
# รายงานการวิจัย

เรื่อง

การเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยสำหรับระบบเครือข่ายสังคมโดยใช้เทคนิคเหมืองข้อมูล
(Enhancement ranking search result for research paper social bookmarking by using data mining technique)

โดย

พิจิตรา จอมศรี

ได้รับทุนอุดหนุนจากมหาวิทยาลัยราชภัฏสวนสุนันทา

ปีงบประมาณ 2555

# รายงานการวิจัย

เรื่อง

การเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยสำหรับระบบเครือข่ายสังคมโดยใช้เทคนิคเหมืองข้อมูล

(Enhancement ranking search result for research paper social bookmarking by using data mining technique)


| คณะผู้วิจัย | สังกัด |
| --- | --- |
| 1.พิจิตรา จอมศรี | คณะวิทยาศาสตร์และเทคโนโลยี |



ได้รับทุนอุดหนุนจากมหาวิทยาลัยราชภัฏสวนสุนันทา

ปีงบประมาณ 2555

ชื่อรายงานการวิจัย : การเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยสำหรับระบบเครือข่ายสังคมโดยใช้เทคนิคเหมืองข้อมูล
ชื่อผู้วิจัย : นางสาวพิจิตรา จอมศรี
ปีที่ทำการวิจัย : 2555

## บทคัดย่อ

ปัจจุบันการสืบค้นข้อมูลผ่านระบบเครือข่ายสังคมด้านงานวิจัยถือเป็นที่นิยมอย่างยิ่งโดยเฉพาะการสืบค้นข้อมูลด้านการศึกษาและงานวิจัย เพื่อเป็นการทบทวนงานวิจัยและศึกษาถึงเทคนิควิธีที่ใช้ในการทำงานวิจัยกับกลุ่มคนที่สนใจในเรื่องเดียวกัน งานวิจัยนี้จึงได้พัฒนาเทคนิคเพื่อใช้ในการเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยด้วยเทคนิคเหมืองข้อมูล โดยการนำกฎการวิเคราะห์ความสัมพันธ์มาทำการวิเคราะห์ความสัมพันธ์ของกลุ่มผู้ใช้บนเครือข่ายสังคมด้านงานวิจัย ทั้งนี้วัตถุประสงค์ของงานวิจัยครั้งนี้คือการสร้างกระบวนการที่ใช้ในการวิเคราะห์ความสัมพันธ์ของกลุ่มผู้ใช้งานในระบบเครือข่ายสังคม โดยการศึกษาถึงความสัมพันธ์ของกลุ่มผู้ใช้ในระบบเครือข่ายสังคมด้านงานวิจัยที่ผู้ใช้แต่ละคนสนใจและเป็นสมาชิก เพื่อนำความสัมพันธ์ที่พบมาเป็นแนวคิดในการจัดอันดับผลการค้นหางานวิจัยที่เหมาะสมกับผู้ใช้แต่ละคน

ซึ่งจากผลการทดลองพบว่าประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยโดยการนำกฎการวิเคราะห์ความสัมพันธ์ของพฤติกรรมผู้ใช้ที่เป็นสมาชิกในแต่ละกลุ่มมาช่วยในการเรียงผลการค้นหานั้นสามารถเพิ่มประสิทธิภาพการค้นหางานวิจัยบนระบบโซเชียลบุ๊กมาร์กได้

## Abstract

Research Title :Enhancement ranking search result for research paper social bookmarking by using data mining technique
Author : Ms. Pijitra Jomsri
Year : 2012

## ABSTRACT

Currently searching through social network is very popular especially in a field of academic because the systems help user to review papers and survey other technique that his/her interest in the groups. Therefore community-base web sites have been developed to help user search information more easily from process of customizing a web site to need each specifies user or set of user. In this paper use association rule to analyzed the community group on research paper bookmarking. A design goal for community group frameworks is developed and discussed. Additionally Researcher analyzes the initial relation by using association rule discovery between the antecedent and the consequent of a rule in the groups of user for generate the idea to improve ranking search result and development recommender system.

This technique analyses the relation of user behavior which is member in each groups on research paper social bookmarking. Furthermore the result can improve efficiency of research paper ranking by using association rule.

## กิตติกรรมประกาศ

เอกสารงานวิจัยฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลือเป็นอย่างดี จากวิชาสาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏสวนสุนันทา และ www.CiteULike.org ที่ให้ความอนุเคราะห์ด้านข้อมูลในการทดลอง

เนื้อหาของเอกสารงานวิจัยเล่มนี้จะมีความสมบูรณ์และถูกต้องไม่ได้ หากไม่ได้รับความอนุเคราะห์จากสำนักวิจัยและพัฒนา และเจ้าหน้าที่จากมหาวิทยาลัยราชภัฏสวนสุนันทาที่กรุณาให้ความเอื้อเฟื้อในการใช้เครื่องมือที่อำนวยความสะดวกในการจัดทำรูปเล่มเอกสารงานวิจัย เพื่อนทุกท่านที่ให้ความช่วยเหลือ และคำแนะนำในการเอกสารงานวิจัย รวมทั้งคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา ที่เอื้ออำนวยคอมพิวเตอร์และอินเทอร์เน็ตเพื่อใช้ในการสืบหาข้อมูลต่างๆ ในการทำเอกสารงานวิจัยเป็นอย่างดี ผู้จัดทำจึงขอขอบคุณมา ณ โอกาสนี้

ท้ายสุดนี้ ขอกราบขอบพระคุณบิดา มารดา และขอขอบคุณเพื่อนร่วมงานที่ได้ช่วยส่งเสริมสนับสนุนกระตุ้นเตือน และเป็นกำลังใจตลอดมาให้ผู้เขียนจัดทำรายงานการวิจัย

พิจิตรา จอมศรี

สิงหาคม 2555

# สารบัญ

# สารบัญ (ต่อ)

# สารบัญตาราง

# สารบัญรูป

# บทที่ 1
# บทนำ

## 1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการใช้งานเว็บไซต์เครือข่ายสังคมเพื่อสืบค้นข้อมูลมีแนวโน้มเพิ่มมากขึ้น โดยเฉพาะอย่างยิ่งการค้นหาข้อมูลทางด้านงานวิจัย นักวิจัยส่วนใหญ่นิยมสืบค้นงานวิจัยที่เกี่ยวข้องผ่านทางเว็บไซต์ที่ให้บริการสืบค้นข้อมูล เช่น IEEE , ACM , Springer นอกจากนี้เว็บไซต์เครือข่ายสังคม (Social network) ซึ่งเป็นรูปแบบของเว็บไซต์ ในการสร้างเครือข่ายสังคม สำหรับผู้ใช้งานในอินเทอร์เน็ต และโซเชียลบุ๊กมาร์ก (Social bookmarking) ซึ่งเป็นการให้บริการบนเว็บที่แบ่งบันการบุ๊กมาร์กบนอินเทอร์เน็ต โดยเว็บไซต์บริการโซเชียลบุ๊กมาร์กเป็นที่นิยมในการจัดเก็บ แบ่งหมวดหมู่ และแบ่งปันข้อมูลให้กับผู้ใช้ที่สนใจข้อมูลในลักษณะเดียวกัน ซึ่งเว็บไซต์ดังกล่าวยังถือเป็นอีกทางเลือกหนึ่งที่ช่วยให้นักวิจัยสามารถแลกเปลี่ยนความรู้ และแบ่งปันบทความวิจัยให้กับกลุ่มนักวิจัยที่ทำงานวิจัยในเรื่องเดียวกัน ปัจจุบันมีเว็บไซต์เครือข่ายสังคมและโซเชียลบุ๊กมาร์ก ที่ให้บริการข้อมูลด้านงานวิจัยและบทความวิชาการ เช่น Connotea , BibSonomy และ CiteULike ทั้งนี้ CiteULike ถือเป็นเว็บไซต์ที่ได้รับความนิยมจากผู้ใช้งานมากที่สุด อย่างไรก็ตามกระบวนการในการการค้นหางานวิจัยของเว็บไซต์ดังกล่าวยังคงมีการพัฒนาอย่างต่อเนื่อง เพื่อเพิ่มประสิทธิภาพในการค้นหาให้ผู้ใช้เกิดความพึงพอใจมากที่สุด

ผู้ใช้งานในระบบโซเชียลบุ๊กมาร์กส่วนใหญ่จะมีการสร้างกลุ่มผู้ใช้ (Group) เพื่อทำการแลกเปลี่ยนข้อมูลข่าวสร้างให้กับสมาชิกภายในกลุ่ม โดยผู้ใช้แต่ละคนสามารถเข้าร่วมเป็นสมาชิกในกลุ่มที่ตนเองต้องการ ทั้งนี้จะต้องได้รับการอนุญาตให้เป็นสมาชิกในกลุ่มจากเจ้าของกลุ่ม จึงจะสามารถแลกเปลี่ยนข้อมูลได้

งานวิจัยนี้จึงได้นำความสำคัญของการเข้าร่วมเป็นสมาชิกในกลุ่มของผู้ใช้ดังกล่าวมาสร้างกฎความสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูล เพื่อทำการการเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยสำหรับระบบเครือข่ายสังคมต่อไป

