การประยุกต์การสืบค้นแบบฮิวริสติคเพื่อการค้นคืนสารสนเทศ ภายใต้ทฤษฎีโครงข่ายเบย์

APPLIED HEURISTIC SEARCH IN INFORMATION RETRIEVAL USING BAYESIAN NETWORKS

วรางคณา เงินแก้ว

WARANGKHANA NGENKAEW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2544

ISBN 974-648-047-2

การประยุกต์การสืบค้นแบบฮิวริสติคเพื่อการค้นคืนสารสนเทศ ภายใต้ทฤษฎีโครงข่ายเบย์

APPLIED HEURISTIC SEARCH IN INFORMATION RETRIEVAL USING BAYESIAN NETWORKS

วรางคณา เงินแก้ว

WARANGKHANA NGENKAEW

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2544

ISBN 974-648-047-2

APPLIED HEURISTIC SEARCH IN INFORMATION RETRIEVAL USING BAYESIAN NETWORKS

WARANGKHANA NGENKAEW

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2001

ISBN 974-648-047-2

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การประยุกต์การสืบค้นแบบฮิวริสติคเพื่อการค้นคืนสารสนเทศภายใต้ทฤษฎีโครงข่ายเบย์

APPLIED HEURISTIC SEARCH IN INFORMATION RETRIEVAL USING BAYESIAN NETWORKS

ชื่อนักศึกษา นางสาววรางคณา เงินแก้ว

รหัสประจำตัว 39067033

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ผู้ควบคุมวิทยานิพนธ์ ผศ.ดร.เอื้อน ปิ่นเงิน

คณะกรรมการสอบวิทยานิพนธ์		ลายมือชื่อ
ผศ.ดร.เอื้อน	ปิ่นเงิน	
รศ.ดร.วิเชียร	เปรมชัยสวัสดิ์	
ดร.อาริต	ธรรมโน	
ดร.วรพจน์	กรีสุระเดช	
ดร.โชติพัชร์	ภรณวลัย	

วัน/เดือน/ปี ที่สอบ 22 ธันวาคม 2543 **เวลา** 9.00 น. เป็นต้นไป

สถานที่สอบ ณ ห้อง LAB 316 ชั้น 3 อาคารสำนักวิจัยและบริการคอมพิวเตอร์

บัณฑิตวิทยาลัยรับรองแล้ว

(รศ.ดร.บุญวัฒน์ อัตชู)

คณบดีบัณฑิตวิทยาลัย

วันที่ 1 **เดือน** มกราคม **พ.ศ.** 2544

หัวข้อวิทยานิพนธ์	การประยุกต์การสืบค้นแบบฮิวริสติคเพื่อการค้นคืนสารสนเทศภายใต้ทฤษฎีโครงข่ายเบย์
นักศึกษา	นางสาววรางคณา เงินแก้ว
รหัสประจำตัว	39067033
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2544
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ. ดร. เอื้อน ปิ่นเงิน

บทคัดย่อ

เนื้อหาของวิทยานิพนธ์เล่มนี้ เป็นการนำเสนอการประยุกต์เทคนิคทางปัญญาประดิษฐ์เพื่อการค้นคืนสารสนเทศ เทคนิคดังกล่าวคือ การสืบค้นแบบฮิวริสติค (Heuristic Search) โดยใช้สืบค้นฐานข้อมูลที่เป็นเอกสาร วิธีการนี้ทำให้ได้เอกสารที่ตรงตามความต้องการของผู้สืบค้นโดยเรียงลำดับจากมากไปหาน้อย เนื่องจากผู้สืบค้นสามารถกำหนดความต้องการของตนเองได้ด้วยการใช้คำเฉพาะ (Keywords) และข้อมูลฮิวริสติค (Heuristic Information) เทคนิคดังกล่าวจะทำการเก็บข้อมูลไว้ในส่วนที่เรียกว่าข้อมูลอธิบายเอกสาร (Document Profile) เพื่อทำการสืบค้นเอกภพของเอกสาร และส่วนของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) เพื่อทำการคัดเลือกเอกสาร จากนั้นได้ใช้หลักการของความน่าจะเป็นในทฤษฎีโครงข่ายเบย์ (Bayesian Networks) ในการจัดเรียงเอกสารเพื่อให้ตรงตามความต้องการของผู้สืบค้น ผลจากการทดลองกับฐานข้อมูลเอกสารจำนวน 850 ระเบียน ของห้องสมุดคณะเทคโนโลยีสารสนเทศ เป็นที่น่าพอใจ

Thesis Title	Applied Heuristic Search in Information Retrieval using Bayesian Networks
Student	Miss Warangkhana Ngenkaew
Student ID.	39067033
Degree	Master of Science
Programme	Information Technology
Year	2001
Thesis Advisor	Asst. Prof. Dr. Ouen Pinngern

ABSTRACT

This thesis presents the using of Artificial Intelligence technique for information retrieval. The technique, called heuristic search, uses for retrieving relevant documents from document space. It helps users to retrieve documents in descending order of relevancy to their needs, since the users can specify their needs through set of keywords of interests along with heuristic information regarding to the required documents. This technique stores heuristic information in the document profile and user profile. Then Bayesian networks were used for ordering documents in descending order of user's relevancy. The experiment consists of 850 document records from the Faculty of Information Technology's library. The result is very promising.

กิตติกรรมประกาศ

งานวิจัยของข้าพเจ้าในครั้งนี้จะสำเร็จลุล่วงด้วยดีไม่ได้ ถ้าปราศจากบุคคลเหล่านี้ ข้าพเจ้าจึงใคร่ขอกล่าวคำขอบพระคุณมา ณ โอกาสนี้

ขอขอบพระคุณบิดา-มารดา ของข้าพเจ้า ผู้ซึ่งให้สติปัญญา อบรมเลี้ยงดูและให้กำลังใจเมื่อยามที่ข้าพเจ้ารู้สึกท้อแท้ ตลอดจนสนับสนุนกำลังทรัพย์มาโดยตลอด

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร. เอื้อน ปิ่นเงิน อาจารย์ที่ปรึกษา ผู้ซึ่งให้คำแนะนำในเรื่องงานวิจัย ให้แนวคิดต่างๆ แก่ข้าพเจ้า ทำให้วิทยานิพนธ์สำเร็จลุล่วงไปได้ด้วยดี

ขอขอบพระคุณ คุณกมลรัตน์ ตัณฑ์เกยูร หัวหน้าฝ่ายวิเคราะห์ทรัพยากรห้องสมุด สำนักหอสมุดกลาง สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ผู้ให้คำปรึกษาในเรื่องเกี่ยวกับงานของบรรณารักษ์ในการกำหนดและตีความการให้คำสำคัญ เพื่อทำการค้นคืนเอกสาร ทำให้ข้าพเจ้าเข้าใจลักษณะการทำงานมากยิ่งขึ้น

ขอขอบคุณ คุณวศิน เสงี่ยมกุล เพื่อนผู้ให้คำปรึกษาเรื่องการเขียนโปรแกรม รวมทั้งรุ่นพี่ คุณนิภาพร ประภาศิริ และ คุณพัฒนพงษ์ ฉันทมิตรโอภาส ศิษย์อาจารย์ที่ปรึกษาเดียวกัน สำหรับคำแนะนำ การช่วยเหลือ และกำลังใจ ในทุกสิ่งทุกอย่างที่ร่วมฝ่าฟันมาด้วยกัน

ท้ายที่สุดขอขอบคุณพี่ๆ และ เพื่อนๆ ทุกคนที่ให้คำปรึกษา เป็นกำลังใจ และมีส่วนช่วยเหลือในการทำวิทยานิพนธ์ครั้งนี้ของข้าพเจ้า

วรางคณา เงินแก้ว

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญตาราง (ต่อ)

สารบัญรูป

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

เนื่องจากระบบค้นคืนสารสนเทศที่ใช้กันแพร่หลายทั่วไปนั้น มีลักษณะเป็น Boolean Model ซึ่งผู้สืบค้นสามารถระบุความต้องการของตนได้จำกัด การเปรียบเทียบคำที่ผู้สืบค้นใส่เข้าไปในระบบ เป็นการเปรียบเทียบกันตรงๆ (Exact match) ผู้สืบค้นจึงได้เอกสารทั้งที่ตรงและไม่ตรงตามความต้องการออกมา ซึ่งในบางครั้งเอกสารที่ไม่ตรงตามความต้องการออกมามากกว่าเอกสารที่ตรงตามความต้องการ [21] งานวิจัยนี้ได้นำการสืบค้นแบบฮิวริสติคมาประยุกต์ใช้ เพื่อให้การสืบค้นได้เอกสารตรงความต้องการของผู้สืบค้น และตัดเอกสารที่ไม่ตรงตามความต้องการของผู้สืบค้นออกไปให้มากที่สุด นอกจากนี้ ข้อจำกัดอีกประการหนึ่งของการสืบค้นแบบเดิม คือผู้สืบค้นสามารถเลือกประเภทการสืบค้นได้เพียงประเภทเดียวในการสืบค้นครั้งหนึ่งๆ ตัวอย่างเช่น หากผู้สืบค้นต้องการสืบค้นชื่อผู้แต่ง และชื่อเรื่อง ก็จำเป็นต้องสืบค้นแยกกันทีละครั้งไม่สามารถนำข้อมูลที่ผู้สืบค้นต้องการมารวมกันได้ในการสืบค้นครั้งเดียว

1.2 วัตถุประสงค์ของการวิจัย

1. ประยุกต์เทคนิคทางปัญญาประดิษฐ์ (Artificial Intelligence) คือการสืบค้นแบบฮิวริสติค (Heuristic Search) โดยการนำข้อมูลฮิวริสติค (Heuristic Information) เช่น ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) และคำตอบจากคำถามที่ระบบสอบถามความต้องการของผู้สืบค้น เป็นต้น มาช่วยในการค้นคืนสารสนเทศ (Information Retrieval)

2. ประยุกต์ทฤษฎีโครงข่ายเบย์ (Bayesian Networks) จากทฤษฎีความน่าจะเป็นของเบย์ (Bayes' theorem) ในการเรียงลำดับเอกสารตามน้ำหนักที่คำนวณได้

3. เพื่อพัฒนาการค้นคืนสารสนเทศให้มีประสิทธิภาพ และได้ผลตรงตามความต้องการของผู้สืบค้นยิ่งขึ้น

1.3 ทฤษฎีและหลักการที่ใช้ในงานวิจัย

การวิจัยนี้ได้นำเอาหลักการและทฤษฎีที่เกี่ยวข้องมาประยุกต์ใช้ดังนี้

1. เทคนิคการสืบค้นแบบฮิวริสติค (Heuristic Search)

การสืบค้นโดยใช้เทคนิคฮิวริสติค มีความแตกต่างกับการสืบค้นที่ไม่มีความรู้ช่วยตรงที่ ฮิวริสติค เป็นกระบวนการที่ใช้ความฉลาดในการช่วยการสืบค้น เพื่อให้ได้คำตอบที่รวดเร็วโดยใช้เวลาและหน่วยความจำน้อยที่สุด และผลลัพธ์ที่ได้เป็นที่น่าพอใจ

2. ข้อมูลหรือสารสนเทศฮิวริสติค (Heuristic Information)

เป็นข้อมูลและรายละเอียดที่ช่วยให้การค้นคืนมีความสะดวก และตรงตามความต้องการมากขึ้น โดยข้อมูลเหล่านี้แบ่งเป็นในส่วนของ ฐานข้อมูลเอกสารซึ่งจัดเก็บไว้ในส่วนที่เรียกว่าข้อมูลอธิบายเอกสาร (Document Profile) เป็นหลักการจัดเก็บฟิลด์ต่างๆ ที่ใช้ในการสืบค้นแหล่งเอกสารข้อมูล รวมถึงบทคัดย่อและคำอธิบายประกอบ ซึ่งฟิลด์ต่างๆ เหล่านี้สามารถใช้บอกความต้องการในการตัดสินใจเลือกเอกสารนั้นได้ และส่วนของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) เป็นส่วนที่จัดเก็บข้อมูลส่วนตัวและความสนใจที่เกี่ยวกับผู้สืบค้น

3. ทฤษฎีโครงข่ายเบย์ (Bayesian Network)

โครงข่ายเบย์ คือกราฟแบบมีทิศทางที่ไม่มีวงวน (Directed Acyclic Graph)โดยที่โหนดแทนเหตุการณ์ และเส้นแทนความสัมพันธ์แบบเป็นเหตุเป็นผล บนพื้นฐานของทฤษฎีความน่าจะเป็นที่คำนวณโดยใช้กฎของเบย์ การใช้งานของโครงข่ายเบย์สามารถนำไปประยุกต์ใช้ในระบบผู้เชี่ยวชาญได้ โดยโครงข่ายเบย์แทนความเชื่อและความรู้เกี่ยวกับเหตุการณ์ต่างๆ ซึ่งทฤษฎีโครงข่ายเบย์นี้สามารถนำมาประยุกต์ใช้ในการเรียงลำดับเอกสารได้

4. การวัดประสิทธิภาพการค้นคืนสารสนเทศ (Retrieval Effectiveness Measures)

การวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ ส่วนใหญ่นิยมใช้วิธีการวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ซึ่งผู้วิจัยได้นำวิธีนี้มาใช้วัดประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้พัฒนาขึ้นเช่นเดียวกัน

1.4 ขอบเขตการวิจัย

1. การวิจัยนี้มุ่งพัฒนาการค้นคืนสารสนเทศโดยใช้แนวคิดทางด้านปัญญาประดิษฐ์ คือ การสืบค้นแบบฮิวริสติค

2. นำทฤษฎีโครงข่ายเบย์มาประยุกต์ใช้ในการเรียงลำดับเอกสาร ให้ตรงตามความต้องการของผู้สืบค้นมากที่สุด

3. การวิจัยนี้ได้ทำการทดลองโดยมีพื้นฐานการทำงานบนเครื่องไมโครคอมพิวเตอร์ และได้สร้างระบบฐานข้อมูลซึ่งสามารถทำงานได้โดยใช้โปรแกรมบอร์แลนด์เดลไฟ (Borland Delphi) เวอร์ชัน 4.0 ของบริษัทบอร์แลนด์ อินเตอร์เนชันเนล จำกัด (Borland International Co.,Ltd.)

4. การจำลองแบบระบบค้นคืนสารสนเทศ พัฒนาขึ้นโดยอาศัยข้อมูลเอกสารจากฐานข้อมูลห้องสมุด คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จำนวน 850 ระเบียน และได้นำมาปรับแต่งข้อมูลของเอกสารให้สมบูรณ์เพื่อการทดลอง

1.5 ขั้นตอนของการวิจัย

1. ศึกษาวิธีการค้นคืนเอกสารแบบดั้งเดิม

2. ศึกษาวิธีการสืบค้นแบบฮิวริสติค เช่น การสร้างแลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets) และคิดวิธีการนำข้อมูลฮิวริสติคมาใช้เพื่อช่วยในการสืบค้น

3. วิเคราะห์และประยุกต์การจัดเก็บข้อมูลอธิบายเอกสาร (Document Profile) และข้อมูลประวัติและความสนใจของผู้ใช้ / ผู้สืบค้น (User Profile) เพื่อนำมาใช้ในการสืบค้น

4. ศึกษาทฤษฎีและหลักการของโครงข่ายเบย์

5. ทำการทดลองสร้างฐานข้อมูลเอกสาร และสร้างระบบสืบค้นเอกสารโดยมีพื้นฐานจากทฤษฎีและหลักการที่ได้ศึกษาไว้

6. สรุปผลและประเมินผลการวิจัย พร้อมทั้งเสนอแนวทางในการวิจัยต่อไป

7. เสนอรายงานการวิจัยในรูปวิทยานิพนธ์

1.6 นิยามศัพท์

เพื่อความเข้าใจที่ตรงกัน ผู้วิจัยขอนิยามศัพท์ที่มักกล่าวถึงในวิทยานิพนธ์ดังนี้

1. ผู้ใช้ หรือ ผู้สืบค้น (User) หมายถึง ผู้ที่ต้องการสืบค้นข้อมูลผ่านระบบค้นคืนสารสนเทศ

2. การสืบค้นแบบดั้งเดิม หมายถึง การสืบค้นที่เป็นลักษณะการจับคู่แบบตรงตัว (Exact match) คือถ้าพบเอกสารที่ตรงกับคำถามสืบค้นระบบจะค้นคืนขึ้นมาแสดง แต่ถ้าไม่พบคำตอบระบบจะแสดงผลลัพธ์เป็นเซ็ตว่าง หมายถึงไม่สามารถค้นคืนเอกสารที่ตรงกับคำถามสืบค้นนั้นได้

3. ระบบค้นคืนสารสนเทศ (Information Retrieval System) มีความหมายเดียวกับระบบสืบค้นเอกสาร

4. เอกสารที่ตรงความต้องการ หรือเอกสารที่เกี่ยวข้อง (Relevance Documents) หมายถึงเอกสารที่ค้นคืนขึ้นมาจากระบบแล้ว ผู้สืบค้นประเมินว่าตรงตามความต้องการของตนเอง

5. เอกสารที่ไม่ตรงความต้องการ หรือเอกสารที่ไม่เกี่ยวข้อง (Non-relevance Documents) หมายถึง เอกสารที่ค้นคืนขึ้นมาจากระบบแล้ว ผู้สืบค้นประเมินว่าไม่ตรงตามความต้องการของตนเอง

6. คำเฉพาะ (Keyword) หรือ คำย่อ (Abbreviation) หรือ คำใกล้เคียง (Synonym) หมายถึง คำที่สามารถอ้างอิงถึงตัวเอกสารได้ จัดเก็บลงในฐานข้อมูล เพื่อให้ผู้สืบค้นได้ทำการสืบค้น

7. คำร้องขอ หรือ คำถามสืบค้น หรือคำสืบค้น (Request / Query) หมายถึง คำที่ผู้สืบค้นใส่ลงไปในระบบ เพื่อบ่งบอกขอบเขตของเอกสารที่ต้องการให้ระบบค้นคืนให้

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1. ทำให้การสืบค้นเอกสารมีประสิทธิภาพยิ่งขึ้น โดยผู้สืบค้นจะได้เอกสารที่ตรงตามความต้องการ มากกว่าเอกสารที่ไม่ตรงตามความต้องการ
2. เอกสารที่เป็นผลลัพธ์จัดเรียงตามลำดับความต้องการจากมากไปน้อย
3. เป็นพื้นฐานของการวิจัยในรูปแบบอื่นต่อไป

บทที่ 2
การค้นคืนสารสนเทศและงานวิจัยที่เกี่ยวข้อง

ตั้งแต่ปี คศ. 1940 ปัญหาของการจัดเก็บข้อมูลและการค้นคืนเป็นเรื่องที่น่าสนใจมากขึ้น [22] เนื่องจากเรามีข้อมูลจำนวนมากที่ต้องการความถูกต้องและความรวดเร็วในการเข้าถึง ซึ่งนับวันยิ่งทวีความยากมากขึ้น ผลกระทบประการหนึ่งคือ ข้อมูลที่ตรงตามความต้องการถูกเพิกเฉยเนื่องจากไม่เคยถูกค้นพบ และต้องทำงานหนักขึ้นในการพยายามจะดึงข้อมูลเหล่านั้นขึ้นมา แต่ต่อมาเมื่อเรามีคอมพิวเตอร์ จึงได้มีการนำคอมพิวเตอร์มาใช้พัฒนาระบบค้นคืนสารสนเทศที่ฉลาดยิ่งขึ้น มีห้องสมุดหลายแห่ง ที่มีปัญหาในเรื่องของการจัดเก็บ และค้นคืนสารสนเทศ งานบางงานที่มนุษย์ทำอยู่ เช่น การจัดบัญชีรายชื่อ (Cataloging) และการบริหารงานทั่วไป (General administration) ได้ถูกนำไปให้คอมพิวเตอร์ทำ และประสบผลสำเร็จอย่างดี อย่างไรก็ตามยังมีปัญหาอีกหลายประการในเรื่องของการค้นคืนสารสนเทศที่ยังไม่ได้รับการแก้ไข

โดยหลักแล้ว การจัดเก็บและค้นคืนสารสนเทศเป็นเรื่องง่าย สมมติว่ามีเอกสารจำนวนหนึ่งถูกจัดเก็บไว้ ผู้สืบค้นได้ป้อนคำสั่ง (คำร้องขอ (Request) หรือ คำถามสืบค้น (Query)) ลงไปเพื่อให้ได้กลุ่มเอกสารที่ให้ข้อมูลตามความต้องการ ตรงตามคำถามหรือคำสืบค้นนั้น ผู้สืบค้นจะทำการอ่านเอกสารทั้งหมดที่จัดเก็บไว้ โดยเลือกเก็บเอาแต่เอกสารที่ตรงตามความต้องการ และทิ้งเอกสารส่วนที่ไม่ตรงตามความต้องการออกไป ซึ่งการอ่านนั้น เป็นความพยายามที่จะคัดข้อมูลทั้งทางด้านโครงสร้าง (Syntactic) และ ความหมาย (Semantic) จากตัวอักษร (Text) และใช้ข้อมูลนั้นเป็นตัวช่วยตัดสินใจว่าเอกสารแต่ละเอกสารนั้นตรงหรือไม่ตรงกับคำร้องขอ หรือความต้องการ นับเป็นการยากทีเดียวที่จะต้องรู้ว่าเราจะคัดข้อมูลที่เกี่ยวข้องออกมาได้อย่างไร รวมทั้งต้องตัดสินใจด้วยว่าข้อมูลที่คัดออกมานั้นตรงตามความต้องการหรือไม่ การกระทำเช่นนี้เป็นการค้นคืนแบบสมบูรณ์ (Perfect retrieval) ซึ่งการแก้ปัญหาแบบนี้เป็นสิ่งที่ทำไม่ได้แน่นอน เนื่องจากคงไม่มีผู้สืบค้นคนใดมีเวลา หรือต้องการใช้เวลามากมายเช่นนี้เพื่อทำการอ่านเอกสารทั้งหมดที่มีอยู่ในกลุ่มเอกสาร และตามความเป็นจริงแล้วนั้น คงไม่มีผู้สืบค้นคนใดที่สามารถทำแบบนี้ได้

ปัญหาเกี่ยวกับข้อมูลอีกประการหนึ่ง คือการไม่เข้าใจในวิธีการจัดเก็บ การจัดการ และการส่งข้อมูล ระบบค้นคืนสารสนเทศสมัยใหม่มีความสามารถช่วยปรับปรุงการเข้าถึงข้อมูลต่างๆ ที่ได้ถูกจัดเก็บไว้ได้ดียิ่งขึ้น ดังนั้นจึงได้เกิดแนวคิดของการค้นคืนแบบอัตโนมัติขึ้น และได้มีการนำมาพัฒนาเป็นระบบค้นคืนสารสนเทศโดยใช้คอมพิวเตอร์ช่วยต่อมาในปัจจุบัน ซึ่งวัตถุประสงค์ของนโยบายการค้นคืนอัตโนมัติคือ เพื่อทำการค้นคืนเอกสารที่ตรงตามความต้องการทั้งหมดออกมา ในขณะที่ได้เอกสารที่ไม่ตรงความต้องการออกมาน้อยที่สุดเท่าที่จะเป็นไปได้

2.1 งานวิจัยที่เกี่ยวข้อง

การจะพัฒนาระบบค้นคืนสารสนเทศให้ดียิ่งขึ้นได้นั้น ต้องอาศัยการศึกษางานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศที่มีผู้วิจัยได้เคยทดลองมาแล้วในอดีต ซึ่งผู้วิจัยจะได้นำเสนออย่างคร่าวในบทนี้

งานวิจัยเกี่ยวกับการพัฒนาระบบเพื่อค้นคืนสารสนเทศในอดีต

1. MEDLARS [14] : National Library of Medicine ได้พัฒนาขึ้นเมื่อปี คศ. 1964 เพื่อนำมาใช้ในห้องสมุดทางการแพทย์ และได้รวบรวมวารสารทางการแพทย์เพื่อการค้นคืน โดยใช้คำสั่งสืบค้นแบบบูลลีนอย่างง่าย

2. The DIALOG System [13] : พัฒนาขึ้นโดย Lockheed Information Systems ที่รัฐแคลิฟอร์เนียในปี คศ. 1980 ระบบทำการสร้างกลุ่มของเอกสารอ้างอิงโดยใช้คำสั่ง SELECT และใช้ตัวดำเนินการทางตรรกศาสตร์เพื่อรวมคำแต่ละคำเข้าด้วยกัน

3. STAIRS (Storage and Information Retrieval System) [13] : เป็นโปรแกรมจัดเก็บและค้นคืนสารสนเทศทางด้านการค้าของบริษัท IBM STAIRS แตกต่างจาก DIALOG ตรงที่ STAIRS ไม่เพียงแต่จัดการเรื่องการค้นคืนเอกสารเท่านั้น แต่ยังสามารถจัดการเกี่ยวกับระบบฐานข้อมูลอีกด้วย

4. The Information Bank [13] : เป็นการสืบค้นชื่อบทความที่ได้ตีพิมพ์ในนิตยสาร The New York Times และสิ่งพิมพ์อื่นที่น่าสนใจ ศัพท์ที่ใช้ในการสืบค้นต้องมีรูปแบบที่แน่นอนและถูกต้อง รูปแบบการสืบค้นใช้ตัวดำเนินการทางตรรกศาสตร์

5. การทดลอง TREC (TREC Experiments) ปี คศ. 1993 - 1995 [13] : TREC ย่อมาจาก Text Retrieval Conference เป็นการประชุมทางวิชาการมุ่งเน้นเรื่องการวิจัยเพื่อทำการพัฒนาประสิทธิภาพของการค้นคืนสารสนเทศ การทดลองนี้ได้ศึกษาการจัดการกับฐานข้อมูลแบบ Full-Text ขนาดใหญ่ รวมถึงการประเมินคำตอบที่ได้จากการทดสอบ ซึ่ง TREC นี้ได้มีการพัฒนาให้มีประสิทธิภาพมากยิ่งขึ้นเป็น TREC-1, TREC-2, TREC-3 และ TREC-4

เมื่อได้ศึกษางานวิจัยที่ได้ทำการพัฒนาไปในอดีตแล้ว จากนั้นต้องศึกษาและทำความเข้าใจเรื่องความหมายและคำจำกัดความต่างๆ ที่เกี่ยวกับการค้นคืนสารสนเทศโดยสังเขปดังนี้

2.2 ความหมายของเอกสาร การจัดเก็บ และการค้นคืน

เอกสาร (Document) หมายถึง ข้อมูลที่ถูกจัดเก็บไว้ในหลายรูปแบบ ซึ่งอาจอยู่ในลักษณะของข้อความ เช่น กระดาษที่เขียนหรือพิมพ์ข้อความเอาไว้ หนังสือ บทความวิชาการ หรือบทความวิจัย หรืออาจอยู่ในรูปแบบของข้อมูลรูปภาพและเสียง เป็นต้น

การจัดเก็บ (Storage) หมายถึง การนำข้อมูล หรือเอกสาร ไปทำการจัดเก็บ ดูแล และรักษาไว้ เพื่อการค้นคืนขึ้นมาใช้ประโยชน์ในภายหลัง

การค้นคืน (Retrieval) หมายถึง การดึงข้อมูลที่ต้องการออกมาโดยอาศัยคำถามสืบค้น

2.3 ความหมายของระบบค้นคืนสารสนเทศ

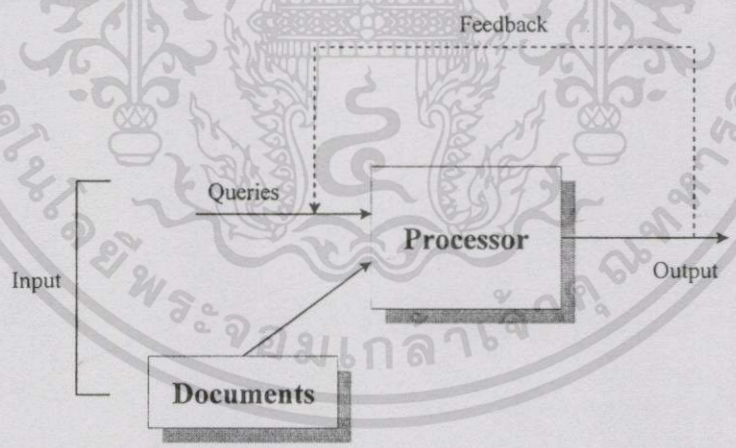
การค้นคืนสารสนเทศ (Information Retrieval : IR) หมายถึง การแสดงข้อมูลที่ผู้สืบค้นต้องการโดยใช้คำถามสืบค้น [13]

การค้นคืนสารสนเทศ หมายถึง การแทน การจัดเก็บ และจัดการ รวมถึงการเข้าถึงข้อมูลในทุกรูปแบบ ซึ่งโดยพื้นฐานแล้วสิ่งที่จะให้ระบบค้นคืนอาจเป็น จดหมาย เอกสารทุกประเภท หนังสือพิมพ์ หนังสือ ผลสรุปทางการแพทย์ บทความวิจัย เป็นต้น [21]

ระบบค้นคืนสารสนเทศ (Information Retrieval System) หมายถึง ระบบใดๆ ซึ่งส่วนใหญ่จะเป็นคอมพิวเตอร์ ที่ทำการค้นคืนค้นข้อมูล [13]

ระบบค้นคืนสารสนเทศ หมายถึง วัตถุ (Object) ที่มีหน้าที่ในการค้นคืนข้อมูล โดยไม่มีมนุษย์เข้ามาเกี่ยวข้อง เป็นเพียงการช่วยเหลือการสืบค้นโดยเครื่องคอมพิวเตอร์เท่านั้น [9]

นอกจากนี้ Van Rijsbergen ได้กล่าวไว้ว่า ระบบค้นคืนสารสนเทศโดยพื้นฐาน มีองค์ประกอบ 3 ประการ ดังรูปที่ 2.1



รูปที่ 2.1 ระบบค้นคืนสารสนเทศโดยพื้นฐานตามที่ Van Rijsbergen ได้กล่าวไว้ [22]

จากรูปที่ 2.1 แสดงให้เห็นองค์ประกอบ 3 ประการที่สำคัญในระบบค้นคืนสารสนเทศโดยทั่วไป คือ Input, Processor และ Output

ส่วนที่ 1 Input เมื่อระบบค้นคืนสารสนเทศเริ่มใช้งาน ผู้สืบค้นจะสามารถเปลี่ยนแปลงคำถามสืบค้นได้ตามความต้องการในการสืบค้นครั้งหนึ่ง (One search session) เพื่อปรับปรุงการค้นคืนในแต่ละครั้งให้ดีขึ้น กระบวนการนี้เรียกว่า Feedback

ส่วนที่ 2 Processor เป็นส่วนหนึ่งของระบบค้นคืนสารสนเทศ ที่เกี่ยวข้องกับกระบวนการค้นคืน ซึ่งกระบวนการนี้เป็นลักษณะการจัดโครงสร้างของข้อมูลในรูปแบบที่เหมาะสม เช่น การจำแนกชนิดของข้อมูล จากไดอะแกรมดังกล่าว เอกสาร (Documents) ได้ถูกแยกออกไปอยู่ที่อื่น แสดงถึงว่า ที่จริงแล้วเอกสารไม่ใช่ Input เพียงอย่างเดียว แต่เป็นสิ่งที่ต้องถูกค้นคืนออกมาด้วย

ส่วนที่ 3 Output คือกลุ่มของเอกสาร หรือจำนวนเอกสารที่ได้

2.4 รูปแบบของคำถามสืบค้น

คำถามสืบค้น (Query) หมายถึงคำที่ผู้สืบค้นเลือก เพื่อมาเป็นคำเฉพาะในการสืบค้นเอกสาร เนื่องจากผู้สืบค้นต้องการเอกสารที่มีคำถามสืบค้นดังกล่าวอยู่ หรืออาจมีเนื้อหาใกล้เคียงกัน ซึ่งคำถามสืบค้นอาจไม่ตรงตามโครงสร้างของภาษา โดยอาจเป็นส่วนหนึ่งของคำได้

ส่วนใหญ่การจับคู่ของคำถามสืบค้นกับข้อมูล (Matching) มักมีลักษณะเป็นการเปรียบเทียบกันโดยตรง (Exact match) ไม่ว่าจะเป็นข้อมูลที่เป็นตัวเลข หรือข้อมูลทางด้านธุรกิจ ดังนั้นข้อมูลใดๆ ก็ตามที่ตรง (Match) กับคำถามสืบค้นจะถูกค้นคืนขึ้นมา ส่วนที่ไม่ตรงจะถูกตัดทิ้งไป ต่อมาได้มีการพัฒนาการจับคู่เป็นช่วง (Range match) ขึ้น ซึ่งนำมาใช้กับช่วงของตัวเลข หรือตัวอักษรเช่นเดียวกัน ดังนั้นจะเห็นได้ว่า หากเรามีข้อมูลที่เป็นข้อความ (Text) หรือรูปภาพ (Image) จะไม่สามารถใช้วิธีทั้ง 2 วิธี ดังกล่าวข้างต้นในการค้นคืนข้อมูลได้ เนื่องจากข้อมูลมีการจัดการที่ไม่ดีนัก รวมทั้งคำถามสืบค้นยังไม่แน่นอนอีกด้วย รูปแบบของคำถามสืบค้นโดยทั่วไปสามารถแบ่งออกได้เป็นดังนี้ [13]

2.4.1 คำถามสืบค้นแบบบูลลีน (Boolean Queries)

คำถามสืบค้นแบบบูลลีนนี้ได้พัฒนามาจากลิสต์ของคำ (Term list) ซึ่งคำถามสืบค้นแบบนี้มีรากฐานมาจากตรรกศาสตร์ หรือพีชคณิตบูลลีน (Boolean algebra) มีลักษณะเป็นการเชื่อมกันของคำโดยใช้ตัวเชื่อมทางตรรกศาสตร์ (Logical connective) ซึ่งโดยพื้นฐานทั่วไปแล้วตัวเชื่อมคือ AND (แทนและ) หมายถึงคำที่อยู่ติดกับคำเชื่อมนี้ต้องปรากฏอยู่ในเอกสารทั้งคู่ ส่วน OR (แทนหรือ) หมายถึงคำที่อยู่ติดกับคำเชื่อมนี้ต้องปรากฏอยู่ในเอกสารอย่างน้อย 1 คำ และ NOT (แทนไม่) หมายถึงคำถามสืบค้นที่อยู่ทางขวาจะต้องไม่ปรากฏอยู่ในเอกสาร การจับคู่เอกสารกับคำถามสืบค้นโดยพื้นฐานทั่วไปมีลักษณะดังตารางที่ 2.1

ตารางที่ 2.1 การจับคู่เอกสารกับคำถามสืบค้น

		เอกสาร	
		ปรากฏคำ	ไม่ปรากฏคำ
คำถามสืบค้น	ปรากฏคำ	จับคู่ตรงกัน	จับคู่ไม่ตรงกัน
	ไม่ปรากฏคำ	จับคู่ไม่ตรงกัน	จับคู่ไม่ตรงกัน

คำถามสืบค้นมักมีลักษณะดังนี้

NOT Query AND Intelligence AND (Retrieval OR Database)

ดังนั้น เอกสารใดก็ตามที่ มีลักษณะ 3 ประการนี้จะถูกค้นคืน คือ

1. ไม่มีคำว่า Query แต่มีคำว่า Intelligence และมีคำว่า Retrieval
2. ไม่มีคำว่า Query แต่มีคำว่า Intelligence และมีคำว่า Database
3. ไม่มีคำว่า Query แต่มีคำว่า Intelligence และมีคำว่า Retrieval และมีคำว่า Database (กรณีนี้ไม่จำเป็น)

อย่างไรก็ตามการสืบค้นโดยใช้คำถามสืบค้นแบบนี้อาจเกิดปัญหาได้หลายประการ เช่น

1. คำถามสืบค้นแบบบูลลีนแท้ๆนั้น มีข้อเสียในเรื่องของการให้น้ำหนักความสำคัญของคำถามสืบค้น ดังนั้นผู้สืบค้นจึงไม่สามารถควบคุมหรือกำหนดได้ว่า คำเหล่านั้นสำคัญเพียงใด
2. อาจมีการเขียนคำถามสืบค้นผิด ซึ่งอาจเป็นเรื่องของการใส่ AND OR ผิด ทำให้คำตอบที่ได้นั้นผิดเพี้ยนไป
3. ลำดับความสำคัญก่อน-หลัง ของตัวเชื่อมทางตรรกศาสตร์ ซึ่งมักใช้เครื่องหมายวงเล็บ เพื่อกำหนดลำดับก่อน-หลังของคำถามสืบค้น ซึ่งถ้าผู้สืบค้นให้ลำดับของวงเล็บผิดตำแหน่งอาจทำให้คำตอบที่ได้ไม่ตรงตามที่คาดหวังไว้
4. ปัญหาเกี่ยวกับเรื่องของการควบคุมขนาดของเซ็ตเอกสารที่ค้นคืน คำถามสืบค้นบางคำ (บางชุด) อาจทำให้ได้เอกสารออกมามากเกินไป หรือคำถามสืบค้นบางคำ (บางชุด) อาจทำให้ได้เอกสารที่น้อยจนเกินไป

Korfhage [13] ได้กล่าวไว้ว่าระบบค้นค้นแบบบูลลีนนี้ ได้รับความนิยมอย่างสูง ซึ่งในงานวิจัยนี้ได้นำพื้นฐานการสืบค้นแบบบูลลีนมาใช้ และได้นำเทคนิคการสืบค้นแบบฮิวริสติกซึ่งเป็นเทคนิคทางด้านปัญญาประดิษฐ์ เข้ามาช่วยแก้ปัญหาที่เกิดขึ้นในการสืบค้นบูลลีนแบบดั้งเดิม ดังที่จะได้กล่าวต่อไปในบทที่ 3

2.4.2 คำถามสืบค้นแบบเวกเตอร์ (Vector Queries)

คำถามสืบค้นแบบบูลลีนมีพื้นฐานมาจากเซ็ตของคำ แต่คำถามสืบค้นแบบเวกเตอร์ แต่ละเอกสารจะถูกแทนด้วยเวกเตอร์ หรือคำต่างๆ ที่เรียงกันมากกว่าจะเป็นกลุ่มคำ ความแตกต่างระหว่างบูลลีนโมเดลกับเวกเตอร์โมเดล เกิดขึ้นเมื่อมีการแทนคำแต่ละคำและวิธีการกำหนดความคล้ายคลึงของเอกสารและคำถามสืบค้น การสืบค้นแบบนี้แทนการปรากฏของคำในเอกสารเป็น 0 และ 1 โดยที่ 0 แทนการไม่ปรากฏคำนั้นในเอกสาร ในขณะที่ 1 แทนการปรากฏคำนั้นในเอกสาร

วิธีการประเมินค่าจะใช้พื้นฐานของเวกเตอร์ 0 และ 1 และอยู่บนพื้นฐานของเวกเตอร์น้ำหนัก (Weighted vector) มีการให้น้ำหนักหรือค่ากับแต่ละคำในเอกสาร ซึ่งการให้น้ำหนักนี้จะให้กับเอกสารโดยอัตโนมัติ โดยอาศัยพื้นฐานของการนับความถี่ของคำ หากมีคำเกิดขึ้นในเอกสารบ่อยครั้ง คือมีความถี่สูง แสดงว่าเอกสารนั้นน่าจะมีความสำคัญมากกว่าเอกสารที่มีความถี่ของคำที่พิจารณาต่ำ

การใช้คำถามสืบค้นแบบเวกเตอร์นี้ผู้สืบค้นสามารถให้น้ำหนักของคำอย่างอิสระ ดังนั้นจึงต้องมีการนอมัลไลซ์ เพื่อให้แน่ใจว่าน้ำหนักที่ได้นั้นเหมาะสมจึงอาศัย Linear transformation ดังนี้

$$S = \frac{s_{\max}(u - u_{\min}) + s_{\min}(u_{\max} - u)}{u_{\max} - u_{\min}} \qquad (2.1)$$

โดยที่ s แทนน้ำหนักที่ระบบให้ และ u แทนน้ำหนักที่ผู้สืบค้นให้

2.4.3 คำถามสืบค้นแบบบูลลีนเพิ่มเติม (Extended Boolean Queries)

คำถามสืบค้นแบบนี้ เป็นการเพิ่มส่วนของการคิดน้ำหนักให้กับการสืบค้นที่ใช้คำถามสืบค้นแบบบูลลีนและเวกเตอร์ ซึ่งน้ำหนักจะอยู่ระหว่าง 0.0 ถึง 1.0 โดยมีสูตรที่ใช้ในการคำนวณ ดังนี้

$$A_{w_1} * B_{w_2} \qquad (2.2)$$

โดยที่ A และ B เป็นคำถามสืบค้น ส่วน w_1 w_2 เป็นน้ำหนัก และ * แทนพีชคณิตบูลลีน หรือการดำเนินการทางตรรกะ การให้น้ำหนักขึ้นอยู่กับแนวคิดของระยะทาง (Distance) ระหว่างกลุ่มของเอกสาร กับคำถามสืบค้น A และ B

2.4.4 คำถามสืบค้นแบบฟัซซี่ (Fuzzy Queries)

คำถามสืบค้นประเภทนี้มีพื้นฐานมาจากฟัซซี่เซ็ต (Fuzzy set) ซึ่งแต่ละ element จะมีการให้ค่าของสมาชิก (Membership grade) ตามเซ็ตที่ให้มา ซึ่งค่าของสมาชิกนี้มีค่าอยู่ระหว่าง 0.0 ถึง 1.0 เช่น สมมติว่ามี U เป็นเอกภพของเซ็ต ซึ่งเป็นเซ็ตที่สมาชิกทุกตัวมีสิทธิถูกพิจารณา ส่วนฟัซซี่เซ็ต S คือ เมื่อ x คือสมาชิกใดๆ ของ U และ μ_s คือฟังก์ชันการให้ค่าของสมาชิก x จากนิยามดังกล่าวจะได้ว่า ทุกๆ สมาชิกใน U ที่ $\mu_s > 0$ จะเป็นสมาชิกของเซ็ต S

อย่างไรก็ตาม การค้นคืนสารสนเทศแบบฟัซซี่นั้น ระบบหรือผู้สืบค้นเองไม่สามารถบอกได้แน่นอนว่าเอกสารที่ค้นคืนมาได้นั้นตรงตามความต้องการเพียงใด แต่ระบบสามารถวัดค่าความเกี่ยวข้องของเอกสารได้โดยเรียงค่าความเกี่ยวข้องจากมากไปหาน้อย ซึ่งบางครั้งอาจเป็นค่าที่ได้ทำการคำนวณมาให้เรียบร้อยแล้ว

2.4.5 คำถามสืบค้นแบบความน่าจะเป็น (Probabilistic Queries)

คำถามสืบค้นประเภทนี้มีลักษณะคล้ายกับฟัซซี่เซ็ต ต่างกันตรงที่ คำถามสืบค้นแบบความน่าจะเป็นนี้มีการคำนวณค่าความน่าจะเป็นของความถี่ของข้อมูลในรูปแบบที่ตายตัว โดยเซ็ตของเอกสารที่ได้ออกมาจากคำถามสืบค้นใดๆ ก็ตาม จะต้องประกอบไปด้วยเอกสารที่ตรงตามคำถามสืบค้นและมีค่าความน่าจะเป็นสูงกว่าค่า threshold ที่ได้กำหนดไว้ รวมทั้งการประเมินค่าเอกสารที่ตรงหรือไม่ตรงกับคำถามสืบค้นของผู้สืบค้น จะมีค่าไม่เกิน 1

2.4.6 คำถามสืบค้นแบบภาษาธรรมชาติ (Natural Language Queries)

คำถามสืบค้นประเภทนี้เป็นการง่ายสำหรับผู้สืบค้นในการหาคำที่ต้องการมาทำการสืบค้น แต่มีข้อเสียตรงที่คำถามสืบค้นนั้นอาจจะไม่ถูกต้อง ไม่แน่นอน และไม่ถูกหลักไวยากรณ์

ดังที่ได้กล่าวมาแล้วว่า งานวิจัยนี้ได้นำการสืบค้นแบบบูลลีนมาใช้โดยได้ปรับปรุงให้ดีขึ้นดังจะได้กล่าวในบทที่ 3 และ บทที่ 4 ต่อไป เมื่อได้ทำการเลือกรูปแบบของคำถามสืบค้นแล้ว ขั้นต่อไปต้องทำการวิเคราะห์และพิจารณาลักษณะของการจัดเก็บข้อมูล หรือโครงสร้างของข้อมูล รวมทั้งรูปแบบของการเข้าถึงข้อมูลที่จัดเก็บไว้ ซึ่งสามารถนำมาใช้ได้กับการสืบค้นแบบบูลลีนนี้ ได้

2.5 วิธีการจัดเก็บและเข้าถึงข้อมูล

ลักษณะของไฟล์ที่จัดเก็บและลักษณะของการเข้าถึงข้อมูลนั้นมีหลายรูปแบบ [8][9] เช่น

1. ไฟล์ที่มีลักษณะเป็นลำดับ และการเข้าถึงแบบเป็นลำดับ (Sequential file and Sequential access)

2. ไฟล์สัมพัทธ์และการเข้าถึงแบบสุ่ม (Relative file and Random access)

3. ไฟล์แบบไม่มีโครงสร้าง (Flat file)

4. ไฟล์ลายเซ็นต์ (Signature file)

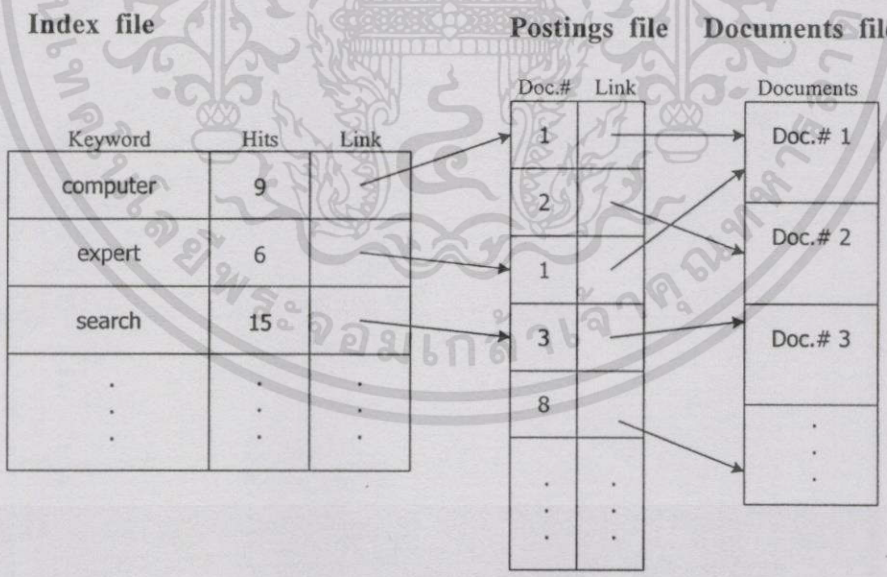
5. แพททรี (Pat tree)

6. กราฟ หรือ โครงข่าย (Graphs or Networks)

7. ไฟล์ผกผัน (Inverted file)

เนื่องจากระบบค้นคืนสารสนเทศแบบบูลลีนนั้น ต้องพยายามหลีกเลี่ยงการเปรียบเทียบทุก Document profile กับคำถามสืบค้นแต่ละตัว หรือแต่ละชุดของคำถามสืบค้น [9] ดังนั้นจึงน่าจะนำโครงสร้างของไฟล์แบบผกผันมาช่วยเพิ่มประสิทธิภาพในการสืบค้น ซึ่งผู้วิจัยได้นำโครงสร้างของไฟล์ผกผันมาใช้ในงานวิจัยด้วย จึงขอลงรายละเอียดเฉพาะไฟล์ผกผันเท่านั้น มีรายละเอียดดังนี้

ไฟล์ผกผัน (Inverted file) คือไฟล์ ที่มีลิสต์ของคำเฉพาะ (Keyword) หรือ แอทริบิวต์ (Attribute) ที่เรียงกันตามดัชนี ซึ่งแต่ละคำเฉพาะจะมีเส้นเชื่อม (Link) ไปยังเอกสารที่มีคำเฉพาะนั้น และมีการบันทึกค่าน้ำหนักความเกี่ยวข้องของแต่ละคำเฉพาะด้วย แสดงดังรูปที่ 2.2



รูปที่ 2.2 โครงสร้างของไฟล์แบบผกผัน

จากรูปที่ 2.2 แสดงให้เห็นถึงโครงสร้างและการเชื่อมถึงกันระหว่าง Index file, Postings file และ Documents file ซึ่งแต่ละ file มีส่วนประกอบดังนี้

1. Index file ประกอบไปด้วย

- Keyword หมายถึง ลิสต์ของคำเฉพาะที่ปรากฏอยู่ในเอกสาร ที่ชี้โดยตัวชี้

- Hits หมายถึง ลิสต์ของจำนวนครั้ง ที่คำเฉพาะทางด้านซ้ายปรากฏอยู่ในเอกสาร

- Link หมายถึง ลิสต์ของตัวชี้ที่อ้างไปยังฟิลด์ Doc# ที่อยู่ในส่วนของ Postings file

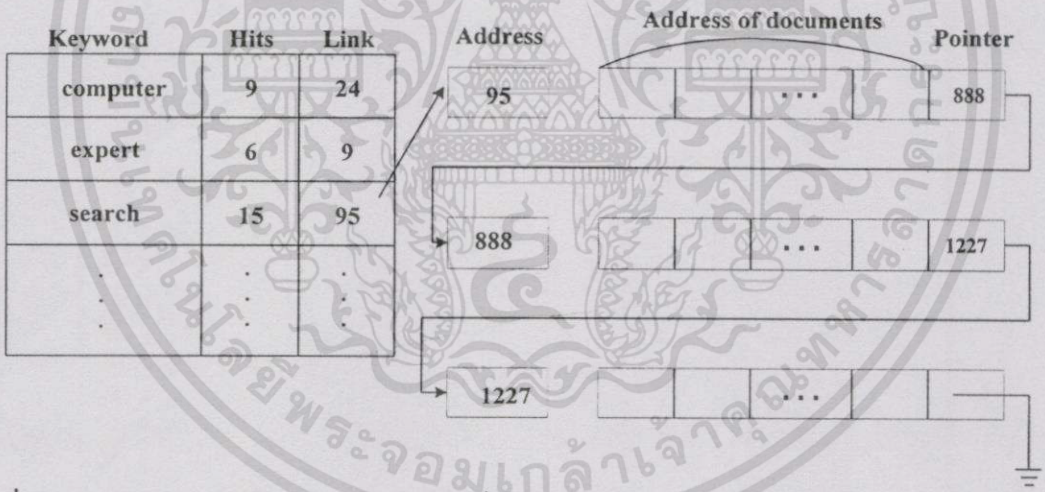
2. Postings file ประกอบไปด้วย

- Doc# หมายถึง ลิสต์ของหมายเลขเอกสารที่ได้จัดเก็บไว้

- Link หมายถึง ลิสต์ของตัวชี้ที่อ้างไปยัง Documents file

3. Documents file ประกอบไปด้วยลิสต์ของเอกสาร ที่มีหมายเลขของเอกสารกำกับอยู่ ซึ่งอาจจะอ้างเป็นเลขตำแหน่งที่จัดเก็บเอกสาร

Frants และคณะได้เสนอรูปแบบของไฟล์ผกผันในลักษณะของการอ้างถึงเลขที่ตำแหน่ง [9] ซึ่งโครงสร้างไฟล์แบบผกผันในอีกลักษณะนี้ แสดงในรูปที่ 2.3



รูปที่ 2.3 โครงสร้างไฟล์ของแบบผกผันตามที่ Frants และคณะได้กล่าวไว้

การใช้ไฟล์แบบผกผันนี้จะช่วยปรับปรุงประสิทธิภาพของการสืบค้นได้โดยเฉพาะไฟล์ข้อความที่มีขนาดใหญ่มาก [8] โดยการสืบค้นของไฟล์นี้จะแบ่งอัลกอริทึมออกเป็น 2 ส่วน คือการสืบค้นคำเฉพาะ ผลลัพธ์ได้เป็นดัชนี จากนั้นอาจนำดัชนีนั้นไปสืบค้นต่อเพื่อจำกัดขอบเขตของผลลัพธ์ให้แคบลงกว่าเดิม ซึ่งท้ายสุดของการสืบค้นโดยใช้โครงสร้างไฟล์แบบผกผันนี้จะได้ผลลัพธ์ออกมาเป็นกลุ่มของเรคอร์ด หรือตัวชี้ไปยังเรคอร์ด

2.6 การวัดประสิทธิภาพของการค้นคืนสารสนเทศ

เมื่อมีการค้นคืนเอกสารออกมาแล้ว จำเป็นต้องตรวจสอบว่าเอกสารนั้นมีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด การวัดประสิทธิภาพของการค้นคืนมีอยู่หลายวิธี ตัวอย่างเช่น Binary Versus N-ary Measures, Precision and Recall , User-Oriented Measures, Average Precision and Recall, Operating Curves and Single Measures เป็นต้น แต่วิธีที่มักนิยมใช้คือ การวัดค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) [21] ซึ่งผู้วิจัยได้นำหลักการนี้มาใช้วัดประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย ดังนั้นจึงขอเสนอรายละเอียดเฉพาะเรื่องการวัดค่าความแม่นยำ และค่าความระลึก

ความหมายและสูตรของการวัดประสิทธิภาพของการค้นคืนสารสนเทศแบบการวัดค่าความแม่นยำ และค่าความระลึก มีดังนี้

ค่าความแม่นยำ (Precision : P) : เป็นอัตราส่วนของเอกสารที่ถูกค้นคืนและตรงตามความต้องการกับเอกสารทั้งหมด ที่ทำการค้นคืนได้ มีสูตรดังนี้

$$P = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ค้นคืนได้}}{\text{จำนวนเอกสารทั้งหมดที่ค้นคืนได้}} \qquad (2.3)$$

ค่าความระลึก (Recall : R) : เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ได้ถูกค้นคืน กับจำนวนเอกสารที่เกี่ยวข้องทั้งหมด มีสูตรดังนี้

$$R = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมดในฐานข้อมูล}} \qquad (2.4)$$

ค่าความแม่นยำ จะวัดว่าเอกสารที่ค้นคืนได้ทั้งหมดนั้นมีเอกสารที่ตรงตามความต้องการจำนวนเท่าใด เช่นการค้นคืนครั้งหนึ่งค้นคืนเอกสารได้ 50 เอกสาร และมี 10 เอกสารที่ได้รับการประเมินว่าตรงตามความต้องการจากผู้ที่ทำการสืบค้น นั่นคือ ค่าอัตราส่วนความแม่นยำ มีค่าเท่ากับ 10/50 หรือ 20% นั่นเอง ส่วนค่าความระลึก หมายถึงการวัดว่าเอกสารที่ต้องการถูกค้นคืนหรือไม่ เช่นผู้สืบค้นต้องการสืบค้นเอกสาร a แล้วเอกสาร a ได้ถูกค้นคืนขึ้นมาจากกลุ่มของเอกสารเช่นเดียวกัน ในกรณีที่ผู้สืบค้นต้องการการสืบค้นที่ครอบคลุมไปทั้งฐานข้อมูล การสืบค้นนั้นจะประสบผลสำเร็จได้ก็ต่อเมื่อเอกสารที่ต้องการได้ถูกค้นคืนขึ้นมา การวัดความสมบูรณ์ของการสืบค้นในฐานข้อมูลจะอ้างถึงอัตราส่วนค่าความระลึก ตัวอย่างเช่น หากมีค่าความระลึก 80% หมายถึง 8/10 ของเอกสารที่ตรงตามความต้องการของผู้สืบค้นที่ปรากฏในฐานข้อมูลได้ถูกค้นพบ

การวัดค่าความแม่นยำ และค่าความระลึก สามารถแสดงได้เป็นตารางที่แสดงถึงผลลัพธ์ที่ได้ในการค้นคืนเอกสาร ดังตารางที่ 2.2

ตารางที่ 2.2 ผลลัพธ์ที่ได้ในการค้นคืนเอกสาร

	ความเกี่ยวข้องกับผู้สืบค้น		
	เกี่ยวข้อง	ไม่เกี่ยวข้อง	รวม
ค้นคืน	A	B	A+B
ไม่ได้ค้นคืน	C	D	C+D
รวม	A+C	B+D	A+B+C+D

