการค้นคืนเว็บเพจ ส่งผลให้ประสิทธิภาพการค้นคืนลดลงอย่างมาก เหตุผลหลักเนื่องจากว่ามีข้อมูลจำนวนมากที่ไม่เกี่ยวข้องที่ปรากฏอยู่ในเว็บเพจ เช่น โฆษณาต่างๆ แถบเมนูนำทาง ซึ่งทำให้เกิดโอกาสบิดเบือนข้อมูลสถิติของคำที่มีความคล้ายคลึงกัน และปรากฏร่วมกัน จึงส่งผลต่อประสิทธิภาพการทำงานของเทคนิคการขยายคำสืบค้น

ในปี ค.ศ. 1998 [Kleinberg 1998] ได้นำเสนอแนวคิดในการนำข้อมูลที่ฝังอยู่ในหน้าเว็บ คือ ข้อมูลการเชื่อมโยงหน้าเว็บ หรือ ลิงค์ มาใช้ในการวิเคราะห์ความสำคัญของหน้าเว็บว่าเกี่ยวข้องกับหัวเรื่องที่ต้องการหรือไม่ และนำเสนออัลกอริทึมในการทำงานที่เรียกว่า Hypertext Inductive

Topic Selection หรือย่อๆ ว่า HITS ซึ่งต่อมา เทคนิคนี้ ถูกจัดเป็นหนึ่งในสองของ เทคนิคการวิเคราะห์เชื่อมโยง (Link Analysis) ที่เป็นที่รู้จักกันดี ซึ่งต่อไปนี้ จะขอเรียกว่า เทคนิคการวิเคราะห์ลิงค์ และได้รับการนำไปประยุกต์ใช้งานด้านต่างๆ ได้แก่ การจัดกลุ่มหน้าเว็บเพจอัตโนมัติ การค้นหาชุมชนออนไลน์ เป็นต้น

ส่วนอีกด้านหนึ่ง เทคนิคการวิเคราะห์เชื่อมโยง อีกเทคนิคหนึ่ง คือ อัลกอรึทึม PageRank ก็ได้รับการยอมรับจากโปรแกรมสืบค้นชื่อดัง คือ google เพื่อใช้ในการจัดลำดับความสำคัญของเซตของหน้าเว็บที่ได้จากการค้นคืนมา ซึ่งนับว่าประสบความสำเร็จเป็นอย่างดี

ในงานวิจัยนี้ จึงนำเสนอแนวทางในการปรับปรุงประสิทธิภาพเทคนิคการขยายคำสืบค้นแบบดั้งเดิม โดย การผสมผสานเอาอัลกอริทึม HITS เข้ามาใช้งาน

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

งานวิจัยนี้มีวัตถุประสงค์ เพื่อศึกษาวิธีการปรับปรุงประสิทธิภาพการสืบค้น โดยนำเทคนิคการวิเคราะห์ลิงค์มาใช้ปรับปรุงเทคนิคขยายคำสืบค้นแบบดั้งเดิม

1.3 สมมติฐานของการศึกษา

1.3.1 ระบบฐานในการทดสอบ (Baseline) คือ ใช้โปรแกรมสืบค้น yahoo โดยเรียกใช้งานผ่าน yahoo API โดยคัดเลือก 30 อันดับแรกของการค้นคืน มาใช้เป็นฐานในการทดลองเปรียบเทียบ

1.3.2 มีการแบ่งเทคนิคการขยายคำสืบค้นแบบ relevance feedback ออกเป็น สองแบบย่อย คือ แบบที่มีผู้ใช้งานเป็นผู้เลือกหน้าเว็บที่เกี่ยวข้อง เพื่อป้อนกลับ ซึ่งจะเรียกว่าวิธี Interactive Query Expansion หรือ ย่อว่า IQE และแบบที่ให้โปรแกรมเลือกหน้าเว็บที่เกี่ยวข้องให้โดยอัตโนมัติ โดยเลือกจากลำดับต้นๆ ที่ค้นคืนมาได้ 30 อันดับแรก ซึ่งจะเรียกว่าวิธี Automatic Query Expansion หรือ ย่อว่า AQE

1.3.3 มีการปรับปรุงเทคนิคการขยายคำสืบค้นแบบ IQE และ AQE ข้างต้นด้วยเทคนิคการวิเคราะห์ลิงค์ และทำการทดสอบเปรียบเทียบประสิทธิภาพโดยมีสมมุติฐานว่า เทคนิคที่มีการปรับปรุงแล้วจะให้ประสิทธิภาพดีกว่าเดิม

1.3.4 ตัววัดประสิทธิภาพที่ใช้ คำนวณจากค่า F-measure ซึ่งได้จากค่า Precision และค่า Recall ดังจะกล่าวต่อไปในบทที่ 2

1.4 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

การวิเคราะห์ลิงค์ (Link Analysis) เป็นการวิเคราะห์โครงสร้างของการเชื่อมโยงกันระหว่างเว็บเพจบนอินเทอร์เน็ต โดยสามารถแทนภาพการเชื่อมโยงเป็นกราฟแบบมีทิศทาง (Directed Graph) เพื่อความสะดวกในการคำนวณค่า อาจแทนกราฟนี้ด้วยเมตริกซ์ ฮิทส์อัลกอริทึม (HITS algorithm) เป็นอัลกอริทึมหนึ่งในเทคนิคการวิเคราะห์ลิงค์ [Kleinberg 1998] ที่มีสมมุติฐานว่า หน้าเว็บเพจใดมีการเชื่อมโยงไปยังหน้าเว็บเพจที่เกี่ยวข้องจำนวนมาก หน้านั้นจัดว่าเป็น "ฮับ" (Hub) ของหัวเรื่องนั้นๆ หมายถึง หน้าฮับเป็นหน้าที่รวบรวมเอาแหล่งข้อมูลที่เกี่ยวข้องกับหัวเรื่องนั้นๆ จำนวนมากเข้าด้วยกัน ส่วนหน้าเว็บเพจใดถูกอ้างอิง หรือถูกชี้มาจากหลายๆหน้าฮับ หน้านั้นจัดเป็นหน้า "ออธอริที" (Authority) หมายถึง เป็นหน้าที่เป็นแหล่งข้อมูลที่มีสำคัญ น่าเชื่อถือได้ การคำนวณน้ำหนักของหน้าเว็บต่างๆ จะคำนวณจากกลุ่มของหน้าเริ่มต้นกลุ่มหนึ่งที่เรียกว่ารูจเพจ (root pages) ซึ่งเบื้องต้นได้มาจากผลลัพธ์การค้นคืนของเสิร์จเอ็นจิ้น โดยกลุ่มหน้าที่อยู่ในรูทเพจเหล่านี้ ไม่มีความจำเป็นต้องเป็นออธอริทีเพจทุกหน้า จากนั้น นำกลุ่มหน้าเหล่านี้มาขยายให้ใหญ่ขึ้น โดยรวมเอาหน้าที่รูทเพจ ชี้ไปหา และ หน้าที่ชี้มายังรูทเพจ มาอยู่ในเซตที่เรียกว่าเบสเซต (base set) ซึ่ง subgraph ที่สร้างมาจากเบสเซตนั้น จะไม่รวมเอาการเชื่อมโยงที่อยู่ในเว็บโดเมนเดียวกันเนื่องจากการเชื่อมโยงเหล่านั้น ส่วนใหญ่ใช้ในการนำทางภายในโดเมนเท่านั้น ผลลัพธ์ที่ได้จะเป็น subgraph ที่เป็นชุดของหน้าที่เชื่อมโยงกันอย่างหนาแน่น แสดงถึงความเกี่ยวข้องกันในหัวข้อที่ต้องการค้นหา

อัลกอริธึมฮิทส์ เป็นอัลกอริธึมที่ทำงานแบบวนซ้ำ โดยแต่ละหน้า จะมีการคำนวณค่าน้ำหนักสองค่า คือ ค่าฮับ และ ค่าออธอริที ดังที่ได้กล่าวมาแล้ว โดยจะกำหนดค่าเริ่มต้นเป็นค่าที่เท่าๆกันหมดก่อน แล้วจึงปรับปรุงค่าดังกล่าว โดยดูความหนาแน่นของการเชื่อมโยงใน subgraph ที่ได้ ค่าน้ำหนักฮับของแต่ละหน้าได้จากผลรวมค่าน้ำหนักของหน้าออธอริที ที่หน้านั้นๆมีการเชื่อมโยงไปในทางกลับกัน ค่าน้ำหนักออธอริที จะได้จากผลรวมของค่าน้ำหนักของหน้าฮับที่ชี้มายังหน้านี้ ในทุกรอบของการคำนวณ จะมีการทำนอลมอไลเซชั่น เพื่อให้ค่าน้ำหนักทั้งหมดของทุกหน้ารวมกันเท่ากับหนึ่ง หรือ หนึ่งร้อยเปอร์เซ็นต์นั่นเอง การทำวนซ้ำ จะดำเนินไปจนกว่า ค่าทั้งสองของทุกหน้า ไม่มีการเปลี่ยนแปลง

ข้อมูลวิเคราะห์ลิงค์ มักมีการใช้งานร่วมกับเนื้อหาในหน้าเว็บ เมื่อประยุกต์ใช้กับงานต่างๆ เช่น การสร้างเครื่องจักรสืบค้น (Search Engine) การค้นหาชุมชนไซเบอร์ (Cybercommunities) การแบ่งกลุ่มหรือการจำแนกหน้าเว็บต่างๆให้เป็นโครงสร้างเชิงลำดับชั้นโดยอัตโนมัติ (Webpage Taxonomy Construction) การวิเคราะห์การอ้างอิงเอกสาร (Citation Analysis) ในงานวิจัยนี้ เป็นการนำเสนอสถาปัตยกรรมที่นำเอาเทคนิคการวิเคราะห์ลิงค์มาปรับปรุงข้อจำกัดของเทคนิคการขยายคำสืบค้นแบบดั้งเดิม

1.5 ขอบเขตการวิจัย

งานวิจัยนี้แบ่งเป็นสองส่วนคือ ส่วนของการปรับปรุงอัลกอริทึมในการขยายคำสืบค้น และส่วนของการทดสอบประสิทธิภาพของระบบที่ปรับปรุง

1.5.1 นำเทคนิคการวิเคราะห์ลิงค์แบบ อัลกอริทึม HITS มาปรับปรุงเทคนิคการขยายคำสืบค้นแบบ relevance feedback

1.5.2 ทำการทดลอง เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการขยายคำสืบค้นแบบที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ กับ แบบที่ยังไม่ได้ทำการปรับปรุง โดยใช้โจทย์การสืบค้นจาก ชุดคำถามที่นิยมใช้ในการวิจัยทางด้านการสืบค้นระบบสารสนเทศ (Text REtrieval Conference) ได้แก่ TREC1, TREC-4 และ TREC-2003 โดยเลือกคำถามที่อยู่ในโดเมนวิทยาศาสตร์และเทคโนโลยี

1.6 ขั้นตอนของการวิจัย

ขั้นตอนในการวิจัยมีดังต่อไปนี้

1.6.1 ศึกษารายละเอียดของเทคนิคการขยายคำสืบค้นและเทคนิคการวิเคราะห์ลิงค์, ศึ กษาวิธีการประเมินประสิทธิภาพของระบบค้นคืนสารสนเทศ และศึกษางานวิจัยต่างๆที่เกี่ยวข้อง

1.6.2 วิเคราะห์ความเป็นไปได้ ของแนวคิดในการนำเอาเทคนิคการขยายคำสืบค้นมาทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

1.6.3 ออกแบบสถาปัตยกรรม เว็บเสิร์จเอนจิ้นแบบเทคนิคการขยายคำสืบค้นทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์และกลไกการทำงานต่างๆ

1.6.4 ออกแบบการทดลองเพื่อใช้ประเมินประสิทธิภาพการค้นคืนระหว่างเทคนิคการขยายคำสืบค้นแบบที่ไม่ได้ปรับปรุงและแบบที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์

1.6.5 ทำการทดลองและประเมินผลที่ได้จากการทดลอง

1.6.6 จัดทำเอกสารสรุปการทำโครงการ

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1.7.1 ได้แนวทางในการปรับปรุงประสิทธิภาพของเทคนิคการขยายคำสืบค้นโดยนำเทคนิคการวิเคราะห์ลิงค์เข้าช่วย

1.7.2 ได้องค์ความรู้ใหม่ในเชิงของการเปรียบเทียบประสิทธิภาพดังนี้

1. เปรียบเทียบระหว่าง 2 วิธีการหลักของเทคนิคการขยายคำสืบค้น

- เทคนิคการขยายคำสืบค้นแบบอัตโนมัติเปรียบเทียบกับเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์

2. เปรียบเทียบ 2 วิธีการหลักของเทคนิคการขยายคำสืบค้นกับเทคนิคผสมระหว่าง 2 วิธีการหลักของการขยายคำสืบค้นทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

- เทคนิคการขยายคำสืบค้นแบบอัตโนมัติเปรียบเทียบกับเทคนิคการขยายคำสืบค้นแบบอัตโนมัติทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

- เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์เปรียบเทียบกับเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

3. เปรียบเทียบเทคนิคการวิเคราะห์ลิงค์กับเทคนิคผสมระหว่าง 2 วิธีการหลักของเทคนิคการขยายคำสืบค้นทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

- เทคนิคการวิเคราะห์ลิงค์เปรียบเทียบกับเทคนิคการขยายคำสืบค้นแบบอัตโนมัติทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

- เทคนิคการวิเคราะห์ลิงค์เปรียบเทียบกับเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

1.8 เครื่องมือที่ใช้

1.8.1 PHP

1.8.2 Apache Web Server

1.8.3 MySQL

1.8.4 Zend Studio

1.8.5 Adobe Macromedia Dreamweaver CS3

1.8.6 Yahoo API

บทที่ 2
ทฤษฎีพื้นฐานที่ใช้ในการวิจัย และงานวิจัยที่เกี่ยวข้อง

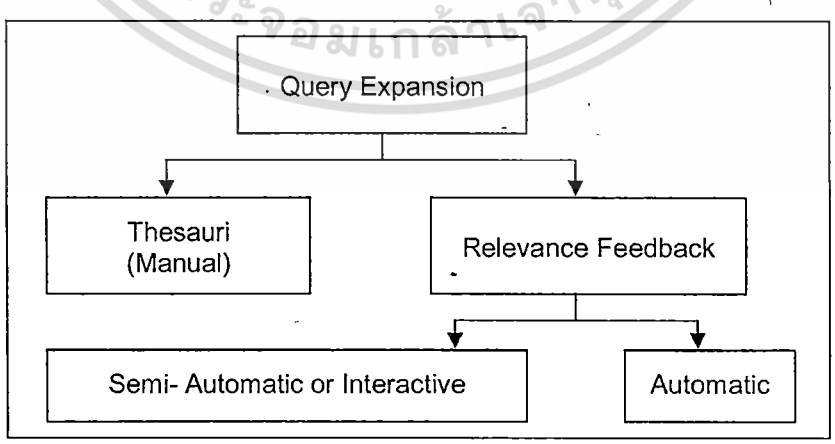
ในบทนี้จะกล่าวถึงเทคนิคที่ใช้ในการปรับปรุงประสิทธิภาพการค้นคืนของเว็บเสิร์จเอนจิ้น ซึ่งได้แก่ เทคนิคการขยายคำสืบค้นและเทคนิคการวิเคราะห์ลิงค์ นอกจากนี้แล้วกล่าวถึงวิธีการประเมินประสิทธิภาพการค้นคืน

2.1 เทคนิคการปรับปรุงประสิทธิภาพการค้นคืนของเสิร์จเอนจิ้น

2.1.1 เทคนิคการขยายคำสืบค้น (Query Expansion Techniques)

เทคนิคการขยายคำสืบค้น [Jian-Fu et. al. 2005, Seher 2006] เป็นวิธีการหนึ่งที่นำมาใช้ปรับปรุงการค้นคืนเว็บเพจหรือเอกสารให้มีประสิทธิภาพมากยิ่งขึ้นโดยการเพิ่มคำใหม่ที่มีความหมายใกล้เคียงกันลงในคำสืบค้นเดิมที่มีอยู่ซึ่งวัตถุประสงค์ของการเพิ่มคำเพื่อปรับปรุงความแม่นยำหรือความครอบคลุมของผลลัพธ์

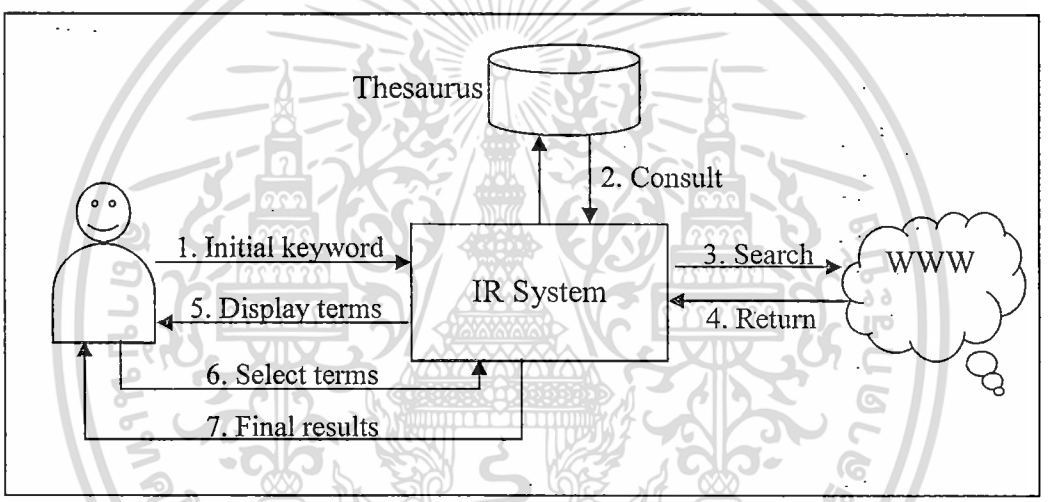
เทคนิคการขยายคำสืบค้นที่มีใช้งานในปัจจุบันสามารถแบ่งตามวิธีการในการหาคำศัพท์ใหม่ได้ 2 วิธีด้วยกันคือ การหาคำศัพท์ใหม่จากพจนานุกรมของคำที่มีความหมายใกล้เคียงกันหรือที่เรียกว่า Thesaurus [Jian-Fu et. al. 2005, Crouch 1990] และการหาคำศัพท์ใหม่จากคำที่ปรากฏในเอกสารที่ถูกค้นคืน [Jian-Fu et. al. 2005, Seher 2006] ซึ่งสามารถแบ่งย่อยได้อีกสองวิธีคือ แบบอัตโนมัติ (Automatic) และแบบกึ่งอัตโนมัติ (Semi-Automatic or Interactive) ดังที่แสดงในรูปที่ 2.1 ซึ่งมีรายละเอียดดังต่อไปนี้



รูปที่ 2.1 เทคนิคการขยายคำสืบค้นแบ่งตามวิธีการหาคำศัพท์ [Efthimiadis 1996]

2.1.1.1 เทคนิคการขยายคำสืบค้นโดยใช้พจนานุกรมคำเหมือน

พจนานุกรมคำเหมือนหรือ Thesaurus [Chen et. al. 2003, Jian-Fu et. al. 2005] คือ พจนานุกรมของคำที่มีความหมายพ้องกันหรือมีความหมายใกล้เคียงกัน ในสถานการณ์ส่วนใหญ่ Thesaurus ถูกใช้ในการปรับปรุงประสิทธิภาพในการค้นคืนโดยการขยายคำสืบค้นด้วยคำศัพท์ที่ใกล้เคียงกับคำสืบค้นเริ่มต้น การทำงานของ Thesaurus จะเริ่มหลังจากที่ผู้ใช้ทำการส่งคำสืบค้นไปยังระบบสืบค้นสารสนเทศ ระบบจะทำการค้นหาข้อมูลจาก Thesaurus โดยอัตโนมัติ จากนั้นจะแสดงคำศัพท์ทั้งหมดที่ใกล้เคียงกับคำสืบค้นเริ่มต้นกลับมาให้ผู้ใช้ทำการเลือก ผู้ใช้สามารถทำการเลือกคำที่ใกล้เคียงเพื่อปรับปรุงคำสืบค้นหรือจะไม่เลือกคำใดๆเลยก็ได้ จากนั้นคำสืบค้นที่ถูกปรับปรุงจะถูกส่งกลับไปค้นหาอีกครั้งเอกสารที่เป็นผลลัพธ์จะถูกค้นคืนกลับมา



รูปที่ 2.2 สถาปัตยกรรมของการขยายคำสืบค้นโดยใช้พจนานุกรมคำเหมือน

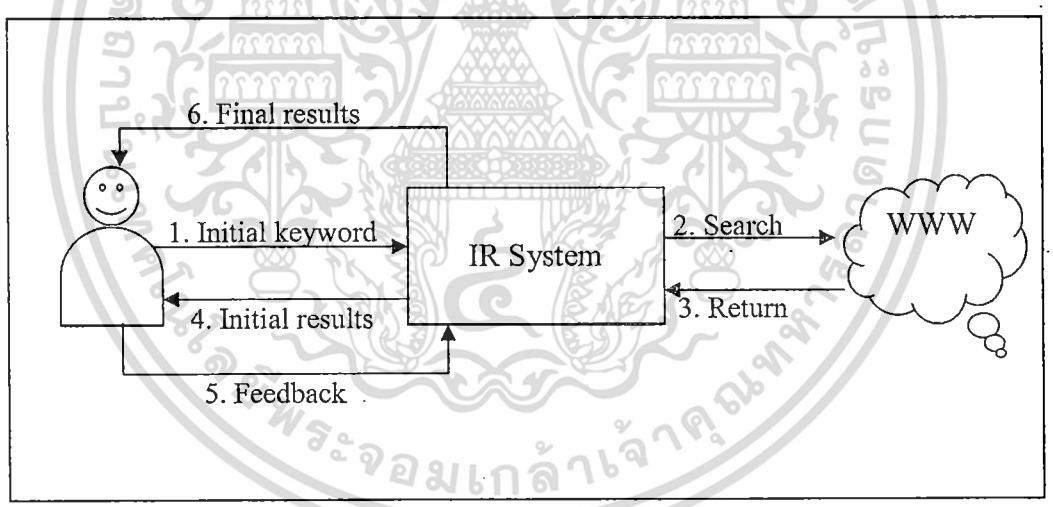
แต่อย่างไรก็ตามเทคนิคการขยายคำสืบค้นโดยใช้พจนานุกรมคำเหมือนยังคงมีข้อเสียบางประการดังนี้ [Jian-Fu et. al. 2005]

1. เนื่องจากพจนานุกรมถูกสร้างขึ้นโดยนักภาษาศาสตร์และผู้เชี่ยวชาญเฉพาะทาง บ่อยครั้งที่ทำให้คำศัพท์ในพจนานุกรมมีความหมายที่กว้างหรือแคบจนเกินไป
2. ในการสร้างพจนานุกรมด้วยคนนั้นค่อนข้างใช้ระยะเวลานาน เพราะต้องพิจารณาทบทวนความรู้ทั้งหมดของโดเมน
3. โครงสร้างของพจนานุกรมที่ถูกสร้างขึ้นอยู่กับลักษณะส่วนตัวของบุคคลที่ทำการสร้าง เช่น ภูมิหลังและประสบการณ์ เป็นต้น
4. เพื่อให้เหมาะกับการเปลี่ยนโดเมนของความรู้ บ่อยครั้งที่ผู้เชี่ยวชาญจะต้องทำการอัพเดทพจนานุกรมอยู่เสมอ และในการอัพเดทนั้นหมายถึงว่าจะต้องมีการปรับปรุงพจนานุกรมทั้งหมด ซึ่งไม่สามารถทำให้สมบูรณ์ได้ด้วยคนเพียงอย่างเดียว

จากข้อเสียบางประการข้างต้นที่กล่าวได้มีการนำเสนอแนวคิดในการปรับปรุงโดยใช้การป้อนกลับความเกี่ยวข้องแบบอัตโนมัติมาทำงานร่วมกับเทคนิคการขยายคำสืบค้นโดยใช้พจนานุกรมคำเหมือน [Jian-Fu et. al. 2005] จุดเด่นของแนวคิดที่นำเสนอนี้คือ ข้อหนึ่งใช้งานได้ง่ายขึ้น ข้อสองสามารถปรับปรุงคำในพจนานุกรมได้โดยอัตโนมัติ ข้อสามโครงสร้างของพจนานุกรมสอดคล้องกับหลักเหตุผลมากขึ้นและข้อสี่ลดระยะเวลาในการค้นคืน

