การค้นคืนเอกสารตามเนื้อหาโดยใช้หลักการของกราฟ

Content-Based Document Retrieval Using Graph-Based Approach

แกมกาญจน์ สมประเสริฐศรี

Gamgarn Somprasertsri

ภาควิชาสารสนเทศศาสตร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

Information Science Department, Faculty of Informatics, Mahasarakham University

E-Mail gamgarns@msu.ac.th

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อ 1) ออกแบบกระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ และ 2) วัดประสิทธิภาพของกระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ ข้อมูลที่ใช้ในการทดลองคือ หนังสืออิเล็กทรอนิกส์ที่มีหัวเรื่อง “database design” วิธีดำเนินการวิจัยประกอบด้วย การรวบรวมข้อมูล การออกแบบกระบวนการค้นคืนเอกสารโดยใช้หลักการของกราฟ และการวัดประสิทธิภาพของการค้นคืน

ผลการวิจัยพบว่า กระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟประกอบด้วยงานหลัก 4 ส่วน ได้แก่ 1) การประมวลผลข้อความ 2) การสร้างดรรชนี 3) การวัดความคล้ายคลึง และ 4) การจัดลำดับผลลัพธ์จากการค้นคืน ผลการวัดประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ พบว่า การคำนวณค่าความคล้ายคลึงระหว่างคำค้นกับดรรชนีในแต่ละโหนดด้วยวิธี Jaccard’s Coefficient รวมกับการใช้ระดับโหนดของดรรชนีในต้นไม้เอกสาร ช่วยในการจัดลำดับผลลัพธ์ที่ได้จากการค้นคืนได้ดีกว่าการไม่ใช้ระดับโหนดของดรรชนีในต้นไม้เอกสาร และผลของการค้นคืนให้ค่าความแม่นยำ 0.75 ค่าการระลึก 1.00 และค่าเอฟ 0.85 ผลการศึกษาครั้งนี้แสดงให้เห็นว่ากระบวนการที่นำเสนอสามารถนำมาใช้ในการค้นคืนเอกสารตามเนื้อหาที่เป็นประเด็นย่อยได้ดี

คำสำคัญ: การค้นคืนเอกสารตามเนื้อหา, กราฟ

Abstract

The purposes of the research were to design method for content-based document retrieval using graph-based approach and to test the effectiveness of the content-based document retrieval using graph-based approach. Experimental data was electronic books entitled database design. The research process consisted of data collection, graph-based document retrieval process design, and retrieval effectiveness evaluation.

The results revealed that the proposed method consists of four major steps; 1) text processing, 2) indexing, 3) similarity measure and 4) document ranking. The experiment showed that using Jaccard’s Coefficient with degree of document tree to measure similarity of documents

can rank the query result more efficient than only using Jaccard's Coefficient. Furthermore, this proposed method was good average precision (0.75), average recall (1.00) and average F-measure (0.846). This results of the study provided the proposed method could be used in content-based document retrieval effectively.

Keywords: Content-based Document Retrieval, Graph-based

บทนำ

การค้นคืนสารสนเทศ (Information retrieval) คือ การกระทำใด ๆ ที่คัดเลือกสารสนเทศจากแหล่งเก็บข้อมูลเพื่อทำให้ได้รับสารสนเทศตามที่ต้องการ อาจเป็นข้อมูลหรือรายการเอกสารซึ่งบรรจุเนื้อหาที่ต้องการ หลักการสำคัญของการค้นคืนสารสนเทศ คือ การค้นหาและนำสารสนเทศที่ตรงตามความต้องการส่งกลับมายังผู้ใช้งานอย่างรวดเร็วและตรงตามความต้องการได้อย่างมีประสิทธิภาพ รูปแบบของการค้นคืนในยุคสารสนเทศดิจิทัลไม่ได้มีเฉพาะลักษณะที่เป็นตัวอักษรเพียงอย่างเดียวแต่มีหลากหลายลักษณะ ได้แก่ ข้อความ รูปภาพ วีดิทัศน์ เสียง หรือในลักษณะของมัลติมีเดีย นอกจากนี้ยังมีการนำไปใช้ในหลายลักษณะ เช่น การค้นคืนแผนที่ การค้นคืนรูปภาพ [1] แหล่งสารสนเทศในปัจจุบันมีหลากหลาย แต่แหล่งสารสนเทศที่สำคัญในสถาบันการศึกษาสำหรับการค้นคืนคือห้องสมุด ซึ่งการค้นคืนสารสนเทศในห้องสมุดมหาวิทยาลัยในปัจจุบันมีการนำคอมพิวเตอร์เข้ามาใช้งานเรียกว่า ระบบห้องสมุดอัตโนมัติ