เมื่อมีการสืบค้นขึ้นในระบบค้นคืนสารสนเทศ โดยทั่วไปแล้วระบบจะแบ่งกลุ่มของเอกสารเป็น 2 ส่วน คือ เอกสารที่ตรงกับวิธีการสืบค้น (ตามคำถามสืบค้น) ระบบจะค้นคืนเท่ากับ (A+B) และเอกสารทั้งหมดที่ไม่ตรงกับคำถามสืบค้นจะไม่ถูกค้นคืนเท่ากับ (C+D) ซึ่งโดยทั่วไปแล้วจำนวนเอกสารที่ถูกค้นคืนในการสืบค้น จะมีขนาดน้อยเมื่อเทียบกับขนาดของกลุ่มเอกสารทั้งหมด คือในการสืบค้น A+B มีค่าน้อย แต่ค่าของ C+D ซึ่งเป็นจำนวนของเอกสารที่ไม่ถูกค้นคืนมีค่ามาก เช่นเอกสารมีการค้นคืน 80 เอกสารจากทั้งหมด 500,000 เอกสาร ในกรณีนี้ ค่า A+B เท่ากับ 80 ในขณะที่ค่าของ C+D เท่ากับ 499,920 เอกสาร

ในอีกด้านของตารางที่ 2.2 (ทางแนวตั้ง) สัมพันธ์กับความเกี่ยวข้องที่ผู้สืบค้นระบบประเมินสำหรับการสืบค้นที่สมบูรณ์แบบจะต้องค้นคืนเอกสารที่ผู้สืบค้นประเมินว่าตรงตามความต้องการทั้งหมดในฐานข้อมูล (A+C) ในกรณีนี้จะมีความสอดคล้องกันระหว่างการประเมินของผู้สืบค้นและการประเมินของระบบค้นคืนสารสนเทศ นั่นคือ การที่ผู้สืบค้นประเมินเอกสารชิ้นหนึ่งว่าตรงตามความต้องการและระบบก็สามารถประเมินได้ว่าตรงตามความต้องการเช่นกัน (โดยการค้นคืนเอกสารดังกล่าวขึ้นมา) นั่นคือค่า $B = 0$ และ $C = 0$ ดังนั้นจะกล่าวได้ว่าระบบค้นคืนสารสนเทศ มีค่าความแม่นยำเท่ากับ 100% และค่าความระลึก เท่ากับ 100% เช่นเดียวกันซึ่งเป็นไปได้ยาก

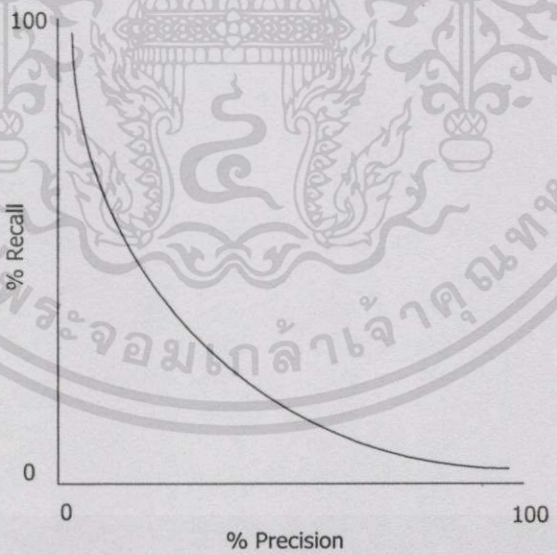
จากตารางที่ 2.2 สังเกตว่าสามารถคำนวณหาค่า P และ R ได้ดังนี้

$$P = \frac{A}{A+B} \qquad \text{และ}$$

$$R = \frac{A}{A+C}$$

ค่าความแม่นยำ และค่าความระลึก จะขึ้นต่อกัน ค่าความระลึก อาจมีค่าถึง 100% ได้ ถ้ามีการค้นคืนเอกสารได้ทั้งหมดในฐานข้อมูล (A+B+C+D) แต่ค่าของความแม่นยำจะต่ำลงอย่างมาก เพราะในการใส่คำถามสืบค้น 1 ชุดคำสั่ง อาจมีเอกสารส่วนใหญ่ที่ไม่ตรงตามความต้องการปะปนออกมาด้วย ดังนั้น ค่าความแม่นยำอาจมองได้ว่าเป็นปัจจัยในเรื่องเวลาที่ผู้สืบค้นใช้ในการประเมินว่าเอกสารชิ้นใดตรงหรือไม่ตรงตามความต้องการของตนเอง ถ้าต้องใช้เวลามาก ค่าความแม่นยำจะลดลง ซึ่งในการสืบค้นครั้งหนึ่งหากได้ ค่าความระลึก 75% และ ค่าความแม่นยำ 50% จะมีประสิทธิภาพมากกว่าการสืบค้นที่ได้ค่าความระลึก 75% และ ค่าความแม่นยำ 25% และจะมีประสิทธิภาพมากกว่า ค่าความระลึก 75% และ ค่าความแม่นยำ 10% ตามลำดับ ค่าความแม่นยำและค่าความระลึก มีลักษณะที่ผกผันกันดังแสดงในรูปที่ 2.4

ผู้สืบค้นที่ต่างกันจะมีความต้องการในค่าความแม่นยำและความระลึกที่แตกต่างกัน และจะมีความต้องการที่แตกต่างกันในเวลาที่ต่างกันด้วยเช่นกัน กรณีที่ผู้สืบค้นต้องการค่าความระลึกสูง เช่นในกรณีที่ผู้สืบค้นต้องการเขียนหนังสือ หรือต้องการหาหัวข้อที่น่าสนใจสำหรับทำการวิจัย ผู้สืบค้นจึงต้องการการสืบค้นที่ได้เอกสารออกมาแบบกว้าง เนื่องจากไม่ต้องการพลาดหัวข้อที่น่าสนใจที่ไม่ได้ถูกค้นคืน แต่ในขณะเดียวกันผู้สืบค้นต้องยอมรับค่าความแม่นยำที่ต่ำ ในทางตรงกันข้าม หากผู้สืบค้นต้องการหัวเรื่องที่เฉพาะเจาะจงในเรื่องใดเรื่องหนึ่ง การสืบค้นนั้นจะได้ค่าความแม่นยำสูง แต่ค่าความระลึกต่ำ



รูปที่ 2.4 กราฟความสัมพันธ์ระหว่างค่าความแม่นยำ และค่าความระลึก

ดังนั้นสามารถสรุปได้ว่าค่าความแม่นยำวัดความสามารถของระบบในการกำจัดเอกสารที่ไม่ต้องการออกไป เพื่อให้ได้ความแม่นยำในการสืบค้น ส่วนค่าความระลึกวัดความสามารถของระบบในการค้นคืนเอกสารที่ตรงตามความต้องการ

เพื่อการง่ายต่อการวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ โดยใช้ค่าทั้ง 2 ค่า จึงมีการรวมค่าความแม่นยำและค่าความระลึกเป็นค่าเพียงค่าเดียว คือ ค่า E เสนอโดย Van Rijsbergen มีสมการเป็นดังนี้

$$E = 1 - \frac{(1+b^2)PR}{b^2P + R} \qquad (2.5)$$

โดยที่ P คือค่าความแม่นยำ และ R คือค่าความระลึก ส่วน b เป็นการวัดความสำคัญเชิงสัมพัทธ์ของค่าทั้ง 2 ค่า ต่อผู้สืบค้น โดยถ้าให้ค่า b = 1 แสดงว่าผู้สืบค้นสนใจค่าทั้งสองเท่ากัน และถ้าให้ค่า b = 2 แสดงว่าผู้สืบค้นมีความสนใจในค่าความระลึกเป็น 2 เท่าของค่าความแม่นยำ เป็นต้น ดังนั้นผลจากการวัดประสิทธิภาพจะเป็นค่าๆเดียว ทำให้สามารถวัดค่าทั้งสองค่าในการค้นคืนครั้งเดียวได้ หรือสามารถเปรียบเทียบประสิทธิภาพของระบบ 2 ระบบ โดยการใช้ชุดคำสืบค้นเพียงชุดเดียวได้ ซึ่งระบบค้นคืนสารสนเทศใดที่มีค่า E น้อยแสดงว่ามีประสิทธิภาพสูง

2.7 ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile)

เนื่องจากงานวิจัยนี้ ได้นำแนวคิดเกี่ยวกับ ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้นมาประยุกต์ใช้เพื่อช่วยในเรื่องของการสืบค้นด้วย ดังนั้นจึงขอกล่าวถึงความหมายของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้นดังนี้

ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User profile) ประกอบไปด้วย ข้อมูลส่วนตัวของผู้สืบค้น และข้อมูลเกี่ยวกับความสนใจที่ผู้สืบค้นต้องการ ซึ่งฐานข้อมูลของระบบค้นคืนสารสนเทศ จะทำการรวบรวมข้อมูลของการสืบค้นครั้งก่อนเพื่อเพิ่มประสิทธิภาพในการค้นคืนให้กับระบบ โดยข้อมูลพื้นฐานของผู้สืบค้นนี้จะช่วยในการกำหนดว่าเอกสารใดบ้าง ที่อยู่ในกลุ่มของเอกสารที่จะถูกค้นคืน ข้อมูลเหล่านี้อาจได้มาจาก บรรณารักษ์อ้างอิง หรือการสัมภาษณ์ผู้สืบค้นซึ่งอาจพบข้อมูลใน 2 ลักษณะใหญ่ซึ่งมีรายละเอียดโดยสังเขปดังนี้ [13]

2.7.1 ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น อย่างง่าย (Simple Profiles)

เป็น User Profile อย่างง่าย คล้ายๆกับคำถามสืบค้น ประกอบไปด้วยกลุ่มของคำที่มีการให้น้ำหนัก มักใช้กับระบบที่มีการกำหนดตัวของผู้สืบค้น และความสนใจของผู้สืบค้นแบบตายตัว

Simple User Profile นี้ โดยธรรมชาติแล้วง่ายสำหรับการจับคู่กับฐานข้อมูลเอกสาร แต่อย่างไรก็ตาม มีการจำกัดลักษณะของเอกสารที่ผู้สืบค้นต้องการ เพราะว่ามีการจำกัดกับคำเฉพาะและวลีสืบค้นที่ต้องการบนข้อความ

2.7.2 ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น อย่างละเอียด (Extended Profiles)

เป็น User Profile ที่ค่อนข้างละเอียด สามารถแบ่งการเก็บข้อมูลเป็น 6 ลักษณะ ดังนี้

2.7.2.1 ระดับการศึกษา (Educational level) : ผู้สืบค้นที่มีระดับการศึกษาต่างกันอาจต้องการกลุ่มเอกสารที่แตกต่างกัน โดยการใช้คำสืบค้นที่เหมือนกัน

2.7.2.2 ความคุ้นเคยกับขอบเขตของการถาม (Familiarity with the area of inquiry) : ต้องรู้ว่าผู้สืบค้นน่าจะถามในขอบเขตอะไร หรือผู้สืบค้นสนใจในขอบเขตเรื่องใด

2.7.2.3 ความสามารถทางด้านภาษาของผู้สืบค้น (Language capabilities) : บทความที่เขียนโดยภาษาที่ผู้สืบค้นไม่คุ้นเคย จะมีประโยชน์น้อยหากผู้สืบค้นไม่สามารถอ่านภาษานั้นได้ เช่น ผู้สืบค้นไม่รู้ภาษาญี่ปุ่นแต่ในฐานข้อมูลมีเอกสารที่เป็นภาษาญี่ปุ่น เอกสารนั้นจะตรงความต้องการน้อย เป็นต้น

2.7.2.4 การเป็นสมาชิกวารสาร (Journal subscriptions) : วารสารที่ผู้สืบค้นเป็นสมาชิก หรืออยู่ในลิตส์ที่เคยอ่านแล้ว จะทำการเก็บไว้เป็นแหล่งอ้างอิง

2.7.2.5 นิสัยการอ่าน (Reading habits) : ถ้าผู้สืบค้นอ่านวารสารฉบับหนึ่งเป็นประจำ ผู้สืบค้นจะรู้เกี่ยวกับบทความที่สำคัญในวารสารนั้น จึงเก็บชื่อวารสารนั้นไว้อ้างอิงว่าผู้สืบค้นคนนี้ชอบอ่านวารสารประเภทนี้ เป็นต้น หากผู้สืบค้นยังไม่เคยอ่านวารสารนั้นมาก่อน และไม่มีชื่อในลิสต์ วารสารนั้นจะถูกนำไปอ้างอิงเป็นเอกสารตัวใหม่

2.7.2.6 สิ่งที่ชอบโดยส่วนตัว (Specific preference) : ผู้สืบค้นอาจมีผู้เขียนที่ชอบหรือมีวารสารที่ชอบและติดตามเป็นประจำอยู่แล้ว ดังนั้นผู้สืบค้นน่าจะสนใจวารสารชิ้นนี้มากกว่าชิ้นอื่น

ข้อมูลนี้ไม่สามารถใช้ได้โดยตรงในกระบวนการค้นคืน แต่ประยุกต์ใช้ในเรื่องการจัดการเกี่ยวกับความเหมาะสมในเรื่องของการสืบค้น เพื่อให้ตรงตามความต้องการของผู้สืบค้นและกำจัดเอกสารบางชิ้นที่ไม่เหมาะสมออกไปเท่าที่จะเป็นไปได้

บทที่ 3
ทฤษฎีและหลักการที่ใช้ในงานวิจัย

เนื่องจากงานวิจัยนี้ได้มาจากการศึกษาแนวทางในการค้นคืนสารสนเทศโดยใช้เทคนิคการสืบค้นทางด้านปัญญาประดิษฐ์ คือ เทคนิคการสืบค้นแบบฮิวริสติค ผนวกเข้ากับการค้นคืนสารสนเทศโดยอาศัยทฤษฎีโครงข่ายเบย์ ดังนั้นในบทนี้ ผู้วิจัยจึงขอกล่าวถึงทฤษฎีและหลักการที่ใช้ในงานวิจัย ซึ่งแบ่งออกเป็น 2 ส่วนใหญ่ คือ การสืบค้นแบบฮิวริสติคและการประยุกต์เทคนิคนี้เข้ากับการค้นคืนสารสนเทศ และการนำทฤษฎีความน่าจะเป็นของเบย์มาใช้ช่วยในการเรียงเอกสาร ในหัวข้อที่ 3.1 และ 3.2 ตามลำดับ

3.1 การสืบค้นแบบฮิวริสติคและการประยุกต์เพื่อการค้นคืนสารสนเทศ

เทคนิคการสืบค้นแบบฮิวริสติค เป็นทฤษฎีที่น่าสนใจสำหรับการนำมาช่วยการค้นคืนสารสนเทศ เพื่อนำไปสู่การช่วยเหลือผู้ใช้หรือผู้สืบค้นในการเลือกคำถามสืบค้นที่เหมาะสม ซึ่งผู้วิจัยขอกล่าวถึงความหมายของการสืบค้นในมุมมองของการสืบค้นแบบฮิวริสติค ประเภทของการสืบค้น การสืบค้นแบบฮิวริสติค และอัลกอริทึม A* ตามลำดับ

3.1.1 ความหมายของการสืบค้น

การสืบค้นในมุมมองทางด้านปัญญาประดิษฐ์ หมายถึง การแก้ไขปัญหาที่มีเทคนิคพื้นฐานใช้ในการตรวจหาคำตอบที่เป็นไปได้ ปัญหาของการสืบค้นมักใช้แผนภาพของการสืบค้นเพื่อบรรยายลักษณะของปัญหา กราฟจะมีโหนดแรก คือโหนด i ซึ่งเป็นโหนดเริ่มต้น และโหนด g ซึ่งเป็นโหนดเป้าหมาย (Goal node) วัตถุประสงค์คือ ทำการค้นหาเส้นทางผ่านเอกภพของการสืบค้น (Search space) ที่อยู่เชื่อมระหว่างโหนดเริ่มต้นกับโหนดเป้าหมาย

3.1.2 รูปแบบของการสืบค้น

การสืบค้นมี 2 รูปแบบ [17] คือ

3.1.2.1 การสืบค้นที่ไม่มีความรู้ช่วย (Blind search)

เป็นวิธีการสืบค้นแบบไม่มีข้อมูลหรือความรู้ช่วยในการสืบค้น การสืบค้นจะทำอย่างมีระบบ เมื่อพบคำตอบจะส่งคำตอบกลับไปจุดเริ่มต้น การสืบค้นแบบนี้ แบ่งออกเป็นหลายลักษณะ เช่น Breadth first search และ Depth first search เป็นต้น

3.1.2.2 การสืบค้นที่มีความรู้ช่วย (Informed search)

เป็นการสืบค้นที่อาศัยข้อมูลและความรู้ ช่วยในการสืบค้น ทำให้การสืบค้นรวดเร็วและตรงตามความต้องการมากขึ้น การสืบค้นแบบนี้ขึ้นกับการใช้ข้อมูลฮิวริสติค (Heuristic information) ตัวอย่างวิธีการสืบค้นแบบนี้ ได้แก่ Hill Climbing, Best-First Search, Branch-and-Bound Search, Optimal Search และ A* Search เป็นต้น ในงานวิจัยนี้ได้นำวิธีการสืบค้นแบบ Best-First Search มาประยุกต์ใช้ เพื่อทำการค้นหาโหนดที่มีค่าใช้จ่ายต่ำที่สุด รวมถึงอัลกอริทึม A* และ อัลกอริทึม IRA ที่ช่วยในการสืบค้นเอกสาร ซึ่งจะกล่าวถึงรายละเอียดต่อไป

3.1.3 ความหมายของการสืบค้นแบบฮิวริสติค

คำว่า "heuristic" มาจากภาษากรีก คือคำว่า heuriskein มีความหมายว่า "เพื่อทำการค้นหา" หรือ "ค้นพบ" ซึ่งความหมายทางเทคนิคของฮิวริสติค ได้มีการเปลี่ยนแปลงหลายครั้งตามประวัติศาสตร์ของ AI [20] ซึ่งส่วนใหญ่ความหมายของฮิวริสติคที่มีการเปลี่ยนแปลงอยู่บ่อยครั้งนั้น ขึ้นอยู่กับการตีความของผู้เขียนและเวลา โดยทั่วไปแล้วมักอธิบายความหมายโดยการยกตัวอย่างเพื่อให้เกิดความเข้าใจมากขึ้น

Feigenbaum และ Feldman กล่าวไว้เมื่อปี ค.ศ.1963 ว่า

> "A *heuristic (heuristic rule, heuristic method)* is a rule of thumb, simplification, or any other kind of device which drastically limits search for solutions in large problem [more correctly, solution] spaces." [1]

การสืบค้นแบบฮิวริสติคนั้นไม่ได้รับประกันว่าผลลัพธ์ที่ออกมาจะต้องเป็นคำตอบที่ดีที่สุด แต่ฮิวริสติคสามารถหาคำตอบที่ดีได้บ่อยครั้ง

ดังที่ได้กล่าวไปแล้วข้างต้น การสืบค้นโดยใช้เทคนิคฮิวริสติค มีความแตกต่างกับการสืบค้นที่ไม่มีความรู้ช่วยตรงที่ **ฮิวริสติค เป็นกระบวนการที่ใช้ความฉลาดในการช่วยการสืบค้นเพื่อให้ได้คำตอบที่รวดเร็ว โดยใช้เวลาและหน่วยความจำน้อยที่สุด และผลลัพธ์ที่ได้เป็นที่น่าพอใจ** โดยทั่วไปฮิวริสติค ใช้แก้ปัญหาใน 2 ลักษณะคือ [16]

1. ปัญหาที่มีวิธีการหาคำตอบที่ไม่แน่นอน เนื่องจากมีความคลุมเครือ เช่นการวินิจฉัยทางการแพทย์ อาการของคนไข้อาจมาจากหลายสาเหตุ ดังนั้นแพทย์จึงใช้ฮิวริสติค เพื่อวินิจฉัยโรคที่น่าจะเป็นไปได้ และวางแผนการรักษา

2. ปัญหาที่มีการหาคำตอบที่แน่นอน เป็นปัญหาที่ Depth first search และ Breadth first search สามารถแก้ไขได้ แต่ต้องใช้เวลา หรือหน่วยความจำมาก หากใช้ฮิวริสติคในการแก้ปัญหาจะทำให้ได้คำตอบที่พึงพอใจ ภายในเวลาและการใช้หน่วยความจำที่เหมาะสม

เหตุผลของการที่ข้อมูลฮิวริสติคสามารถนำมาประยุกต์ใช้ในการสืบค้นมีดังนี้ [2]

1. ข้อมูลฮิวริสติคสามารถนำมาใช้ช่วยในการตัดสินใจได้ว่า ควรจะขยาย (Expand) โหนดใดต่อไป แทนที่จะทำการขยายโหนดแบบวิธีที่จำกัดตามลำดับของการสืบค้นแบบ Depth first search และ Breadth first search โดยแนวคิดแบบนี้ทำการขยายโหนดที่มีความเหมาะสม หรือ โหนดที่คาดว่าดีที่สุด (Most promising) ซึ่งการสืบค้นที่นำเอาหลักการนี้ไปประยุกต์นั้นเรียกว่า Ordered search หรือ Best-first search

2. การขยายโหนด ในแต่ละเส้นทาง การสืบค้นแบบฮิวริสติคจะใช้ข้อมูลฮิวริสติคทำการตัดสินใจว่าตัว successor ใดที่ควรจะ generate แล้วกระทำเฉพาะตัวที่เลือกเท่านั้น แต่สำหรับการสืบค้นแบบไม่มีความรู้ช่วยจะ generate ทุกๆ ตัว successor ที่เป็นไปได้พร้อมๆ กัน ในเวลาเดียวกัน โดยโหนดที่ถูกเลือกเพื่อการ generate ต่อไป เรียกว่าโหนดนั้นได้ถูกทำการ Partially developed หรือ Partially expanded

3. การใช้ข้อมูลฮิวริสติคช่วยตัดสินใจได้อย่างแน่นอนว่า โหนดใดควรถูกตัดทิ้งออกไปจาก แผนภูมิต้นไม้เพื่อการสืบค้น (Search tree) อันเนื่องมาจากโหนดดังกล่าวไม่เป็นส่วนหนึ่งของคำตอบ หรือควรที่จะขยายโหนดใดบ้าง นอกจากนั้น วิธีการสืบค้นแบบ Best-first search ยังช่วยประหยัดเนื้อที่ที่ควรจะเสียไปในหน่วยความจำที่สงวนเอาไว้สำหรับโหนดที่ไม่เหมาะสม

3.1.3.1 อัลกอริทึม Best-First Search [24]

อัลกอริทึม Best-first search มีลักษณะดังนี้

```
Best-First Search(root)
begin
    open ∈ ϕ ; n ← root
    WHILE (n is not a goal node)
        EXPAND n, generating and evaluating all its children
        INSERT all its children into open
        DELETE n from open
    n ← a minimum - cost node in open
end.
```

หลักการทำงานของ Best-first search คือมีการใช้ลิสต์ (List) เพื่อจัดเก็บสถานะ ซึ่งมีอยู่ 2 ประเภท คือ ลิสต์ OPEN และ ลิสต์ CLOSED ลิสต์ OPEN ใช้สำหรับเก็บโหนดที่จะดำเนินการค้นหาในสถานะปัจจุบัน และ ลิสต์ CLOSED ใช้สำหรับบันทึกโหนดที่ได้ทำการสืบค้น (Expanded) เรียบร้อยแล้ว ในแต่ละวงรอบการทำซ้ำ Best-first search ย้ายจากสมาชิกตัวแรก (First element) ไปยังตัวถัดไป ถ้าพบสภาวะที่เป็นเป้าหมาย อัลกอริทึมส่งค่าเส้นทางของคำตอบที่นำไปสู่เป้าหมายกลับไป แต่ถ้าสมาชิกตัวแรกในที่อยู่ในลิสต์ OPEN ไม่ใช่โหนดเป้าหมาย อัลกอริทึมจะดูที่ตัว children ซึ่งเป็นทางเลือกถัดไป ถ้าสถานะของ children อยู่ในลิสต์ OPEN หรือ ลิสต์ CLOSED อยู่แล้ว อัลกอริทึมจะทำการตรวจสอบเพื่อให้แน่ใจว่า ทั้ง 2 สถานะ ที่เกิดขึ้นนั้น เป็นเส้นทางที่สั้นที่สุด

Best-first search ให้ค่าของฮิวริสติค ในสถานะที่ลิสต์ OPEN และลิสต์ถูกเรียงลำดับตามค่าของฮิวริสติค อันจะนำมาซึ่งสถานะที่ดีที่สุดก่อนจัดเก็บในลิสต์ OPEN เนื่องมาจากการประมาณค่านั้น เป็นธรรมชาติของฮิวริสติค สถานะถัดมาที่พิจารณาอาจมาจากระดับขั้น (Level) ใดๆ ก็ตามใน เอกภพของสถานะ (State space) และการจัดเก็บโหนดในลิสต์ OPEN ทำการเรียงลำดับลิสต์เป็นคิวตามความสำคัญ (Priority queue)

เป้าหมายของ Best-first search คือหาสถานะเป้าหมาย โดยใช้การสืบค้นสถานะให้น้อยครั้งที่สุดเท่าที่จะเป็นไปได้

3.1.4 อัลกอริทึม A*

เป็นเทคนิคการสืบค้นแบบ optimal search สำหรับ optimal solution อัลกอริทึม A* มีลักษณะคล้าย Best-first search ซึ่งมีลักษณะดังนี้

ให้ค่าเริ่มต้นของลิสต์ของเส้นทางมีความยาวเท่ากับ 0 , เป็น step ที่ 0 ถ้าลิสต์ว่าง แสดงว่ายังไม่พบคำตอบ

-> ถ้าเส้นทางแรกในลิสต์ตรงกับเป้าหมาย, จบการทำงาน

- ถ้าไม่ตรงกับเป้าหมาย : ลบเส้นทางแรกออกจากลิสต์
- ขยายโหนด n ของเส้นทาง จากนั้นสร้างเส้นทางเพิ่ม
- รวมเส้นทางใหม่เข้าไปยังลิสต์
- เรียงลิสต์โดยใช้ฟังก์ชัน $f^*(n) = g^*(n) + h^*(n)$ ใส่ค่าใช้จ่ายน้อยที่สุดไว้ด้านหน้า
- ถ้า 2 เส้นทางหรือมากกว่านั้นผ่านโหนดร่วมกัน ให้ลบเส้นทางที่มีโหนดนั้นอยู่ยกเว้นเส้นทางที่มีค่าใช้จ่ายน้อยที่สุด

-> ทำซ้ำ

จากอัลกอริทึม A* ได้ฟังก์ชันฮิวริสติคที่มีสมการดังนี้

$$f^*(n) = g^*(n) + h^*(n) \qquad (3.1)$$

โดยที่

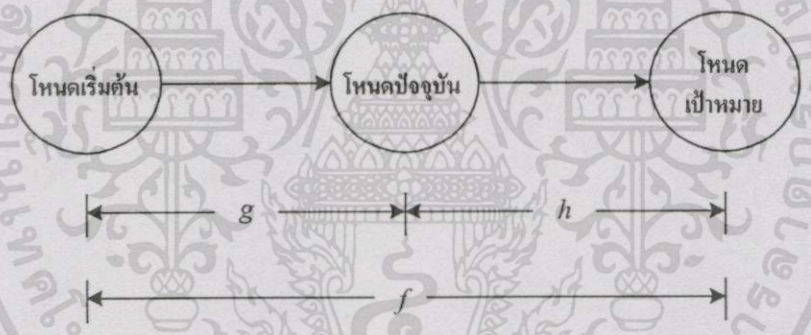
$g^*(n)$ เป็นค่าใช้จ่ายที่น้อยที่สุดของเส้นทางจากโหนดเริ่มต้นไปยังโหนด n

$h^*(n)$ เป็นค่าประมาณที่น้อยที่สุดจากโหนด n ไปยังโหนดเป้าหมาย

ดังนั้น

$f^*(n)$ เป็นค่าใช้จ่ายโดยประมาณที่น้อยที่สุด ตามเส้นทางจากโหนดเริ่มต้น ไปยังโหนดเป้าหมายโดยผ่านโหนด n

เพื่อให้เห็นภาพที่ชัดเจนขึ้น ดูแผนภาพประกอบการสืบค้นโดยใช้อัลกอริทึม A* ในรูปที่ 3.1



รูปที่ 3.1 แผนภาพประกอบการสืบค้นโดยใช้อัลกอริทึม A*

จากรูปที่ 3.1 สามารถคำนวณหาเส้นทางที่ดีที่สุดโดย $f = g + h$ และมี f^*, g^* และ h^* เป็นการคาดการณ์ค่าที่แท้จริงของ f , g และ h ตามลำดับ โดยจำเป็นต้องรู้ค่าใช้จ่ายในการหาเส้นทางแต่ละครั้งเพื่อทำการคำนวณค่า g^* และ การประมาณค่าฟังก์ชันฮิวริสติค เพื่อพิจารณาว่ายากหรือง่ายเพียงใดในการเดินทางจากสถานะปัจจุบันไปสถานะเป้าหมาย (h^*) ทั้งนี้มีข้อกำหนดของ g^* และ h^* ไว้ดังนี้

3.1.4.1 ข้อกำหนดและรายละเอียดของ g^*

1). g^* ทำการคำนวณโดยใช้ค่าใช้จ่ายที่แท้จริงของของเส้นทางที่ได้สำรวจเท่านั้น

2). g^* ได้จากการรวมค่าใช้จ่ายของเส้นทางทุกเส้นทาง จากโหนดเริ่มต้นไปยังโหนดปัจจุบัน

3). หากว่าเอกภพของการสืบค้นเป็นแผนภูมิต้นไม้แล้ว $g^* = g$ โดยที่จะมีเส้นทางระหว่างโหนดเริ่มต้นไปยังโหนดปัจจุบันเพียงเส้นทางเดียวเท่านั้น

4). โดยทั่วไปแล้วเอกภพของการสืบค้นมักเป็นกราฟ ซึ่งในกรณีนี้จะได้ว่า $g^* >= g$ และ g^* จะไม่น้อยกว่าค่าใช้จ่ายของเส้นทางที่สั้นที่สุด

5). g^* สามารถเท่ากับ g ในกราฟถ้าเลือกอย่างถูกต้อง

3.1.4.2 ข้อกำหนดและรายละเอียดของ h^*

1). h^* เป็นข้อมูลฮิวริสติคที่แทนการคาดเดาถึงความยาก ในการเดินทางจากโหนดปัจจุบันไปยังโหนดเป้าหมาย

2). h^* อาจทำการประมาณโดยใช้ฟังก์ชันประเมินค่า (Evaluation function) ที่วัดความดีหรือ ความเหมาะสม (Goodness) ของโหนด (คล้าย Best-first search)

3). ใช้สำหรับอัลกอริทึม A* ในการหาคำตอบของค่าใช้จ่ายที่น้อยที่สุด โดยมีเงื่อนไขดังนี้ :

- $h^*(n)$ ต้องมีค่ามากกว่า หรือเท่ากับ 0

- $h^*(n)$ ต้องมีค่าน้อยกว่าหรือเท่ากับ $h(n)$; h^* ไม่ควรมากกว่าค่าใช้จ่ายที่จะไปยังโหนดเป้าหมายของโหนดใดๆ ดังนั้นจึงเรียกว่าสภาวะที่ยอมรับได้ (Admissability Condition) (ถ้าอัลกอริทึม เป็นที่รับรองว่าสามารถหาเส้นทางของคำตอบที่มีค่าใช้จ่ายน้อยที่สุด (ถ้ามี) ก็จะสามารถยอมรับได้)

4). ถ้า $h^* = h$ (ฮิวริสติคแบบสมบูรณ์) แสดงว่าไม่ต้องมีโหนดใดที่ต้องขยายต่อไป

5). ในสถานะของโหนดเป้าหมาย จะทำให้ h^* ใกล้กับ h ที่สุดเท่าที่จะเป็นไปได้ โดยไม่มีการประมาณค่า h ที่มากเกินไป

6). สภาวะที่ยอมรับได้ ไม่ได้เน้นเรื่องประสิทธิภาพมากนักแต่จะได้คำตอบออกมาเช่นเดียวกัน ซึ่งคำตอบนั้นอาจไม่ใช่คำตอบที่ดีที่สุด

จากอัลกอริทึม A* ที่กล่าวข้างต้น สามารถนำมาประยุกต์เป็นอัลกอริทึมทางปัญญาประดิษฐ์ที่ใช้เพื่อการค้นคืนสารสนเทศ เรียกว่า อัลกอริทึม IRA ดังจะได้กล่าวต่อไปในหัวข้อที่ 3.1.5

3.1.5 การประยุกต์การสืบค้นแบบฮิวริสติคเพื่อการค้นคืนสารสนเทศ

จากอัลกอริทึม A* ที่ได้กล่าวไปแล้วในหัวข้อที่ 3.1.4 สามารถนำมาประยุกต์สำหรับการสืบค้นเอกสาร [11] ซึ่งมีรายละเอียดดังนี้

3.1.5.1 อัลกอริทึม A* กับการสืบค้นเอกสาร

การประยุกต์อัลกอริทึม A* จากปัญญาประดิษฐ์เพื่อสืบค้นเอกภพของเอกสาร เทคนิคการสืบค้นที่ใช้ฮิวริสติค นับเป็นเครื่องมือที่มีประโยชน์และสามารถนำไปปฏิบัติได้จริง สำหรับการแก้ปัญหา เอกภพของปัญหาจำนวนมากสามารถแทนได้ด้วยกราฟ อัลกอริทึม Best-First Search (ตามที่ได้เสนอในหัวข้อ 3.1.3.1) ที่มีลักษณะคล้ายกับ A* (ตามที่ได้เสนอในหัวข้อ 3.1.4) มักใช้สำหรับหาเส้นทางของค่าใช้จ่ายที่น้อยที่สุดของกราฟ โดยใช้ข้อมูลที่เป็นฮิวริสติค ที่เกี่ยวข้องกับค่าใช้จ่ายของส่วนที่ยังไม่ทราบคำตอบสำหรับโหนดใดๆ ของกราฟ นิยามฟังก์ชันดังนี้

$g(n)$ = ค่าใช้จ่ายที่น้อยที่สุดตามเส้นทางจากโหนดเริ่มต้น (Start node) ถึงโหนด n

$h(n)$ = ค่าใช้จ่ายประมาณการที่น้อยที่สุดตามเส้นทางจากโหนด n ไปยังโหนดเป้าหมาย

$f(n) = g(n) + h(n)$ แทนค่าใช้จ่ายที่คาดว่าน้อยที่สุดตามเส้นทางจากโหนดเริ่มต้นไปยังโหนดเป้าหมายโดยผ่านโหนด n

เนื่องจากในระหว่างการสืบค้น เรายังไม่ทราบค่าใช้จ่ายที่แท้จริงของเส้นทางการค้นหาแต่ละเส้นทาง จึงจำเป็นต้องทำการประมาณค่าใช้จ่าย เพื่อให้อัลกอริทึมสามารถทำงานต่อไปได้ โดยให้

$g^*(n)$ = ค่าใช้จ่ายประมาณที่น้อยที่สุดตามเส้นทางจากโหนดเริ่มต้นไปยังโหนด n

$h^*(n)$ = ค่าประมาณของ $h(n)$ ที่ได้จากข้อมูลหรือความรู้ที่เป็นฮิวริสติค

$f^*(n) = g^*(n) + h^*(n)$ เป็นค่าใช้จ่ายโดยประมาณที่น้อยที่สุด ตามเส้นทางจากโหนดเริ่มต้นไปยังโหนดเป้าหมายโดยผ่านโหนด n

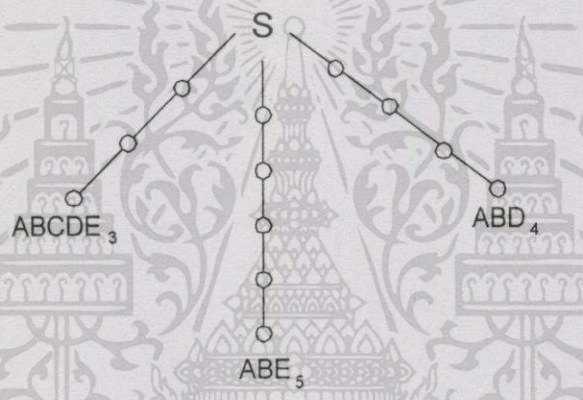
3.1.5.2 การสืบค้นเอกภพของเอกสารโดยใช้ฮิวริสติค

การสืบค้นเอกสารปกติเรามักใช้คำถามสืบค้น ทำการสืบค้นกลุ่มของเอกสารที่มีดัชนี เป็นคำเฉพาะแทนเอกสารแต่ละฉบับ อัลกอริทึม A* ช่วยให้ผู้สืบค้นสืบค้นเอกสารได้ดีขึ้น โดยมีการปรับแต่งค่าความน่าจะเป็นของเอกสารที่ตรงตามความต้องการอยู่อย่างต่อเนื่อง โดยการใช้ฮิวริสติคที่ผู้สืบค้นสามารถอธิบาย หรือกำหนดความต้องการด้วยคำเฉพาะที่ดีที่ใช้สำหรับสืบค้น ปัญหาในการสืบค้นเอกสาร คือถ้ามีคำถามสืบค้นจำนวน m คำ จะมีเซ็ตของเอกสารจำนวน 2^m-1 เซ็ต ที่จะสืบค้นได้ ซึ่งจำนวนมากขนาดนี้ทำให้ผู้สืบค้นไม่สามารถกำหนดการจัดกลุ่มของคำได้ ว่า

กลุ่มของคำแบบใดจึงจะทำให้การสืบค้นมีประสิทธิภาพที่สุด และเพื่ออธิบายอัลกอริทึมได้ชัดเจนขึ้น จึงขอนิยามคำศัพท์เฉพาะที่จำเป็นต้องใช้ดังนี้

นิยาม 1 : เทอมและโหนดสืบค้น

สมมติว่าคำเฉพาะที่จะใช้สำหรับสืบค้น 5 คำคือ {A,B,C,D,E} ส่วนย่อยของกลุ่มของเอกสารจะแทนด้วยโหนดสืบค้น ดังรูปที่ 3.2 S คือ เอกภพของเอกสาร (Document Space) ซึ่งแทนเอกสารทั้งหมดที่ยังไม่ได้รับการประเมินจากผู้สืบค้น วงกลมเล็กแทนการประเมินของผู้สืบค้นที่ประเมินโหนดนั้นๆ และกลุ่มของเอกสารจะแทนด้วยโหนดสืบค้น ตัวอย่างเช่น โหนดสืบค้น ABE_5 จะแทนส่วนหนึ่งของเอกสารที่ชี้โดยเทอม A,B และ E (อาจรวมถึง C และ D ด้วยแต่ไม่จำเป็น) และเป็นโหนดที่ผู้สืบค้นได้รับและผ่านการประเมินมาแล้ว 5 ครั้ง



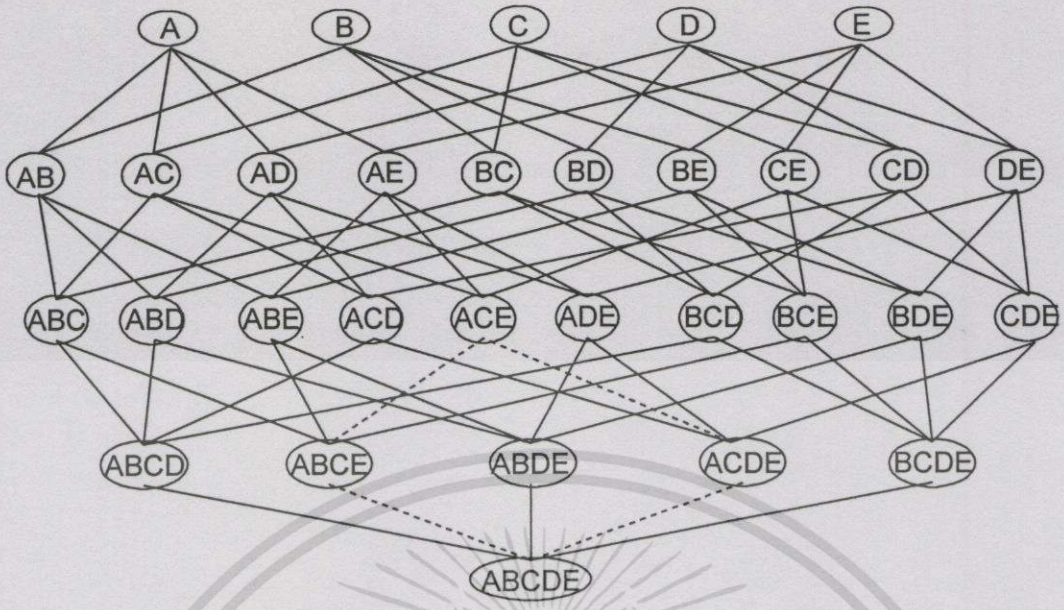
รูปที่ 3.2 สถานะในเอกภพของเอกสาร (Document state space)

นิยาม 2 : แลททิสโหนด (Lattice nodes)

แลททิสโหนด คือ กลุ่มของโหนดที่แทนเอกสารที่สร้างขึ้น ยกตัวอย่างดังรูปที่ 3.3 แต่ละโหนดแสดงถึงความลึกหรือระดับของกลุ่มตัวอย่างเอกสารที่ได้รับการประเมินว่าตรงตามความต้องการมากน้อยเพียงใด

นิยาม 3 : การสุ่มตัวอย่างแลททิสโหนด (Lattice - node - sampling)

การสุ่มตัวอย่าง คือ วิธีการได้มาซึ่งตัวอย่างที่แยกจากเอกสารที่สืบค้นและถูกนำเสนอให้ผู้สืบค้นพิจารณาว่าตรงตามคำสั่งที่สืบค้นหรือไม่ การสุ่มเลือกแลททิสโหนดจะใช้เป็นแบบใส่กลับคืน (Sampling with replacement) เอกสารใดที่ถูกเลือกเป็นครั้งที่สอง ไม่จำเป็นที่จะต้องนำเสนอให้ผู้สืบค้นดูอีก



รูปที่ 3.3 แลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets)

นิยาม 4 : การสุ่มตัวอย่างที่แท้จริง (Actual - sampling)

เนื่องจากแลททิสโหนดมีส่วนที่คาบเกี่ยวกันอยู่ ดังนั้นเอกสารที่สืบค้นได้อาจมีคุณสมบัติที่สอดคล้องกับหลายๆแลททิสโหนด ถ้ากลุ่มตัวอย่างแลททิสโหนด เป็นโหนด X และเอกสาร doc ได้รับการสืบค้นมาแล้ว ดังนั้นแลททิสโหนด Y ทั้งหมดที่ $Y \subseteq X$ และ $doc \in Y$ จะเป็นกลุ่มตัวอย่างที่แท้จริง ขอยกตัวอย่างในรูปที่ 3.3 แลททิสโหนดที่เชื่อมกันด้วยเส้นประ (ABCE, ACDE และ ABCDE) ล้วนแต่เป็นโหนดที่เป็นกลุ่มตัวอย่างที่แท้จริง ส่วนแลททิสโหนด ACE เป็นกลุ่มตัวอย่างแลททิสโหนด เอกสารที่มีคุณลักษณะตรงกับโหนด ABCDE จะถูกสืบค้นและนำเสนอ

นิยาม 5 : ฟังก์ชันปรับค่าความน่าจะเป็น (Update probability function)

เมื่อเอกสาร doc ที่มีคุณลักษณะตรงกับโหนด ABCDE ได้รับการสืบค้นแล้ว (ดังรูปที่ 3.3) ดังนั้น ค่าความน่าจะเป็นของโหนด ACE, ABCE, ACDE และ ABCDE จะต้องได้รับการปรับค่า (Update) ถ้าเอกสาร doc ตรงตามที่ต้องการและโหนด ABCE ได้รับการเลือกมาแล้ว 4 ครั้ง ซึ่งพบว่า 2 ครั้งตรงตามที่ต้องการ ดังนั้นจึงต้องปรับค่าความน่าจะเป็น $P(Ref \mid ABCE)$ จาก 2/4 เป็น 3/5 (โหนดอื่นกระทำเช่นเดียวกัน)

3.1.5.3 อัลกอริทึม IRA

IRA ย่อมาจาก Information Retrieval A*_Algorithm ต่างจากอัลกอริทึม A* ตรงที่ IRA พยายามสืบค้นเพื่อหาโหนดเป้าหมายหลายโหนด (Multiple goal nodes) จากเซ็ตของ

โหนดที่แทนเอกสารทั้งหมด เมื่อพบเอกสารที่อยู่ในข่ายของความสนใจมากกว่า 1 เอกสาร อัลกอริทึม IRA จะจัดลำดับเพื่อนำเสนอผู้สืบค้นตามค่าความน่าจะเป็นที่เอกสารตรงตามความต้องการโดยเรียงจากมากไปหาน้อย และเพื่อจำกัดจำนวนโหนดในการเลือกครั้งแรกให้เหมาะสม จากตัวอย่างแลททิสโหนดดังรูปที่ 3.3 จึงใช้ฮิวริสติคดังนี้

> *"แลททิสโหนดที่มีตัวอักษรน้อยกว่า 3 ตัว จะไม่ได้รับการพิจารณา เพราะว่าโหนดเหล่านี้มีค่าความน่าจะเป็นที่ตรงตามความต้องการน้อยกว่าโหนดที่มีตัวอักษรมากกว่า"*

อัลกอริทึม IRA แสดงได้ดังนี้

begin IRA

create_initial_search_tree;

{update : แต่ละโหนดที่มีอักษรมากกว่า 3 ตัวอักษรให้ทำการประเมิน i ครั้ง}

OPEN = [start_node];

CLOSED = [ ];

while OPEN < > [ ] **do**

begin

· n = select_doc_node (OPEN);

· {ทำการสุ่มกลุ่มตัวอย่างแลททิสโหนด กับเอกสารจากแลททิสโหนดที่มีความสัมพันธ์กับ n}

· OPEN = OPEN - [n];

· CLOSED = CLOSED ∪ [n];

· ถ้า n = โหนดเป้าหมายแล้วให้ออกจากลูป;

· successor_nodes = expand_doc_node(n);

· OPEN = OPEN ∪ (successor_nodes - CLOSED);

· {ทำการปรับค่าของแต่ละโหนดจากการสุ่มตัวอย่างจริง;

· จัดรูปแบบ search tree ของเอกสารใหม่ :

· สำหรับโหนด ที่ทำการปรับค่าทั้งหมดให้เชื่อมโยงโหนดที่เกี่ยวข้องเข้ากับ search tree;

·เชื่อมโยงทุกๆ leaf nodes เข้ากับ nodes ที่เป็น predecessors ของมันพร้อมกับลบโหนดที่มี successor อยู่ใน search tree ออกจาก OPEN แล้วนำไปผนวกกับ CLOSED}

end while;

end IRA.

3.1.5.4 ฟังก์ชันฮิวริสติคสำหรับ IRA

ดังที่กล่าวมาแล้วในหัวข้อ 3.1.5.1 ถ้าเป็นเรื่องของการค้นคืนเอกสาร สามารถกำหนดความหมายของค่า $g^*(n)$ และค่า $h^*(n)$ ได้โดยให้ $g^*(n)$ แทนค่าใช้จ่าย (จำนวนครั้ง) ที่เอกสารซึ่งแทนด้วยโหนด n ถูกประเมิน $h^*(n)$ แทนค่าประมาณจำนวนเอกสารที่ต้องทำการสืบค้นจากโหนด n จนกว่าจะพบเอกสารที่ต้องการ (สมมติว่าเป็นโหนดที่ G) ซึ่งโหนด n ที่มีค่า f^* ที่น้อยที่สุดจะถูกกระจายออก ถ้าโหนด n มีความลึกเท่ากับ b และมีเอกสารจำนวน a ฉบับที่ถูกสืบค้นแล้ว จะได้ฟังก์ชันฮิวริสติคสำหรับ IRA [11] ดังนี้

$$f^*(n) = b + (G - a) * (b/a) \qquad (3.2)$$

หลังจากพบโหนดที่ต้องการโหนดหนึ่งแล้ว อัลกอริทึม IRA จะหยุดทำการคำนวณในส่วนนั้น และทำการหาโหนดอื่นต่อไปจนกว่าครบจำนวนที่กำหนดไว้ในตอนแรก หรือจนกระทั่งได้จำนวนเอกสารทั้งหมดที่ต้องการสอดคล้องตามเงื่อนไขที่กำหนด

3.2 การประยุกต์ทฤษฎีโครงข่ายเบย์เพื่อการเรียงลำดับเอกสาร

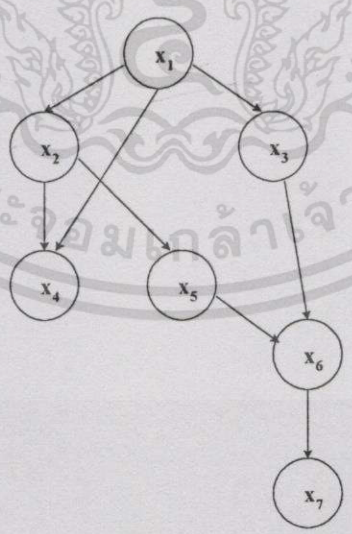
แหล่งที่มาของความน่าจะเป็นที่รู้จักกันโดยทั่วไปคือ กฎของเบย์ (Bayes' theorem) ที่มีรากฐานมาจากความจริง ของความน่าจะเป็นแบบเกี่ยวข้องกันของเหตุการณ์ 2 เหตุการณ์ ซึ่งสามารถเขียนได้เป็นผลลัพธ์ของความน่าจะเป็นของเหตุการณ์เดียว และความน่าจะเป็นแบบมีเงื่อนไขของเหตุการณ์ที่ 2 ที่ทำให้เกิดเหตุการณ์แรก [15] จากพื้นฐานของความน่าจะเป็นนี้ ได้มีการพัฒนาโครงข่ายของความรู้ หรือเรียกว่าโครงข่ายเบย์ขึ้น ต่อมาได้มีการนำมาประยุกต์เข้ากับการค้นคืนสารสนเทศในหลายแง่ สำหรับในงานวิจัยนี้ ผู้วิจัยสนใจเรื่องของการให้น้ำหนักของเทอม (Term weighting) ซึ่งมีรายละเอียดดังที่จะได้กล่าวต่อไป

3.2.1 ทฤษฎีโครงข่ายเบย์ (Bayesian Networks)

ทฤษฎีโครงข่ายเบย์นี้ได้มาจากพื้นฐานของทฤษฎีความน่าจะเป็น [15] และ การอนุมานความน่าจะเป็นแบบเบย์ (Bayesian Probabilistic Inference) [17] มีรายละเอียดดังนี้

โครงข่ายเบย์ เป็นการแทนโครงข่ายของความรู้ แสดงโดยรูปกราฟแบบมีทิศทางไม่เป็นวงวน (Directed acyclic graph) ที่มีความสัมพันธ์ระหว่างกัน ซึ่งเกิดขึ้นระหว่างความสัมพันธ์ของแต่ละส่วนของความรู้ การแทนโครงข่ายจะอธิบายระดับขั้นของความเชื่อของข้อวินิจฉัย (Proposition) และสาเหตุของความเกี่ยวข้องที่เกิดขึ้น การอนุมานจำนวนในโครงข่าย ใช้สำหรับอธิบายความน่าจะเป็นของการเกิดขึ้นและความสัมพันธ์ผ่านโครงข่ายของโหนดเพียงโหนดเดียวหรือหลายโหนด

การแทนโครงข่ายสำหรับความเกี่ยวข้องที่ไม่แน่นอน เราจะต้องแทนความรู้ที่ไม่แน่นอน ที่สัมพันธ์กับกลุ่มของตัวแปร $x_1,\ldots,x_n$ โดยการทำการกระจายร่วม (Joint distribution) $P(x_1,\ldots,x_n)$ และต้องใช้รูปแบบของการกระจายทั้งหมด 2^n รูปแบบ สามารถอธิบายปัญหาโดยใช้โหนดในเครือข่ายที่แทนตัวแปร x_i ซึ่งเป็นตัวแปรสุ่ม (กลุ่มของ Mutually exclusive และ Collectively exhaustive proposition) เชื่อมกับเส้นที่แทนสาเหตุหรือความน่าจะเป็นที่ขึ้นต่อกันระหว่างโหนดและพ่อ/แม่ (Parent) ของโหนดนั้น หรือความเกี่ยวข้องกันระหว่างโหนด การแสดงความสัมพันธ์จะเป็นค่าของความน่าจะเป็นแบบมีเงื่อนไขของตัวแปรแต่ละตัว ซึ่งตัวอย่างความสัมพันธ์ระหว่างตัวแปร $x_1,\ldots,x_n$ (ในที่นี้ $n = 7$) ได้แสดงไว้ดังรูปที่ 3.4



รูปที่ 3.4 ตัวอย่างของโครงข่ายเบย์ที่แสดงความสัมพันธ์ระหว่างตัวแปร $x_1,\ldots,x_7$

จากรูปที่ 3.4 สามารถเขียนความน่าจะเป็นแบบร่วม (Joint probability) : $P(x_1,\ldots,x_n)$ โดยมีผลของความน่าจะเป็นแบบมีเงื่อนไขดังนี้

$$P(x_1,\ldots,x_7) = P(x_7|x_6)P(x_6|x_5,x_3)P(x_5|x_2)P(x_4|x_2,x_1)P(x_3|x_1)P(x_2|x_1)P(x_1) \quad (3.3)$$

การใช้งานของโครงข่ายเบย์คล้ายกับเทคโนโลยีระบบผู้เชี่ยวชาญ โครงข่ายเบย์จะแทนความเชื่อและความรู้เกี่ยวกับกลุ่มของเหตุการณ์ โดยให้โครงข่ายเบย์เป็นฐานความรู้สำหรับกลุ่มของเหตุการณ์และข้อพิสูจน์ (เช่นการสังเกต หรือหลักความจริง) เกี่ยวกับเหตุการณ์เฉพาะในกลุ่มเดียวกัน

ระบบค้นคืนสารสนเทศที่ใช้เครื่องคอมพิวเตอร์ส่วนใหญ่ มักไม่สามารถอ่านและเข้าใจเอกสารเหมือนที่มนุษย์ทำได้ จึงจำเป็นต้องมีข้อสรุปเกี่ยวกับเอกสารจากวิธีการคำนวณลักษณะเฉพาะของเอกสาร ซึ่งเป็นที่น่าเชื่อถือ เช่น การแสดงว่าเอกสารนั้นมีหรือไม่มีคำ หรือวลีใดบ้าง ตัวอย่างเช่น เอกสารที่เกี่ยวข้องกับหัวข้อ "Applied Heuristic Search in IR" มีความน่าจะเป็นที่จะปรากฏคำว่า "IR" แต่ไม่จำเป็นต้องปรากฏคำว่า "Heuristic" ในเอกสารนั้น และในทางกลับกัน เอกสารที่ปรากฏคำว่า "IR" ไม่จำเป็นจะต้องเกี่ยวข้องกับหัวข้อ "Applied Heuristic Search in IR" ระบบค้นคืนสารสนเทศต้องสามารถอธิบายความสัมพันธ์ที่ไม่แน่นอนเหล่านี้ได้ เพื่อทำการกำหนดว่าเอกสารมีความสัมพันธ์กับสิ่งที่ผู้สืบค้นต้องการมากน้อยเพียงใด จากการสรุปเหตุการณ์ดังกล่าวข้างต้น สามารถใช้ทฤษฎีการแสดงถึงเหตุและผล (Evidential Reasoning) โดยที่ทฤษฎีดังกล่าวคือทฤษฎีความน่าจะเป็นนั่นเอง

ดังนั้นจึงมีการประยุกต์ใช้โครงข่ายเบย์ในการแทนค่าความน่าจะเป็นและการอ้างอิงเพื่อใช้ในการค้นคืนสารสนเทศ ซึ่งข้อดีของโครงข่ายเบย์สำหรับการค้นคืนสารสนเทศ คือความสามารถในการแทนความสัมพันธ์ที่ไม่แน่นอน มีรายละเอียดดังนี้ [10]