2.1.1.2 การขยายคำสืบค้นโดยใช้การป้อนกลับความเกี่ยวข้อง

การป้อนกลับความเกี่ยวข้องหรือ Relevance Feedback [Harman 1988, Chung and Lee 2004] เป็นวัฏจักรขั้นตอนโดยอาศัยผู้ใช้ป้อนกลับเข้าสู่ระบบเพื่อตัดสินความเกี่ยวข้องของเอกสารที่ถูกค้นคืนและจากนั้นระบบจะทำการประเมินเพื่อดำเนินการแก้ไขการสืบค้น แนวคิดหลักของวิธีการนี้ประกอบไปด้วยการเลือกคำศัพท์ที่มีความสำคัญจากเอกสารที่เกี่ยวข้องและนำคำศัพท์ที่มีการคำนวณค่าน้ำหนักแล้วไปทำการสร้างคำสืบค้นขึ้นมาใหม่ ซึ่งในการป้อนกลับความเกี่ยวข้องมีด้วยกันสองวิธีการย่อย [Seher 2006] ดังนี้



รูปที่ 2.3 สถาปัตยกรรมของการขยายคำสืบค้นโดยใช้การป้อนกลับความเกี่ยวข้อง

1. การขยายคำสืบค้นแบบอัตโนมัติ (Automatic Query Expansion)

การขยายคำสืบค้นแบบอัตโนมัติ หรือ Automatic Query Expansion ซึ่งถูกเรียกสั้นๆ ว่า AQE [Qiu and Frei 1993] โดยวิธีการนี้ระบบสืบค้นสารสนเทศจะทำการประเมินความเกี่ยวข้องของเอกสารที่ถูกค้นคืนให้แบบอัตโนมัติโดยไม่ขึ้นกับผู้ใช้ แนวคิดของวิธีการนี้อยู่บนสมมติฐานที่ว่าลำดับสูงสุด n ลำดับของเอกสาร (เช่น เอกสาร 20 ลำดับแรก) ที่ถูกค้นคืนในครั้งแรกและมีความเกี่ยวข้อง ซึ่งการเพิ่มคำศัพท์จะถูกเลือกจะใช้หลักการการเรียนรู้ทางสถิติจากลำดับ

สูงสุด n ลำดับของเอกสาร อย่างไรก็ตามข้อเสียของ AQE ที่พบคือจำนวนเอกสารที่เหมาะสมที่จะนำไปทำการคำนวณหลังจากที่ผู้ใช้ทำการส่งคำสืบค้นไปแล้วรวมถึงการเลือกเทอมที่เหมาะสมและการคำนวณค่าน้ำหนักของคำที่จะนำไปเพิ่มลงในคำสืบค้น [Qiu and Frei 1993]

จากข้อเสียข้างต้นที่กล่าวมาของ AQE จึงได้มีการนำเสนอแนวคิดต่างๆในการปรับปรุงประสิทธิภาพของ AQE เช่น การปรับปรุงเทคนิคการขยายคำสืบค้นที่ใช้หลักการขอบเขตของความรู้ หรือ Domain Knowledge [Qiu and Frei 1993] โดยที่ขอบเขตความรู้จะถูกสร้างโดยอัตโนมัติซึ่งมีลักษณะคล้ายคลึงกับพจนานุกรมคำเหมือน (Thesaurus) เทคนิคนี้ใช้ในการแก้ปัญหาที่สำคัญสองประการของการขยายคำสืบค้นก็คือการเลือกเทอมที่เหมาะสมและการคำนวณค่าน้ำหนักของเทอมที่ถูกเลือกเพื่อนำไปเพิ่มลงในคำสืบค้น โดยการเลือกเทอมนั้นขึ้นอยู่กับความคล้ายคลึงระหว่างความหมายของคำสืบค้นโดยรวมกับกลุ่มของเทอมที่ถูกรวบรวมไว้แทนที่จะเลือกความคล้ายคลึงระหว่างเทอมของคำสืบค้นและกลุ่มของเทอมที่ถูกรวบรวมไว้ นอกจากนี้แล้วยังมีวิธีการปรับปรุงเทคนิคการขยายคำสืบค้นโดยใช้การจัดแบ่งประเภทของเอกสาร (Crouch, 1990) โดยใช้อัลกอริทึมในการจัดแบ่งประเภทของเอกสาร เทอมที่ไม่พบบ่อยในแต่ละประเภทของเอกสารจะถูกพิจารณาความคล้ายคลึงกันและรวบรวมเป็นกลุ่มตามประเภทของเทอมที่เหมือนกัน รวมทั้งยังปรับปรุงการทำดัชนีของเอกสารและคำสืบค้นโดยการแทนที่เทอมด้วยพจนานุกรมคำเหมือน (Thesaurus) หรือการเพิ่มคำในพจนานุกรมคำเหมือนลงในข้อมูลดัชนี เป็นต้น

2. การขยายคำสืบค้นแบบปฏิสัมพันธ์ (Interactive Query Expansion)

การขยายคำสืบค้นแบบปฏิสัมพันธ์ หรือ Interactive Query Expansion ซึ่งถูกเรียกสั้นๆ ว่า IQE [Efthimiadis 1996] หลักการของวิธีนี้คือให้ผู้ใช้ทำการพิจารณาเลือกข้อมูลที่ถูกค้นคืนโดยระบบค้นคืนสารสนเทศ ความเกี่ยวข้องจะขึ้นอยู่กับการประเมินของผู้ใช้ซึ่งวิธีการนี้จะให้ประสิทธิภาพที่ดีมากถ้าหากผู้ใช้มีเวลาในการพิจารณาทุกๆเอกสารที่ถูกสืบค้นขึ้นมาและสามารถระบุให้กับระบบได้ว่าเอกสารใดเกี่ยวข้องเอกสารใดไม่เกี่ยวข้อง ในการประเมินของผู้ใช้ถ้าผู้ใช้มีประสบการณ์ต่อการเลือกเทอมคือรู้ว่ามีคำใดที่มีความเกี่ยวข้องหรือใกล้เคียงกับกลุ่มเอกสารหรือเว็บเพจที่ต้องการค้นหาจะมีส่วนช่วยให้ผลลัพธ์ที่ได้มีการปรับปรุงที่ดีขึ้นมากกว่าเดิม ในทางตรงกันข้ามถ้าผู้ใช้ที่ไม่มีประสบการณ์ในการเลือกเทอมก็จะส่งผลให้ผลลัพธ์ที่ได้ในบางโอกาสไม่ดีไปกว่าเดิม [Magennis and Rijsbergen 1997]

จากงานวิจัยต่างๆที่ศึกษาเกี่ยวกับการเปรียบเทียบประสิทธิภาพของ AQE และ IQE พบว่า ข้อดีของ AQE ที่เหนือกว่า IQE คือผู้ใช้ไม่ต้องเสียเวลามาทำการนั่งอ่านทุกๆเอกสารและประเมินความเกี่ยวข้อง ในขณะที่ IQE นั้นผู้ใช้จำเป็นต้องทำการพิจารณาเลือกผลลัพธ์ว่ามีความเกี่ยวข้องหรือไม่เกี่ยวข้องจึงส่งผลให้ต้องใช้ระยะเวลานานในการพิจารณาเลือกผลลัพธ์ทั้งหมดซึ่ง

จุดนี้กลายเป็นการรบกวนและเป็นการเพิ่มภาระให้กับผู้ใช้ที่ต้องทำ [Jian-Fu et. al. 2005, Seher 2006] นอกจากนี้แล้ว AQE มีประสิทธิภาพในการปรับปรุงผลลัพธ์ได้ดีขึ้นเป็นอย่างมากโดยเฉพาะในกรณีที่มีการทำซ้ำหลายๆครั้งกับข้อมูลจำนวนมากที่ต้องการจะค้นหา [Magennis and Rijsbergen 1997] สำหรับในกรณีที่ต้องการค้นคืนในเรื่องที่เฉพาะเจาะจงนั้น IQE ให้ประสิทธิภาพที่ดีกว่า AQE [Koenemann and Belkin 1996]

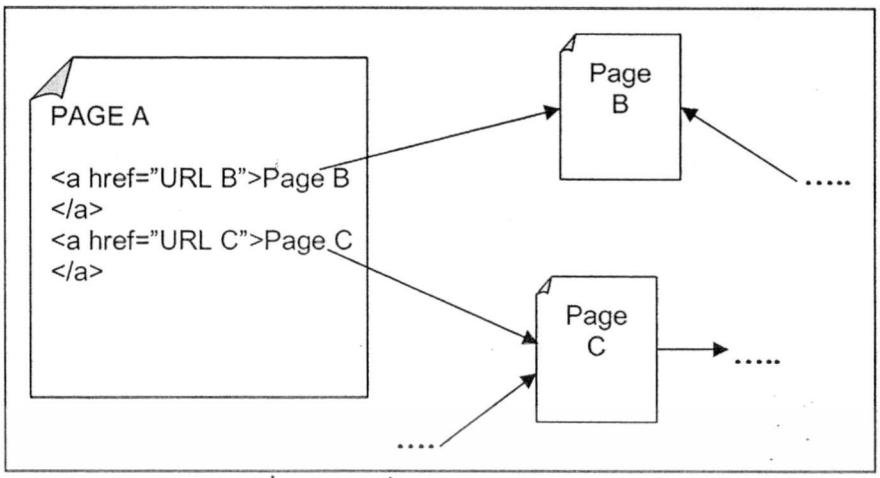
2.1.2. เทคนิคการวิเคราะห์ลิงค์ (Link Analysis Techniques)

การวิเคราะห์ไฮเปอร์ลิงค์และโครงสร้างกราฟของเว็บ [Allan et. al. 2005] ถูกนำมาใช้เป็นเครื่องมือในการพัฒนาเว็บเสิร์จเอนจิ้นโดยเน้นการนำไฮเปอร์ลิงค์มาใช้สำหรับการจัดเรียงลำดับของผลลัพธ์ที่ได้จากการค้นคืน การวิเคราะห์ลิงค์เป็นหนึ่งในหลายๆปัจจัยที่เสิร์จเอนจิ้นนำมาใช้ในการพิจารณาเป็นส่วนประกอบในการคำนวณคะแนนเว็บเพจที่ได้จากการค้นคืนด้วยคำสืบค้น การวิเคราะห์ลิงค์อาศัยแนวคิดพื้นฐานของการนำเว็บกราฟมาใช้ในการวิเคราะห์ โดยมองลักษณะการเชื่อมโยงกันของเว็บเพจเป็นกราฟซึ่งโหนดแทนด้วยเว็บเพจและกราฟแบบมีทิศทางแทนด้วยไฮเปอร์ลิงค์ ดังที่แสดงในรูป 2.4 สมมติฐาน 2 ข้อ [Henzinger 2000] ที่ถูกนำมาใช้ในการวิเคราะห์คือ

- ข้อแรก ถ้ามีลิงค์จากเพจ A ไปยังเพจ B แสดงว่ามีการแนะนำเพจ B โดยเพจ A และ
- ข้อที่สอง ถ้าเพจ A และเพจ B มีการเชื่อมโยงต่อกันด้วยไฮเปอร์ลิงค์ โอกาสความน่าจะเป็นที่ทั้งเพจ A และเพจ B จะอยู่ในหัวข้อเดียวกันมีมากกว่าถ้าทั้งเพจ A และเพจ B ที่ไม่ได้เชื่อมต่อกันด้วยลิงค์

จากสมมติฐานทั้ง 2 ข้อจะช่วยแสดงให้เห็นถึงความนิยมของเว็บเพจที่แตกต่างกัน นอกจากนี้แล้วเว็บกราฟยังถูกนำเอาไปใช้ในการบ่งบอกถึงความสอดคล้องของคำสืบค้นต่างๆที่ใช้สืบค้นโดยนำเอาการอ้างอิงของเว็บเพจต่างๆมาใช้วิเคราะห์

เทคนิคการวิเคราะห์ลิงค์นอกจากพัฒนาเพื่อวัตถุประสงค์ในการจัดลำดับความสัมพันธ์ของเว็บเพจและเพื่อแก้ข้อจำกัดของการจัดลำดับเว็บเพจแบบเดิมที่ใช้การคำนวณถ่วงน้ำหนักของคำ รวมถึงปัญหาที่เว็บเพจไม่สามารถอธิบายตัวมันเองได้อย่างเพียงพอแล้วยังสามารถนำมาประยุกต์ใช้ในด้านอื่นๆได้ เช่น การจัดกลุ่มหรือจัดประเภทของเว็บเพจ การหาเว็บเพจที่มีเนื้อหาใกล้เคียงกัน การหาเว็บเพจที่มีลักษณะซ้ำกัน และการตัดสินใจเลือกกลุ่มเว็บเพจที่จะใช้ครอลเลอร์ (Crawler) หรือสไปเดอร์ (Spider) ไปเก็บรวบรวม เป็นต้น



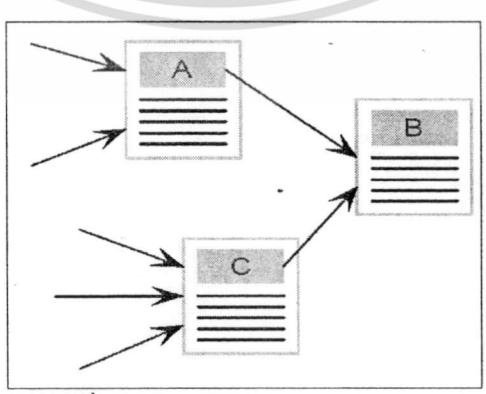
รูปที่ 2.4 การเชื่อมโยงของไฮเปอร์ลิงค์

อัลกอริทึมหลักของเทคนิคการวิเคราะห์ลิงค์ที่นิยมใช้กันในปัจจุบันได้แก่ PageRank และ Hypertext Induce Topic Search หรือ HITS อัลกอริทึม ซึ่งทั้งสองอัลกอริทึมมีรายละเอียดดังต่อไปนี้

2.1.2.1 PageRank

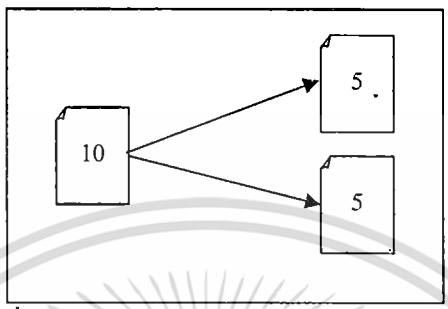
PageRank [Brin and Page 1998] เป็นขั้นตอนวิธีเพื่อใช้ในการวัดระดับค่าความสำคัญของแต่ละเว็บเพจจากเว็บเพจทั้งหมดที่มีอยู่ PageRankใช้แนวคิดที่ว่าเว็บเพจจะมีค่าลำดับสูงถ้าผลรวมของลำดับของเพจที่ชี้มาหามีค่าสูงด้วยซึ่ง PageRank ของเพจที่มีค่าสูงอาจจะขึ้นอยู่กับเว็บเพจจำนวนมากที่ชี้มาหาหรือเว็บเพจจำนวนไม่มากแต่มีค่า PageRank สูงๆที่ชี้มาหา หลักการหาค่า PageRank มีดังต่อไปนี้

1. เพจใดๆ ที่ถูกอ้างอิงจากหลายๆเพจ ดังนั้นเพจดังกล่าวน่าจะมีแนวโน้มว่ามีความสำคัญ ดังแสดงในรูปที่ 2.5



รูปที่ 2.5 การหาความสำคัญของเพจ

2. ถ้าเพจใดๆ ที่ถูกอ้างอิงด้วยเพจที่มีความสำคัญ ดังนั้นเพจดังกล่าวน่าจะมีความเป็นไปได้ว่ามีความสำคัญเท่ากันถึงแม้ว่าจะมีเพจจำนวนไม่มากอ้างอิงถึงก็ตาม
3. ความสำคัญของเพจจะถูกแบ่งโดยเท่าๆกันและกระจายไปยังเพจที่ถูกชี้ด้วยตัวมันเองดังที่แสดงในรูปที่ 2.6



รูปที่ 2.6 แสดงการแบ่งค่าความสำคัญของเพจ

การคำนวณหาค่าความสำคัญของ PageRank มีสมการดังต่อไปนี้

$$PR(A) = (1-d) + d(PR(T1)/C(T1) + ... + PR(T_n)/C(T_n)) \quad (2.1)$$

โดยที่

PR(A) คือ PageRank ของเพจ A (เพจที่ต้องการคำนวณ)

d คือ dampening factor โดยปกติจะถูกนำกำหนดค่าเป็น 0.85

PR(T1) คือ PageRank ของเว็บเพจที่ชี้ไปยังเพจ A

C(T1) คือ จำนวนของลิงค์จากหน้าที่กำลังพิจารณา

PR(Tn)/C(Tn) หมายถึง การหาค่า PageRank ของแต่ละหน้าที่มีการชี้ไปยังเพจ A

ค่า PageRank ของเพจใดๆจะมีค่าสูงก็ต่อเมื่อ ผลรวมของค่า PageRank ของเว็บเพจที่อ้างอิงถึงเพจนั้นมีค่าสูง

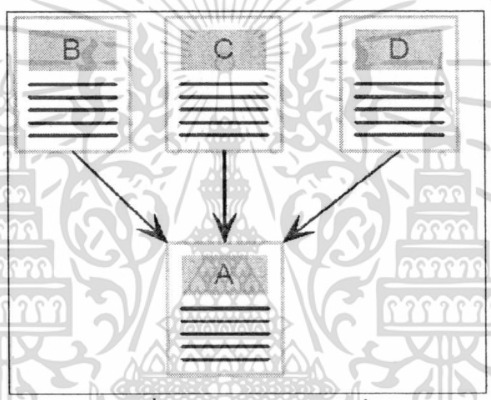
ตัวอย่างการประยุกต์ใช้งานของ PageRank (Henzinger, 2000) เช่น

- การจัดเรียงลำดับผลการสืบค้นด้วยเว็บเสิร์จเอนจิ้น (Web Search Ranking)
- การประมาณการเข้าใช้งานเว็บ (Estimate web traffic)
- การทำนายเว็บเพจที่มีการชี้โยงมาหา (Backlink predictor)

2.1.2.2 Hypertext Induce Topics Search (HITS Algorithm)

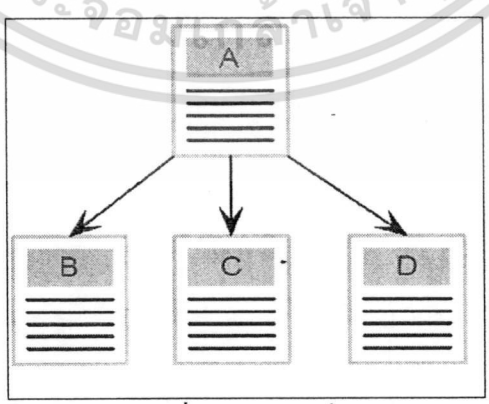
HITS [Kleinberg 1998] ถูกพัฒนาขึ้นเพื่อใช้ในการวัดคุณภาพของเว็บเพจที่มีความเกี่ยวข้องกับหัวข้อหรือหัวเรื่องผ่านทางการวิเคราะห์ความสัมพันธ์ของกราฟย่อย (subgraph) ของเว็บเพจ โดยมีการจำแนกเว็บเพจ 2 ประเภทได้แก่ ฮับเพจ (Hub page) คือเพจที่ทำหน้าที่ชี้ไปยังเว็บเพจที่เป็นประโยชน์หรือเพจที่ให้แหล่งข้อมูลที่ดีและออธอริทีเพจ (Authority page) คือเพจที่มีข้อมูลที่เกี่ยวข้องหรือเพจที่เป็นแหล่งข้อมูลที่ดี แนวคิดเบื้องต้นของ HITS ประกอบไปด้วย

1. เพจที่เป็นออธอริเพจทีที่ดีโดยมีความเกี่ยวข้องกับคำสืบค้น ถ้าออธอริทีเพจนั้นถูกอ้างอิงจากหลายๆเพจ (เพจที่เป็นฮับเพจที่ดี) และฮับเพจเหล่านั้นเกี่ยวข้องกับคำสืบค้น ดังแสดงในรูปที่ 2.7



รูปที่ 2.7 ออธอริทีเพจที่ดี

2. เพจที่เป็นฮับเพจที่ดีโดยมีความเกี่ยวข้องกับคำสืบค้น ถ้าฮับเพจนั้นอ้างอิงไปยังออธอริทีเพจที่ดีหลายๆเพจและออธอริทีเพจเหล่านั้นเกี่ยวข้องกับคำสืบค้น ดังแสดงในรูปที่ 2.8

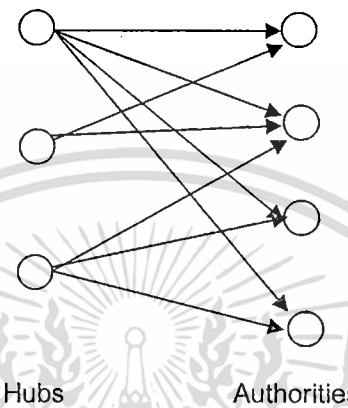


รูปที่ 2.8 ฮับเพจที่ดี

3. เพจที่เป็นออธอริทีเพจที่ดีและเป็นฮับเพจที่ดีจะช่วยสนับสนุนซึ่งกันและกัน (Mutual Reinforcement)

4. ออธอริทีเพจและฮับเพจ ที่มีความเกี่ยวข้องกับคำสืบค้นเดียวกันมีแนวโน้มมาจากกราฟย่อย (sub-graph) ที่แบ่งออกเป็นสองส่วนของเว็บกราฟ (web graph) ดังที่แสดงในรูปที่ 2.9

5. เว็บเพจสามารถเป็นได้ทั้งออธอริทีเพจที่ดีและฮับเพจที่ดี



รูปที่ 2.9 Bipartite sub-graph

ตัวอย่าง Psuedo-code ของ HITS algorithm

Let *p*, *q* and *r* are pages

Line 1: *G* := set of page

Line 2: **for each** page *p* in *G* **do**

Line 3: *p.auth* = 1

Line 4: *p.hub* = 1

Line 5: **function** HITS (*G*)

Line 6: **for** step **from** 1 **to** k **do**

Line 7: **for each** *page p* in *G* **do**

Line 8: **for each** *page q* in *p.incomingNeighbors* **do**

Line 9: *p.auth* += *q.hub*

Line 10: **for each** *page p* in *G* **do**

Line 11: **for each** *page r* in *p.outgoingNeighbors* **do**

Line 12: *p.hub* += *r.auth*

Line 13: *p.auth* = *p.auth* / |*p.auth*|

Line 14: *p.hub* = *p.hub* / |*p.hub*|

Line 15: sort (*p.auth*)

Line 16: sort (*p.hub*)

Line 17: return (p.auth, p.hub)

อธิบายการทำงานในแต่ละขั้นตอนของ Psuedo-code

กำหนดให้ *p, q* และ *r* เป็นเพจใดๆ

บรรทัดที่ 1 กำหนดให้ *G* คือเซตของผลลัพธ์ที่ได้จากการค้นคืนด้วยคำสืบค้น

บรรทัดที่ 2-4 กำหนดให้ค่าออธอริตีและฮับของเพจทุกๆ *p* ที่อยู่ใน *G* มีค่าเริ่มต้นเท่ากับ 1

บรรทัดที่ 5 เป็นฟังก์ชั่นในการคำนวณ HITS อัลกอริทึม

บรรทัดที่ 6 จำนวนรอบที่จะทำซ้ำ โดยกำหนดค่าของ *k*

บรรทัดที่ 7 คำนวณค่าออธอริทีของทุกๆ เพจ *p* ที่อยู่ใน *G*

บรรทัดที่ 8-9 หาผลรวมจำนวนของเพจ *q* ที่ชี้มาหาเพจ *p* โดยที่ *p.incomingNeighbors* คือเซตของกลุ่มเพจที่ชี้มาหาเพจ *p*

บรรทัดที่ 10 คำนวณค่าฮับของทุกๆ เพจ *p* ที่อยู่ใน *G*

บรรทัดที่ 11-12 หาผลรวมจำนวนของเพจ *r* ที่เพจ *p* ชี้ไปหา โดยที่ *p.outgoingNeighbors* คือเซตจองกลุ่มเพจที่เพจ p ชี้ไปหา

บรรทัดที่ 13-14 ในแต่ละรอบของการทำซ้ำจะมีการนอมัลไลซ์ทั้งค่าออธอริทีและค่าฮับ

บรรทัดที่ 15-16 เมื่อทำซ้ำครบตามจำนวนรอบที่กำหนด จากนั้นจะทำการเรียงลำดับค่าออธอริทีและค่าฮับที่มีค่ามากที่สุดไปยังค่าที่น้อยที่สุด

บรรทัดที่ 17 รีเทิร์นค่า *p.auth* และ *p.hub*

ตัวอย่างการประยุกต์ใช้งานของ HITS อัลกอริทึม [Henzinger 2000] เช่น

- การจัดเรียงลำดับผลการสืบค้นด้วยเว็บเสิร์จเอนจิ้น (Web Search Ranking)
- การหาเว็บเพจที่มีเนื้อหาคล้ายคลึงหรือใกล้เคียงกัน (Finding related pages)
- การหาหมวดหมู่ที่ได้รับความนิยมในเว็บไดเรกทอรี่ (Popularity categories in the web directories)
- การวิเคราะห์การอ้างอิง (Citation Analysis)