ระบบห้องสมุดอัตโนมัติที่นำมาใช้งานในห้องสมุดในปัจจุบันโดยเฉพาะห้องสมุดระดับมหาวิทยาลัยที่นิยมใช้ เช่น ระบบห้องสมุด VTLS ระบบห้องสมุด INNOPAC ระบบห้องสมุด WALAI AutoLib ระบบห้องสมุดอัตโนมัติเหล่านี้จะประกอบด้วยชุดโปรแกรมหรือโมดูล (Module) ต่าง ๆ ที่ใช้กับงานในห้องสมุด [2] ได้แก่ 1) โมดูลสำหรับการบริหารจัดการทรัพยากรสารสนเทศของห้องสมุดเป็นการจัดซื้อจัดหา รวมทั้งการบริหารงบประมาณ (Acquisition module) 2) โมดูลในการกำหนดและจัดหมวดหมู่ทรัพยากรสารสนเทศในห้องสมุด (Cataloguing module) 3) โมดูลบริหารจัดการสิ่งพิมพ์ต่อเนื่อง ได้แก่ การบอกรับสิ่งพิมพ์ต่อเนื่อง เช่น วารสาร หนังสือพิมพ์ (Serial Control module) 4) โมดูลในการจัดการเกี่ยวกับการสืบค้นทรัพยากรสารสนเทศในห้องสมุด (OPAC & WEBPAC module) 5) โมดูลในการให้บริการยืม-คืนทรัพยากรสารสนเทศของห้องสมุด (Circulation module) และ 6) โมดูลที่ใช้ในการควบคุมกำหนดเงื่อนไขการทำงานของโมดูลต่าง ๆ ให้ทำงานตามเงื่อนไขที่ต้องการ (System module) การค้นคืนทรัพยากรสารสนเทศในห้องสมุดผ่านระบบห้องสมุดอัตโนมัติ ผลลัพธ์ของการค้นคืนจะแสดงเป็นรายการบรรณานุกรม และการค้นคืนมีหลายช่องทาง ได้แก่ ชื่อผู้แต่ง ชื่อเรื่อง เลขเรียก หัวเรื่อง และคำสำคัญ ซึ่งเป็นการสืบค้นในระดับกว้าง (General terms) ที่ทำหน้าที่แทนเฉพาะในระดับตัวเล่ม ทำให้ผู้ใช้ไม่สามารถสืบค้นไปในระดับประเด็นย่อย ๆ เช่น บท หรือ ตอน ที่มีอยู่ภายในเล่มได้ และคำสืบค้นในระดับกว้างมักเป็นคำศัพท์ควบคุม (Controlled vocabulary) ซึ่งผู้ใช้งานส่วนใหญ่ไม่คุ้นเคยกับคำศัพท์ดังกล่าว ส่งผลให้เกิดความผิดพลาดในการค้น ทั้งนี้ผู้ใช้งานส่วนใหญ่มักสืบค้นด้วยคำศัพท์อิสระซึ่งเป็นคำที่มักปรากฏอยู่ในเนื้อหา และเป็นศัพท์ที่มีความทันสมัยกว่าศัพท์ควบคุม ทำให้ไม่สามารถตอบสนองความต้องการในการเข้าถึงเนื้อหาในสื่อสิ่งพิมพ์ได้ [3]

จากปัญหาดังกล่าวผู้วิจัยจึงสนใจที่จะศึกษาและออกแบบกระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยอาศัยหลักของกราฟเข้ามาช่วยในการค้นคืนเอกสารให้มีประสิทธิภาพมากขึ้น

1. วัตถุประสงค์การวิจัย

1.1 เพื่อออกแบบกระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ

1.2 เพื่อศึกษาประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 ตัวแทนสารสนเทศ

ตัวแทนสารสนเทศ (Information representation) เป็นสัญลักษณ์หรือสารสนเทศที่มีโครงสร้างที่อธิบายลักษณะต่าง ๆ ที่สำคัญของทรัพยากรสารสนเทศ เพื่อทำหน้าที่เป็นตัวแทนของทรัพยากรสารสนเทศนั้น สัญลักษณ์หรือสารสนเทศที่เป็นตัวแทนนี้อาจอยู่ในรูปแบบของคำ วลี ข้อความ ตัวเลข หรือรหัส อย่างใดอย่างหนึ่งหรือประกอบกัน [4]

2.2 กราฟต้นไม้และเทคนิคการค้นหาข้อมูล

ต้นไม้ (Tree) เป็นกราฟชนิดพิเศษที่ถูกนำไปใช้ประโยชน์มากที่สุดชนิดหนึ่ง ซึ่งเป็นกราฟเชื่อมต่อไม่มีทิศทาง ซึ่งไม่มีวงวนที่จุดใดและไม่สามารถมีเส้นเชื่อมหลายเส้นเชื่อมเกิดขึ้นระหว่างจุดยอดคู่ใด ๆ กราฟชนิดนี้เมื่อวาดรูปจุดยอดและเส้นเชื่อมของกราฟแล้วจะมีลักษณะคล้ายต้นไม้ [5] โดยทั่วไปแล้วการนำโครงสร้างต้นไม้ไปใช้ประโยชน์ มักจะกำหนดจุดยอดจุดหนึ่งให้เป็นราก (Root) ของต้นไม้ สำหรับต้นไม้กำหนดราก เราเรียกจุดยอดใด ๆ ที่มีเส้นเชื่อมมีทิศทางจากจุดยอดนั้น ๆ ไปยังจุดยอด v จุดยอดหนึ่งว่าเป็นแม่ (Parent) ของจุดยอด v นั้น และจุดยอด v จะถูกเรียกว่าเป็นลูก (Child) สำหรับจุดยอดที่มีแม่เป็นจุดยอดเดียวกันจะถูกเรียกว่าเป็น พี่น้อง (Siblings) ส่วนจุดยอดที่ไม่มีลูกจะถูกเรียกว่าใบ (Leaf)