3.2.2 การค้นคืนสารสนเทศโดยใช้โครงข่ายเบย์

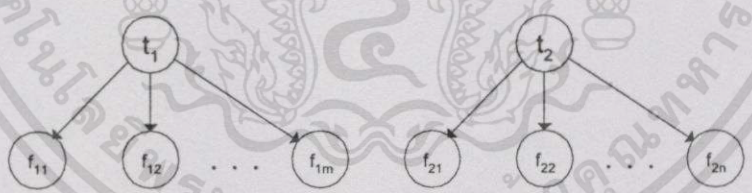
จากการประยุกต์ทฤษฎีโครงข่ายเบย์กับการค้นคืนสารสนเทศ จะได้ว่าสูตรของความน่าจะเป็นคือ การให้น้ำหนักของเทอมและการประเมินค่าของการค้นคืนสารสนเทศ ส่วนการรวมสูตรเหล่านี้เข้ากับหลักการของโครงข่ายเบย์นั้น Robert Fung และคณะ ได้ทำการปรับปรุงโดยใช้เทคนิคใหม่ เพื่อให้ผู้สืบค้นได้กำหนดหัวข้อที่สนใจและระบุลักษณะเฉพาะของเอกสารที่ชัดเจนขึ้น ทฤษฎีโครงข่ายเบย์สามารถแบ่งออกเป็นขั้นตอนใหญ่ 3 ขั้นตอนดังนี้ [10]

1. สร้างโครงข่ายที่แทนคำถาม
2. ให้ค่าของแต่ละเอกสาร
   2.1 ขยาย feature จากเอกสาร
   2.2 แทนค่า feature ในโครงข่ายการค้นคืน
   2.3 ทำการคำนวณค่าความน่าจะเป็น
3. ทำการเรียงเอกสารตามค่าความน่าจะเป็นที่ได้

ในขั้นตอนที่ 2.1 และ 3 เป็นลักษณะของ การทำงานย่อยของการค้นคืนสารสนเทศ ในขณะที่ขั้นตอนที่ 2.2 และ 2.3 เป็นลักษณะของ การทำงานย่อยของการอนุมาน ส่วนขั้นตอนแรกเป็นการทำตามการประยุกต์โครงข่ายเบย์ สำหรับการค้นคืนสารสนเทศ ซึ่งในงานวิจัยนี้ผู้วิจัยได้นำขั้นตอนที่ 3 มาใช้ในการเรียงลำดับเอกสารเท่านั้น ส่วนในขั้นตอนที่ 1 และ ขั้นตอนที่ 2 ของการค้นคืนสารสนเทศโดยใช้โครงข่ายเบย์ ผู้วิจัยได้นำเทคนิคการสืบค้นแบบฮิวริสติคมาช่วยในการค้นคืนสารสนเทศแทน ดังนั้นผู้วิจัยจึงขอกล่าวถึงเพียงหัวข้อของการเรียงเอกสารตามค่าความน่าจะเป็นเท่านั้น

3.2.3 ความน่าจะเป็นแบบดั้งเดิมในโครงข่ายของเบย์ เพื่อใช้ในการค้นคืนสารสนเทศ

จากความน่าจะเป็นแบบดั้งเดิม ได้มีการนำมาประยุกต์ใช้ในการค้นคืนสารสนเทศในแง่ของการให้น้ำหนักของเทอมเพื่อการเรียงลำดับเอกสารตามความน่าจะเป็น โดยพื้นฐานการให้น้ำหนักของเทอมในโครงข่ายเบย์ แสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 แบบจำลองโครงข่ายเบย์สองระดับแทนการค้นคืนสารสนเทศ

จากรูปที่ 3.5 ให้

t_i = หัวเรื่องที่สนใจซึ่งมีมากกว่า 1 หัวเรื่อง และ

f_{ij} = feature ของเอกสาร ซึ่งมีมากกว่า 1 ค่า

โหนด t_i แทนเหตุการณ์ที่ "เอกสารมีความเกี่ยวข้องกับหัวเรื่อง t_i" และ

โหนด f_{ij} แทนเหตุการณ์ที่ "feature f_{ij} มีความเกี่ยวข้องกับเอกสาร"

โครงข่ายในรูปที่ 3.5 มีสมมติฐานของความน่าจะเป็นดังนี้ [10] :

<u>สมมติฐานที่ 1</u> : สมมติว่าเรารู้ว่าเอกสารใดบ้างที่เกี่ยวข้องกับหัวเรื่องนั้นๆ การที่รู้ว่ามี feature ใดที่เกิดขึ้นหรือไม่เกิดขึ้นแล้วนั้น ไม่มีผลต่อความเชื่อเกี่ยวกับการเกิดขึ้นหรือไม่เกิดขึ้นของ feature ตัวอื่น

<u>สมมติฐานที่ 2</u> : เอกสารใดที่มีความเกี่ยวข้องกับหัวเรื่องหนึ่ง จะไม่มีผลกระทบต่อความเชื่อหรือความน่าจะเป็นเกี่ยวกับความเกี่ยวข้องของเอกสารกับหัวเรื่องอื่น

จากสมมติฐานที่ 1 ระบุถึงความเป็นอิสระของ feature กับหัวเรื่องที่มีให้ ซึ่งเป็นลักษณะของ Binary Independence นั่นคือ ความอิสระระหว่าง feature สามารถแทนด้วยรูปที่ 3.5 และมีข้อสรุปประการหนึ่ง ที่ได้จากโครงสร้างของโครงข่ายรูปที่ 3.5 คือจากสมมติฐานที่ 1 จะมีความไม่สมเหตุสมผลถ้าคำถามนั้นมี feature ที่มีความเกี่ยวข้องกัน เช่น คำพ้อง (Synonyms) หรือคำตรงข้าม (Antonyms) ซึ่งในความเป็นจริงแล้วมีโอกาสที่จะเกิดคำพ้องขึ้นได้

ดังนั้นโครงข่ายรูปที่ 3.5 จึงต้องกำหนดเซ็ตของความน่าจะเป็นดังนี้ :

1. Prior probability $P(t_i)$: ค่าความน่าจะเป็นที่เอกสารนั้นๆ มีความเกี่ยวข้องกับหัวเรื่อง t_i

2. Conditional probability $P(f_{ik}|t_i)$: ค่าความน่าจะเป็นของแต่ละ feature โดยมีหัวเรื่องนั้นๆ เกิดขึ้นแล้ว หมายถึงความน่าจะเป็นที่ feature f_{ij} เกิดขึ้นในเอกสาร โดยให้ข้อกำหนดว่าเอกสารนั้นเกี่ยวข้องกับหัวเรื่อง t_i

จากนั้น สามารถคำนวณค่าความน่าจะเป็นแบบ posterior $P(t_i|f_{i1},\ldots,f_{im})$ เป็นค่าความน่าจะเป็นที่เอกสารเกี่ยวข้องกับ t_i โดยให้เงื่อนไขว่าได้เกิดหรือไม่เกิด feature f_{ij} ของแต่ละเอกสารขึ้นก่อนแล้ว จึงสามารถใช้กฎของเบย์มาแทนโดยตรงได้ โดยเรียกว่า การอ้างอิงกฎของเบย์อย่างง่ายดังสมการที่ (3.4)

$$P(t_i|f_{i1},\ldots,f_{im}) = \frac{P(t_i)P(f_{i1},\ldots,f_{im}|t_i)}{P(f_{i1},\ldots,f_{im})} \qquad (3.4)$$

ถ้าจำเป็นต้องรู้ค่าของตัวเลขของความน่าจะเป็นแบบ posterior เพื่อนำมาเรียงลำดับเอกสารตามค่าความน่าจะเป็นนั้น สามารถใช้กฎของเบย์คำนวณหาตัวเลขนั้นได้ ซึ่งเรียกว่า Linear decision rule ดังสมการที่ (3.5)

$$g(t_i|f_{i1},\ldots,f_{im}) = \sum_k I(f_{ik})w(f_{ik},t_i) \qquad (3.5)$$

โดยให้

$I(f_{ik})$ = ตัวแปรอินดิเคเตอร์มีค่าเท่ากับ 1 ถ้าเกิด f_{ik} ในเอกสาร และมีค่าเป็น 0 ถ้าไม่เกิด f_{ik} ในเอกสาร

w = ค่า coefficient ของแต่ละคู่ของ feature และหัวเรื่องการเรียงลำดับของเอกสารจากค่ามากไปหาน้อยของ $g(\ )$ ซึ่งได้ผลลัพธ์เหมือนกับการเรียงเอกสารจากมากไปหาน้อยของความน่าจะเป็นแบบ posterior

จากค่า coefficient w เราสามารถแปลงเป็นน้ำหนักของแต่ละ feature หรือเทอม และในทำนองเดียวกัน ฟังก์ชัน $g(\ )$ สามารถแปลงเป็นผลรวมของน้ำหนักของ feature ที่เกิดขึ้นหรือมีอยู่ในเอกสาร วิธีการดังกล่าวเรียกว่า การให้น้ำหนักของเทอม (Term weighting)

บทที่ 4
การออกแบบระบบค้นคืนสารสนเทศแบบฮิวริสติค

จากทฤษฎีและแนวทางการประยุกต์ใช้ทั้งหมดที่ได้กล่าวไป ผู้วิจัยได้นำมาประยุกต์เข้ากับการสืบค้น ซึ่งจากที่ได้ศึกษาเรื่องที่นักวิจัย Hofferer [11] เสนอเกี่ยวกับแนวคิดการประยุกต์การสืบค้นแบบฮิวริสติคมาใช้สำหรับการค้นคืนสารสนเทศ ผู้วิจัยคิดว่าในส่วนของการเพิ่มเติมรายละเอียดเกี่ยวกับข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) เข้ามามีส่วนช่วยในการสืบค้นนับเป็นสิ่งที่ดีและน่าจะเพิ่มประสิทธิภาพของระบบค้นคืนสารสนเทศอีกทางหนึ่ง ผู้วิจัยจึงได้นำข้อมูลฮิวริสติค (Heuristic Information) มาผสมผสานกัน ทั้งส่วนของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น และคำตอบที่ผู้สืบค้นตอบคำถามที่ระบบถาม มาช่วยในส่วนของการตัดเอกสารที่ไม่ตรงตามความต้องการออก รวมทั้งนำมาช่วยในเรื่องของการให้น้ำหนักเอกสาร เพื่อการเรียงลำดับเอกสารอีกด้วย ส่วนเรื่องของการใช้ทฤษฎีโครงข่ายเบย์มาช่วยการสืบค้นนั้น ผู้วิจัยได้นำ Linear decision rule มาประยุกต์ใช้กับการเรียงลำดับเอกสาร ซึ่งผู้วิจัยได้วางโครงสร้างระบบค้นคืนสารสนเทศแบบฮิวริสติคประยุกต์กับการเรียงเอกสารโดยทฤษฎีโครงข่ายเบย์ มีลักษณะดังที่จะได้กล่าวต่อไป

4.1 ภาพรวมและองค์ประกอบของระบบค้นคืนสารสนเทศที่วิจัย

การจำลองแบบระบบค้นคืนสารสนเทศแบบฮิวริสติค ตามที่ผู้วิจัยได้ทำการวิจัยนี้ มีภาพรวมของระบบค้นคืนสารสนเทศสามารถแสดงได้ดังรูปที่ 4.1

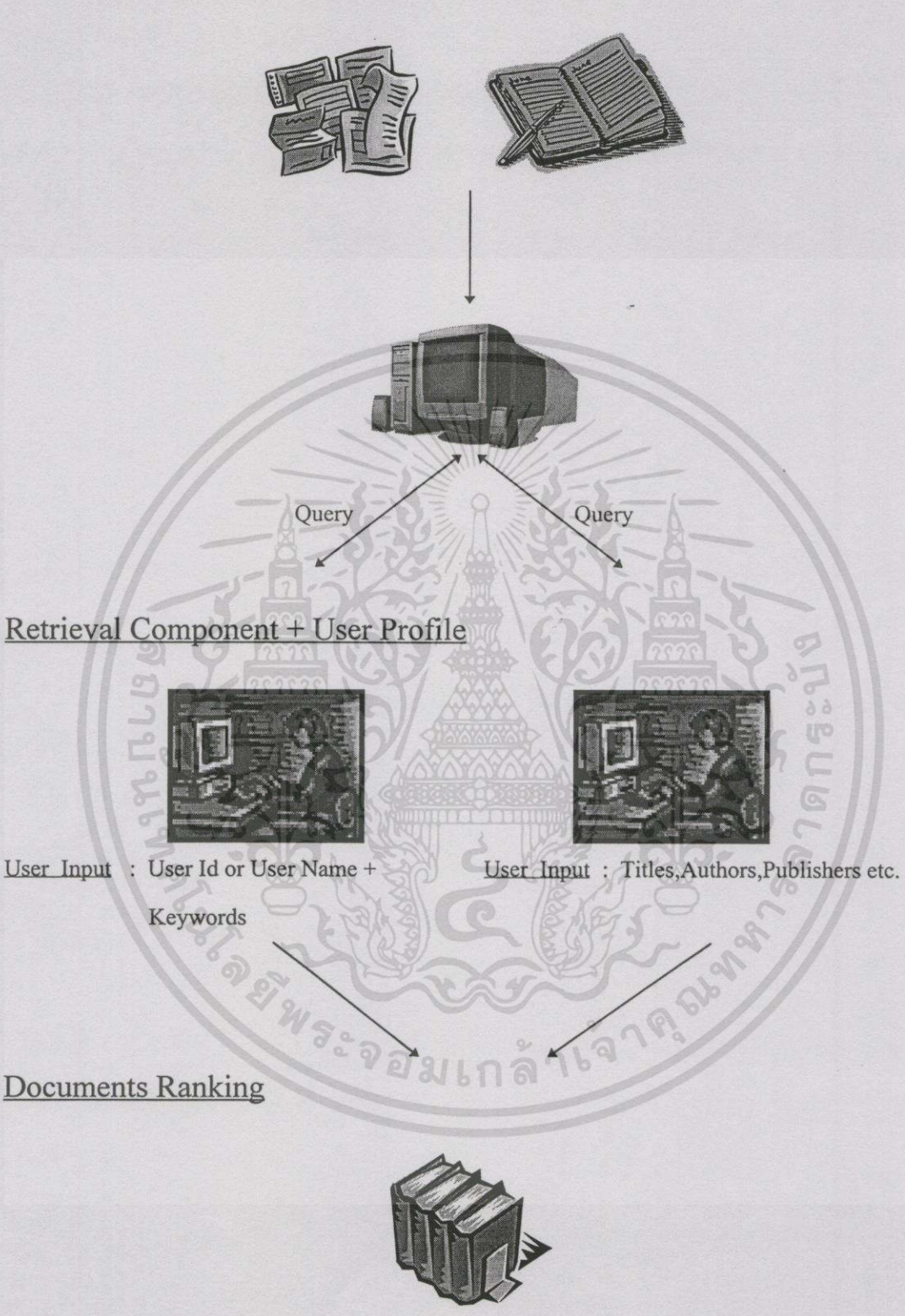
จากรูปที่ 4.1 สามารถแบ่งองค์ประกอบของระบบค้นคืนสารสนเทศออกเป็น 4 ส่วนหลักดังนี้

1. ส่วนจัดเก็บข้อมูลเอกสาร (Indexing Component)
2. ส่วนที่ช่วยในการค้นคืน (Retrieval Component)
3. ส่วนของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User profile)
4. ส่วนของการเรียงลำดับเอกสารตามความต้องการ (Documents Ranking)

รายละเอียดของแต่ละองค์ประกอบ จะได้กล่าวต่อไปในหัวข้อที่ 4.1.1 ถึงหัวข้อที่ 4.1.4

4.1.1 ส่วนจัดเก็บข้อมูลเอกสาร (Indexing Component)

ระบบค้นคืนสารสนเทศใช้ข้อมูลอธิบายเอกสาร (Document Profile) เป็นดัชนีเพื่อการสืบค้นเอกสาร โดยบรรณารักษ์เป็นผู้ป้อนข้อมูลอธิบายเอกสารให้กับระบบ



รูปที่ 4.1 ภาพรวมและองค์ประกอบของระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ได้ทำการวิจัย

ข้อมูลอธิบายเอกสาร คือการรวมฟิลด์ที่ใช้ในการสืบค้นและจัดรายชื่อของแหล่งเอกสาร รวมถึงบทคัดย่อ และหมายเหตุประกอบ เพื่อใช้อ้างอิงถึงเอกสารชิ้นนั้น เมื่อบรรณารักษ์หรือผู้ที่จัดการเอกสารได้ใส่ข้อมูลของเอกสารลงเก็บในฐานข้อมูล โดยการเติมข้อมูลลงในช่องว่าง จากนั้นระบบจึงจัดเก็บฟิลด์ทั้งหมดไว้ในตาราง (Table)

ข้อมูลอธิบายเอกสารมีรายละเอียดดังนี้

1. เลข ISBN ของเอกสาร (กรณีของหนังสือ) หรือ เลข ISSN ของเอกสาร (กรณีของวารสาร)
2. ชื่อ / หัวเรื่องของเอกสาร (Title) และหัวเรื่องย่อยของเอกสาร (Sub-Title) (ถ้ามี)
3. ชื่อผู้แต่งเอกสาร (Authors)
4. ชื่อสำนักพิมพ์ (Publisher)
5. ชื่อสาขาวิชาที่เกี่ยวข้อง (Subjects)
6. ประเทศที่พิมพ์ (Country)
7. ปีที่พิมพ์ (Year)
8. ครั้งที่พิมพ์ (Edition)
9. ชนิดของฐานข้อมูล (Database)
10. ภาษาของเอกสาร (Language)
11. จำนวนหน้า (Pages)
12. คำอธิบายเพิ่มเติม หรือ บทคัดย่อ (Description or Abstract)
13. คำเฉพาะ/คำใกล้เคียง/คำย่อ ที่ใช้ในการสืบค้นเอกสาร (Keywords/ Synonyms/ Abbreviations)

การติดต่อของระบบสำหรับผู้ที่รับผิดชอบในการใส่ข้อมูลอธิบายเอกสาร ของเอกสารแต่ละตัว จะมีรูปแบบของการติดต่อเป็นแบบกราฟฟิค เพื่อสะดวกในการใช้งานมากขึ้น

สำหรับการสืบค้น ระบบจะทำการสืบค้นโดยใช้ข้อมูลอธิบายเอกสารเข้าช่วย ซึ่งค่าหลายฟิลด์ในข้อมูลอธิบายเอกสารสามารถได้มาโดยตรงจากแหล่งของเอกสาร และถ้าต้องการแก้ไขเพิ่มเติมข้อมูลของเอกสารในระบบ ผู้ที่ต้องการแก้ไขจะต้องมีรหัสผ่าน เพื่อให้แน่ใจว่าความเปลี่ยนแปลงที่เกิดขึ้นของข้อมูลอธิบายเอกสารนั้น ถูกจัดการโดยผู้ที่ได้รับอนุญาตเท่านั้น

4.1.2 ส่วนที่ช่วยในการค้นคืน (Retrieval Component)

เป็นส่วนของการสืบค้นเอกสาร การติดต่อกับผู้สืบค้นเป็นแบบกราฟฟิค คือให้ผู้สืบค้นใส่สิ่งที่ต้องการทำการสืบค้น ระบบค้นคืนสารสนเทศจะแบ่งความต้องการของการสืบค้นของผู้สืบค้นออกเป็น 2 ลักษณะ คือ

4.1.2.1 การสืบค้นเอกสารโดยใช้คำเฉพาะ และมีการ AND OR NOT กันของคำเฉพาะ ซึ่งในส่วนของการดำเนินการ AND OR NOT นี้ระบบจะจัดการให้ผู้สืบค้นแบบอัตโนมัติ โดยการสร้างกลุ่มของคำเฉพาะที่เป็นไปได้ การสร้างกลุ่มคำดังกล่าวนี้เรียกว่า การสร้างแลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets) มีขั้นตอนการทำงานดังนี้

อัลกอริทึมในการสร้างแลททิสโหนด

สำหรับคำเฉพาะ* ตั้งแต่ 1 คำ จนถึง n คำ ให้ทำดังนี้

1. สร้างกลุ่มของ ISBN เพื่อแทนเอกสาร (ใช้ SQL Statement) ที่มีคำเฉพาะตั้งแต่ตัวที่ 1 ถึงตัวที่ n
2. สร้างโหนดใหม่ให้ NodeID = $2^{(\text{จำนวนคำเฉพาะ} - 1)}$, ค้นหากลุ่มของเอกสารที่เกี่ยวข้องกับคำเฉพาะนั้น และกำหนดว่าโหนดนั้นเกี่ยวข้องกับคำเฉพาะใด
3. รวมโหนดที่ได้จากข้อ 2 ใน Search tree

   สำหรับค่า ID ตั้งแต่ $(2^{(\text{จำนวนคำเฉพาะ} - 1)} + 1)$ ถึง $(2^{\text{จำนวนคำเฉพาะ}} - 1)$ ให้ทำดังนี้

   3.1 กำหนดค่าเลขของโหนดให้กับโหนดใดๆใน level 2 และ level ถัดไป

   3.2 หาโหนดที่เป็นพ่อ-แม่ ที่มีคำเฉพาะ = จำนวนคำเฉพาะ - 1 คำ ให้เป็น super set ของโหนดนี้

   3.3 เก็บกลุ่มของ ISBN ให้กับโหนดใหม่นี้

   3.4 กำหนดคำเฉพาะที่ปรากฏในโหนด

   3.5 รวมโหนดที่ได้ใน Search tree

ระบบจะสร้าง Search tree ให้ หลังจากที่ได้มีการใส่คำเฉพาะเพื่อทำการสืบค้นลงในระบบแล้ว เพื่อให้เห็นภาพชัดเจนขึ้นขอยกตัวอย่างการทำงานดังนี้

สมมติผู้สืบค้นใส่ คำเฉพาะ 4 คำตามลำดับก่อน/หลัง ดังนี้

1. Network	2. Computer	3. Intelligence	4. Robot

ระบบจะเริ่มจากการสร้างโหนดใน Level ที่ 1 ซึ่งมีคำสืบค้นเพียง 1 คำ ก่อน จากนั้นจับคำแต่ละคำมา AND กัน ทำเป็นลักษณะของเซ็ต ดังรูปที่ 4.2

เมื่อสร้างแลททิสโหนดได้แล้ว ระบบจะทำการสุ่มกลุ่มตัวอย่างแลททิสโหนด โดยกำหนดให้

---

* หมายถึง คำเฉพาะ (Keywords) คำใกล้เคียง (Synonyms) และคำย่อ (Abbreviations)

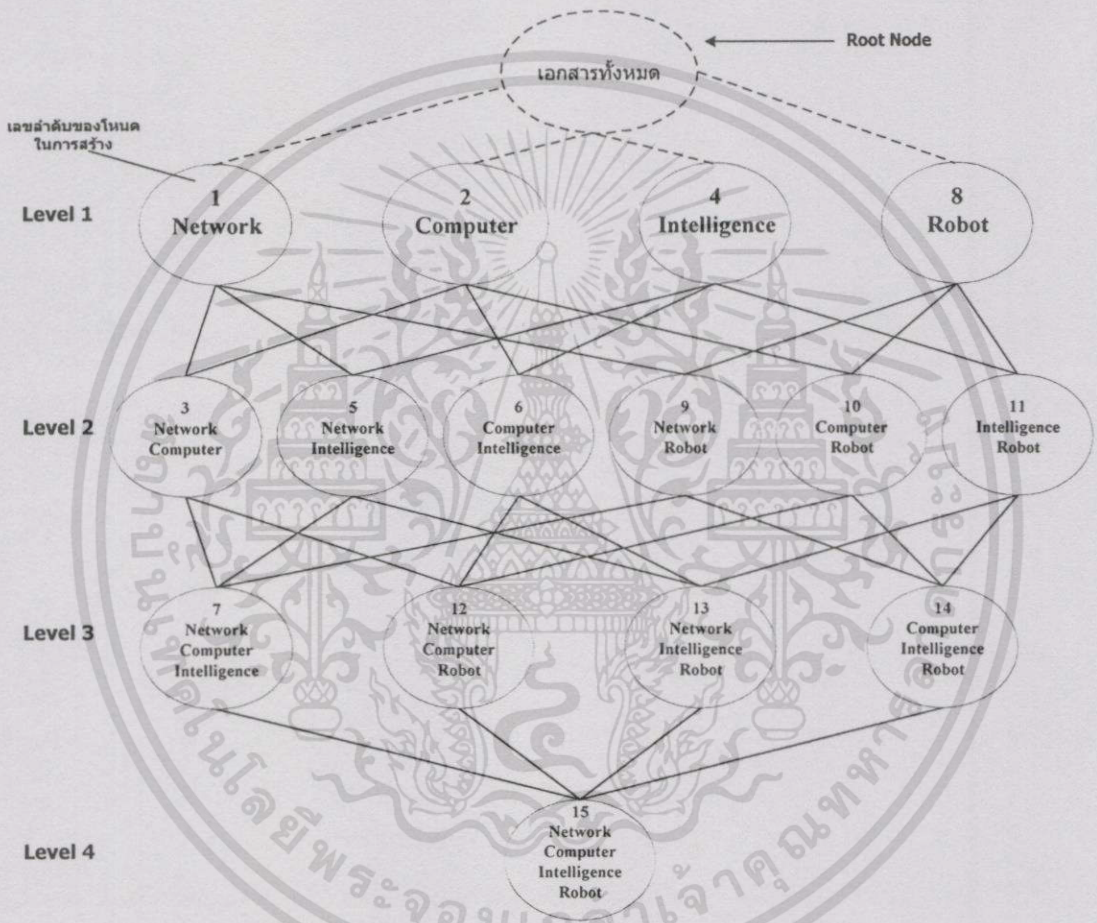
Level : ระดับความลึกของโหนด โดยกำหนดดังนี้

Level 1 : โหนดทุกโหนดที่มีคำเฉพาะ 1 คำ

Level 2 : โหนดทุกโหนดที่มีคำเฉพาะ 2 คำ

Level 3 : โหนดทุกโหนดที่มีคำเฉพาะ 3 คำ

Level 4 : โหนดทุกโหนดที่มีคำเฉพาะ 4 คำ



รูปที่ 4.2 ตัวอย่างการสร้างกลุ่มแลททิสซับเซ็ตของเอกสารของชุดคำถามสืบค้นจำนวน 4 คำได้แก่ {Network, Computer, Intelligence, Robot}

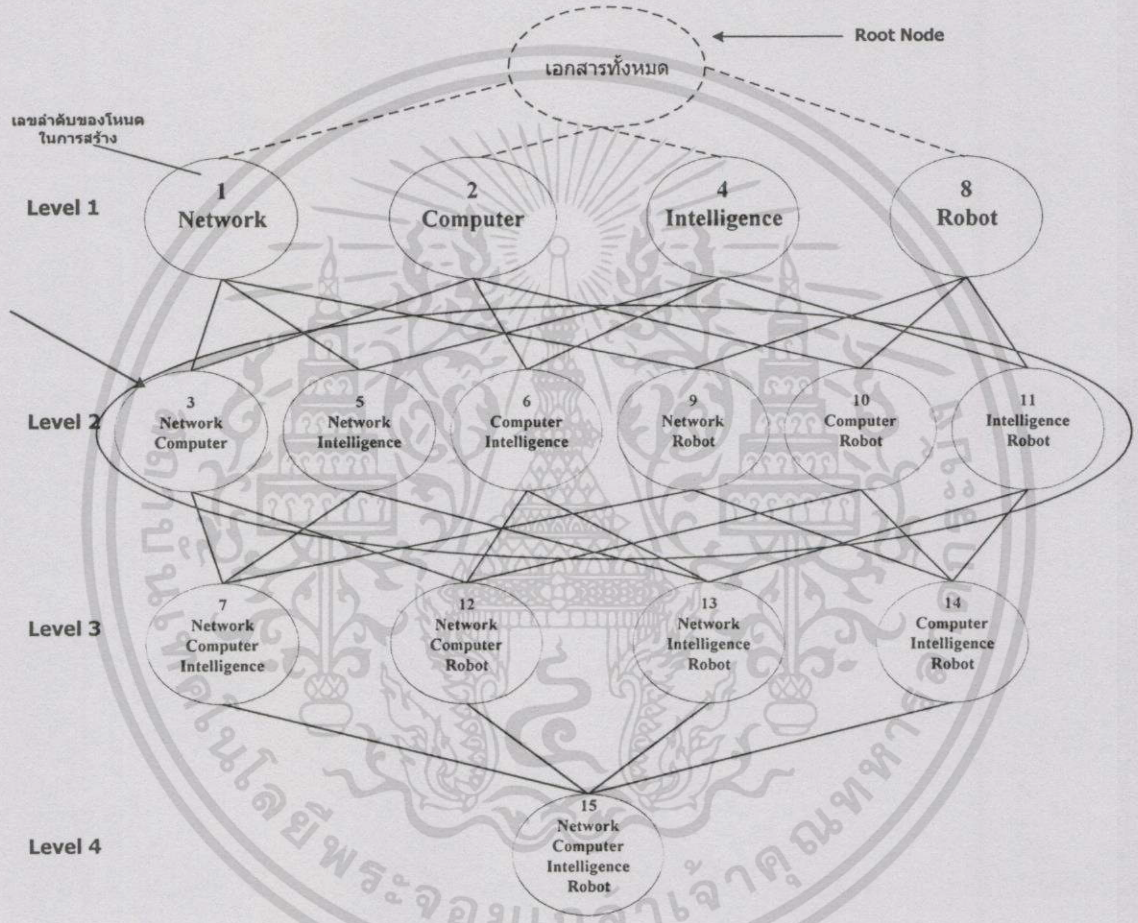
หลักเกณฑ์ในการคำนวณ Level ที่จะเริ่มสุ่มกลุ่มตัวอย่างแลททิสโหนด คือ

$$\text{Level (เริ่มต้น)} = \left\lceil \frac{\text{จำนวนคำสืบค้น}}{2} \right\rceil \qquad (4.1)$$

เพื่อจำกัดจำนวนโหนดในการเลือกครั้งแรกให้เหมาะสมจึงใช้หลักเกณฑ์ต่อไปนี้ :

แลททิสโหนดที่มีคำเฉพาะน้อยกว่าค่า Level ที่กำหนดไว้ ไม่ต้องพิจารณา เพราะว่าโหนดเหล่านั้น มีค่าความน่าจะเป็นที่ตรงตามความต้องการน้อยมาก เมื่อเทียบกับโหนดที่มีคำเฉพาะมากกว่า

ซึ่งในกรณีนี้ ระบบจะเริ่มสุ่มกลุ่มตัวอย่างจากโหนดที่มีคำเฉพาะ 2 คำ คือใน Level ที่ 2 ออกมาให้ผู้สืบค้นพิจารณา ดังรูปที่ 4.3



รูปที่ 4.3 Level ของการเริ่มสุ่มเลือกโหนด ของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Computer, Intelligence, Robot}

จากรูปที่ 4.3 ระบบจะเริ่มทำการสืบค้นเอกสารจากโหนดที่มีลูกศรชี้อยู่ ซึ่งอยู่ในส่วนที่เป็นบริเวณวงรีก่อน โดยการสืบค้นทำตามอัลกอริทึม IRA ดังที่ได้กล่าวไปแล้วในบทที่ 3 หัวข้อที่ 3.1.5.3 และอาศัยฟังก์ชัน $f^*(n)$ ในการหาทางเดินที่ดีที่สุดเพื่อไปยังโหนดเป้าหมายตามที่ได้กล่าวไว้ในบทที่ 3 หัวข้อที่ 3.1.5.4

ในส่วนนี้ผู้วิจัยได้นำฟังก์ชัน $f^*(n) = g^*(n) + h^*(n)$ มาใช้สำหรับการให้ค่าฮิวริสติคของโหนดแต่ละโหนดเพื่อนำไปสู่โหนดเป้าหมาย (กลุ่มคำสืบค้นที่เหมาะสม) โดยได้กำหนดนิยามขึ้นใหม่ดังนี้

Goal node (โหนดเป้าหมาย) = โหนดที่มีคำสืบค้นเหมาะสมกับความต้องการของผู้สืบค้น ซึ่งอาจมีโหนดเป้าหมายเดียว หรือ หลายโหนดก็ได้ (Multiple Goal node)

$g^*(n)$ = อัตราส่วนเอกสารที่ทำการสุ่มเลือกจากโหนด n กับจำนวนเอกสารทั้งหมดในโหนด n

$h^*(n)$ = ค่าความน่าจะเป็นที่เอกสารภายในโหนด n ยังไม่ตรงตามความต้องการ

$f^*(n) = g^*(n) + h^*(n)$ เป็นค่าประมาณอัตราส่วนในการสุ่มเลือกเอกสารที่คาดว่าตรงตามความต้องการ

$$f^*(n) = g^*(n) + h^*(n)$$

$$= \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right) \quad (4.2)$$

กำหนดให้

n = โหนดที่พิจารณาในขณะใดขณะหนึ่ง

$Ret\#$ = จำนวนเอกสารที่สุ่มขึ้นมาจากโหนด n

$Rel\#$ = จำนวนเอกสารที่ระบบสุ่มขึ้นมาแล้วตรงตามความต้องการของผู้สืบค้น

$Doc\#$ = จำนวนเอกสารทั้งหมดในโหนด n

โดยการสุ่มเลือกเอกสารใช้สูตรการหาขนาดของกลุ่มตัวอย่าง (Sample size = n) [27] และวิธีการคำนวณ $f^*(n)$ ผู้วิจัยได้แสดงการคำนวณเพิ่มเติมไว้ในภาคผนวก ข.

ดังนั้นตามอัลกอริทึม IRA ระบบทำการ OPEN โหนดที่อยู่ในบริเวณวงรี และ CLOSED โหนดที่อยู่ Level เหนือขึ้นไป คือ Level ที่ 1 ส่วน Level 3 กับ 4 นั้นเป็น Successor nodes ซึ่งยังไม่ถูกประเมินในตอนนี้

การเลือกโหนด ต้องเลือกโหนดที่อยู่ในเซ็ต OPEN เท่านั้น และสามารถพิจารณาตามลำดับดังนี้

1. การพิจารณาโหนดที่ถูกประเมินแล้วมีค่า $f^*(n)$: โดยเลือกโหนดที่มีค่า $f^*(n)$ น้อยที่สุด หากโหนดที่น้อยที่สุดนั้น มีค่า $f^*(n)$ เท่ากันให้เลือกโหนดที่อยู่ซ้ายสุด

2. การพิจารณาโหนดที่ถูกประเมินแล้วมีค่า $f^*(n)$ ซึ่งหาค่าไม่ได้ หรือ โหนดที่มีค่า a = 0 : ไม่ต้องพิจารณา

3. การพิจารณาโหนดที่ไม่เคยถูกประเมินมาก่อน : ให้เลือกโหนดที่ อยู่ซ้ายสุดก่อน

ในตัวอย่างนี้ เป็นการเลือกโหนดในกรณีที่ 3 ดังนั้น การสุ่มตัวอย่างจะเริ่มตั้งแต่โหนดที่ลูกศรชี้อยู่คือโหนดที่มีคำว่า Network และ Computer (โหนดหมายเลข 3) ซึ่งเป็น*กลุ่มตัวอย่างแลททิสโหนด* ส่วนโหนดที่เป็น*กลุ่มตัวอย่างที่แท้จริง* คือโหนดที่มีเส้นเชื่อมกับโหนด ที่เป็นกลุ่มตัวอย่างแลททิสโหนด (โหนดใน Level 3 และ 4 ทุกโหนด ที่มีคำว่า Network และ Computer) (ดูบทที่ 3 หัวข้อที่ 3.1.5.2 ประกอบ)

เพื่อความเข้าใจที่สอดคล้องกันในเบื้องต้น ผู้วิจัยขอแสดงการคำนวณค่า $f^*(n)$ ตัวอย่างจากรูปที่ 4.3 ดังนี้

การสุ่มตัวอย่างจากกลุ่มตัวอย่างแลททิสโหนด

$$f^*(3) = g^*(3) + h^*(3)$$

$$= \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$

$$= \frac{55}{111} + \left(1 - \frac{30}{55}\right)$$

$$= \frac{55}{111} + \frac{25}{55} = 0.95$$

หากเอกสารในโหนดที่ 3 เปรียบเทียบกับ User Profile แล้วตรง 30 ฉบับ และเอกสารดังกล่าวอยู่ในโหนดหมายเลข 7 จำนวน 15 ฉบับ อยู่ในโหนดหมายเลข 12 จำนวน 13 ฉบับ ดังนั้นสามารถคำนวณการสุ่มตัวอย่างจากกลุ่มตัวอย่างที่แท้จริง เพื่อทำการหาทางเดินไปสู่โหนดเป้าหมายโหนดต่อไปได้ดังนี้

การสุ่มตัวอย่างจากกลุ่มตัวอย่างที่แท้จริง

$$f^*(7) = g^*(7) + h^*(7)$$

$$= \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$

$$= \frac{15}{56} + \left(1 - \frac{10}{15}\right)$$

$$= \frac{15}{56} + \frac{5}{15} \qquad = 0.601$$

$$f^{\cdot}(12) = g^{\cdot}(12) + h^{\cdot}(12)$$

$$= \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$

$$= \frac{13}{72} + \left(1 - \frac{7}{13}\right)$$

$$= \frac{13}{72} + \frac{6}{13} \qquad = 0.642$$

จากนิยามที่ได้กล่าวไปแล้ว การหาทางเดินไปยังโหนดถัดไป ต้องพิจารณาโหนดที่มีค่า $f^{*}(n)$ ที่น้อยที่สุด ดังนั้นจึงเลือกโหนดหมายเลข 7 เนื่องจากมีค่า $f^{*}(n)$ น้อยกว่าโหนดหมายเลข 12

4.1.2.2 การสืบค้นเอกสารโดยใช้รายละเอียดอื่นที่อ้างถึงเอกสาร เช่น ISBN,Title, Authors,Publisher,Edition ฯลฯ ในส่วนนี้ผู้วิจัยได้ทำการพัฒนาเพิ่มเติมจากระบบค้นคืนสารสนเทศโดยทั่วไปคือผู้สืบค้นสามารถใส่สิ่งที่ต้องการสืบค้นได้หลายส่วนรวมกัน และทำการสืบค้นพร้อมกัน ตัวอย่างเช่น ต้องการสืบค้นเอกสารที่มีรายละเอียดดังนี้

Author = 'A' และ Publisher = 'B' และ Edition = 'C'

ผู้สืบค้นสามารถระบุความต้องการลงไปได้ในการสืบค้นเพียงครั้งเดียว ซึ่งแตกต่างจากระบบค้นคืนสารสนเทศที่ใช้กันอยู่ทั่วไปตรงที่ ระบบค้นคืนสารสนเทศทั่วไปมักให้เลือกสืบค้นรายละเอียดเพียงอย่างใดอย่างถึงเท่านั้น

4.1.3 ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User profile)

ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น ประกอบไปด้วย ข้อมูลส่วนตัวของผู้สืบค้นและข้อมูลเกี่ยวกับความสนใจที่ผู้สืบค้นต้องการ (รายละเอียด) ซึ่งฐานข้อมูลของระบบค้นคืนสารสนเทศ จะทำการรวบรวมข้อมูลของการสืบค้นครั้งก่อนเพื่อเพิ่มประสิทธิภาพในการค้นคืนให้กับระบบ โดยข้อมูลพื้นฐานของผู้สืบค้นนี้จะช่วยในการกำหนดว่าเอกสารใดบ้าง ที่อยู่ในกลุ่มของเอกสารที่จะถูกค้นคืน ข้อมูลเหล่านี้อาจได้มาจากบรรณารักษ์อ้างอิง หรือการสัมภาษณ์ผู้

สืบค้น ข้อมูลทั้งหมดสามารถปรับปรุงแก้ไข (Update) ได้เมื่อผู้สืบค้นต้องการ ซึ่งอาจพบข้อมูลของผู้สืบค้นที่จัดเก็บในตาราง โดยมีรายละเอียดของแต่ละฟิลด์ดังนี้

1. รหัสผู้ใช้ /ผู้สืบค้น (User ID) - ใช้ในการสืบค้นเอกสาร

2. ชื่อผู้ใช้ /ผู้สืบค้น (User Name) - ในกรณีที่ผู้สืบค้น สืบค้นข้อมูลในระบบเป็นประจำจะได้ไม่ต้องใส่ข้อมูลอีกในภายหลัง

3. ระดับการศึกษา (Educational level) - เพื่อเป็นข้อมูลให้กับระบบในการคัดเลือกเอกสาร

4. สาขาวิชาที่จบการศึกษา (Major) - เพื่อเป็นข้อมูลให้กับระบบในการคัดเลือกเอกสารให้สอดคล้องกับสาขาวิชาที่ผู้สืบค้นสำเร็จการศึกษา

5. อาชีพ (Occupation) - เพื่อเป็นข้อมูลให้กับระบบในการคัดเลือกเอกสารให้เหมาะสมกับอาชีพ

6. ประเภทเอกสารที่สนใจ (Areas of interested) - หมายถึง เรื่องหรือขอบเขต ที่ผู้สืบค้นสนใจ และต้องการสืบค้น

7. ผู้เขียนที่ชอบโดยส่วนตัว (Specific authors preference) - เพื่อเป็นข้อมูลให้กับระบบว่าผู้สืบค้นชอบวิธีการเขียนของผู้เขียนคนใดเป็นพิเศษ หากพบเอกสารของผู้เขียนที่ผู้สืบค้นระบุ ระบบจะให้ความสำคัญมากกว่าเอกสารชิ้นอื่น

8. ผู้เขียนที่ไม่ชอบโดยส่วนตัว (Non-specific authors preference) - เพื่อเป็นข้อมูลให้กับระบบว่าผู้สืบค้นไม่ชื่นชอบวิธีการเขียนของผู้เขียนคนใดเป็นพิเศษ หากพบเอกสารของผู้เขียนที่ผู้สืบค้นระบุ ระบบจะตัดออกไปโดยไม่ต้องนำมาพิจารณา

4.1.4 ส่วนของการเรียงลำดับเอกสารตามความต้องการ (Documents Ranking)

เมื่อเอกสารถูกค้นคืนออกมาจากฐานข้อมูลตามคำถามสืบค้นที่ผู้สืบค้นได้ใส่ลงไปแล้ว ขั้นตอนต่อมาคือการให้น้ำหนักของเอกสารเหล่านั้นและเรียงเอกสารตามน้ำหนักจากมากไปหาน้อย เอกสารที่มีน้ำหนักมากแสดงว่าเอกสารนั้นน่าจะตรงตามความต้องการของผู้สืบค้นมากกว่าเอกสารที่มีน้ำหนักน้อยกว่า การให้น้ำหนักเอกสารนี้ ผู้วิจัยได้ศึกษาทฤษฎีโครงข่ายเบย์ (Bayesian Networks) ซึ่งได้กล่าวไปแล้วในบทที่ 3 และนำทฤษฎีนี้มาประยุกต์ใช้ในแนวคิดดังกล่าวข้างต้น

ขั้นตอนแรก ผู้สืบค้นจะต้องระบุว่าต้องการเรียงเอกสารตามอะไร ซึ่งสามารถแยกออกเป็น 2 ลักษณะ ตามที่ได้กล่าวไปในภาคทฤษฎี คือ

1. เรียงตามฟิลด์ใดๆ ที่อยู่ใน Document Profile

2. เรียงตามน้ำหนักของคำเฉพาะที่ใช้สืบค้น

เมื่อผู้สืบค้นระบุแล้ว ระบบจะทำการเรียงเอกสารโดยอาศัย สมการ Linear dicision rule ดังนี้

$$g(t_i|f_{i1},\ldots,f_{im}) = \sum_k I(f_{ik})w(f_{ik},t_i)$$

ตัวอย่างการคำนวณหาน้ำหนักของเทอม แบบที่ 1 Term = ชนิดของฟิลด์ใน Document profile สมมติว่าผู้สืบค้นใส่คำถามสืบค้นดังนี้

Title = A | Subject = B | Author = C

Editon = 1 | Publisher = D

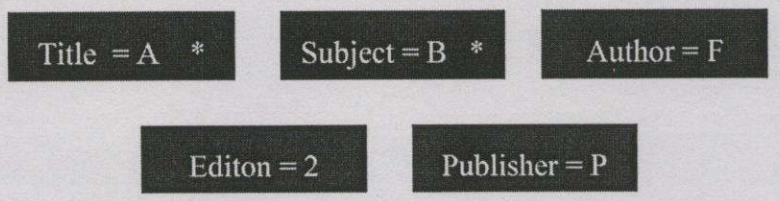
และต้องการเรียงเอกสารตาม

1. Author
2. Title
3. Publisher
4. Subject
5. Edition

จึงให้น้ำหนักดังนี้

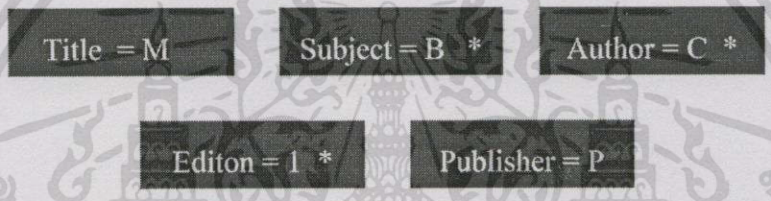
1. Author = 5
2. Title = 4
3. Publisher = 3
4. Subject = 2
5. Edition = 1

เอกสาร Doc #1 มีค่าที่เก็บในฐานข้อมูลดังนี้



หมายเหตุ * หมายถึงเอกสารในฐานข้อมูล มีข้อมูลซึ่งตรงกับข้อมูลที่ผู้สืบค้นใส่ลงไปในกระบวนการสืบค้น

เอกสาร Doc #2 มีค่าที่เก็บในฐานข้อมูลดังนี้



ทำการคำนวณค่าได้ดังนี้

$$g(t_i|f_{i1},\ldots,f_{im}) = \sum_k I(f_{ik})w(f_{ik},t_i)$$

$$g(Doc\#1|f_1,\ldots,f_5) = (1x4) + (1x2) + (0x5) + (0x1) + (0x3)$$

$$= 6$$

$$g(Doc\#2|f_1,\ldots,f_5) = (0x4) + (1x2) + (1x5) + (1x1) + (0x3)$$

$$= 8$$

แสดงว่า Doc#2 จะถูกเรียงลำดับขึ้นก่อน Doc#1 เนื่องจากมีน้ำหนักรวมมากกว่า

ซึ่งการคำนวณหาน้ำหนักของเทอม แบบที่ 2 Term = คำสืบค้นที่อยู่ในชุดของคำถามสืบค้น เช่น Query1 = {keyword1,keyword2.....} จะมีลักษณะการคำนวณที่คล้ายกัน เพียงแต่ผู้สืบค้นจะเลือกให้น้ำหนักเพื่อเรียงเอกสารตามความสำคัญของคำเฉพาะ เช่น ในคำถามสืบค้น Query1 ที่ผู้สืบค้นใส่ลงไปมีคำเฉพาะดังนี้ (ความสำคัญของคำเฉพาะเรียงตามการใส่คำนั้นก่อน/หลัง) = {Network, Communication, Digital, Image, Sound} การให้น้ำหนักจึงเป็นดังนี้

1. Network = 5
2. Communication = 4

3. Digital = 3

4. Image = 2

5. Sound = 1 เป็นต้น

วิธีการคำนวณคล้ายกับการคำนวณหาน้ำหนักของเทอม แบบที่ 1 คือ หากเอกสาร Doc#1 มีคำเฉพาะ 3 คำ เช่น {Digital, Image, Sound} และ Doc#2 มีคำเฉพาะ 2 คำ เช่น {Network, Communication} สามารถคำนวณได้ดังนี้

ค่า I {Network, Communication, Digital, Image, Sound} => Doc#1 {0,0,1,1,1}

=> Doc#2 {1,1,0,0,0}

คำนวณหาน้ำหนักรวม

$$g(t_i|f_{i1},\ldots,f_{im}) = \sum_k I(f_{ik})w(f_{ik},t_i)$$

$$g(Doc\#1|f_1,\ldots,f_5) = (0x5) + (0x4) + (1x3) + (1x2) + (1x1)$$

$$= 6$$

$$g(Doc\#2|f_1,\ldots,f_5) = (1x5) + (1x4) + (0x3) + (0x2) + (0x1)$$

$$= 9$$

แสดงว่า Doc#2 จะถูกเรียงลำดับขึ้นก่อน Doc#1 เนื่องจากมีน้ำหนักรวมมากกว่า ท้ายที่สุด ผู้สืบค้นจะได้เอกสารที่ถูกเรียงตามความต้องการออกมาตามน้ำหนักของเทอม

4.2 การออกแบบฐานข้อมูลเพื่อจัดเก็บเอกสารและคำสืบค้น

การจัดเก็บข้อมูลอธิบายเอกสารรวมถึงตาราง (Table) ต่างๆที่ต้องใช้ในการค้นคืนสารสนเทศ มีลักษณะของการจัดเก็บออกเป็น 5 Table ได้แก่ Table ชื่อ Book_Semantic, Table ชื่อ Book_Keyword, Table ชื่อ Book_Synonym, Table ชื่อ Book_Abbreviation,Table ชื่อ Book_Image และ Table ชื่อ User_Profile ซึ่ง Table แต่ละ Table มีรายละเอียดดังนี้

1. Table Book_Semantic

เปรียบเสมือน Document Profile (ข้อมูลอธิบายเอกสาร) ใช้เก็บข้อมูลเพื่ออธิบายเอกสารแต่ละฉบับ โดยอ้างถึงรายละเอียดทั้งหมดที่สามารถอ้างอิงตัวเอกสารฉบับนั้นได้ ดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 การจัดเก็บข้อมูลอธิบายเอกสาร

Table ข้อมูลอธิบายเอกสาร (Book_Semantic)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	ISBN	เลขที่เอกสาร	P.K. (Primary key)	เลข 13 หลัก เป็นเลขประจำตัวของเอกสารประเภทหนังสือ
2.	CALLNO	เลขเรียกเอกสาร		ใช้ในระบบห้องสมุด
3.	TITLE	ชื่อเอกสาร		
4.	AUTHORS	ชื่อผู้แต่ง		
5.	PUBLISHERS	ชื่อสำนักพิมพ์		
6.	BOOKYEAR	ปีที่พิมพ์		
7.	SUBJECTS	ชื่อวิชาที่เกี่ยวข้อง		
8.	EDITION	ครั้งที่พิมพ์		
9.	DBTYPE	ชนิดของฐานข้อมูล		
10.	COUNTRY	ประเทศที่พิมพ์		
11.	LANGUAGE	ภาษาของเอกสาร		
12.	PAGENO	จำนวนหน้า		
13.	DESCRIPTION	คำอธิบายเพิ่มเติม		เช่น บทคัดย่อ

2. Table Book_Keyword

เป็น table ที่ใช้จัดเก็บคำเฉพาะโดยมีเลข ISBN เป็นคีย์ ซึ่งโดยทั่วไปแล้วหนังสือ 1 เล่ม จะมีคำเฉพาะมากกว่า 1 คำ ซึ่งการจัดเก็บคำเฉพาะได้อาศัยหลักการตีความการให้คำสำคัญของบรรณารักษ์โดยอาศัยหนังสือชื่อ Subject Headings [25] (ดูภาคผนวก ง ประกอบ) รายละเอียดที่จัดเก็บนั้นแสดงดังตารางที่ 4.2

ตารางที่ 4.2 การจัดเก็บคำเฉพาะ (Keyword)

Table คำเฉพาะ (Book_Keyword)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	ISBN	เลขที่เอกสาร	P.K.	เลข 13 หลัก เป็นเลขประจำตัวของเอกสารประเภทหนังสือ
2.	KEYWORD	คำเฉพาะ	P.K.	

3. Table Book_Synonym

เป็น table ที่ใช้จัดเก็บคำใกล้เคียงที่สามารถนำไปสู่คำถามสืบค้นได้ ในกรณีที่ผู้สืบค้นไม่ได้ใส่คำที่เป็นคำเฉพาะที่ฐานข้อมูลระบบค้นคืนสารสนเทศจัดเก็บอยู่ แต่ผู้สืบค้นใส่คำที่มีความหมายใกล้เคียงลงไป ระบบจะนำคำที่มีความหมายใกล้เคียงนั้นไปหาคำเฉพาะที่เกี่ยวข้องกัน จากนั้นระบบจะนำคำเฉพาะที่ได้ไปค้นหาเอกสารที่เกี่ยวข้องเป็นขั้นตอนต่อไป รายละเอียดของการจัดเก็บคำใกล้เคียง (Synonym) แสดงได้ดังตารางที่ 4.3

ตารางที่ 4.3 การจัดเก็บคำใกล้เคียง (Synonym)

Table คำใกล้เคียง (Book_Synonym)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	KEYWORD	คำเฉพาะ	P.K.	
2.	SYNONYM	คำใกล้เคียง		
3.	DBTYPE	ชนิดของฐานข้อมูล	P.K.	

4. Table Book_Abbreviation

เป็น table ที่ใช้จัดเก็บคำย่อที่สำคัญ ใช้ในการอ้างอิงถึงเอกสารนั้นได้ หากผู้สืบค้นต้องการหาเอกสารที่มีคำย่อดังที่ตนต้องการ สามารถใส่คำย่อที่ต้องการค้นหาลงในระบบได้ ระบบมีโครงสร้างในการจัดเก็บคำย่อดังตารางที่ 4.4

ตารางที่ 4.4 การจัดเก็บคำย่อ (Abbreviation)

Table คำย่อ (Book_Abbreviation)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	DBTYPE	ชนิดของฐานข้อมูล	P.K.	
2.	ABBREVIATION	คำย่อ		เช่น AI,IR
3.	KEYWORD	คำเฉพาะ	P.K.	

5. Table Book_Image

เป็น table ที่ใช้จัดเก็บรูปภาพ หน้าปกหนังสือหรือเอกสาร รวมไปถึง สารบัญ (Contents) แสดงโครงสร้างการจัดเก็บได้ตามตารางที่ 4.5

ตารางที่ 4.5 การจัดเก็บรูปภาพ (Image)

Table รูปภาพ (Book_Image)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	ISBN	เลขที่เอกสาร	P.K.	เลข 13 หลัก เป็นเลขประจำตัวของเอกสารประเภทหนังสือ
2.	PHOTO	รูปภาพ	P.K.	