2.2 การประเมินประสิทธิภาพระบบค้นคืนสารสนเทศ

การประเมินประสิทธิภาพการค้นคืนเป็นกระบวนการที่สำคัญของระบบค้นคืนสารสนเทศ ทั้งนี้เพื่อต้องการทราบถึงสมรรถนะของระบบและคุณภาพของการบริการ กระบวนการทำงาน ตลอดจนนโยบายของการค้นคืน สำหรับการประเมินประสิทธิภาพที่จะกล่าวถึงคือการประเมินประสิทธิภาพของเว็บเสิร์จเอนจิ้นโดยส่วนประกอบในการประเมินประสิทธิภาพจะประกอบด้วยหัวเรื่องหรือชุดคำถามที่ใช้ในการทดสอบและการวัดประสิทธิภาพในการค้นคืนซึ่งมีรายละเอียดดังต่อไปนี้

2.2.1 หัวเรื่องหรือชุดคำถามที่ใช้ในการทดสอบ

หัวเรื่องหรือชุดคำถามที่ถูกนำมาใช้ในการทดสอบหรือวัดประสิทธิภาพด้านการค้นคืนมักนำมาจาก TREC (Text REtrieval Conference) ตัวอย่างของงานทางด้านนี้ เช่น [White et. al. 2002, Nemeth et. al. 2004] ซึ่งในการทดสอบจะมีการกำหนดหัวเรื่องเฉพาะด้านหรือกำหนดคำถามขึ้นมาเพื่อใช้ในการค้นคืนแล้วนำผลลัพธ์ที่ได้ไปทำการประเมิน

2.2.2 การวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ

ในการประเมินผลการค้นคืนของเว็บเสิร์จเอนจิ้นนั้นจะมีการวัดประสิทธิภาพที่สำคัญสามค่าด้วยกันคือ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเฉลี่ย (F-measure) ซึ่งปกติใช้กับการค้นคืนเอกสารจากชุดเอกสารจำนวนมาก อย่างไรก็ตามในการค้นคืนจากอินเทอร์เน็ตการวัดประสิทธิภาพของค่าความระลึกจะใช้วิธีการวัดแบบสัมพัทธ์ ตัวอย่างเช่น ในงานวิจัยของ [Shafi and Rather 2005] ซึ่งการวัดประสิทธิภาพทั้งสามค่ามีรายละเอียดดังต่อไปนี้

1. ความแม่นยำ (Precision) คำนวณได้จากสมการที่ 2.2

$$\text{ความแม่นยำ} = \frac{(\text{จำนวนเว็บเพจที่ค้นคืนกลับมาได้และมีความเกี่ยวข้อง})}{(\text{จำนวนเว็บเพจที่ค้นคืนกลับมาได้})} \qquad (2.2)$$

2. ความระลึกสัมพัทธ์ (Relative-Recall) คำนวณได้จากสมการที่ 2.3

$$\text{ความระลึกสัมพัทธ์} = \frac{(\text{จำนวนเอกสารทั้งหมดที่เกี่ยวข้องและถูกค้นคืนโดยเสิร์จเอนจิ้นนั้นๆ})}{(\text{จำนวนเอกสารทั้งหมดที่เกี่ยวข้องและถูกค้นคืนโดยทุกเสิร์จเอนจิ้น})} \qquad (2.3)$$

โดยการวัดค่าความระลึกแบบสัมพัทธ์จะแบ่งเป็น 2 กรณี คือ

กำหนดให้ a, b, c, d และ e แทนเสิร์จเอนจิ้น และ

a1, b1, c1, d1 และ e1 เป็นจำนวนผลลัพธ์ที่ถูกค้นคืนโดยเสิร์จเอนจิ้น a, b, c, d และ e ตามลำดับ

- กรณีแรกคือ ผลลัพธ์ที่ได้จากการค้นคืนในแต่ละเสิร์จเอนจิ้นไม่มีผลลัพธ์ที่ซ้ำกัน เช่น ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ไม่ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น b, ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ไม่ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น c, ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ไม่ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น d และผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ไม่ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น e เป็นต้น ดังนั้นการหาค่าความระลึกสัมพัทธ์ของเสิร์จเอนจิ้น a คือ

Relative Recall ของ a = (a1) / (a1+b1+c1+d1+e1) (2.4)

- กรณีที่สองคือ ผลลัพธ์ที่ได้จากการค้นคืนในแต่ละเสิร์จเอนจิ้นมีผลลัพธ์บางส่วนที่ซ้ำกัน เช่น ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น b จำนวน b2 ผลลัพธ์, ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น c จำนวน c2 ผลลัพธ์, ผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น d จำนวน d2 ผลลัพธ์และผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น a ซ้ำกับผลลัพธ์ที่ถูกค้นคืนด้วยเสิร์จเอนจิ้น e จำนวน e2 ผลลัพธ์ เป็นต้น ดังนั้นการหาค่าความระลึกสัมพัทธ์ของเสิร์จเอนจิ้น a คือ

Relative Recall ของ a = (a1) / (a1+b2+c2+d2+e2) (2.5)

3. ค่าเฉลี่ย (F-measure) คือ เป็นการวัดค่าเฉลี่ยประสิทธิภาพของผลการค้นคืนโดยนำค่าความแม่นยำและค่าความระลึกมาเฉลี่ยรวมเป็นค่าเดียว ซึ่งหาได้จาก

กำหนดให้

P คือความแม่นยำ

R คือความครอบคลุม

α เป็นค่าสัมประสิทธิ์น้ำหนักที่ผู้สืบค้นให้ความสำคัญในค่าความแม่นยำหรือค่าความระลึก

$$F = \frac{1}{\alpha \times \frac{1}{P} + (1-\alpha) \times \frac{1}{R}} \quad (2.6)$$

เมื่อให้ค่าสัมประสิทธิ์เท่ากับ $\alpha = 0.5$ จะได้ว่า

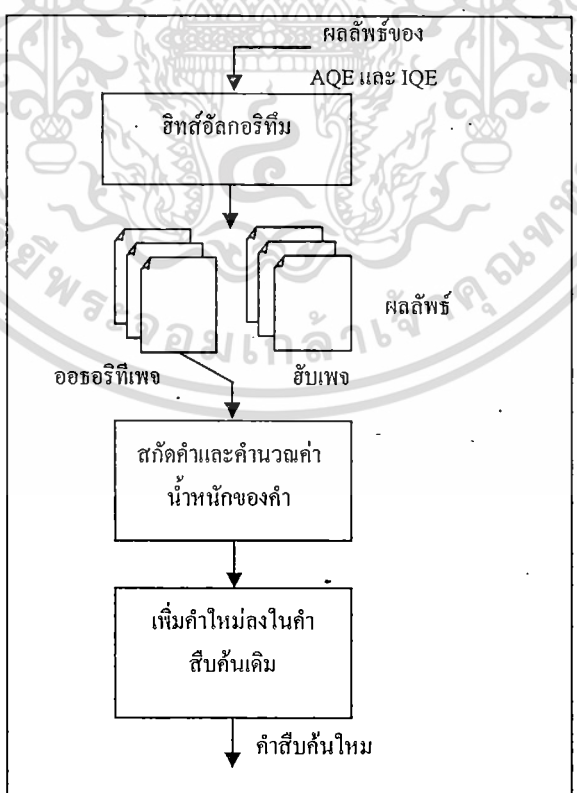
$$F = \frac{2 \times P \times R}{P + R} \tag{2.7}$$

บทที่ 3
การปรับปรุงประสิทธิภาพเทคนิคการขยายคำสืบค้นโดยใช้เทคนิคการวิเคราะห์ลิงค์

ในบทนี้จะกล่าวถึงวิธีการที่นำเสนอและการออกแบบสถาปัตยกรรมเว็บเสิร์จเอนจิ้นเพื่อใช้ในการปรับปรุงประสิทธิภาพของเทคนิคการขยายคำสืบค้น นอกจากนี้จะกล่าวถึงแนวทางการออกแบบการทดลองเพื่อใช้ในการประเมินประสิทธิภาพของวิธีการที่นำเสนอ

3.1 แนวคิดที่นำเสนอ

เพื่อปรับปรุงประสิทธิภาพของเทคนิคการขยายคำสืบค้นให้ดีขึ้น วิธีการที่นำเสนอคือการนำเอาเทคนิคการวิเคราะห์ลิงค์มาบูรณาการเข้ากับเทคนิคการขยายคำสืบค้นโดยนำผลลัพธ์ที่ได้จากเทคนิคการขยายคำสืบค้นมาทำการวิเคราะห์ลิงค์เพื่อหากลุ่มเว็บเพจที่มีเนื้อหาเกี่ยวข้องที่ใกล้เคียงกัน จากนั้นนำกลุ่มเว็บเพจดังกล่าวมาทำการหาคำศัพท์ร่วมแล้วนำคำศัพท์ร่วมที่ได้เพิ่มลงในคำสืบค้นเดิมที่มีอยู่แล้วนำกลับไปสืบค้นใหม่อีกครั้งดังแสดงในรูปที่ 3.1

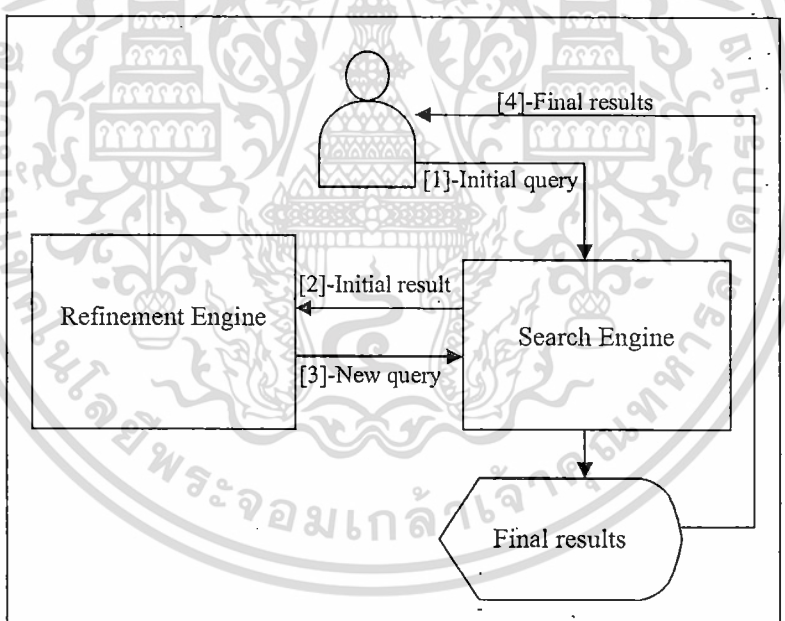


รูปที่ 3.1 ขั้นตอนของแนวคิดที่นำเสนอ

3.2 การออกแบบสถาปัตยกรรมและการทำงานของส่วนประกอบต่างๆ

จากแนวคิดที่นำเสนอในหัวข้อ 3.1 จึงเป็นที่มาของการออกแบบสถาปัตยกรรมเว็บเสิร์จเอนจิ้นแบบขยายคำสืบค้นทำงานร่วมกับการวิเคราะห์ลิงค์โดยมีส่วนประกอบหลักที่เรียกว่ากลไกการปรับปรุง (Refinement Engine) ดังรูปที่ 3.2 และส่วนประกอบของกลไกการปรับปรุง (Refinement Engine) แสดงในรูปที่ 3.3

การทำงานของกลไกการปรับปรุงเริ่มจากคำสืบค้นเริ่มแรกของผู้ใช้ เสิร์จเอนจิ้นทำการประมวลผลคำสืบค้นและค้นคืนผลลัพธ์เบื้องต้นกลับมา ผลลัพธ์เบื้องต้นที่ได้จะถูกนำมาคำนวณหาความสัมพันธ์ของไฮเปอร์ลิงค์ระหว่างเซตผลลัพธ์เบื้องต้นและลิงค์ที่ถูกขยายซึ่งจะถูกเพิ่มลงในเซตโดยการใช้ฮิทส์อัลกอริทึม (HITS algorithm) ผลลัพธ์ที่ได้จากการคำนวณด้วยฮิทส์อัลกอริทึม (HITS algorithm) คือเซตของฮับเพจ (Hubs) และเซตของออธอริทีเพจ (Authorities) พร้อมกับค่าคะแนนของฮับเพจและออธอริทีเพจที่สัมพันธ์กันจากนั้นคำค้นหาใหม่จะถูกดึงออกมาและถูกรวมเข้ากับคำค้นหาเริ่มต้น และขั้นสุดท้ายคำค้นหาใหม่ที่ได้จะถูกส่งกลับไปยังเสิร์จเอนจิ้นอีกครั้ง



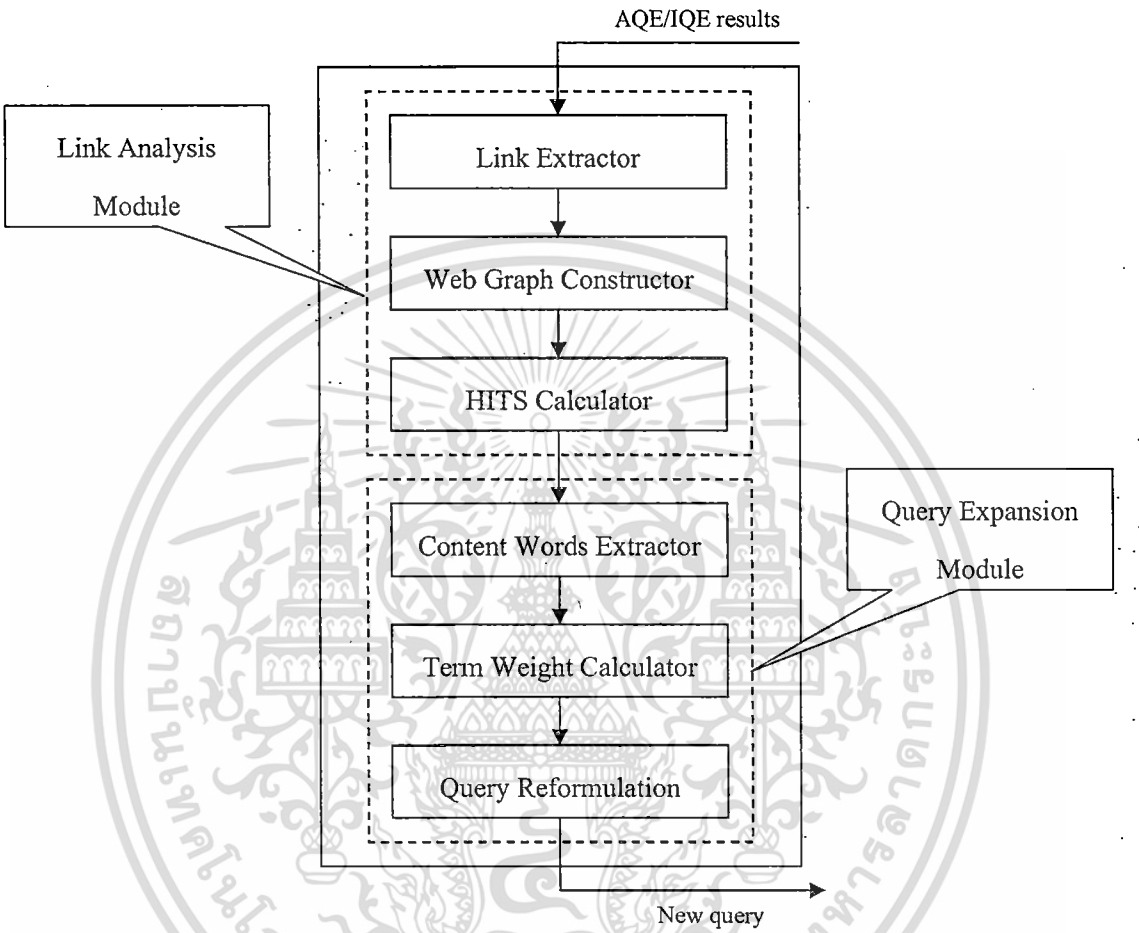
รูปที่ 3.2 สถาปัตยกรรมเว็บเสิร์จเอนจิ้นแบบขยายคำสืบค้นทำงานร่วมกับการวิเคราะห์ลิงค์

กลไกการปรับปรุงดังที่แสดงในรูป 3.3 ประกอบด้วยโมดูลย่อย 2 โมดูลคือ โมดูลการวิเคราะห์ลิงค์ (Link Analysis Module) และโมดูลการขยายคำสืบค้น (Query Expansion Module)

3.2.1 โมดูลการวิเคราะห์ลิงค์ (Link Analysis Module)

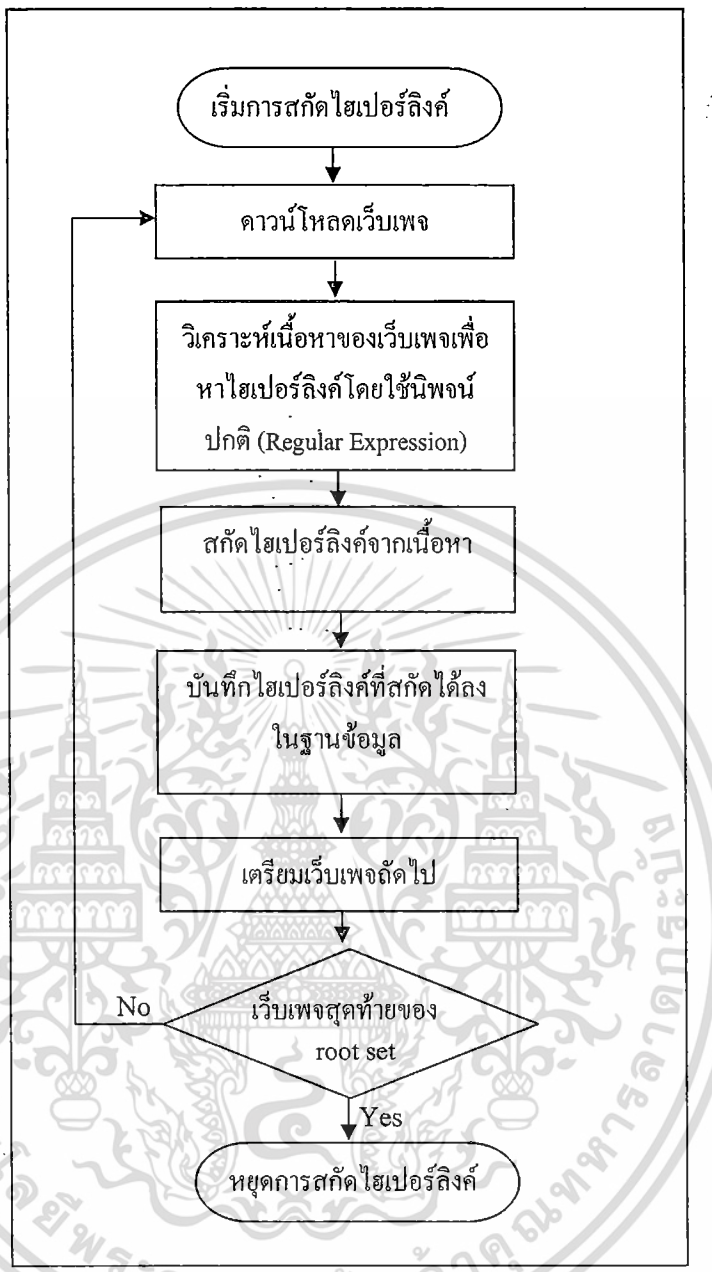
การทำงานย่อยของโมดูลการวิเคราะห์ลิงค์ประกอบด้วย 3 โมดูล ได้แก่ โมดูลสกัดไฮเปอร์ลิงค์ทำหน้าที่สกัดไฮเปอร์ลิงค์จากชุดผลลัพธ์เริ่มต้น โมดูลสร้างเว็บกราฟทำหน้าที่ใน

การสร้างซับกราฟของไฮเปอร์ลิงค์ที่ได้จากโมดูลสกัดไฮเปอร์ลิงค์ และโมดูลคำนวณฮิทส์อัลกอริทึมทำหน้าที่ในการคำนวณค่าคะแนนของเว็บเพจเพื่อใช้ในการจัดลำดับความเกี่ยวข้อง ซึ่งรายละเอียดของทั้ง 3 โมดูลมีดังต่อไปนี้



รูปที่ 3.3 แสดงส่วนประกอบของกลไกการปรับปรุง (Refinement Engine)

3.2.1.1 โมดูลสกัดไฮเปอร์ลิงค์ (Link Extractor Module) ทำหน้าที่ในการสกัดไฮเปอร์ลิงค์จากชุดผลลัพธ์เริ่มต้น ขั้นตอนการสกัดไฮเปอร์ลิงค์เริ่มจากการดาวน์โหลดเว็บเพจโดยใช้ไฮเปอร์ลิงค์ที่ได้จากผลลัพธ์เริ่มต้น เมื่อดาวน์โหลดเว็บเพจมาแล้วจะทำการอ่านและวิเคราะห์เว็บเพจเพื่อหาไฮเปอร์ลิงค์ที่เชื่อมโยงไปยังเว็บเพจอื่นโดยใช้นิพจน์ปกติหรือที่เรียกว่า Regular Expression เมื่อได้ผลลัพธ์ของไฮเปอร์ลิงค์ที่เชื่อมโยงไปยังเว็บเพจอื่นแล้วจะถูกบันทึกเก็บไว้ จากนั้นก็ทำการดาวน์โหลดเว็บเพจถัดไปแล้วทำตามขบวนการข้างต้นที่ได้กล่าวมา โดยขั้นตอนการสกัดไฮเปอร์ลิงค์สามารถแสดงเป็นแผนผังการทำงานได้ดังรูปที่ 3.4 ไฮเปอร์ลิงค์ผลลัพธ์ที่ได้ถูกส่งต่อไปยังโมดูลสร้างเว็บกราฟ (Web Graph Constructor Module)



รูปที่ 3.4 แผนผังการสกัดไฮเปอร์ลิงค์

3.2.1.2 โมดูลสร้างเว็บกราฟ (Web Graph Constructor Module) ทำหน้าที่ในการสร้างกราฟย่อย (sub graph) เพื่อหาความสัมพันธ์ของกลุ่มไฮเปอร์ลิงค์ทั้งหมดที่ประกอบไปด้วย กลุ่มไฮเปอร์ลิงค์ที่เป็นผลลัพธ์เริ่มต้น กลุ่มไฮเปอร์ลิงค์ที่มีการชี้มายังกลุ่มไฮเปอร์ลิงค์ที่อยู่ในผลลัพธ์เริ่มต้นหรือกลุ่มไฮเปอร์ลิงค์ที่ถูกชี้โดยกลุ่มไฮเปอร์ลิงค์ที่อยู่ในผลลัพธ์เริ่มต้น ผลลัพธ์ที่ได้จากการสร้างกราฟย่อยแสดงในรูปที่ 3.5 ซึ่งขั้นตอนในการสร้างกราฟย่อยมีดังต่อไปนี้

1. นำเซตไฮเปอร์ลิงค์ของผลลัพธ์ที่มีลำดับสูงสุด n ลำดับ (ในงานวิจัยนี้เลือกใช้ค่า n เท่ากับ 30) ที่ได้จากการสืบค้นด้วยเทคนิคการขยายคำสืบค้นแบบอัตโนมัติหรือเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์กับผู้ใช้ โดยเรียกเซตไฮเปอร์ลิงค์นี้ว่าเซตเริ่มต้น (Initial link set)

2. สร้างเซตขยายลิงค์ (Expanded link set) โดยการเพิ่มกลุ่มไฮเปอร์ลิงค์ที่มีความสัมพันธ์กับกลุ่มไฮเปอร์ลิงค์ที่อยู่ในเซตเริ่มต้นดังนี้