เทคนิคที่ใช้ในการค้นหาข้อมูลจากโครงสร้างต้นไม้ที่ เป็นการค้นหาเพียงบางส่วนโดยการค้นหาแบบบอด (Blind search) สามารถแบ่งได้เป็น 2 ประเภท คือ [6] 1) การค้นหาแนวลึกก่อน (Depth-first search) ในการค้นหาแนวลึกก่อนเป็นการค้นหาที่กำหนดทิศทางจากรูปของโครงสร้างต้นไม้ ที่เริ่มต้นจากโหนดราก (Root node) ที่ อยู่บนสุด แล้วเดินลงมาให้ลึกที่สุด เมื่อถึงโหนดล่างสุด (Terminal node) และยังไม่ได้คำตอบก็จะทำการย้อนกลับ ขึ้นมาที่จุดสูงสุดของกิ่งเดี่ยวกันที่มีกิ่งแยกและยังไม่ได้เดินผ่าน แล้วเริ่มเดินลงจนถึงโหนดลึกสุดอีก เช่นนี้สลับไป เรื่อยจนพบโหนดที่ต้องการหาหรือสำรวจครบทุกโหนด และ 2) การค้นหาแนวกว้างก่อน (Breadth-first search) ในการค้นหาแนวกว้างก่อนนี้ สถานะของโหนดทุกตัวที่อยู่ในระดับเดียวกันของต้นไม้จะถูกตรวจสอบก่อนสถานะที่อยู่ในระดับถัดไป วิธีการของการค้นหาแบบนี้จะเริ่มจากการสร้างสถานะลูกของสถานะเริ่มต้นก่อน แล้วตรวจสอบว่า มีสถานะใดที่เป็นสถานะสุดท้ายหรือไม่ ถ้าหากว่ามีก็เป็นอันว่าการค้นหาสิ้นสุด ถ้าไม่มีก็จะสร้างสถานะลูกของสถานะเหล่านั้น แล้วทำการตรวจสอบสถานะลูกทุกตัวของสถานะเหล่านั้น ถ้าพบสถานะสุดท้ายการค้นหาก็สิ้นสุด ถ้าไม่พบก็สร้างสถานะลูกของสถานะเหล่านั้นต่อไปอีก ทำเช่นนี้ไปเรื่อย ๆ จนกว่าจะพบสถานะสุดท้ายหรือจนไม่สามารถสร้างสถานะลูกได้ใหม่

2.3 เอ็นแกรม

เอ็นแกรม (N-Gram) [7] คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของชุดอักขระ (Character Sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงกัน (Word Sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของชุดอักขระหรือคำ ประมาณได้จากคลังข้อมูลที่สร้างไว้ แกรม (Gram) คือ หน่วยที่ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียงคำ หรืออักขระก็ได้ ซึ่งแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึง เอ็น (n) ในแบบจำลองเอ็นแกรมนี้ใช้ความยาวของสายอักขระหรือสายคำแตกต่างกัน ได้แก่ 1-แกรม 2-แกรม 3-แกรม และ 4-แกรม เป็นต้น โดย n=1 จะเรียกว่ายูนิแกรม (Unigram) n=2 จะเรียกว่าไบแกรม (Bigram) หรือ n=3 จะเรียกว่าไตรแกรม (Trigram) เทคนิคเอ็นแกรมถูกนำมาประยุกต์ใช้กับงานได้หลายลักษณะ เช่น การบีบอัดข้อความ การค้นคืนสารสนเทศ การสกัดสารสนเทศ และการประมวลผลภาษาธรรมชาติ

งานวิจัยนี้นำเทคนิคเอ็นแกรมมาใช้ในการประมวลผลดรรชนีหรือตัวแทนเอกสารที่สร้างจากหน้าสารบัญและคำสอบถามหรือคำค้น (Query) โดยแบ่งกลุ่มคำหรือข้อความออกมาเป็นหน่วยคำ (Term) หรือแกรมตามขนาดของแกรมที่กำหนด โดยใช้จำนวนแกรม N เท่ากับ 3 (n = 1, 2, 3) การใช้ข้อมูลจากหน้าสารบัญเป็นตัวแทนของเอกสาร ในลักษณะคำสำคัญ (Keyword) เนื่องจากเป็นองค์ประกอบที่บ่งชี้ถึงเนื้อหาและความเกี่ยวเนื่องกันของเนื้อหาได้ดี สามารถช่วยเพิ่มประสิทธิภาพของการค้นคืนเอกสารได้ นอกจากนี้ยังไม่จำเป็นต้องอาศัยผู้เชี่ยวชาญในการวิเคราะห์ งานวิจัยนี้ใช้โครงสร้างต้นไม้ในการนำเสนอศัพท์ดรรชนีของเอกสารตามหน้าสารบัญ และใช้เทคนิคการค้นหาแนวกว้างก่อน เพื่อเปรียบเทียบคำค้นกับดรรชนีที่เป็นตัวแทนเอกสารที่อยู่ในรูปแบบโครงสร้างต้นไม้ โดยการวัดค่าความคล้ายคลึงจะใช้ Jaccard’s Coefficient ระหว่างสายคำของคำค้นกับดรรชนี