6. Table User_Profile

เป็น table ใช้สำหรับจัดเก็บข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) อธิบายถึงรายละเอียดและความชอบส่วนตัว ซึ่งสามารถอ้างอิงไปยังบุคคลนั้นได้ แสดงโครงสร้างการจัดเก็บได้ตามตารางที่ 4.6

ตารางที่ 4.6 การจัดเก็บข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile)

Table ข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User_Profile)

ลำดับ	ชื่อ Column	ความหมาย	Key	Note
1.	USER_ID	รหัสผู้ใช้/ผู้สืบค้น	P.K.	เป็นตัวเลข
2.	USER_NAME	ชื่อผู้ใช้/ผู้สืบค้น		
3.	EDUCATION	การศึกษา		1: ต่ำกว่าปริญญาตรี 2: ปริญญาตรี 3: ปริญญาโท 4: ปริญญาเอก
4.	MAJOR	สาขาวิชาที่สำเร็จการศึกษา		
5.	OCCUPATION	อาชีพ		
6.	AREA_INTERESTED	เรื่องที่สนใจ		
7.	SPEC_AUTHOR	ผู้เขียนที่ชอบเป็นพิเศษ (ถ้ามี)		
8.	NOTSPEC_AUTHOR	ผู้เขียนที่ไม่ชอบเป็นพิเศษ (ถ้ามี)		

4.3 ขั้นตอนการทำงานของระบบค้นคืนสารสนเทศและการทดลอง

เพื่อให้เป็นไปตามทฤษฎีที่ผู้วิจัยได้วิจัยมาแล้วตั้งแต่ต้น จึงได้มีการสร้างระบบค้นคืนสารสนเทศขึ้นมาเพื่อทดลองทำการสืบค้นให้เป็นไปตามทฤษฎีที่ได้ศึกษาและวิจัย ดังที่ได้กล่าวไปแล้วว่า ระบบค้นคืนสารสนเทศที่ได้ทำการวิจัยและทดลองนี้แบ่งเป็น 4 ส่วนหลัก แต่ในหัวข้อนี้ขอเน้นในเรื่องของกระบวนการสืบค้น (Retrieval Component) ว่าผู้สืบค้นต้องทำอย่างไรบ้าง และ

ระบบมีบทบาทอย่างไรในการตอบสนองหรือหาคำตอบที่ผู้สืบค้นต้องการ เพื่อการอธิบายที่ชัดเจนยิ่งขึ้น ผู้วิจัยจึงขอเสนอตัวอย่างของการสืบค้นประกอบกับการอธิบายขั้นตอนการทำงานไปพร้อมกัน

4.3.1 แผนผังการทำงานของระบบค้นคืนสารสนเทศ

ระบบจะแยกการสืบค้นออกเป็น 2 ส่วน คือสืบค้นเอกสารโดยใช้คำเฉพาะ และสืบค้นเอกสารโดยใช้รายละเอียดอื่น ดังที่ได้กล่าวไปแล้วในหัวข้อที่ 4.1.2.1 และ 4.1.2.2 ส่วนรายละเอียดของการทำงานแสดงได้ดังแผนภาพ Flow Chart รูปที่ 4.4

4.3.2 ขั้นตอนการทำงานและการทดลอง

ภายในฐานข้อมูลของระบบค้นคืนสารสนเทศที่ทำการทดลอง มีเอกสารทั้งหมด 850 ฉบับ และมีคำเฉพาะ 720 คำ ข้อกำหนดของการพัฒนาโปรแกรมสืบค้นมีดังนี้

1. ซอฟแวร์ : Delphi
2. ฮาร์ดแวร์ : CPU 333 MHz. RAM 126 MB.
3. ฐานข้อมูล : Standard Paradox
4. รูปภาพเป็นไฟล์ชนิด : Bitmap

ขั้นตอนการทำงานของระบบตามแผนผังการทำงานของระบบค้นคืนสารสนเทศในรูปที่ 4.4 มีดังนี้

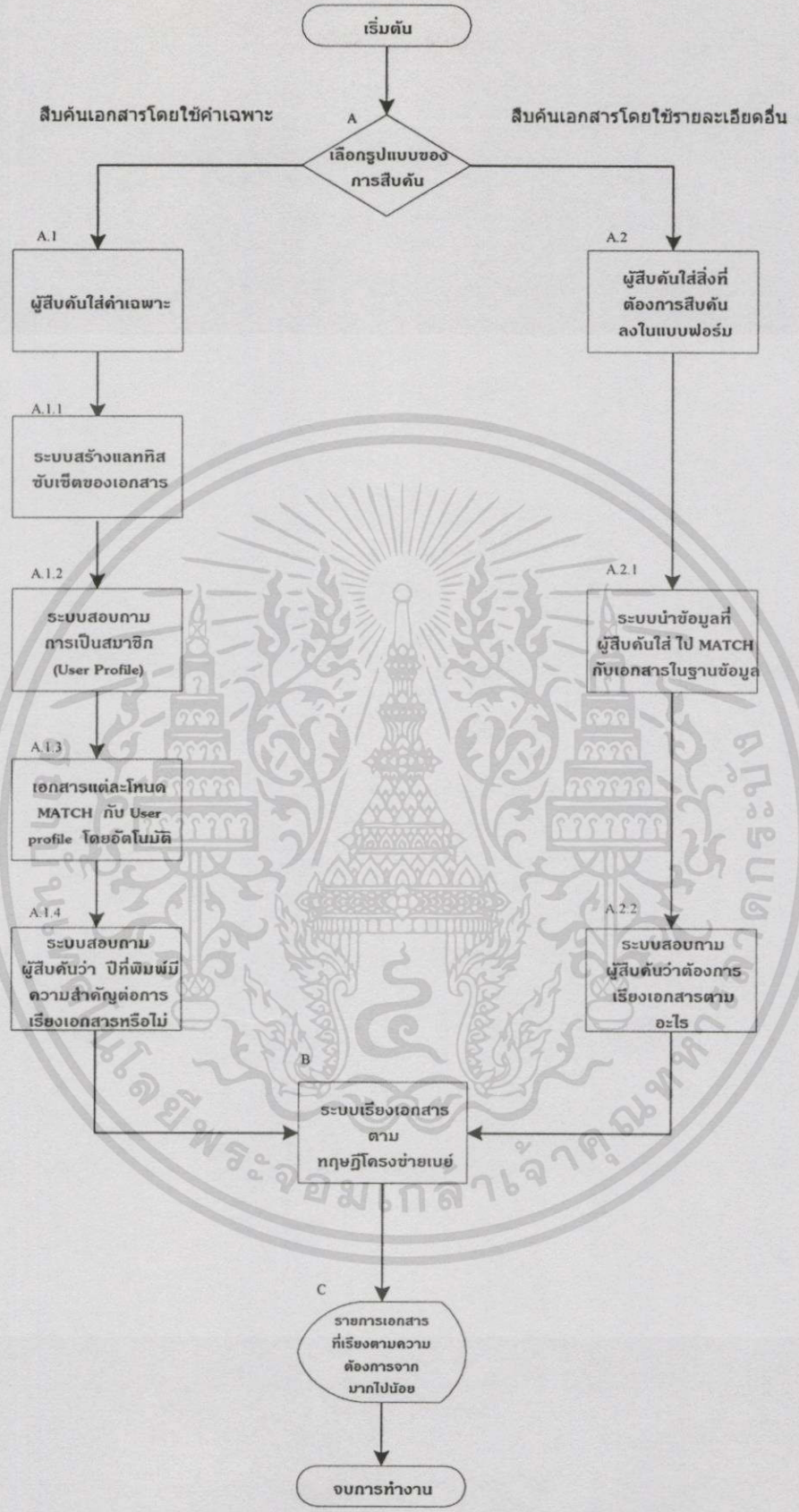
4.3.2.1 ขั้นตอนที่ A ผู้สืบค้นเลือกรูปแบบของการสืบค้น

รูปแบบของการสืบค้นมีอยู่ 2 ลักษณะ ดังที่ได้กล่าวไปแล้วในตอนต้น ซึ่งตามรูปที่ 4.4 จะเป็นขั้นตอนที่ A.1 และ A.2 ผู้วิจัยขออธิบายขั้นตอนที่ A.1 และ ขั้นตอนที่ A.2 ตามลำดับ

- ขั้นตอนที่ A.1 ผู้สืบค้นใส่คำเฉพาะ

สมมติว่าผู้สืบค้นต้องการสืบค้นโดยใส่คำเฉพาะ คำย่อ หรือคำใกล้เคียง และผู้สืบค้นใส่คำเฉพาะ 4 คำ ตามลำดับก่อน/หลังดังนี้

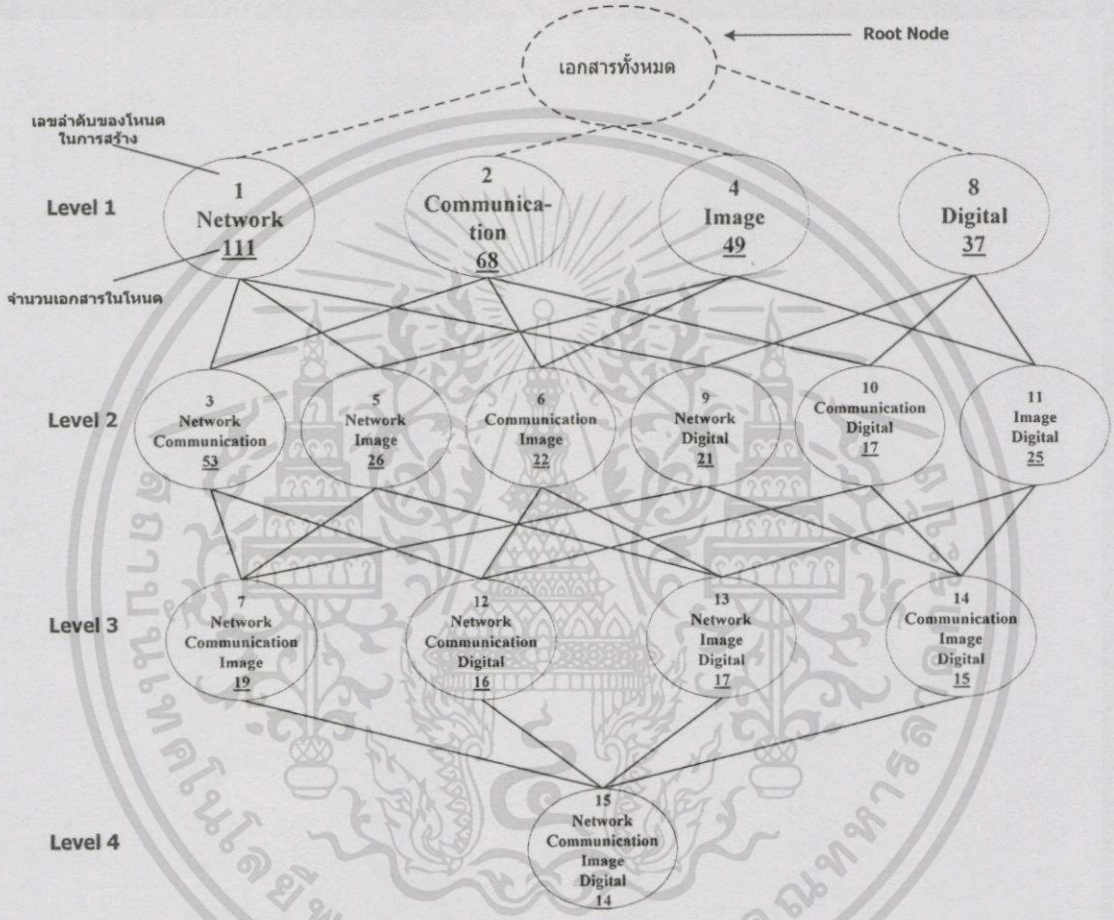
1. Network	2. Communication	3. Image	4. Digital



รูปที่ 4.4 แผนผังการทำงาน (Flow Chart) ของระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ทำการวิจัย

- ขั้นตอนที่ A.1.1 ระบบสร้างแลททิสซับเซ็ตของเอกสาร

เมื่อใส่คำเฉพาะที่ใช้สืบค้นเรียบร้อยแล้ว ระบบจะทำการสร้างกลุ่มของคำสืบค้นและกลุ่มของเอกสารที่มีคำสืบค้นเหล่านั้นอ้างถึงอยู่ เรียกว่า การสร้างแลททิสซับเซ็ตของเอกสาร (Lattice of document subsets) โดยมีอัลกอริทึมในการสร้างดังที่ได้กล่าวไปแล้วในหัวข้อที่ 4.1.2.1 ดังนั้นจึงได้ผลลัพธ์ดังรูปที่ 4.5



รูปที่ 4.5 ลำดับการสร้างแลททิสซับเซ็ตของเอกสารของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

จากนั้นระบบจึงทำการสุ่มเลือกโหนดขึ้นมา เพื่อนำมา Match กับข้อมูลใน User Profile โดยมีหลักการเลือก Level ดังที่ได้กล่าวไปแล้วในหัวข้อที่ 4.1.2.1 ดังนั้นจะได้ค่า Level ดังนี้

จากสูตร

$$\text{Level (เริ่มต้น)} = \left\lceil \frac{\text{จำนวนคำสืบค้น}}{2} \right\rceil$$

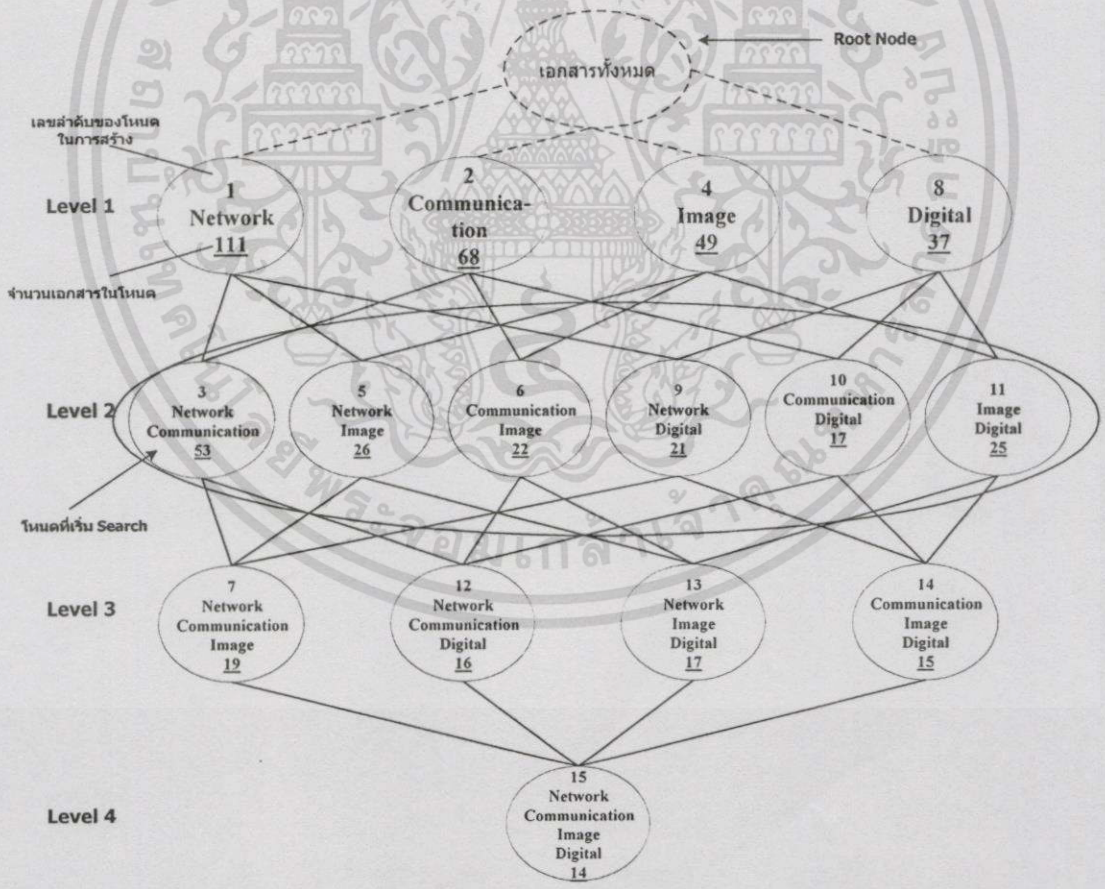
คำนวณได้ Level ดังนี้

$$\text{Level (เริ่มต้น)} = \left\lceil \frac{4}{2} \right\rceil = 2$$

ดังนั้นระบบจะทำการสืบค้นเอกสารในโหนด Level ที่ 2 คือโหนดที่อยู่บริเวณวงรี เริ่มต้นจากโหนดที่มีลูกศรชี้อยู่ ดังรูปที่ 4.6

จากนั้น ให้พิจารณาค่า $f^*(n)$ ของโหนดใน Level ที่ 2 นี้ กรณีตัวอย่างนี้เป็นการพิจารณาโหนดที่ไม่เคยถูกประเมินมาก่อน ดังนั้นให้เลือกโหนดที่อยู่ซ้ายสุดก่อน

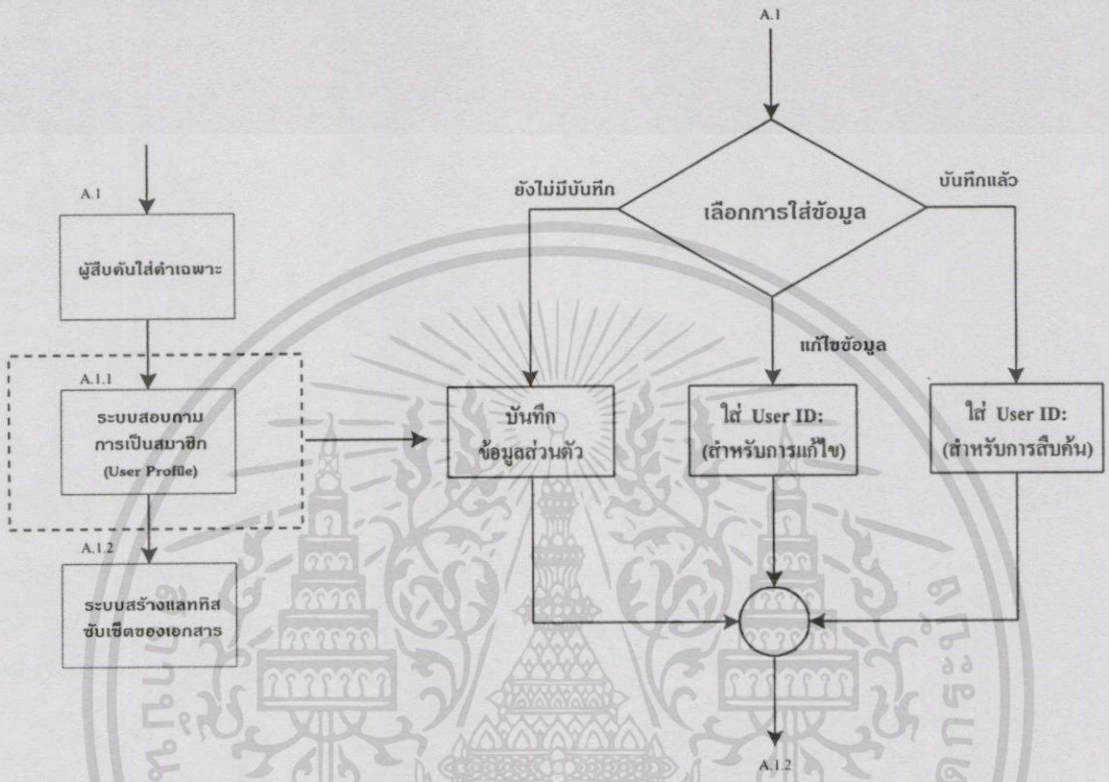
สำหรับการสืบค้น ผู้วิจัยได้ทดลองสืบค้นจากล่างขึ้นบน (Bottom up) ของ Search tree ด้วย ซึ่งการพิจารณาค่า $f^*(n)$ ทำในลักษณะเดียวกันกับการสืบค้นจากบนลงล่าง (Top down) โดยเริ่มจาก Level ที่อยู่ลึกสุด ไปยัง Level ที่ระดับเริ่มต้นของการสืบค้นจากบนลงล่าง คือ Level ที่ 2 นั่นเอง



รูปที่ 4.6 Level ของการเริ่มสุ่มเลือกโหนด ของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

- ขั้นตอนที่ A.1.2 ระบบสอบถามการเป็นสมาชิก (User Profile)

ระบบจะสอบถามการเป็นสมาชิกของผู้สืบค้น เพื่อจัดเก็บหรือแก้ไขข้อมูลในส่วนของข้อมูลประวัติของผู้ใช้/ผู้สืบค้น (User Profile) ส่วนนี้สามารถขยายเป็นขั้นตอนย่อยดัง Flow Chart ในรูปที่ 4.7



รูปที่ 4.7 รูปแบบของการสอบถามการเป็นสมาชิกของระบบค้นคืนสารสนเทศแบบฮิวริสติค

จากรูปที่ 4.7 การใส่ข้อมูลผู้สืบค้น มีตัวเลือกอยู่ 3 ประเภท คือ

1). บันทึกข้อมูลส่วนตัว : กรณีที่ยังไม่เคยมีการบันทึกข้อมูลประวัติของผู้สืบค้นมาก่อน

2). ใส่ User ID สำหรับการแก้ไข : ผู้สืบค้นต้องใส่รหัสของตนเอง เพื่อที่ระบบจะดึงข้อมูลที่ได้เคยป้อนไว้แล้วมาให้ผู้สืบค้นทำการแก้ไข

3). ใส่ User ID สำหรับการสืบค้น : ผู้สืบค้นต้องใส่รหัสของตนเอง จากนั้นระบบจะดึงข้อมูลที่ได้เคยป้อนไว้แล้วมาเปรียบเทียบกับส่วนของฐานข้อมูลเอกสาร หรือ Document profile แต่ละเรคอร์ด เพื่อคัดเลือกเอกสารที่ตรงกับคุณสมบัติของเอกสารที่ผู้สืบค้นต้องการ

สมมติว่าผู้สืบค้นเคยป้อนข้อมูลไว้แล้วมีรายละเอียดดังนี้

1. รหัสผู้ใช้ / ผู้สืบค้น : 10000002
2. ชื่อผู้ใช้ /ผู้สืบค้น : นางสาววรางคณา เงินแก้ว
3. ระดับการศึกษา : ปริญญาตรี

4. สาขาวิชาที่จบการศึกษา : Computer Science
5. อาชีพ : นักวิชาการคอมพิวเตอร์
6. ประเภทเอกสารที่สนใจ : Network
7. ผู้เขียนที่ชอบโดยส่วนตัว : -
8. ผู้เขียนที่ไม่ชอบโดยส่วนตัว : -

● ขั้นตอนที่ A.1.3 เอกสารแต่ละโหนด Match กับ User Profile โดยอัตโนมัติ ในขั้นตอนนี้ระบบทำการเปรียบเทียบเอกสารทุกเอกสารกับ User Profile ตัวที่ได้ป้อนรหัสเข้าไปในระบบ ดังรูปที่ 4.8

รูปที่ 4.8 ลักษณะของการประเมินเอกสารของระบบค้นคืนสารสนเทศแบบฮิวริสติค

เพื่อให้เห็นภาพชัดเจน เมื่อระบบทำขั้นตอนที่ A.1.2 และ A.1.3 เป็นที่เรียบร้อยแล้ว จะได้จำนวนเอกสารที่เปรียบเทียบกับ User Profile แล้วตรงกัน ในแต่ละโหนด และจะได้ค่า $f^*(n)$ เพื่อการหาทางเดินของ Search Tree ครั้งต่อไปดังตารางที่ 4.7

จากตารางที่ 4.7 แสดงผลลัพธ์ของการ Match เอกสารกับข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น และค่า $f^*(n)$ ของโหนดแต่ละโหนด โดยให้

Node No. หมายถึง หมายเลขโหนดตามลำดับการสร้างแลททิสโหนด
Keyword หมายถึง คำเฉพาะในโหนด
Doc# หมายถึง จำนวนเอกสารในโหนด
Doc# Match หมายถึง จำนวนเอกสารในโหนดที่ Match กับ User Profile
Rel หมายถึง จำนวนเอกสารในโหนดที่ Match กับ User Profile แล้วตรงกัน

โดยโหนดที่มี $f^*(n)$ = Close หมายถึง โหนดนั้นไม่ใช่โหนดเป้าหมาย ส่วนโหนดที่มี $f^*(n)$ = G(Gold) แสดงว่าโหนดนั้นเป็นโหนดเป้าหมาย และมีได้หลายโหนด

ผู้วิจัยขอยกตัวอย่างการคำนวณค่า ฟังก์ชันฮิวริสติค $f^*(n)$ ที่ได้จากตารางที่ 4.7 โดยยกตัวอย่างการคำนวณของโหนดที่ 3 และโหนดที่ 5 ดังนี้

ตารางที่ 4.7 ผลลัพธ์ของการ Match เอกสารกับข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) และค่า $f^*(n)$ ในแต่ละโหนด

Node No.	Keywords	Doc#	Doc# Match	Rel	$f^*(n)$
1.	Network	111	0	0	Close
2.	Communication	68	0	0	Close
3.	Network + Communication	53	53	32	0.95 G
4.	Image	49	0	0	Close
5.	Network + Image	26	26	15	0.77 G
6.	Communication + Image	22	22	12	0.55 G
7.	Network + Communication + Image	19	19	11	0.58 G
8.	Digital	37	0	0	Close
9.	Network + Digital	21	21	10	0.48 G
10.	Communication + Digital	17	17	9	0.53 G
11.	Network + Communication + Digital	16	16	9	0.56 G
12.	Image + Digital	25	25	9	0.36 G
13.	Network + Image + Digital	17	17	9	0.53 G
14.	Communication + Image + Digital	15	15	9	0.60 G
15.	Network + Communication + Image + Digital	14	14	9	0.64 G

โหนดที่ 3 : Network + Communication สามารถคำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(n) = \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$

$$= 40/53 + (1\text{-}32/40) = 0.95$$

โหนดที่ 5 : Network + Image สามารถคำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(n) = \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$

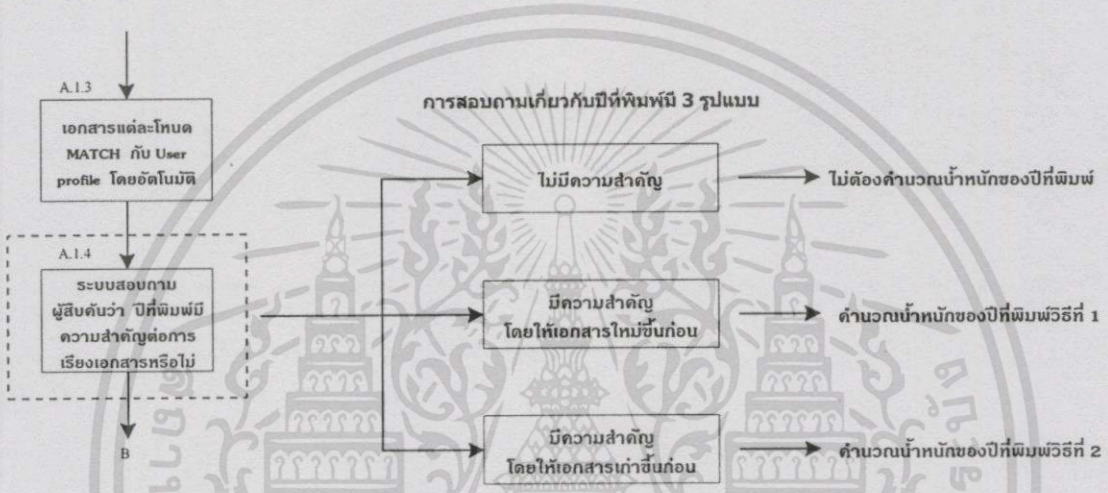
$$= 17/26 + (1\text{-}15/17) = 0.77$$

● ขั้นตอนที่ A.1.4 ระบบสอบถามผู้สืบค้นว่าปีที่พิมพ์มีความสำคัญต่อการเรียงเอกสารหรือไม่ โดยมีตัวเลือก 3 รูปแบบ ดังรูปที่ 4.9

1). ไม่มีความสำคัญ หมายถึง ไม่ต้องคำนวณน้ำหนักของปี ให้เรียงเอกสารตามน้ำหนักของคำสืบค้นที่ใส่เท่านั้น

2). มีความสำคัญ โดยต้องการให้เอกสารใหม่ขึ้นก่อน ให้เรียงเอกสารตามน้ำหนักของคำสืบค้น + น้ำหนักที่ได้จากการคำนวณความสำคัญของปี

กรณีนี้ต้องคำนวณน้ำหนักของปี โดยมีสูตรที่ใช้ในการคำนวณดังนี้



รูปที่ 4.9 การสอบถามผู้สืบค้นเรื่องการเรียงเอกสารตามปีที่พิมพ์ (พศ. หรือ คศ.) โดยมีตัวเลือก 3 รูปแบบ

<u>วิธีที่ 1</u>

$$W_{Year} = \text{ผลต่างของ } [Y_{Present} - Y_{BookLast}] - [Y_{Present} - Y_{Book} - 1] \qquad (4.3)$$

โดยที่

W_{Year}	หมายถึง	น้ำหนักของปี
$Y_{Present}$	หมายถึง	ปีที่พิมพ์ (พศ. หรือ คศ.) ปัจจุบัน
$Y_{BookLast}$	หมายถึง	ปีที่พิมพ์ (พศ. หรือ คศ.) ของเอกสารเล่มที่น้อยที่สุดในลิสต์ของผลลัพธ์
Y_{Book}	หมายถึง	ปีที่พิมพ์ (พศ. หรือ คศ.) ของเอกสาร

วิธีการคำนวณ

ให้ ปีปัจจุบัน = 2000 (กรณี คศ.)

ปีที่พิมพ์ของเอกสารฉบับน้อยที่สุด = 1995

ผลต่างระหว่างปีปัจจุบัน = 2000-1995 = 5

ดังนั้น

W_{Year} ของเอกสารพิมพ์ปีคศ. 1999 = 5- (2000-1999 -1)

= 5

W_{Year} ของเอกสารพิมพ์ปีคศ. 1997 = 5- (2000-1997 -1)

= 3

W_{Year} ของเอกสารพิมพ์ปีคศ. 1995 = 5- (2000-1995 -1)

= 1

3). มีความสำคัญ โดยต้องการให้เอกสารเก่าขึ้นก่อน

กรณีนี้ต้องคำนวณน้ำหนักของปี โดยมีสูตรที่ใช้ในการคำนวณดังนี้

วิธีที่ 2

$$W_{Year} = \left[Y_{Present} - Y_{Book}\right] \qquad (4.4)$$

วิธีการคำนวณ

นิยามเหมือนวิธีที่ 1 สามารถคำนวณได้ดังนี้

W_{Year} ของเอกสารพิมพ์ปีคศ. 1999 = 2000-1999 = 1

W_{Year} ของเอกสารพิมพ์ปีคศ. 1997 = 2000-1997 = 3

W_{Year} ของเอกสารพิมพ์ปีคศ. 1995 = 2000-1995 = 5

ผลลัพธ์จะเป็นรายการเอกสารที่เรียงตามน้ำหนักที่คำนวณได้จาก น้ำหนักที่ได้จากเทอมสืบค้น (คำเฉพาะ หรือคำใกล้เคียง หรือคำย่อ) + น้ำหนักที่ได้จากปีที่พิมพ์ของเอกสาร ซึ่งอยู่ในตารางที่ 5.1 ในบทที่ 5

ในอีกกรณี หากผู้สืบค้นต้องการสืบค้นโดยใช้รายละเอียดอื่นที่สามารถอ้างอิงถึงตัวเอกสารที่ต้องการได้ นอกเหนือจากคำเฉพาะ ผู้สืบค้นสามารถเลือกกระทำได้ในขั้นตอนที่ A.2 (ดูรูปที่ 4.4 ประกอบ)

- ขั้นตอนที่ A.2 ผู้สืบค้นใส่สิ่งที่ต้องการสืบค้นลงในแบบฟอร์ม (ดูรูปที่ ก.15 ในภาคผนวก ก. ประกอบ)

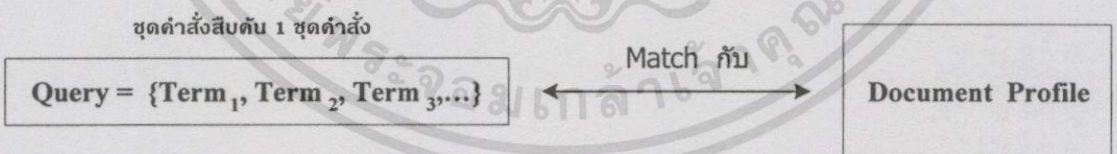
แบบฟอร์มดังกล่าวจะให้ผู้สืบค้นระบุสิ่งที่ต้องการสืบค้นจากระบบค้นคืนสารสนเทศ เช่น ชื่อหนังสือ/วารสาร ชื่อผู้แต่ง ชื่อสำนักพิมพ์ เป็นต้น (รายละเอียด ดูในหัวข้อที่ 4.1.1) ซึ่งผู้สืบค้นสามารถใส่รายละเอียดที่ทราบได้มากที่สุดเท่าที่ต้องการลงในช่องว่าง

สมมติว่าผู้สืบค้นต้องการสืบค้นเอกสารที่มีชื่อเรื่องเกี่ยวกับ "Database" และพิมพ์ปีคศ. "1997"

ชื่อหนังสือ	=	"Database"
ปีที่พิมพ์	=	"1997"

- ขั้นตอนที่ A.2.1 ระบบนำข้อมูลที่ผู้สืบค้นใส่ไป Match กับเอกสารในฐานข้อมูล

การจับคู่กระทำโดยวิธีการจับคู่ทุกๆ รายละเอียดของเอกสารกับคำสั่งสืบค้น 1 ชุดคำสั่ง ดังรูปที่ 4.10



รูปที่ 4.10 ลักษณะการทำงานของระบบค้นคืนสารสนเทศแบบฮิวริสติค ในการจับคู่ระหว่างเทอมที่ใช้สืบค้นกับข้อมูลอธิบายเอกสาร

ดังนั้นระบบจึงค้นคืนเอกสาร ที่ชื่อเรื่องหรือชื่อเอกสารมีคำว่า "Database" และพิมพ์ปีคศ. "1997" นอกจากนั้นระบบยังสามารถค้นคืนเอกสารที่มีชื่อเรื่องว่า "Database" แต่พิมพ์ในปีอื่นๆ มาให้ผู้สืบค้นพิจารณาด้วย

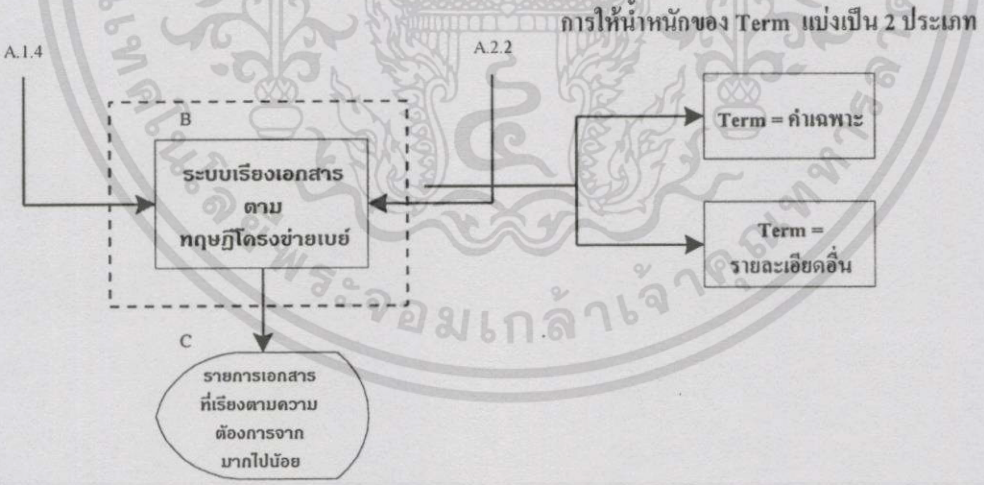
- ขั้นตอนที่ A.2.2 ระบบสอบถามผู้สืบค้นว่าต้องการเรียงเอกสารตามอะไร

ในขั้นตอนนี้ระบบถามคำถามผู้สืบค้นเรื่องของการจัดเรียงเอกสารว่า ผู้สืบค้นต้องการเรียงเอกสารตามอะไร หรือฟิลด์ใด ดังต่อไปนี้

1. ชื่อเอกสาร (ในกรณีนี้เป็นหนังสือ)
2. ชื่อสำนักพิมพ์
3. ปีที่พิมพ์
4. ครั้งที่พิมพ์
5. ชื่อผู้แต่ง
6. ชื่อสาขาวิชาที่เกี่ยวข้อง

4.3.2.2 ขั้นตอนที่ B ระบบเรียงเอกสารตามทฤษฎีโครงข่ายเบย์

จากนั้นระบบจะเรียงเอกสารตามทฤษฎีโครงข่ายเบย์ ที่ได้เสนอไปแล้วในบทที่ 3 หัวข้อที่ 3.2 โดยไม่ว่าผู้สืบค้นจะเลือกรูปแบบการสืบค้นเป็นแบบสืบค้นโดยใช้คำเฉพาะ หรือสืบค้นโดยใช้รายละเอียดอื่นก็ตาม ระบบจะทำการเรียงลำดับเอกสารตามความต้องการของผู้สืบค้นโดยอาศัยทฤษฎีโครงข่ายเบย์เช่นเดียวกัน หากดูจาก Flow Chart ของระบบค้นคืนสารสนเทศ ในรูปที่ 4.11 จะเป็นขั้นตอนที่ B มีรายละเอียดของการเรียงลำดับเอกสารดังนี้



รูปที่ 4.11 การเรียงเอกสารตามทฤษฎีโครงข่ายเบย์ โดยให้น้ำหนักของเทอมแบ่งเป็น 2 ประเภท

การให้น้ำหนักของเทอมจะแบ่งเป็น 2 ประเภท ดังนี้

1). เรียงตามฟิลด์ใดๆ ที่อยู่ใน Document Profile

ผู้สืบค้นสามารถเลื่อนตำแหน่งของฟิลด์ที่ต้องการเรียงลำดับได้ โดยฟิลด์ที่ใช้เรียงลำดับอยู่ในขั้นตอนที่ A.2.2 ซึ่งฟิลด์ที่อยู่ตำแหน่งบนสุดระบบจะให้น้ำหนักสูงที่สุด ลดหลั่นกันลงมา ดังนี้

สมมติผู้สืบค้นต้องการเรียงเอกสารโดยเรียงตามชื่อหนังสือก่อน แล้วตามด้วยชื่อสำนักพิมพ์ และรายละเอียดอื่นตามลำดับ ระบบจะทำการกำหนดน้ำหนักดังนี้

1. ชื่อหนังสือ w = 6
2. ชื่อสำนักพิมพ์ w = 5
3. ปีที่พิมพ์ w = 4
4. ครั้งที่พิมพ์ w = 3
5. ชื่อผู้แต่ง w = 2
6. ชื่อสาขาวิชาที่เกี่ยวข้อง w = 1

ผู้วิจัยขอยกตัวอย่างเปรียบเทียบระหว่าง เอกสารลำดับที่ 5 และ 6 จากตารางผลลัพธ์ตารางที่ 5.2 ในบทที่ 5 ผลการทดลอง ดังนั้นการคำนวณน้ำหนักของเอกสารจึงเป็นดังนี้

<u>วิธีการคำนวณน้ำหนัก</u>

1. 0-13-079661-1 : DB2 Universal database certification guide

Score = 0.476

เอกสารฉบับนี้มีชื่อเอกสาร และ ปีที่พิมพ์ตรงตามความต้องการ ผู้สืบค้นต้องการเรียงเอกสารตามชื่อหนังสือ ชื่อสำนักพิมพ์ ปีที่พิมพ์ ฯลฯ ตามที่ได้เสนอไป แต่เนื่องจากผู้สืบค้นได้ระบุความต้องการเพียงชื่อเอกสาร และปีที่พิมพ์ เท่านั้น ดังนั้นคำนวณได้ดังนี้

เอกสารลำดับที่ 5 มีรายละเอียดเอกสารที่เก็บในฐานข้อมูลดังนี้

{ชื่อหนังสือ, ชื่อสำนักพิมพ์, ปีที่พิมพ์, ครั้งที่พิมพ์,ชื่อผู้แต่ง, ชื่อสาขาวิชาที่เกี่ยวข้อง}

= {1,0,1,0,0,0}

= {(6x1)+(5x0)+(4x1)+(3x0)+(2x0)+(1x0)}/21

= 0.476

2. 974-512-474-5 : นำทางสู่ระบบฐานข้อมูลแบบไคลเอนต์/เซิร์ฟเวอร์=Guide to client/server database

Score = 0.286

เอกสารลำดับที่ 6 มีรายละเอียดเอกสารที่เก็บในฐานข้อมูลดังนี้

{ชื่อหนังสือ, ชื่อสำนักพิมพ์, ปีที่พิมพ์, ครั้งที่พิมพ์,ชื่อผู้แต่ง, ชื่อสาขาวิชาที่เกี่ยวข้อง}

= {1,0,0,0,0,0}

= {(6x1)+(5x0)+(4x0)+(3x0)+(2x0)+(1x0)}/21*

= 0.286

หมายเหตุ 21* => 21 หมายถึงผลรวมของน้ำหนักทั้งหมด

สรุปได้ว่าเอกสารลำดับที่ 5 มีน้ำหนักรวมมากกว่าเอกสารลำดับที่ 6 จึงถูกเรียงอยู่ในอันดับที่สูงกว่า

2). เรียงตามน้ำหนักของคำเฉพาะที่ใช้สืบค้น

เนื่องจากการเรียงลำดับตามน้ำหนักของคำเฉพาะนี้ อยู่ในขั้นตอนของการสืบค้นโดยใช้คำเฉพาะ คำใกล้เคียง หรือคำย่อ ซึ่งระบบได้เพิ่มการสอบถามผู้สืบค้น ถึงการให้ความสำคัญของปีที่พิมพ์มาช่วยในการเรียงเอกสารด้วย ดังนั้นหากผู้สืบค้นต้องการเรียงเอกสารตามปีที่พิมพ์ด้วย จึงต้องกำหนดการเรียงเอกสารในวิธีนี้เป็น

$$W_{Rank} = W_{Keyword} + W_{Year} \qquad (4.5)$$

ระบบจะให้น้ำหนัก ซึ่งได้อธิบายวิธีการคำนวณแล้วในขั้นตอนที่ A.1.4 และจะใช้ Linear dicision Rule ในการคำนวณหาน้ำหนักรวมเพื่อเรียงเอกสาร รายละเอียดของสมการและตัวอย่างได้กล่าวไว้แล้วในหัวข้อที่ 4.1.4

จากขั้นตอนที่ A.1 เรื่องของการใส่คำเฉพาะ ซึ่งผู้สืบค้นใส่คำเฉพาะตามลำดับดังนี้

1. Network
2. Communication
3. Image
4. Digital

ในตัวอย่างนี้ผู้สืบค้นไม่ต้องการเรียงเอกสารตามปีที่พิมพ์ และเพื่อความเข้าใจง่ายขึ้น ผู้วิจัยขอยกตัวอย่างเปรียบเทียบระหว่าง เอกสารลำดับที่ 1 7 และ 10 จากตารางผลลัพธ์ตารางที่ 5.1 ในบทที่ 5 ผลการทดลอง ดังนั้นการคำนวณน้ำหนักของเอกสารจึงเป็นดังนี้

<u>วิธีการคำนวณน้ำหนัก</u>

1. 0-201-42270-0 :Wide area network performance and optimization : practical strategies for …

Score = 0.722 มีวิธีคำนวณดังนี้

การให้น้ำหนักคำเฉพาะ {Network, Communication, Image ,Digital} = {8,7,6,5}

ได้มาจากการที่ ระบบได้กำหนดการใส่คำสืบค้นได้ไม่เกิน 8 คำ ดังนั้น น้ำหนักของคำสืบค้นที่ใส่เข้ามาในระบบ คำแรกในที่นี้คือคำว่า Network จึงมีน้ำหนักเท่ากับ 8 และน้ำหนักจะลดลงมาตามลำดับการใส่คำสืบค้น

ถ้าเอกสารลำดับที่ 1 นี้ไม่มีคำสืบค้นใดเป็นคำเฉพาะในฐานข้อมูล ให้ค่าเป็น 0 ดังนั้นจาก Linear dicision rule ค่า I = 0 ในทางตรงข้ามถ้าเอกสารมีคำสืบค้นใดเป็นคำเฉพาะในฐานข้อมูลให้ ค่า I = 1

เอกสารลำดับที่ 1 มีคำเฉพาะที่เก็บในฐานข้อมูลดังนี้

{Network, Communication, Image ,Digital} = {1,1,1,1}

= {(8x1)+(7x1)+(6x1)+(5x1)}/36*

= 0.722

2. 0-12-691395-1 : Network design essentials

Score = 0.583

{Network, Communication, Image ,Digital} = {1,1,1,0}

= {(8x1)+(7x1)+(6x1)+(5x0)}/36*

= 0.583

3. 0-13-474321-0 : Internetworking with TCP/IP

Score = 0.556

{Network, Communication, Image ,Digital} = {1,1,0,1}

= {(8x1)+(7x1)+(6x0)+(5x1)}/36*

= 0.556

<u>หมายเหตุ</u> 36* => 36 หมายถึงผลรวมของน้ำหนักทั้งหมด

สรุปได้ว่าเอกสารลำดับที่ 1 มีน้ำหนักรวมมากกว่าเอกสารลำดับที่ 7 และ เอกสารลำดับที่ 10 จึงถูกเรียงอยู่ในอันดับที่สูงกว่า

4.3.2.3 ขั้นตอนที่ C รายการเอกสารที่เรียงตามความต้องการจากมากไปน้อย

เมื่อระบบทำการเรียงเอกสารเป็นที่เรียบร้อยแล้ว ขั้นตอนสุดท้ายคือ การที่ระบบแสดงผลลัพธ์หรือรายการเอกสาร ที่เรียงลำดับตามความต้องการของผู้สืบค้น โดยอาศัยการให้น้ำหนักของเทอม ซึ่งรายการเอกสารที่ได้ออกมา ซึ่งเป็นผลลัพธ์ของการสืบค้นมีรายละเอียดเช่น

1. เลข ISBN/ISSN ของเอกสาร
2. ชื่อหนังสือ/วารสาร
3. ชื่อผู้แต่ง
4. ชื่อสำนักพิมพ์
5. ปีที่พิมพ์ เป็นต้น

นอกจากนั้นผู้สืบค้นสามารถดูรูปภาพของหนังสือหรือวารสารที่ตนเองสนใจได้อีกด้วย

รายการเอกสารที่ได้จากการสืบค้นตามขั้นตอน A.1 และรายการเอกสารที่ได้จากการสืบค้นตามขั้นตอน A.2 ได้ผลลัพธ์ดังตารางที่ 5.1 และตารางที่ 5.2 ในบทที่ 5 ตามลำดับ

บทที่ 5
ผลการทดลองและการวัดประสิทธิภาพในการสืบค้น

ตามขั้นตอนการทดลองจำลองระบบค้นคืนสารสนเทศแบบที่ผู้วิจัยได้ศึกษาและพัฒนาขึ้น ที่ได้กล่าวไปแล้วในบทที่ 4 รวมไปถึงตัวอย่างของการทดลองบางส่วน ในบทนี้ ผู้วิจัยขอกล่าวถึงผลของการวิจัยที่ได้พัฒนาขึ้นจากการเขียนโปรแกรมพัฒนาการจำลองการสืบค้นแบบใหม่นี้ โดยมีขั้นตอนการทำงานและรายละเอียด ดังที่ได้กล่าวไปแล้วในบทที่ 4

5.1 ผลการทดลอง

เพื่อความสอดคล้องกับงานวิจัยที่ได้พัฒนาขึ้น ผู้วิจัยขอเสนอผลการทดลองตามการสืบค้นของผู้สืบค้นซึ่งแบ่งเป็น 2 ประเภท คือ

ผู้สืบค้นรายที่ 1 สืบค้นโดยใช้คำสืบค้นที่เป็นคำเฉพาะ 4 คำดังต่อไปนี้

1. Network	2. Communication	3. Image	4. Digital

จากนั้น เป็นการติดตามผลว่าระบบจัดการกับสิ่งที่เป็น Input อย่างไร ซึ่งขั้นตอนการทำงานทั้งหมด ผู้วิจัยได้กล่าวไว้แล้วในบทที่ 4 ผลลัพธ์ที่ได้จากการสืบค้นเรียงตามลำดับความต้องการของผู้สืบค้นได้ดังตารางที่ 5.1

ผู้สืบค้นรายที่ 2 สืบค้นโดยใช้รายละเอียดอื่น ต้องการเอกสารดังนี้

ชื่อหนังสือ	=	"Database"
ปีที่พิมพ์	=	"1997"

ขั้นตอนการทำงานทั้งหมด ผู้วิจัยได้กล่าวไว้แล้วในบทที่ 4 เช่นเดียวกัน ดังนั้นผลลัพธ์ที่ได้จากการสืบค้นเรียงตามลำดับความต้องการของผู้สืบค้น ได้ดังตารางที่ 5.2

ตารางที่ 5.1 ตัวอย่างรายการเอกสารผลลัพธ์ที่ได้จากการสืบค้น โดยใช้คำเฉพาะ {Network, Communication, Image ,Digital} โดยเรียงลำดับตามน้ำหนักของการตรงต่อความต้องการของผู้สืบค้น

Doc No.	ISBN	Title	Score
1.	0-201-42270-0	Wide area network performance and optimization : practical strategies for …	0.722
2.	0-534-20244-6	Understanding Data Communications and Networks	0.722
3.	0-07-021422-0	The intelligent network standards: their application to services	0.722
4.	156604-329-8	The Internet power toolkit: cutting-edge tool & techniques for power users	0.722
5.	0-07-020346-6	TCP/IP: architecture, protocols and implementation	0.722
6.	0-07-005593-9	The X series recommendations: standards for data communications	0.583
7.	0-12-691395-1	Network design essentials	0.583
8.	0-201-62745-0	Network and distributed systems management	0.583
9.	0-201-56741-5	Internet system handbook	0.583
10.	0-13-474321-0	Internetworking with TCP/IP	0.556
11.	0-201-56506-4	Data communications, computer networks and open systems	0.556
12.	1-57521-113-0	Web programming with Java	0.528
13.	1-56830-300-9	Internet publishing with Acrobat	0.528
14.	0-8053-7724-7	Local area networks	0.500
15.	0-672-48440-4	Unix networking	0.417
16.	0-471-54845-6	Business data communications: basic concepts,security, and design	0.417
17.	0-471-12365-x	Business data communications and networking	0.417
18.	0-07-072220-X	Building communication networks with distributed objects	0.417
19.	1-57521-051-7	Web publishing unleashed: HTML, CGI, SGML, VRML Java	0.389
20.	0-201-56318-5	Unix System V network programming	0.389
21.	1-56830-307-6	Web page scripting Techniques	0.361
22.	0-201-50803-6	Digital image processing	0.306

ตารางที่ 5.2 ตัวอย่างรายการเอกสารผลลัพธ์ที่ได้จากการสืบค้น โดยใช้รายละเอียดอื่น (ชื่อเอกสาร = "Database" และปีที่พิมพ์ = "1997") โดยเรียงลำดับตามน้ำหนักของการตรงต่อความต้องการของผู้สืบค้น

Doc No.	ISBN	Title	Score
1.	1-57231-342-0	Microsoft Jet database engine programmer's guide	0.476
2.	0-7600-4904-1	Database systems: design,implementation, and management	0.476
3.	0-07-044756-x	Database system concepts	0.476
4.	1-56276-530-2	Database backed web sites: the thinking person's guide to web publishing	0.476
5.	0-13-079661-1	DB2 Universal database certification guide	0.476
6.	974-512-474-5	นำทางสู่ระบบฐานข้อมูลแบบไคลเอนต์/เซิร์เวอร์=Guide to client/server database	0.286
7.	1-57169-032-8	Web database construction kit: a step-by-step guide to linking Microsoft Access	0.286
8.	1-57169-070-0	Web Database Primer Plus: everything you need to know about to make database	0.286
9.	0-256-13438-3	The science of database management	0.286
10.	0-471-14718-4	The object database handbook: how to select, implement, and use object-oriented	0.286
11.	0-201-50881-8	Relational database writings, 1985-1989	0.286
12.	0-13-771791-1	Relational database design: an introduction	0.286
13.	0-8186-5452-X	Query processing in parallel relational database systems	0.286

5.2 การวัดประสิทธิภาพของระบบค้นคืนสารสนเทศแบบฮิวริสติคที่วิจัย กับระบบค้นคืนสารสนเทศแบบทั่วไป

การวัดประสิทธิภาพระหว่างระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย กับระบบค้นคืนสารสนเทศที่ใช้หลักการสืบค้นแบบทั่วไป คือการจับคู่แบบตรงตัวนี้ ผู้วิจัยจะทำการวัดประสิทธิภาพเฉพาะในส่วนของการสืบค้นโดยใช้คำเฉพาะเท่านั้น ทั้งนี้เนื่องจากส่วนของการสืบค้นโดยใช้รายละเอียดอื่น เช่น ชื่อเรื่อง หัวเรื่อง ชื่อผู้แต่ง เป็นต้น มีลักษณะของการสืบค้นคล้ายกับการสืบค้นแบบตรงตัวอยู่แล้ว ผู้วิจัยจึงเห็นว่าไม่จำเป็นที่จะต้องวัดประสิทธิภาพอีก ดังนั้นจึงได้เน้นในเรื่องของการสืบค้นโดยใช้คำสืบค้นที่มีลักษณะเป็น คำเฉพาะ รวมถึง คำใกล้เคียง หรือคำย่อ ซึ่งการสืบค้นแบบนี้มีความแตกต่างกับระบบทั่วไปอย่างมาก

ผู้วิจัยได้ทดลองสืบค้นโดยกำหนดพารามิเตอร์เป็นจำนวนคำสืบค้น ที่ใช้สืบค้นในครั้งหนึ่งๆ จำนวนคำถามสืบค้น (Query) มีจำนวนทั้งหมด 20 ชุดคำถาม โดยมีคำสืบค้นและจำนวนชุดคำถามดังนี้

ประเภทที่ 1 คำสืบค้นจำนวน 3 คำ 6 ชุดคำถาม
ประเภทที่ 2 คำสืบค้นจำนวน 4 คำ 6 ชุดคำถาม
ประเภทที่ 3 คำสืบค้นจำนวน 5 คำ 4 ชุดคำถาม
ประเภทที่ 4 คำสืบค้นจำนวน 6 คำ 2 ชุดคำถาม
ประเภทที่ 5 คำสืบค้นจำนวน 7 คำ 2 ชุดคำถาม

เพื่อการวัดประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย ผู้วิจัยจึงได้ทำการสร้างระบบค้นคืนสารสนเทศที่ใช้หลักการสืบค้นแบบทั่วไปขึ้นอีก 1 ระบบ เพื่อทำการเปรียบเทียบระบบทั้งสอง โดยระบบค้นคืนสารสนเทศทั้ง 2 ประเภท ต้องทำการสืบค้นภายใต้คำสืบค้นชุดเดียวกัน และฐานข้อมูลตัวเดียวกัน โดยให้ผู้สืบค้นจำนวนหนึ่ง เป็นผู้ดำเนินการสืบค้น

เนื่องจากคำสืบค้นที่ใช้ทดลองมีจำนวนหลายคำ สำหรับระบบค้นคืนสารสนเทศแบบทั่วไปนั้น การระบุความสัมพันธ์หรือตัวดำเนินการระหว่างคำให้สอดคล้องกับความต้องการของผู้สืบค้น เป็นการยากที่จะกระทำได้ ดังนั้น ผู้วิจัยจึงขอกำหนดตัวดำเนินการระหว่างคำสืบค้นให้เป็น 2 ลักษณะคือ AND และ OR ทั้งหมด

ในแต่ละชุดคำถาม เมื่อผู้สืบค้นได้ทำการสืบค้นเอกสารที่ต้องการเป็นที่เรียบร้อยแล้ว ผู้วิจัยทำการบันทึกจำนวนเอกสารที่แต่ละระบบ (นั่นคือ ระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND และระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR) ค้นคืนได้ ทำการคำนวณหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และประสิทธิภาพของระบบโดยรวม (E-Measure) ตามสูตรที่ได้กล่าวไปแล้วในภาคทฤษฎี บทที่ 2 หัวข้อที่ 2.6 ได้ผลการทดลองดังตารางที่ 5.3 และ ตารางที่ 5.4 และกราฟรูปที่ 5.1 และ รูปที่ 5.2 ตามลำดับ ดังนี้

จากตารางที่ 5.3 และ ตารางที่ 5.4 Relevance หมายถึง เอกสารที่ตรงตามความต้องการของผู้สืบค้น Retrieve หมายถึง เอกสารทั้งหมดที่ระบบค้นคืนได้ RetRel หมายถึง เอกสารที่เกี่ยวข้องที่ค้นคืนได้ RetNRel หมายถึง เอกสารที่ไม่เกี่ยวข้องที่ถูกค้นคืน

ตารางที่ 5.3 เปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย กับระบบค้นคืนสารสนเทศโดยทั่วไป กรณีให้ระบบค้นคืนสารสนเทศโดยทั่วไปมีตัวดำเนินการเป็น AND

			ระบบค้นคืนสารสนเทศตามที่ได้ทำการวิจัย					ระบบค้นคืนสารสนเทศแบบทั่วไป (AND)				
Query No.	Users' queries	Relevance	Retrieve	RelRet	RetNRel	Precision	Recall	Retrieve	RelRet	RetNRel	Precision	Recall
1.	TCP/IP+digital+network	17	29	17	12	0.586	1.000	0	0	0	0.000	0.000
2.	html+internet+web	22	31	22	9	0.710	1.000	4	4	0	1.000	0.182
3.	database+sql+table	17	22	17	5	0.723	1.000	3	3	0	1.000	0.176
4.	business+management+market	7	9	7	2	0.778	1.000	2	2	0	1.000	0.286
5.	network+communication+image	39	63	31	32	0.492	0.795	17	17	0	1.000	0.436
6.	design+database+implement	15	13	11	2	0.846	0.733	3	3	0	1.000	0.200
7.	client+server+database+guide	19	29	12	17	0.414	0.632	2	2	0	1.000	0.105
8.	object-oriented+programming+ database+language	15	23	13	10	0.565	0.867	0	0	0	0.000	0.000
9.	window+os+tool+internet	10	10	8	2	0.800	0.800	0	0	0	0.000	0.000
10.	3D+design+graphic+image	13	13	9	4	0.692	0.692	1	1	0	1.000	0.077
11.	programming+data structure+ algorithm+C++	18	24	15	9	0.625	0.833	2	2	0	1.000	0.111
12.	telecommunication+network+ mobile+telephone	3	3	3	0	1.000	1.000	1	1	0	1.000	0.333

ตารางที่ 5.3 (ต่อ) เปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย กับระบบค้นคืนสารสนเทศโดยทั่วไป กรณีให้ระบบค้นคืนสารสนเทศโดยทั่วไปมีตัวดำเนินการเป็น AND