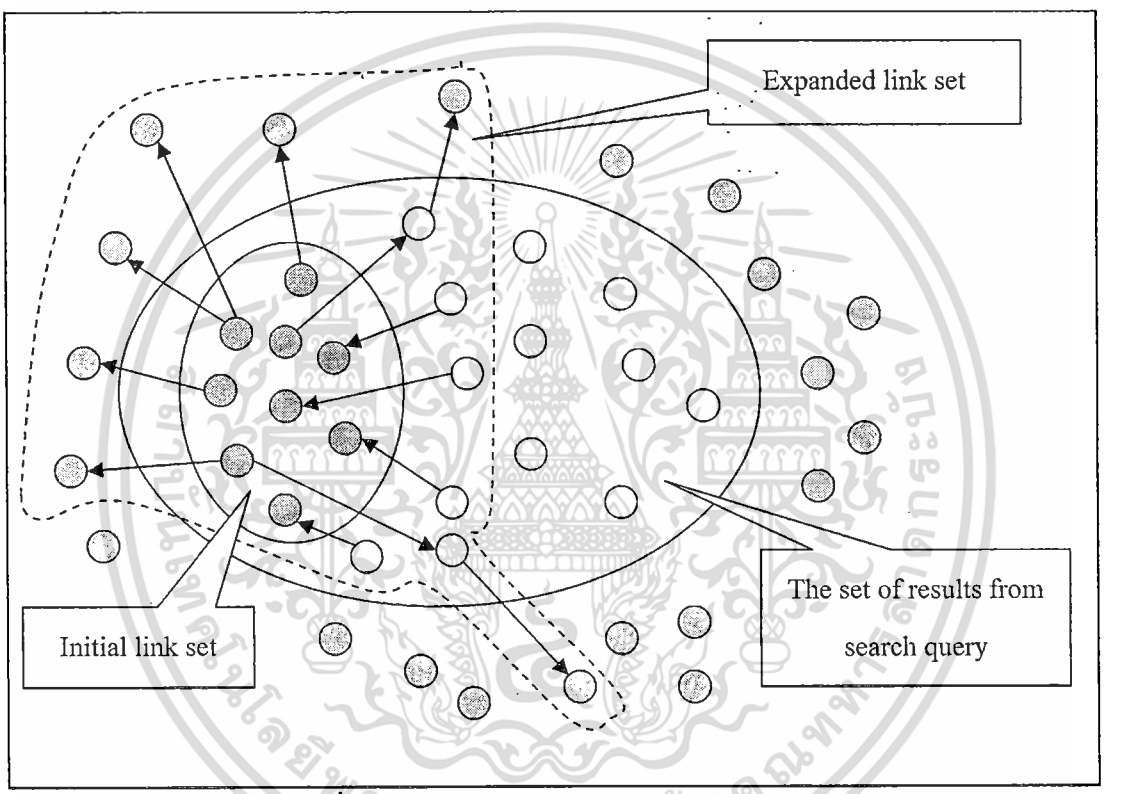
สำหรับทุกไฮเปอร์ลิงค์ p ที่อยู่ในเซตเริ่มต้น

2.1 เพิ่มทุกไฮเปอร์ลิงค์ที่ไฮเปอร์ลิงค์ p ชี้ไปหาลงในเซตขยายลิงค์ (outlink)

หรือ

2.2 เพิ่มทุกไฮเปอร์ลิงค์ที่ชี้มาหาไฮเปอร์ลิงค์ p ลงในเซตขยายลิงค์ (inlink)

3. ทำการลบลิงค์ทั้งหมดที่อยู่ภายในโดเมนเดียวกันในเซตขยายลิงค์



รูปที่ 3.5 แสดงการสร้างกราฟย่อย (Sub-graph)

จากรูปที่ 3.5 แสดงให้เห็นถึงเซตขยายลิงค์ของเซตเริ่มต้นที่มีอยู่ โดยรวมเอาโหนดที่มีลิงค์จากเซตเริ่มต้นซึ่งประกอบด้วย เซตของโหนดที่ถูกชี้โดยเซตเริ่มต้นและเซตของโหนดที่ชี้กลับมาหาเซตเริ่มต้น

ผลลัพธ์ที่ได้คือเซตของลิงค์ที่ถูกขยาย (Expanded link set) ซึ่งจะถูกส่งต่อไปยังโมดูลคำนวณฮิทส์อัลกอริทึม (HITS Calculator Module)

3.2.1.3 โมดูลคำนวณฮิทส์อัลกอริทึม (HITS Calculator Module) ทำหน้าที่ในการคำนวณค่าคะแนนของฮับเพจและออธอริทีเพจเพื่อใช้ในการจัดเรียงลำดับเว็บเพจที่มีความเกี่ยวข้องและได้รับความนิยม ในการคำนวณฮิทส์อัลกอริทึมจะอาศัยความสัมพันธ์ของการสนับสนุนซึ่งกัน

และกัน (Mutual reinforcing relation) ระหว่างฮับเพจและออธอริทีเพจที่ว่าฮับเพจที่ดีจะเชื่อมโยงไปยังออธอริทีเพจที่ดีจำนวนมากๆและออธอริทีเพจที่ดีจะถูกเชื่อมโยงโดยฮับเพจที่ดีจำนวนมากๆ ซึ่งขั้นตอนในการคำนวณฮิทส์อัลกอริทึม มีดังต่อไปนี้

คำนวณหาค่าคะแนนออธอริทีเพจและค่าคะแนนฮับเพจของแต่ละเว็บเพจที่อยู่ในเซตของลิงค์ที่ถูกขยาย (Expanded link set) ที่อยู่บนกราฟย่อย (sub-graph) $SG(V,E)$

ให้ p,q แทน เพจใดๆ และกำหนด

$A(p)$ คือค่าคะแนน Authority เพจของเพจ p

$H(p)$ คือค่าคะแนน Hub เพจของเพจ p

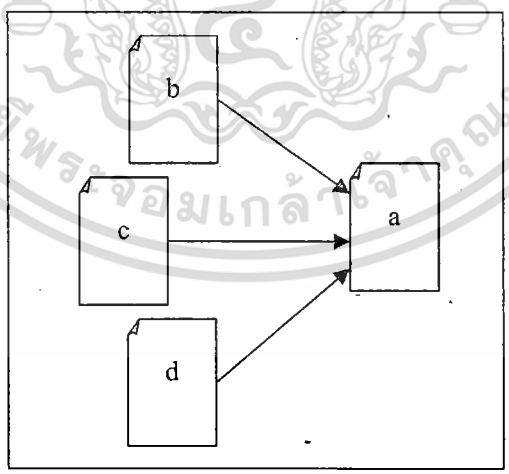
(p,q) คือกราฟที่มีทิศทางซึ่งเป็นสมาชิกใน E และชี้จาก p ไปยัง q

กำหนดค่าเริ่มต้น $A(p)$ และ $H(p)$ ของทุกเว็บเพจมีค่าเท่ากับ 1

วนรอบ p ใน V จนกระทั่งค่าคะแนนเริ่มไม่เปลี่ยนแปลง

$$A(p) = \sum_{q:(q,p)\in E} H(q) \tag{3.1}$$

ในสมการที่ 3.1 เป็นการคำนวณหาค่า $A(p)$ ของแต่ละเว็บเพจ p โดยหาจากผลรวม $H(q)$ ของทุกเว็บเพจ q ที่ชี้เข้าหาเว็บเพจ p ดังแสดงในรูปที่ 3.6

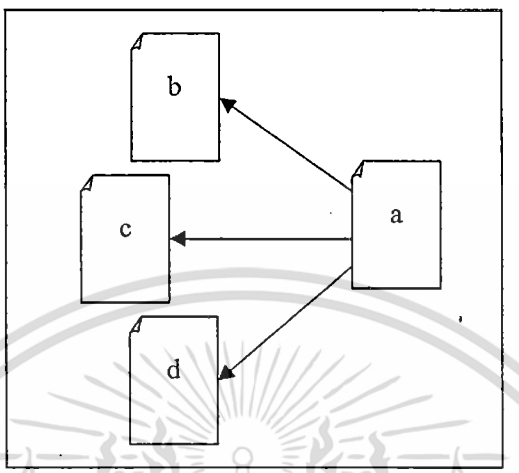


$$A(a) = H(b) + H(c) + H(d)$$

รูปที่ 3.6 การคำนวณหาค่าคะแนนออธอริทีเพจของเพจ a

$$H(p) = \sum_{q:(p,q)\in E} A(q) \tag{3.2}$$

ในสมการที่ 3.2 เป็นการคำนวณหาค่า $H(p)$ ของแต่ละเว็บเพจ p โดยหาจากผลรวม $A(q)$ ของทุกเว็บเพจ q ที่ถูกชี้โดยเว็บเพจ p ดังรูปที่ 3.7



$$H(a) = A(b) + A(c) + A(d)$$

รูปที่ 3.7 การคำนวณหาค่าคะแนนฮับเพจของเพจ a

นำค่าคงที่ค่าหนึ่งมาทำการหารทั้ง $A(p)$ และ $H(p)$ เพื่อทำการ Normalize ไม่ให้ค่า $A(p)$ และ $H(p)$ เยอะมากจนเกินไป

เรียงลำดับคะแนนของ $A(p)$และ $H(p)$ จากมากสุดไปหาน้อยสุด

ผลลัพธ์ที่ได้จากการคำนวณนี้คือกลุ่มของฮับเพจและออธอริทีเพจ ซึ่งจะถูกส่งต่อไปยังโมดูลแยกคำจากเนื้อหา (Content Words Extractor Module)

3.2.2 โมดูลการขยายคำสืบค้น (Query Expansion Module)

การทำงานย่อยของโมดูลการขยายคำสืบค้น (Query Expansion Module) ประกอบไปด้วย 3 โมดูล ได้แก่ โมดูลสกัดคำจากเนื้อหาทำหน้าที่ในการสกัดคำจากเนื้อหาของเว็บเพจ โมดูลคำนวณน้ำหนักของคำทำหน้าที่ในการคำนวณค่าน้ำหนักของคำที่ได้จากการสกัดคำจากเนื้อหา และโมดูลปรับปรุงคำสืบค้นทำหน้าที่ในเพิ่มคำใหม่ที่ได้ลงในคำสืบค้นเดิมที่มีอยู่ รายละเอียดการทำงานทั้ง 3 โมดูลมีดังต่อไปนี้

3.2.2.1 โมดูลสกัดคำจากเนื้อหา (Content Words Extractor Module) ทำหน้าที่สกัดคำและตัดคำจากเนื้อหาในแท็กชื่อเรื่อง (Title) และแท็กคำอธิบาย (Description) ของเว็บเพจ จากนั้นทำการกรองคำที่ไม่สำคัญหรือคำที่มีความสำคัญน้อย (Stopwords) ออก สำหรับเนื้อหาของเว็บเพจที่นำมาสกัดคำนั้นจะเลือกใช้เว็บเพจที่อยู่ในกลุ่มประเภทออธอริทีเพจเนื่องจากเว็บเพจที่อยู่ในกลุ่ม

ดังกล่าวจะมีเนื้อหาที่สำคัญ มีความเกี่ยวข้องโดยตรงกับหัวเรื่องและเป็นเว็บเพจที่ได้รับความนิยม โดยขั้นตอนของการสกัดคำจากเนื้อหาจะมีแผนผังการทำงานดังรูปที่ 3.8

รูปที่ 3.8 แผนผังการสกัดคำจากเนื้อหา

ผลลัพธ์ที่ได้คือคำซึ่งจะถูกส่งต่อไปยังโมดูลคำนวณน้ำหนักของคำ (Term Weight Calculator Module)

3.2.2.2 โมดูลคำนวณน้ำหนักของคำ (Terms Weight Calculator Module) ทำหน้าที่คำนวณค่าน้ำหนักของคำซึ่งพิจารณาจากจำนวนครั้งของคำที่ปรากฏร่วมระหว่างเว็บเพจที่เกิดจากการนำเว็บเพจทั้งหมดที่อยู่ในเซตมาทำการเปรียบเทียบกัน ขั้นตอนวิธีการเปรียบเทียบเทียบเว็บเพจในเซตเพื่อหาคำร่วมมีดังต่อไปนี้

กำหนดให้

CoW แทนอะเรย์ที่ใช้เก็บคำที่ปรากฏร่วมระหว่างเว็บเพจ

N แทนจำนวนเว็บเพจทั้งหมดที่อยู่ในเซต

W แทนคำใดๆที่ปรากฏในเว็บเพจ

doc แทนเว็บเพจใดๆที่อยู่ในเซต

For i = 1 To N

For j = (i+1) To N

If W co-occur in doc_i and doc_j Then add W into CoW

เมื่อได้คำที่ปรากฏร่วมแล้ว จากนั้นนำมาคำนวณค่าน้ำหนักโดยใช้สูตรการคำนวณได้ดังต่อไปนี้

กำหนดให้ W_i เป็นเทอม

[(จำนวนเว็บเพจที่อยู่ในเซตและมีคำ W_i ปรากฏอยู่)-1] / [(จำนวนเว็บเพจที่อยู่ในเซต)] (3.3)

ผลลัพธ์ที่ได้คือคำและค่าน้ำหนักของคำ ข้อสังเกตประการหนึ่งของสูตรการคำนวณนี้คือตัวหารจะเป็นค่าคงที่ ดังนั้นในการพิจารณาเพื่อเลือกคำใหม่นั้นจะขึ้นอยู่กับค่าตัวตั้งเป็นหลัก สำหรับคำ W_i ใดๆที่ไม่ปรากฏในเว็บเพจอื่นๆที่อยู่ในเซต ค่าน้ำหนักของคำดังกล่าวจะมีค่าเป็นศูนย์ ในทางทฤษฎีแล้วอัลกอริทึมจะเลือกเฉพาะคำที่มีค่าน้ำหนักมากกว่าค่าเฉพาะค่าหนึ่งที่กำหนดไว้ แต่ในการทดลองนี้จะใช้จำนวนคำที่ได้รับแนะนำโดย Magenis และ Rijsbergen [Magenis and Rijsbergen 1997] ซึ่งจะเลือกคำที่มีค่าน้ำหนักสูงสุดทั้งหมด 6 คำ

3.2.2.3 โมดูลปรับปรุงคำสืบค้น (Query Reformulation Module) ทำหน้าที่ในการปรับปรุงคำสืบค้นโดยการเพิ่มคำใหม่ที่มีค่าน้ำหนักสูงสุด 6 คำแรกลงในคำสืบค้นเดิมที่มีอยู่โดยใช้ตัวดำเนินการ “AND” เชื่อมระหว่างคำสืบค้นเดิมกับคำใหม่ที่ได้ จากนั้นส่งคำสืบค้นที่ปรับปรุงแล้วกลับไปสืบค้นที่เว็บเสิร์จเอนจิ้นอีกครั้ง

3.3 การออกแบบการทดลองเพื่อใช้เปรียบเทียบประสิทธิภาพ

ในหัวข้อนี้จะกล่าวถึงปัจจัยต่างๆที่เกี่ยวข้องกับการออกแบบการทดลองเพื่อใช้เปรียบเทียบประสิทธิภาพของเทคนิคการสืบค้นต่างๆ ได้แก่ รูปแบบการทดลอง สมมติฐานของการทดลอง ตัวแปรที่ต้องการศึกษาและตัวแปรอื่นๆ เครื่องมือที่ใช้ในการทดลอง ชุดคำถามที่ใช้ในการทดลอง ผู้เข้าร่วมทำการทดลอง กระบวนการหรือขั้นตอนในการทดลอง และจำนวนการทำรายการทั้งหมดที่ได้จากการทดลอง ซึ่งมีรายละเอียดดังต่อไปนี้

3.3.1 รูปแบบการทดลอง

รูปแบบการทดลองที่ใช้เปรียบเทียบประสิทธิภาพของเทคนิคการขยายคำสืบค้นที่ยังไม่ได้ทำการปรับปรุงและเทคนิคการขยายคำสืบค้นที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ จะประกอบด้วย 6 รูปแบบด้วยกัน ได้แก่

1. เสิร์จเอนจิ้นพื้นฐานทั่วๆไป เป็นการค้นคืนโดยไม่ใช้เทคนิคใดๆในการปรับปรุงประสิทธิภาพ

2. เทคนิคการขยายคำสืบค้นแบบอาศัยการปฏิสัมพันธ์ (Interactive Query Expansion: IQE) เป็นเทคนิคในการเพิ่มคำใหม่ลงในคำสืบค้นเดิมที่มีอยู่โดยอาศัยผู้สืบค้นปฏิสัมพันธ์กับระบบค้นคืนสารสนเทศ

3. เทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (Automatic Query Expansion: AQE) เป็นเทคนิคในการเพิ่มคำใหม่ลงในคำสืบค้นเดิมที่มีอยู่แบบอัตโนมัติ

4. เทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบอาศัยการปฏิสัมพันธ์ (Link Analysis collaborative with Interactive Query Expansion: LIQE) เป็นเทคนิคในการเพิ่มคำใหม่ลงในคำสืบค้นเดิมที่มีอยู่โดยใช้เทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบอาศัยการปฏิสัมพันธ์

5. เทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (Link Analysis collaborative with Automatic Query Expansion: LAQE) เป็นเทคนิคในการเพิ่มคำใหม่ลงในคำสืบค้นเดิมที่มีอยู่โดยใช้เทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ

6. เทคนิคการวิเคราะห์ลิงค์ (Link Analysis Techniques) เป็นเทคนิคในการปรับปรุงประสิทธิภาพการค้นคืนโดยอาศัยการวิเคราะห์โครงสร้างไฮเปอร์ลิงค์ของเว็บเพจ ซึ่งอัลกอริทึมที่เลือกใช้คือฮิทส์อัลกอริทึม (HITS algorithm)

โดยวิธีการที่นำเสนอคือรูปแบบที่ 4 คือเทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์และแบบที่ 5 คือเทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นแบบอัตโนมัติตามลำดับ

3.3.2 สมมติฐานของการทดลอง

ในการทดลองนี้ได้ตั้งสมมติฐานของการทดลองซึ่งมีดังต่อไปนี้

1. เทคนิคการขยายคำสืบค้นแบบอัตโนมัติให้ค่า F-measure ดีกว่าเสิร์จเอนจิ้นพื้นฐาน

2. เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ให้ค่า F-measure ดีกว่าเสิร์จเอนจิ้นพื้นฐาน

3. เทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ให้ค่า F-measure ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ

4. เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ให้ค่า F-measure ดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์

5. เทคนิคการวิเคราะห์ลิงค์ให้ค่า F-measure ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

6. เทคนิคการวิเคราะห์ลิงค์ให้ค่า F-measure ดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

3.3.3 ตัวแปรที่ต้องการศึกษาและตัวแปรอื่นๆ

ตัวแปรต่างๆที่ต้องการศึกษาในการทดลองมีดังต่อไปนี้

1. ตัวแปรอิสระ (Independent Variable)

การค้นคืนในรูปแบบต่างๆ ได้แก่ BSE, AQE, LAQE, IQE, LIQE และ LA

2. ตัวแปรตาม (Dependent Variable)

2.1 จำนวนเว็บเพจผลลัพธ์ที่ค้นคืนกลับมาได้ในแต่ละรูปแบบการค้นคืน

2.2 จำนวนเว็บเพจที่เกี่ยวข้องกับคำถามและถูกค้นคืนขึ้นมาได้ในแต่ละรูปแบบการค้นคืน

2.3 ประสิทธิภาพในการค้นคืน ซึ่งประกอบไปด้วย ความแม่นยำในการค้นคืน ความระลึกในการค้นคืนและค่าถ่วงดุลของการค้นคืนซึ่งมีรายละเอียดดังต่อไปนี้

ค่าถ่วงดุลของการสืบค้น เป็นการวัดค่าเฉลี่ยความแม่นยำของระบบโดยใช้ทั้งค่าความแม่นยำในการสืบค้นและค่าความระลึกของการสืบค้น โดยค่าถ่วงดุลของการสืบค้นหาได้จากค่าเฉลี่ยฮาร์โมนิคของความแม่นยำในการสืบค้นและความระลึกของการสืบค้นดังนี้

(2*ความแม่นยำ*ความระลึก) / (ความแม่นยำ+ความระลึก)

3. ตัวแปรอื่นๆที่ต้องการศึกษา

เวลาที่ใช้ในการประมวลผลของทั้ง 6 รูปแบบการค้นคืน

4. ตัวแปรควบคุม (Control Variable)

4.1 การค้นคืนแบบเสิร์จเอนจิ้นพื้นฐานโดยที่ไม่ได้ใช้เทคนิคใดๆในการปรับปรับปรุงประสิทธิภาพการค้นคืน

4.2 ชุดคำถามที่ใช้ในการทดลอง

3.3.4 ชุดคำถามที่ใช้ในการทดลอง

ชุดข้อมูลที่ใช้ในการทดลองนี้เป็นชุดคำถามที่นิยมใช้ในการวิจัยทางด้านการสืบค้นระบบสารสนเทศซึ่งนำมาจาก TREC1, TREC-4 และ TREC-2003 ประกอบไปด้วยคำถามจำนวน

10 คำถามที่เกี่ยวข้องกับเรื่องทางด้านวิทยาการคอมพิวเตอร์ซึ่งอยู่ในโดเมนวิทยาศาสตร์และเทคโนโลยี ตัวอย่างคำถามที่นำมาใช้ในการทดลองนี้แสดงในตารางที่ 3.1 ชุดคำถามทั้ง 10 ข้อสามารถดูได้ในภาคผนวก ก.

ตารางที่ 3.1 ตัวอย่างคำถามที่เป็นภาษาอังกฤษและแปลเป็นภาษาไทยที่นำมาจาก TREC-1, TREC-4 และ TREC 2003 เพื่อใช้ในการทดลอง

Domain	Science and Technology
Question No.1	How Rewritable Optical Disks Work?
Description	Document describes the principles and mechanisms behind rewritable optical disk technology.
Hint	To be relevant, a document must describe how rewritable optical disk technology works at length and in significant and comprehensive technical detail.

หัวเรื่อง	วิทยาศาสตร์และเทคโนโลยี
คำถามที่ 1	แผ่นเก็บข้อมูล สามารถเขียนข้อมูลซ้ำหลายๆรอบได้อย่างไร
คำอธิบาย	เป็นเอกสารที่กล่าวถึงหลักการและกลไกลที่อยู่เบื้องหลังของเทคโนโลยีในการเขียนข้อมูลซ้ำของแผ่นเก็บข้อมูล
คำแนะนำ	เอกสารที่จะต้องมีการกล่าวถึงเทคโนโลยีการเขียนข้อมูลซ้ำของแผ่นเก็บข้อมูลว่ามีการทำงานอย่างไรถึงใช้ได้ โดยมีรายละเอียดทางเทคนิคมากพอสมควร

3.3.5 ผู้เข้าร่วมทำการทดลอง

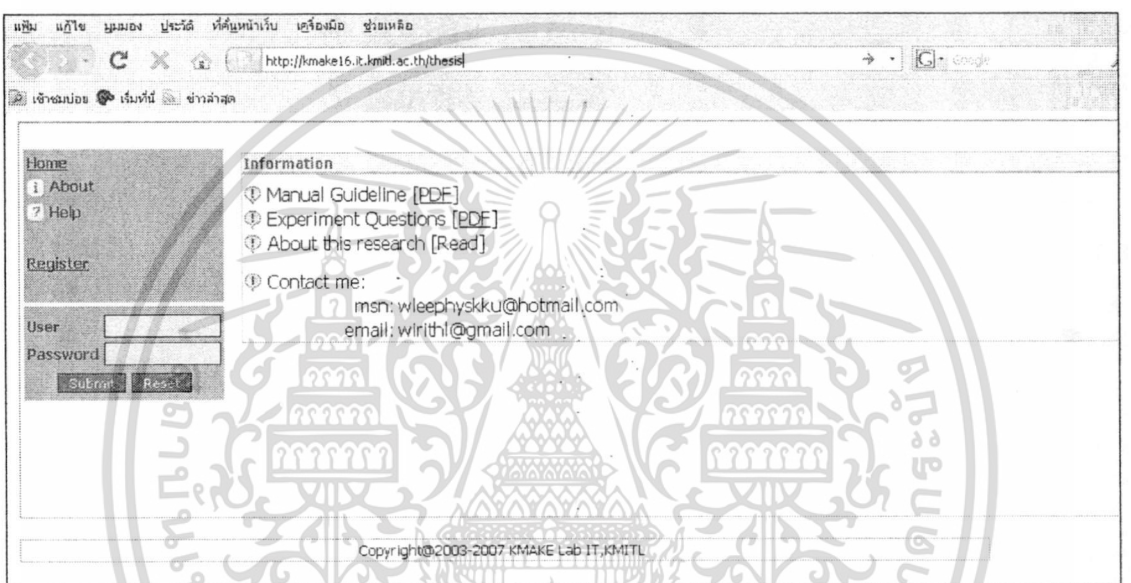
ในส่วนของผู้ร่วมทำการทดลองนั้นได้ใช้ผู้ร่วมทำการทดลองทั้งหมดจำนวน 24 คนซึ่งเป็นนักศึกษาระดับบัณฑิตศึกษาของคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

3.3.6 สิ่งแวดล้อมในการทดลอง

สิ่งแวดล้อมหลักในการทดลองที่ผู้ร่วมทำการทดลองต้องพบคือเว็บเสิร์จเอนจิ้น ดังนั้นผู้ทำการทดลองควรรู้ถึงขั้นตอนและวิธีการใช้งานของเว็บเสิร์จเอนจิ้น ซึ่งขั้นตอนและวิธีการใช้งานเว็บเสิร์จเอนจิ้นมีดังต่อไปนี้