## 1.2 วัตถุประสงค์การวิจัย

1.2.1 เพื่อวิเคราะห์ความสัมพันธ์ของกลุ่มผู้ใช้บนระบบเครือข่ายสังคมโดยใช้เทคนิคเหมืองข้อมูล สำหรับผลการค้นหางานวิจัยบนระบบเครือข่ายสังคม

1.2.2 เพื่อเพิ่มประสิทธิภาพผลการค้นหาด้านงานวิจัยบนระบบเครือข่ายสังคม

## 1.3 ขอบเขตการวิจัย

1.3.1. งานวิจัยนี้มุ่งเน้นในการเพิ่มประสิทธิภาพผลการค้นหาบนระบบเครือข่ายสังคมทางด้านงานวิจัยเท่านั้น

1.3.2. งานวิจัยนี้ใช้กลุ่มตัวอย่างในการทดลองจากผู้ใช้งานในระบบระบบเครือข่ายสังคมด้านงานวิจัยเท่านั้น

## 1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 มีอัลกอริทึมที่มีเหมาะสมในการจัดอันดับผลการค้นหาสำหรับระบบเครือข่ายสังคมทางด้านงานวิจัย

1.4.2 เพิ่มประสิทธิภาพการค้นหางานวิจัย ให้กับระบบเครือข่ายสังคมที่ให้บริการด้านงานวิจัย

## 1.5 คำจำกัดความที่ใช้

1.5.1 เครือข่ายสังคม (Social network) มีผู้อธิบายไว้หลายท่านดังนี้

- ประดนเดช นีละคุปต์ (2551) อธิบายว่า เครือข่ายสังคม คือ การเชื่อมโยงผู้คนเข้าด้วยกันโดยทางใดทางหนึ่ง โดยอาศัยเทคโนโลยีเว็บ
- อนงค์นาฎ ศรีวิหค (2551) อธิบายว่า เครือข่ายสังคม คือ การเชื่อมโยงประชากรเข้าด้วยกัน

1.5.2 เอกสารงานวิจัย (Research paper) หมายถึงรายงานวิจัย และวิทยานิพนธ์ และเอกสารสนับสนุนงานวิจัยที่เผยแพร่ภายในห้องสมุดงานวิจัย (ระเบียบสำนักงานคณะกรรมการวิจัยแห่งชาติ, 2547)

1.5.3 การค้นคืนสารสนเทศ (Information retrieval) หมายถึง การสืบค้นสารสนเทศ กระบวนการค้นหาสารสนเทศที่ต้องการ โดยใช้เครื่องมือสืบค้นสารสนเทศที่สถาบันบริการสารสนเทศจัดเตรียมไว้ให้

1.5.4 เหมืองข้อมูล (Data mining)

คือ กระบวนการที่กระทำกับข้อมูล(โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบแนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่องและหลักคณิตศาสตร์

**1.6 ระยะเวลาทำการวิจัย และแผนการดำเนินงานตลอดโครงการวิจัย**

ตารางที่ 1.1 ระยะเวลาดำเนินการโครงการ 12 เดือน

| กิจกรรม | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.ศึกษาหาข้อมูลด้านวิชาการและงานวิจัยที่เกี่ยวข้อง | ← | — | — | — | — | — | — | — | — | — | → | |
| 2. จัดเก็บข้อมูล | ← | — | — | → | | | | | | | | |
| 3.เตรียมข้อมูล | | ← | — | — | → | | | | | | | |
| 4. ออกแบบและทดสอบ Data mining Model | | | | ← | — | — | — | → | | | | |
| 5.วิเคราะห์ผลการทดลองและสรุปผลการทดลอง | | | | | | | ← | — | — | — | → | |
| 6.จัดทำรายงานฉบับสมบูรณ์ | | | | | | | | | | | ← | → |

**1.7 อุปกรณ์การวิจัย**

1.7.1.ซอฟต์แวร์ (Software)

- Mysql server
- Tomcat
- Apache lucene
- Java Servlet
- Netbean
- SAS

1.7.2. ฮาร์ดแวร์ (Hardware)

- เครื่องคอมพิวเตอร์ที่ทำหน้าที่เป็น Server จำนวน 1 เครื่อง
- เครื่องคอมพิวเตอร์ที่ทำหน้าที่เป็น Client สำหรับผู้ทดสอบระบบจำนวน 5 เครื่อง

# บทที่ 2
# ทฤษฎีที่เกี่ยวข้อง

การทบทวนวรรณกรรมที่เกี่ยวข้องกับการพัฒนาอัลกอริทึมสำหรับการจัดอันดับผลการค้นหางานวิจัยบนระบบเครือข่ายสังคม แบ่งออกเป็น 2 หัวข้อ คือ ทฤษฎีที่เกี่ยวข้อง และ การทบทวนวรรณกรรมหรือสารสนเทศที่เกี่ยวข้อง

## 2.1 ทฤษฎีที่เกี่ยวข้อง

### 2.1.1 ขั้นตอนการประมวลผลของการค้นคืนสารสนเทศ

ขั้นตอนการประมวลผลของการค้นคืนสารสนเทศประกอบด้วยขั้นตอนจำนวน 5 ขั้นตอนหลักเป็นดังนี้(ดร.ศุภชัย ตั้งวงศ์ศานต์ ,2551)

2.1.1.1 การทำดรรชนี (Indexing) เป็นการสร้างตัวแทนเอกสาร

วัตถุประสงค์ของการทำดรรชนีก็เพื่อทำเป็นตัวแทนเอกสาร โดยจัดเป็นหมวดหมู่อย่างเป็นระบบ อันเป็นการเตรียมการในขั้นต้นของ การค้นคืนสารสนเทศ เพื่อการสืบค้นต่อไปให้ดำเนินการได้อย่างมีประสิทธิภาพและประสิทธิผล ขั้นตอนการทำดรรชนีสามารถสร้างด้วยแรงหรือสร้างอย่างอัตโนมัติได้ รูปแสดงขั้นตอนการทำดรรชนีจากต้นแหล่งที่เป็นเอกสาร

ขั้นตอกแรกเป็น Lexical Analysisโดยทำการแปลงสายตัวอักษรในเอกสารหรือสิ่งเป้าหมายกลายเป็นสายของคำศัพท์ เมื่อได้คำแต่ละคำในข้อความแล้ว ให้จัดการคำที่เป็นตัวเลข คำที่เชื่อมด้วยอักษรตัวใหญ่เล็ก หรือคำที่เชื่อมด้วยเครื่องหมาย Punctuation ตามข้อกำหนด และส่งต่อในขั้นต่อไป ลำดับต่อไปเป็นการจำกัดคำโหล (Stop-words) อันหมายถึง คำที่ปรากฏบ่อยครั้งมากจนไม่มีผลในการสืบค้นทางปฏิบัติ จึงไม่จำเป็นต้องนำมาสร้างเป็นดรรชนี ขั้นต่อมาคือ Stemming เป็นการหารากศัพท์ของคำในข้อความ เพื่อใช้คำ Stem นั้นเป็นตัวแทนในการสร้างดรรชนีต่อไป ติดตามด้วยการเลือกคำศัพท์ เช่น กลุ่มของนามเพื่อสร้างดรรชนี จากนั้นเป็นการสร้างคำศัพท์สัมพันธ์ (Thesaurus Construction) เพื่อจัดกลุ่มคำที่มีความหมายเดียวกันให้เชื่อมต่อกันอย่างเป็นระบบและกำหนดคำศัพท์ควบคุม (Controlled Vocabulary) เมื่อได้คำศัพท์ตามกระบวนการข้างต้น ผลลัพธ์ที่ได้นำมาจัดทำดรรชนีและจัดเก็บตามการออกแบบของโครงสร้าง ที่สำคัญก็จะมีรูปแบบของแฟ้มข้อมูลผกผัน(Inverted File) เพื่อการสืบค้นต่อไป

2.1.1.2 การจัดรูปแบบคำสอบถาม (Query Formulation) เป็นการสร้างตัวแทนคำสอบถาม

การเขียนคำสอบถามเพื่อการค้นหาสารสนเทศที่ต้องการทำได้ในหลากหลายรูปแบบที่ง่ายและสะดวกคือการใส่คำสำคัญ (Keywords) เป็นหนึ่งคำศัพท์หรือหลายๆคำเรียงต่อกันได้ตามที่ต้องการ เช่น ต้องการค้นหาเรื่อง Quantum Computing เขียนโดย Joseph Stelmach ก็เพียงเขียนคำสอบถามว่า Quantum Computing Joseph Stelmach ป้อนถาม

Search Engine ก็จะได้ผลลัพธ์ตามที่เทียบเคียงได้ การเขียนรูปแบบข้างต้นสามารถขยายผลโดยเชื่อมต่อด้วย Boolean Operators อันได้แก่ AND,OR และ NOT เช่น(Quantum OR Computing) AND (Not Stelmach) การเขียนคำสอบถามด้วยตัวเชื่อมเหล่านี้ จะทำให้การเขียนมีความซับซ้อน แต่ในขณะเดียวกันก็มีความหมายที่เจาะจงมากขึ้น อันทำให้ผลลัพธ์สอดคล้องตามความต้องการมากขึ้น อีกวิธีหนึ่งเป็นการสืบค้นแบบ pattern โดยการอนุญาตให้ใช้ตัว Wild Card Character แทนตัวอักษรที่เป็น Prefix หรือ Suffix ของคำเช่น ใช้สัญลักษณ์ * แทน Wild Card ตัวอย่างการเขียนที่เป็น Pattern คือ compu*,*ters, *กุมาร*เป็นต้น