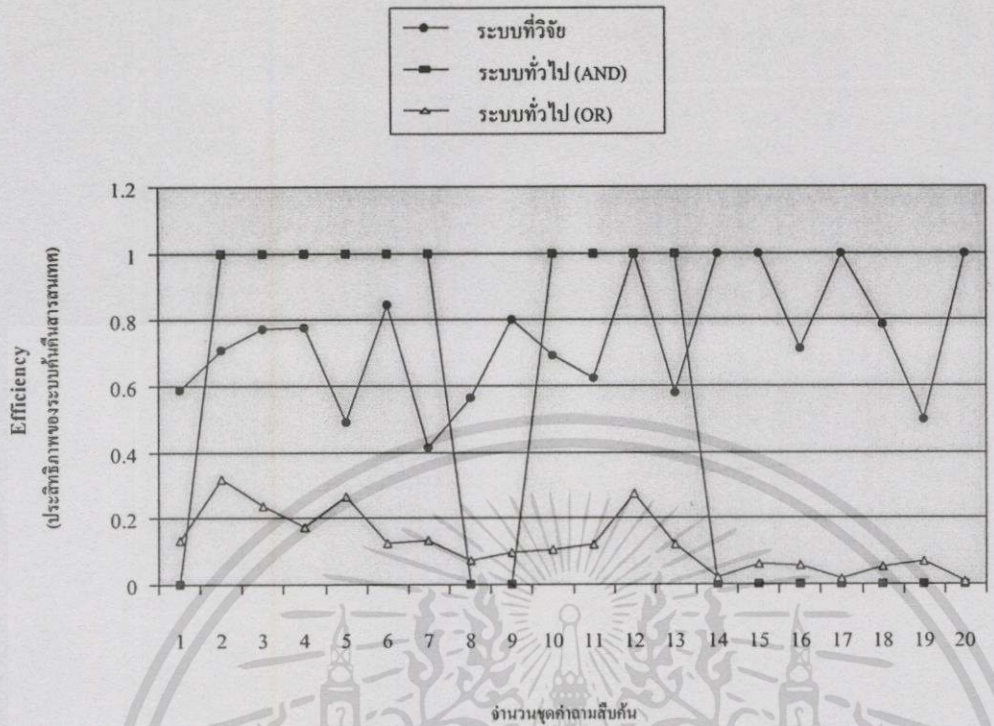
			ระบบค้นคืนสารสนเทศตามที่ได้ทำการวิจัย					ระบบค้นคืนสารสนเทศแบบทั่วไป (AND)				
Query No.	Users' queries	Relevance	Retrieve	RelRet	RetNRel	Precision	Recall	Retrieve	RelRet	RetNRel	Precision	Recall
13.	communication+image+network+ digital+internet	22	31	18	13	0.581	0.818	2	2	0	1.000	0.091
14.	programming+UNIX+OS+ network+administrator	5	2	2	0	1.000	0.400	0	0	0	0.000	0.000
15.	database+oracle+sql+guide+administrator	7	5	5	0	1.000	0.714	0	0	0	0.000	0.000
16.	internet+network+image+web+homepage	10	14	10	4	0.714	1.000	0	0	0	0.000	0.000
17.	network+communication+programming+ application+layer+protocol **	4	2	2	0	1.000	0.500	0	0	0	0.000	0.000
18.	network+communication+programming+ application+layer+protocol **	12	14	11	3	0.786	0.917	0	0	0	0.000	0.000
19.	network+communication+image+digital+ internet+web+html	13	22	11	11	0.500	0.846	0	0	0	0.000	0.000
20.	programming+network+internet+java+ application+window+interface	3	2	2	0	1.000	0.667	0	0	0	0.000	0.000

ตารางที่ 5.4 เปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย กับระบบค้นคืนสารสนเทศโดยทั่วไป กรณีให้ระบบค้นคืนสารสนเทศโดยทั่วไปมีตัวดำเนินการเป็น OR

			ระบบค้นคืนสารสนเทศตามที่ได้ทำการวิจัย					ระบบค้นคืนสารสนเทศแบบทั่วไป (OR)				
Query No.	Users' queries	Relevance	Retrieve	RelRet	RetNRel	Precision	Recall	Retrieve	RelRet	RetNRel	Precision	Recall
1.	TCP/IP+digital+network	17	29	17	12	0.586	1.000	127	17	110	0.134	1.000
2.	html+internet+web	22	31	22	9	0.710	1.000	69	22	47	0.319	1.000
3.	database+sql+table	17	22	17	5	0.723	1.000	72	17	55	0.236	1.000
4.	business+management+market	7	9	7	2	0.778	1.000	40	7	33	0.175	1.000
5.	network+communication+image	39	63	31	32	0.492	0.795	146	39	107	0.267	1.000
6.	design+database+implement	15	13	11	2	0.846	0.733	122	15	107	0.123	1.000
7.	client+server+database+guide	19	29	12	17	0.414	0.632	144	19	125	0.132	1.000
8.	object-oriented+programming+ database+language	15	23	13	10	0.565	0.867	209	15	194	0.718	1.000
9.	window+os+tool+internet	10	10	8	2	0.800	0.800	102	10	92	0.098	1.000
10.	3D+design+graphic+image	13	13	9	4	0.692	0.692	124	13	111	0.105	1.000
11.	programming+data structure+ algorithm+C++	18	24	15	9	0.625	0.833	149	18	131	0.121	1.000
12.	telecommunication+network+ mobile+telephone	3	3	3	0	1.000	1.000	11	3	8	0.273	1.000

ตารางที่ 5.4 (ต่อ) เปรียบเทียบประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย กับระบบค้นคืนสารสนเทศโดยทั่วไป กรณีให้ระบบค้นคืนสารสนเทศโดยทั่วไปมีตัวดำเนินการเป็น OR

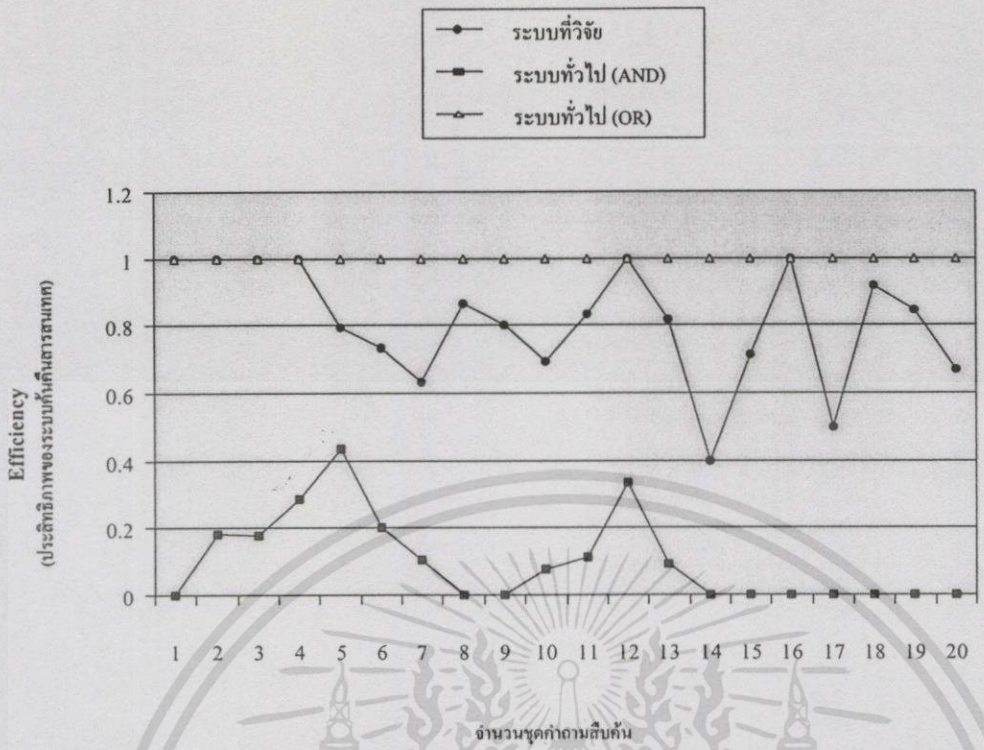
			ระบบค้นคืนสารสนเทศตามที่ได้ทำการวิจัย					ระบบค้นคืนสารสนเทศแบบทั่วไป (OR)				
Query No.	Users' queries	Relevance	Retrieve	RelRet	RetNRel	Precision	Recall	Retrieve	RelRet	RetNRel	Precision	Recall
13.	communication+image+network+ digital+internet	22	31	18	13	0.581	0.818	182	22	160	0.121	1.000
14.	programming+UNIX+OS+ network+administrator	5	2	2	0	1.000	0.400	252	5	247	0.020	1.000
15.	database+oracle+sql+guide+administrator	7	5	5	0	1.000	0.714	114	7	107	0.061	1.000
16.	internet+network+image+web+homepage	10	14	10	4	0.714	1.000	176	10	166	0.057	1.000
17.	network+communication+programming+ application+layer+protocol **	4	2	2	0	1.000	0.500	223	4	219	0.018	1.000
18.	network+communication+programming+ application+layer+protocol **	12	14	11	3	0.786	0.917	223	12	211	0.054	1.000
19.	network+communication+image+digital+ internet+web+html	13	22	11	11	0.500	0.846	195	13	182	0.067	1.000
20.	programming+network+internet+java+ application+window+interface	3	2	2	0	1.000	0.667	309	3	306	0.010	1.000



รูปที่ 5.1 การเปรียบเทียบค่าความแม่นยำ (Precision) ของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND และระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR ในแต่ละชุดคำถามสืบค้นจำนวนทั้งหมด 20 ชุดคำถาม

จากกราฟรูปที่ 5.1 ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND จะมีค่าความแม่นยำมากในชุดของคำถามที่มีจำนวนคำสืบค้นอยู่ในช่วง 3-5 คำ ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่าความแม่นยำต่ำมาก ในขณะที่ระบบค้นคืนสารสนเทศที่ได้ทำการวิจัยมีค่าความแม่นยำขึ้น-ลงอยู่ในระดับเฉลี่ยที่สูงกว่าค่า 0.400

จากกราฟรูปที่ 5.2 ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND จะมีความระลึกต่ำมากในชุดของคำถามที่มีจำนวนคำสืบค้นอยู่ในช่วง 5-7 คำ ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่าความระลึกสูงมากและคงที่ตลอด ในขณะที่ระบบค้นคืนสารสนเทศที่ได้ทำการวิจัยมีค่าความระลึกขึ้น-ลงอยู่ในระดับเฉลี่ยที่สูงกว่าค่า 0.400



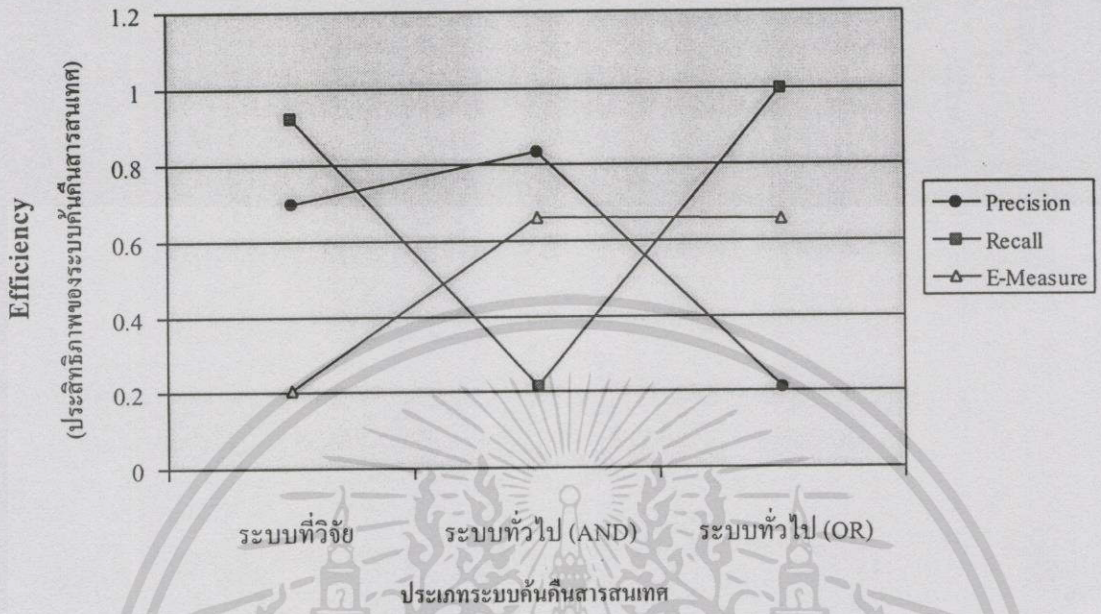
รูปที่ 5.2 การเปรียบเทียบค่าความระลึก (Recall) ของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND และระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR ในแต่ละชุดคำถามสืบค้นจำนวนทั้งหมด 20 ชุดคำถาม

นอกจากนี้ผู้วิจัยได้เปรียบเทียบว่าจำนวนคำสืบค้นที่ใส่ในแต่ละครั้งมีผลกับประสิทธิภาพของระบบค้นคืนสารสนเทศแต่ละประเภทอย่างไร แสดงเป็นตารางเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ และกราฟเปรียบเทียบระบบค้นคืนสารสนเทศทั้ง 3 ประเภท ซึ่งได้มาจากค่าเฉลี่ยของจำนวนคำสืบค้นแต่ละประเภท จากตารางที่ 5.3 และตารางที่ 5.4 ดังนี้

ตารางที่ 5. 5 เปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 3 คำ

	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.6975	0.8333	0.2090
Recall	0.9213	0.2133	1.0000
E-Measure	0.2061	0.6603	0.6543

สามารถเขียนกราฟได้ดังรูปที่ 5.3



รูปที่ 5.3 กราฟเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 3 คำ

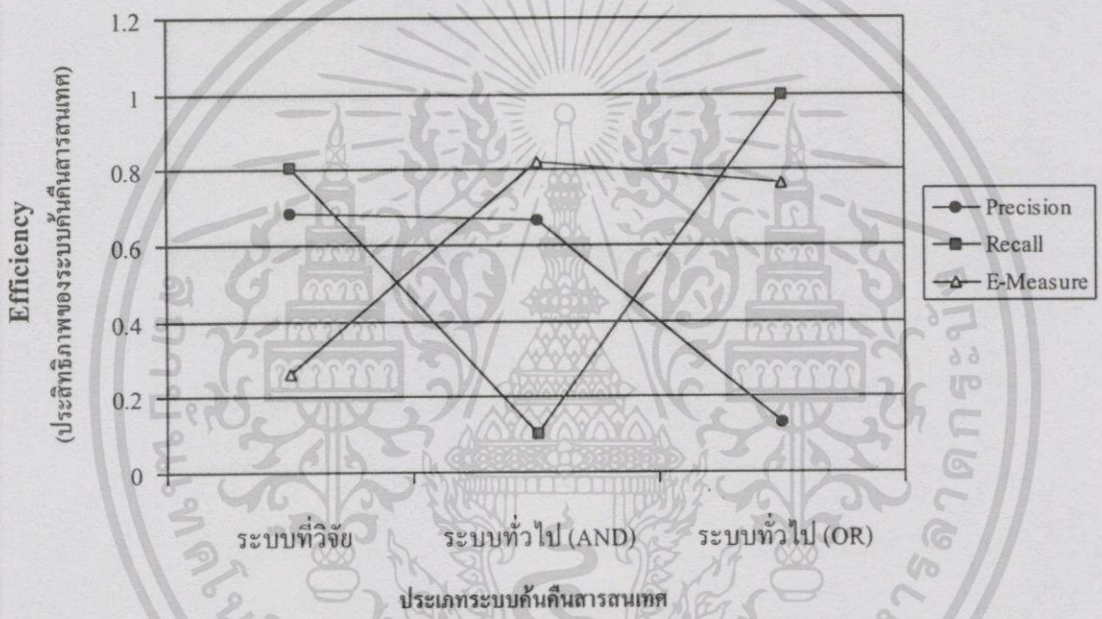
จากตารางที่ 5.5 และกราฟรูปที่ 5.3 จะเห็นได้ว่า ระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ทำการวิจัยนั้น มีค่า E ที่ต่ำกว่า ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND และ ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR แสดงว่าระบบที่ทำการวิจัยมีประสิทธิภาพที่ดีกว่า ส่วนค่าความแม่นยำ ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND จะมีค่าสูงที่สุด เนื่องจาก ระบบนี้จะค้นคืนเฉพาะเอกสารที่มีคำสืบค้นทุกคำที่ผู้สืบค้นได้ใส่ลงไป ส่วนค่าความแม่นยำของระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR จะมีค่าต่ำที่สุดเนื่องจากระบบนี้ ค้นคืนเอกสารทั้งหมดออกมาจึงทำให้ได้เอกสารจำนวนมาก แต่ที่ตรงต่อความต้องการมีน้อย

สำหรับค่าความระลึก ระบบที่ทำการวิจัยมีค่าความระลึกสูงมากพอควร แต่ไม่เท่าระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR เนื่องจากระบบนี้ค้นคืนเอกสารออกมาทั้งหมด ทั้งที่เกี่ยวข้องและไม่เกี่ยวข้องกับความต้องการของผู้สืบค้น

ตารางที่ 5.6 เปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 4 คำ

	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.6827	0.6667	0.1335
Recall	0.8040	0.1043	1.0000
E-Measure	0.2616	0.8196	0.7644

สามารถเขียนกราฟได้ดังรูปที่ 5.4



รูปที่ 5.4 กราฟเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 4 คำ

จากตารางที่ 5.6 และกราฟรูปที่ 5.4 ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND เริ่มมีค่า E สูงมากขึ้น ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่า E คงอยู่ในระดับสูง ในขณะที่ระบบที่ทำการวิจัยมีค่า E สูงขึ้นเพียงเล็กน้อย แต่ก็ยังอยู่ในระดับที่ต่ำ

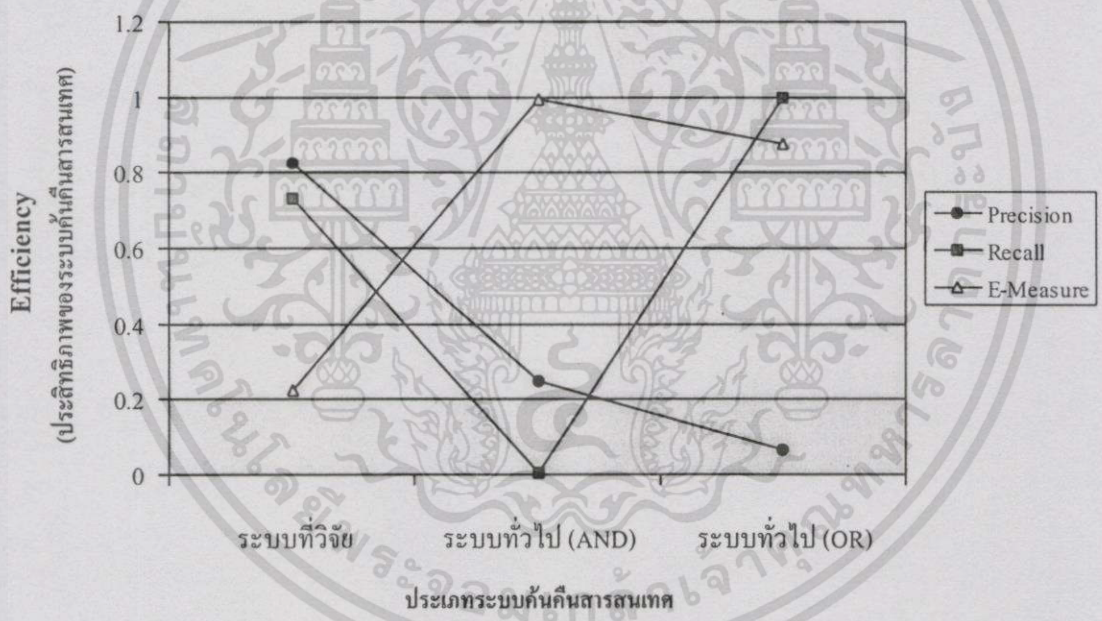
ส่วนค่าความแม่นยำ สำหรับระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND เริ่มมีค่าต่ำลงกว่าการสืบค้นโดยใช้คำสืบค้น 3 คำ ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่าความแม่นยำต่ำลงเล็กน้อย ส่วนระบบที่ทำการวิจัย มีค่าความแม่นยำใกล้เคียงกับการสืบค้นโดยใช้คำสืบค้น 3 คำ

สำหรับค่าความระลึก ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR ยังมีค่าสูงที่สุด ในขณะที่อีก 2 ระบบ ลดลงเล็กน้อย

ตารางที่ 5.7 เปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 5 คำ

	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.8238	0.2500	0.0648
Recall	0.7330	0.0028	1.0000
E-Measure	0.2242	0.9945	0.8783

สามารถเขียนกราฟได้ดังรูปที่ 5.5



รูปที่ 5.5 กราฟเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 5 คำ

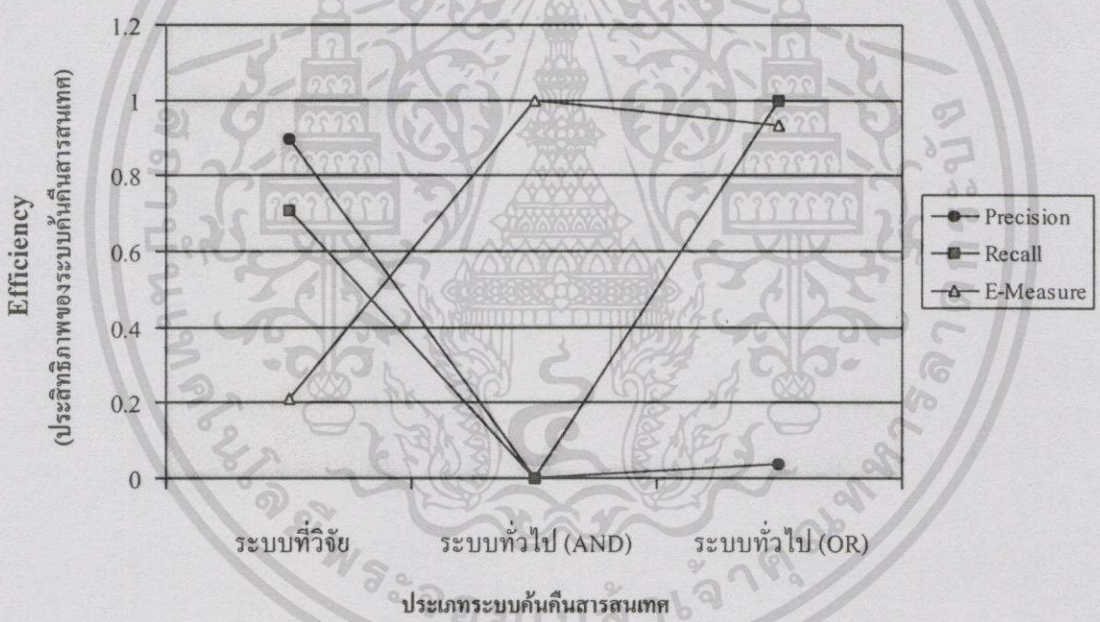
จากตารางที่ 5.7 และกราฟรูปที่ 5.5 ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND มีค่า E ที่สูงมาก และสูงกว่าระบบอื่น ส่วนค่าความระลึกมีค่าต่ำที่สุดเช่นเดียวกัน แสดงว่าระบบนี้ไม่มีประสิทธิภาพในการสืบค้นสำหรับกรณีคำสืบค้นจำนวน 5 คำ ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่า E ที่สูงมากเช่นเดียวกัน ในขณะที่ค่า E ของระบบที่ทำการวิจัย อยู่ในระดับใกล้เคียงกับการสืบค้นที่มีจำนวนคำสืบค้น 4 คำ และยังคงอยู่ใน

ระดับที่ต่ำแสดงว่าระบบมีประสิทธิภาพในการค้นคืนที่ดีกว่าอีก 2 ระบบ เช่นเดียวกันกับค่าความแม่นยำและค่าความระลึกที่ยังคงอยู่ในระดับที่สูง

ตารางที่ 5.8 เปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 6 คำ

	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.8930	0.0000	0.0360
Recall	0.7085	0.0000	1.0000
E-Measure	0.2099	1.0000	0.9305

สามารถเขียนกราฟได้ดังรูปที่ 5.6



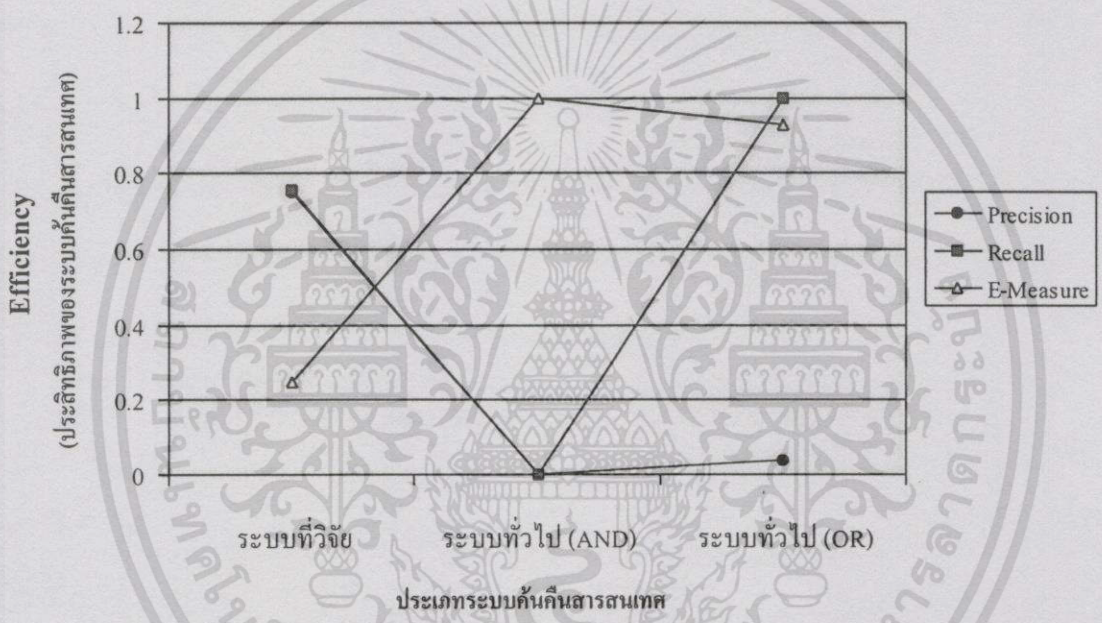
รูปที่ 5.6 กราฟเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 6 คำ

จากตารางที่ 5.8 และกราฟรูปที่ 5.6 แสดงให้เห็นว่าระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND ไม่สามารถทำงานได้อีกต่อไปเนื่องจากค่าความแม่นยำ และค่าความระลึก เป็น 0 ทั้งคู่ ส่วนค่า E มีค่าสูงที่สุดคือ 1 ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR มีค่าความแม่นยำที่ต่ำมากอย่างคงที่ และค่าความระลึกที่สูงมากอย่างคงที่เช่นกัน ในขณะที่ระบบค้นคืนสารสนเทศที่ได้ทำการวิจัยยังคงรักษาสมดุลของค่าทั้ง 3 ค่าได้อย่างดี

ตารางที่ 5.9 เปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 7 คำ

	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.7500	0.0000	0.0385
Recall	0.7565	0.0000	1.0000
E-Measure	0.2468	1.0000	0.9259

สามารถเขียนกราฟได้ดังรูปที่ 5.7



รูปที่ 5.7 กราฟเปรียบเทียบการสืบค้น 3 รูปแบบ กับ จำนวนคำสืบค้นที่ใส่ให้กับระบบ กรณีคำสืบค้น 7 คำ

สำหรับการสืบค้นโดยใช้คำสืบค้นจำนวน 7 คำนี้ จะมีรูปกราฟที่คล้ายคลึงกับการสืบค้นโดยใช้คำสืบค้นจำนวน 6 คำ ค่าความแม่นยำและค่าความระลึกของระบบที่ได้ทำการวิจัยเริ่มขยับเข้าใกล้กันมาก ส่วนค่าทั้ง 3 ค่า ของอีก 2 ระบบ มีลักษณะกราฟคล้ายกับกราฟรูปที่ 5.6

จากตารางและกราฟที่ได้เสนอมาทั้งหมด สามารถสรุปได้ว่า เมื่อจำนวนคำสืบค้นมากขึ้นจนถึง 6 คำ ระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND ไม่สามารถทำงานได้อีกต่อไป ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR ยังคงทำงานได้ แต่จะมีความแม่นยำที่น้อยมาก เนื่องจากระบบนี้จะค้นคืนเอกสารออกมาทั้งหมดทำให้ผู้สืบค้นได้เอกสารที่ไม่ต้องการออกมามากกว่าที่ต้องการ ในขณะที่ระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ผู้วิจัยได้

ทำการวิจัยนั้นสามารถรักษาสมดุลของค่าทั้ง 3 ค่าได้ แม้ว่าจะมีจำนวนคำสืบค้นที่เพิ่มมากขึ้น แสดงว่าระบบค้นคืนเอกสารที่ตรงต่อความต้องการของผู้สืบค้นได้ และระบบสามารถควบคุมให้ผู้สืบค้นได้จำนวนเอกสารที่เป็นผลลัพธ์ไม่น้อยจนเกินไปและไม่มากจนเกินไป

5.3 การเปรียบเทียบการเรียงเอกสารตามทฤษฎีโครงข่ายเบย์ ของระบบค้นคืนสารสนเทศแบบฮิวริสติคที่วิจัย กับการเรียงเอกสารตามลำดับชื่อเอกสาร

ระบบค้นคืนสารสนเทศแบบทั่วไปหากมีการจัดเรียงเอกสาร มักจะเรียงตามลำดับชื่อเอกสาร ซึ่งผู้วิจัยเห็นว่าวิธีการจัดเรียงเอกสารแบบนี้ ผลลัพธ์รายการเอกสารที่ได้ไม่เรียงตามลำดับความต้องการของผู้สืบค้น เนื่องจากผู้สืบค้นไม่ได้ระบุความต้องการของตนเองลงไปด้วย แต่การเรียงเอกสารโดยการใช้ทฤษฎีโครงข่ายเบย์เข้ามาช่วยนั้น ทำให้ผู้สืบค้นได้ผลลัพธ์รายการเอกสารที่เรียงลำดับตามความต้องการ เพื่อให้เห็นภาพชัดเจนยิ่งขึ้น ผู้วิจัยขอแสดงผลลัพธ์เพื่อทำการเปรียบเทียบวิธีการจัดเรียงเอกสาร 2 วิธีการดังกล่าว

ยกตัวอย่างการเรียงลำดับเอกสารโดยอาศัยการให้น้ำหนักของเทอมจากทฤษฎีโครงข่ายเบย์ เปรียบเทียบกับการเรียงลำดับเอกสารแบบเรียงตามชื่อเอกสาร การทดลองใช้ชุดคำสืบค้น {Network, Image, Digital, Database} ซึ่งก่อนหน้าที่จะทำการเรียงลำดับ ได้ให้ผู้สืบค้นพิจารณาเอกสารที่ตนเองต้องการจำนวนหนึ่งจากรายการเอกสาร โดยที่เอกสารที่ผู้สืบค้นต้องการจะปรากฏคำว่า "Rel" อยู่ในคอลัมน์ "JUDGED" (Manually judged) ของหน้าจอผลลัพธ์ ซึ่งมีจำนวนเอกสาร 39 รายการ ดังรูปที่ 5.8 และ รูปที่ 5.9

รูปที่ 5.8 เป็นการแสดงรายการผลลัพธ์เอกสารที่ได้จากการสืบค้น โดยใช้การเรียงเอกสารตามน้ำหนักของเทอมจากทฤษฎีโครงข่ายเบย์ จะสังเกตเห็นว่าเอกสารที่ผู้สืบค้นต้องการอยู่ในลำดับต้นๆ ส่วนรูปที่ 5.8 เป็นการแสดงรายการผลลัพธ์เอกสารที่ได้จากการสืบค้น โดยใช้การเรียงเอกสารตามลำดับตัวอักษรที่เป็นชื่อเอกสาร ซึ่งเอกสารที่ต้องการจะกระจายอยู่ทั่วไปมีทั้งอยู่ในลำดับต้นๆ และลำดับท้ายๆ

ดังนั้นจึงสามารถสรุปได้ว่าการเรียงเอกสารโดยใช้น้ำหนักของเทอมตามทฤษฎีโครงข่ายเบย์ สามารถจัดลำดับเอกสารที่ผู้สืบค้นต้องการให้อยู่ในลำดับต้นๆ ของรายการเอกสารผลลัพธ์ได้ ทั้งนี้เนื่องจากได้อาศัยข้อมูลฮิวริสติคที่ได้จากผู้สืบค้นมาช่วยในการเรียงลำดับเอกสาร

SEQ	ISBN	TITLE	SCORE	JUDGED
1	0-13-243700-7	ATM & MPEG-2: integrating digital video into broadband networks	9.528	
2	0-07-021422-0	The intelligent network standards: their application to services	8.528	Rel
3	1-56205-603-4	Internetworking technologies handbook	8.528	
4	0-13-571274-2	Data and computer communications	8.528	Rel
5	0-13-259193-6	ISDN & SS7: architectures for digital signaling networks	8.361	Rel
6	0-201-42270-0	Wide area network performance and optimization: practical strategies for	7.528	Rel
7	1-57521-113-0	Web programming with Java	7.528	Rel
8	156604-329-8	The Internet power toolkit: cutting-edge tools & techniques for power users	7.528	Rel
9	0-471-14274-3	Reengineering IBM networks	7.528	
10	1-56830-300-9	Internet publishing with Acrobat	7.528	
11	1-57521-051-7	Web publishing unleashed: HTML, CGI, SGML, VRML Java	7.389	
12	1-883577-88-8	Internet protocols handbook	7.389	
13	1-55851-443-0	ATM: the key to high-speed broadband networking	7.389	
14	1-56830-307-6	Web page scripting Techniques	7.361	
15	0-13-394338-9	Digital signal processing: principles, algorithms, and applicatios	7.306	
16	0-534-20244-6	Understanding Data Communications-and Networks	6.528	
17	0-07-024043-4	Introduction to ATM networking	6.528	
18	0-07-005593-9	The X series recommendations: standards for data communications	6.389	
19	0-8493-2516-1	The image processing handbook	6.306	
20	0-07-057240-2	Image processing:theory,algorithms,and architectures	6.306	
21	0-13-190075-7	Digital video processing	6.306	
22	0-02-415441-5	Data and computer communications	5.722	Rel
23	0-201-42274-3	ATM networks concepts, protocols, applications	5.528	
24	0-12-691395-1	Network design essentials	5.389	
25	0-201-62745-0	Network and distributed systems management	5.389	
26	0-8053-7724-7	Local area networks	5.306	
27	0-471-00949-0	Digital image processing: principles and applications	5.306	
28	0-07-020346-6	TCP/IP: architecture, protocols and implementation	4.528	
29	0-13-090853-3	Data communications and distributed networks	4.528	
30	0-201-56318-5	Unix System V network programming	4.389	
31	0-201-56741-5	Internet system handbook	4.389	
32	0-13-036252-2	Distributed computing a practical synthesis of network, and client-server syster	4.389	
33	0-02-946399-8	Isdn: An Introduction	3.528	
34	0-07-157673-8	Distributed databases, cooperative processing and networking	3.417	
35	0-201-56506-4	Data communications, computer networks and open systems	3.361	
36	0-201-50803-6	Digital image procesing	3.306	
37	0-13-474321-0	Internetworking with TCP/IP	2.361	
38	0-07-707323-1	Image processing	2.306	
39	0-02-415465-2	Local and metropolitan area networks	1.528	

รูปที่ 5.8 ผลลัพธ์รายการเอกสารที่เรียงลำดับตามการให้น้ำหนักของเทอมจากทฤษฎีโครงข่ายเบย์

SEQ	ISBN	TITLE	SCORE	JUDGED
1	0-13-243700-7	ATM & MPEG-2: integrating digital video into broadband networks	9.528	
2	0-201-42274-3	ATM networks concepts, protocols, applications	5.528	
3	1-55851-443-0	ATM: the key to high-speed broadband networking	7.389	
4	0-13-571274-2	Data and computer communications	8.528	Rel
5	0-02-415441-5	Data and computer communications	5.722	Rel
6	0-13-090853-3	Data communications and distributed networks	4.528	
7	0-201-56506-4	Data communications, computer networks and open systems	3.361	
8	0-201-50803-6	Digital image procesing	3.306	
9	0-471-00949-0	Digital image processing: principles and applications	5.306	
10	0-13-394338-9	Digital signal processing: principles, algorithms, and applicatios	7.306	
11	0-13-190075-7	Digital video processing	6.306	
12	0-13-036252-2	Distributed computing a practical synthesis of network, and client-server syster	4.389	
13	0-07-157673-8	Distributed databases, cooperative processing and networking	3.417	
14	0-13-259133-6	ISDN & SS7: architectures for digital signaling networks	8.361	Rel
15	0-07-707323-1	Image processing	2.306	
16	0-07-057240-2	Image processing:theory,algorithms,and architectures	6.306	
17	1-883577-88-8	Internet protocols handbook	7.389	
18	1-56830-300-9	Internet publishing with Acrobat	7.528	
19	0-201-56741-5	Internet system handbook	4.389	
20	1-56205-603-4	Internetworking technologies handbook	8.528	
21	0-13-474321-0	Internetworking with TCP/IP	2.361	
22	0-07-024043-4	Introduction to ATM networking	6.528	
23	0-02-946399-8	Isdn: An Introduction	3.528	
24	0-02-415465-2	Local and metropolitan area networks	1.528	
25	0-8053-7724-7	Local area networks	5.306	
26	0-201-62745-0	Network and distributed systems management	5.389	
27	0-12-691395-1	Network design essentials	5.389	
28	0-471-14274-3	Reengineering IBM networks	7.528	
29	0-07-020346-6	TCP/IP: architecture, protocols and implementation	4.528	
30	156604-329-8	The Internet power toolkit: cutting-edge tools & techniques for power users	7.528	Rel
31	0-07-005593-9	The X series recommendations: standards for data communications	6.389	
32	0-8493-2516-1	The image processing handbook	6.306	
33	0-07-021422-0	The intelligent network standards: their application to services	8.528	Rel
34	0-534-20244-6	Understanding Data Communications-and Networks	6.528	
35	0-201-56318-5	Unix System V network programming	4.389	
36	1-56830-307-6	Web page scripting Techniques	7.361	
37	1-57521-113-0	Web programming with Java	7.528	Rel
38	1-57521-051-7	Web publishing unleashed: HTML, CGI, SGML, VRML Java	7.389	
39	0-201-42270-0	Wide area network performance and optimization: practical strategies for	7.528	Rel

รูปที่ 5.9 ผลลัพธ์รายการเอกสารที่เรียงลำดับตามชื่อเอกสาร

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

ปัญหาของการค้นคืนสารสนเทศให้ได้ประสิทธิภาพดีนั้น ขึ้นอยู่กับว่าเอกสารที่ระบบค้นคืนให้ ตรงต่อความต้องการของผู้สืบค้นเพียงใด ซึ่งเป็นการยากที่จะวัดค่าเหล่านี้ออกมาได้อย่างแม่นยำ เนื่องจากตัวแปรสำคัญที่จะตัดสินว่าระบบค้นคืนสารสนเทศนั้นทำงานได้ถูกต้องตรงตามความต้องการ คือผู้สืบค้นเอกสารเท่านั้น

การค้นคืนสารสนเทศโดยใช้เทคนิคการสืบค้นแบบฮิวริสติค โดยอาศัยข้อมูลฮิวริสติค (Heuristic Information) และประยุกต์ทฤษฎีโครงข่ายเบย์ในเรื่องการจัดเรียงเอกสาร ดังที่ผู้วิจัยได้ทำการศึกษาและวิจัยในครั้งนี้ นับว่าเป็นอีกทางเลือกหนึ่งของการค้นคืนสารสนเทศ และการปรับปรุงเทคนิคการสืบค้นให้ก้าวหน้ายิ่งขึ้น โดยมีเป้าหมายเพื่อกำจัดเอกสารที่ไม่ตรงตามความต้องการออกไป ในอีกทางหนึ่งระบบต้องสามารถค้นคืนเอกสารที่ตรงตามความต้องการ หรือใกล้เคียงกับความต้องการออกมาให้ได้มากที่สุด โดยท้ายที่สุดแล้วผู้สืบค้นจะได้รายชื่อเอกสารเรียงลำดับตามน้ำหนักของความต้องการจากมากไปหาน้อย

ในงานวิจัยนี้ปัจจัยหลักของระบบค้นคืนสารสนเทศคือ ตัวระบบ ฐานข้อมูลเอกสาร และผู้ใช้ระบบหรือผู้สืบค้นเอกสาร ผู้วิจัยเห็นว่าข้อมูลฮิวริสติคมีส่วนช่วยเหลือระบบในการสืบค้นอย่างมาก ยิ่งระบบมีข้อมูลเกี่ยวกับเอกสารและผู้สืบค้นมากเพียงใด ยิ่งช่วยให้ระบบทำงานได้ดีขึ้นมากเท่านั้น เริ่มจากตัวระบบ ผู้สืบค้นสามารถสืบค้นเอกสารได้โดยอาศัยคำสืบค้นใน 2 รูปแบบ คือสืบค้นโดยใช้คำสืบค้นที่อ้างถึงตัวเอกสาร และสืบค้นโดยใช้รายละเอียดอื่นที่มีอยู่ในเอกสาร ดังที่ได้กล่าวไปแล้วนั้น เรื่องของการสืบค้นโดยใช้คำสืบค้น ผู้สืบค้นไม่ต้องเสียเวลาในการใส่ตัวดำเนินการ (AND OR และ NOT) ระหว่างคำสืบค้นให้กับระบบ ระบบจะนำคำที่ผู้สืบค้นใส่ ไปสร้างความสัมพันธ์ของคำสืบค้นโดยอัตโนมัติ โดยอาศัยหลักการสร้างแลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets) โดยยึดอัลกอริทึม IRA เป็นหลัก และอาศัยฮิวริสติคฟังก์ชัน $f^*(n)$ ที่ผู้วิจัยได้นำเสนอไปแล้วในบทที่ 4 ซึ่งฟังก์ชันนี้ เป็นตัวกำหนดทางเดินเพื่อหาโหนดเป้าหมายใน Search Tree นั่นเอง ส่วนเรื่องของการสืบค้นโดยใช้รายละเอียดอื่นที่มีอยู่ในเอกสาร ผู้สืบค้นสามารถใส่รายละเอียดของเอกสารที่ต้องการ (คำถามสืบค้น) ได้หลายส่วนพร้อมกันเพียงครั้งเดียวเท่านั้น

อย่างไรก็ดีผู้วิจัยเห็นว่าการสืบค้นในหลายลักษณะ ไม่ว่าจะเป็นการสืบค้นเอกสารตามห้องสมุด หรือการค้นหาข้อมูลผ่านเครือข่ายใยแมงมุม (World Wide Web) ปัจจัยสำคัญคือความพอใจของผู้ใช้หรือผู้สืบค้น ในการพิจารณาว่าระบบค้นคืนสารสนเทศดังกล่าวสามารถค้นคืน

เอกสารให้ได้ตรงตามความต้องการหรือไม่ แต่ปัญหาสำคัญอีกประการหนึ่งคือระบบค้นคืนสารสนเทศแบบทั่วไป มักแสดงรายชื่อเอกสารออกมาโดยไม่ได้เรียงลำดับเอกสารว่าฉบับใดน่าจะตรงตามความต้องการมากและน้อยตามลำดับ หากมีการเรียงเอกสาร ส่วนมากมักเรียงตามลำดับตัวอักษร ซึ่งวิธีการนี้ไม่สามารถบ่งบอกได้ว่าเอกสารที่เรียงลำดับนั้นเรียงตามความต้องการของผู้สืบค้น ทำให้ผู้สืบค้นต้องเสียเวลาในการเลือกเอกสาร หากบางครั้งเอกสารที่ต้องการอยู่ในรายการเอกสารผลลัพธ์รายการต้นๆ ผู้สืบค้นจึงไม่เสียเวลาในการคัดเลือกเอกสารที่ต้องการ แต่หากผู้สืบค้นต้องเลือกดูเอกสารจนถึงรายการลำดับท้ายๆ จากรายการเอกสารผลลัพธ์ที่มีจำนวนมากนั้น ถือว่าเป็นการสิ้นเปลืองเวลา บางระบบอาจใช้เทคนิคการจัดเรียงเอกสารที่ซับซ้อนกว่านั้น คือใช้วิธีการนับจำนวนความถี่ของคำสืบค้นที่ปรากฏอยู่ในเอกสาร ถ้ามีจำนวนคำสืบค้นปรากฏอยู่ในเอกสารมาก แสดงว่าเอกสารนั้นตรงต่อความต้องการมาก ซึ่งอาจเป็นไปได้ในบางกรณีเท่านั้น เนื่องจากความต้องการของผู้สืบค้นมีความซับซ้อนอยู่ไม่น้อย ดังนั้นเอกสารที่มีจำนวนคำสืบค้นปรากฏอยู่มาก ไม่จำเป็นเสมอไปว่าต้องตรงต่อความต้องการของผู้สืบค้นมาก เพราะผู้สืบค้นไม่ได้ระบุความต้องการนี้โดยตรง เป็นเพียงการใส่คำสืบค้นเพื่อให้ระบบกระทำการสืบค้นเท่านั้น ซึ่งในส่วนนี้ ผู้วิจัยได้แก้ปัญหาโดยการนำทฤษฎีโครงข่ายเบย์ ที่ประยุกต์มาจากความน่าจะเป็นของเบย์มาใช้ โดยได้พิจารณาถึงการให้น้ำหนักของเทอม (Term weighting) เป็นหลักสำคัญในการเรียงเอกสารตามลำดับความต้องการ ซึ่งการกำหนดน้ำหนักของเทอมนี้ถือเป็นการได้รับข้อมูลฮิวริสติคอีกประการหนึ่งจากผู้สืบค้นเอกสาร เนื่องจากผู้สืบค้นจะเป็นผู้กำหนดความต้องการของตนเอง ทำให้เอกสารที่เป็นผลลัพธ์ถูกจัดเรียงตามความต้องการของผู้สืบค้นได้ดียิ่งขึ้น

เรื่องการนำข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น (User Profile) มาใช้เป็นข้อมูลฮิวริสติคให้กับระบบนั้น ก็เป็นอีกส่วนหนึ่งในการช่วยให้ระบบพิจารณาเอกสารให้ตรงตามความต้องการมากขึ้นเช่นเดียวกัน

6.1 สรุปผลการทดลองและการวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ

การวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ ยังคงต้องอาศัยการตัดสินจากผู้สืบค้นเอง ว่าเอกสารที่เป็นผลลัพธ์นั้น แท้จริงแล้วมีความตรงต่อความต้องการหรือตรงใจผู้สืบค้นมากน้อยเพียงใด จึงเป็นการวัดค่าออกมาได้ยากพอควร ถ้าวัดจากการสอบถามหรือสัมภาษณ์ผู้สืบค้นเองน่าจะวัดผลได้ดีกว่า

แม้ว่าการสืบค้นด้วยวิธีการที่ผู้วิจัยได้ทำการวิจัยในครั้งนี้ ผู้สืบค้นจะต้องให้เวลากับการสืบค้นมากกว่าระบบค้นคืนสารสนเทศแบบทั่วไปเพียงเล็กน้อยก็ตาม ซึ่งเวลาที่เสียไปใช้ในการตอบคำถามและให้ข้อมูลแก่ระบบ แต่ผู้วิจัยได้เล็งเห็นแล้วว่าผลลัพธ์ของรายการเอกสารที่ได้ เป็นที่น่าพอใจกว่าระบบค้นคืนสารสนเทศแบบเก่า อย่างน้อยที่สุดก็ช่วยลดเวลาผู้สืบค้น ในการคัดเลือก

เอกสารที่เป็นผลลัพธ์ได้ ดังนั้นเมื่อพิจารณาเวลาที่เสียไปในระหว่างการสืบค้นกับระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ได้ทำการวิจัย เปรียบเทียบกับ เวลาที่เสียไปให้กับระบบค้นคืนสารสนเทศทั่วไปในขั้นตอนของการคัดเลือกผลลัพธ์ที่ตรงต่อความต้องการแล้ว ผู้วิจัยเห็นว่าน่าจะคุ้มค่าและสามารถยอมรับได้ในทางปฏิบัติ

จากผลการทดลองและการวัดประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้กล่าวไปแล้วในบทที่ 5 สามารถสรุปผลการเปรียบเทียบประสิทธิภาพระหว่างระบบค้นคืนสารสนเทศทั้ง 3 ลักษณะ คือ ระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ทำการวิจัย ระบบค้นคืนสารสนเทศแบบทั่วไป (ใช้ตัวดำเนินการเป็น : AND) และ ระบบค้นคืนสารสนเทศทั่วไป (ใช้ตัวดำเนินการเป็น : OR) ได้เป็น 2 หลักใหญ่คือ

1. การเปรียบเทียบการวัดประสิทธิภาพระหว่างระบบค้นคืนสารสนเทศทั้ง 3 ระบบ โดยคิดค่าเฉลี่ยจากชุดคำถามสืบค้นทั้ง 20 ชุดคำถาม

2. การเปรียบเทียบผลกระทบของจำนวนคำถามสืบค้นที่เพิ่มขึ้น ต่อประสิทธิภาพของระบบค้นคืนสารสนเทศทั้ง 3 ระบบ

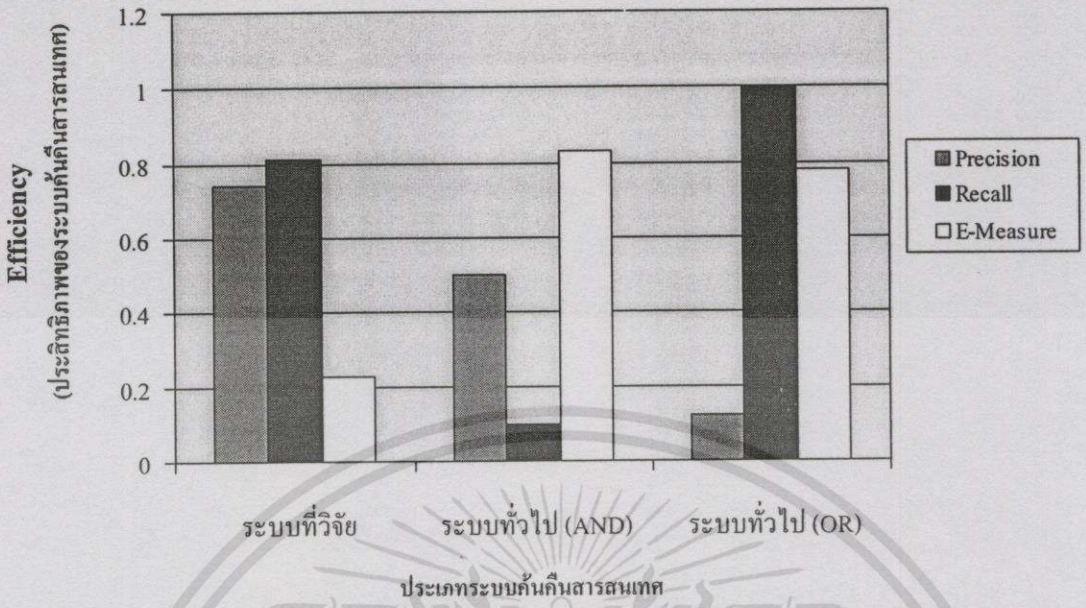
ผลของการเปรียบเทียบสามารถแสดงได้ดังตาราง และรูปกราฟ ดังนี้

ตารางที่ 6.1 เปรียบเทียบการวัดประสิทธิภาพระหว่าง ระบบค้นคืนสารสนเทศทั้ง 3 ลักษณะ โดยคิดค่าเฉลี่ยจากชุดคำถามสืบค้นทั้ง 20 ชุดคำถาม

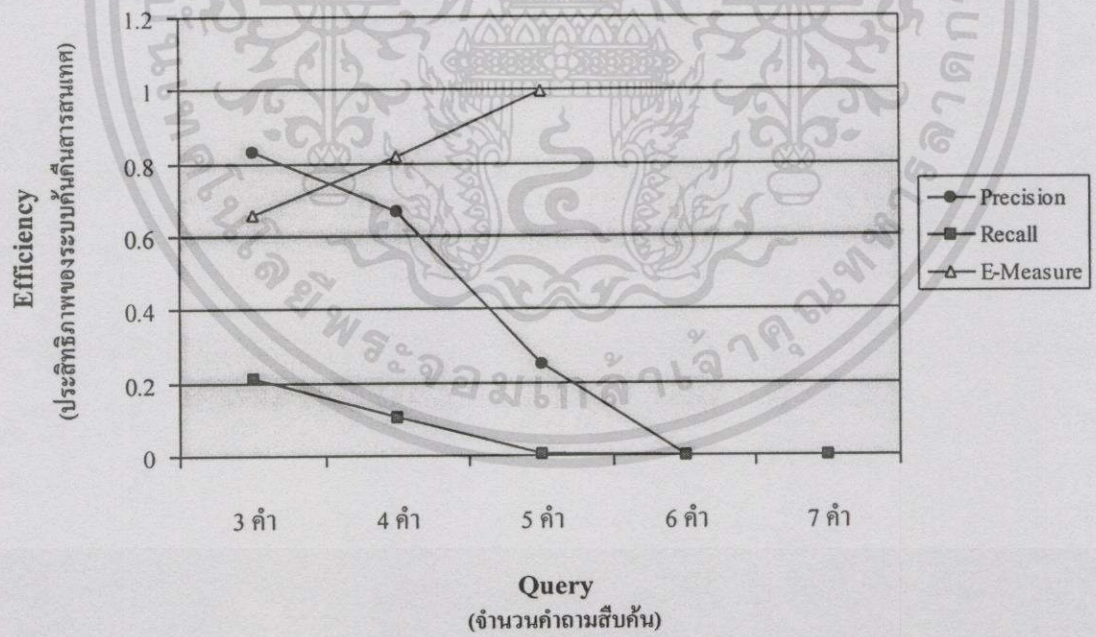
	ระบบที่วิจัย	ระบบทั่วไป (AND)	ระบบทั่วไป (OR)
Precision	0.7431	0.5000	0.1232
Recall	0.8107	0.0999	1.0000
E-Measure	0.2246	0.8335	0.7806

จากตารางที่ 6.1 จะเห็นได้ว่าค่าความแม่นยำ ค่าความระลึกและค่า E ของระบบค้นคืนสารสนเทศทั้ง 3 ลักษณะ มีความแตกต่างกัน เพื่อให้เห็นความแตกต่างอย่างชัดเจน สามารถนำค่าทั้ง 3 ค่า ของทั้ง 3 ระบบมาพล็อตกราฟได้ดังรูปที่ 6.1

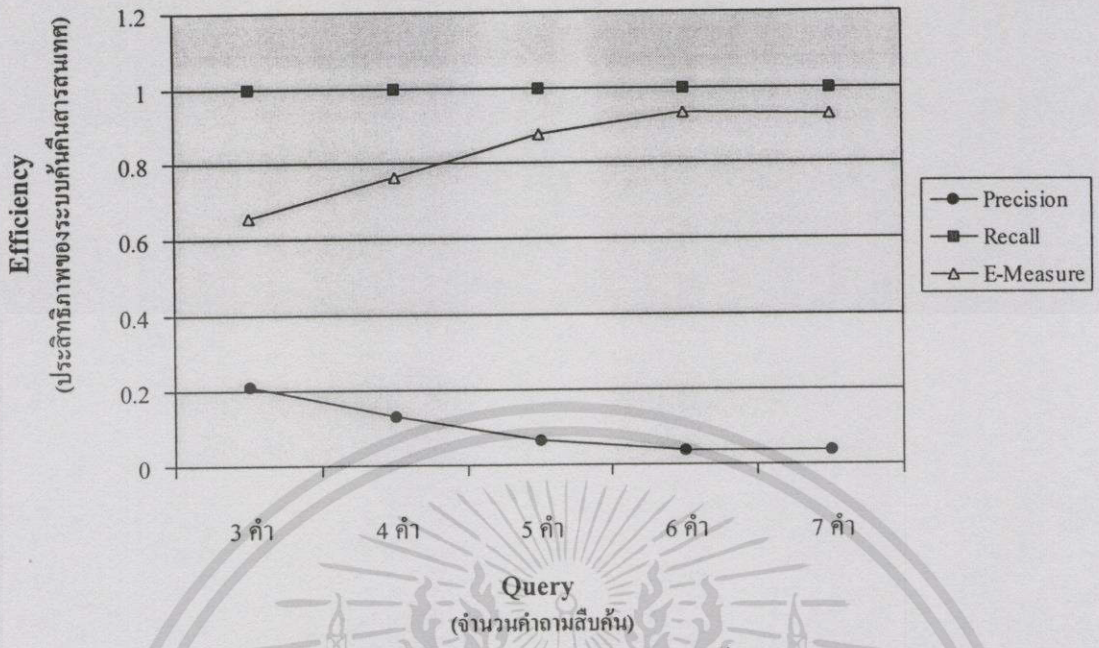
นอกจากนี้ ผู้วิจัยยังพบว่าเมื่อเราให้จำนวนคำสืบค้นในชุดของคำถามสืบค้นเป็นตัวพารามิเตอร์ เมื่อจำนวนคำสืบค้นในแต่ละชุดเพิ่มขึ้นจะมีผลการทบกับค่าความแม่นยำ ค่าความระลึก และค่า E ของระบบค้นคืนสารสนเทศแต่ละประเภทด้วย ดังแสดงในกราฟรูปที่ 6.2 - รูปที่ 6.4