1. เว็บเสิร์จเอนจิ้นที่ใช้สำหรับทำการทดลองสามารถเข้าใช้งานได้สองวิธี วิธีแรกคือเข้าใช้งานเว็บเสิร์จเอนจิ้นผ่านทางเครือข่ายอินทราเน็ตในห้องปฏิบัติการการจัดการองค์ความรู้และวิศวกรรมองค์ความรู้โดยใช้ URL ดังนี้ http://kmake16.it.kmitl.ac.th/thesis ส่วนวิธีที่สองคือเข้าใช้งานเว็บเสิร์จเอนจิ้นจากภายนอกสถาบันโดยผ่านทางเครือข่ายอินเตอร์เน็ตซึ่งมี URL ดังนี้ http://www.wirith.bkksmarthost.com/index.php

2. เมื่อเข้าใช้งานเว็บเสิร์จเอนจิ้นด้วยวิธีแรกหรือวิธีที่สองแล้วจะปรากฏหน้าหลักของเว็บเสิร์จเอนจิ้นดังที่แสดงในรูป 3.9



รูปที่ 3.9 หน้าหลักของเว็บเสิร์จเอนจิ้น

3. ทำการลงทะเบียนก่อนเริ่มเข้าใช้งานเว็บเสิร์จเอนจิ้น

4. ทำการล็อกอินเข้าใช้งาน เมื่อล็อกอินผ่านแล้วจะปรากฏหน้าหลักของการใช้งานดังที่แสดงในรูป 3.10 ซึ่งมีส่วนประกอบหลักดังนี้

4.1 ส่วนที่หนึ่ง (หมายเลข 1) ใช้ในการแสดงข้อมูลของผู้ใช้

4.2 ส่วนที่สอง (หมายเลข 2) ใช้ในการเลือกคำถาม (Question Search)

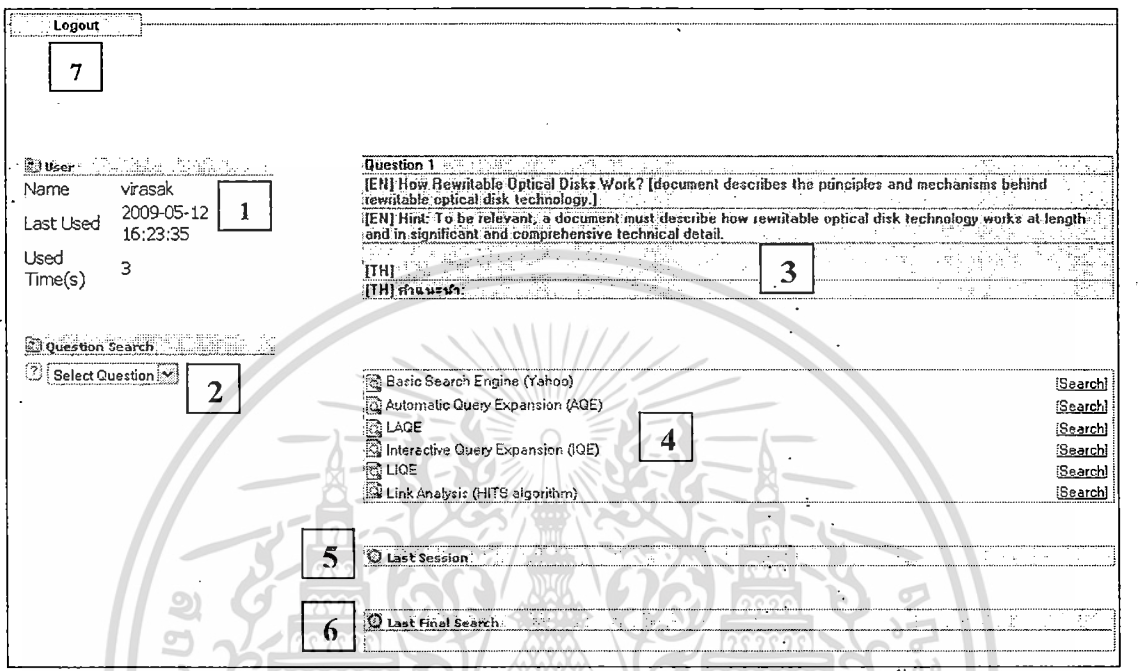
4.3 ส่วนที่สอง (หมายเลข 3) ใช้ในการแสดงรายละเอียดของคำถาม เช่น คำถาม, คำอธิบายของคำถามและคำแนะนำสำหรับเลือกเอกสารที่เกี่ยวข้อง

4.4 ส่วนที่สี่ (หมายเลข 4) ใช้ในการเลือกรูปแบบการค้นคืนซึ่งมีด้วยกันทั้งหมด 6 รูปแบบ

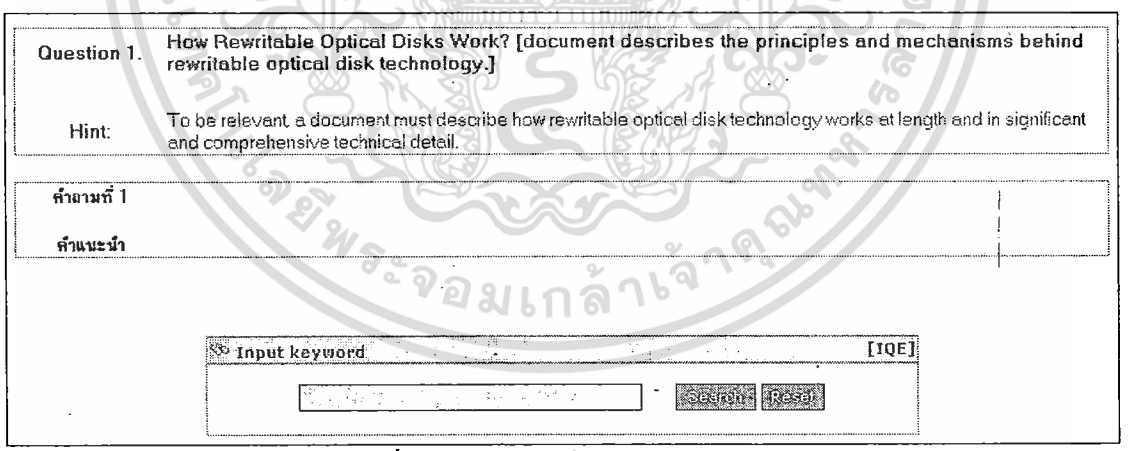
4.5 ส่วนที่ห้า (หมายเลข 5) ใช้แสดงผลการค้นคืนครั้งล่าสุดที่ยังไม่เสร็จสมบูรณ์ โดยผู้ร่วมทำการทดลองสามารถเรียกคืนกลับมาทำการทดลองต่อได้

4.6 ส่วนที่หก (หมายเลข 6) ใช้แสดงผลลัพธ์การค้นคืนที่เสร็จสมบรูณ์แล้วโดยผู้ร่วมทำการทดลองสามารถเรียกคืนกลับมาดูผลลัพธ์ได้

4.7 ส่วนที่เจ็ด (หมายเลข 7) ใช้ล็อกเอาทออกจากโปรแกรมเว็บเสิร์จเอนจิ้น



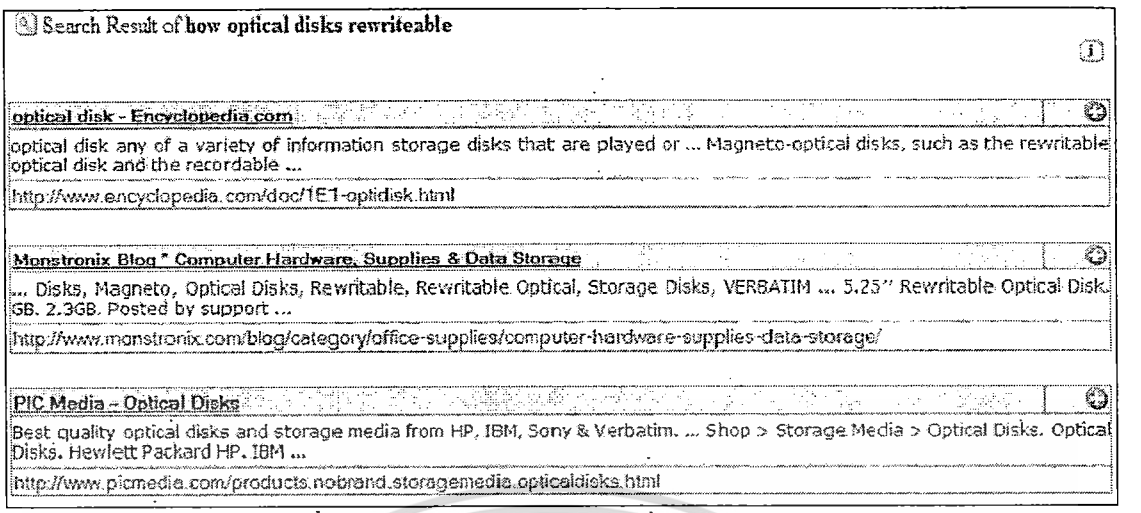
รูปที่ 3.10 ส่วนประกอบหน้าหลักของการใช้งานเว็บเสิร์จเอนจิ้น



รูปที่ 3.11 หน้าต่างที่ใช้ป้อนคำสืบค้น

5. การเลือกคำถามสามารถเลือกได้จาก Question Search (หมายเลข 2 ในรูปที่ 3.10)

6. เมื่อต้องการค้นคืนให้ทำการเลือกรูปแบบการค้นคืน (หมายเลข 4 ในรูปที่ 3.10) แล้วคลิกที่ [Search] เมื่อคลิกเลือกแล้วจะปรากฏหน้าต่างดังรูปที่ 3.11 จากนั้นให้ทำการป้อนคำสืบค้นลงในช่อง Input keyword แล้วกดปุ่มค้นหา (Search) ผลลัพธ์ที่ได้จากการค้นคืนด้วยเสิร์จเอนจิ้นจะแสดงผลออกมาดังรูปที่ 3.12



รูปที่ 3.12 แสดงผลลัพธ์บางส่วนที่ได้จากการค้นคืน

ผลลัพธ์ที่แสดงจะประกอบด้วย คำสืบค้นที่ใช้ (Keyword), หัวเรื่อง (Title) ของเว็บเพจ, คำอธิบาย (Description) ของเว็บเพจและที่อยู่ (Address) ของเว็บเพจ ในแต่ละเว็บเพจที่เป็นผลลัพธ์จะมีปุ่ม ปรากฏอยู่โดยปุ่มดังกล่าวจะใช้สำหรับเลือกเว็บเพจที่มีความเกี่ยวข้อง

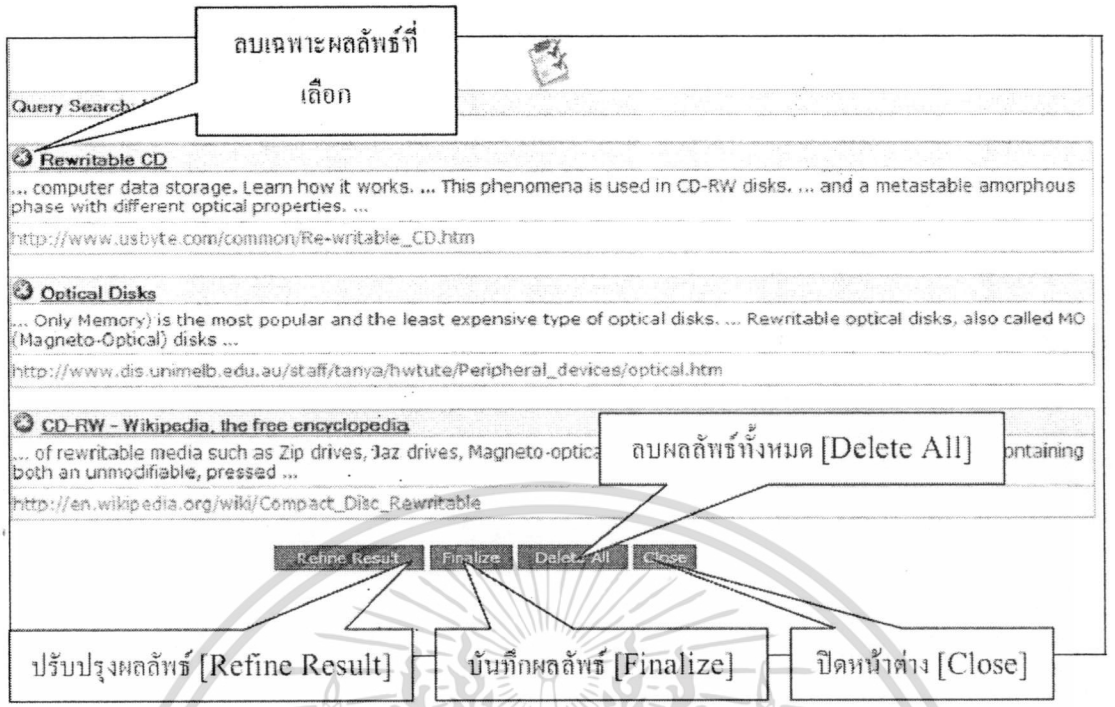
7. ในการเลือกเว็บเพจสามารถพิจารณาได้จากคำอธิบาย (Description) ของเว็บเพจหรือสามารถเข้าไปดูเว็บเพจได้โดยการคลิกที่หัวเรื่อง (Title) ของเว็บเพจ

8. ในกรณีที่กดปุ่ม เพื่อทำการเลือกเว็บเพจที่มีความเกี่ยวข้อง เมื่อทำการกดปุ่มแล้วจะปรากฏหน้าต่างแสดงเว็บเพจที่ถูกเลือกดังแสดงในรูปที่ 3.13

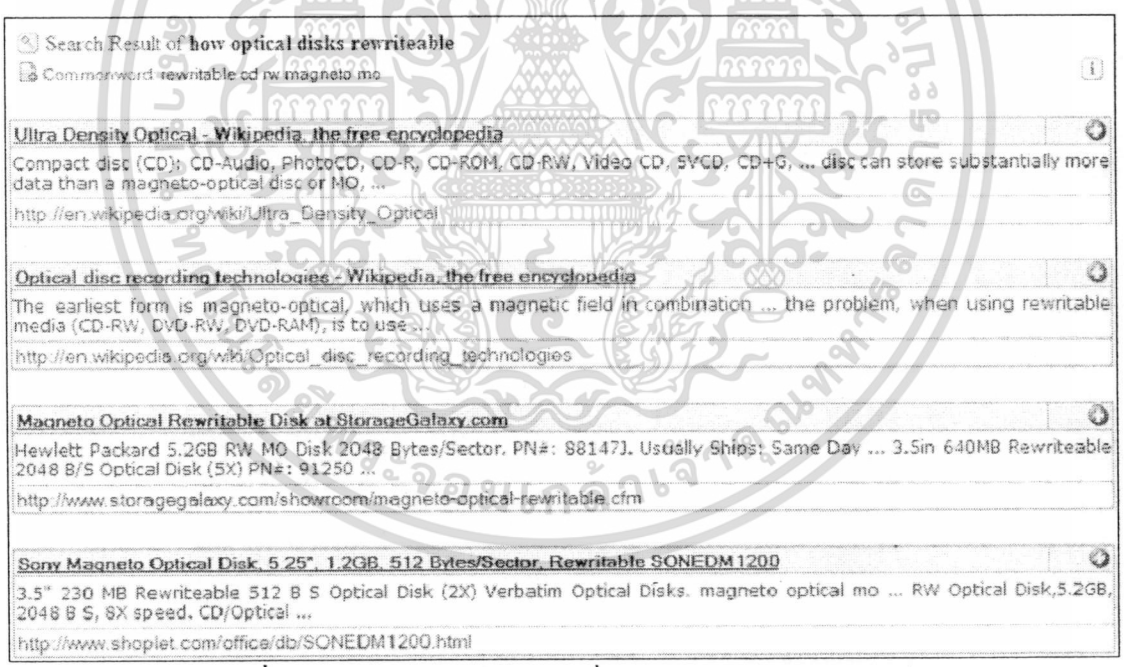
9. ในกรณีที่ต้องการลบเว็บเพจออกจากรายการสามารถทำได้โดยการกดปุ่ม (ในรูปที่ 3.13) ที่อยู่ด้านหน้าหัวเรื่อง (Title) ของเว็บเพจ

10. เมื่อทำการเลือกเว็บเพจที่มีความเกี่ยวข้องได้ตามจำนวนที่ต้องการแล้ว ให้ทำการกดปุ่มปรับปรุงผลลัพธ์ [Refine Result] (ในรูปที่ 3.13) จากนั้นผลลัพธ์ใหม่จะปรากฏ ดังแสดงในรูปที่ 3.14 ผลลัพธ์ที่แสดงจะประกอบด้วย คำสืบค้นที่ใช้ (Keyword), คำขยายที่ได้ (Common word), หัวเรื่อง (Title) ของเว็บเพจ, คำอธิบาย (Description) ของเว็บเพจและที่อยู่ (Address) ของเว็บเพจ ในแต่ละเว็บเพจที่เป็นผลลัพธ์จะมีปุ่ม ปรากฏอยู่โดยปุ่มดังกล่าวจะใช้สำหรับเลือกเว็บเพจที่มีความเกี่ยวข้อง

11. ในการปรับปรุงผลลัพธ์สามารถทำซ้ำได้หลายครั้งจนกว่าผลลัพธ์ที่ได้จะเป็นที่พอใจหรือจนกว่าจะไม่พบเว็บเพจใหม่ๆที่มีความเกี่ยวข้องปรากฏอยู่ในผลลัพธ์



รูปที่ 3.13 หน้าต่างแสดงรายการเว็บที่เกี่ยวข้องที่ถูกเลือกไว้

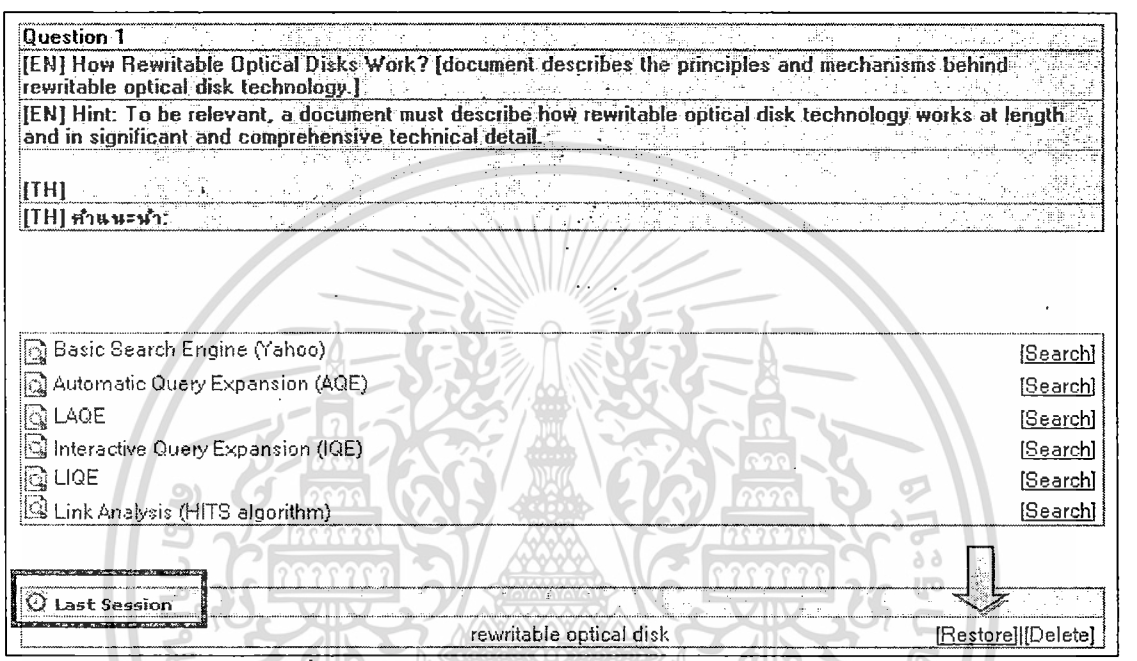


รูปที่ 3.14 แสดงผลลัพธ์บางส่วนที่ได้จากการปรับปรุงผลลัพธ์

12. เมื่อได้ผลลัพธ์ตามที่ต้องการแล้วให้ทำการบันทึกผลลัพธ์โดยการกดปุ่มบันทึก [Finalize]

13. กดปุ่มปิดหน้าต่าง [Close] (ในรูปที่ 3.13) เพื่อปิดหน้าต่างแสดงรายการเว็บเพจที่ถูกเลือกและหน้าต่างที่แสดงผลลัพธ์การค้นคืน

14. ในกรณีที่ผู้ร่วมทำการทดลองยังทำการทดลองไม่เสร็จสมบูรณ์ เสิร์จเอนจิ้นจะทำการบันทึกผลลัพธ์การทดลองล่าสุดเก็บไว้ เมื่อผู้ร่วมทำการทดลองเข้ามาใช้งานในครั้งถัดไป สามารถจะเรียกผลลัพธ์ล่าสุดขึ้นมาทำการทดลองต่อได้ดังรูปที่ 3.15 โดยที่กรอบ Last Session จะมีชื่อผลลัพธ์การทดลองที่ทำค้างไว้ เมื่อต้องการเรียกคืนผลลัพธ์ดังกล่าวให้คลิกที่ [Restore] ผลลัพธ์ที่ทำค้างไว้จะปรากฏขึ้นมา (ในรูปที่ 3.14)



รูปที่ 3.15 แสดงการเรียกคืนผลลัพธ์การทดลองล่าสุด

3.3.7 กระบวนการหรือขั้นตอนในการทดลอง

ขั้นตอนหรือกระบวนการในการทดลองมีดังต่อไปนี้

1. จัดหาผู้เข้าร่วมทำการทดลอง ซึ่งในการทดลองนี้ใช้นักศึกษาที่ลงเรียนวิชาการจัดการองค์ความรู้ประจำปีการศึกษา 2/2550

2. สถานที่ที่ใช้ทำการทดลองคือห้องปฏิบัติการการจัดการความองค์รู้และวิศวกรรมองค์ความรู้ (ถ้าผู้ร่วมทำการทดลองมาทำการทดลองที่คณะ) หรือทำการทดลองผ่านทางอินเตอร์เน็ตตาม URL ที่กำหนด (ถ้าผู้ร่วมทำการทดลองไม่สะดวกมาทำการทดลองที่คณะ)

3. ผู้ร่วมทำการทดลองทุกคนจะได้รับแจกคู่มือซึ่งประกอบด้วยชุดคำถามและวิธีการใช้งานเว็บเสิร์จเอนจิ้น

4. ผู้ร่วมทำการทดลองแต่ละคนทำความเข้าใจชุดคำถามและวิธีการใช้งานเว็บเสิร์จเอนจิ้น

5. ผู้ร่วมทำการทดลองเปิดโปรแกรมเว็บบราวเซอร์แล้วพิมพ์ URL ที่กำหนด

6. ผู้ร่วมทำการทดลองล็อกอินเข้าใช้งานเว็บเสิร์จเอนจิ้น

7. ผู้ร่วมทำการทดลองเลือกคำถามและอ่านคำถาม

8. ผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนซึ่งในแต่ละรูปแบบการค้นคืนจะมีกระบวนการหรือขั้นตอนดังนี้

8.1 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเสิร์จเอนจิ้นพื้นฐาน

8.1.1 ผู้ร่วมทำการทดลองป้อนคำสืบค้นจากนั้นส่งคำสืบค้นไปที่เว็บเสิร์จเอนจิ้น จากนั้นเว็บเสิร์จเอนจิ้นจะทำการประมวลผลและแสดงผลลัพธ์ที่ค้นคืนกลับมาได้

8.1.2 ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูล

8.2 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ

8.2.1 ผู้ร่วมทำการทดลองป้อนคำสืบค้นจากนั้นส่งคำสืบค้นไปที่เว็บเสิร์จเอนจิ้น เว็บเสิร์จเอนจิ้นจะประมวลผลและค้นคืนผลลัพธ์กลับมา

8.2.2 ผลลัพธ์ที่ค้นคืนกลับมาจะถูกนำไปประมวลผลโดยเริ่มจากการสกัดคำจากนั้นนำคำที่ได้ไปคำนวณหาค่าน้ำหนักและนำคำที่ผ่านการคำนวณค่าน้ำหนักไปเพิ่มลงในคำสืบค้นเดิมแล้วส่งคำสืบค้นใหม่ที่ได้ไปยังเสิร์จเอนจิ้น

8.2.3 เสิร์จเอนจิ้นทำการประมวลผลและแสดงผลลัพธ์ที่ค้นคืนกลับมาได้

8.2.4 ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูล

8.3 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบอัตโนมัติทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

8.3.1. เสิร์จเอนจิ้นนำผลลัพธ์ที่ได้จาก 8.2.4 มาทำการสกัดไฮเปอร์ลิงค์จากเนื้อหาของเว็บเพจ