2.4 งานวิจัยที่เกี่ยวข้อง

จากงานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศตามเนื้อหา สมจิน เปียโคกสูง และนิศาชล จำนงศรี [3] ได้นำเสนองานวิจัยเรื่องระบบนำทางความรู้เพื่อการเข้าถึงเนื้อหาในสื่อสิ่งพิมพ์ โดยมีวัตถุประสงค์เพื่อออกแบบและพัฒนากระบวนการระบบนำทางความรู้ และประเมินประสิทธิภาพของระบบนำทางความรู้เพื่อเข้าถึงเนื้อหาในทรัพยากรสารสนเทศประเภทสื่อสิ่งพิมพ์ ใช้รายการสารบัญและดรรชนีท้ายเล่มของสื่อสิ่งพิมพ์เป็นตัวแทนของประเด็นเนื้อหา มีการประยุกต์ใช้เทคนิคเอ็นแกรม ปริภูมิเวกเตอร์สเปซโมเดล และโครงสร้างกราฟเอ็มแอล เพื่อสืบค้น นำเสนอและเชื่อมโยงผลการค้นคืน ผลการวิจัยพบว่า ระบบนำทางความรู้มีค่าความแม่นยำ 0.81 ค่าความระลึก 0.64 และค่าเอฟเมเชอร์ 0.71 ผู้ใช้ระบบมีความพึงพอใจภาพรวมในระดับมาก โดยมีความพึงพอใจด้านประโยชน์ในการส่งเสริมการเรียนรู้ และด้านคุณภาพความรู้และสารสนเทศที่ได้ในรับระดับมากเช่นกัน ส่วนด้านสมรรถนะในการให้บริการของระบบผู้ใช้มีความพึงพอใจในระดับปานกลาง นอกจากนี้สมชาย ประสิทธิ์จูตระกูล [8] ได้ศึกษาวิจัยเรื่อง การออกแบบแฟ้มผกผันเพื่อการค้นคืนข้อความภาษาไทย โดยนำเสนอขั้นตอนวิธีการหาคำเพื่อจัดทำดัชนีสำหรับระบบการค้นคืนข้อความไทยที่ใช้โครงสร้างแฟ้มผกผัน งานวิจัยนี้ใช้พจนานุกรมและกฎการแบ่งพยางค์ข้อความภาษาไทยในการแยกคำ ทำให้สามารถจัดการกับคำที่ไม่ปรากฏในพจนานุกรม เช่น คำทับศัพท์ หรือคำที่สะกดผิดได้ ในการค้นคืนใช้กราฟการต่อและซ้อนกันของคำ ซึ่งมีโหนดแทนคำและเส้นเชื่อมแทนการต่อหรือซ้อนกันของคำโดยมีเส้นทางสั้นสุดจากซ้ายไปขวาในกราฟแทนรายการคำพื้นฐานที่ควรถูกจัดทำดัชนีสำหรับแฟ้มผกผัน เวลาการทำงานของการหาคำเป็น O(n2) โดยที่ n คือความยาวของข้อความ ขั้นตอนวิธีนี้ถูกใช้ทั้งในขั้นตอนการเตรียมเอกสารก่อนการทำดัชนี และการประมวลข้อคำถามก่อนการสืบค้น ผลการวิจัยพบว่า จำนวนคำที่หาได้

เพื่อทำดัชนีนั้นมีจำนวนประมาณ 30-50% ของจำนวนคำที่เป็นไปได้ทั้งหมดที่ปรากฏในข้อความทดสอบ ระบบสามารถค้นคืนเอกสารที่มีคำสำคัญภาษาอังกฤษ หรือคำทับศัพท์เป็นภาษาไทยของคำอังกฤษนั้นได้ โดยมีค่าความระลึกและความแม่นยำมากกว่า 80% สำหรับงานวิจัยที่เกี่ยวกับการค้นคืนสารสนเทศแบบกราฟ พบว่า Thammasut and Sornil [9] ได้ทำการศึกษาวิจัย เรื่องการค้นคืนสารสนเทศแบบกราฟ โดยคำค้นหรือข้อคำถามและเอกสารจะถูกแสดงในลักษณะของกราฟแทนแบบจำลองทั่วไป ซึ่งเอกสารทั้งหมดจะนำเสนอในลักษณะของกราฟเอกสาร (Document graph) ที่ไม่มีทิศทางแต่เส้นเชื่อมระหว่างเอกสารมีน้ำหนัก คำนวณโดยใช้ tf-idf กระบวนการที่นำเสนอในงานวิจัยนี้มีทั้งหมด 6 ขั้นตอน คือ 1) การประมวลผลข้อความ (Text processing) 2) การทำดรรชนี (Indexing) 3) การสร้างกราฟเอกสาร (Document graph construction) 4) ส่วนติดต่อผู้ใช้ (User Interface) 5) การรวมกราฟ (Graph merging) และ 6) การจัดอันดับเอกสาร (Document ranking) ใช้ Random Walk Restarts algorithm ในการคำนวณการจัดอันดับของเอกสารที่คล้ายคลึงกับคำค้น การวัดประสิทธิภาพทำการทดลองโดยใช้ข้อมูลจาก TREC 3 ชุดข้อมูล คือ AP, FBIS และ WSJ ทำการเปรียบเทียบประสิทธิภาพการค้นคืนสารสนเทศกับแบบจำลองเวกเตอร์ ผลการวิจัยพบว่า การค้นคืนโดยใช้กราฟมีค่าความแม่นยำสูงกว่าการใช้ตัวแบบปริภูมิเวกเตอร์ โดยให้ค่าความแม่นยำของการทดลองกับ AP, FBIS และ WSJ เท่ากับ 0.42, 0.34 และ 0.39 ตามลำดับ