2.1.1.3 การเทียบเคียงจับคู่ (Matching) ตัวแทนคำสอบถามกับตัวแทนเอกสาร

เป็นการจับคู่ระหว่างคำสอบถามกับเอกสารโดยใช้ตัวแทนจากที่ได้มาในขั้นตอนก่อนหน้านี้ วิธีการขึ้นอยู่กับโมเดลของ การค้นคืนสารสนเทศ หากเป็น Boolean Model หรือโมเดลที่เป็นลักษณะเดียวกันนี้ ก็จะให้ค่าการเทียบเคียงเป็นเพียง 2 ค่า คือ สอดคล้องหรือไม่ไม่สอดคล้อง จะไม่มีค่าระหว่างกลางทำให้ผลของการสืบค้นได้ค่า Recall ที่ต่ำ สำหรับ Vector Model หรือโมเดลในทำนองเดียวกัน ค่าการเทียบเคียงจะได้จาก Inner Product ของเวกเตอร์ของคำสอบถามและของเอกสาร ซึ่งเป็นการวัดความใกล้เคียงสอดคล้องของแต่ละคู่ หากได้ค่าสูง แสดงถึงความเกี่ยวข้องสอดคล้องกันมาก ในทางตรงกันข้าม หากได้ค่าต่ำ หมายถึง ความสอดคล้องน้อย จากค่าที่ได้ ยังสามารถนำไปจัดลำดับของเอกสารในผลลัพธ์ตามความสำคัญก่อนหลัง ค่าสูงอยู่ก่อน ค่าต่ำอยู่หลัง อนึ่ง ค่าการจับคู่ยังอาจปรับได้ในทางคำนวณโดยการให้น้ำหนักของเทอมในคำสอบถามรวมทั้งค่าน้ำหนักใน Term-Document Matrix ที่สอดคล้องกันอย่างเหมาะสม

การจัดลำดับของเอกสารที่นิยมในระบบ Search Engine ที่สำคัญ เช่น Google ยังได้จัดลำดับที่เรียกว่า PageRank โดยมีกระบวนการคำนวณจากความสำคัญของเว็บไซต์นั้นมีการถูกอ้างอิงมากน้อย โดยดูจากจำนวนสายเชื่อมโยง (Hyperlink)

2.1.1.4 การเลือก(Selection) รายการผลลัพธ์ที่ตรงประเด็น

เป็นการเลือกของผู้ใช้ในผลลัพธ์ที่ปรากฏของเอกสารที่สอดคล้องตรงประเด็น ซึ่งผลลัพธ์ของการสืบค้นอาจจะเรียงลำดับความสำคัญในกลุ่มหัวเรื่องๆต่างๆ หรือกลุ่มประเภทต่างๆอย่างอัตโนมัติ(Classification หรือ Clustering) นอกจากนี้ เพื่อช่วยผู้ใช้ในการเลือกที่ชัดเจน ผลลัพธ์อาจจะแสดงเป็นหัวเรื่อง (Title) ส่วนของบทคัดย่อ (Abstract) พร้อมทั้งเน้นเป็นแถบสว่าง (Highlight) ที่ชัดเจนในคำศัพท์หรือเทอมของคำสอบถามที่ต้องการค้นหา

2.1.1.5 การปรับเปลี่ยนคำสอบถามใหม่(Query Reformulation) ในรอบต่อไป

ในระบบการค้นคืนสารสนเทศ ผู้ใช้มักจะมีปัญหาการตั้งคำสอบถาม ปัญหาไม่ใช่ตั้งไม่ได้แต่การตั้งที่ให้ผลการสืบค้นที่มีประสิทธิภาพนั้นทำได้ไม่ง่ายนัก บ่อยครั้งที่ผู้ใช้ต้องเสียเวลาในการเลือกดูชุดเอกสารที่เป็นผลลัพธ์ที่ได้มามากมายแต่มีขยะปะปนมาก็มาก แนวทางหนึ่ง

ในการแก้ปัญหานี้เพื่อสร้างความพึงพอใจแก่ผู้ใช้คือ การปรับเปลี่ยนคำสอบถามใหม่ (Query Reformulation) ด้วยเทคนิคของ Query Expansion หรือของ Relevance Feedback หรือรวมกันทั้งสองเทคนิค

Query Expansion หมายถึงการเพิ่มเทอมในคำสอบถามของระบบการค้นคืนสารสนเทศ เพื่อให้การสืบค้นในรอบต่อไปมีผลลัพธ์ที่ดีขึ้น ซึ่งเทอมที่เพิ่มขึ้นนั้นอาจจะมาจากการ Feedback หรือไม่ก็แล้วแต่กรณี รวมทั้งอาจเปลี่ยนค่าน้ำหนักของเทอมเพื่อความเหมาะสม ส่วน Relevance Feedback หมายถึง การป้อนความเกี่ยวพันธ์ย้อนกลับ อันเป็นการใช้ประโยชน์จากข้อมูลย้อนกลับนี้ไปปรับเปลี่ยนคำสอบถามเก่ามาเป็นคำสอบถามใหม่ของการสืบค้นในแต่ละรอบ วิธีการเป็นไปได้ทั้งอัตโนมัติและกึ่งอัตโนมัติโดยใช้แรงคนร่วมปฏิบัติการด้วย

สำหรับแหล่งข้อมูลเพื่อการปรับเปลี่ยนคำสอบถามใหม่เป็นไปได้ใน 3 แหล่งคือ จาก Local Analysis, Global Analysisและ/หรือ Query Reuse ตามที่แสดงในรูป ใน Local Analysis หมายถึงแหล่งข้อมูลที่ทำการวิเคราะห์มาจากชุดเอกสารรวมทั้งหมดและดึงเทอมที่เป็นคำสัมพันธ์ (Thesaurus) มาพิจารณา ส่วน Query Reuse หมายถึงการใช้ซ้ำของคำสอบถามจากแหล่งฐานข้อมูล Query Base

### 2.1.2 การสร้างโมเดล

โมเดล หมายถึง รูปแบบที่แสดงในเชิงตรรกะ(Logical View) เพื่อจำลององค์ประกอบในระบบ รวมทั้งจำลองการปฏิบัติการ โดยทั่วไปรูปแบบอาจจะเป็นรูปภาพ หรือเป็นสัญลักษณ์และมีลายเส้นโยงต่อกันไปมา รวมทั้งอาจเขียนเป็นสัญลักษณ์ทางคณิตศาสตร์และสมการคณิตศาสตร์เพื่อแสดงขบวนการปฏิบัติการที่ปรากฏทางกายภาพ

สำหรับโมเดลของระบบการสืบค้น มีรูปแบบเป็นการเฉพาะที่แสดงเชิงตรรกะเพื่อจำลององค์ประกอบต่างๆในระบบ เช่นตัวเอกสาร คลังเอกสาร ข้อสนเทศที่ผู้ใช้ต้องการหรือคำสอบถาม รวมทั้งปฏิบัติการเทียบเคียง(Matching) เพื่อหาผลลัพธ์ ปฏิบัติการการจัดกลุ่ม การจัดหมวดหมู่ของคลังเอกสาร และอื่นๆเป็นต้น

ในหัวข้อนี้ จะบรรยายโมเดลพื้นฐานหลักของระบบการค้นคืนข้อมูล จำนวน 2 รูปแบบคือ Classical Boolean Model และ Vector Space Model

#### 2.1.2.1 Classical Boolean Model (CBM)

Classical Boolean Model เป็นต้นแบบของระบบการสืบค้นในยุคแรก โดยเอกสารประกอบด้วยเทอมต่างๆ ที่อยู่ภายในและเพื่อความสะดวกในการสืบค้นจึงสร้างเป็นดรรชนีของเทอม เมื่อกำหนดคำสอบถามเป็นชุดของเทอมต่างๆ การสืบค้นจึงเป็นเรื่องการ Match ดังตัวอย่างต่อไปนี้ (ดร.ศุภชัย ตั้งวงศ์ศานต์, 2551)

ให้เอกสารชุดหนึ่งมีข้อความดังนี้

| Document | Text |
|---|---|
| 1 | Please porridge hot, please porridge cold. |
| 2 | Please porridge in the pot. |
| 3 | Nine days old. |
| 4 | Some like it hot, some like it cold. |
| 5 | Some like it in the pot. |
| 6 | Nine days old. |

สำหรับการสอบถามเพื่อการค้นหาเทอมใน CBM เป็นไปอย่างตรงไปตรงมา ตัวอย่างเช่นใน ต้องการสอบถามเทอมว่า hot ปรากฏที่เอกสารใด ก็เพียงแต่ไล่ตรวจสอบจากเอกสารที่ 1 จนถึง N (ในที่นี่ N = 6 สำหรับเอกสาร 6 ชุด และการไล่ทีละเอกสาร ไม่ได้เป็นวิธีที่ดี มีวิธีการค้นหาที่มีประสิทธิภาพสูง เช่น การค้นจากดรรชนีของเทอม ซึ่งจะได้นำเสนอต่อไป) เมื่อสิ้นสุดจะมีเพียงเอกสารชุดที่ 1 และ 4 ที่มีเทอม hot ปรากฏ คำตอบจึงเป็น Match และผลลัพธ์เป็น {1,4}