8.3.2. เสิร์จเอนจิ้นนำไฮเปอร์ลิงค์ที่ได้มาทำการสร้างกราฟย่อย จากนั้นจะคำนวณค่าคะแนนความเกี่ยวข้องระหว่างเว็บเพจกับหัวเรื่องเพื่อใช้ในการจัดเรียงลำดับของเว็บเพจ

8.3.3. เสิร์จเอนจิ้นนำออธอริเพจที่มีค่าคะแนนสูงๆมาทำการสกัดคำจากเนื้อหาต่อจากนั้นนำคำที่ได้มาคำนวณหาค่าน้ำหนักของคำ

8.3.4. เสิร์จเอนจิ้นนำคำที่ได้จากการคำนวณค่าน้ำหนักมาเพิ่มลงในคำสืบค้นเดิมที่มีอยู่ จากนั้นส่งคำสืบค้นใหม่ที่ได้กลับไปค้นคืนอีกครั้ง

8.3.5. เสิร์จเอนจิ้นทำการประมวลผลและแสดงผลลัพธ์ที่ค้นคืนกลับมาได้

8.3.6. ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูล

8.4 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์

8.4.1 ผู้ร่วมทำการทดลองป้อนคำสืบค้นจากนั้นส่งคำสืบค้นไปที่เว็บเสิร์จเอนจิ้น เว็บเสิร์จเอนจิ้นจะประมวลผลและค้นคืนผลลัพธ์กลับมา

8.4.2 ผู้ร่วมทำการทดลองพิจารณาเลือกเว็บเพจที่มีความเกี่ยวข้องจากผลลัพธ์ที่ถูกค้นคืนกลับมา เมื่อเลือกครบตามจำนวนที่ต้องการแล้วจากนั้นส่งเว็บเพจที่เลือกกลับไปยังเสิร์จเอนจิ้น

8.4.3 เสิร์จเอนจิ้นนำเว็บเพจที่มีความเกี่ยวข้องที่ผู้ร่วมทำการทดลองส่งกลับเข้ามาไปประมวลผลโดยการสกัดคำจากเนื้อหาของเว็บเพจที่เกี่ยวข้อง เมื่อได้คำแล้วจะถูกนำไปคำนวณเพื่อหาค่าน้ำหนักของคำ

8.4.4 เสิร์จเอนจิ้นนำคำที่ได้จากการคำนวณค่าน้ำหนักมาเพิ่มลงในคำสืบค้นเดิมที่มีอยู่จากนั้นส่งคำสืบค้นใหม่ที่ได้กลับไปสืบค้นอีกครั้ง

8.4.5 เสิร์จเอนจิ้นทำการประมวลผลและแสดงผลลัพธ์ที่ค้นคืนกลับมาได้

8.4.6 ถ้าผลลัพธ์ที่ได้ยังไม่เป็นที่พอใจผู้ร่วมทำการทดลองสามารถทำซ้ำในขั้นตอนที่ 8.4.2 ได้หลายครั้งจนกว่าผลลัพธ์ที่ได้จะเป็นที่พอใจหรือจนกว่าจะไม่พบเว็บเพจใหม่ๆที่มีความเกี่ยวข้องปรากฏอยู่ในผลลัพธ์

8.4.7 เมื่อได้ผลลัพธ์เป็นที่พอใจแล้วให้ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูล

8.5 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์

8.5.1 เสิร์จเอนจิ้นนำผลลัพธ์ที่ได้จากข้อ 8.4.7 มาทำการสกัดไฮเปอร์ลิงค์จากเนื้อหาของเว็บเพจ

8.5.2 เสิร์จเอนจิ้นนำไฮเปอร์ลิงค์ที่ได้มาทำการสร้างกราฟย่อย จากนั้นจะคำนวณค่าคะแนนความเกี่ยวข้องระหว่างเว็บเพจกับหัวเรื่องเพื่อใช้ในการจัดเรียงลำดับของเว็บเพจ

8.5.3 เสิร์จเอนจิ้นนำออธอริเพจที่มีค่าคะแนนสูงๆมาทำการสกัดคำจากเนื้อหาต่อจากนั้นนำคำที่ได้มาคำนวณหาค่าน้ำหนักของคำ

8.5.4 เสิร์จเอนจิ้นนำคำที่ได้จากการคำนวณค่าน้ำหนักมาเพิ่มลงในคำสืบค้นเดิมที่มีอยู่ จากนั้นส่งคำสืบค้นใหม่ที่ได้กลับไปสืบค้นอีกครั้ง

8.5.5 เสิร์จเอนจิ้นทำการประมวลผลและแสดงผลลัพธ์ที่ค้นคืนกลับมาได้

8.5.6 ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูล

8.6 ถ้าผู้ร่วมทำการทดลองเลือกรูปแบบการค้นคืนด้วยเทคนิคการวิเคราะห์ลิงค์

8.6.1 เสิร์จเอนจิ้นนำผลลัพธ์ที่ได้จากข้อ 8.1.2 มาทำการสกัดไฮเปอร์ลิงค์จากเนื้อหาของเว็บเพจ

8.6.2 เสิร์จเอนจิ้นนำไฮเปอร์ลิงค์ที่ได้มาทำการสร้างกราฟย่อย จากนั้นคำนวณค่าคะแนนความเกี่ยวข้องระหว่างเว็บเพจกับหัวเรื่องเพื่อใช้ในการจัดเรียงลำดับเว็บเพจ

8.6.3 เสิร์จเอนจิ้นแสดงผลลัพธ์ของออธอริทีเพจ

8.6.4 ผู้ร่วมทำการทดลองบันทึกผลลัพธ์ที่ได้ลงในฐานข้อมูล

9. ทำตามขั้นตอน 8.1-8.6 จนครบทั้ง 10 คำถาม แต่ในการทดลองนี้กำหนดให้ผู้ร่วมทำการทดลองทำการทดลองเฉพาะในรูปแบบการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์เท่านั้น โดยข้อมูลของรูปแบบการค้นคืนแบบอื่นๆจะใช้ข้อมูลที่ได้จากการค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์สร้างขึ้นมา

10. ในการทำการทดลองทั้ง 10 คำถามนั้น ผู้ทำการทดลองไม่จำเป็นต้องทำให้เสร็จภายในครั้งเดียว แต่สามารถทำต่อเนื่องได้เรื่อยๆ จนกระทั่งเสร็จครบทั้ง 10 คำถาม โดยให้ระยะเวลาในการทำการทดลอง 3 สัปดาห์ สำหรับเสิร์จเอนจิ้นที่ใช้ในการทดลองนี้สามารถทำการบันทึกผลลัพธ์ล่าสุดที่ทำค้างไว้ได้และเมื่อกลับเข้ามาใช้งานอีกครั้งสามารถจะเรียกผลลัพธ์ล่าสุดขึ้นมาทำการทดลองต่อได้ (ดูได้ในหัวข้อ 3.3.7 สภาพแวดล้อมการทดลอง)

11. นำ 30 เว็บเพจแรกของผลลัพธ์ที่ได้จากข้อ 8.1.2, 8.2.4, 8.3.6, 8.4.7, 8.5.6 และ 8.6.4 ของผู้ร่วมทำการทดลองแต่ละคนของทั้ง 10 คำถามส่งให้ผู้เชี่ยวชาญทำการประเมินเพื่อหาจำนวนเว็บเพจที่มีความเกี่ยวข้องกับคำถาม

12. นำผลลัพธ์ที่ได้จากข้อ 10 มาทำการคำนวณเพื่อหาประสิทธิภาพการค้นคืนของแต่ละรูปแบบการค้นคืน ประสิทธิภาพของการค้นคืนจะพิจารณาจากค่าเฉลี่ยหรือค่า F-measure ซึ่งค่าทั้งสามสามารถหาได้ดังนี้

กำหนดให้

P แทน ค่าความแม่นยำ

R แทน ค่าความระลึก

F-measure แทน ค่าเฉลี่ย

R_A แทน จำนวนเว็บเพจที่เกี่ยวข้องกับคำถามและถูกสืบค้นขึ้นมาได้

N แทน จำนวนเว็บเพจที่นำมาประเมิน

R_C แทน จำนวนเว็บเพจที่เกี่ยวข้องกับคำถามทั้งหมดที่พบในแต่ละคำถาม

R_M แทน จำนวนเว็บเพจที่เกี่ยวข้องกับคำถามที่พบในแต่ละ

รูปแบบการสืบค้น

$$P = \frac{R_A}{N} \tag{3.4}$$

$$R = \frac{R_M}{R_C} \tag{3.5}$$

$$F - measure = \frac{2 \times P \times R}{P + R} \tag{3.6}$$

3.3.8 จำนวนการทำรายการทั้งหมดในการทดลอง

จำนวนการทำรายการทั้งหมดที่ได้จากการทดลองมีจำนวนทั้งสิ้น 1338 รายการ โดยแต่ละรายการ มีข้อมูลเพื่อนำมาวิเคราะห์ดังต่อไปนี้

1. จำนวนเว็บเพจที่เกี่ยวข้องกับคำถามและถูกคืนค้นขึ้นมาได้ในแต่ละรูปแบบการค้นคืน
2. เวลาที่ใช้ในการประมวลผลในแต่ละรูปแบบการสืบค้น

บทที่ 4
ผลการทดลองและวิเคราะห์เปรียบเทียบประสิทธิภาพการค้นคืน

ในบทนี้จะกล่าวถึงผลการทดลองและการอภิปรายผลการเปรียบเทียบประสิทธิภาพของการค้นคืนทั้ง 6 รูปแบบ ได้แก่ การค้นคืนโดยเสิร์จเอนจิ้นทั่วไป (BSE), การค้นคืนโดยใช้เทคนิคการขยายคำสืบค้นทั้งแบบอัตโนมัติและแบบปฏิสัมพันธ์ (AQE, IQE), การค้นคืนโดยใช้เทคนิคการวิเคราะห์ลิงค์ทำงานร่วมกับเทคนิคการขยายคำสืบค้นทั้งแบบอัตโนมัติและแบบปฏิสัมพันธ์ (LAQE, LIQE) และการค้นคืนโดยใช้เทคนิคการวิเคราะห์ลิงค์ (LA) โดยมีลำดับการนำเสนอดังนี้ หัวข้อแรกเป็นการนำเสนอการเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลของทั้ง 6 รูปแบบการค้นคืน การนำเสนอผลการเปรียบเทียบค่าเฉลี่ยหรือค่า F-measure ของทั้ง 6 รูปแบบการค้นคืนถูกนำเสนอในหัวข้อที่สอง ในหัวข้อที่สามนำเสนอเกี่ยวกับการทดสอบสมมติฐานการทดลอง (Hypothesis testing) และหัวข้อสุดท้ายคือหัวข้อที่สี่เป็นการอภิปรายผลการทดลอง

4.1 การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลของทั้ง 6 รูปแบบการค้นคืน

การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลของทั้ง 6 รูปการค้นคืน ได้แก่ การค้นคืนด้วยเสิร์จเอนจิ้นพื้นฐาน (BSE), การค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE), การค้นคืนด้วยเทคนิคการขยายคำสืบค้นที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE), การค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE), การค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LIQE) และเทคนิคการวิเคราะห์ลิงค์ (LA) แสดงในตารางที่ 4.1 เมื่อพิจารณาจากตารางที่ 4.1 พบว่าเวลาที่ใช้ในการประมวลผลของเทคนิคการขยายคำสืบค้นที่ถูกปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์จะใช้เวลามากกว่าเสิร์จเอนจิ้นพื้นฐาน, เทคนิคการขยายคำสืบค้นที่ยังไม่ได้ถูกปรับปรุงและเทคนิคการวิเคราะห์ลิงค์เนื่องจากเทคนิคการขยายคำสืบค้นที่ถูกปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์จะต้องใช้เวลาในการประมวลผล 3 ขั้นตอนหลักได้แก่

1. การคำนวณฮิทส์อัลกอริทึม
2. การสกัดหรือการดึงคำจากเว็บเพจและการคำนวณค่าน้ำหนักของคำ
3. การสร้างคิวรี่ใหม่

ตารางที่ 4.1 แสดงเวลาเฉลี่ยที่ใช้ในการประมวลผลของทั้ง 6 รูปแบบการค้นคืนใน 10 คำถาม

	รูปแบบการค้นคืน					
	BSE	AQE	LAQE	IQE	LIQE	LA
เวลาเฉลี่ย (วินาที)	8	14	47	201	230	30

หมายเหตุ เวลาเฉลี่ยในการประมวลผลของ IQE ที่ปรากฏในตารางเป็นเวลาเฉลี่ยที่รวมช่วงระยะเวลาที่ผู้ใช้ทำการเลือกเว็บเพจที่มีความเกี่ยวข้องก่อนที่จะป้อนกลับเข้าสู่ระบบค้นคืน

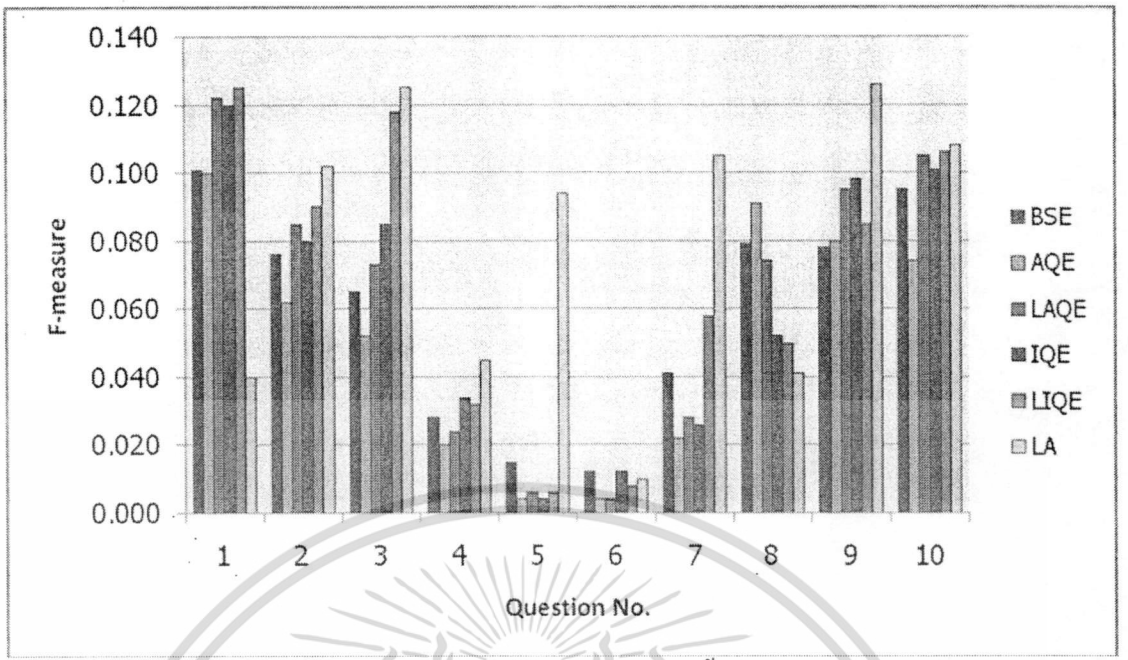
4.2 ผลการเปรียบเทียบประสิทธิภาพในการค้นคืนของทั้ง 6 รูปแบบการค้นคืน

ในการเปรียบเทียบประสิทธิภาพการค้นคืนของทั้ง 6 รูปแบบการค้นคืนจะใช้ค่าเฉลี่ยหรือที่เรียกว่า F-measure ซึ่งเป็นค่าเฉลี่ยโดยรวมระหว่างความแม่นยำและความระลึก ซึ่งในการคำนวณค่า F-measure จะใช้สมการที่ 3.6 โดยผลของการเปรียบเทียบค่า F-measure แสดงในตารางที่ 4.2

ตารางที่ 4.2 แสดงค่า F-measure ของทั้ง 6 รูปแบบการค้นคืนใน 10 คำถาม

ค่า F-measure						
คำถามข้อที่	รูปแบบการค้นคืน					
	BSE	AQE	LAQE	IQE	LIQE	LA
1	0.101	0.100	0.122	0.120	0.125	0.040
2	0.076	0.062	0.085	0.080	0.090	0.102
3	0.065	0.052	0.073	0.085	0.118	0.125
4	0.028	0.020	0.024	0.034	0.032	0.045
5	0.015	0.004	0.006	0.004	0.006	0.094
6	0.012	0.004	0.004	0.012	0.008	0.010
7	0.041	0.022	0.028	0.026	0.058	0.105
8	0.079	0.091	0.074	0.052	0.050	0.041
9	0.078	0.080	0.095	0.098	0.085	0.126
10	0.095	0.074	0.105	0.101	0.106	0.108
เฉลี่ย	0.059	0.051	0.062	0.061	0.068	0.080

เมื่อนำค่า F-measure จากตารางที่ 4.2 มาสร้างกราฟเพื่อเปรียบเทียบในแต่ละรูปแบบการค้นคืนของทั้ง 10 คำถามจะได้ดังรูปที่ 4.1



รูปที่ 4.1 กราฟแสดงการเปรียบเทียบค่าเฉลี่ย F-measure ของทั้ง 6 รูปแบบการค้นคืนใน 10 คำถาม

เมื่อพิจารณาจากตารางที่ 4.2 และกราฟรูปที่ 4.1 พบว่า

- เมื่อเปรียบเทียบค่าเฉลี่ยระหว่างเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ยังไม่ได้ปรับปรุง (AQE) และเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE) จะเห็นได้ว่า 8 คำถามใน 10 คำถาม (คำถามที่ 1, 2, 3, 4, 5, 7, 9, 10) เทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE) ให้ค่าเฉลี่ย F-measure ที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ยังไม่ได้ปรับปรุง (AQE)
- เมื่อเปรียบเทียบค่าเฉลี่ย F-measure ของทั้ง 10 คำถามพบว่าค่าเฉลี่ย F-measure ของเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE) มีค่ามากกว่าค่าเฉลี่ย F-measure ของเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ยังไม่ได้ปรับปรุง (AQE) ประมาณ 21.57 % ซึ่งแสดงว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE) มีประสิทธิภาพในการค้นคืนที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ยังไม่ได้ปรับปรุง (AQE)
- เมื่อเปรียบเทียบค่าเฉลี่ย F-measure ระหว่างเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ยังไม่ได้ปรับปรุง (IQE) และเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LIQE) จะเห็นได้ว่า 6 คำถามใน 10 คำถาม (คำถามที่ 1, 2, 3, 5, 7, 10) เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการ

วิเคราะห์ลิงค์ (LIQE) ให้ค่าเฉลี่ย F-measure ที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ยังไม่ได้ปรับปรุง (IQE)

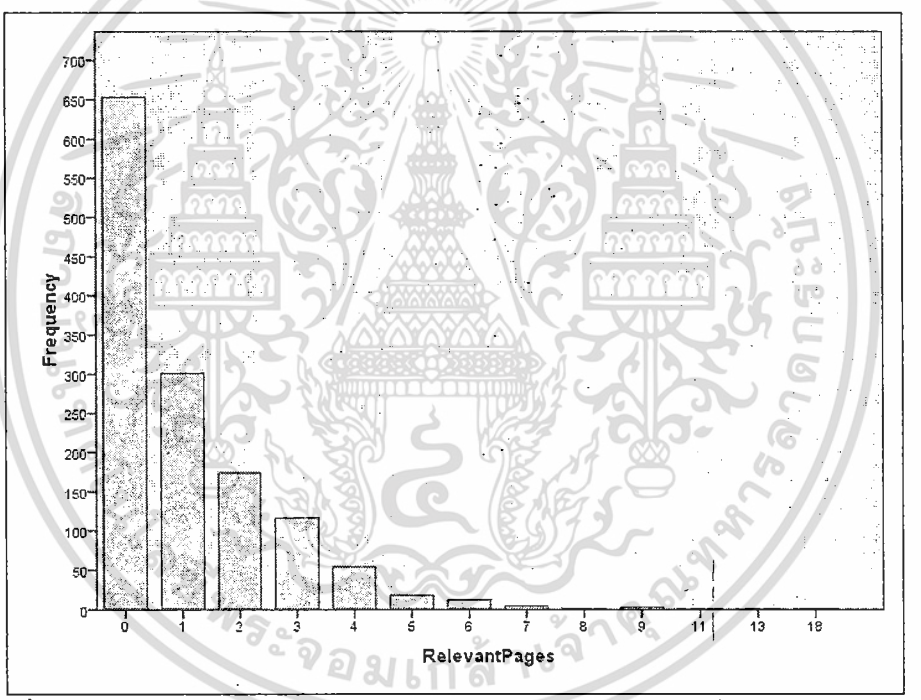
- เมื่อเปรียบเทียบค่าเฉลี่ย F-measure ของทั้ง 10 คำถามพบว่าค่าเฉลี่ย F-measure ของเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LIQE) มีค่ามากกว่าค่าเฉลี่ย F-measure ของเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ยังไม่ได้ปรับปรุง (IQE) ประมาณ 11.48 % ซึ่งแสดงว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LIQE) มีประสิทธิภาพในการค้นคืนที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ยังไม่ได้ปรับปรุง (IQE)
- เทคนิคการวิเคราะห์ลิงค์ (LA) ให้ค่า F-measure ดีที่สุดเป็นอันดับแรกใน 7 คำถาม (คำถามที่ 2, 3, 4, 5, 7, 9, 10) และผลจากการคำนวณค่าเฉลี่ย F-measure ทั้ง 10 คำถาม พบว่าเทคนิคการวิเคราะห์ลิงค์ (LA) ให้ค่า F-measure ดีที่สุดเป็นอันดับหนึ่งโดยมีค่าเท่ากับ 0.080 ซึ่งแสดงว่าเทคนิคการวิเคราะห์ลิงค์ (LA) ให้ประสิทธิภาพในการค้นคืนดีที่สุด
- เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LIQE) ให้ค่าเฉลี่ย F-measure ของทั้ง 10 คำถามที่ดีที่สุดรองลงมาเป็นอันดับที่สองโดยมีค่าเท่ากับ 0.068 ในขณะที่เทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ (LAQE) ให้ค่าเฉลี่ย F-measure ที่ดีที่สุดรองลงมาเป็นอันดับที่สามโดยมีค่าเท่ากับ 0.062

4.3 การทดสอบสมมติฐาน (Hypothesis Testing)

การทดสอบสมมติฐานเป็นขั้นตอนหนึ่งที่ใช้ในการตรวจสอบเพื่อแสดงให้เห็นว่าคำตอบหรือข้อค้นพบที่คาดคะเนไว้ตรงกับคำตอบที่ได้จากข้อมูลที่มีอยู่จริงหรือไม่โดยอาศัยการเขียนอธิบายข้อเท็จจริงในรูปของสัญลักษณ์ทางคณิตศาสตร์ที่เกี่ยวข้องกับค่าพารามิเตอร์ของประชากร (Population parameter) ข้อสมมติที่กำหนดขึ้นอาจจะจริงหรือเท็จไม่สามารถทราบได้อย่างแน่นอน ดังนั้นจึงจำเป็นต้องมีการสุ่มตัวอย่างมาทำการทดสอบแล้วนำค่าสถิติที่ได้จากกลุ่มตัวอย่างมาใช้ในการตัดสินใจว่าข้อสมมติฐานที่กำหนดนั้นถูกต้องหรือไม่ ในการทดสอบสมมติฐานจะแทนสมมติฐานด้วย H ซึ่งสมมติฐานที่จะทดสอบเรียกว่าสมมติฐานเพื่อการทดสอบหรือสมมติฐานหลัก (Null hypothesis) และแทนด้วย H_0 ส่วนสมมติฐานที่แย้งกับสมมติฐานหลักและนำมาพิจารณาในการทดสอบ H_0 ด้วยเรียกว่าสมมติฐานแย้งหรือสมมติฐานรอง (Alternative hypothesis) ซึ่งแทนด้วย H_1

สถิตินอนพาราเมตริก (Nonparametric Statistics) เป็นสถิติอนุมานแบบไม่ใช้พารามิเตอร์โดยที่ไม่จำเป็นต้องมีรูปแบบการแจกแจงความน่าจะเป็นของประชากรในรูปแบบใดรูปแบบหนึ่ง สามารถใช้ได้กับข้อมูลที่มีระดับการวัดต่ำตั้งแต่มาตรานามบัญญัติ (Nominal Scale) ที่นับเป็นความถี่ได้และเป็นมาตราเรียงอันดับ (Ordinal Scale) หรือตัวเลขใดๆ (Interval Scale or Ratio Scale) ที่สามารถนำมาจัดอันดับที่ (Rank) ได้และใช้ได้ผลดีกับตัวอย่างที่มีขนาดเล็กอีกทั้งยังคำนวณได้ง่ายไม่ยุ่งยากซับซ้อน สามารถคำนวณได้รวดเร็ว