วิธีดำเนินการวิจัย

1. เครื่องมือการวิจัย

1.1 การวิจัยครั้งนี้ศึกษาจากหนังสืออิเล็กทรอนิกส์ภาษาอังกฤษที่มีลักษณะเป็นไฟล์ PDF ซึ่งไม่ได้เกิดจากการสแกนเอกสาร เป็นหนังสือที่มีเนื้อหาเกี่ยวกับการออกแบบฐานข้อมูล และมีหัวเรื่อง “Database design” ที่ค้นจากฐานข้อมูลที่ให้บริการของสำนักวิทยบริการ มหาวิทยาลัยมหาสารคาม ซึ่งมีเนื้อหาประเด็นย่อย คือ การปรับบรรทัดฐาน (Normalization) หรือแบบจำลองความสัมพันธ์เอนทิตี (Entity Relationship Diagram หรือ Entity Relationship Model : E-R Diagram) จำนวน 5 รายการ (D_1, D_2, D_3, D_4, D_5) และหนังสือที่ไม่มีเนื้อหาประเด็นย่อยที่เกี่ยวกับการปรับบรรทัดฐาน หรือแบบจำลองความสัมพันธ์เอนทิตีสำหรับการออกแบบฐานข้อมูล แต่เป็นผลลัพธ์ที่ได้จากหัวเรื่อง “Database design” จำนวน 3 รายการ (D_6, D_7, D_8)

1.2 ในการศึกษาประสิทธิภาพของการค้นคืนสารสนเทศ ผู้วิจัยใช้คำค้นที่เป็นทั้งคำเดี่ยวและวลี 6 ชุด ดังนี้ 1) database design (q_1) 2) normalization (q_2) 3) entity relationship model (q_3) 4) database design normalization (q_4) 5) database design entity relationship model (q_5) และ 6) normalization entity relationship model (q_6) และสร้างการทดลองด้วยโปรแกรม MatLab

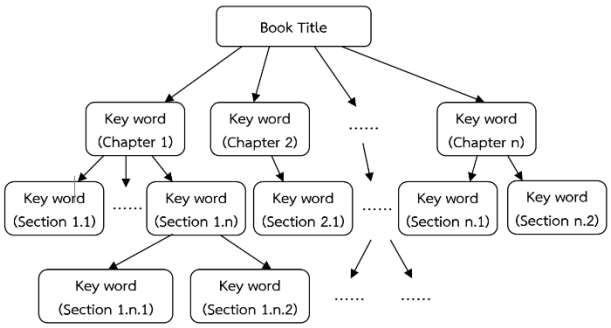
2. ขั้นตอนการดำเนินการวิจัย

2.1 รวบรวมข้อมูล เป็นการศึกษาและวิเคราะห์เอกสารที่เป็นหนังสืออิเล็กทรอนิกส์ พบว่าหน้าสารบัญเป็นข้อมูลที่เหมาะสำหรับการทำดรรชนีเพื่อเป็นตัวแทนเนื้อหาในหนังสือ

2.2 ออกแบบกระบวนการค้นคืนเอกสารตามเนื้อหาโดยใช้หลักการของกราฟ ประกอบด้วยงานหลักดังนี้

2.2.1 การประมวลผลข้อความ เป็นการประมวลผลข้อมูลก่อนการทำดรรชนีเพื่อสร้างต้นไม้เอกสาร โดยเริ่มจากการแปลงไฟล์ภาพหน้าสารบัญให้เป็นไฟล์ข้อความด้วยโปรแกรมโอซีอาร์แล้วจึงทำการประมวลผลข้อความ ประกอบไปด้วยขั้นตอนย่อย ๆ คือ 1) การวิเคราะห์หน่วยคำ 2) การกำจัดคำหยุด และ 3) การหารากศัพท์

2.2.2 การสร้างดรรชนี เป็นนำข้อมูลที่ได้ผ่านการประมวลผลข้อความแล้วมาสร้างดรรชนีในรูปแบบโครงสร้างต้นไม้ แสดงดังภาพที่ 1



ภาพที่ 1 โครงสร้างต้นไม้เอกสาร

2.2.3 การวัดความคล้ายคลึง (Similarity measure) จะนำเอาคำค้นมาเปรียบเทียบกับดรรชนีในโหนดของต้นไม้เอกสาร โดยใช้วิธีค้นแบบกว้างก่อนแสดงดังภาพที่ 2 พร้อมทั้งคำนวณค่าความคล้ายคลึงระหว่างคำค้นกับดรรชนีในแต่ละโหนดด้วยวิธี Jaccard's Coefficient ตัวอย่างการวัดความคล้ายคลึงระหว่างคำค้นและดรรชนีในแต่ละโหนด ด้วยวิธี Jaccard's Coefficient