Classical Boolean Model เป็นรูปแบบพื้นฐานในระยะแรกของการพัฒนาระบบการค้นคืนสารสนเทศ และมีข้อจำกัดในตัวระบบหลายลักษณะ จึงได้มีการพัฒนารูปแบบอื่นที่มีประสิทธิภาพและประสิทธิผลมากกว่า ดังจะได้บรรยายในหัวข้อต่อไป

**2.1.2.1 Vector Space Model (VSM)**

Vector Space Model เป็นรูปแบบที่พัฒนาต่อจาก Classical Boolean Model ด้วยการสร้างโมเดลของคลังเอกสารเป็นแบบ Matrix ของ Term-Document โดยเขียนคำค้นหาหรือคำสอบถาม และตัวเอกสารอยู่ในรูปเวกเตอร์ การวัดความเหมือนจึงไม่ได้วัดเพียง Match และไม่ Match ดังเช่นใน CBM แต่จะวัดความใกล้ของคู่เวกเตอร์ของคำสอบถามกับเอกสารนั้นว่าใกล้กันมากน้อยเพียงใด หากใกล้มากก็ควรจะมีความสอดคล้องมาก หรือ Match มาก หากไกลมากก็ควรจะมีความสอดคล้องน้อยหรือ Match น้อย วิธีนี้จำทำให้ระบบสามารถจัดอันดับ Ranking ของผลลัพธ์โดยเรียงเอกสารที่ได้ตามลำดับความสอดคล้องกับคำสอบถาม สำหรับวิธีการวัดความสอดคล้องของเวกเตอร์ ก็จะใช้ Inner Product เป็นเกณฑ์ในการคำนวณโดยมีรายละเอียดดังนี้

กระบวนการของ VSM จะมีการคำนวณค่าความสอดคล้องกับความถี่ที่ปรากฏของเทอมในเอกสารค่าดังกล่าวจึงเป็นน้ำหนักของเทอมเขียนแทนด้วย $w_{i,j}$ กำหนดให้ $w_{i,j}$ มาจากค่า 2 ค่า ค่าหนึ่งเป็นความถี่ของเทอมที่ปรากฏ(Term Frequency) เขียนแทนด้วย tf เทอมใดที่มีปรากฏบ่อย ค่า tf ย่อมจะสูง ตรงกันข้าม ค่า tf ก็จะมีค่าต่ำ หรือหากไม่ปรากฏเลย ค่า tf จะมีค่าเป็น 0 ส่วนอีกค่าหนึ่งเป็นค่าผกผันของความถี่ของเอกสารที่มีเทอมนั้นปรากฏ(Inverse Document Frequency) เขียนเป็น idf ดังนั้นในเอกสาร$d_j$ จึงเขียนเป็นเวกเตอร์ที่มีค่าเป็นน้ำหนักของเทอมดังสมการที่ ( 2.1 )และ (2.2)

$$d_j = \left[ w_{1,j} \; w_{2,j} \; \ldots . w_{i,j} \; w_{M,j} \right] \qquad (2.1)$$

โดย

$$w_{i,J} = if_{i,J} \times idf_i \quad (2.2)$$

สำหรับ tf แม้จะเป็นค่าความถี่ของการปรากฏ แต่ด้วยจำนวนเทอมในแต่ละเอกสารมีมากน้อยต่างกัน ค่าความถี่ 10 ของเอกสารหนึ่ง ที่มีเทอมจำนวนเพียง 100 อาจจะดูว่ามีค่าสูง แต่ในขณะค่าความถี่ของอีกเอกสารหนึ่งที่มีเทอมจำนวนเป็น 10000 จะดูเป็นค่าต่ำ ดังนั้นเพื่อให้สามารถนำมาเปรียบเทียบกันได้ โดยไม่ขึ้นกับขนาดเอกสาร จึงได้กำหนด tf เป็น Normalized Form ดังสมการที่ (2.3)

$$tf_{i,J} = \frac{f_{i,J}}{\max_k f_{k,j}} \quad (2.3)$$

กำหนดให้ $f_{i,j}$ เป็นข้อมูลดิบของความถี่ของเทอม$t_i$ ในเอกสาร $j$

ขณะเดียวกัน การคำนวณ idf ขึ้นอยู่กับค่าความถี่ของเอกสารที่มีเทอมนั้นปรากฏอยู่ เนื่องจากเป็นค่าผกผันหากมีความถี่สูง ค่า idf จะต่ำ ในทางกลับกัน ค่า idf จะสูง เช่นมี เทอม “retrieval” ปรากฏในเอกสารจำนวน 1,000 เอกสารจากจำนวนเอกสารทั้งหมด 10,000,000 เอกสาร ดังนั้นค่าความถี่ของเอกสารเป็น 0.0001 (1,000/10,000,000) ค่าผกผันจะกลายเป็น 10,000 อย่างไรก็ตาม เพื่อไม่ให้ค่าที่ได้สูงโด่งจนเกินไป ค่า idf จะใช้เป็นค่า log แทนค่าโดดๆดังดังสมการที่ (2.4)

$$idf_{i,} = \log \frac{N}{n_i} \quad (2.4)$$

กำหนดให้ $N$ เป็นจำนวนเอกสารทั้งหมดในระบบ

$n_1$ เป็นจำนวนเอกสารทั้งหมดที่มีเทอม I ปรากฏ

log จะใช้ log ฐาน 2 หรือฐาน e ที่เขียนว่า ln ก็เลือกได้

สำหรับเอกสารในองค์รวมทั้งระบบให้ D เป็น Matrix ของเอกสารโดยรวบรวมจากเอกสารที่ 1 จนถึง N เขียนได้ว่า

$$D = \begin{bmatrix} d_1 & d_2 & d_3 & \cdots & d_j \cdots d_N \end{bmatrix} \quad (2.5)$$

$$= \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,N} \\ w_{2,1} & w_{2,2} & & w_{2,N} \\ .. & & w_{i,j} & \cdots \\ w_{M1} & w_{M,2} & & w_{M,N} \end{bmatrix} \quad (2.6)$$

D จะเป็น Matrix ขนาด M * N ส่วนค่า $w_{i,j}$ จะแสดงถึงความสำคัญของเทอม t นั้น ในเอกสาร j ในภาพรวมทั้งระบบ ต่างจาก D ใน Boolean Model ที่มีค่า 1 หรือ 0 ซึ่งค่อนข้างจะไม่ยืดหยุ่น

อย่างไรก็ตาม รูปแบบ TF-IDF ได้มีการพัฒนาอย่างต่อเนื่อง โดยเฉพาะการจัดอันดับ ranking ของเอกสารได้มีการพัฒนาไปอย่างมีนัยสำคัญ เมื่อนำระบบการสืบค้นมาใช้บนอินเตอร์เน็ต ทำให้มีความแปลกใหม่ในการให้น้ำหนักของเทอมในเอกสาร ตัวอย่างกลไกการสืบค้น (Search Engine) ที่สำคัญ ได้ลำดับผลลัพธ์ด้วย PageRank ใน Google ทำให้ผลงานมีประสิทธิภาพเหนือคู่แข่ง กลายเป็นผู้นำการให้บริการ Search Engine

2.1.2.3 ตัวพิมพ์ (Font) และขนาดของตัวพิมพ์

รายงานการวิจัยภาษาไทย ให้ใช้ตัวพิมพ์ Crodia โดยใช้ตัวอักษรธรรมดา ขนาด 16 เท่านั้น

2.1.2.4 ให้ใช้เครื่องพิมพ์ (Printer) แบบ Letter Quality หรือเครื่องพิมพ์เลเซอร์ (Laser Printer) หรือ Inkjet และไม่ให้ใช้วิธีพิมพ์แบบ Draft จากเครื่องพิมพ์ Dot Matrix

**2.1.3 การประเมินประสิทธิภาพ**

การประเมินผลการสืบค้นเป็นกระบวนการหนึ่งที่สำคัญเพื่อวัดสมรรถนะของระบบคุณภาพของการบริการ กระบวนการทำงาน รวมทั้งอาจใช้เพื่อการเปรียบเทียบระบบต่อระบบ โดยหลักการแล้ว ระบบการสืบค้นที่ดีจะต้องเป็นระบบที่สามารถดึงเอาสารสนเทศที่เกี่ยวพันธ์อันเป็นความต้องการและเป็นประโยชน์ได้อย่างรวดเร็ว ถูกต้องแม่นยำ ครบถ้วนสมบูรณ์ในทุกแห่งที่ปรากฏ ขณะเดียวกัน ก็จะไม่ดึงเอาสารสนเทศที่ไม่เกี่ยวพันธ์ หรือที่เป็นขยะให้เห็นนอกจากนี้สารสนเทศที่ปรากฏหากมีเป็นจำนวนมากก็ควรจัดเรียงลำดับตามลำดับตามความสำคัญก่อนหลัง หรือจัดกลุ่มอย่างเป็นสาระ เพื่อสร้างความพึงพอใจต่อผู้ใช้งาน ซึ่งมีวิธีประเมินประสิทธิภาพที่สำคัญจำ 2 วิธีคือ Precision/Recall และ Normalized discounted cumulative gain (NDCG)