ในการทดสอบสมมติฐานของการทดลองนี้เลือกใช้สถิตินอนพาราเมตริกเนื่องจากข้อมูลที่ได้จากการทดลองเมื่อนำมาวาดกราฟเพื่อดูลักษณะการแจกแจงของข้อมูลพบว่ามีลักษณะการแจกแจงไม่เป็นแบบโค้งปกติดังรูปที่ 4.2 นอกจากนั้นในการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของประชากรที่มากกว่าสองกลุ่มยังไม่ทราบถึงค่าความแปรปรวนของกลุ่มประชากรที่แน่ชัด



รูปที่ 4.2 แสดงการแจกแจงข้อมูลไม่เป็นแบบโค้งปกติของข้อมูลที่ได้จากการทดลอง สามารถใช้ได้กับข้อมูลที่อยู่ในมาตรวัดตั้งแต่นามบัญญัติ (Nominal) ขึ้นไป

ขั้นตอนในการทดสอบสมมติฐานมีดังต่อไปนี้

ขั้นตอนในการทดสอบสมมติฐานจะประกอบด้วยการตั้งสมมติฐานทางสถิติ การกำหนดนัยสำคัญของการทดสอบ สถิติที่เลือกใช้ในการทดสอบ กฎการตัดสินใจ การคำนวณด้วยสถิติวิธีที่เลือกใช้และการตัดสินใจเกี่ยวกับการทดสอบสมมติฐาน ซึ่งแต่ละขั้นตอนมีรายละเอียดดังต่อไปนี้

4.3.1 สมมติฐาน

1. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย BSE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE คือ $\mu_{BSE} \geq \mu_{AQE}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย BSE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE คือ $\mu_{BSE} < \mu_{AQE}$

2. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย BSE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE คือ $\mu_{BSE} \geq \mu_{IQE}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย BSE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE คือ $\mu_{BSE} < \mu_{IQE}$

3. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE คือ $\mu_{AQE} \geq \mu_{IQE}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE คือ $\mu_{AQE} < \mu_{IQE}$

4. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LAQE คือ $\mu_{AQE} \geq \mu_{LAQE}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย AQE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LAQE คือ $\mu_{AQE} < \mu_{LAQE}$

5. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LIQE คือ $\mu_{IQE} \geq \mu_{LIQE}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย IQE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LIQE คือ $\mu_{IQE} < \mu_{LIQE}$

6. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LAQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LA คือ $\mu_{LAQE} \geq \mu_{LA}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LAQE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LA คือ $\mu_{LAQE} < \mu_{LA}$

7. H_0 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LIQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LA คือ $\mu_{LIQE} \geq \mu_{LA}$

   H_1 : จำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LIQE มีค่าน้อยกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกค้นคืนโดย LA คือ $\mu_{LIQE} < \mu_{LA}$

4.3.2 กำหนดนัยสำคัญของการทดสอบ

α= 0.1, 0.2, 0.3

4.3.3 สถิติที่เลือกใช้ในการทดสอบ

สถิติทดสอบวิลคอกซันจับคู่เครื่องหมายตำแหน่ง (The Wilcoxon Matched Pairs Signed-Rank Test) เป็นสถิตินอนพาราเมตริกวิธีหนึ่งที่พัฒนามาจาก Sign test เพื่อใช้ในการทดสอบความแตกต่างระหว่างสองกลุ่มที่ไม่เป็นอิสระต่อกันหรือมีความสัมพันธ์กัน โดยนำเอาขนาดของความแตกต่างของข้อมูลแต่ละคู่มาคิดอันดับและทำการคำนวณค่าสถิติ

4.3.4 กฎการตัดสินใจ

1. ที่ $\alpha = 0.1$ ถ้า Z ที่คำนวณได้มีค่าน้อยกว่า $-Z_{0.1}$ จะปฏิเสธ H_0 และยอมรับ H_1
2. ที่ $\alpha = 0.2$ ถ้า Z ที่คำนวณได้มีค่าน้อยกว่า $-Z_{0.2}$ จะปฏิเสธ H_0 และยอมรับ H_1
3. ที่ $\alpha = 0.3$ ถ้า Z ที่คำนวณได้มีค่าน้อยกว่า $-Z_{0.3}$ จะปฏิเสธ H_0 และยอมรับ H_1

4.3.5 ผลการคำนวณสถิติทดสอบวิลคอกซันจับคู่เครื่องหมายตำแหน่ง

การคำนวณสถิติทดสอบวิลคอกซันจับคู่เครื่องหมายตำแหน่งมีขั้นตอนดังนี้

1. หาความแตกต่างของข้อมูลแต่ละคู่โดยคิดเครื่องหมาย (กำหนดเป็นค่า d_i เมื่อ i =1,2,3,..,N และ N เป็นจำนวนคู่หรือขนาดของกลุ่มตัวอย่าง)
2. นำค่าความแตกต่าง (d_i) มาจัดอันดับ โดยพิจารณาตัวเลขค่าสัมบูรณ์ของ

ความแตกต่างของข้อมูลแต่ละคู่ (คือไม่คิดเครื่องหมาย)

- ให้อันดับความแตกต่างของข้อมูลที่น้อยที่สุดเป็นอันดับ 1
- กรณีที่ความแตกต่างของข้อมูลมีค่าเท่ากัน ให้ใช้การเฉลี่ยอันดับ
- สำหรับคู่ของข้อมูลที่มีความแตกต่างเท่ากับค่าศูนย์ ($d_i = 0$) จะไม่นำมาคิดอันดับ

3. บันทึกเครื่องหมายของอันดับตามเครื่องหมายของ d_i
4. หาผลรวมของอันดับ โดยแยกเป็น ผลรวมของอันดับที่มีเครื่องหมายบวก และผลรวมของอันดับที่มีเครื่องหมายลบ
5. ให้ค่าผลรวมของอันดับที่มีค่าน้อยกว่า (ไม่คิดเครื่องหมาย) เป็นค่า T ที่จะใช้ในการทดสอบ
6. นับจำนวนอันดับที่มีอยู่ทั้งหมด ให้เป็น N
   - ในกรณีที่คู่ลำดับใดๆ มีค่าความแตกต่างเท่ากับศูนย์ ($d_i = 0$) จะไม่นับคู่ลำดับนั้นๆ
7. กรณีที่กลุ่มตัวอย่างมีขนาดใหญ่ (N > 25) การแจกแจงของกลุ่มตัวอย่างจะมีลักษณะใกล้เคียงกับการแจกแจงปกติ จะทำการเปลี่ยนค่า T เป็นค่า Z ดังนี้

ค่าเฉลี่ย (Mean) หรือ μ

$$\mu = \frac{N*(N+1)}{4} \tag{4.1}$$

ค่าส่วนเบี่ยงเบนมาตรฐาน (STDEV)

$$\sigma = \sqrt{\frac{N*(N+1)*(2N+1)}{24}} \tag{4.2}$$

ค่ามาตรฐาน (Z)

$$Z = \frac{T-\mu}{\sigma} = \frac{T - \left(\frac{N*(N+1)}{4}\right)}{\sqrt{\frac{N*(N+1)*(2N+1)}{24}}} \tag{4.3}$$

โดยที่ N คือ จำนวนอันดับที่นำมาใช้ในการคำนวณ

T คือ ผลรวมอันดับที่มีผลรวมของอันดับที่นำมาใช้ในการทดสอบคือค่าผลรวมอันดับที่มีค่าน้อยสุด

ตารางที่ 4.3 แสดงค่าต่างๆที่ใช้ในการคำนวณสถิติทดสอบวิลคอกซันจับคู่เครื่องหมายตำแหน่ง

H_1	N	Sum Rank		R_{min}	Mean	STDEV
		R_+	R_-			
$\mu_{BSE} < \mu_{AQE}$	116	2483.5	4302.5	2483.5	3393	362.9897
$\mu_{BSE} < \mu_{IQE}$	108	2510	3376	2510	2943	326.2491
$\mu_{AQE} < \mu_{IQE}$	109	3292	2703	2703	2997.5	330.7699
$\mu_{AQE} < \mu_{LAQE}$	116	3936.5	2849.5	2849.5	3393	362.9897
$\mu_{IQE} < \mu_{LIQE}$	106	3061.5	2609.5	2609.5	2835.5	317.2700
$\mu_{LAQE} < \mu_{LA}$	143	6515.5	3780.5	3780.5	5148	496.2318
$\mu_{LIQE} < \mu_{LA}$	142	6550.5	3602.5	3602.5	5076.5	491.0537

ผลการคำนวณค่าผลรวมของอันดับที่มีเครื่องหมายบวกและผลรวมรวมของอันดับที่มีเครื่องหมายลบ (Sum Rank: R+, R-) ผลรวมของอันดับที่นำมาใช้ในการทดสอบคือค่าผลรวมอันดับที่มีค่าน้อยสุด (R_{min}) แสดงไว้ในตารางที่ 4.3 ส่วนผลการคำนวณค่ามาตรฐาน (Z) และค่า p-value แสดงไว้ในตารางที่ 4.4

ตารางที่ 4.4 แสดงผลลัพธ์ค่า Z_{cal} และค่า p-value ที่ได้จากการคำนวณ

H_1	Zcal	p-value	Z		
			α=-0.1	α=-0.2	α=-0.3
$\mu_{BSE} < \mu_{AQE}$	-2.5056	0.0061	-1.2815	-0.8416	-0.5245
$\mu_{BSE} < \mu_{IQE}$	-1.3272	0.0922			
$\mu_{AQE} < \mu_{IQE}$	-0.8903	0.1867			
$\mu_{AQE} < \mu_{LAQE}$	-1.4973	0.0672			
$\mu_{IQE} < \mu_{LIQE}$	-0.7123	0.2381			
$\mu_{LAQE} < \mu_{LA}$	-2.7558	0.0029			
$\mu_{LIQE} < \mu_{LA}$	-3.0017	0.0013			

ค่า P-value

P-value หรือ P คือระดับนัยสำคัญ α ที่น้อยที่สุด ที่ H_0 จะถูกปฏิเสธ จากนิยามดังกล่าว P-value อาจเป็นพื้นที่ทางปลายหางด้านซ้ายหรือด้านขวา หรือทั้งสองของปลายหางของการแจกแจงของตัวสถิติทดสอบก็ได้ แล้วแต่สมมติฐานของการทดสอบครั้งนั้นๆ โดยจะต้องปฏิเสธ H_0 ถ้า P-value $\leq \alpha$

$$\text{P-value} = 0.5000 - Z_{cal} \qquad (4.4)$$

8. นำค่า Z_{cal} ที่คำนวณได้ไปเปรียบเทียบค่า Z ที่เปิดจากตารางแล้วพิจารณาตามกฎการตัดสินใจ

4.3.6 การตัดสินใจเกี่ยวกับการทดสอบสมมติฐานและการตีความหมาย

ตารางที่ 4.5 แสดงผลการตัดสินใจเกี่ยวกับการทดสอบสมมติฐาน

H_1	Z	H_0 at α=0.1		H_0 at α=0.2		H_0 at α=0.3	
		ยอมรับ H_0	ปฏิเสธ H_0	ยอมรับ H_0	ปฏิเสธ H_0	ยอมรับ H_0	ปฏิเสธ H_0
$\mu_{BSE} < \mu_{AQE}$	-2.50558		✓		✓		✓
$\mu_{BSE} < \mu_{IQE}$	-1.32721		✓		✓		✓
$\mu_{AQE} < \mu_{IQE}$	-0.89035	✓			✓		✓
$\mu_{AQE} < \mu_{LAQE}$	-1.49729		✓		✓		✓
$\mu_{IQE} < \mu_{LIQE}$	-0.25845	✓		✓			✓
$\mu_{LAQE} < \mu_{LA}$	-2.75577		✓		✓		✓
$\mu_{LIQE} < \mu_{LA}$	-3.00171		✓		✓		✓

จากตารางที่ 4.5 สามารถสรุปผลการตัดสินใจเกี่ยวกับการทดสอบสมมติฐานได้ดังนี้

1. $\mu_{BSE} < \mu_{AQE}$: ค่า Z ที่คำนวณได้คือ -2.50558 ซึ่งมีค่าน้อยกว่า $-Z_\alpha$=0.3 คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1 ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย AQE มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย BSE

2. $\mu_{BSE} < \mu_{IQE}$: ค่า Z ที่คำนวณได้คือ -1.32721 ซึ่งมีค่าน้อยกว่า $-Z_\alpha$=0.3 คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1 ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย IQE มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย BSE

3. $\mu_{AQE} < \mu_{IQE}$: ค่า Z ที่คำนวณได้คือ -0.89035 ซึ่งมีค่าน้อยกว่า $-Z_\alpha$=0.3 คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1 ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย IQE มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย AQE

4. $\mu_{AQE} < \mu_{LAQE}$: ค่า Z ที่คำนวณได้คือ -1.49729 ซึ่งมีค่าน้อยกว่า $-Z_\alpha$=0.3 คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1 ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LAQE มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย AQE

5. $\mu_{IQE} < \mu_{LIQE}$: ค่า Z ที่คำนวณได้คือ -0.71232 ซึ่งมีค่าน้อยกว่า $-Z_\alpha$=0.3 คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1 ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LIQE มีค่ามากกว่าหรือเท่ากับจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย IQE

6. $\mu_{LAQE} < \mu_{LA}$: ค่า Z ที่คำนวณได้คือ -2.7557 มีค่าน้อยกว่า $-Z_{\alpha}=0.3$ คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LA มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LAQE

7. $\mu_{LIQE} < \mu_{LA}$: ค่า Z ที่คำนวณได้คือ -3.0017 มีค่าน้อยกว่า $-Z_{\alpha}=0.1$ คือ -0.5245 จึงทำให้ปฏิเสธ H_0 และยอมรับ H_1ซึ่งหมายความว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LA มีค่ามากกว่าจำนวนเฉลี่ยของเว็บเพจที่เกี่ยวข้องกับคำถามที่ถูกสืบค้นโดย LIQE

4.4 การอภิปรายผลการทดลอง

4.4.1 การเปรียบเทียบเวลาที่ใช้ในการประมวลผล

1. การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลระหว่างเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LIQE) และเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) พบว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LIQE) จะใช้เวลาในการประมวลผลมากกว่าทั้งนี้เนื่องจากต้องใช้เวลาในสองขั้นตอนคือการเลือกเว็บเพจที่มีความเกี่ยวข้องจากผู้ใช้เพื่อป้อนกลับเข้าสู่ระบบรวมถึงขั้นตอนการวิเคราะห์ลิงค์เพื่อหาเว็บเพจที่มีความเกี่ยวข้องในขณะที่เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) จะใช้เวลานานในขั้นตอนของการเลือกผลลัพธ์ที่เกี่ยวข้องโดยผู้ใช้เพียงขั้นตอนเดียว

2. การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลระหว่างเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) และเทคนิคการขยายคำสืบค้นแบบอัตโนมัติทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LAQE) พบว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) ใช้เวลาเฉลี่ยในการประมวลผลเร็วกว่าทั้งนี้เพราะไม่ต้องผ่านขั้นตอนในการวิเคราะห์ลิงค์เพื่อหาเว็บเพจที่มีความเกี่ยวข้อง

3. การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลระหว่างเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) และเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) พบว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) ใช้เวลาเฉลี่ยน้อยกว่าเนื่องจากไม่มีขั้นตอนในการเลือกผลลัพธ์ที่เกี่ยวข้องโดยผู้ใช้แต่ระบบค้นคืนจะนำผลลัพธ์เริ่มต้นไปประมวลผลให้โดยอัตโนมัติซึ่งช่วยลดระยะเวลาในการประมวลผลลงได้อย่างมาก

4. การเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลระหว่างเทคนิคการวิเคราะห์ลิงค์ (LA) และเทคนิคการขยายคำสืบค้นที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (ทั้ง LAQE และ LIQE) พบว่าเทคนิคการวิเคราะห์ลิงค์ (LA) ใช้เวลาเฉลี่ยที่น้อยกว่าทั้งนี้เนื่องจากเทคนิคการวิเคราะห์ลิงค์ (LA) จะนำผลลัพธ์ที่ได้จากการค้นคืนในครั้งแรกไปผ่านขั้นตอนการวิเคราะห์ลิงค์เพื่อหาเว็บเพจที่มีความเกี่ยวข้องได้ทันที ในขณะที่เทคนิคการขยายคำสืบค้นที่ทำงานร่วมกับ

เทคนิคการวิเคราะห์ลิงค์ (LAQE และ LIQE) นั้นจะต้องนำผลลัพธ์ที่ได้จากการค้นคืนครั้งแรกหรือที่ถูกเลือกโดยผู้ใช้ไปประมวลผลเพื่อหาคำใหม่ก่อนแล้วนำคำใหม่ที่ได้เพิ่มลงในคำสืบค้นเดิมแล้วส่งไปสืบค้นอีกครั้งซึ่งผลลัพธ์ในครั้งนี้จึงจะถูกนำไปวิเคราะห์ลิงค์เพื่อหาเว็บเพจที่มีความเกี่ยวข้อง

4.4.2 การเปรียบเทียบประสิทธิภาพการค้นคืน

1. เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) ให้ประสิทธิภาพที่ดีกว่าการขยายคำสืบค้นแบบอัตโนมัติ (AQE) เนื่องจากเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) ผู้ใช้ต้องทำการพิจารณาเลือกเว็บเพจที่มีความเกี่ยวข้องป้อนกลับเข้าระบบค้นคืนซึ่งการคัดเลือกเว็บเพจของผู้ใช้จะช่วยในการคัดเลือกเทอมจากเว็บเพจที่มีความเกี่ยวข้องโดยเทอมที่ได้นั้นจะช่วยเพิ่มโอกาสในการค้นคืนเว็บเพจที่มีความเกี่ยวข้องกับหัวข้อหรือหัวเรื่องได้เพิ่มมากยิ่งขึ้น ซึ่งผลจากการทดลองพบว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) ให้ประสิทธิภาพที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) ประมาณ 19.61 %

2. เทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LAQE) ให้ประสิทธิภาพในการค้นคืนที่ดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) เนื่องจากกระบวนการในการวิเคราะห์ลิงค์จะช่วยในการจัดเรียงลำดับความเกี่ยวข้องของเว็บเพจผลลัพธ์ที่ได้จากเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) โดยจะมีการจัดเรียงลำดับตามความเกี่ยวข้องมากที่สุดไปยังน้อยที่สุดและเมื่อนำเว็บเพจที่มีความเกี่ยวข้องมากที่สุดมาสกัดคำรวมทั้งคำนวณหาค่าน้ำหนักของคำจึงทำให้มีโอกาสที่จะได้คำหรือกลุ่มคำที่แตกต่างจากคำสืบค้นแต่ยังคงมีความเกี่ยวข้องกับหัวข้อหรือหัวเรื่องเดียวกันซึ่งจะช่วยแก้ปัญหาในเรื่องความคลุมเครือของคำที่ใช้ในการอธิบายหัวเรื่องที่แตกต่างกันได้ ซึ่งผลจากการทดลองพบว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LAQE) ให้ประสิทธิภาพที่ดีกกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (AQE) ประมาณ 21.57 %

3. เทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LIQE) ให้ประสิทธิภาพในการค้นคืนดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) เนื่องจากกระบวนการในการวิเคราะห์ลิงค์จะช่วยในการจัดเรียงลำดับความเกี่ยวข้องของเว็บเพจผลลัพธ์ที่ได้จากเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) โดยจะมีการจัดเรียงลำดับตามความเกี่ยวข้องมากที่สุดไปยังน้อยที่สุดและเมื่อนำเว็บเพจที่มีความเกี่ยวข้องมากที่สุดมาสกัดคำรวมทั้งคำนวณหาค่าน้ำหนักของคำจึงทำให้มีโอกาสที่จะได้คำหรือกลุ่มคำที่แตกต่างจากคำสืบค้นแต่ยังคงมีความเกี่ยวข้องกับหัวข้อหรือหัวเรื่องเดียวกันซึ่งจะช่วยแก้ปัญหาในเรื่องความคลุมเครือของคำที่ใช้ในการอธิบายหัวเรื่องที่แตกต่างกันได้ ซึ่งผลจากการทดลองพบว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LIQE) ให้ประสิทธิภาพดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ (IQE) ประมาณ 11.48 %

4. เทคนิคการวิเคราะห์ลิงค์ (LA) ให้ประสิทธิภาพในการค้นคืนดีที่สุดเนื่องจากเทคนิคการวิเคราะห์ลิงค์ (LA) ใช้วิธีการคำนวณหาเว็บเพจที่มีความเกี่ยวข้องกับหัวเรื่องโดยผ่านทางการวิเคราะห์กราฟย่อยของกลุ่มเว็บเพจที่มีความเกี่ยวข้อง การวิเคราะห์กราฟย่อยจะช่วยทำให้ทราบถึงกลุ่มของเว็บเพจอื่นๆที่มีความเกี่ยวข้อง โดยที่กลุ่มเว็บเพจเหล่านั้นอาจจะใช้คำสืบค้นหรือคำที่ใช้อธิบายตัวเว็บเพจเองที่แตกต่างกันออกไปซึ่งจะช่วยเพิ่มโอกาสในการค้นคืนเว็บเพจที่มีความเกี่ยวข้องกับหัวเรื่องหรือหัวเรื่องได้มากยิ่งขึ้นซึ่งผลจากการทดลองพบว่าเทคนิคการวิเคราะห์ลิงค์ (LA) ให้ประสิทธิภาพดีกว่าเทคนิคการขยายคำสืบค้นแบบอัตโนมัติที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LAQE) ประมาณ 29.03 % และดีกว่าเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์ที่ทำงานร่วมกับเทคนิคการวิเคราะห์ลิงค์ (LIQE) ประมาณ 17.65 %

4.4.3 กลุ่มเว็บเพจผลลัพธ์ที่มีความเกี่ยวข้องที่ได้จากการค้นคืน

จากการเปรียบเทียบและพิจารณากลุ่มเว็บเพจที่มีความเกี่ยวข้องที่ถูกค้นคืนจะเห็นได้ว่าส่วนใหญ่ของกลุ่มเว็บเพจที่มีความเกี่ยวข้องที่ถูกค้นคืนด้วยเทคนิคการขยายคำสืบค้นแบบที่ยังไม่ได้ปรับปรุงจะเป็นคนละกลุ่มกับกลุ่มเว็บเพจที่มีความเกี่ยวข้องที่ถูกค้นคืนด้วยเทคนิคการขยายคำสืบค้นที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์ ตัวอย่างการเปรียบเทียบกลุ่มผลลัพธ์บางส่วนที่ค้นคืนได้และมีความเกี่ยวข้องแสดงในรูปที่ 4.3 โดยที่ U คือ Unique หมายถึงเว็บเพจไม่ปรากฏซ้ำในแต่ละรูปแบบการค้นคืน และ D คือ Duplicate หมายถึงมีเว็บเพจปรากฏซ้ำมากกว่าหนึ่งรูปแบบการค้นคืน

คำถามที่ 3			
หมายเลขผู้ทดลอง	รูปแบบการค้นคืน	เว็บเพจ	
29	AQE	http://www.umsl.edu/services/govdocs/ooh20022003/ocor009.pdf	U
		http://www.answers.com/topic/magnetic-resonance-imaging	U
	LAQE	http://www.broadlane.com/services/capital_equipment/equiptionary.pdf	U
		http://www.altera.com/end-markets/medical/diagnostic/med-diagnostic.html	D
	IQE	http://en.wikipedia.org/wiki/Diagnostic_imaging	U
		http://www.altera.com/end-markets/medical/diagnostic/med-diagnostic.html	D
	LIQE	http://www.fdimedical.com/products/body-fat.html	U
	LA	http://en.wikipedia.org/wiki/Medical_imaging	U
		http://www.afcindustries.com/Diagnostic_Imaging.htm	U
		http://www.collegeboard.com/csearch/majors_careers/profiles/majors/51.0910.html	U
		http://www.medscope.co.uk/	U
43	AQE	http://www.wma.net/e/publications/pdf/2000/giger.pdf	U
	LAQE	http://scienceline.org/2008/01/04/doctor%E2%80%99s-diagnosis-version-20/	D
		http://en.wikipedia.org/wiki/Computer-aided_diagnosis	D
		http://www.enotalone.com/article/8315.html	U
	IQE	http://www.emedicine.com/neuro/topic722.htm	U
	LIQE	http://www.easydiagnosis.com/articles/technology.html	U
	LA	http://scienceline.org/2008/01/04/doctor%E2%80%99s-diagnosis-version-20/	D
		http://www.news-medical.net/?id=17340	U
		http://en.wikipedia.org/wiki/Computer-aided_diagnosis	D
		http://www.med.umich.edu/opm/newspage/2004/computer.htm	U

รูปที่ 4.3 ตัวอย่างการเปรียบเทียบกลุ่มผลลัพธ์บางส่วนที่มีความเกี่ยวข้องและถูกค้นคืนโดย AQE, LAQE, IQE, LIQE และ LA ในคำถามที่สาม