ตัวอย่าง คำค้น = database design

ดรรชนี = data database management

1-แกรมของคำค้นและดรรชนี

คำค้น = {database} {design}

ดรรชนี = {data} {database} {management}

$$|X \cap Y| / |X \cup Y| = 1/4 = 0.25$$

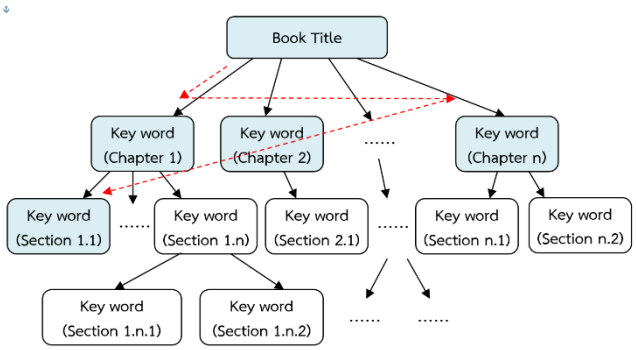
2-แกรมของคำค้นและดรรชนี

คำค้น = {database design}

ดรรชนี = {data database} {database management}

$$|X \cap Y| / |X \cup Y| = 0$$

ดังนั้น ได้ค่าความคล้ายคลึงระหว่างคำค้นและดรรชนีมีค่าเท่ากับ 0.25



ภาพที่ 2 การค้นแบบกว้างก่อนในต้นไม้เอกสาร

2.2.4 การจัดอันดับผลลัพธ์จากการค้นคืน ใช้ค่าความคล้ายคลึงระหว่างเอกสาร D_j และคำค้น Q_i หรือ $Sim(Q_i, D_j)$ ซึ่งได้จากผลรวมของค่าความคล้ายคลึงแต่ละโหนดดรรชนีคูณกับระดับ (Level) ของโหนดดรรชนีในต้นไม้เอกสาร แสดงได้ดังสูตร

$$Sim(Q_i, D_j) = \sum_{k=1}^{n} Sv_k \times Lv_k$$

โดย Sv_k = ค่าความคล้ายคลึงของโหนด k กับคำค้น

Lv_k = ระดับของโหนด k ในต้นไม้เอกสาร

2.3 วัดประสิทธิภาพของกระบวนการค้นคืนเอกสารโดยใช้หลักการของกราฟ โดยใช้ค่าการระลึก (Recall) ค่าความแม่นยำ (Precision) และค่าเอฟ (F-measure)

ผลการวิจัย

ผู้วิจัยได้ดำเนินการออกแบบกระบวนการค้นคืนเอกสารโดยใช้หลักการของกราฟ และได้แบ่งการทดลองเพื่อศึกษาประสิทธิภาพของกระบวนการค้นคืนเอกสารโดยใช้หลักการของกราฟ ออกเป็น 2 ส่วน คือ 1) การศึกษาประสิทธิภาพของการจัดอันดับผลลัพธ์ของการค้นคืนโดยใช้ต้นไม้เอกสารร่วมกับ Jaccard's Coefficient 2) การศึกษาประสิทธิภาพของการค้นคืนเอกสารโดยใช้หลักการของกราฟ แสดงรายละเอียดดังนี้

1. ผลของการจัดอันดับผลลัพธ์ของการค้นคืนโดยใช้ต้นไม้เอกสารร่วมกับ Jaccard's Coefficient แบ่งออกเป็น 2 ส่วน คือ การวัดความคล้ายคลึงจากเอ็นแกรมและใช้ระดับของโหนดในต้นไม้เอกสารในการจัดอันดับผลลัพธ์ และการวัดความคล้ายคลึงจากเอ็นแกรม แต่ไม่ใช้ระดับของโหนดในต้นไม้เอกสารในการจัดอันดับผลลัพธ์ ผลการทดลองแสดงค่าความสอดคล้อง (Relevant) ของแต่ละเอกสารกับแต่ละคำค้นแสดงได้ดังตารางที่ 1 และตารางที่ 2

ตารางที่ 1 ค่าความคล้ายคลึงของแต่ละเอกสารกับแต่ละคำค้นโดยใช้เอ็นแกรมและระดับของโหนด

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
q_1	5.63	4.42	3.90	6.92	5.95	15.25	0.74	0.42
q_2	3.74	3.41	0.29	2.25	5.77	0	0	0
q_3	5.09	9.28	6.65	6.90	36.39	6.99	0.33	0.96
q_4	9.37	7.88	4.19	9.22	11.72	15.24	0.74	0.42
q_5	10.72	13.71	10.56	13.82	42.49	22.3	1.08	1.38
q_6	8.83	12.70	6.94	9.15	42.16	6.99	0.33	0.96

จากตารางที่ 1 ค่าความคล้ายคลึงของแต่ละเอกสารกับแต่ละคำค้นโดยใช้เอ็นแกรมและระดับของโหนดในกราฟต้นไม้ พบว่า คำค้น q_1 จะมีค่าคล้ายคลึงกับเอกสาร D_6 มากที่สุด รองลงคือ เอกสาร D_4 และ เอกสาร D_5 ตามลำดับ และคำค้น q_2 q_3 q_5 และ q_6 จะมีค่าคล้ายคลึงกับเอกสาร D_5 มากที่สุด ส่วนคำค้น q_4 จะมีค่าคล้ายคลึงกับเอกสาร D_6 มากที่สุด

ตารางที่ 2 ค่าความคล้ายคลึงของแต่ละเอกสารกับแต่ละคำค้นโดยใช้เอ็นแกรมและไม่ใช้ระดับของโหนด

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8
q_1	3.12	2.14	2.29	3.93	2.96	7.07	0.38	0.31
q_2	2.01	1.40	0.14	1.21	2.40	0	0	0
q_3	2.41	3.84	3.74	3.66	14.26	2.70	0.11	0.53
q_4	5.13	3.56	2.43	5.17	5.36	7.07	0.38	0.31
q_5	5.54	5.98	6.03	7.59	17.28	9.81	0.49	0.83
q_6	4.43	5.24	3.88	4.86	16.66	2.70	0.11	0.53