2.1.3.1 Precision-Recall

ดร.ศุภชัย ตั้งวงศ์ศานต์ (2551) กล่าวว่า ค่า Recall เป็นการวัดความสามารถของระบบในการค้นหาเอกสารที่เกี่ยวพันธ์ ในขณะที่ค่า Precision เป็นการวัดความแม่นยำของระบบในการค้นหาเอกสารที่เกี่ยวพันธ์ได้ถูกต้อง หากการสืบค้นจากการสอบถามได้ผลตามตารางที่2.1

ตารางที่ 2.1แสดงผลการสืบค้น

| | Retrieved | Not Retrieved |
|---|---|---|
| Relevant | tp | fn |
| Not Relevant | fp | tn |

ค่า Recall = tp/(tp+fn)

และ ค่า Precision = tp/(tp+fp)

ระบบการสืบค้นข้อมูลในอุดมคติ ควรจะได้ทั้งค่า Recall และค่า Precision ที่สูง แต่ในความเป็นจริงเป็นเช่นนั้นไม่ และมักจะได้ผลในลักษณะที่ค่าหนึ่งสูง อีกค่าหนึ่งก็จะต่ำ เมื่อนำค่าทั้งสองมาจับคู่กันในการสืบค้นครั้งต่างๆและแสดงเป็นกราฟของ Recall-Precision ก็จะเห็นถึงความสัมพันธ์ของเส้นกราฟที่ค่าทั้งสองเป็นผกผันกัน ซึ่งทั้งนี้การประเมินด้วย Precision-recall ไม่ได้นำลำดับ (Ranking) ของผลการค้นหาเข้ามาพิจารณาในการประเมินผล

2.1.3.2 Normalized discounted cumulative gain (NDCG)

Normalized discounted cumulative gain เป็นทฤษฎีที่ใช้ในการประเมินผล search engine มีการให้ระดับคะแนนความเกี่ยวข้องของเอกสาร รวมทั้งพิจารณาลำดับของผลการค้นหา (Ranking) และ DCG เป็นการวัดความเหมาะสมของเอกสารโดยสนใจตำแหน่งหรือลำดับของเอกสาร ค่าระดับคะแนนที่ได้รับเป็นสะสมจากลำดับบนของรายการผลลัพธ์การค้นหาไปยังลำดับล่างของผลลัพธ์การค้นหา โดยค่าคะแนนจะลดลงเมื่อผลความพึงพอใจอยู่ในลำดับที่ต่ำ (Jarvelin, K., and Kekalainen, J. 2006) โดยสมการ NDCG เป็นดังนี้

$$NDCG_q = M_q \sum_{j=1}^{k} \frac{\left(2^{r(j)} - 1\right)}{\log\left(1 + j\right)} \qquad (2.7)$$

เมื่อ *k* คือ ระดับหรือเกณฑ์ที่ใช้ , *r(j)* คือค่าลำดับความเกี่ยวข้องของเอกสารที่ได้จากกระประเมินโดยผู้ใช้ , $M_q$ คือ ค่าคงที่ที่เกิดจากความสมบูรณ์ในการจัดลำดับโดยมีค่ามากที่สุดคือ 1 ทั้งนี้ NDCG จะให้รางวัลกับเอกสารที่เกี่ยวข้องที่ปรากฏในลำดับของการจัดอันดับผลการค้นหาและลงโทษเอกสารที่ไม่เกี่ยวข้องโดยการลดคะแนน NDCG

**2.1.4 เทคนิคเหมืองข้อมูล**

**2.1.4.1 กฎการวิเคราะห์ความสัมพันธ์ (Association rule)**

กฎการวิเคราะห์ความสัมพันธ์เป็นการหาความสัมพันธ์ระหว่างข้อมูลด้วยกันเองซึ่งมีพื้นฐานมาจากการเกิดขึ้น ร่วมกันหรือพร้อมกันในฐานข้อมูล

<u>รูปแบบของการค้นหากฎความสัมพันธ์</u>

รูปแบบทั่วไปของการค้นหากฎความสัมพันธ์ คือ A→ B

โดยที่ A : เป็นเงื่อนไข หรือ LHS (Left - Hand Side)

และ B : เป็นผลลัพธ์ที่เกิดขึ้น หรือ RHS (Right – Hand Side)

หรืออยู่ในรูปของ “ถ้า.....แล้ว” (If.....Then....) เช่น

A→B ; if A Then B เป็นกฎที่ 1

B→A ; if B Then A เป็นกฎที่ 2

การประเมินค่าของกฎจะใช้ค่าสนับสนุน(Support) และค่าความเชื่อมั่น (Confidence)

โดยที่ค่าสนับสนุน คือ เปอร์เซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

$$\text{ค่าสนับสนุน}(A,B) = \frac{\text{จำนวนของ Transaction } (A,B)}{\text{จำนวน Transaction ทั้งหมด}}$$

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการหาผลลัพธ์

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์

Transaction(A,B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

ค่าความเชื่อมั่น คือเปอร์เซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

$$\text{ค่าความเชื่อมั่น } (A,B) = \frac{\text{จำนวนของ Transaction } (A,B)}{\text{จำนวน Transaction } (A)}$$

โดยที่ Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว

ในการเลือกว่าจะกฎใดนั้นจะต้องพิจารณาค่าสนับสนุนและค่าความเชื่อมั่นที่มีค่าสูงกว่าค่า Threshold ที่ตั้งไว้ นอกจากนี้จะต้องกำหนดค่าสนับสนุนต่ำสุด (Minimum Support) และค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) โดยทั่วไปจะกำหนดค่าสนับสนุนต่ำสุดเป็น 5-10% และค่าความเชื่อมั่นต่ำสุดเป็น 50-100%

<u>อัลกอริทึมกฎการวิเคราะห์ความสัมพันธ์</u>

มีหลายอัลกอริทึมที่ใช้ในการค้นหาความสัมพันธ์เช่น

- อัลกอริทึม Apriori
- อัลกอริทึม AprioriTid
- อัลกอริทึม AIS
- อัลกอริทึม SETM

อัลกอริทึม Apriori ( agrawal and Srikant 1994) เป็นประเภทของการค้นหาความสัมพันธ์แบบ Boolean Association Rule ซึ่งเป็นอัลกอริทึมที่ใช้กันอย่างแพร่หลาย คือเทคนิควิธีที่ใช้สำหรับค้นหาสิ่งที่ปรากฏเด่นชัด (Frequent Itemsets)จากฐานข้อมูลที่ กำหนดโดยมีหลักการทำงานคืออัลกอริทึม Apriori ทำหน้าที่สร้างเซ็ตไอเท็มหรือกลุ่มข้อมูลที่ต้องการวิเคราะห์ที่เป็นไปได้ทั้งหมดที่มีค่าสนับสนุน มากกว่าค่าสนับสนุนขั้นต่ำโดยอัลกอริทึม Apriori เป็นการทำงานในแบบล่างขึ้นบน (bottom up) โดยมีขั้นตอนดังนี้

ขั้นตอนที่1 อัลกอริทึม Apriori อ่านฐานข้อมูลทั้งหมดและสร้างไอเท็มเซ็ตที่ผ่านค่าสนับสนุนขั้นต่ำความยาว 1 ไอเท็ม (Frequent 1-itemset)

ขั้นตอนที่2 อัลกอริทึม Apriori สร้างเซ็ตไอเท็มทดสอบ (Candidate Itemset) ที่มีความยาว 2 ไอเท็มจากเซ็ตไอเท็มที่ปรากฏเด่นชัดความยาว 1 ไอเท็มในขั้นตอนแรกและนำไปหาค่าสนับสนุนเพื่อ

ค้นหาไอเท็มเซ็ตที่ ปรากฏเด่นชัดความยาว 2 ไอเท็ม กรรมวิธีอพิออรีจะวนรอบทำงานจนกระทั่งไม่พบไอเท็มเซ็ตที่ผ่านค่าสนับสนุนขั้นต่ำจึงจบการทำงานไอเท็มเซ็ตที่ผ่านค่าสนับสนุน ขั้นต่ำในแต่ละรอบคือสิ่งที่ปรากฏเด่นชัดจากฐานข้อมูล

**2.1.4.2 การแบ่งกลุ่มข้อมูล (Cluster Analysis)**

การแบ่งกลุ่มเป็นเทคนิคที่ใช้จำแนกหรือแบ่งเป็นกรณี (คน สัตว์ สิ่งของ หรือองค์กร ฯลฯ) หรือแบ่งตัวแปรออกเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไป โดยกรณีที่อยู่กลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วนกรณีที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกัน ดังนั้นการพิจารณาเลือกลักษณะตัวแปรที่จะนำมาใช้แบ่งกลุ่ม Caseจึงมีความสำคัญ นอกจากนั้น Case ใด Case หนึ่งจะอยู่ในกลุ่มหนึ่งเพียงกลุ่มเดียว (ดร. กัลยา วินิชย์บัญชา., 2001 : 125)