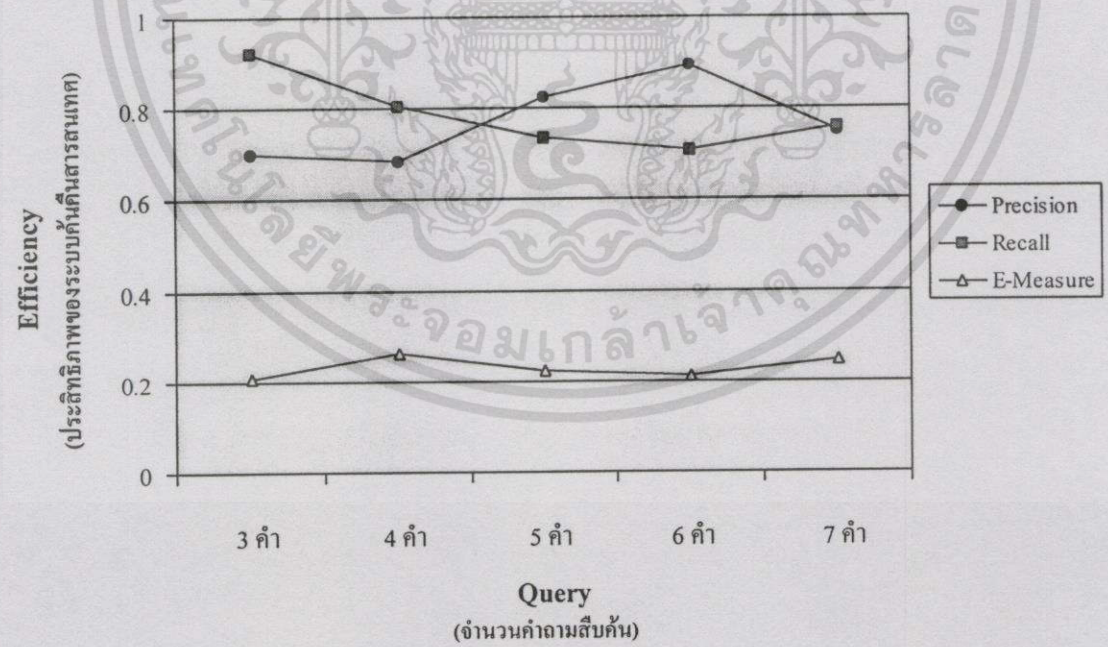
รูปที่ 6.1 กราฟเปรียบเทียบการวัดประสิทธิภาพระหว่าง ระบบค้นคืนสารสนเทศทั้ง 3 ลักษณะ โดยคิดค่าเฉลี่ยจากชุดคำถามสืบค้นทั้ง 20 ชุดคำถาม



รูปที่ 6.2 กราฟแสดงผลกระทบของจำนวนคำถามสืบค้นที่เพิ่มขึ้น ต่อประสิทธิภาพของระบบค้นคืนสารสนเทศทั่วไป (ใช้ตัวดำเนินการเป็น : AND)



รูปที่ 6.3 กราฟแสดงผลกระทบของจำนวนคำถามสืบค้นที่เพิ่มขึ้น ต่อประสิทธิภาพของระบบค้นคืนสารสนเทศแบบทั่วไป (ใช้ตัวดำเนินการเป็น : OR)



รูปที่ 6.4 กราฟแสดงผลกระทบของจำนวนคำถามสืบค้นที่เพิ่มขึ้น ต่อประสิทธิภาพของระบบค้นคืนสารสนเทศที่ได้ทำการวิจัย

จากรูปที่ 6.2 ถึง รูปที่ 6.4 เห็นได้ว่า จำนวนของคำสืบค้นมีผลกระทบต่อระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น AND มากที่สุด เห็นได้ชัดว่า เมื่อมีจำนวนคำถามสืบค้นมากขึ้น ค่าความแม่นยำ และค่าความระลึกจะลดต่ำลง ในขณะที่ค่า E ยิ่งเพิ่มมากขึ้น แสดงว่าระบบมีประสิทธิภาพลดลงเมื่อมีจำนวนคำถามสืบค้นมากขึ้น ส่วนระบบค้นคืนสารสนเทศแบบทั่วไปที่มีตัวดำเนินการเป็น OR ได้รับผลกระทบเช่นเดียวกัน โดยมีค่าความแม่นยำลดลง ในขณะที่ค่า E เพิ่มขึ้น ส่วนค่าความระลึกมีลักษณะคงที่คือมีค่าเท่ากับ 1 เนื่องจากระบบค้นคืนสารสนเทศได้ค้นคืนเอกสารขึ้นมาทั้งหมด โดยไม่ตัดเอกสารที่ไม่มีความเกี่ยวข้องออกไปเลยแม้แต่ฉบับเดียว ซึ่งถ้าอ้างอิงตามนิยามจะเห็นว่าค่าความแม่นยำ และค่าความระลึกยิ่งมากจะยิ่งดี แต่อันจริงแล้วต้องดูค่าทั้ง 2 ค่าประกอบกัน ซึ่งในกรณีนี้ถึงแม้ค่าความระลึกจะมีค่าสูงสุดตลอด ก็ไม่ถือว่าเป็นการดี เนื่องจากค่าความแม่นยำมีแนวโน้มลดลงเรื่อยๆ ตามจำนวนคำถามสืบค้นที่เพิ่มขึ้น

สำหรับระบบที่ได้ทำการวิจัยนั้น ค่าความแม่นยำมีแนวโน้มสูงขึ้นหากมีจำนวนคำถามสืบค้นมากขึ้น ส่วนค่าความระลึกมีแนวโน้มลดลงอยู่บ้างแต่ปลายกราฟยังคงมีลักษณะเชิดขึ้น ในขณะที่ค่า E ไม่ได้มีแนวโน้มว่าจะสูงขึ้นมากนัก แสดงว่าระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ได้ทำการวิจัย สามารถรักษาค่าความแม่นยำ ค่าความระลึก รวมถึงค่า E ไว้ในระดับที่สมดุลไม่ให้เพิ่มขึ้นมากจนเกินไปหรือลดลงมากจนเกินไป นั่นหมายถึงจำนวนคำถามสืบค้นมีผลกระทบกับระบบค้นคืนสารสนเทศแบบฮิวริสติคที่ได้ทำการวิจัยน้อยกว่าระบบอื่น

ในเรื่องของการค้นคืนสารสนเทศที่มีขนาดของฐานข้อมูลใหญ่กว่าที่ได้ทำการทดลองไว้นั้น ไม่มีปัญหาสำหรับระบบค้นคืนสารสนเทศที่ได้พัฒนานี้ เนื่องจากหลักการสร้างแลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets) เหมาะสำหรับฐานข้อมูลที่มีเอกสารจำนวนมากอยู่แล้ว ปัญหาที่อาจจะเกิดขึ้นกับบรรณารักษ์ คือเรื่องการกำหนดคำเฉพาะที่ใช้สืบค้นให้เหมาะสมในการหาความสัมพันธ์ระหว่างแต่ละเอกสาร เพื่อให้ได้การค้นคืนที่ตรงต่อความต้องการของผู้สืบค้นมากที่สุด

6.2 ข้อเสนอแนะเพื่อการวิจัยในครั้งต่อไป

สิ่งที่ผู้วิจัยขอเสนอแนะเพื่อการวิจัยในครั้งต่อไป มีดังนี้

1. การเพิ่มเติมส่วนของข้อมูลประวัติและความสนใจของผู้ใช้/ผู้สืบค้น ให้มีรายละเอียดเพิ่มมากขึ้น เพื่อที่ระบบจะได้นำข้อมูลที่ได้ ไปประมวลผลในการคัดเลือกเอกสารในฐานข้อมูลให้ตรงต่อความต้องการของผู้สืบค้นมากยิ่งขึ้น

2. ใช้เทคนิคการเรียงลำดับเอกสารเทคนิคอื่นนอกเหนือจากการให้น้ำหนักของเทอมตามทฤษฎีโครงข่ายเบย์

3. มีการใช้เทคนิคอื่นเข้ามาช่วยในการสืบค้น เช่น Neutal network หรือ Genetic Algorithm เป็นต้น

4. ประยุกต์ใช้การสืบค้นแบบฮิวริสติค โดยติดต่อฐานข้อมูลเอกสารผ่านเครือข่ายใยแมงมุม (World Wide Web)

เอกสารอ้างอิง

[1] Andrew, A. M. **Continuous Heuristic The Prelinguistic Basis of Intelligence.** New York : Ellis Horwood. 1993.

[2] Barr, A. and Feigenbaum, E. A. **The handbook of Artificial Intelligence Volume I.** 2nd Ed. Addison-Wesley Publishing Company,Inc. 1982.

[3] Bloemeke, M. "An Algorithm for the Recovery of Both Target Joint Beliefs and Full Belief from Bayesian Networks." ACM. 1998. pp. 136-142.

[4] Callan, J. P. "Fast Incremental Indexing for Full-Text Information retrieval." Proceedings of the 20th VLDB Conference. 1994. pp. 1-11.

[5] Chander,P.G. et.al. "An Expert System to Aid Cataloging and Searching Electronic Documents on Digital Libraries." Expert Systems With Applications,vol. 12,no. 4,1997. pp. 405-416.

[6] Chang, K.C., and Fung, R.M. "Node aggregation for distributed inference in Bayesian Networks." In Proceedings of IJCAI 89 (Detroit, Michi.). Morgan Kaufmann, San Mateo, Calif., 1989. pp. 265-270.

[7] Croft, W. B. "What Do People Want from Information Retrieval?." [Online]. Available : http://www.dlib.org/dlib/november95/11croft.html. 1995.

[8] Frakes, W.B. and Baeza-Yates, R. **Information retrieval Data Structures & Algorithms.** New Jersey : Prentice-Hall, Inc. 1992.

[9] Frants, V. I. et.al. **Automated Information Retrieval Theory and Methods.** New York : Academic Press. 1997.

[10] Fung, R and Del Favero, B. "Applying Bayesian Networks to Information Retrieval." Communication of the ACM, vol. 38, no. 3,March 1995. pp. 42-57.

[11] Hofferer, M. "Heuristic Search in Information Retrieval." **Information Retrieval : New Systems and Current Research**. London : Taylor Graham. 1994.

[12] Inversen, G.R. **Bayesian Statistical Inference.** 2nd Ed. California : Sage Publications. 1985.

[13] Korfhage, R. R.**Information Storage and Retrieval.** New York : Wiley Computer Publishing. 1997.

[14] Lancaster, F. W. **Information Retrieval Systems:Characteristics, Testing and Evaluation.** A Wiley-Interscience Publication. 1979.

[15] Lindley, D.V. **Introduction to Probability and Statistics from a Bayesian Viewpoint.** Cambridge : Cambridge University Press. 1969.

[16] Luger, G. F. and Stubblefield, W. A. **Artificial Intelligence Structures and Strategies for complex problem solving.** Redwood City, California : The Benjamin/Cummings Pubblishing Company, Inc. 1993.

[17] Patterson, D. W. **Introduction to Artificial Intelligence and Expert Systems.** New Jersey: Prentice-Hall, Inc. 1991.

[18] Ram, A. College of Computing Georgia Institute of Technology, Atlanta, Georgia [Online]. Available : http://www.netg.se/~kerfor/aikey.html. 1990-1993.

[19] Rich, E. and Knight, K. **Artificial Intelligence.** International Edition, 2nd Ed. Singapore : McGraw-Hill,Inc. 1991.

[20] Russell, S.J. and Norvig, P. **Artificial Intelligence A Modern Approach.** New Jersey : Prentice-Hall, Inc. 1995.

[21] Salton, G. and McGill, M.J. **Introduction to Modern Information Retrieval.** 2nd Ed. Singapore : McGraw-Hill, Inc. 1984.

[22] Van Rijsbergen, C. J. **Information retrieval.** 2nd Ed. London : Butterworths. 1979.

[23] Weitzel, J. R. and Kerschberg, L. "Developing knowledge-based systems: Reorganizing the system development cycle." Communication of the ACM, vol. 32 ,no. 4, 1989. pp.482-488.

[24] Zhang, W. and Korf, E.R. "Performance of linear-space search algorithms." Artificial Intelligence,vol. 79,no.2,1995. pp.275.

[25] Library of Congress. **Subject Headings.** 21st Ed. Washing DC. : Library of Congress, Cataloging Distribution Service. 1998.

[26] นิพนธ์ เจริญกิจการ. "การจัดเก็บและค้นคืนสารสนเทศ (Information Storage and Retrieval) ฉบับปรับปรุงครั้งที่ 1." [Online].Available : http://web.it.kmutt.ac.th/nipon/syllabus_temp.html. 2542.

[27] บุญธรรม กิจปรีดาบริสุทธิ์ (ผู้รวบรวม). "สูตรการคำนวณทางสถิติและตารางสถิติ." ภาควิชาศึกษาศาสตร์ คณะสังคมศาสตร์และมนุษยศาสตร์ มหาวิทยาลัยมหิดล. 2530.

ภาคผนวก

ภาคผนวก ก.

การจำลองระบบค้นคืนสารสนเทศตามทฤษฎีที่ได้ทำการวิจัย

ระบบค้นคืนสารสนเทศที่ได้จำลองขึ้น ตามทฤษฎีและหลักการต่างที่ได้ทำการวิจัย มีหน้าจอที่ใช้ติดต่อกับผู้สืบค้นเป็นลักษณะแบบกราฟฟิค (Graphic User Interface) ดังนี้

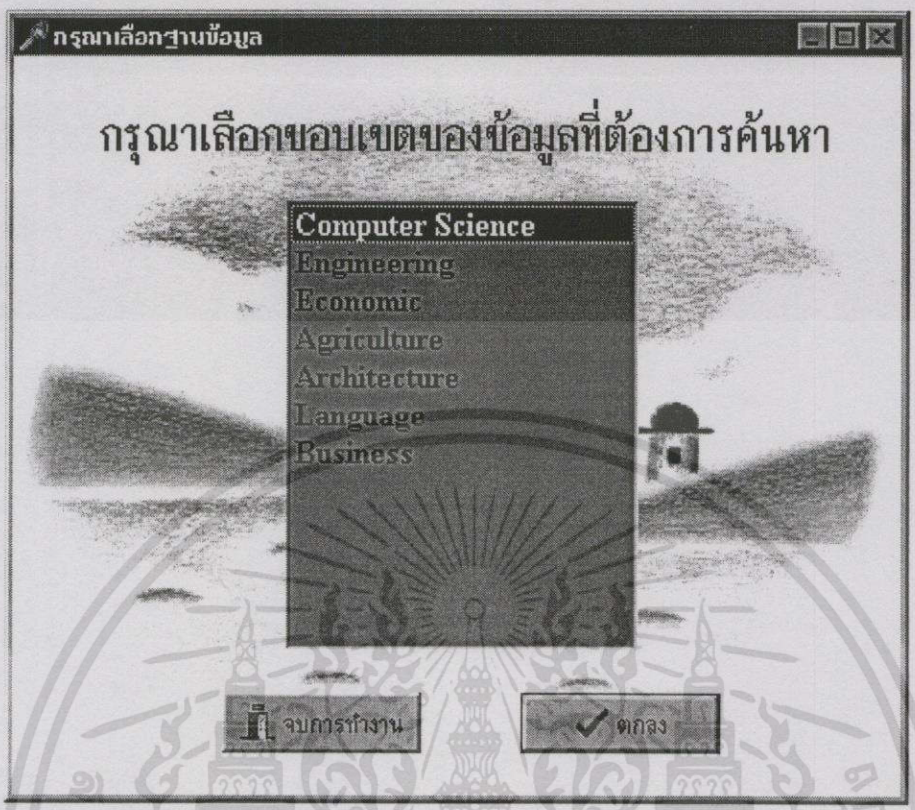
เริ่มต้นการเข้าสู่ระบบโดยการ Run ไฟล์ ชื่อ Retrieve.exe ปรากฏหน้าจอแรกที่เข้าสู่ระบบดังรูปที่ ก.1

รูปที่ ก.1 หน้าจอแรกของระบบเมื่อทำการ Run โปรแกรม

หน้าจอแรกมีปุ่มทั้งหมด 4 ปุ่ม

1. ปุ่ม "จบโปรแกรม" : ใช้เมื่อต้องการออกจากโปรแกรม
2. ปุ่ม "เกี่ยวกับ" : แสดงรายละเอียดเกี่ยวกับผู้พัฒนาระบบ
3. ปุ่ม "ตรวจสอบข้อมูล" : สำหรับบรรณารักษ์ หรือผู้ดูแลฐานข้อมูลเอกสารของระบบควบคุมโดยใช้รหัสผ่าน
4. ปุ่ม "สืบค้นข้อมูล" : สำหรับผู้สืบค้นใช้สืบค้นเอกสาร

ผู้สืบค้นกดปุ่ม "สืบค้นข้อมูล" ปรากฏหน้าจอถัดไปดังนี้



รูปที่ ก.2 หน้าจอให้เลือกฐานข้อมูล

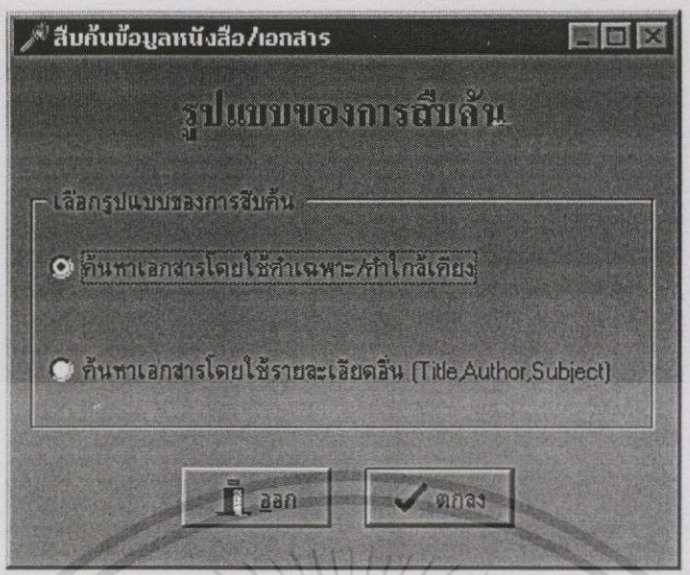
จากรูปที่ ก.2 ผู้สืบค้นต้องทำการเลือกฐานข้อมูล หรือขอบเขตของเอกสารที่ต้องการสืบค้น เนื่องจากฐานข้อมูลเอกสารจะแบ่งเป็น 7 ชนิด จากรูปผู้สืบค้นเลือกขอบเขตของการสืบค้นเป็นทางด้านวิทยาการคอมพิวเตอร์ (Computer Science) จากนั้นกดปุ่ม "ตกลง" ปรากฏหน้าจอถัดไปดังรูปที่ ก.3

จากรูปที่ ก.3 จะปรากฏรูปแบบของการสืบค้นให้ผู้สืบค้นเลือกคือ

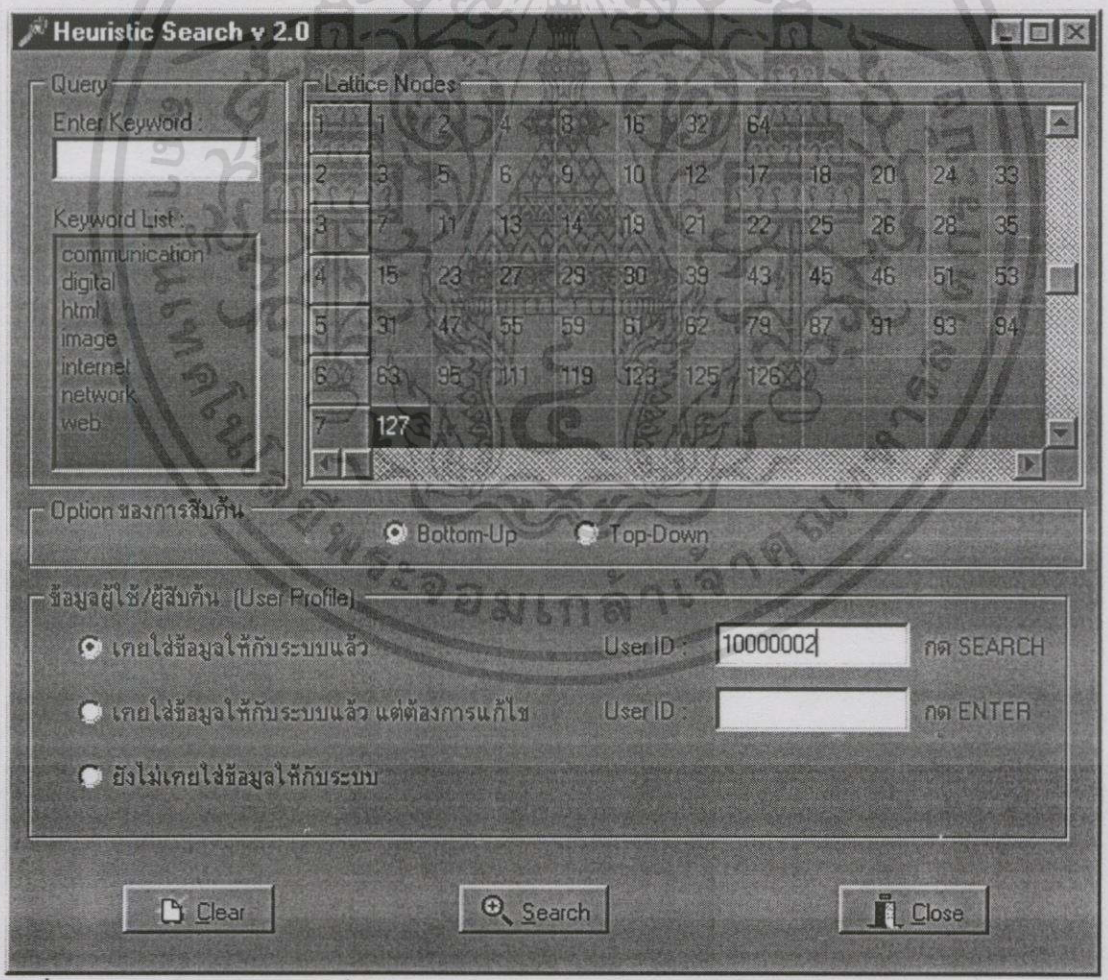
ตัวเลือกที่ 1 ค้นหาเอกสารโดยใช้คำเฉพาะ หรือคำใกล้เคียง และรวมไปถึงคำย่อด้วย

ตัวเลือกที่ 2 ค้นหาเอกสารโดยใช้รายละเอียดอื่น เช่น ชื่อเอกสาร (Title) ชื่อผู้แต่ง (Author) หรือ ชื่อสาขาวิชาที่เกี่ยวข้อง/หัวเรื่อง (Subject) เป็นต้น

ในกรณีแรก ผู้สืบค้นเลือกตัวเลือกที่ 1 เมื่อกดปุ่ม "ตกลง" ปรากฏหน้าจอถัดไปดังรูปที่ ก.4



รูปที่ ก.3 หน้าจอการเลือกรูปแบบของการสืบค้น กรณีเลือกค้นหาเอกสารโดยใช้คำเฉพาะ/คำใกล้เคียง และคำย่อ



รูปที่ ก.4 หน้าจอใส่คำสืบค้นที่เป็นลักษณะของคำเฉพาะ และใส่ข้อมูลประวัติและความสนใจของผู้สืบค้น

จากรูปที่ ก.4 ผู้สืบค้นได้ใส่คำสืบค้นตรงช่องว่างใต้คำว่า "Enter Keyword :" เมื่อใส่แล้วกดคีย์ Enter ระบบจะรับคำดังกล่าวไปทำการสร้างแลททิสของซับเซ็ตของเอกสาร (Lattice of document subsets) ตามขั้นตอนที่ได้กล่าวไว้แล้วในหัวข้อที่ 4.1.2.1 ส่วนของอัลกอริทึมในการสร้างแลททิสโหนด โดยในชุดคำถามสืบค้นนี้ ผู้สืบค้นได้ใส่คำสืบค้นทั้งหมดจำนวน 7 คำ เรียงลำดับดังนี้

1. network 2. communication 3. image
4. digital 5. internet 6. web 7. html

ระบบสร้างโหนดได้ทั้งหมด 127 โหนด ตามสูตรที่ได้กล่าวไว้ในอัลกอริทึมในการสร้างแลททิสโหนด

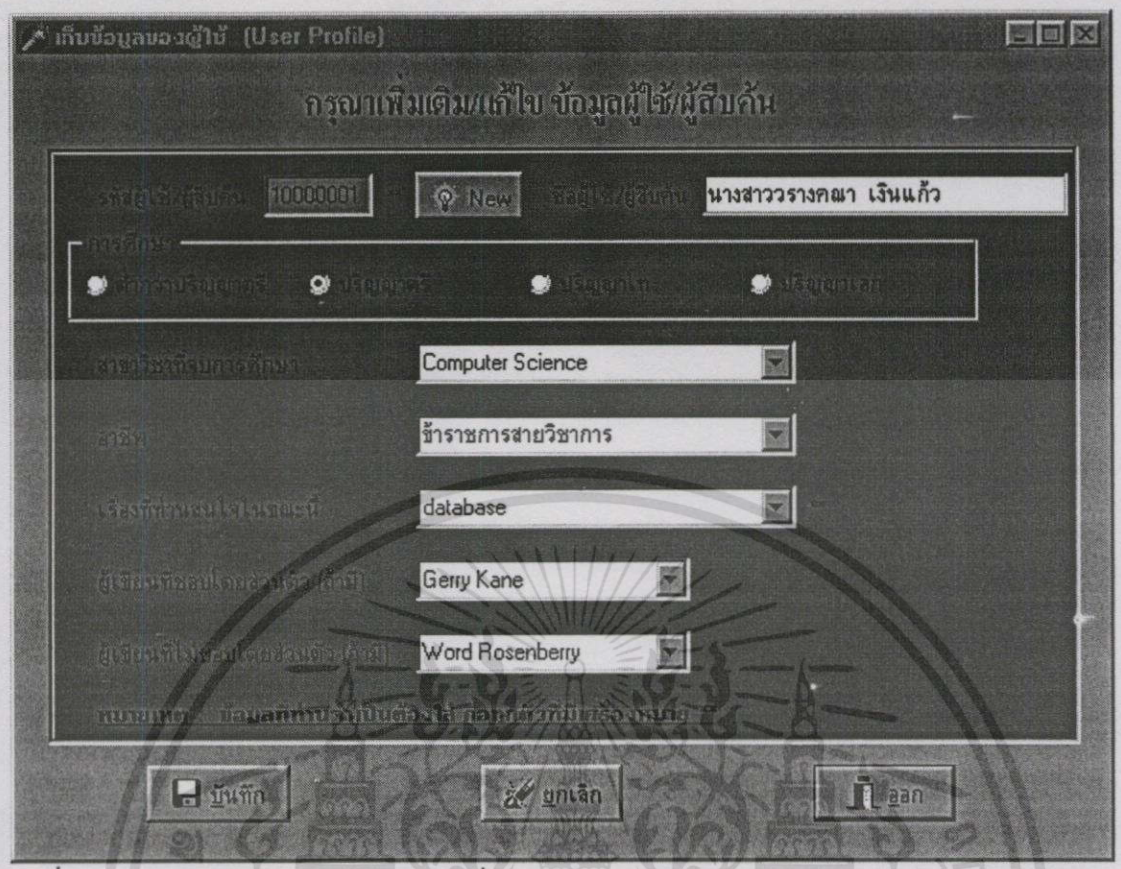
จากนั้นผู้สืบค้นใส่ข้อมูลในส่วนของข้อมูลผู้ใช้/ผู้สืบค้น (User Profile) ซึ่งมีตัวเลือกให้เลือก 3 แบบ

ตัวเลือกที่ 1 เคยใส่ข้อมูลให้กับระบบแล้ว : ผู้สืบค้นเพียงแค่ใส่รหัสของตนเองแล้วกดปุ่ม "Search"

ตัวเลือกที่ 2 เคยใส่ข้อมูลให้กับระบบแล้วแต่ต้องการแก้ไข : กรณีนี้ใช้สำหรับแก้ไขหรือเปลี่ยนแปลงข้อมูลของผู้สืบค้น หากผู้สืบค้นเลือกตัวเลือกที่ 2 จะปรากฏหน้าจอดังนี้

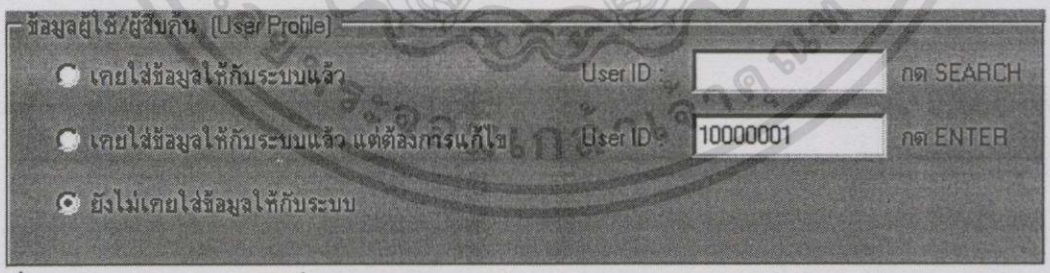
รูปที่ ก.5 การเลือกตัวเลือก เมื่อผู้สืบค้นต้องการแก้ไขข้อมูลประวัติและความสนใจที่เคยใส่ไว้แล้วในฐานข้อมูล

เมื่อกดคีย์ Enter แล้วจะปรากฏข้อมูลของผู้สืบค้นรหัส 10000001 ออกมาเพื่อทำการแก้ไขและบันทึก ดังรูปที่ ก.6

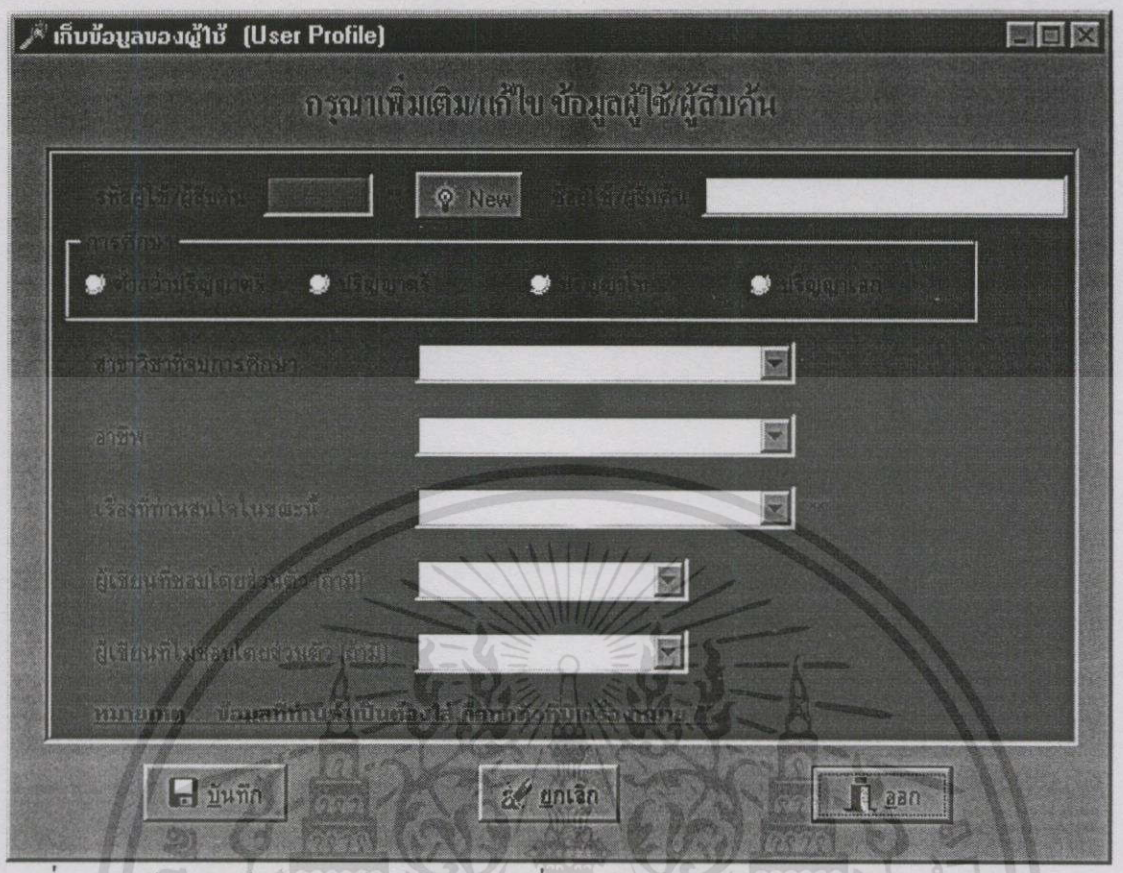


รูปที่ ก.6 หน้าจอให้ผู้สืบค้นแก้ไขหรือเปลี่ยนแปลงข้อมูลของตนเอง

ตัวเลือกที่ 3 ยังไม่เคยใส่ข้อมูลให้กับระบบ : กรณีนี้ใช้สำหรับผู้สืบค้นที่ต้องการบันทึกข้อมูลประวัติและความสนใจของตนเอง ลงฐานข้อมูลของผู้สืบค้นเพื่อการค้นคืนครั้งต่อไป หากเลือกตัวเลือกนี้จะปรากฏหน้าจอดังรูปที่ ก.7 และ รูปที่ ก.8



รูปที่ ก.7 การเลือกตัวเลือก เมื่อผู้สืบค้นต้องการใส่ข้อมูลประวัติและความสนใจของตนเองเข้าสู่ฐานข้อมูลของระบบ



รูปที่ ก.8 หน้าจอใส่ข้อมูลของผู้สืบค้นในกรณีที่ไม่เคยใส่ข้อมูลลงในระบบมาก่อน

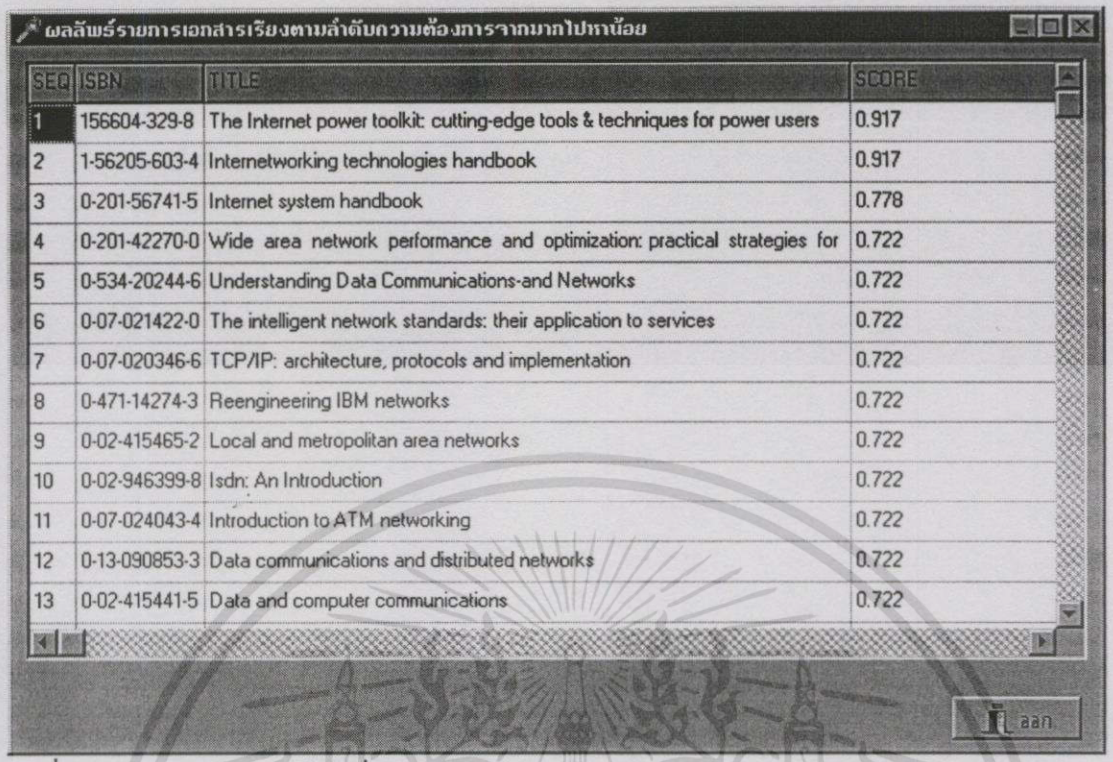
กดปุ่ม "New" เพื่อให้ระบบทำการ Run เลขรหัสของผู้สืบค้นให้

เนื่องจากตัวอย่างของการสืบค้นข้างต้นผู้สืบค้นใส่ข้อมูลเป็นที่เรียบร้อยแล้วโดยมีรหัสเป็น 10000002 (ผู้สืบค้นได้ใส่ว่าสนใจเรื่องเกี่ยวกับ Network) จึงย้อนกลับไปที่รูปที่ ก.4 ดังนั้นเมื่อกดปุ่ม "Search" แล้วจะปรากฏหน้าจอคำถามดังรูปที่ ก.9

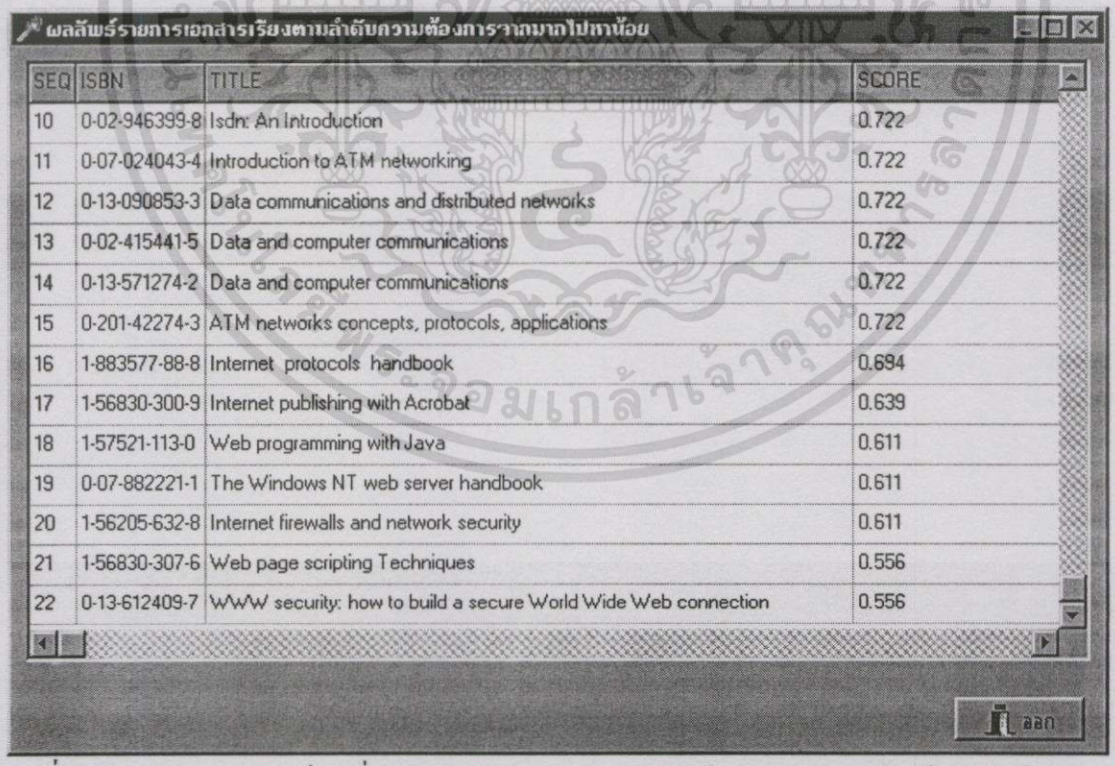
จากรูปที่ ก.9 ระบบจะสอบถามผู้สืบค้นในเรื่องของการให้ความสำคัญของปีที่พิมพ์ เพื่อนำมาเรียงลำดับเอกสาร ว่า "ปีที่พิมพ์สำคัญต่อการเรียงลำดับเอกสารหรือไม่" และมีตัวเลือก 3 ตัวเลือกดังที่แสดงในรูป ในที่นี้ผู้สืบค้นไม่ต้องการเรียงลำดับตามปีที่พิมพ์ จึงเลือกตัวเลือกที่ 1

รูปที่ ก.9 หน้าจอคำถามที่ระบบถามผู้สืบค้น เรื่องของการให้ความสำคัญของปีในการเรียงเอกสาร

จากนั้นจึงกดปุ่ม "ตกลง" หน้าจอของผลลัพธ์จึงปรากฏขึ้นดังรูปที่ ก.10 และ รูปที่ ก.11 โดยจากรูปทั้งสอง จะปรากฏรายชื่อเอกสาร และเลข ISBN รวมถึงน้ำหนักที่ได้จากการเรียงเอกสารด้วย ซึ่งอันที่จริงแล้วหากนำระบบไปใช้จริง น้ำหนักนี้จะไม่แสดงให้ผู้สืบค้นได้เห็น

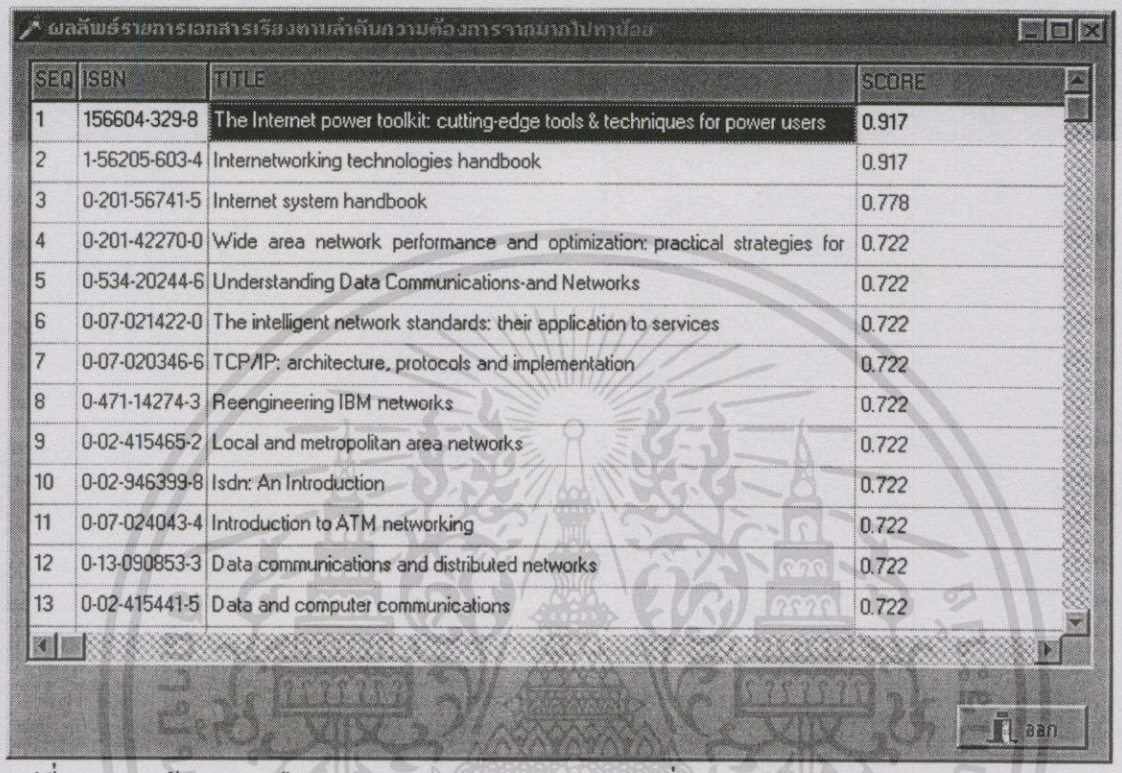


รูปที่ ก.10 หน้าจอผลลัพธ์รายชื่อเอกสารเรียงตามลำดับความต้องการของการสืบค้นโดยใช้คำเฉพาะ7 คำ หน้าที่ 1



รูปที่ ก.11 หน้าจอผลลัพธ์รายชื่อเอกสารเรียงตามลำดับความต้องการของการสืบค้นโดยใช้คำเฉพาะ 7 คำหน้าที่ 2

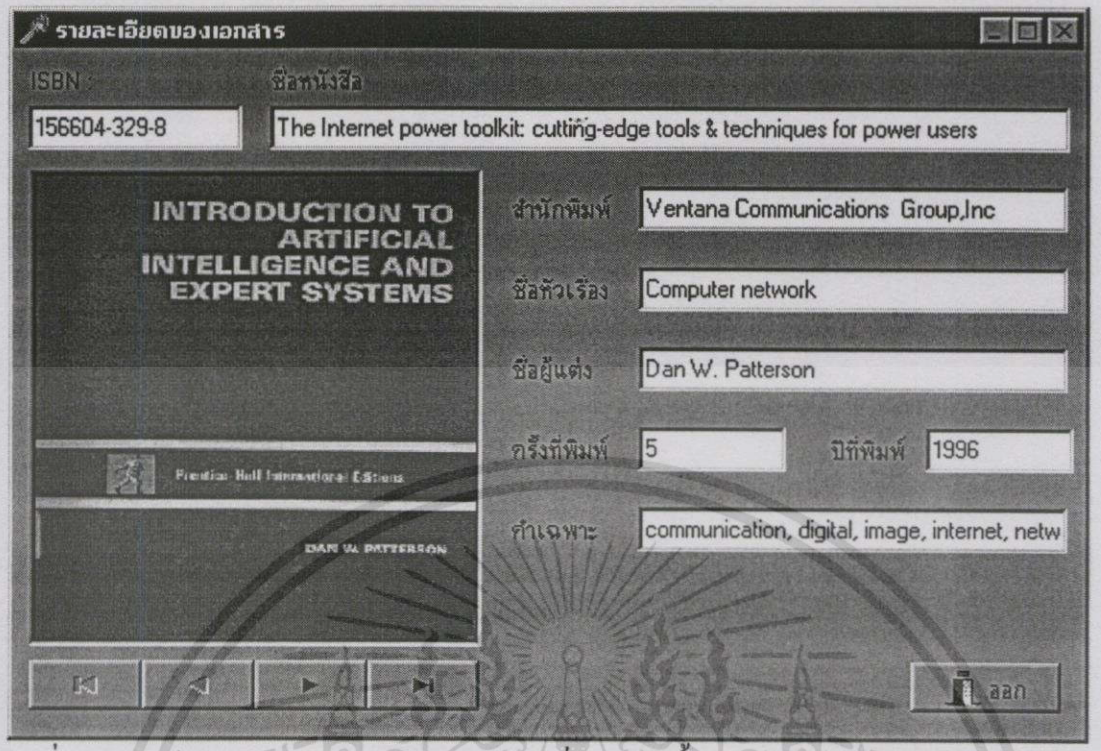
นอกจากผู้สืบค้นได้ทราบรายชื่อเอกสารที่เป็นผลลัพธ์ของการสืบค้นแล้ว ผู้สืบค้นยังสามารถดับเบิ้ลคลิ๊ก (Double Click) เข้าไปดูรายละเอียดของเอกสารพร้อมรูปหน้าปกและสารบัญได้อีกด้วย เพื่อให้เห็นภาพชัดเจนผู้วิจัยจึงขอยกตัวอย่างดังรูปที่ ก.12 และ รูปที่ ก.13 ดังนี้



SEQ	ISBN	TITLE	SCORE
1	156604-329-8	The Internet power toolkit: cutting-edge tools & techniques for power users	0.917
2	1-56205-603-4	Internetworking technologies handbook	0.917
3	0-201-56741-5	Internet system handbook	0.778
4	0-201-42270-0	Wide area network performance and optimization: practical strategies for	0.722
5	0-534-20244-6	Understanding Data Communications-and Networks	0.722
6	0-07-021422-0	The intelligent network standards: their application to services	0.722
7	0-07-020346-6	TCP/IP: architecture, protocols and implementation	0.722
8	0-471-14274-3	Reengineering IBM networks	0.722
9	0-02-415465-2	Local and metropolitan area networks	0.722
10	0-02-946399-8	Isdn: An Introduction	0.722
11	0-07-024043-4	Introduction to ATM networking	0.722
12	0-13-090853-3	Data communications and distributed networks	0.722
13	0-02-415441-5	Data and computer communications	0.722

รูปที่ ก.12 หน้าจอการเลือกดูรายละเอียดของเอกสารฉบับที่ 1 ของผู้สืบค้น

รูปที่ ก.12 แสดงให้เห็นแถบสีน้ำเงินคลุมชื่อเอกสาร หมายถึง ผู้สืบค้นต้องการเลือกดูรายละเอียดของเอกสารลำดับที่ 1 เมื่อผู้สืบค้นทำการดับเบิ้ลคลิ๊ก (Double Click) ตรงชื่อเอกสารฉบับที่ 1 แล้ว จะปรากฏหน้าจอถัดไปดังรูปที่ ก.13

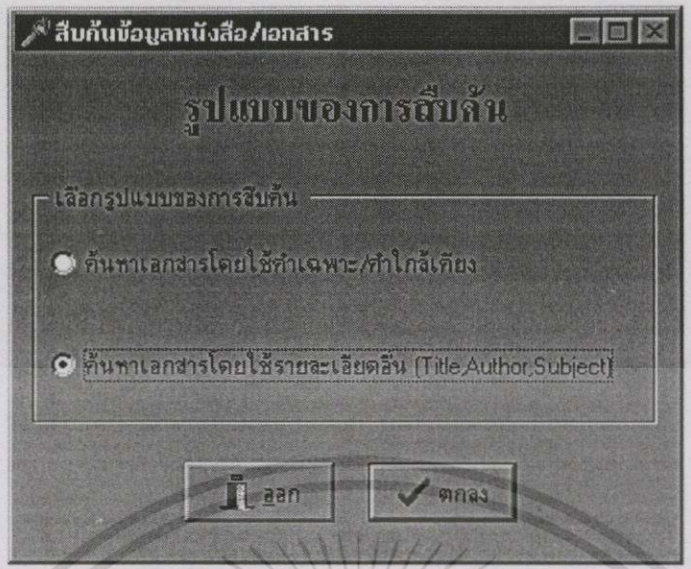


รูปที่ ก.13 หน้าจอรายละเอียดของเอกสารลำดับที่ 1 พร้อมทั้งภาพเอกสาร

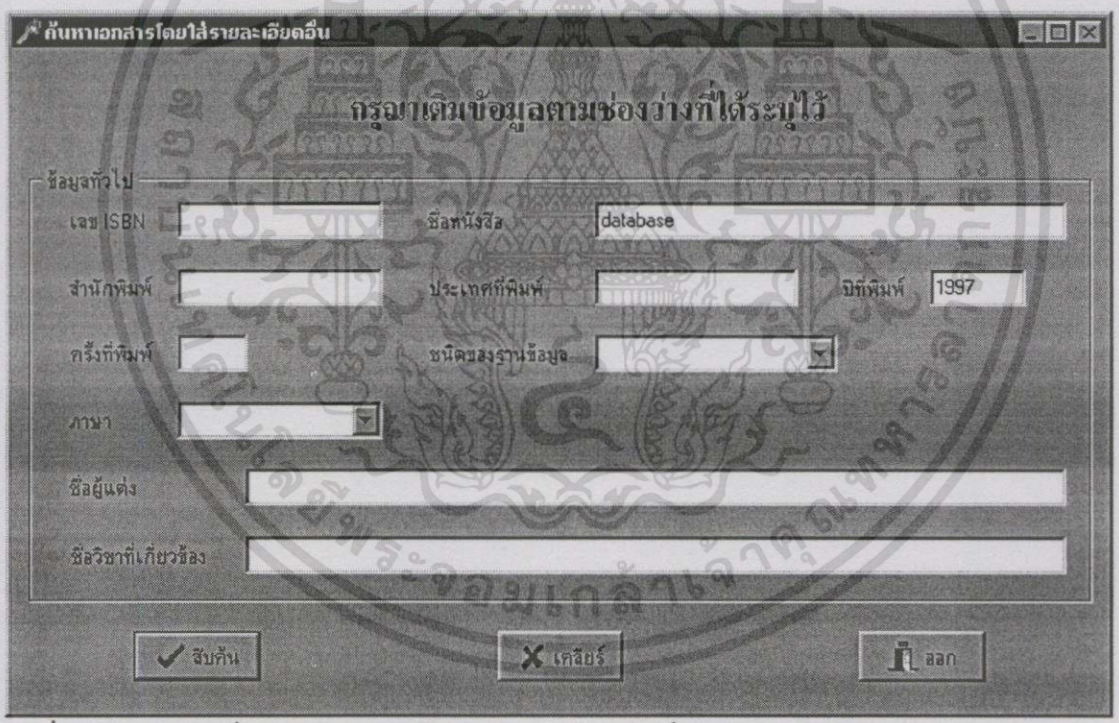
จากรูปที่ ก.13 ผู้สืบค้นจะทราบรายละเอียดทั้งหมดในเอกสาร พร้อมทั้งรูปหน้าปกเอกสาร และสารบัญ

จากรูปที่ ก.3 ที่ได้แสดงไปแล้วในข้างต้น ผู้สืบค้นขอกล่าวถึงกรณีที่ผู้สืบค้นเลือกรูปแบบของการสืบค้น ตัวเลือกที่ 2 คือการค้นหาเอกสารโดยใช้รายละเอียดอื่น เช่น ชื่อเอกสาร (Title) ชื่อผู้แต่ง (Author) หรือ ชื่อสาขาวิชาที่เกี่ยวข้อง/หัวเรื่อง (Subject) เป็นต้น ดังรูปที่ ก. 14

เมื่อกดปุ่ม "ตกลง" จะปรากฏหน้าจอถัดไปดังรูปที่ ก.15 ซึ่งเป็นหน้าจอที่ให้ผู้สืบค้นใส่รายละเอียดที่ต้องการค้นหา สามารถเติมช่องว่างได้ทุกช่อง หากต้องการ