จากตารางที่ 2 ค่าความคล้ายคลึงของแต่ละเอกสารกับแต่ละคำค้นโดยใช้เอ็นแกรมและไม่ใช้ ระดับของโหนดในกราฟต้นไม้ พบว่า คำค้น q_1 จะมีค่าคล้ายคลึงกับเอกสาร D_6 มากที่สุด รองลงคือ เอกสาร D_4 และ เอกสาร D_1 ตามลำดับ และคำค้น q_2 q_3 q_5 และ q_6 จะมีค่าคล้ายคลึงกับเอกสาร D_5 มากที่สุด ส่วนคำค้น q_4 จะมีค่าคล้ายคลึงกับเอกสาร D_6 มากที่สุด เมื่อพิจารณาเปรียบเทียบค่าความคล้ายคลึงของแต่ละเอกสารกับแต่ละคำค้นจะพบว่าการใช้และไม่ใช้โหนดในกราฟต้นไม้จะให้ค่าสูงสุดกับเอกสารที่เหมือนกัน แต่เมื่อพิจารณาค่าความคล้ายคลึงในลำดับถัดไปจะพบความแตกต่างกัน และเมื่อนำค่าความคล้ายคลึงจากการใช้และไม่ใช้ระดับของโหนดในกราฟต้นไม้มาจัดอันดับผลลัพธ์ เพื่อเปรียบเทียบกับการจัดอันดับผลลัพธ์โดยผู้ใช้ จะแสดงได้ดังตารางที่ 3

ตารางที่ 3 เปรียบเทียบการจัดอันดับผลลัพธ์ของการค้นคืน

	D_1			D_2			D_3			D_4			D_5			D_6			D_7			D_8		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
q_1	5	4	3	4	5	6	6	6	5	2	2	2	3	3	4	1	1	1	7	7	7	8	8	8
q_2	3	2	2	4	3	3	2	5	5	5	4	4	1	1	1	-	-	-	-	-	-	-	-	-
q_3	4	6	6	3	2	2	2	5	3	5	4	4	1	1	1	6	3	5	8	8	8	7	7	7
q_4	3	3	4	4	5	5	2	6	6	5	4	3	1	2	2	6	1	1	7	7	7	8	8	8
q_5	4	5	6	3	4	5	2	6	4	5	3	3	1	1	1	6	2	2	8	8	8	7	7	7
q_6	4	4	4	3	2	2	2	6	5	5	3	3	1	1	1	6	5	6	8	8	8	7	7	7

หมายเหตุ: A จัดอันดับโดยผู้ใช้ B จัดอันดับโดยใช้ระดับโหนด B คือ จัดอันดับโดยไม่ใช้ระดับโหนด

จากตารางที่ 3 พบว่าอันดับผลลัพธ์ที่ได้จากคำค้น q_2 “normalization” การจัดอันดับโดยการใช้ระดับโหนดและไม่ใช้ระดับโหนด ผลลัพธ์ที่ได้จะมีอันดับที่เหมือนกัน และส่วน D_7 และ D_8 จะไม่ถูกจัดอันดับ เนื่องจากเป็นหนังสือที่ไม่มีประเด็นเนื้อหาย่อยเกี่ยวกับ “normalization” เมื่อเปรียบเทียบการจัดอันดับผลลัพธ์ของการค้นคืน ทั้งการใช้ระดับของโหนดและไม่ใช้ระดับของโหนดกับการจัดอันดับโดยผู้ใช้ พบว่า การใช้ระดับโหนดของดรรชนีในต้นไม้เอกสาร จะช่วยให้ลำดับผลลัพธ์ที่ได้จากการค้นคืนสอดคล้องกับคำค้นและการจัดอันดับโดยผู้ใช้มากกว่าการไม่ใช้ระดับโหนดของดรรชนีในต้นไม้เอกสาร

2. ผลการศึกษาประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ ใช้ค่าการระลึก ค่าความแม่นยำ และค่าเอฟ แสดงดังตารางที่ 4

ตารางที่ 4 ประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ

	ค่าความแม่นยำ	**ค่าการระลึก**	**ค่าเอฟ**
q_1	1.00	1.00	1.00
q_2	1.00	1.00	1.00
q_3	0.63	1.00	0.77
q_4	0.63	1.00	0.77
q_5	0.63	1.00	0.77
q_6	0.63	1.00	0.77
เฉลี่ย	**0.75**	**1.00**	**0.85**

จากตารางที่ 4 พบว่าประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้ต้นไม้เอกสารร่วมกับ Jaccard’s Coefficient ให้ค่าความแม่นยำเฉลี่ยเท่ากับ 0.75 ค่าการระลึกเฉลี่ยเท่ากับ 1.00 และค่าเอฟเฉลี่ยเท่ากับ 0.846