เทคนิคCluster Analysisแบ่งเป็นหลายประเภทโดยเทคนิคที่ใช้กันมากมี 2 เทคนิค คือ

1. Hierachical Cluster Analysis
2. K-Means Cluster Analysis

เทคนิค Hierachical Cluster Analysisเป็นเทคนิคที่นิยมใช้กันมากในการแบ่งกลุ่ม Case หรือแบ่งกลุ่มตัวแปรที่มีจำนวนไม่มาก สามารถใช้ได้กับตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ ซึ่งไม่จำเป็นต้องทราบว่าตัวแปรนั้นอยู่กลุ่มใด และมีจำนวนกลุ่มมากเท่าใด

เทคนิค K-Mean Clustering เป็นเทคนิคการจำแนก Case ออกเป็นกลุ่มย่อย โดยจะใช้กับ Case ที่มีจำนวนมากและต้องกำหนดจำนวนกลุ่ม ตัวแปรที่ใช้จะต้องเป็นตัวแปรเชิงปริมาณคือเป็นสเกลอันตรภาค (Interval Scale) หรือ สเกลอัตราส่วน (Ratio Scale) เท่านั้น ผู้วิเคราะห์จะต้องทำข้อมูลให้เป็นมาตรฐานก่อน

## 2.2 การทบทวนวรรณกรรม/สารสนเทศ (Information) ที่เกี่ยวข้อง

**การทบทวนวรรณกรรมด้านการเพิ่มประสิทธิภาพผลการค้นหา**

งานวิจัยจำนวนมากมุ่งพัฒนาประสิทธิภาพของผลการค้นหาโดยพยายามที่จะเปรียบเทียบและประเมินผลโดยทำการเทียบกับ Search engine ที่มีการให้บริการอยู่ในปัจจุบัน ดังดารวิจัยต่อไปนี้Gordon และ Pathank (1999) ได้ดำเนินการศึกษาและเปรียบเทียบประสิทธิผลของ search engines จำนวน 8 เว็บไซต์รวมทั้งศึกษาความสัมพันธ์ของแต่ละเว็บ, Thomas (2006) พยายามที่จะประเมินคุณภาพของ search engine ซึ่งผลการประเมินแสดงให้เห็นว่าการจัดอันดับตามคุณภาพจะนำไปสู่ผลลัพธ์ที่ดีขึ้น โดยคุณภาพของโมเดลจะเป็นประโยชน์ในการระบุคุณลักษณะที่สำคัญอาจเกิดขึ้นและลักษณะคุณภาพของเว็บ ,Long และเพื่อน ( 2007) ได้ทำการประเมินและเปรียบเทียบผลของ search engines จำนวน 3 เว็บไซต์ที่พัฒนาขึ้นด้วยภาษาจีนโดยทำการทดลองที่ใช้การประเมินจากผู้ใช้ระบบ

Sun และ Lee Giles (2007) ได้นำเสนออัลกอริทึม ในการจัดอันดับผลการค้นหาโดยการถ่วงน้ำหนัก (weight) สำหรับการค้นหาบทความวิชาการโดยใช้ข้อมูลจาก CiteSeer usage log ซึ่งใช้ปัจจัยด้านการตีพิมพ์เข้ามาเกี่ยวข้อง การทดลองได้ทำการเปรียบเทียบกับ PageRank ซึ่งเป็น

อัลกอริทึม ที่ใช้กับ google อย่างไรก็ตามผลการทดลองดังการเป็นการทดลองโดยวัดจากการพฤติกรรมการคลิกของผู้ใช้ (click through) ไม่ได้มาจากการประเมินของผู้ใช้

Richardson, Prakash, และ Brill (2006) ได้สร้าง RankNet ซึ่งเป็นอัลกอริทึมสำหรับการจัดอันดับโดยอาศัยกระบวนการการเรียนรู้ พร้อมทั้งยังทำการรวมกระบวนการ RankNet กับปัจจัยอื่นๆ โดยใช้ anchor text จากผลการทดลองพบว่าวิธีการนี้มีประสิทธิภาพดีกว่า PageRank อย่างมีนัยสำคัญ ในปี 2008 Hosein Keyhanipour, Piroozmand, และ Badie (2008) ได้พัฒนา GPRank ซึ่งมีประสิทธิภาพมากกว่าวิธีการจัดอันดับด้วยวิธีอื่น

Choochaiwattana และ Spring (2009) ทำการสำรวจการใช้งานระบบบันทึกย่อสำหรับเครือข่ายสังคม (Social annotation) เพื่อปรับปรุงคุณภาพของการค้นหา โดยพัฒนาวิธีในการจัดอันดับจำนวน 4 วิธี คือ1) Popularity Count (PC), 2) Query weighted Popularity Count (QWPC), 3) Matched Tag Count (MTC) และ 4) Normalized Matched Tag Count (NMTC) ซึ่งผลการประเมินจากผู้ใช้งานจริงพบว่าประสิทธิภาพในการจัดอันดับของ Normalized Matched Tag Count (NMTC) ดีกว่าการจัดอันดับของ google

งานวิจัยบางงานได้ประยุกต์ใช้ปัจจัยที่เกี่ยวข้องกับเวลาเข้ามาเพิ่มประสิทธิภาพของการจัดอันดับผลการค้นหาดังนี้ Berberichl, Vazirgiannis, และ Weikum (2004) ได้พัฒนา T-Rank ซึ่งเป็นการจัดอันดับโดยใช้วิธี link analysis เช่น timestamps ที่ใช้ และกิจกรรมที่ถูกใช้ เช่น การปรับปรุงหน้าเว็บและการลิงค์ ผลการวิเคราะห์พบว่า T-Rank เพิ่มประสิทธิภาพของกรจัดอันดับเว็บเพจได้

นอกจากนี้ยังมีงานวิจัยที่ศึกษาด้านการค้นหางานวิจัยหรือบทความวิชาการบนเว็บเครือข่ายสังคมและ โซเชียลบุ๊กมาร์ก ดังนี้ Capocci and Caldarelli(2007) วิเคราะห์คุณสมบัติของ folksonomy บน CiteUlike และ Santos-Neto, Ripeanu, และ Iamnitchi (2007) ได้ทำการสำรวจสามเส้นทางหลักสำหรับนำเสนอลักษณะเฉพาะของ CiteULike และ Bibsonomy ที่มีเป้าหมายเพื่อการจัดการทางด้านวรรณกรรมที่เกี่ยวข้องทางวิทยาศาสตร์ ทั้งนี้เทคนิคของ Citeulike ได้ถูกนำมาประยุกต์ใช้กับเว็บอื่น เช่น CiteSeer เพื่อใช้ในการค้นหาบทความวิชาการ (Farooq et al. 2007a, 2007b)

ในปี 2008 Toine Bogers และ Van den Bosch (2008) ได้ใช้ CiteULike ในการสร้างรายการที่อ่านบทความทางวิทยาศาสตร์สำหรับผู้ใช้งานออนไลน์ และใช้พัฒนา CF algorithms จำนวน 3 วิธี ที่แตกต่างกันคือ และพบว่าวิธีที่ดีที่สุดคือ User-model

ในปี 2009 นั้น พิจิตรา จอมศรี (2009a, 2009b) ได้ตัวสร้างดัชนีจำนวน 3 ดัชนี คือ 1) “tag”(T), “title, 2) abstract”(TA) , 3) “tag, title and abstract”(TTA) และทำการเปรียบเทียบกับ CiteULike ผลการทดลองพบว่า ดัชนีที่สร้างโดยใช้ “tag, title and abstract”(TTA) มีประสิทธิภาพดีที่สุด

**การทบทวนวรรณกรรมด้านการนำเทคนิคเหมืองข้อมูลมาประยุกต์ใช้กับข้อมูลบนเว็บ**

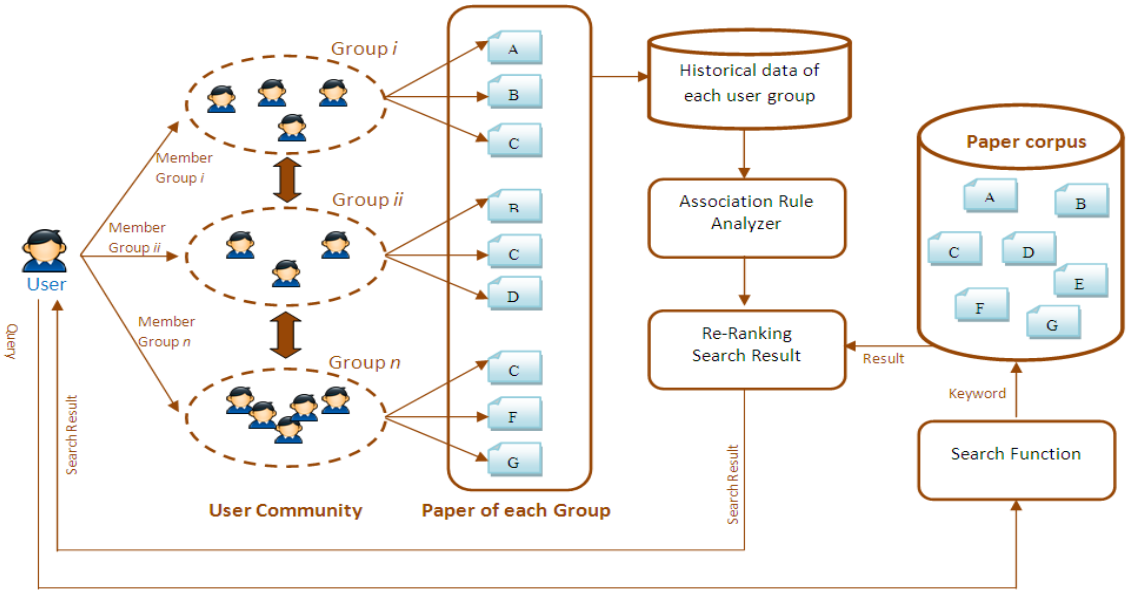
Wenge Rong, Kecheng Liu และ Lin Liang (2009) ได้นำเทคนิค Association rule ซึ่งเป็นเทคนิคเหมืองข้อมูลมาทำการวิเคราะห์ความสำพันธ์ของกลุ่มผู้ใช้เพื่อเพิ่มประสิทธิภาพของเว็บเซอร์วิสแบบปัจจเจกบุคคล ซึ่งพบว่าการทดลองดังกล่าวทำให้เว็บมีประสิทธิภาพเพิ่มขึ้น