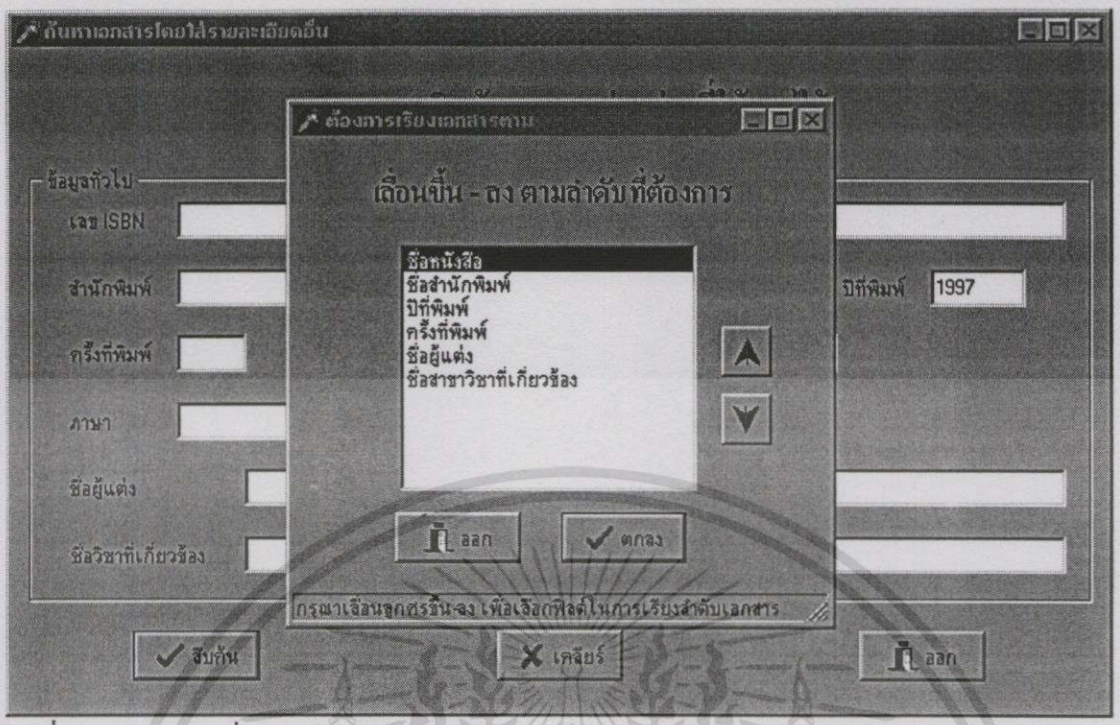


รูปที่ ก.14 แสดงหน้าจอการเลือกรูปแบบของการสืบค้น กรณีเลือกค้นหาเอกสารโดยใช้รายละเอียดอื่น



รูปที่ ก.15 หน้าจอที่ให้ผู้สืบค้นใส่รายละเอียดของเอกสารที่ต้องการสืบค้น

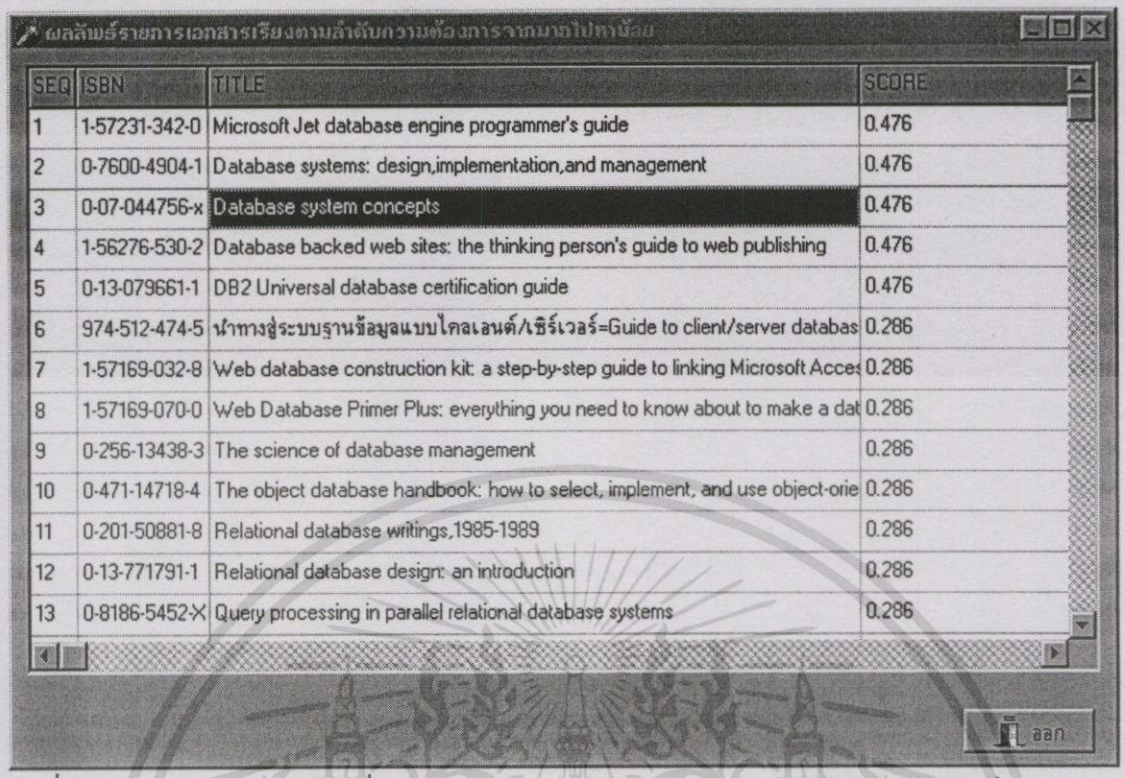
ในตัวอย่างรูปที่ ก.15 นี้ สมมติว่าผู้สืบค้น ต้องการเอกสารที่เกี่ยวกับ “Database” และต้องการปีที่พิมพ์เป็นปี “1997” ด้วย เมื่อกดปุ่ม “สืบค้น” แล้วระบบจะถามคำถามดังรูปที่ ก.16



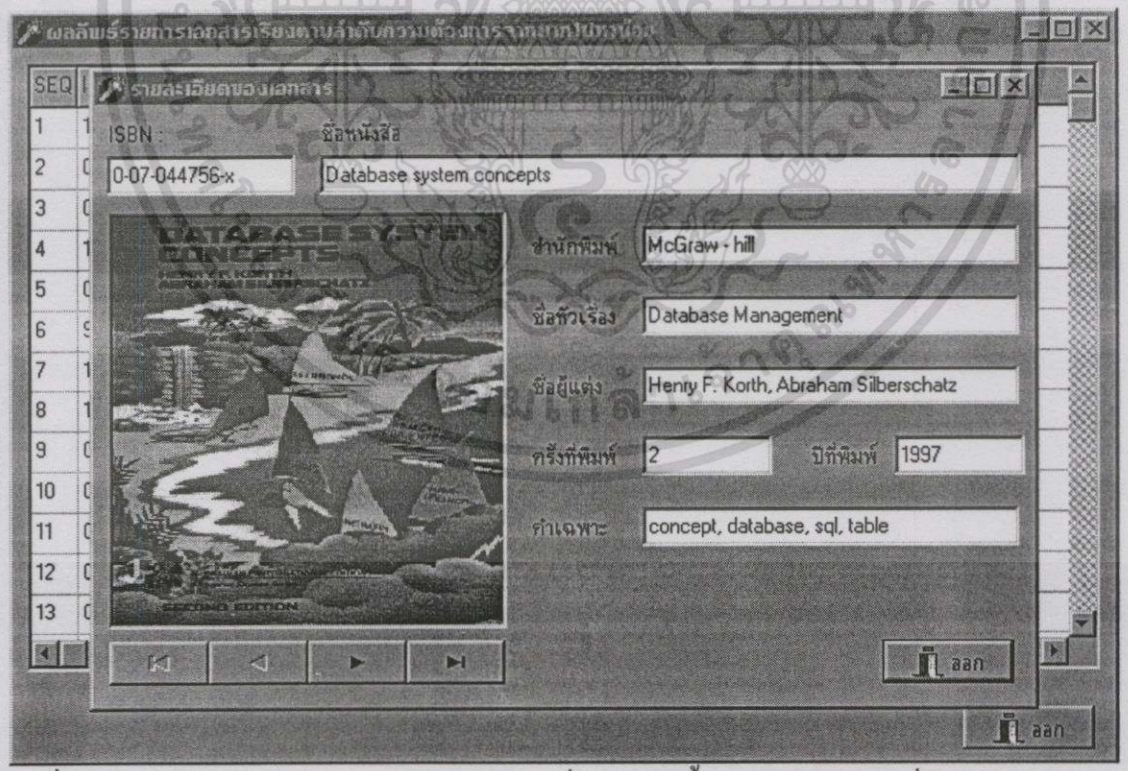
รูปที่ ก.16 หน้าจอที่ระบบถามผู้สืบค้นว่าต้องการเรียงเอกสารตามอะไร

จากรูปที่ ก.16 เห็นได้ว่าระบบจะถามผู้สืบค้นว่าต้องการเรียงเอกสารตามอะไร ให้เลื่อนแถบน้ำเงินขึ้น-ลง โดยใช้ลูกศรทางขวามือ ได้ตามความต้องการ สมมติว่าผู้สืบค้นต้องการเรียงเอกสารตาม ชื่อหนังสือ ชื่อสำนักพิมพ์ ปีที่พิมพ์ ตามลำดับ

เมื่อผู้สืบค้นกดปุ่ม "ตกลง" แล้วจะปรากฏผลลัพธ์ เป็นรายชื่อเอกสารดังรูปที่ ก.17 จากรายการผลลัพธ์ดังรูป ผู้สืบค้นสนใจดูรายละเอียดของเอกสารลำดับที่ 3 ชื่อ "Database System Concepts" เมื่อดับเบิ้ลคลิ๊ก (Double Click) ตรงชื่อเอกสาร จะปรากฏหน้าจอของรายละเอียดเอกสารเล่มนี้ดังรูปที่ ก.18

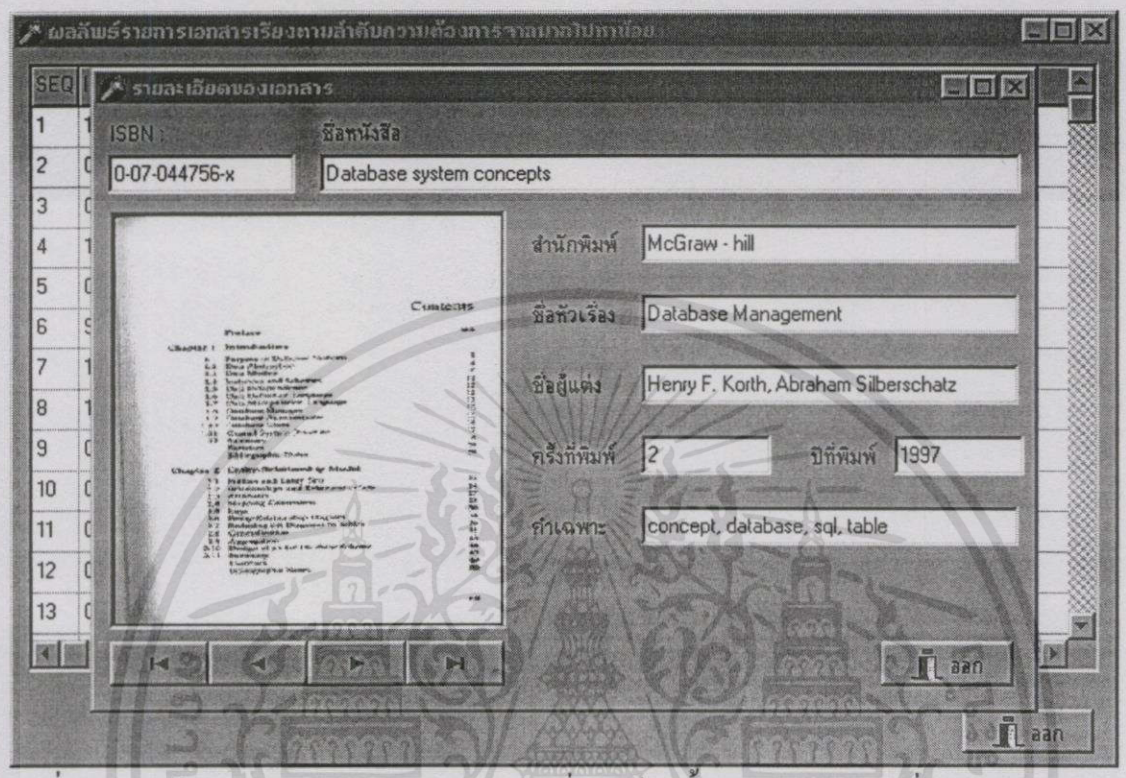


รูปที่ ก.17 หน้าจอผลลัพธ์รายชื่อเอกสารเรียงตามลำดับความต้องการ โดยใช้คำสืบค้นเป็น "Database" และ "1997"



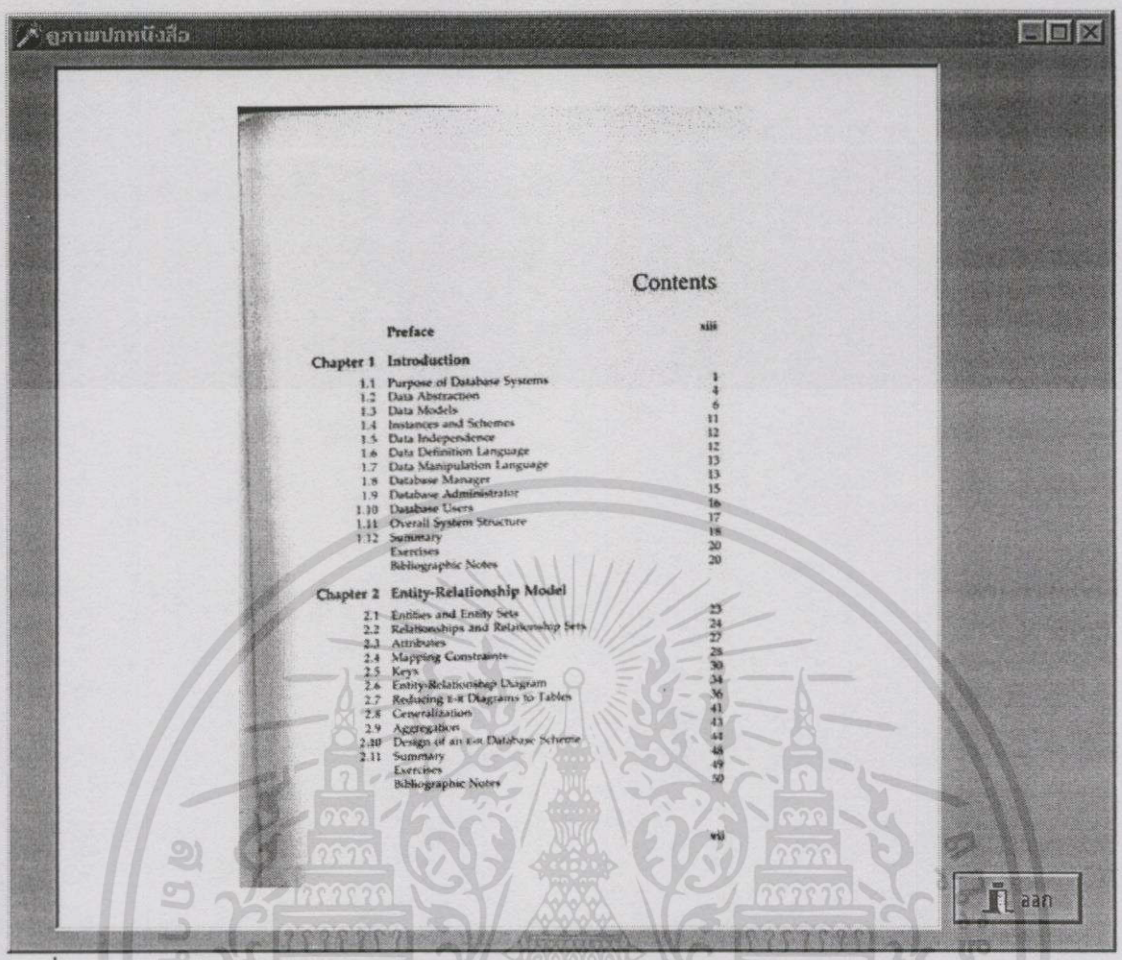
รูปที่ ก.18 หน้าจอรายละเอียดของเอกสารลำดับที่ 3 พร้อมทั้งภาพเอกสาร ภาพที่1

หากผู้สืบค้นต้องการดูรูปของเอกสารรูปถัดไป สามารถกดแถบคำสั่งด้านล่างรูปภาพที่มีลักษณะเป็นรูปลูกศรชี้ไปทางขวา จะปรากฏผลลัพธ์ดังรูปที่ ก.19



รูปที่ ก.19 หน้าจอรายละเอียดของเอกสารลำดับที่ 3 พร้อมทั้งภาพเอกสาร ภาพที่ 2

นอกจากนั้น ผู้สืบค้นยังสามารถดูภาพขยายได้โดยคลิ๊ก (Click) ตรงรูปภาพเอกสาร จะปรากฏภาพขยายของเอกสาร ดังรูปที่ ก.20

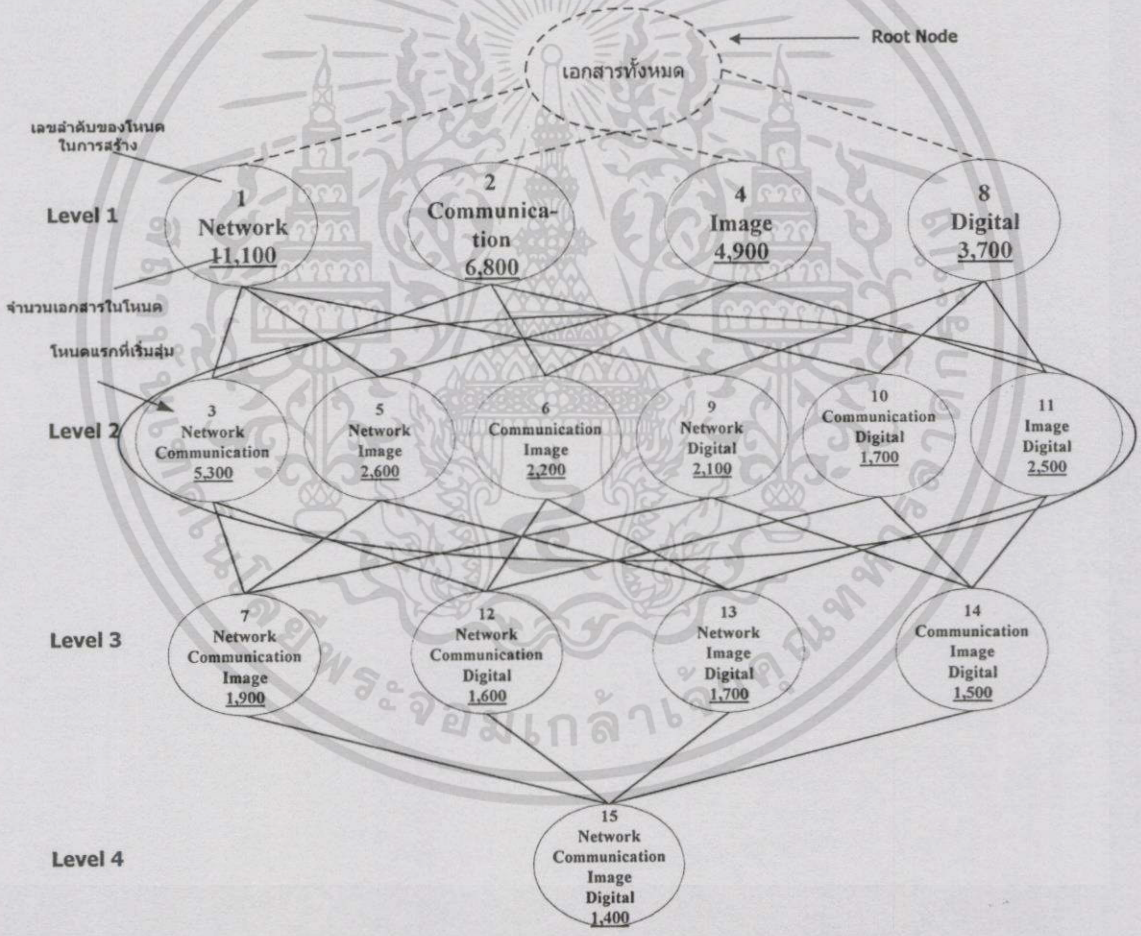


รูปที่ ก.20 แสดงรูปภาพเอกสารแบบขยาย

ภาคผนวก ข.

ตัวอย่างวิธีการคำนวณค่า $f^*(n)$ เพื่อเลือกโหนดที่จะทำการสืบค้น

ค่า $f^*(n)$ คำนวณเพื่อหาโหนดที่จะทำการสืบค้นถัดไป ซึ่งผู้วิจัยได้กล่าวไปแล้วในบทที่ 3 และ บทที่ 4 ดังนั้นในภาคผนวกนี้ เพื่อความเข้าใจที่สอดคล้องกัน ผู้วิจัยขอแสดงวิธีการคำนวณค่า $f^*(n)$ โดยการยกตัวอย่างแลททิสโหนดที่ได้จากการทดลองจริง โดยนำจำนวนเอกสารในโหนดแต่ละโหนด ที่ทำการทดลองคูณด้วย 100 ให้ได้จำนวนเอกสารในแต่ละโหนดมากขึ้น เพื่อทำการสุ่มตัวอย่างของเอกสารต่อไป แลททิสโหนดของคำที่นำมายกตัวอย่างมีลักษณะดังรูปที่ ข.1



รูปที่ ข.1 Level ของการเริ่มสุ่มเลือกโหนด และจำนวนเอกสารในแต่ละโหนด ของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

รูปที่ ข.1 แสดง Level ของการเริ่มสุ่มเลือกโหนด จากแลททิสโหนดของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

หลักการสุ่มเอกสารใช้สูตร การหาขนาดกลุ่มตัวอย่างของประชากร [27] (Sample size = n) ดังนี้

$$n = \frac{N}{1 + Ne^2}$$

โดยที่

ประชากรในที่นี้หมายถึงเอกสาร

n คือขนาดของกลุ่มตัวอย่างเอกสาร

N คือจำนวนเอกสารทั้งหมดในโหนดที่ทำการพิจารณาในขณะใดขณะหนึ่ง

e คือค่าความคลาดเคลื่อนของกลุ่มตัวอย่าง

ถ้าจำนวนเอกสารทั้งหมดในโหนดมีน้อยกว่า 1,500 เอกสาร ที่ความคลาดเคลื่อน 3% อาจไม่ต้องทำการสุ่ม

ในกรณีตัวอย่างนี้ ให้ค่าความคลาดเคลื่อนเป็น 3% แสดงว่ามีค่าความเชื่อมั่น 97%

<u>รอบการทำงานที่ 1</u>

1.1 โหนดที่จะพิจารณาอยู่ในเซ็ต OPEN = {3, 5, 6, 9, 10, 11} ตามที่อยู่ในบริเวณวงรีดังรูปที่ ข.1

1.2 ทำการเลือกโหนดที่จะสืบค้นโดยพิจารณาค่า $f^*(n)$ แต่เนื่องจากรอบแรกนี้ยังไม่มีโหนดใดที่ถูกประเมิน ค่า $f^*(n)$ จึงไม่สามารถคำนวณได้ เราจึงเลือกโหนดซ้ายสุดก่อน ได้แก่ โหนดหมายเลข 3

1.3 ทำการคำนวณหาขนาดของกลุ่มตัวอย่างเอกสาร ได้ดังนี้

$$n = \frac{5{,}300}{1 + 5{,}300 \times \left(0.03^2\right)}$$

$$= 919 \text{ ฉบับ}$$

ทำการสุ่มเอกสารจากโหนดหมายเลข 3 จำนวน 919 ฉบับ เปรียบเทียบกับ User Profile แล้วตรง 555 ฉบับ *(การสุ่มตัวอย่างแลททิสโหนด : Lattice - node - sampling)*

1.4 จากนั้นทำการปรับค่า *Ret#* และ *Rel#* ของโหนดที่เป็นกลุ่มตัวอย่างที่แท้จริง ของโหนดที่ 3 (ได้แก่โหนด 7, 12, 15) *(การสุ่มตัวอย่างที่แท้จริง : Actual - sampling)* สามารถแสดงการคำนวณได้ดังนี้

- <u>โหนดหมายเลข 7</u> จากเอกสารที่สุ่มจากโหนดหมายเลข 3 (919 ฉบับ) ปรากฏในโหนดหมายเลข 7 จำนวน 300 ฉบับ ($Ret\# = 300$) และได้ถูกประเมินจากโหนดหมายเลข 3 ว่าตรงความต้องการเป็นจำนวน 200 ฉบับ ($Rel\# = 200$) คำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(7) = g^*(7) + h^*(7)$$
$$= \frac{Ret\#}{Doc\#} + \left(1 - \frac{Rel\#}{Ret\#}\right)$$
$$= \frac{300}{1{,}900} + \left(1 - \frac{200}{300}\right)$$
$$= 0.491$$

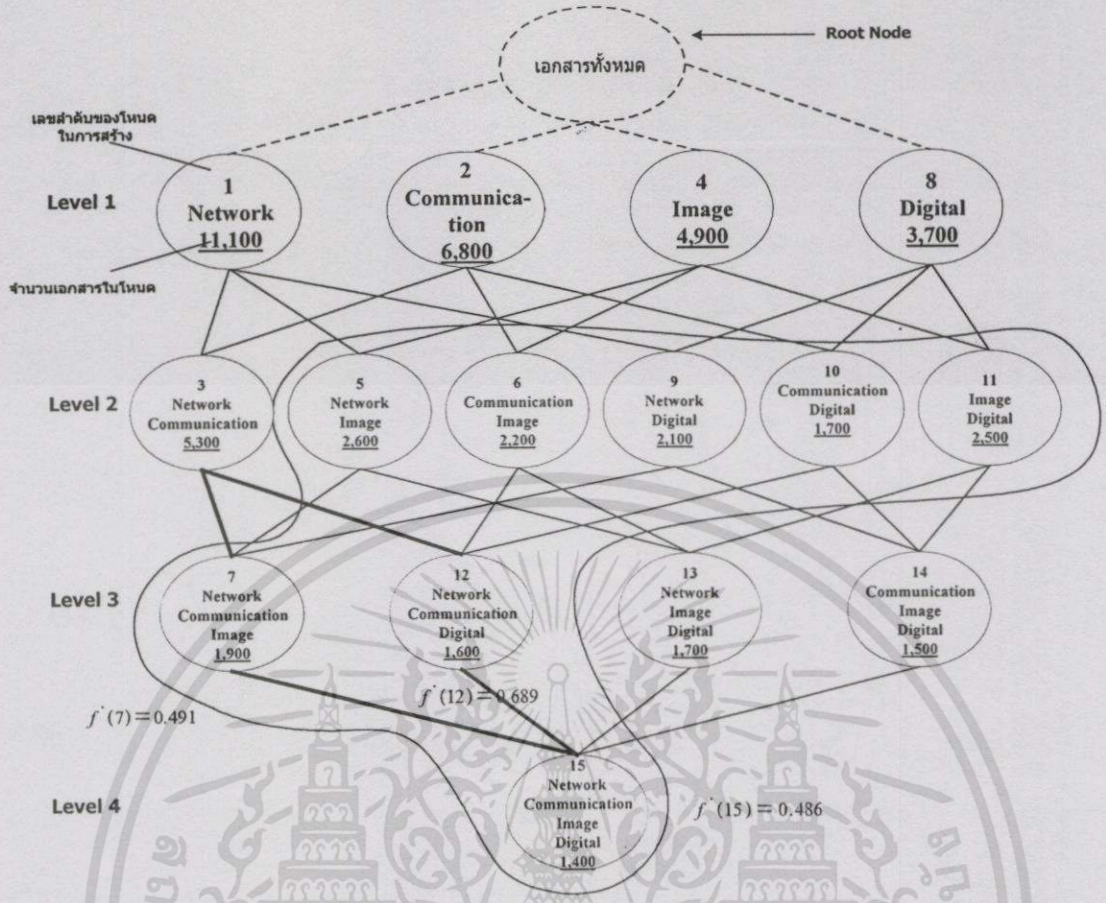
- <u>โหนดหมายเลข 12</u> ในทำนองเดียวกัน เอกสารที่เป็นกลุ่มตัวอย่างที่แท้จริงในโหนดหมายเลข 12 จำนวน 356 ฉบับ ($Ret\# = 356$) และตรงความต้องการ 190 ฉบับ ($Rel\# = 190$) คำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(12) = g^*(12) + h^*(12)$$
$$= \frac{356}{1{,}600} + \left(1 - \frac{190}{356}\right)$$
$$= 0.689$$

- <u>โหนดหมายเลข 15</u> สำหรับโหนดหมายเลข 15 มีเอกสารตรงกับโหนดหมายเลข 3 จำนวน 280 ฉบับ ($Ret\# = 280$) และมีเอกสารที่ตรงความต้องการจำนวน 200 ฉบับ ($Rel\# = 200$)

$$f^*(15) = g^*(15) + h^*(15)$$
$$= \frac{280}{1{,}400} + \left(1 - \frac{200}{280}\right)$$
$$= 0.486$$

หลังจากปรับค่าความเกี่ยวข้องและคำนวณค่า $f^*(n)$ ในโหนดที่เกี่ยวข้องกับโหนดหมายเลข 3 ที่อยู่ใน Level ล่างแล้ว จากนั้นอัลกอริทึม IRA จะทำการรวมโหนดหมายเลข 3 ไว้ใน CLOSED เซ็ต สามารถแสดงได้ดังรูปที่ ข.2



รูปที่ ข.2 แสดงการคำนวณหาค่า $f^*(n)$ ของโหนดหมายเลข 3 และการปรับค่า $f^*(n)$ ของโหนดหมายเลข 7 12 และ 15 ของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

<u>รอบการทำงานที่ 2</u>

2.1 โหนดที่จะพิจารณาอยู่ในเซ็ต OPEN = {5, 6, 7, 9, 10, 11, 12, 15} ตามที่อยู่ในบริเวณวงรีดังรูปที่ ข.2

2.2 ทำการเลือกโหนดที่จะสืบค้นโดยพิจารณาค่า $f^*(n)$ แต่เนื่องจากยังมีโหนดที่ยังไม่สามารถคำนวณค่า $f^*(n)$ เราจึงเลือกโหนดซ้ายสุดก่อน ได้แก่โหนดหมายเลข 5

2.3 ทำการคำนวณหาขนาดของกลุ่มตัวอย่างเอกสาร ได้ดังนี้

$$n = \frac{2,600}{1 + 2,600 \times \left(0.03^2\right)}$$

$$= 779 \text{ ฉบับ}$$

ทำการสุ่มเอกสารจากโหนดหมายเลข 5 จำนวน 779 ฉบับ เปรียบเทียบกับ User Profile แล้วตรง 450 ฉบับ *(การสุ่มตัวอย่างแลททิสโหนด : Lattice node - sampling)*

2.4 จากนั้นทำการปรับค่า *Ret#* และ *Rel#* ของโหนดที่เป็นกลุ่มตัวอย่างที่แท้จริง ของโหนดหมายเลข 5 (ได้แก่โหนด 7, 13, 15) *(การสุ่มตัวอย่างที่แท้จริง : Actual - sampling)* สามารถแสดงการคำนวณได้ดังนี้

- <u>โหนดหมายเลข 7</u> มีเอกสารตรงกับเอกสารที่สุ่มจากโหนดหมายเลข 5 จำนวน 105 ฉบับ (*Ret#* = 300+105) และได้ถูกประเมินว่าตรงความต้องการ 75 ฉบับ (*Rel#* = 200+75) คำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(7) = g^*(7) + h^*(7)$$
$$= \frac{405}{1,900} + \left(1 - \frac{275}{405}\right)$$
$$= 0.534$$

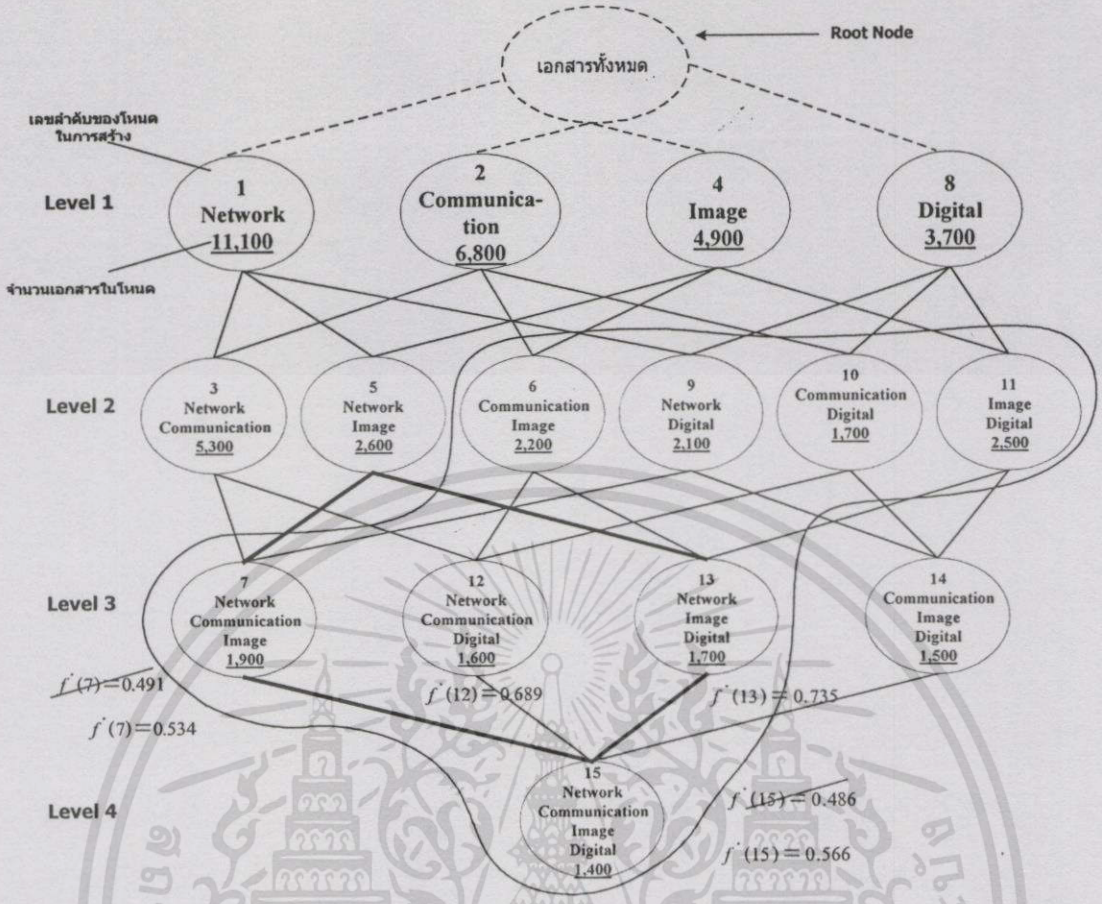
- <u>โหนดหมายเลข 13</u> ในทำนองเดียวกัน เอกสารที่เป็นกลุ่มตัวอย่างที่แท้จริงในโหนดหมายเลข 13 จำนวน 400 ฉบับ (*Ret#* = 400) และตรงความต้องการ 200 ฉบับ (*Rel#* = 200) คำนวณค่า $f^*(n)$ ได้ดังนี้

$$f^*(13) = g^*(13) + h^*(13)$$
$$= \frac{400}{1,700} + \left(1 - \frac{200}{400}\right)$$
$$= 0.735$$

- <u>โหนดหมายเลข 15</u> มีเอกสารตรงกับโหนดหมายเลข 5 จำนวน 50 ฉบับ (*Ret#* = 280+50) และมีเอกสารที่ตรงความต้องการจำนวน 21 ฉบับ (*Rel#* = 200+21)

$$f^*(15) = g^*(15) + h^*(15)$$
$$= \frac{330}{1,400} + \left(1 - \frac{221}{330}\right)$$
$$= 0.566$$

เช่นเดียวกันกับโหนดหมายเลข 3 หลังจากปรับค่าความเกี่ยวข้องและคำนวณค่า $f^*(n)$ ในโหนดที่เกี่ยวข้องที่อยู่ใน Level ล่างแล้ว อัลกอริทึม IP.A จะทำการรวมโหนดหมายเลขที่ 5 ไว้ใน CLOSED เซ็ต สามารถแสดงได้ดังรูปที่ ข.3



รูปที่ ข.3 แสดงการคำนวณหาค่า $f^*(n)$ ของโหนดหมายเลข 5 และการปรับค่า $f^*(n)$ ของโหนดหมายเลข 7 13 และ 15 ของชุดคำถามสืบค้นจำนวน 4 คำ ได้แก่ {Network, Communication, Image, Digital}

รอบการทำงานทั้งหมดต้องกระทำเป็นจำนวน 11 รอบ ซึ่งมีลักษณะของการคำนวณคล้ายๆ กันในแต่ละรอบ ดังที่ได้ยกตัวอย่างไปแล้ว ดังนั้นจึงขอสรุปว่าในแต่ละรอบต้องทำการสุ่มตัวอย่างและปรับค่า $f^*(n)$ ของโหนดใดบ้าง ดังนี้

<u>รอบการทำงานที่ 3</u>

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 6

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 12, 13 และ 15

<u>รอบการทำงานที่ 4</u>

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 9

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 7, 14 และ 15

รอบการทำงานที่ 5

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 10

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 12, 14 และ 15

รอบการทำงานที่ 6

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 11

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 13, 14 และ 15

รอบการทำงานที่ 7

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 7

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 15

รอบการทำงานที่ 8

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 12

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 15

รอบการทำงานที่ 9

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 13

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 15

รอบการทำงานที่ 10

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 14

การสุ่มตัวอย่างที่แท้จริง ได้แก่ โหนดหมายเลข 15

รอบการทำงานที่ 11

การสุ่มตัวอย่างแลททิสโหนด ได้แก่ โหนดหมายเลข 15

การสุ่มตัวอย่างที่แท้จริง ไม่มีการปรับค่า

ภาคผนวก ค.

ตารางแสดงการคำนวณหาขนาดของกลุ่มตัวอย่าง สำหรับการสุ่มตัวอย่าง

Sample Size for Specified Confidence Limits and Precision When Sampling Attributes in Percent

A. 2σ Confidence Interval
($\pi = 0.5$)[a]

Size of Population (N)	Sample Size (n) for Precision (e) of ±1%	±2%	±3%	±4%	±5%	±10%
500	[b]	[b]	[b]	[b]	222	83
1,000	[b]	[b]	[b]	385	286	91
1,500	[b]	[b]	638	441	316	94
2,000	[b]	[b]	714	476	333	95
2,500	[b]	1,250	769	500	345	96
3,000	[b]	1,364	811	517	353	97
3,500	[b]	1,458	843	530	359	97
4,000	[b]	1,538	870	541	364	98
4,500	[b]	1,607	891	549	367	98
5,000	[b]	1,667	909	556	370	98
6,000	[b]	1,765	938	566	375	98
7,000	[b]	1,842	959	574	378	99
8,000	[b]	1,905	976	580	381	99
9,000	[b]	1,957	989	584	383	99
10,000	5,000	2,000	1,000	588	385	99
15,000	6,000	2,143	1,034	600	390	99
20,000	6,667	2,222	1,053	606	392	100
25,000	7,143	2,273	1,064	610	394	100
50,000	8,333	2,381	1,087	617	397	100
100,000	9,091	2,439	1,099	621	398	100
→ ∞	10,000	2,500	1,111	625	400	100

[a] Formula for sample size when population proportion π is

$$n_0 = \frac{z^2\pi(1-\pi)N}{z^2\pi(1-\pi) + Ne^2}$$

This table assumes $\pi = 0.5$, $z = 2$:

$$n = \frac{2^2(0.5)^2N}{2^2(0.5)^2 + Ne^2} = \frac{N}{1 + Ne^2}$$

$$n \geqq n_e$$

[b] In these cases the assumption of normal approximation is poor, and the formula does not apply.

ภาคผนวก ง.

ตัวอย่างหนังสืออ้างอิงการให้คำสำคัญเอกสาร สำหรับบรรณารักษ์

การกำหนดคำสำคัญให้กับหนังสือ หรือเอกสารในห้องสมุดนั้น เป็นหน้าที่ของบรรณารักษ์ที่ต้องเป็นผู้กำหนด และตีความเอกสารเพื่อกำหนดคำสำคัญ สำหรับการสืบค้นตามกระบวนการของการค้นคืนสารสนเทศ ซึ่งการกำหนดคำสำคัญโดยทั่วไปที่กำหนดมาจากหัวเรื่อง (Subject) โดยอาศัยหนังสืออ้างอิงที่ใช้เป็นมาตรฐานสากล มีทั้งสำหรับเอกสารภาษาอังกฤษรวมทั้งภาษาไทยด้วย แต่การอ้างหัวเรื่องภาษาไทยนั้นต้องใช้หนังสือคู่มืออ้างอิงจำนวนหลายเล่ม ซึ่งต่างจากภาษาอังกฤษ ในส่วนของการกำหนดคำสำคัญโดยใช้หัวเรื่องที่เป็นภาษาอังกฤษนั้น ได้อาศัยหนังสือ "Subject Headings" ของ Library of Congress, Cataloging Distribution Service (ระบบห้องสมุดรัฐสภาอเมริกัน) สำหรับหอสมุดกลาง สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังใช้อยู่นั้น เป็นการตีพิมพ์ครั้งที่ 21 ในปี คศ. 1998 ซึ่งหนังสือคู่มือการให้หัวเรื่องที่เป็นภาษาอังกฤษนี้มีจำนวนทั้งสิ้น 5 เล่ม แบ่งตามตัวอักษร คือหมวด A-C, D-H, I-M, N-R และ S-Z ที่ผู้วิจัยนำมาเป็นตัวอย่างนี้ เป็นเพียงส่วนหนึ่งของหมวดอักษร C กลุ่มหัวเรื่อง "Computer" และหมวดอักษร H กลุ่มหัวเรื่อง "Heuristic"

BT Computer arithmetic
Electronic digital computers—Circuits

Computer art *(May Subd Geog)*
[N7433.8-N7433.85 (Visual arts)]
[TT869.5 (Handicraft)]
UF Art, Computer
Computer craft
BT Art, Modern—20th century
Computers
NT Stereograms

Computer-assisted design
USE Computer-aided design

Computer-assisted film making
USE Computer animation

Computer-assisted instruction
(May Subd Geog)
[LB1028.5-LB1028.7]
Here are entered works on an automated method of instruction in which a student interacts directly with instructional materials stored in a computer. Works on the use of computers to assist teachers and administrators in coordinating the instructional process, e.g. retrieving and summarizing performance records and curriculum files, are entered under Computer managed instruction.
UF CAI
Computer-aided instruction
Computer-assisted learning
Computer based instruction
Electronic data processing in programmed instruction
Microcomputer-aided instruction
Microcomputer-assisted instruction
Microcomputer-assisted learning
Microcomputer-based instruction
Teaching—Data processing
BT Education—Data processing
Educational technology
Programmed instruction
Telematics
SA *subdivision* Computer-assisted instruction *under topical headings, e.g.* Chemistry—Computer-assisted instruction *and subdivisions* Computer-assisted instruction, Computer-assisted instruction for foreign speakers, *and* Computer-assisted instruction for French, [Spanish, etc.] speakers *under individual languages, e.g.* English language—Computer-assisted instruction
NT Intelligent tutoring systems
LEKTOR (Computer system)
Mathematical statistics—Computer-assisted instruction
TICCIT (Computer system)
— Authoring programs
[LB1028.66]
Here are entered works on computer programs that allow the user, with relatively little expertise, to create customized computer programs for educational purposes.
UF Authoring programs for computer-assisted instruction
BT Computer managed instruction
Computer programs
NT ALT Authoring System (Computer system)
— Computer programs
— Law and legislation *(May Subd Geog)*
BT Educational law and legislation
— Programming
[LB1028.65]
BT Programming (Electronic computers)

Computer-assisted learning
USE Computer-assisted instruction

Computer assisted logic design
USE Logic design—Data processing

Computer-Assisted Maintenance Planning and Control System
USE CAMCOS System

Computer Assisted Makeup and Imaging System
USE CAMIS System

Computer-assisted neurosurgery
(May Subd Geog)
[RD593.5]
UF Computerized neurosurgery
BT Nervous system—Surgery

Computer-assisted videokeratography
(May Subd Geog)
UF Computerized videokeratography
Videokeratography, Computer-assisted
BT Cornea—Examination
Medical cinematography

Computer audits
USE Auditing—Data processing
Electronic data processing—Auditing

Computer aviation
USE Computer flight games

Computer-based conferencing
USE Computer conferencing

Computer-based information systems
USE Information storage and retrieval systems
Management information systems

Computer based instruction
USE Computer-assisted instruction

Computer-based multimedia information systems
USE Multimedia systems

Computer bulletin boards *(May Subd Geog)*
[QA76.9.B84]
Here are entered works on services that allow a computer user to post messages to, and read messages from, a group of people who have a common interest, via a dedicated telephone line established for the purpose. Works on services, commonly called newsgroups and listservs, that allow a computer user to post messages to, and read messages from, a group of people who have a common interest, usually by means of the Internet, a commercial online service, or electronic mail are entered under Electronic discussion groups. Works on services that allow a person, using a computer, to engage in an actual "conversation" with other people in real time are entered under Online chat groups.
UF BBSs (Computer bulletin boards)
Bulletin board systems (Computers)
Electronic bulletin boards
RT Electronic discussion groups
Online chat groups
NT PharmNet (Information retrieval system)
— Law and legislation *(May Subd Geog)*
— Telephone directories
BT Telephone—Directories

Computer camps *(May Subd Geog)*
[QA76.33]
BT Camps
Electronic data processing—Study and teaching
Microcomputers—Study and teaching

Computer capacity
[QA76.9.C63]
UF Capacity, Computer
BT Electronic data processing

Computer cartography
USE Digital mapping

Computer centers
USE Computation laboratories
Data processing service centers
Electronic data processing departments

Computer checkers *(May Subd Geog)*
UF Checkers—Data processing
BT Checkers

Computer chess *(May Subd Geog)*
[GV1449.3]
UF Chess—Data processing
[Former heading]
BT Chess

Computer chips
USE Integrated circuits

Computer circuits
USE Computers—Circuits

Computer color graphics
USE Color computer graphics

Computer coloring of motion pictures
USE Colorization of motion pictures

Computer communication systems
USE Computer networks

Computer composition
BT Computer sound processing
Electronic composition

Computer conferencing *(May Subd Geog)*
[HF5734.7 (Business communication)]
UF Computer-based conferencing
Conferencing, Computer
Desktop conferencing
BT Business communication
Telematics

Computer consultants
USE Electronic data processing consultants

Computer contracts *(May Subd Geog)*
BT Contracts
NT Computer leases

Computer control
USE Automation

Computer controlled instruments
USE Computerized instruments

Computer craft
USE Computer art

Computer crimes *(May Subd Geog)*
[HV6773]
UF Computer fraud
Computers and crime
BT Crime
RT Privacy, Right of
NT Computer viruses
— Investigation *(May Subd Geog)*
[HV8079.C65]
BT Criminal investigation

Computer department security measures
USE Electronic data processing departments—Security measures

Computer-directed trading (Securities)
USE Program trading (Securities)

Computer documentation
USE Electronic data processing documentation

Computer drawing
Here are entered works on the use of computer graphics to create artistic designs. Works on the technique for producing line drawings, including particularly engineering drawings, by use of current digital computing and plotting equipment are entered under Computer graphics.
BT Drawing
Image processing—Digital techniques
NT Computer animation

Computer engineering *(May Subd Geog)*
[TK7885-TK7895]
Here are entered works on the design of computer hardware and circuitry. Works on the logical structure that determines the way a computer executes programs are entered under Computer architecture. Works on the way a computer is constructed to implement its architecture, including what components are used and how they are connected, are entered under Computer organization.
UF Computers—Design and construction
— Data processing
NT CONLAN (Computer hardware description language)

Computer engineers *(May Subd Geog)*
BT Engineers

Computer file conversion
USE File conversion (Computer science)

Computer files *(May Subd Geog)*
UF Machine-readable data files
[Former heading]
Machine-readable files

Computer files *(Continued)*
BT Files (Records)
NT Cataloging of computer files
Computer programs
Databases
Image files
Text files
— Conversion
USE File conversion (Computer science)
— Copyright
USE Copyright—Computer files
— Law and legislation *(May Subd Geog)*

Computer firmware
[QA76.765]
UF Firmware, Computer
BT Computer programs

Computer flight games *(May Subd Geog)*
[TL712.8]
UF Computer aviation
Flight—Computer games
Flight games
BT Airplanes—Piloting—Computer simulation
Computer games
Flight simulators
SA *subdivision* Computer games *under subjects*

Computer fonts *(May Subd Geog)*
UF Fonts, Computer
BT Type and type-founding
RT Font editors
NT PostScript fonts
TrueType fonts

Computer fraud
USE Computer crimes

Computer furniture *(May Subd Geog)*
[TT197.5.C65 (Woodworking)]
UF Furniture, Computer
BT Furniture
Microcomputers—Equipment and supplies

Computer game programming
USE Computer games—Programming

Computer games *(May Subd Geog)*
[GV1469.15-GV1469.25]
Here are entered works on games played on a computer. Works on the application of computers and data processing techniques to games in general, including recording statistics, setting up tournaments, etc., are entered under Games—Data processing.
BT Application software
Electronic games
SA *subdivision* Computer games *under subjects*
NT Antagonists (Game)
Balance of Power (Game)
Carmen Sandiego (Game)
Computer adventure games
Computer flight games
Computer war games
Harpoon Battlebook (Game)
MicroLeague Baseball (Game)
MicroLeague Football (Game)
Populous (Game)
Quest for Glory (Game)
Railroad Tycoon (Game)
Simlife (Game)
Super munchers (Game)
Timelost (Game)
Ultima (Game)
— Programming
[QA76.76.C672]
UF Computer game programming
Game programming (Computer games)
BT Programming (Electronic computers)

Computer-generated metamorphosis
USE Morphing (Computer animation)

Computer graphics
[T385]
Here are entered works on the technique for producing line drawings, including particularly engineering drawings, by use of current digital computing and plotting equipment. Works on the use of computer graphics to create artistic designs are entered under Computer drawing.
UF Automatic drafting
Graphic data processing
Graphics, Computer
BT Electronic digital computers
Engineering graphics
Image processing—Digital techniques
NT Bit-mapped graphics
Color computer graphics
Digital incremental plotters
Digital video
DRAGON (Computer system)
DSS/A (Computer system)
DSS/F (Computer system)
GIML (Computer program language)
GRAIL (Electronic computer system)
GRASS (Electronic computer system)
HIRASP (Computer system)
Radiosity
ROBOCAR (Electronic computer system)
— Computer programs
— Equipment and supplies
— Standards *(May Subd Geog)*
NT GKS (Computer system)

Computer graphics equipment industry *(May Subd Geog)*
[HD9696.C6-HD9696.C64]
BT Computer industry

Computer hackers *(May Subd Geog)*
UF Hackers, Computer
BT Criminals
Persons

Computer hardware
USE Computer input-output equipment
Computers

Computer hardware description languages
[TK7885.7]
UF Hardware description languages, Computer
Languages, Computer hardware description
BT Electronic digital computers—Design and construction—Data processing
NT CONLAN (Computer hardware description language)
ELLA (Computer hardware description language)
STREAM (Computer hardware description language)
Verilog (Computer hardware description language)
VHDL (Computer hardware description language)

Computer-human interaction
USE Human-computer interaction

Computer I/O equipment
USE Computer input-output equipment

Computer industry *(May Subd Geog)*
[HD9696.C6-HD9696.C64]
BT Electronic industries
NT Automatic data collection equipment industry
Computer access control equipment industry
Computer graphics equipment industry
Computer printer supplies industry
Computer service industry
Computer storage device industry
Computer vision equipment industry
Computers
Data disk drives industry
Data tape drives industry
Data transmission equipment industry
Internet industry
Wide area networks industry
— Customer services *(May Subd Geog)*
NT Computer technical support
— Employees
NT Electronic data processing personnel
Women computer industry employees
— Public relations
USE Public relations—Computer industry

Computer input design
USE Input design, Computer

Computer input-output equipment *(May Subd Geog)*
UF Computer hardware
Computer I/O equipment
Computers—Input-output equipment
Electronic analog computers—Input-output equipment
Electronic digital computers—Input-output equipment
Hardware, Computer
I/O equipment (Computers)
Input equipment (Computers)
Input-output equipment (Computers)
Output equipment (Computers)
NT Analog-to-digital converters
Automatic speech recognition
Computer interfaces
Computer output microfilm
Computer output microfilm devices
Computer output optical disk devices
Computer output optical disks
Computer peripherals
Computer storage devices
Computer terminals
Computers—Optical equipment
Data disk drives
Data tape drives
Electronic data processing—Data preparation
Information display systems
Input design, Computer
Intel 8089 (Microprocessor)
Keyboards (Electronics)
Keypunches
Mice (Computers)
Modems
Printers (Data processing systems)
Punched card systems
Reading machines (Data processing equipment)
Speech synthesis

Computer insurance
USE Insurance, Computer

Computer integrated manufacturing systems *(May Subd Geog)*
[TS155.63]
UF CIM systems
Manufacturing, Computer integrated
BT Computer-aided engineering
Industrial engineering
RT Flexible manufacturing systems
NT CAD/CAM systems
Solid freeform fabrication

Computer interfaces
[TK7887.5]
Here are entered works on the connections or links between two or more computer systems or devices. Works on the point where a user interacts with a computer system, either directly through a terminal or indirectly through a data processing department, are entered under User interfaces (Computer systems).
UF Interfaces, Computer
BT Computer input-output equipment
Interface circuits
NT Expansion boards (Microcomputers)
Line drivers (Integrated circuits)
Line receivers (Integrated circuits)

Heteropogon *(Continued)*
NT Black spear grass
Heteropogon contortus
USE Black spear grass
Heteropsammia *(May Subd Geog)*
[QL377.C7]
BT Dendrophylliidae
Heteropsylla *(May Subd Geog)*
[QL527.P88 (Zoology)]
BT Jumping plant-lice
NT Leucaena psyllid
Heteropsylla cubana
USE Leucaena psyllid
Heteroptera
USE Hemiptera
Heteroptera, Fossil
USE Hemiptera, Fossil
Heterorhabditidae *(May Subd Geog)*
[QL391.N4]
BT Rhabditida
NT Heterorhabditis
Heterorhabditis *(May Subd Geog)*
[QL391.N4]
BT Heterorhabditidae
Heterorthina *(May Subd Geog)*
BT Dalmanellidae
Heterosarus *(May Subd Geog)*
[QL568.A4]
BT Andrenidae
Heteroscedasticity
BT Analysis of variance
Econometrics
Least squares
RT Homoscedasticity
Heterosexism *(May Subd Geog)*
Here are entered works on prejudicial attitudes or assumptions held by heterosexuals concerning homosexuals or homosexuality. Works on active discrimination against, or aversion to, homosexuals by heterosexuals are entered under Homophobia.
UF Heterocentrism
BT Sexism
RT Homophobia
Heterosexual men *(May Subd Geog)*
UF Straight men (Sexual orientation)
BT Heterosexuals
Men
Heterosexual mothers *(May Subd Geog)*
Here are entered works on mothers that emphasize their heterosexuality, usually in contrast to lesbians as mothers. General works on mothers without regard to their sexuality are entered under Mothers.
BT Heterosexual parents
Mothers
Heterosexual parents *(May Subd Geog)*
Here are entered works on parents that emphasize their heterosexuality, usually in contrast to gay men or lesbians as parents. General works on parents without regard to their sexuality are entered under Parents.
BT Parents
NT Children of heterosexual parents
Heterosexual mothers
Heterosexual persons
USE Heterosexuals
Heterosexual teachers *(May Subd Geog)*
BT Teachers
Heterosexual women *(May Subd Geog)*
UF Straight women (Sexual orientation)
BT Heterosexuals
Women
Heterosexuality *(May Subd Geog)*
BT Sexual orientation
Heterosexuality in art *(Not Subd Geog)*
Heterosexuality in literature *(Not Subd Geog)*
Heterosexuals *(May Subd Geog)*
UF Heterosexual persons
Straight persons (Sexual orientation)
Straights (Sexual orientation)
BT Persons
NT Heterosexual men
Heterosexual women
Heterosis
[QH421 (Biology)]
[S494 (Breeding)]
[SB123 (Plant breeding)]
[SF105 (Animal breeding)]
UF Hybrid vigor
BT Breeding
Fertility
Growth
Growth (Plants)
Hybridization
Heterosomata
USE Flatfishes
Heterosporium *(May Subd Geog)*
[QK625.D4]
BT Dematiaceae
Heterostraci *(May Subd Geog)*
[QE852.A33]
BT Agnatha, Fossil
Heterostructures *(May Subd Geog)*
BT Crystals
Superlattices as materials
Heterostylism
[QK926]
BT Plants—Reproduction
Plants, Sex in
Pollination
Heterotardigrada *(May Subd Geog)*
[QL447.5]
BT Tardigrada
NT Halechiniscidae
Heterotetrarhynchus
USE Grillotia
Heterotherms
USE Poikilotherms
Heterotopic pain
USE Referred pain
Heterotricha
USE Heterotrichida
Heterotrichida *(May Subd Geog)*
[QL368.H55]
UF Heterotricha
BT Spirotricha
NT Climacostomidae
Nyctotheridae
Spirostomidae
Stentoridae
Heterotrichum
USE Saussurea
Heterotrophic bacteria
USE Bacteria, Heterotrophic
Heterotropia
USE Strabismus
Heterozygosis
[QH21-QH25 (Biology)]
[SB123 (Plant breeding)]
Heth family
USE Heath family
Hetherington family *(Not Subd Geog)*
RT Harrington family
Hetherley family
USE Heatherly family
Hetherlie family
USE Heatherly family
Hetherly family
USE Heatherly family
Hethman family
USE Heathman family
Hethumid family *(Not Subd Geog)*
Hetizel family
USE Hutzel family
Hetmans *(May Subd Geog)*
BT Cossacks
Hetol
[RM666.H]
Hetrick family
USE Hedderich family
Hetterley family
USE Heatherly family
Hettinger family
USE Hardinger family
Hetzel family *(Not Subd Geog)*
Hetzendorf Castle (Vienna, Austria)
USE Schloss Hetzendorf (Vienna, Austria)
Heubes family *(Not Subd Geog)*
Heuchera *(May Subd Geog)*
[QK495.S3]
UF Alumroot
BT Saxifragaceae
Heuchert family *(Not Subd Geog)*
Heudebourg family
USE Hudiburgh family
Heugh family
USE Hughes family
Heukelom family
USE Van Heukelom family
Heulandite *(May Subd Geog)*
[QE391.H55]
BT Zeolites
Heuman family *(Not Subd Geog)*
Heune family
USE Hohn family
Heuneburg Site (Germany)
BT Germany (West)—Antiquities
Heurich family *(Not Subd Geog)*
Heuristic
[BD260]
BT Methodology
Philosophy
Heuristic programming
[T57.84]
BT Artificial intelligence
Programming (Mathematics)
Heurn family *(Not Subd Geog)*
Heuscheuer Mountains (Poland)
USE Stołowe Mountains (Poland)
Heuschkel family
USE Heiskell family
Heuschkell family
USE Heiskell family
Heuschle family
USE Heisley family
Heuser family
USE Hauser family
Heuss family *(Not Subd Geog)*
RT Heiss family
Heutmacher family
USE Hatmaker family
Hevajra (Buddhist deity) *(Not Subd Geog)*
[BQ4860.H47]
BT Gods, Buddhist
Hève, Cap de la (France)
UF Cap de la Hève (France)
BT Capes (Coasts)—France
Heve language
USE Eudeve language
Hevea *(May Subd Geog)*
[QK495.E9 (Botany)]
[SB291.H4 (Culture)]
UF Hevea brasiliensis
Para rubber tree
Siphonia ridleyana
BT Euphorbiaceae
Rubber plants
— Diseases and pests *(May Subd Geog)*
NT Brown root disease of hevea
White root disease of hevea
— Weed control *(May Subd Geog)*
[SB608.H5]
Hevea brasiliensis
USE Hevea
Hevea seed oil *(May Subd Geog)*
[TP684.R83 (Chemical technology)]
UF Pará rubber seed oil
Rubber-seed oil
BT Vegetable oils
Hevel (The Hebrew word)
BT Hebrew language—Etymology
Hevel Elot (Israel)
USE Elot Region (Israel)

ภาคผนวก จ.