อภิปรายผลการวิจัย

การวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบกระบวนการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ และศึกษาประสิทธิภาพของการค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟ โดยทำการทดสอบประสิทธิภาพการค้นคืนกับหนังสืออิเล็กทรอนิกส์ที่มีหัวเรื่อง “database design” ที่ปรากฏในฐานข้อมูลของสำนักวิทยบริการ มหาวิทยาลัยมหาสารคาม และทดลองใช้คำค้นทั้งหมด 6 ชุด ผลการวิจัยพบว่าการใช้ดรรชนีจากข้อมูลสารบัญหนังสือในรูปแบบโครงสร้างต้นไม้ และการวัดความคล้ายคลึงด้วย Jaccard’s Coefficient รวมกับการใช้ระดับโหนดของดรรชนีในต้นไม้เอกสารช่วยให้อันดับผลลัพธ์ที่ได้จากการค้นคืนสอดคล้องกับคำค้นมากกว่าการไม่ใช้ระดับโหนดของดรรชนีในต้นไม้เอกสาร นอกจากนี้ การค้นคืนเอกสารตามเนื้อหาของเอกสารโดยใช้หลักการของกราฟที่นำเสนอนี้ยังให้ค่าความแม่นยำ เท่ากับ 0.75 ค่าการระลึก เท่ากับ 1.00 และค่าเอฟ เท่ากับ 0.85 แสดงให้เห็นว่าข้อมูลสารบัญสามารถใช้เป็นตัวแทนเนื้อหาในเอกสารได้ดี เนื่องจากเป็นส่วนที่ผู้เขียนได้กำหนดขึ้นมาแทนเนื้อหาทุกประเด็นของหนังสือและเป็นการวิเคราะห์ในระดับลึก ซึ่งช่วยส่งเสริมประสิทธิภาพของการค้นคืนมากกว่าการใช้ชื่อเรื่อง หรือหัวเรื่องซึ่งเป็นการวิเคราะห์เนื้อหาในระดับกว้างที่วิเคราะห์เฉพาะประเด็นหลัก เพื่อกำหนดคำที่ครอบคลุมเนื้อหาทั้งหมดของเอกสารคร่าว ๆ แต่ไม่คำนึงถึงประเด็นย่อยของเนื้อหา และการจัดอันดับผลลัพธ์ของการค้นคืนที่พบว่า การใช้ระดับโหนดของดรรชนีในต้นไม้เอกสารจะช่วยให้ลำดับผลลัพธ์ที่ได้จากการค้นคืนสอดคล้องกับคำค้น ทั้งนี้เนื่องมาจากการใช้ดรรชนีในลักษณะโครงสร้างต้นไม้ ทำให้สามารถระบุค่าน้ำหนักความสำคัญของดรรชนีได้ตามระดับความลึกของโหนดในต้นไม้เอกสาร และนำมาใช้ในการคำนวณหาค่าความสอดคล้องของแต่ละเอกสารกับแต่ละคำค้น

ข้อเสนอแนะ

งานวิจัยนี้สามารถเป็นแนวทางเพื่อนำไปประยุกต์ใช้กับการค้นคืนสารสนเทศจากหนังสืออิเล็กทรอนิกส์ภาษาไทย หรือเอกสารอิเล็กทรอนิกส์รูปแบบอื่น ๆ ที่ต้องการค้นคืนสารสนเทศในระดับเชิงลึกของเนื้อหาในเอกสาร แต่กระบวนการที่ออกแบบยังไม่รองรับการค้นคืนเชิงความหมาย และมีข้อจำกัดสำหรับเอกสารที่มีการใช้คำย่อในเนื้อหา เช่น ER แทน entity relationship ในเอกสาร ซึ่งมีผลต่อการคำนวณค่าความคล้ายคลึงของตัวแทนเอกสารกับคำค้น การวิจัยในครั้งต่อไปควรมีพิจารณาความหมายของคำร่วมด้วย หรือมีการนำวิธีการประมวลผลภาษาธรรมชาติมาใช้เพื่อแก้ไขปัญหา หรือข้อจำกัดดังกล่าว

เอกสารอ้างอิง

[1] เสกศักดิ์ ปราบพาลา, และภัทรา นามเมือง. (2559). การพัฒนาระบบสืบค้นเชิงแผนที่โรงเรียนมัธยมศึกษาในจังหวัดชัยภูมิ. *วารสารวิชาการการจัดการเทคโนโลยีสารสนเทศและนวัตกรรม*, 3(2), 81-87.

[2] มหาวิทยาลัยสุโขทัยธรรมาธิราช สาขาวิชาศิลปศาสตร์. (2554). *เอกสารการสอนชุดวิชา 13723 หน่วยที่ 8-15 การจัดโครงสร้างสารสนเทศและการค้นคืน.* นนทบุรี: ผู้แต่ง.

[3] สมจิน เปียโคกสูง, และนิศาชล จำนงศรี. (2553). ระบบนำทางความรู้เพื่อการเข้าถึงเนื้อหาในสื่อสิ่งพิมพ์. *วารสารสารสนเทศศาสตร์*, 28(3), 9-20.

[4] มหาวิทยาลัยสุโขทัยธรรมาธิราช สาขาวิชาศิลปศาสตร์. (2554). *เอกสารการสอนชุดวิชา 13723 หน่วยที่ 1-7 การจัดโครงสร้างสารสนเทศและการค้นคืน.* นนทบุรี: ผู้แต่ง.

[5] อำพล ธรรมเจริญ. (2551). *กราฟและการประยุกต์.* กรุงเทพฯ: โรงพิมพ์พิทักษ์การพิมพ์.

[6] บุญเสริม กิจศิริกุล. (2546). *ปัญญาประดิษฐ์.* กรุงเทพฯ: ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

[7] NECTEC PEDIA. (2014). *N-Gram.* Retrieved from http://wiki.nectec.or.th/runewwiki/bin/ view/IT630_11 _Assignment/N-Gram

[8] สมชาย ประสิทธิ์จูตระกูล. (2541). *การออกแบบแฟ้มผกผันเพื่อการค้นคืนข้อความไทย*. กรุงเทพฯ: สถาบันวิจัยและพัฒนา คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย.

[9] Thammasut, D., & Sornil, O. (2006). A Graph-Based Information Retrieval System. *Proceeding of International Symposium on Communication and Information Technologies* (pp. 743 – 748). Bangkok, Thailand: IEEE.