R. Forsati, M. R. Meybodi และ A. Ghari Neiat (2009) ทำการเพิ่มประสิทธิภาพการนำเสนอเว็บเพจโดยศึกษาจากพฤติกรรมการค้นหาของผู้ใช้แต่ละคน โดยการทำการถ่วงน้ำหนักด้วยเวลาการใช้งานเว็บเพจ และใช้กฎAssociation rule มาทำการค้นหาความสัมพันธ์ของการใช้งานเว็บเพจของผู้ใช้แต่ละคน โดยผลการทดลองพบว่าการถ่วงน้ำหนักดังกล่าวทำให้เพิ่มประสิทธิภาพการนำเสนอเว็บเพจเมื่อทำการเปรียบเทียบกับการวิเคราะห์ด้วย Association rule เพียงอย่างเดียว

จากการทบทวนวรรณกรรมข้างต้น ทำให้พบว่างานวิจัยครั้งนี้มีความแต่งต่างจากงานวิจัยอื่นโดยการนำเทคนิคเหมืองข้อมูลมาประยุกต์ใช้กับการทำงานของเว็บเครือข่ายสังคม โดยการวิเคราะห์ความสัมพันธ์ของผู้ใช้ในแต่ละกลุ่ม เพื่อนำมาเพิ่มประสิทธิภาพการค้นหา

# บทที่ 3
# วิธีการดำเนินงาน

ในการจัดทำเอกสารงานวิจัยฉบับนี้ มีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหาสำหรับงานวิจัยโดยใช้เทคนิคเหมืองข้อมูล วิธีการดำเนินการวิจัย แบบออกเป็น 4 กระบวนการหลัก คือ 1) การจัดเก็บข้อมูล 2) ออกแบบ Data Mining Model 3) ทดสอบ Data Mining Model 4) กระบวนการประเมินผลตามทฤษฎี ดังนี้




**ภาพที่ 3.1** กระบวนการในการพัฒนาอัลกอริทึม ของการวิเคราะห์ความสัมพันธ์

## 3.1. การจัดเก็บข้อมูล

การจัดเก็บข้อมูลขั้นตอนนี้ได้ทำการพัฒนาโปรแกรมทาง Search Engine ที่มีหน้าที่ในการดึงข้อมูลจากเว็บเครือข่ายสังคม และ โซเชียลบุ๊กมาร์ก มาจัดเก็บในฐานข้อมูล โดยตัว Crawler จะเก็บข้อมูลโดยผ่านทาง Link ที่เชื่อมโยงกันไปมาของเว็บต่างๆที่ต้องการดึงข้อมูล ซึ่งโปรแกรมรวบรวมข้อมูลงานวิจัยมีหน้าที่ในการดึงข้อมูลบทความวิจัยจากเครื่องแม่ข่ายของระบบโซเชียลบุ๊กมาร์ก เช่น CiteULike โดยมีการดึงข้อมูลต่อไปนี้ ชื่อผู้แต่ง, tag, เวลาที่โพสต์ , ปีที่ตีพิมพ์ , ความสำคัญของบทความ และอื่นๆ ซึ่งข้อมูลเหล่านี้มีประโยชน์ต่อระบบในการตรวจสอบความสนใจของผู้ใช้และยังช่วยให้ระบบการดัชนีสำหรับแต่ละบทความวิจัย

*1 Crawler:* คือการพัฒนาโปรแกรมทาง Search Engine ที่มีหน้าที่ในการดึงข้อมูลจากเว็บเครือข่ายสังคม และ โซเชียลบุ๊กมาร์ก มาจัดเก็บในฐานข้อมูล โดยตัว Crawler จะเก็บข้อมูลโดยผ่านทาง Link ที่เชื่อมโยงกันไปมาของเว็บต่างๆที่ต้องการดึงข้อมูล ซึ่งโปรแกรมรวบรวมข้อมูลงานวิจัยมี

หน้าที่ในการดึงข้อมูลบทความวิจัยจากเครื่องแม่ข่ายของระบบโซเชียลบุ๊กมาร์ก เช่น CiteUlike โดยมีการดึงข้อมูลต่อไปนี้ ชื่อผู้แต่ง, tag, เวลาที่โพสต์ , ปีที่ตีพิมพ์ , ความสำคัญของบทความ และอื่นๆ ซึ่งข้อมูลเหล่านี้มีประโยชน์ต่อระบบในการตรวจสอบความสนใจของผู้ใช้และยังช่วยให้ระบบการดัชนีสำหรับแต่ละบทความวิจัย

ตารางที่ 3.1 รายการข้อมูลที่ใช้ในการทดลอง

| **รายการ** | **รายละเอียด** |
|---|---|
| แหล่งข้อมูล | CiteULike |
| จำนวนงานวิจัย | 102,242 เอกสารงานวิจัย |
| ข้อมูลกลุ่มผู้ใช้ | 200 กลุ่มผู้ใช้งาน |

*2 Historical data of each user group:* คือ การเก็บข้อมูลการใช้งานของผู้ใช้ และข้อมูลการเป็นสมาชิกในแต่ละกลุ่มของผู้ใช้แต่ละคน เพื่อเป็นการเตรียมข้อมูลสำหรับการสร้างโมเมลโดยใช้เทคนิคเหมืองข้อมูล

3. Association rule analyzer เป็นขั้นตอนที่นำเทคนิคเหมืองข้อมูลมาทำการวิเคราะห์หาความสัมพันธ์ของกลุ่มของกลุ่มผู้ใช้ ซึ่งถูกอธิบายในหัวข้อ 3.2.2 และ 3.2.3

*4.Ranking:* การจัดอันดับแบ่งออกเป็น 2 ประเภทคือ *Similarity ranking* และ *Re-ranking*:

- *Similarity ranking*: คือการเปรียบเทียบคำค้นหากับดัชนีที่สร้างขึ้น คะแนนที่ได้จากการเปรียบเทียบถูกคำนวณจากสมการที่ (3.1) ซึ่งเป็นฟังก์ชั่นของ Apache lucene พัฒนาโดย Hatcher และ Gospodnetic จากสมการ *q* หมายถึงคำค้นหา และ *d* หมายถึงเอกสารงานวิจัย

กำหนดให้

$$score(q,d) = \sum_{t \in q} \left( tf\left(t \in d\right) \times idf\left(t\right)^2 \times B_q \times B_d \times L \right) \times C \qquad (3.1)$$

*tf(t in d)* คือ Term frequency หรือ ความถี่ของ term (t) ที่ปรากฏใน document (d)
*idf(t)* คือ Inverse document frequency หรือ เป็นค่าที่ใช้บ่งบอกความสำคัญของ term ที่เทียบกับ ทุกๆ document ที่เก็บอยู่ในฐานข้อมูล
*boost(t.field in d)* คือ ค่า boost ที่ถูกสร้างขึ้นจากการสร้างดัชนี
*lengthNorm(t.field in d)* คือ กระบวนการ Normalization value of a field,
*coord(q, d)* คือ กระบวนการ Coordination factor

*queryNorm(q)* คือ กระบวนการ Normalization value for a query

● *Re-ranking:* คือผลของการวิเคราะห์ความสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูลมาทำการจัดอันดับใหม่ ทั้งนี้กฎความสัมพันธ์ที่ได้จะต้องมีค่าสนับสนุน และค่าความเชื่อมั่นตามเกณฑ์ที่กำหนด

*5. Search Result* : คือ ขั้นตอนการแสดงผลลัพธ์การค้นหาได้มีการสร้างอินเตอร์เฟสใหม่เพื่อป้องกันความอคติของผู้ใช้หรือผู้ประเมิน โดยผู้ประเมินระบบสามารถค้นหาผลงานวิจัยผ่านอินเตอร์เฟสที่สร้างขึ้นจากอัลกอริทึมในการจัดอันดับแต่ละวิธี ผู้ประเมินสามารถดูผลลัพธ์การค้นหาได้ ซึ่งผลลัพธ์แต่ละลำดับประกอบด้วย ชื่อเรื่อง, บทคัดย่อ และ การเชื่อมโยงไปยังเอกสารฉบับเต็ม