บทความและผลงานวิจัยที่ได้รับการตีพิมพ์

1. วรางคณา เงินแก้ว และเอื้อน ปิ่นเงิน. "การประยุกต์การสืบค้นแบบฮิวริสติคสำหรับการค้นคืนสารสนเทศ (Applied Heuristic Search in Information Retrieval)." สารสนเทศลาดกระบัง, ปีที่ 5, ฉบับที่ 1, 2543.

วารสาร

กรกฎาคม 2543 ปีที่ 5 ฉบับที่ 1

สารสนเทศลาดกระบัง

Ladkrabang Information Journal

ISSN 0859 - 5208

July 2000 Vol.5 No.1

คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Faculty of Information Technology, King Mongkut's Institute of Technology, Ladkrabang Bangkok 10520

บทความวิจัย

บทความวิชาการ

บทความทั่วไป

การประยุกต์การสืบค้นแบบฮิวริสติคสำหรับการค้นคืนสารสนเทศ

Applied Heuristic Search in Information Retrieval

นางสาววรางคณา เงินแก้ว*
Miss Warangkhana Ngenkaew

ผศ. ดร.เอื้อน ปิ่นเงิน**
Asst. Prof. Dr. Ouen Pinngern

บทคัดย่อ

บทความนี้ เป็นการนำเสนอการประยุกต์เทคนิคทางปัญญาประดิษฐ์ (Artificial Intelligence) เพื่อการค้นคืนสารสนเทศ (Information Retrieval) เทคนิคดังกล่าวคือ การสืบค้นแบบฮิวริสติค (Heuristic search) โดยใช้สืบค้นฐานข้อมูลที่เป็นเอกสาร วิธีการนี้ทำให้การสืบค้นเอกสารเป็นไปได้อย่างรวดเร็ว และได้เอกสารที่ตรงตามความต้องการของผู้ใช้โดยเรียงลำดับจากมากไปหาน้อย เนื่องจากผู้ใช้สามารถกำหนดความต้องการของตนเองได้ด้วยการใช้คำเฉพาะ (Keywords) เทคนิคดังกล่าวอาศัยหลักการของอัลกอริทึม A* และ อัลกอริทึม IRA ซึ่งใช้หลักการของความน่าจะเป็น ทำการสืบค้นเอกภพของเอกสาร อัลกอริทึมดังกล่าวมีการปรับค่าความน่าจะเป็นตลอดเวลาของการสืบค้น เพื่อให้ตรงตามความต้องการของผู้ใช้

Abstract

This paper presents the using of Artificial Intelligence technique for information retrieval. The technique, called heuristic search, uses for retrieving relevant documents from document space. It helps users to retrieve documents in descending order of relevancy to their needs, since the users can specify their needs through the set of keywords of interests. A used heuristic search is based on A* algorithm and IRA algorithm which use probability technique for retrieving the relevant document. These two algorithms update the probability function along the way the search proceeds.

---

* นักศึกษาปริญญาโท คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

** อาจารย์ประจำภาควิชา วิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

1. บทนำ

การค้นคืนสารสนเทศ (Information Retrieval : IR) เป็นการกระทำที่เกี่ยวกับ การแทน การจัดเก็บ การจัดรูปแบบ และการเข้าถึง สารสนเทศที่มีลักษณะที่เป็นเอกสาร (document) ปัญหาพื้นฐานของ IR คือ การระบุเอกสารที่ตรงตามความต้องการ (relevant) ของผู้ใช้จากเอกสารที่มีอยู่ทั้งหมด แต่ระบบสืบค้นโดยทั่วไปมักใช้หลักการสืบค้นแบบตรงตัว (exact match) โดยใช้คำเฉพาะเป็นตัวค้นคืนเอกสารที่ต้องการ แล้วนำคำเฉพาะนั้นไปจับคู่กับชื่อเรื่องหรือ หัวเรื่องของเอกสาร ซึ่งวิธีการนี้มีข้อด้อยที่สำคัญในเรื่องของการป้อนคำสืบค้น คือการที่ผู้ใช้ป้อนคำสืบค้นมากเกินไป หรือน้อยเกินไปนั้นอาจมีผลกับคำตอบที่ได้รับ หากผู้ใช้ป้อนคำสืบค้นที่เฉพาะเจาะจงมากเกินไป อาจได้เอกสารออกมาเป็นจำนวนน้อย เป็นผลให้เอกสารที่ผู้ใช้ต้องการนั้นถูกตัดออกไป หรือหากผู้ใช้ใส่คำเฉพาะน้อยคำเกินไป อาจทำให้ได้เอกสารออกมาจำนวนมาก จนทำให้ต้องเสียเวลาในการคัดเลือกว่าเอกสารที่ต้องการนั้นคือเอกสารใด

การสืบค้นแบบฮิวริสติคช่วยผู้ใช้แก้ปัญหาโดยการหากลุ่มของคำสืบค้นที่เหมาะสมสำหรับสืบค้นเอกสาร เพื่อแก้ปัญหาเรื่องการได้เอกสารออกมามากหรือน้อยเกินไป

2. งานวิจัยที่เกี่ยวข้อง

งานวิจัยเกี่ยวกับการพัฒนาระบบเพื่อค้นคืนสารสนเทศในอดีต ตัวอย่างเช่น

1. MEDLARS [4] : National Library of Medicine ได้พัฒนาขึ้นเมื่อปี คศ. 1964 เพื่อนำมาใช้ในห้องสมุดทางการแพทย์ และได้รวบรวมวารสารทางการแพทย์เพื่อการค้นคืน โดยใช้คำสั่งสืบค้นแบบบูลีนอย่างง่าย

2. The DIALOG System [3] : พัฒนาขึ้นโดย Lockheed Information Systems ที่รัฐแคลิฟอร์เนียในปี คศ. 1980 ระบบทำการสร้างกลุ่มของเอกสารอ้างอิงโดยใช้คำสั่ง SELECT และใช้ตัวดำเนินการทางตรรกศาสตร์เพื่อรวมคำแต่ละคำเข้าด้วยกัน

3.STAIRS (Storage and Information Retrieval System) [3] : เป็นโปรแกรมจัดเก็บและค้นคืนสารสนเทศทางด้านการค้าของบริษัท IBM STAIRS แตกต่างจาก DIALOG ตรงที่ STAIRS ไม่เพียงแต่จัดการเรื่องการค้นคืนเอกสารเท่านั้น แต่ยังสามารถจัดการเกี่ยวกับระบบฐานข้อมูลอีกด้วย

4. The Information Bank [3] : เป็นการสืบค้นชื่อบทความที่ได้ตีพิมพ์ในนิตยสาร The New York Times และสิ่งพิมพ์อื่นที่น่าสนใจ ศัพท์ที่ใช้ในการสืบค้นต้องมีรูปแบบที่แน่นอนและถูกต้อง รูปแบบการสืบค้นใช้ตัวดำเนินการทางตรรกศาสตร์

5. การทดลอง TREC (TREC Experiments) ปี คศ. 1993 - 1995 [3] : TREC ย่อมาจาก Text Retrieval Conference เป็นการประชุมทางวิชาการมุ่งเน้นเรื่องการวิจัยเพื่อทำการพัฒนาประสิทธิภาพของการค้นคืนสารสนเทศ การทดลองนี้ได้ศึกษาการจัดการกับฐานข้อมูลแบบ Full-Text ขนาดใหญ่ รวมถึงการประเมินคำตอบที่ได้จากการทดสอบ ซึ่ง TREC นี้ได้มีการพัฒนาให้มีประสิทธิภาพมากยิ่งขึ้นเป็น TREC-1, TREC-2, TREC-3 และ TREC-4

จากงานวิจัยที่ได้กล่าวข้างต้น ระบบสืบค้นส่วนมากมักใช้หลักการสืบค้นแบบตรงตัว และการเรียงลำดับเอกสารทำโดยการนับจำนวนความถี่ของคำสืบค้นที่พบในเอกสาร โดยระบบดังกล่าวไม่ได้นำข้อมูลที่ได้จากผู้ใช้มาช่วย ในเรื่องของการสืบค้นและเรียงลำดับ

3. การสืบค้นแบบฮิวริสติค (Heuristic Search)

คำว่า "heuristic" มาจากภาษากรีก คือคำว่า heuriskein มีความหมายว่า "เพื่อทำการค้นหา" หรือ "ค้นพบ" ซึ่งความหมายทางเทคนิคของฮิวริสติค ได้มีการเปลี่ยนแปลงหลายครั้ง ตามประวัติศาสตร์ของ AI [8]

Feigenbaum และ Feldman [1] กล่าวไว้เมื่อปี ค.ศ.1963 ว่า

"A *heuristic (heuristic rule, heuristic method)* is a rule of thumb, simplification, or any other kind of device which drastically limits search for solutions in large problem [more correctly, solution] spaces."

ดังที่ได้กล่าวไปแล้วข้างต้น การสืบค้นโดยใช้เทคนิคฮิวริสติค มีความแตกต่างกับการสืบค้นที่ไม่มีความรู้ช่วยตรงที่ ฮิวริสติค เป็นกระบวนการที่ใช้ความฉลาดในการช่วยการสืบค้น เพื่อให้ได้คำตอบที่รวดเร็ว โดยใช้เวลาและหน่วยความจำน้อยที่สุด และผลลัพธ์ที่ได้เป็นที่น่าพอใจ โดยทั่วไปฮิวริสติค ใช้แก้ปัญหาใน 2 ลักษณะคือ [5]

1. ปัญหาที่มีวิธีการหาคำตอบที่ไม่แน่นอน เนื่องจากมีความคลุมเครือ เช่นการวินิจฉัยทางการแพทย์ อาการของคนไข้อาจมาจากหลายๆสาเหตุ ดังนั้น แพทย์จะใช้ฮิวริสติค เพื่อวินิจฉัยโรคที่น่าจะเป็นไปได้ และวางแผนการรักษา

2. ปัญหาที่มีการหาคำตอบที่แน่นอน เป็นปัญหาที่ DFS และ BFS สามารถแก้ไขได้ แต่ต้องใช้เวลา หรือหน่วยความจำมาก หากใช้ฮิวริสติค ในการแก้ปัญหาจะทำให้ได้คำตอบที่พึงพอใจ ภายในเวลาและการใช้หน่วยความจำที่เหมาะสม

ในบทความนี้จะกล่าวถึง อัลกอริทึม IRA ที่มีพื้นฐานมาจากอัลกอริทึม A* [2] ซึ่งเป็นอัลกอริทึมทางปัญญาประดิษฐ์ที่ถูกประยุกต์ขึ้นมาเพื่อใช้กับการค้นคืนสารสนเทศ และฟังก์ชันฮิวริสติคของทั้ง 2 อัลกอริทึม ซึ่งอัลกอริทึมทั้ง 2 ชนิดนี้มีข้อแตกต่างกันดังจะได้กล่าวต่อไป

4. อัลกอริทึม A* กับการสืบค้นเอกสาร

การประยุกต์อัลกอริทึม A* จากปัญญาประดิษฐ์เพื่อสืบค้นเอกภพของเอกสาร เทคนิคการสืบค้นใช้ฮิวริสติค นับเป็นเครื่องมือที่มีประโยชน์และสามารถนำไปปฏิบัติได้จริงสำหรับการแก้ปัญหาประเภทต่าง ๆ เอกภพของปัญหาจำนวนมากสามารถแทนได้ด้วยกราฟ อัลกอริทึม A* จะใช้สำหรับหาเส้นทางของค่าใช้จ่ายที่น้อยที่สุดของโหนด จากจุดเริ่มต้นไปยังโหนดเป้าหมาย โดยใช้ข้อมูลที่เป็นฮิวริสติคที่เกี่ยวข้องกับค่าใช้จ่ายของส่วนที่ยังไม่ทราบคำตอบสำหรับโหนดใดๆ ของกราฟ นิยามฟังก์ชันดังนี้

$g(n)$ = ค่าใช้จ่ายที่น้อยที่สุดตามเส้นทางจากโหนดเริ่มต้น ถึงโหนด n

$h(n)$ = ค่าใช้จ่ายประมาณการที่น้อยที่สุดตามเส้นทางจากโหนด n ไปยังโหนดเป้าหมาย

$f(n) = g(n) + h(n)$ แทนค่าใช้จ่ายที่คาดว่าน้อยที่สุดตามเส้นทางจากโหนดเริ่มต้นไปยังโหนดเป้าหมายโดยผ่านโหนด n

เนื่องจากในระหว่างการสืบค้น เรายังไม่ทราบค่าใช้จ่ายที่แท้จริงของเส้นทางการค้นหาแต่ละเส้นทาง จึงจำเป็นที่จะต้องทำการประมาณค่าใช้จ่าย เพื่อให้อัลกอริทึมสามารถทำงานต่อไปได้

ให้ $g^*(n)$ = ค่าใช้จ่ายประมาณที่น้อยที่สุดตามเส้นทางจากโหนดเริ่มต้นไปยังโหนด n

$h^*(n)$ = ค่าประมาณของ $h(n)$ ที่ได้จากข้อมูลหรือความรู้ที่เป็นฮิวริสติค

$f^*(n) = g^*(n)+h^*(n)$ เป็นค่าใช้จ่ายโดยประมาณที่น้อยที่สุด ตามเส้นทางจากโหนดเริ่มต้นไปยังโหนดเป้าหมายโดยผ่านโหนด n

5. การสืบค้นเอกภพของเอกสารโดยใช้ฮิวริสติค

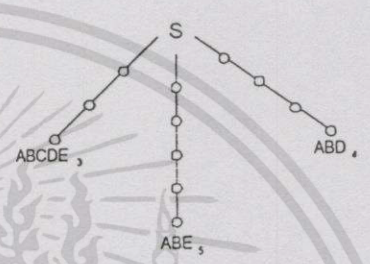
การสืบค้นเอกสารปกติเราจะใช้คำสืบค้นเฉพาะ ทำการสืบค้นกับกลุ่มของเอกสารที่มีดัชนีเป็นคำเฉพาะแทนเอกสารแต่ละฉบับ อัลกอริทึม A* ช่วยให้ผู้ใช้สืบค้นเอกสารได้ดีขึ้น โดยมีการปรับแต่งค่าความน่าจะเป็นของเอกสารที่ตรงตามความต้องการอยู่อย่างต่อเนื่อง โดยการใช้ฮิวริสติคที่ผู้ใช้สามารถอธิบาย หรือกำหนดความต้องการด้วยคำเฉพาะที่ดีที่ใช้สำหรับสืบค้น ปัญหาในการสืบค้นเอกสาร คือถ้ามีคำสืบค้นจำนวน m คำ จะมีเซ็ตของเอกสารจำนวน 2^m-1 เซ็ต ที่จะสืบค้นได้ ซึ่งจำนวนมากขนาดนี้ทำให้ผู้ใช้ไม่สามารถกำหนดการจัดกลุ่มของคำ ว่าเป็นแบบใดจึงจะทำให้การสืบค้นมีประสิทธิภาพที่สุด และเพื่อให้การอธิบายอัลกอริทึมได้ชัดเจนขึ้น จะขอนิยามคำศัพท์เฉพาะที่จำเป็นต้องใช้ดังนี้ [2]

นิยาม 1 : เทอมและโหนดสืบค้น

สมมติว่าคำเฉพาะที่จะใช้สำหรับสืบค้น 5 คำ คือ {A,B,C,D, และ E} ดังรูปที่ 1 S คือ เอกภพของเอกสาร (Document Space) แทนเอกสารทั้งหมด ที่ยังไม่ได้รับการประเมินจากผู้ใช้ วงกลมเล็กแทนการประเมินของผู้ใช้ที่ประเมินโหนดดังกล่าว และกลุ่มของเอกสารจะแทนด้วยโหนดสืบค้น ตัวอย่างเช่น โหนดสืบค้น ABE_5 จะแทนส่วนหนึ่งของเอกสารที่ชี้โดยเทอม A,B,และ E (อาจรวมถึง C และ D ด้วยแต่ไม่จำเป็น) และเป็นโหนดที่ผู้ใช้ได้รับและผ่านการประเมินมาแล้ว 5 ครั้ง

นิยาม 2 : แลททิสโหนด (Lattice nodes)

แลททิสโหนด คือ กลุ่มของโหนดที่แทนเอกสารที่สร้างขึ้น ดังรูปที่ 2 แต่ละโหนดแสดงถึงความลึกหรือระดับของกลุ่มตัวอย่างเอกสารที่ได้รับการประเมิน ว่าตรงตามความต้องการมากน้อยเพียงใด



รูปที่ 1 : แสดงสถานะในเอกภพของเอกสาร

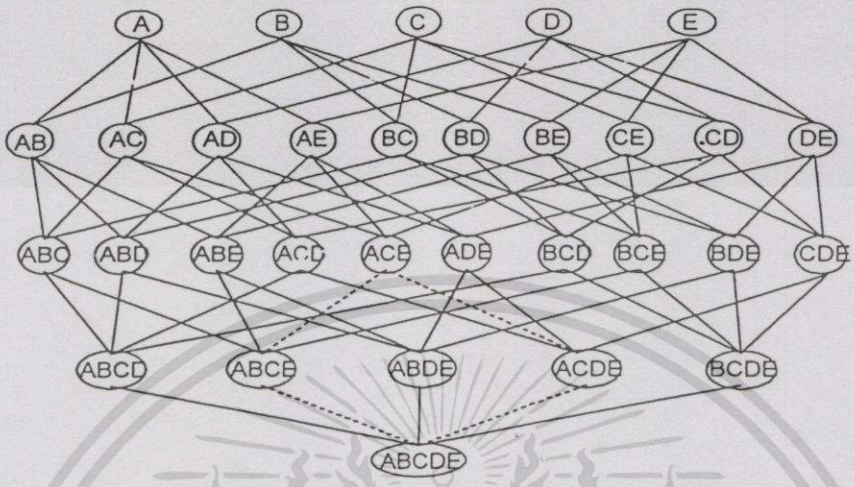
นิยาม 3 : การสุ่มตัวอย่างแลททิสโหนด (Lattice - node - sampling)

การสุ่มตัวอย่าง คือ วิธีการได้มาซึ่งตัวอย่างที่แยกจากเอกสารที่สืบค้นและถูกนำเสนอให้ผู้ใช้พิจารณาว่า ตรงตามคำสั่งที่สืบค้นหรือไม่

การสุ่มเลือกแลททิสโหนดจะใช้เป็นแบบใส่กลับคืน (sampling with replacement) เอกสารใดที่ถูกเลือกเป็นครั้งที่สอง ไม่จำเป็นที่จะต้องนำเสนอให้ผู้ใช้ดูอีก

นิยาม 4 : การสุ่มตัวอย่างที่แท้จริง (Actual - sampling)

เนื่องจากแลททิสโหนดมีส่วนที่คาบเกี่ยวกันอยู่ ดังนั้นเอกสารที่สืบค้นมาอาจมีคุณสมบัติที่สอดคล้องกับหลายๆ แลททิสโหนด ถ้ากลุ่มตัวอย่างแลททิสโหนด เป็นโหนด X และเอกสาร doc ได้รับการสืบค้นมาแล้วแลททิสโหนด Y ทั้งหมดที่ $Y \subseteq X$ และ $doc \in Y$ จะเป็นกลุ่มตัวอย่างที่แท้จริง ของเอก



รูปที่ 2 : แลททิสของซับเซ็ตของเอกสาร

ตัวอย่างในรูปที่ 2 แลททิสโหนดที่เชื่อมกันด้วยเส้นประ (ABCE, ACDE, ABCDE) ล้วนแต่เป็นโหนดที่เป็น กลุ่มตัวอย่างที่แท้จริง ส่วนแลททิสโหนด ACE เป็นกลุ่มตัวอย่างแลททิสโหนด เอกสารที่มีคุณลักษณะตรงกับโหนด ABCDE จะถูกสืบค้นและนำเสนอ

นิยาม 5 : ฟังก์ชั่นปรับค่าความน่าจะเป็น (Update probability function)

เมื่อเอกสาร doc ที่มีคุณลักษณะตรงกับโหนด ABCDE ได้รับการสืบค้น (ดังรูปที่ 2) แล้วค่าความน่าจะเป็นของโหนด ACE, ABCE, ACDE, และ ABCDE จะต้องได้รับการปรับค่า (update) ถ้าเอกสาร doc ตรงตามที่ต้องการและโหนด ABCE ได้รับการเลือกมาแล้ว 4 ครั้ง ซึ่งพบว่า 2 ครั้งตรงตามที่ต้องการ ดังนั้น จึงต้องปรับค่าความน่าจะเป็น P(Rel | ABCE) จาก 2/4 เป็น 3/5 (โหนดอื่นก็เช่นเดียวกัน)

6. อัลกอริทึม IRA [2]

IRA ย่อมาจาก Information Retrieval A*_Algorithm ต่างจากอัลกอริทึม A* ตรงที่ IRA พยายามสืบค้นเพื่อหาโหนดเป้าหมายหลายโหนด (multiple goal nodes) จากเซ็ตของโหนดที่แทนเอกสารทั้งหมด เมื่อพบเอกสารที่อยู่ในข่ายของความสนใจมากกว่า 1 เอกสาร IRA จะจัดลำดับในการนำเสนอต่อผู้ใช้ตามค่าความน่าจะเป็นที่เอกสารจะตรงตามความต้องการ โดยเรียงจากมากไปหาน้อย

อัลกอริทึม IRA มีการใช้ลิสต์ (list) เพื่อจัดเก็บสถานะ ซึ่งมีอยู่ 2 ประเภท คือ ลิสต์ OPEN และลิสต์ CLOSED ลิสต์ OPEN ใช้สำหรับเก็บโหนดที่จะดำเนินการค้นหาในสถานะปัจจุบัน และลิสต์ CLOSED ใช้สำหรับบันทึกโหนดที่ได้ทำการสืบค้น (expanded) เรียบร้อยแล้ว

สิ่งที่นำเข้า (input) ไปในอัลกอริทึมนี้ คือกลุ่มคำสืบค้นที่อยู่ในโหนดที่พิจารณา เอกสารที่เกี่ยวข้อง

กับกลุ่มคำที่อยู่ในโหนด ค่า Threshold ที่ผู้ใช้ใส่ และการประเมินเอกสารว่าตรงตามความต้องการหรือไม่จากผู้ใช้ เมื่ออัลกอริทึมนี้ได้ถูกดำเนินการจนเสร็จสิ้นแล้วจะได้สิ่งที่เป็นผลลัพธ์ (output) คือโหนดเป้าหมาย ซึ่งมีจำนวนเอกสารที่เกี่ยวข้องมากกว่าค่า Threshold อัลกอริทึม IRA มีรายละเอียดดังนี้

```
begin IRA
    create_initial_search_tree
    OPEN = start_node;
    CLOSED = [ ];
    while OPEN <> [ ] และ
        (จำนวนเอกสารที่เกี่ยวข้อง < ค่า Treshold
    หรือ จำนวนโหนดเป้าหมายที่พบ < ค่าที่กำหนด)
    do  begin
        ·n = select_doc_node (OPEN);
        ·ทำการสุ่มกลุ่มตัวอย่างแลททิสโหนด
        กับเอกสารจากแลททิสโหนดที่มีความ
          สัมพันธ์กับ n
        ·OPEN = OPEN - [n];
        ·CLOSED = CLOSED ∪ [n];
        ·นับจำนวนโหนดเป้าหมาย หรือจำนวน
          เอกสาร ที่เกี่ยวข้อง (ที่ได้ถูกค้นพบแล้ว)
        ·successor_nodes = expand_doc_node(n);
        ·OPEN = OPEN ∪ (successor_nodes -
          CLOSED);
        ·ทำการปรับค่าของแต่ละโหนดจากการสุ่ม
          ตัวอย่างจริง
        ·จัดรูปแบบ search tree ของเอกสารใหม่ :
            ·สำหรับโหนด ที่ทำการปรับค่าทั้งหมด
            ให้เชื่อมโยงโหนดที่เกี่ยวข้องเข้ากับ
            search tree
            ·เชื่อมโยงทุกๆ leaf nodes เข้ากับโหนด
            ที่เป็น predecessors ของมันพร้อมกับ
            ลบโหนดที่มี successor อยู่ใน search
            tree ออกจาก OPEN แล้วนำไปผนวก
            กับ CLOSED
    end while;
end IRA.
```

7. ฟังก์ชั่นฮิวริสติกสำหรับ IRA [2]

ดังที่กล่าวมาแล้วข้างต้นในอัลกอริทึม A* ว่า $g^*(n)$ แทนค่าใช้จ่าย(จำนวนครั้ง) ที่เอกสารซึ่งแทนด้วยโหนด n ถูกประเมิน $h^*(n)$ แทนค่าประมาณจำนวนเอกสารที่จะต้องทำการสืบค้นจากโหนด n จนกว่าจะพบเอกสารที่ต้องการ (สมมติว่าเป็นโหนดที่ G) โหนด n ที่มีค่า f^* ที่น้อยที่สุดจะถูกกระจายออก ถ้าโหนด n มีความลึกเท่ากับ b และมีเอกสารจำนวน a ฉบับที่ถูกสืบค้นแล้ว จะได้ว่าฟังก์ชันฮิวริสติกสำหรับ IRA คือ

$$f^*(n) = b + (G-a)*(b/a) \qquad (1)$$

<u>พิสูจน์</u>

ถ้า b เป็นความลึกของโหนด n แล้ว $g^*(n)$ จะเป็นค่าประมาณของค่าความน่าจะเป็นที่เอกสาร ที่แทนด้วย n จะตรงตามความต้องการ โดยมีค่า p เท่ากับ a / b และคาดว่าจะมีจำนวนของเอกสารที่ยังไม่ได้รับการสืบค้นอีก G-a ฉบับ โดยที่มีเอกสารที่ไม่ตรงตามความต้องการ k ฉบับที่พบก่อนหน้าเอกสารที่ตรงตามความต้องการฉบับที่ (G-a)th ดังนั้นจะได้ว่า

$$h^*(n) = E(G-a+k) = G-a+E(k) \qquad (2)$$

และเนื่องจากค่า k สอดคล้องตามการกระจายแบบไบโนเมียล (binomial distribution) จึงได้

$$h^*(n) = G - a + (G - a)*(q/p) \quad (3)$$

โดยที่ p = 1 - q เป็นค่าประมาณของความน่าจะเป็นที่โหนด n จะตรงตามความต้องการ ดังนั้นฟังก์ชันฮิวริสติคดังกล่าวจะเป็นดังนี้

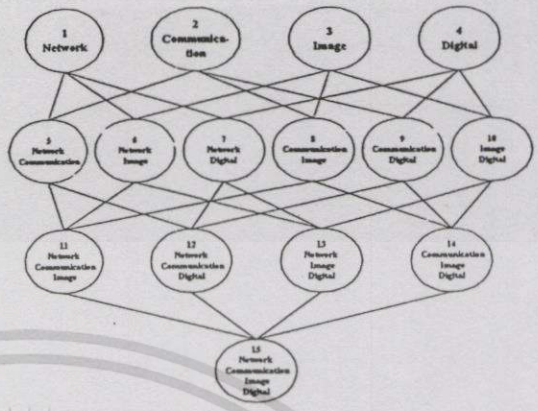
$$\begin{aligned} f^*(n) &= g^*(n) + h^*(n) \\ &= b + (G - a) + (G - a)*(q/p) \\ &= b - (G - a) + (G - a)*[(b - a)/a] \\ &= b + (G - a)*(b/a) \end{aligned} \quad (4)$$

8. การทดลอง

จากอัลกอริทึม A* และ อัลกอริทึม IRA ดังที่กล่าวมาแล้วข้างต้น ได้นำมาประยุกต์ใช้เพื่อการสืบค้นเอกภพของเอกสาร โดยสร้างแบบจำลอง ที่ใช้เทคนิคฮิวริสติคเข้ามาช่วยการค้นคืนสารสนเทศ โดยยึดอัลกอริทึม IRA เป็นหลัก

ตัวอย่างการทดลอง

ภายในฐานข้อมูลของระบบสืบค้นที่ทำการทดลอง มีเอกสารทั้งหมด 500 ฉบับ และมีคำเฉพาะ 400 คำ ผู้ใช้ได้ใส่คำเฉพาะเข้าไปในระบบจำนวน 4 คำ ดังนี้ Network, Communication, Image, Digital ระบบสามารถนำคำเฉพาะทั้ง 4 คำ มาสร้างแลทิสโหนด และกำหนดความสัมพันธ์ของแต่ละโหนด กับเอกสารที่มีอยู่ในฐานข้อมูล ซึ่งเอกสารที่ได้ทั้งหมดจากคำสืบค้นดังกล่าวมีจำนวน 215 ฉบับ ดังรูปที่ 3 และกำหนดจำนวนเอกสารในแต่ละโหนดดังแสดงไว้ในตารางที่ 2



รูปที่ 3 : แสดงแลทิสโหนดที่สร้างจากคำเฉพาะ 4 คำ ที่ใส่เข้าไปในระบบ

เมื่อสร้างแลทิสโหนดได้แล้ว ระบบจะทำการสุ่มกลุ่มตัวอย่างแลทิสโหนด โดยกำหนดให้

Level : ระดับความลึกของโหนด โดยจะให้ค่าดังนี้

Level 1 : โหนดทุกโหนดที่มีคำเฉพาะ 1 คำ

Level 2 : โหนดทุกโหนดที่มีคำเฉพาะ 2 คำ

Level 3 : โหนดทุกโหนดที่มีคำเฉพาะ 3 คำ

Level 4 : โหนดทุกโหนดที่มีคำเฉพาะ 4 คำ

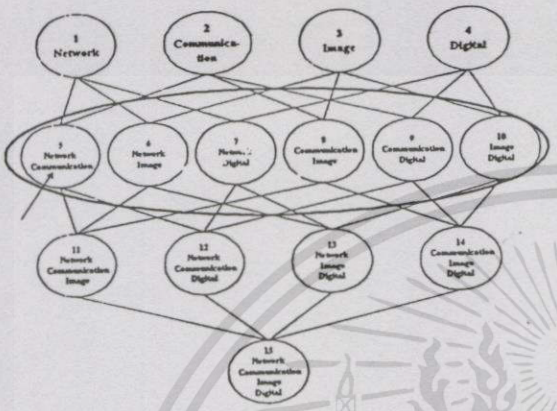
หลักเกณฑ์ในการคำนวณ Level ที่จะเริ่มสุ่มกลุ่มตัวอย่างแลทิสโหนด คือ

$$\text{Level (เริ่มต้น)} = \left\lceil \frac{\text{จำนวนคำสืบค้น}}{2} \right\rceil \quad (5)$$

เพื่อที่จะจำกัดจำนวนโหนดในการเลือกครั้งแรกให้เหมาะสมจึงใช้หลักการต่อไปนี้ :

แลทิสโหนดที่มีคำเฉพาะน้อยกว่าค่า Level ที่กำหนดไว้ ไม่ต้องพิจารณา เพราะว่าโหนดเหล่านั้น มีค่าความน่าจะเป็นที่ตรงตามความต้องการน้อยมาก เมื่อเทียบกับโหนดที่มีคำเฉพาะมากกว่า

ซึ่งในกรณีนี้ เราจะเริ่มสุ่มกลุ่มตัวอย่างจากโหนดที่มีคำเฉพาะ 2 คำ ดังรูปที่ 4



รูปที่ 4 : แสดงถึง Level ที่เริ่มทำการสุ่ม และโหนดแรกที่เริ่มสุ่ม

จากรูปที่ 4 ตามอัลกอริทึม IRA ระบบทำการ OPEN โหนดที่อยู่ในบริเวณวงรี และ CLOSED โหนดที่อยู่ Level เหนือขึ้นไป คือ Level ที่ 1 ส่วน Level 3 กับ 4 นั้นเป็น Successor nodes ซึ่งยังไม่ถูกประเมินในตอนนี้

การเลือกโหนด ต้องเลือกโหนดที่อยู่ในเซ็ต OPEN เท่านั้น ซึ่งสามารถพิจารณาตามลำดับดังนี้

1. การพิจารณาโหนดที่ถูกประเมินแล้วมีค่า $f^*(n)$: โดยเลือกโหนดที่มีค่า $f^*(n)$ น้อยที่สุด หากโหนดที่น้อยที่สุดนั้น มีค่า $f^*(n)$ เท่ากันให้เลือกโหนดที่อยู่ซ้ายสุด

2. การพิจารณาโหนดที่ถูกประเมินแล้วมีค่า $f^*(n)$ ซึ่งหาค่าไม่ได้ หรือ โหนดที่มีค่า a = 0 : ไม่ต้องพิจารณา

3. การพิจารณาโหนดที่ไม่เคยถูกประเมินมาก่อน : ให้เลือกโหนดที่อยู่ซ้ายสุดก่อน

ในการทดลองนี้ เป็นการเลือกโหนดในกรณีที่ 3 ดังนั้น การสุ่มตัวอย่างจะเริ่มตั้งแต่โหนดที่ลูกศรชี้อยู่คือโหนดที่มีคำว่า Network และ Communication (โหนดหมายเลข 5) ซึ่งเป็น*กลุ่มตัวอย่างแลททิสโหนด* ส่วนโหนดที่เป็น*กลุ่มตัวอย่างที่แท้จริง* คือ โหนดที่มีเส้นเชื่อมกับโหนด ที่เป็นกลุ่มตัวอย่างแลททิสโหนด (โหนดใน Level 3 และ 4 ทุกโหนด ที่มีคำว่า Network และ Communication ซึ่งได้แก่ โหนดหมายเลข 11, 12 และ 15)

ทำการสุ่มเอกสาร ในแลททิสโหนดดังกล่าวขึ้นมาจำนวน b ฉบับ (กำหนด b เท่ากับ 25% ของเอกสารในโหนด) ให้ผู้ใช้ประเมินว่าตรงตามความต้องการของคนเองหรือไม่ ตัวอย่างเช่น

โหนดหมายเลข 5 มีเอกสารทั้งหมด 78 ฉบับ สุ่มขึ้นมา 25% จะได้ 19 เอกสาร โดยให้ผู้ใช้ทำการประเมินความต้องการ ซึ่งได้ผลดังตารางที่ 1

การสุ่มตัวอย่างแลททิสโหนด (Lattice - node - sampling)			การสุ่มตัวอย่างที่แท้จริง (Actual - sampling)		
เอกสาร	ผล	a/b (5)	a/b (11)	a/b (12)	a/b (15)
Doc#1	ตรง	1/1	0/0	1/1	0/0
Doc#2	ไม่ตรง	1/2	0/1	1/1	0/0
Doc#3	ตรง	2/3	0/1	2/2	1/1
...	...	...	...	...	...
Doc#17	ตรง	6/17	0/3	5/7	1/1
Doc#18	ไม่ตรง	6/18	0/3	5/8	1/2
Doc#19	ไม่ตรง	6/19	0/3	5/9	1/2

หมายเหตุ a : จำนวนเอกสารที่ตรงตามความต้องการ
b : จำนวนเอกสารที่สุ่ม

ตารางที่ 1 : แสดงการปรับค่าความน่าจะเป็นที่กลุ่มตัวอย่างแลททิสโหนดหมายเลข 5 และกลุ่มตัวอย่างที่แท้จริงของโหนดหมายเลข 5

ในระบบจะมีค่า G เป็นค่า Threshold ของระบบโดยมีข้อกำหนดคือ จำนวนเอกสารที่ผู้ใช้คาดว่าจะตรงตามความต้องการ เพื่อนำไปเป็นเกณฑ์ในการพิจารณาโหนดเป้าหมาย (ในที่นี้ผู้ใช้กำหนดว่า ต้องการเอกสาร 25 ฉบับ)

การพิจารณาโหนดเป้าหมาย ทำได้โดยพิจารณาค่าประมาณของจำนวนเอกสารที่ตรงความต้องการ ว่ามากกว่าหรือเท่ากับค่า G หรือไม่ ในที่นี้ค่าประมาณของโหนดที่ 5 เท่ากับ 6/19*78 = 24.63 ซึ่งน้อยกว่า 25 ดังนั้นโหนดที่ 5 นี้ยังไม่ใช่โหนดเป้าหมายของผู้ใช้

เมื่อระบบทำการปรับค่าความน่าจะเป็นของโหนดทั้งหมดแล้ว ระบบจะนำค่าที่ประเมินความเกี่ยวข้องนั้น มาปรับค่าฮิวริสติคประจำโหนด ทุกโหนดจะถูกปรับค่าความน่าจะเป็น ตามฟังก์ชัน $f^*(n)$ ที่ได้กล่าวไปแล้วในสมการที่ (4) ดังนั้นค่า f^* ของโหนดหมายเลข 5, 11, 12 และ 15 คือ

$f^*(5) = 19 + (25 - 6)*(19/6)$
$= 79.16$

$f^*(11) = 3 + (25 - 0)*(5/0)$
$= -$

$f^*(12) = 9 + (25 - 5)*(15/5)$
$= 69.00$

$f^*(15) = 2 + (25 - 1)*(3/1)$
$= 74.00$

ระบบจะบันทึกค่า $f^*(n)$ ของแต่ละโหนดเพื่อใช้พิจารณาในกรณีอื่นๆ ต่อไป

จากนั้นโหนดหมายเลข 5 จะถูกลบออกจากเซ็ต OPEN และระบบจะทำการขยายเซ็ต OPEN ออกไป โดยการรวมเอาโหนดที่เป็น Successor nodes ของโหนดหมายเลข 5 เข้ามารวมด้วยได้แก่ โหนดหมายเลข 11 และ 12 ดังนั้น เซ็ต OPEN ในตอนนี้มีโหนดหมายเลข 6, 7, 8, 9, 10, 11 และ 12 เป็นสมาชิก

ในรอบต่อไปของการสืบค้น โหนดหมายเลข 6 จะถูกเลือกขึ้นมาเพื่อทำการประเมินและปรับค่าความน่าจะเป็นของเอกสารในโหนด และปรับค่าฟังก์ชันของโหนด ดังวิธีการที่ได้กล่าวไปแล้วตามลำดับ จนกระทั่งพบโหนดหมายเลข 12 ซึ่งเป็นโหนดเป้าหมาย คือมีจำนวนเอกสารที่ตรงตามความต้องการ 29 ฉบับ ซึ่งมีค่ามากกว่าค่า G

ท้ายที่สุดผู้ใช้ หรือผู้สืบค้นจะได้รับผลลัพธ์สุดท้ายเป็นเอกสารทั้งหมดที่อยู่ในโหนดหมายเลข 12 ซึ่งมีคำสืบค้นเกี่ยวข้อง 3 คำคือ Network, Communication และ Digital ดังนั้น คำสืบค้นทั้ง 3 คำนี้ จึงเป็นกลุ่มคำที่เหมาะสมมากที่สุดสำหรับการสืบค้นในครั้งนี้

อัลกอริทึมนี้จะจบการทำงานก็ต่อเมื่อได้โหนดเป้าหมาย หรือได้เอกสารที่เกี่ยวข้องครบตามที่ผู้ใช้ต้องการ

การวัดประสิทธิภาพของระบบสืบค้น

การสืบค้นแบบฮิวริสติคมีเป้าหมายในการค้นคืนเอกสารให้ตรงตามความต้องการของผู้ใช้มากที่สุด และให้ได้เอกสารที่ไม่ตรงตามความต้องการออกมาให้น้อยที่สุด และเมื่อมีการค้นคืนเอกสารออกมาได้แล้ว จากนั้นเราจะมาทำการเปรียบเทียบว่าเอกสารนั้นมีความถูกต้องตรงกับความต้องการมากน้อยเพียงใด การวัดประสิทธิภาพของการค้นคืนมีอยู่หลายวิธี แต่สองวิธีที่มักนิยมใช้กันคือ การวัดค่า Precision และค่า Recall [9]

Precision (P) : เป็นอัตราส่วนของเอกสารที่ถูกค้นคืนและตรงตามความต้องการกับเอกสารทั้งหมดที่ทำการค้นคืนได้ มีสูตรดังนี้

$$P = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ค้นคืนได้}}{\text{จำนวนเอกสารทั้งหมดที่ค้นคืนได้}} \quad (6)$$

Recall (R) : เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ได้ถูกค้นคืนกับจำนวนเอกสารที่เกี่ยวข้องทั้งหมด มีสูตรดังนี้

$$R = \frac{\text{จำนวนเอกสารที่เกี่ยวข้องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่เกี่ยวข้องทั้งหมดในฐานข้อมูล}} \quad (7)$$

จะเห็นได้ว่าค่า Precision วัดความสามารถของระบบในการกำจัดเอกสารที่ไม่ต้องการออกไป เพื่อให้ได้ความแม่นยำในการสืบค้น ส่วนค่า Recall วัดความสามารถของระบบในการค้นคืนเอกสารที่ตรงตามความต้องการ หากจะกล่าวว่าระบบควรมีค่าใดมากหรือน้อยนั้น ไม่สามารถสรุปได้ เนื่องจากค่าทั้ง 2 ค่านี้ ขึ้นอยู่กับผู้ใช้ว่าต้องการเอกสารที่มีลักษณะอย่างไร ดังนั้นถ้าผู้ใช้สนใจค่า Precision ที่สูงจะต้องใส่คำสืบค้นที่เฉพาะเจาะจงลงไป (ใส่คำสืบค้นหลายคำ) เพื่อให้ได้เอกสารออกมาจำนวนน้อยแต่น่าจะตรงตามความต้องการมาก ในขณะที่หากผู้ใช้ต้องการค่า Recall ที่สูงจะต้องยอมรับคำสืบค้นที่ให้เอกสารออกมาจำนวนมาก (ใส่คำสืบค้นน้อยคำ)

ผลการทดลอง

ผลการทดลองที่ได้ แสดงถึงการสุ่มจำนวนเอกสารในแต่ละโหนด ที่ประกอบอยู่ในแลททิสโหนดตัวอย่าง ค่า Precision และ Recall ของโหนด ดังตารางที่ 2

จากตารางที่ 2 จะเห็นได้ว่าโหนดที่ประกอบไปด้วยกลุ่มคำที่เหมาะสมที่สุดสำหรับการสืบค้น คือ โหนดหมายเลข 12 มีค่า Precision เท่ากับ 0.46 และค่า Recall เท่ากับ 0.80 ซึ่งการสืบค้นแบบฮิวริสติค มุ่งเน้นในเรื่องของการเลือกเอกสารที่ตรงต่อความต้องการของผู้ใช้มากที่สุด จึงพิจารณาค่า Recall เป็นสำคัญ ดังนั้นให้พิจารณาโหนดที่มีค่า Recall มากที่สุดก่อน ซึ่งได้แก่โหนดหมายเลข 5, 9 และ 12 เมื่อทั้ง 3 โหนดมีค่า Recall เท่ากัน จึงเลือกโหนดที่มีค่า Precision มากที่สุด ได้แก่โหนดหมายเลข 12 นั่นเอง

9. บทสรุปและข้อเสนอแนะ

อัลกอริทึม A* และ อัลกอริทึม IRA ได้นำมาประยุกต์ใช้เพื่อการสืบค้นเอกภพของเอกสาร ซึ่งการประยุกต์การสืบค้น โดยใช้เทคนิคฮิวริสติคเข้ามาช่วยการค้นคืนสารสนเทศนั้น ผลที่ได้ทำให้การสืบค้นได้เอกสารที่ตรงตามความต้องการของผู้ใช้ เนื่องจากอัลกอริทึมดังกล่าวทำให้ผู้ใช้สามารถเลือกกลุ่มคำสืบค้นที่เหมาะสมที่สุดเพื่อทำให้ได้เอกสารที่ต้องการ

จากผลการทดลองสรุปได้ว่า ผู้ใช้จะต้องประเมินเอกสารในกลุ่มตัวอย่างแลททิสโหนดที่ระบบสุ่มขึ้นมาว่าตรงตามความต้องการหรือไม่ จากนั้นระบบจะนำคำตอบที่ได้จากผู้ใช้ไปปรับค่าความน่าจะเป็นของเอกสารในแต่ละโหนด ที่เป็นทั้งกลุ่มตัวอย่างแลททิสโหนด และกลุ่มตัวอย่างที่แท้จริง มีปัจจัย 2 ประการที่มีผลกระทบกับผลลัพธ์ที่ได้คือ การสุ่มเอกสารและการกำหนดค่า G ของผู้ใช้ ซึ่งในขณะนี้ทางผู้วิจัยกำลังพัฒนาวิธีการสุ่มให้มีความน่าเชื่อถือและลดปริมาณเอกสารที่ผู้ใช้ต้องพิจารณา ให้อยู่ในช่วงที่เหมาะสม รวมถึงการกำหนดค่า G ที่ใกล้เคียงกับเอกสารที่ควรจะเป็น

โหนด	คำสืบค้น	เอกสารในโหนด (ฉบับ)	จำนวนเอกสารที่สุ่ม (25%)	X (ฉบับ)	ค่า Precision	ค่า Recall
1.	Network	80	20	3	0.15	0.40
2.	Communication	123	31	5	0.16	0.66
3.	Image	98	25	0	0.00	0.00
4.	Digital	150	38	5	0.13	0.66
5.	Network + Communication	78	20	6	0.30	0.80
6.	Network + Image	23	6	0	0.00	0.00
7.	Network + Digital	65	7	4	0.57	0.53
8.	Communication + Image	30	16	4	0.25	0.53
9.	Communication + Digital	74	19	6	0.31	0.80
10.	Image + Digital	90	23	4	0.17	0.53
11.	Network + Communication + Image	20	5	2	0.40	0.26
12.	Network + Communication + Digital	60	15	7	0.46	0.80
13.	Network + Image + Digital	23	6	1	0.16	0.13
14.	Communication + Image + Digital	25	6	2	0.33	0.26
15.	Network + Communication + Image + Digital	12	3	2	0.66	0.26

X = จำนวนเอกสารที่ตรงความต้องการที่ค้นคืนได้โดยระบบ *จำนวนเอกสารที่ตรงตามความต้องการที่แท้จริง = 30 ฉบับ

ตารางที่ 2 : แสดงค่า Precision และ Recall ในแต่ละโหนด

ในเรื่องของการวัดประสิทธิภาพแบบวัดค่า Precision และ ค่า Recall ของระบบการสืบค้นแบบฮิวริสติคนี้ระบบจะเน้นเรื่องของการหากลุ่มคำสืบค้นที่เหมาะสมให้กับผู้ใช้ ทำให้ผู้ใช้ได้เอกสารที่ตรงตามความต้องการเพียงพอ โดยที่มีเอกสารที่ไม่ตรงความต้องการน้อยที่สุด

ส่วนเรื่องของเวลาที่ใช้ของระบบสามารถแบ่งออกได้เป็น 3 ลักษณะ ดังนี้

1. เวลาของการค้นคืนข้อมูล : ทำการค้นคืนข้อมูลจากฐานข้อมูลปกติทั่วไป โดยโปรแกรมจะส่งคำสั่ง SQL เพื่อดึงข้อมูลที่ต้องการออกมา ซึ่งของความเร็วเรื่องการอ่าน/เขียน ฮาร์ดดิสก์ เหมือนการ run โปรแกรมโดยทั่วไป

2. เวลาของอัลกอริทึม : การ expand node ขึ้นกับจำนวนคำเฉพาะที่ผู้ใช้ใส่ลงไป ถ้าใส่คำเฉพาะมากคำ จะช้ากว่าใส่คำเฉพาะน้อยคำ แต่จะแตกต่างกันเพียงเล็กน้อยเท่านั้น ทั้งนี้ขึ้นอยู่กับความเร็วของ CPU ของเครื่องคอมพิวเตอร์ที่ใช้ด้วย

3. เวลาในการประเมินเอกสารของผู้ใช้ : ขึ้นอยู่กับผู้ใช้ว่าต้องการประเมินเอกสารจำนวนมากหรือน้อย

ข้อเสนอแนะบางประการที่น่าสนใจ ซึ่งผู้วิจัยได้ทำการศึกษาระหว่างทำการทดลองคือ การนำข้อมูลประวัติและความสนใจของผู้ใช้ (User profile) มาช่วยในเรื่องของการสืบค้นด้วย ซึ่งมีรายละเอียดโดยสังเขปดังนี้ [3]

ข้อมูลประวัติและความสนใจของผู้ใช้ประกอบไปด้วย ข้อมูลส่วนตัวของผู้ใช้และข้อมูลเกี่ยวกับความสนใจที่ผู้ใช้ต้องการ ซึ่งฐานข้อมูลของระบบสืบค้น จะทำการรวบรวมข้อมูลของการสืบค้นครั้งก่อนเพื่อเพิ่มประสิทธิภาพในการค้นคืนให้กับระบบ โดยข้อมูลพื้นฐานของผู้ใช้นี้จะช่วยในการกำหนดว่าเอกสารใดบ้างที่อยู่ในกลุ่มของเอกสารที่จะถูกค้นคืน ข้อมูลเหล่านี้อาจได้มาจาก บรรณารักษ์อ้างอิง หรือการสัมภาษณ์ผู้ใช้ซึ่งอาจพบข้อมูลในลักษณะดังต่อไปนี้

1. ระดับการศึกษา (Educational level)

ผู้ใช้ที่มีระดับการศึกษาต่างกันอาจต้องการกลุ่มเอกสารที่แตกต่างกัน โดยการใช้คำสืบค้นที่เหมือนกัน

2. ความคุ้นเคยกับขอบเขตของการถาม (Familiarity with the area of inquiry) คือ

ต้องรู้ว่าผู้ใช้น่าจะถามในขอบเขตอะไร หรือผู้ใช้สนใจในขอบเขตเรื่องใด

3. ความสามารถทางด้านภาษาของผู้ใช้ (Language capabilities)

บทความที่เขียนโดยภาษาที่ผู้ใช้ไม่คุ้นเคย จะมีประโยชน์น้อยหากผู้ใช้ไม่สามารถอ่านภาษานั้นได้ เช่น ผู้ใช้ไม่รู้ภาษาญี่ปุ่นแต่ในฐานข้อมูลมีเอกสารที่เป็นภาษาญี่ปุ่น เอกสารนั้นจะตรงความต้องการน้อย เป็นต้น

4. การเป็นสมาชิกวารสาร (Journal subscriptions)

วารสารที่ผู้ใช้เป็นสมาชิก หรืออยู่ในลิสต์ที่เคยอ่านแล้ว จะทำการเก็บไว้เป็นแหล่งอ้างอิง

5. นิสัยการอ่าน (Reading habits)

ถ้าผู้ใช้อ่านวารสารฉบับหนึ่งเป็นประจำ ผู้ใช้จะรู้เกี่ยวกับบทความที่สำคัญในวารสารนั้น จึงเก็บชื่อวารสารนั้นไว้อ้างอิงว่าผู้ใช้คนนี้ชอบอ่านวารสารประเภทนี้ เป็นต้น หากผู้ใช้ยังไม่เคยอ่านวารสารนั้นมาก่อน และไม่มีชื่อในลิสต์ วารสารนั้นจะถูกนำไปอ้างอิงเป็นเอกสารตัวใหม่

6. สิ่งที่ชอบโดยส่วนตัว (Specific preference)

ผู้ใช้อาจมีผู้เขียนที่ชอบ หรือมีวารสารที่ชอบและติดตามเป็นประจำอยู่แล้ว ดังนั้นผู้ใช้น่าจะสนใจวารสารชิ้นนี้มากกว่าชิ้นอื่น

ข้อมูลนี้ ไม่สามารถใช้ได้โดยตรงในกระบวนการค้นคืน แต่ประยุกต์ใช้ในเรื่องการจัดการเกี่ยวกับความเหมาะสมในเรื่องของการสืบค้น เพื่อให้ตรงตามความต้องการของผู้ใช้และกำจัดเอกสารบางชิ้นที่ไม่เหมาะสมออกไปเท่าที่จะเป็นไปได้

เอกสารอ้างอิง

[1] ANDREW, A. M. Continuous Heuristic The Prelinguistic Basis of Intelligence. New York:Ellis Horwood,1993.

[2] HOFFERER, M. "Heuristic Search in Information Retrieval." , Information Retrieval : New Systems and Current Research. Taylor Graham, London,pp. 81-90.,1994.

[3] KORFHAGE, R. R.Information Storage and Retrieval. Wiley Computer Publishing,USA.,1997.

[4] LANCASTER, F. W. Information Retrieval Systems:Characteristics, Testing and Evaluation. A Wiley-Interscience Publication,US.,1979.

[5] LUGER, G. F. and STUBBLE FIELD,W. A. Artificial Intelligence Structures and Strategies for complex problem solving. The

Benjamin/Cummings Pubblishing Company, Inc. Redwood City, California.,1993.

[6] MERKL,W. et.al. "A hypertexi oriented user-interface for an intelligent legal fulltext information retrieval system",Proceedings International Conference on Database and Expert Systems Applications,A.M. Tjoa and R.Wagner (eds). Vienna,August 1990.

[7] PATTERSON, D. W. Introduction to Artificial Intelligence and Expert Systems. Prentice-Hall, US.,1991.

[8] RUSSELL,S.J. and NORVIG, P. Artificial Intelligence A Modern Approach. Prentice-Hall,US,pp.94.,1995.

[9] SALTON, G. and McGill, M.J. Introduction to Modern Information Retrieval. McGraw-Hill, Singapore,1984.

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาววรางคณา เงินแก้ว
วัน/เดือน/ปี เกิด	15 มกราคม พศ. 2518
วุฒิการศึกษาระดับปริญญาตรี	วิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ปีการศึกษา 2539
ประสบการณ์ในการทำงาน	นักวิชาการคอมพิวเตอร์ ฝ่ายระบบและโปรแกรม สำนักวิจัยและบริการคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหาร ลาดกระบัง (สจล.) ปีพศ. 2540 - ปัจจุบัน