4.4.4 ปัจจัยที่มีผลกระทบต่อการทดลอง

ในการทดลองพบว่ามีบางปัจจัยที่ส่งผลกระทบต่อผลของการทดลองซึ่งปัจจัยดังกล่าวมีดังต่อไปนี้

1. ชุดคำถามที่เลือกใช้

ชุดคำถามที่เลือกใช้ในการทดลองมีความยากจนเกินไปจึงส่งผลต่อจำนวนเว็บเพจที่มีความเกี่ยวข้องที่ค้นคืนกลับมาได้

2. ทักษะทางด้านภาษาของผู้ร่วมทำการทดลอง

ผู้ร่วมทำการทดลองยังขาดความชำนาญในภาษาอังกฤษเนื่องจากในการสืบค้นเพื่อหาหัวเรื่องที่ต้องการจำเป็นที่จะต้องใช้ความรู้ที่เกี่ยวข้องกับคำศัพท์มาช่วยในการสืบค้น

เมื่อพิจารณาเปรียบเทียบเวลาเฉลี่ยที่ใช้ในการประมวลผลและประสิทธิภาพในการค้นคืนแล้วพบว่าเทคนิคการวิเคราะห์ลิงค์เป็นทางเลือกที่ดีที่เหมาะกับการนำมาใช้งาน อย่างไรก็ตามเนื่องจากคำถามที่นำมาใช้ในการทดสอบเป็นโดเมนเฉพาะทางจึงทำให้จำนวนเว็บเพจที่เกี่ยวข้องและถูกค้นคืนกลับมาได้มีจำนวนน้อยมากจึงมีข้อแนะนำสำหรับการทดลองในอนาคต ให้ทำการทดสอบกับโดเมนทั่วไปมากขึ้น

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยนี้ได้ปรับปรุงประสิทธิภาพของเทคนิคการขยายคำสืบค้นแบบดั้งเดิม โดยใช้เทคนิคการวิเคราะห์ลิงค์แบบฮิทส์อัลกอริทึมเข้าร่วมในการขยายคำสืบค้นแบบอัตโนมัติและแบบปฏิสัมพันธ์กับผู้ใช้งาน โดยสถาปัตยกรรมที่ออกแบบสามารถนำไปจัดสร้างระบบสืบค้น เพื่อใช้ทดลองเปรียบเทียบประสิทธิภาพการค้นคืนได้ ผลการทดลอง พบว่า เทคนิคที่ปรับปรุงแล้วมีประสิทธิภาพในการค้นคืนที่ดีกว่าเดิม แต่ใช้เวลาในการทำงานที่นานกว่าเดิม เปรียบเทียบระหว่างการปรับปรุงเทคนิคการขยายคำสืบค้นแบบอัตโนมัติ (LAQE) กับการปรับปรุงเทคนิคการขยายคำสืบค้นแบบปฏิสัมพันธ์กับผู้ใช้งาน (LIQE) พบว่า การปรับปรุงให้ผลลัพธ์ที่ดีขึ้นทั้งคู่ โดยมีค่าเฉลี่ยถ่วงดุลเพิ่มขึ้น 21.57% และ 11.48% ตามลำดับ อย่างไรก็ตาม ในงานวิจัยนี้ ได้เพิ่มเติมการทดลองในส่วนการทดสอบประสิทธิภาพการค้นคืน โดยใช้เทคนิคการวิเคราะห์ลิงค์เพียงอย่างเดียว (LA) พบว่าให้ประสิทธิภาพดีที่สุด และใช้เวลาไม่มาก คือ ให้ค่าเฉลี่ยถ่วงดุล 0.080 เมื่อเทียบกับ 0.068 และ 0.062 ซึ่งเป็นค่าเฉลี่ยถ่วงดุลของ LIQE และ LAQE ตามลำดับ

แนวโน้มอีกประการหนึ่งที่พบ คือกลุ่มเว็บเพจผลลัพธ์ส่วนใหญ่ที่ได้จากการค้นคืนด้วยเทคนิคที่ปรับปรุงด้วยเทคนิคการวิเคราะห์ลิงค์จะไม่พบในกลุ่มเว็บเพจผลลัพธ์ที่ได้จากการค้นคืนด้วยเทคนิคการขยายคำสืบค้นทั้งแบบอัตโนมัติและแบบปฏิสัมพันธ์ที่ยังไม่ได้ทำการปรับปรุง ดังนั้นจึงสรุปได้ว่าเทคนิคการวิเคราะห์ลิงค์ที่นำมาใช้ปรับปรุงเทคนิคการขยายคำสืบค้นนั้นสามารถช่วยในการค้นคืนเว็บเพจที่มีความเกี่ยวข้องเพิ่มมากยิ่งขึ้น

5.2 ข้อเสนอแนะ

สำหรับแนวทางของงานวิจัยที่จะทำต่อเนื่องในอนาคตได้แก่ การศึกษาลักษณะของคำถามที่เทคนิคการวิเคราะห์ลิงค์ทำงานได้ดี เนื่องจากผลลัพธ์ของการค้นคืนในงานวิจัยนี้มีผลลัพธ์ที่เกี่ยวข้องจำนวนน้อยมาก ซึ่งอาจเกิดจากลักษณะคำถามที่เจาะจง และเป็นโดเมนเฉพาะทางมากเกินไป ในอนาคต อาจจะทำการทดลองกับชุดคำถามในโดเมนอื่นๆ หรือเปลี่ยนจากการใช้ Yahoo Search Engine มาใช้ชุดเอกสารของ TREC แทน เพื่อให้สอดคล้องกับส่วนของคำถามที่นำมาจาก TREC

ในส่วนของกระบวนการทำงานของเทคนิคที่นำเสนอนั้น อาจจะทำการทดลองปรับเปลี่ยนจำนวนเว็บเพจตั้งต้น เพื่อเปรียบเทียบประสิทธิภาพเพิ่มเติม เนื่องจากในงานวิจัยนี้ใช้เว็บเพจตั้งต้นเพียงจำนวน 30 เว็บเพจ ในขณะที่ฮิทส์อัลกอริทึม จะแนะนำเซตเริ่มต้นไว้ที่ประมาณ 200 หน้าเว็บ

ฮิทส์อัลกอริทึม ยังสามารถประยุกต์ใช้งานต่างๆ อีกมาก เช่น การค้นหาชุมชนไซเบอร์ การจัดหมวดหมู่หน้าเว็บโดยอัตโนมัติ การวิเคราะห์การอ้างอิง และงานอื่นๆ นอกจากนี้ ที่ผ่านมา ยังไม่มีงานวิจัยในเชิงเปรียบเทียบประสิทธิภาพในแง่มุมของการประยุกต์ใช้งานต่างๆ ของเทคนิคการวิเคราะห์ลิงค์ทั้งสองวิธี คือ เปรียบเทียบเทคนิค เพจแรงค์ (PageRank) กับ ฮิทส์ ซึ่งก็นับว่าเป็นเรื่องที่น่าสนใจเช่นกัน

เอกสารอ้างอิง

[1] Allan B., et. al. "Link Analysis Ranking: Algorithms, Theory and Experiments." *ACM Trans. Internet Technol.*, Vol.5, No.1, Feb. 2005. pp. 231-297

[2] Brin S., Page L. "The Anatomy of a Large-scale Hypertextual Web Search Engine." *Computer Networks and ISDN Systems.*, Vol.30, 1998. pp. 107-117

[3] Chen Z., et. al. "Building a Web Thesaurus from Web Link Structure." *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada*, May 29, 2003. pp. 48-55.

[4] Chung Y. M., Lee Y. J. "Optimization of Some Factors Affecting the Performance of Query Expansion." *An International Journal: Information Processing and Management.*, Vol. 40, No.6, Nov. 2004. pp. 891-917

[5] Crouch C. J. "An Approach to the Automatic Construction of Global Thesauri." *An International Journal: Information Processing and Management.*, Vol.26, No.5, 1990. pp. 629-640

[6] Efthimiadis E. N. "Query Expansion." In Williams, Martha E., ed. *Annual Review of Information Science and Technology.*, Vol. 31. 1996. pp. 121-187

[7] Harman D. "Towards interactive query expansion." *in Proc. Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, Grenoble, France*, 1988. pp. 321-331.

[8] Henzinger M. "Link Analysis in Web Information Retrieval." *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.*, Vol.23, No.3, Sep. 2000. pp. 3-8

[9] Jian-Fu L., et. al. "A New Approach to Query Expansion." *Proceedings of International Conference on Machine Learning and Cybernetics, Guangzhou*, August 18-21, 2005. pp. 2302-2306

[10] Kleinberg M. J. "Authoritative Sources in a Hyperlinked Environment." *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, San Francisco, California, United States*, January 25-27, 1998. pp. 668-677

[11] Koenemann J., Belkin J. N. "A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness." *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground, Vancouver, British Columbia, Canada*, April 13-18, 1996. pp. 205-212

[12] Magennis M., Rijsbergen J. v. C. "The Potential and Actual Effectiveness of IQE." *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, Philadelphia, Pennsylvania, United States*, July 27-31, 1997. pp. 324-332

[13] Nemeth Y., et. al. "Evaluation of the Real and Perceived Value of Automatic and Interactive Query Expansion." *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom*, July 25-29, 2004. pp. 526-527

[14] Peat H.J., Willett P. "The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval System." *Journal of the ASIS.*, Vol.42, No.5, 1997. pp. 378-383

[15] Qiu Y., Frei P. H. "Concept Based Query Expansion." *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, United States*, 1993. pp. 160-169

[16] White R.W., et. al. "Comparing Explicit and Implicit Feedback Techniques for Web Retrieval: TREC-10 Interactive Track Report." *Proceedings of the Tenth Text Retrieval Conference (TREC-10), Gaithersburg, Maryland, USA*, May 6, 2002. pp. 534-538.

[17] Seher I. "Query Expansion in Personal Queries." *IADIS Virtual Multi Conference on Computer Science and Information Systems 2006, Lisbon, Portugal,* May 15-19, 2006.

[18] Shafi S. M., Rather R. A. *Precision and Recall of Five Search Engine for Retrieval of Scholarly Information in the Field of Biotechnology*. Webology, Vol.2, No.2, Article 12. [Online]. Available at: http://www.webology.ir/2005/v2n2/a12.html. 2005.

Improving Query Expansion Using Link Analysis

Wirith Leelapatra and Ponrudee Netisopakul
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, 10520 Thailand
wirithl@gmail.com, ponrudee@it.kmitl.ac.th

***Abstract*-Query expansion techniques aim to improve a user's search by adding new query terms to an existing query. This can be done in two ways: by having a human user refines the first result set or by using automatic extraction of terms from the first result set. Both techniques usually depend on terms frequency in web pages only. This paper proposes to improve upon query expansion technique using a link analysis technique called Hypertext Induce Topic Selection algorithm (HITS). The preliminary result using questions set from TREC is presented.**

I. INTRODUCTION

Some important problems with search engine is the unsatisfied results set when using words or keywords search. Usually these systems employ information retrieval techniques by relying on term frequency derived from webpage title, description and content. Since same term or same words can be used to describe different concept, at the same time, the same concept can naturally be described by different terms, the relevant websites are not retrieved and the irrelevant websites are retrieved. Consequently, the precision and recall of the results sets are low.

Query Expansion [1], [4] techniques partially solve the above problem. The general concept is to overcome the limitation of initial keywords provided by users, by having search terms expanded from the first result set. A query is expanded by adding other terms closely related to the original query terms. In some work [3], expansion terms are added to the query by the user (Interactive Query Expansion-IQE), in other works expansion terms are added by the retrieval system (Automatic Query Expansion-AQE) [5]. For the past years, a wide range of methods for query expansion have been proposed, from manual techniques such as thesauri to automatic techniques such as automatic relevance feedback [4]. These methods have shown to be effective on different extent in improving the performance of IR system, but they are still far from being satisfactory.

One of the limitations with the traditional query expansion techniques is that a query is often expanded only by importance of terms [6]. This paper proposes concept of improving effectiveness of query expansion by integrating link analysis technique. Combination of both techniques should improve efficiency of search results.

II. RELATED RESEARCH

A number of research works related to query expansion and link analysis techniques are briefly described here.

Jian, Mao-zu and Shu-Hong [4] presented new approach to query expansion. Their approach combined two traditional methods thesauri and automatic relevance feedback. The advantages of their approach are simplicity and speed. The thesaurus can be automatically updated using information from automatic relevance feedback. In their experiments, the use of new approach gave better retrieval results than query expansion methods that use only WordNET and Local Context Analysis (LCA).

Magennis and Van Rijsbergen [7] presented the experiments aim to determine both the potential and the actual effectiveness of multiple iterations of interactive query expansion in large scale realistic search context. In their experiments, the retrieval effectiveness of automatic, experienced user interactive and inexperienced user interactive query expansion were compared. The conclusion of their experiments was that interactive query expansion performed by experienced user showed consistent improvement across a range of topics, while interactive query expansion performed by inexperienced user consistently showed lack of improvement. In contrast, automatic query expansion showed overall improvement but varied across topics.

Kleinberg [2] presented Hypertext Induced Topic Selection (HITS). It is a method for computational determining hubs and authorities on a particular topic through analysis of a relevant sub-graph of web. The idea of this method is to identifying good hub and good authority. The good hub is a page that points to many good authorities. The good authority is a page pointed to by many good hubs. Pages with high authority are results of query.

III. PROPOSED CONCEPT

To improve the effectiveness of query expansion, we proposed a technique that integrating link analysis with traditional query expansion. We design component called refinement engine as shown in figure 1. The working process

978-1-4244-2101-5/08/$25.00 ©2008 IEEE

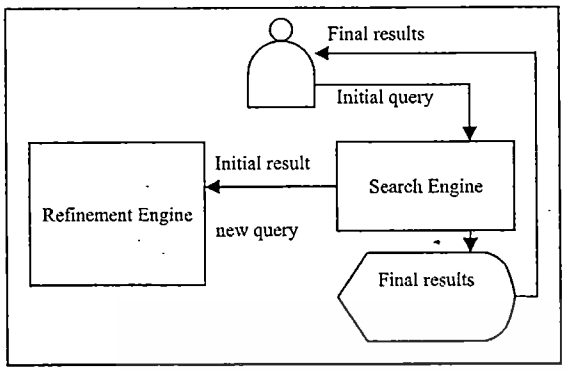


Figure 1. The context of refinement engine

of refinement engine starts with the user initial query. The search engine processes the query and returns the initial result set. These initial result set is used to calculate relevancy of the hyperlinks among the initial set and their extended links which have been added to set by using HITS algorithm. The result of the calculation are sets of hubs and authorities associated with their relevancy scores. Then, new search terms are extracted from this new set and combined with the initial query. Finally, the new query is send to the search engine again.

The refinement engine is shown in figure 2. It consists of two sub-modules. These are Link Analysis module and Query Expansion module.

The sub functions of Link Analysis are explained here.

- Link Expander module extracts hyperlinks from the initial result set. The resulting hyperlinks are then sent to the web graph constructor module.

- Web Graph Constructor module constructs a sub-graph. Steps of constructing sub-graph are:

1. Collecting the top t pages (say t =30) based on the input query; call this **initial link set**

2. Extending the initial links set into a larger set. For all pages p in the initial link set:

   2.1 Adding to the initial link set all pages that p point to, and

   2.2 Adding to the initial link set all pages that point to p.

called this **expanded link set**.

3. Delete all links within the same web site in the expanded link set.

The steps of constructing web-graph are depicted in figure 3. The extended link set is then sent to HITS calculator module.

- HITS Calculator module calculates hub and authority scores as followed

Compute the authority score and hub score of each web page in expanded link set on the sub-graph SG (V, E)

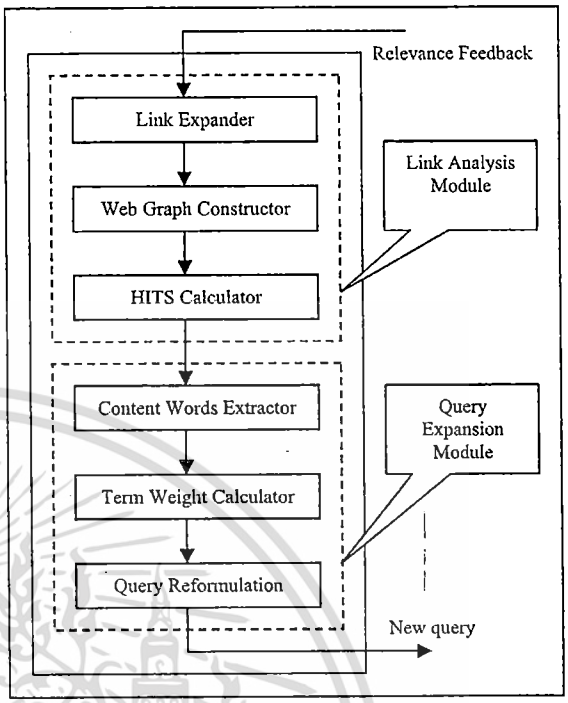


Figure 2. The refinement engine

Given a page p, let

A (p) be the authority score of p

H (p) be the hub score of p

(p, q) be a direct edge in E from p to q

Initialize a (p) = h (p) = 1 for all p in V

For each p in V until the score converge

$$A(p) = \sum_{q:(q,p)\in E} H(q) \quad (1)$$

$$H(p) = \sum_{q:(p,q)\in E} A(q) \quad (2)$$

normalize a (p) and h (p)

The result of this calculation is sets of hubs and authorities pages that will be send to the content words extractor module.

The sub functions of Query Expansion are explained next.

- Content Words Extractor module extracts words and filters non-content words from webpages in authority set. The result is terms that will be sent to the terms weight calculator module.

- Terms Weight Calculator module calculates term weight as define below.

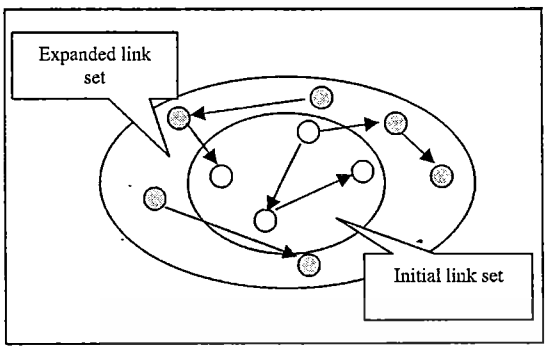


Figure 3. Constructing sub-graph.

Let w_i be a term;

[(The number of webpages in the set that has term w_i)-1] / [(The number of webpages in the set)] (3)

The results are terms and their associated weight. Notice that the denominator is constant, therefore, we can use only the numerator to choose the new terms. For the term w_i that does not appear in other webpage in the set, the weight is zero. Ideally, the algorithm should select only terms that weight greater than some significant threshold value. In this experiment, following the number suggested by [7], we select the best six terms.

- Query Formulation module formulates a new query by adding new terms into existing query and then send the new query to the search engine again.

IV. EXPERIMENTAL METHOD

The design of experimental methods include five following modes

1. Basic search engine. (General search, search without using any improvement technique).
2. Interactive Query Expansion (IQE). (A technique of adding new terms to an existing query by using user interaction information with the search engine)
3. Automatic Query Expansion (AQE) (A technique of adding new terms to an existing query automatically)
4. Link Analysis collaborative with Interactive Query Expansion (LIQE). (A technique of adding new terms to an existing query by combining link analysis and IQE)
5. Link Analysis collaborate with Automatic Query Expansion (LAQE). (A technique of adding new terms to an existing query by combining link analysis and AQE)

The experimental participants consist of 20 computer science graduate students. Each participant is assigned searching tasks consist of question related to computer science topics based on science and technology domain from TREC-1, TREC-4 and TREC-2003. Figure 4 shows an example of searching task from TREC-1. In each question, participants need to contribute information used in IQE mode only. We then use this information to generate results compare to other modes. Yahoo is used as the basic search engine. The comparison of all five modes as shown in next section.

V. EXPERIMENT RESULT

In this section we show preliminary results of experiment using integrating technique link analysis into query expansion.

The partial details of experiments are presented in TABLE I. The partial comparison results of relevant webpages retrieved in AQE and LAQE mode for question 2 are presented in TABLE II. From the preliminary experiment, we have found two good trends:

- AQE mode or IQE mode found zero relevant webpages for some questions queried by some participants, while LIQE mode and LAQE mode found relevant webpages with the same questions and the same participants as shown in TABLE I.
- Most of relevant webpages retrieved by AQE/IQE and LIQE/LAQE are not duplicated as shown in TABLE II.

From the two good trends above we can conclude that link analysis has a potential to retrieve more relevant pages. However, we found some factors affected the experiment. These factors are described here.

- Participants

The experiment participants were inexperience in using search engine and rarely used search engine. This reason affects the search method, how to start the search and which word should be used in query. Furthermore, they also lacked the ability of understanding and interpreting questions. These were evidenced in their initial query.

- The number of expansion terms

The number of expansion terms added to the existing query affected the performance. In this experiment, we selected six best terms suggested by [7]. But from TABLE I

Domain: Science and Technology
Topic: How Rewritable Optical Disks Work.
Description: Document describes the principles and mechanisms behind rewritable optical disk technology.
Narrative: To be relevant, a document must describe how rewritable optical disk technology works at length and in significant and comprehensive technical detail.

Figure 4. The example of TREC-1 searching task

TABLE I
THE PARTIAL COMPARISON THE NUMBER OF RELEVANT WEBPAGES RETRIEVED IN AQE, LAQE, IQE AND LIQE.

Question	User No.	The number of relevant webpages			
		AQE	LAQE	IQE	LIQE
1	2	0	3	1	4
	3	0	4	0	3
	8	1	3	0	2
	11	0	1	0	2
	17	0	3	0	3
2	2	0	2	1	3
	7	1	2	0	2
	12	0	2	0	3
	19	1	3	0	3
3	3	0	2	0	3
	11	0	2	2	4
	12	1	3	0	3
	13	0	2	0	3
	19	1	3	0	2

TABLE II
THE PARTIAL COMPARISON SET OF RELEVANT WEBPAGES RETRIEVED IN AQE MODE AND LAQE MODE FOR QUESTION 2.

Mode	set of relevant webpages
AQE	http://www.translationsoftware4u.com/machine-translation.htm
	http://www.answers.com/topic/machine-translation
	http://www.openinternetlexicon.com/MTSystems/MTSystems.html
LAQE	http://www.e-promt.com/
	http://www.systransoft.com/
	http://www.soget.com/Translation.asp

the number of relevant webpages retrieved in LIQE mode and LAQE mode were less than expectation results. We assumed that the best expansion terms in this experiment should not be six terms.

- Source of extract content

The results from HITS algorithm consist of the authority pages and the hub pages. In this experiment, we selected the authority page because we had an assumption that it contained significant, reliable, and useful information on the topics. However, in this experiment we have not compared whether the source of extract content from the authority page or the hub page gives better results.

VI. CONCLUTIONS

This paper proposed to improve query expansion technique by integrating link analysis technique. From the preliminary experiment, link analysis was shown to be useful for retrieving more relevant pages when integrated with traditional query expansion.

The future works includes improving experiment design. For experimental participants, we will divide participants into groups based on experiences of using and frequency of using search engine. As for the number of expansion terms, we will perform experiment by varying the number of terms added to initial query to find the best number of expansion terms in this experiment. Finally, we will extend the methodology to determine whether the authority set or the hub set gives the best expansion terms.

ACKNOWLEDGMENT

The work was funded by Faculty of Information Technology research funding fiscal year 2008, King Mongkut's Institute of Technology Ladkrabang Bangkok, Thailand.

REFERENCES

[1] Chung Young Mere, Lee Jake Yun, "Optimization of some factors affecting the performance of query expansion", Information Processing and Management, Vol 40, No. 6, 2004. pp. 891-917.

[2] Kleinberg, J. "Authoritative sources in a hyperlinked environment". In Proceeding of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, January 1998, pages 668-667

[3] Koenemann, J. and Belkin, N.J. "A case for interaction: A study of interactive information retrieval behavior and effectiveness". Proceedings of CHI 96 International conference on Human Computer Interaction, Vancouver, B.C., Canada, 1996, 205-212.

[4] Li Jain-Fu., Guo Mao-Zu. And Tian Shu-Hong. "A New Approach to query expansion". Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005

[5] Seher Indra, "Query. Expansion in Personal Queries". IADIS Virtual Multi Conference on. Computer. Science. and. Information. Systems (MCCSIS2006), 2006.

[6] Song Min., Song, Il-Yeol, Hu Xiaohua, and Allen, B. Robert., "Semantic Query Expansion Combining Association Rules with Ontologies and Information Retrieval Techniques". Proceeding of Data Warehousing and Knowledge Discovery, 7th International Conference, DaWaK 2005, Copenhagen, Denmark, August 22-26, 2005., p 326-335

[7] Magennis M., Van Rijbergen, C.J. "The potential and actual effectiveness of Interactive Query Expansion. Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval. Philadelphia, Pennsylvania, United States. Pages: 324-332