**3. 2 การออกแบบ Data Mining Model**

งานวิจัยนี้เลือกใช้เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule) เพื่อใช้ในการหาความสัมพันธ์ของการเป็นสมาชิกกลุ่มผู้ใช้แต่ละคน เพื่อใช้เป็นแนวทางในการแนะนำเอกสารงานวิจัยที่เกี่ยวข้องกับสิ่งที่ผู้ใช้สนใจมากที่สุด โดยพิจารณาจากร้อยละของค่าความเชื่อมั่น และค่าสนับสนุนที่เกิดขึ้น สามารถหาค่าความเชื่อมั่น และค่าสนับสนุนจากสมการดังนี้

$$\text{ค่าความเชื่อมั่น (A,B)} = \frac{\text{จำนวนของ Transaction (A,B)}}{\text{จำนวน Transaction (A)}} \quad (3.2)$$

$$\text{ค่าสนับสนุน (A,B)} = \frac{\text{จำนวนของ Transaction (A,B)}}{\text{จำนวน Transaction ทั้งหมด}} \quad (3.3)$$

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการทำนาย

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์ที่ได้จากการทำนาย

Transaction (A, B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว

การหากฎความสัมพันธ์ทั้งหมดจะต้องมีค่าสนับสนุนมากกว่าค่าสนับสนุนต่ำสุดที่กำหนดไว้ และมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นต่ำสุดที่กำหนดไว้

**3. 3 ทดสอบ** Data Mining Model

การทดสอบและตรวจสอบความถูกต้องของโมเดลที่สร้างขึ้น จะเปรียบเทียบกับผลลัพธ์ที่ได้จากการใช้เครื่องมือของ SAS

**1)** การตรวจสอบผลการทำนาย

นำผลที่ได้จากการวิเคราะห์ความสัมพันธ์ ไปเปรียบกับพฤติกรรมการใช้งานจริงของผู้ใช้

2) ประเมินผลจากการตรวจสอบการทำนาย

นำผลที่ได้จากการตรวจสอบมาประเมินผล โดยวัดจากประสิทธิภาพและความน่าเชื่อถือที่ตั้งไป

**3. 4 ประเมินผลการจัดอันดับผลการค้นหา**

งานวิจัยนี้ได้นำเทคนิค Precision/Recall มาทำการประเมินผล โดย Recall นั้นถูกกำหนดให้เป็นอัตราส่วนของเอกสารที่เกี่ยวข้องที่ถูกดึงออกมาจาก จำนวนเอกสารที่เกี่ยวข้องทั้งหมด ในขณะที่ Precision เป็นอัตราส่วนของเอกสารที่ถูกดึงออก มาแล้วข้อความ จากจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด ในทางปฏิบัติ, สารสนเทศที่ ต้องการของผู้ใช้แต่ละคนนั้นย่อมแตกต่างกันไป ซึ่งผู้ใช้บางคนอาจจะต้องการ Recall ที่สูง, กล่าวคือ, ทุกสิ่งที่ถูกดึงออกมาเป็นเรื่องที่น่าสนใจ ในขณะที่อีกคนหนึ่งอาจจะต้องการ ถ้าหากมีเส้นตัดผ่าน Document Collection เพื่อที่จะแยกแยะรายการเอกสารที่ถูกดึง ออกมา ให้ออกจากรายการเอกสารที่ไม่ได้ถูกดึงออกมา ดังแสดงในรูป 6.3 และถ้าหากเรา สามารถมีวิธีการที่จะแยกแยะเอกสารที่เกี่ยวข้องออกจากเอกสารที่ไม่เกี่ยวข้อง Recall มาตรฐาน R และ Precision มาตรฐาน P อาจจะกำหนดได้โดย

$$recall = \frac{\textit{จำนวนเอกสารที่เกี่ยวข้องทั้งหมดและถูกดึงออกมา}}{\textit{จำนวนเอกสารที่ถูกต้องทั้งหมด}} \quad (3.4)$$

$$\text{Pr}\,ecision = \frac{\textit{จำนวนเอกสารที่ถเกี่ยวข้องและถูกดึงออกมา}}{\textit{จำนวนเอกสารที่ถูกดึงออกมาทั้งหมด}} \quad (3.5)$$

**3.5 สถานที่ทำการทดลอง/เก็บข้อมูล**

ห้องปฏิบัติการ สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา

# บทที่ 4
# ผลการดำเนินงานและการวิเคราะห์

ในบทนี้อธิบายผลการดำเนินงานและการวิเคราะห์ ซึ่งประกอบด้วย1) การจัดเก็บและเตรียมข้อมูล 2) ศึกษาตัวแบบเพื่อใช้ในการวิเคราะห์ความสัมพันธ์ 3) ทดสอบตัวแบบเพื่อใช้ในการวิเคราะห์ความสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูล 4) ทดสอบความถูกต้องของตัวแบบ 5) ผลการทดลองและการประเมินผล โดยมีรายละเอียดดังนี้

## 4.1 การจัดเก็บข้อมูลจาก CiteULike

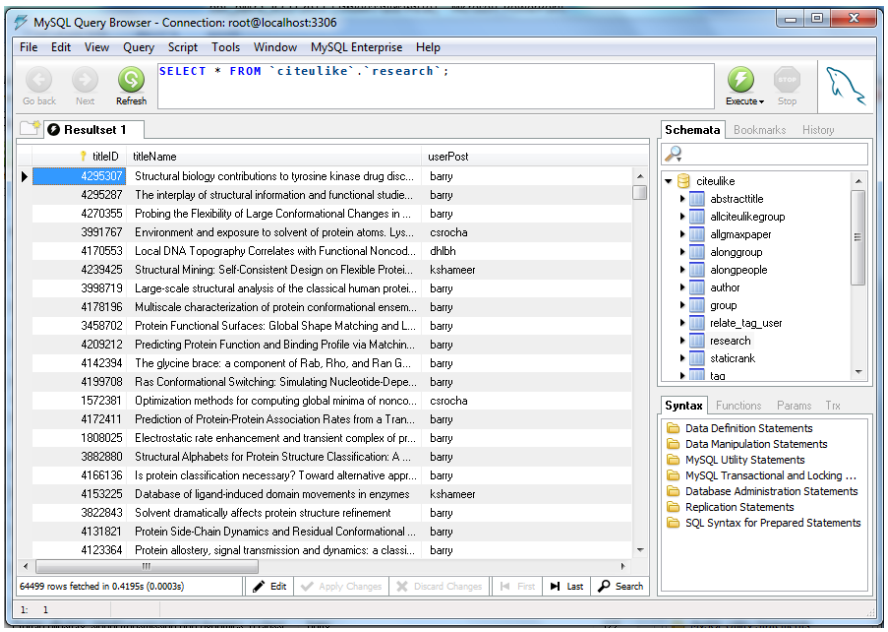
CiteULike ถือเป็นระบบโซเชียลบุ๊กมาร์กด้านงานวิจัยที่ได้รับความนิยมและมีฐานข้อมูลมากที่สุด ผู้วิจัยจึงได้ดำเนินการจัดเก็บข้อมูลจาก CiteULike โดยทำการพัฒนาโปรแกรมเพื่อทำการดึงข้อมูลผ่าน CiteULike ซึ่งดำเนินการจัดเก็บข้อมูลระหว่างเดือนมีนาคม – พฤษภาคม 2553 โดยมีตัวอย่างภาษา HTML ที่นำมาใช้ดังภาพที่ 4.1 และทำการจัดเก็บไว้ฐานข้อมูล ดังภาพที่ 4.2 โดยมีการจัดเก็บข้อมูลได้ทั้งสิ้น 62,192 เอกสารงานวิจัย

```
</td></tr>
<tr class="list">

                    <td class="selectable" >
                          <input id="cb_3063696" style="display:inline"
class="faded"
                          type="checkbox" name="article_id" value="3063696"
onclick="return toggle_cb(3063696)"/> 
                    </td>

<td class="list_item " id="li_3063696">
<span style="clear:right">
<div
      class="article-item-pending is_owner {username:'noo_au9',
article_id:'3063696', group_id:'', base: '/user/noo_au9', link:
'/user/noo_au9/article/3063696', own_pub:0}">
<span class='item-arrows'><img
      id="adon_3063696" class="item-arrow show_details_on {article_id:
'3063696'}" src="/static/img/arrow-collapsed.gif" /><img
      id="adoff_3063696" class="item-arrow show_details_off {article_id:
'3063696'}" src="/static/img/arrow-open.gif" /></span><h2 class="title"><a
class="title" href="/user/noo_au9/article/3063696">Efficient top-k querying
over social-tagging networks</a></h2> <span class='quick-edit'
style="display:none"><a>[Quick Edit]</a></span>
</div><div id="adt_3063696" class="article_details">
<div class="vague">In Proceedings of the 31st annual international ACM
SIGIR conference on Research and development in information retrieval
(2008), pp. 523-530.</div>
<div class="vague">by <a class="author"
href="/user/noo_au9/author/Schenkel:R">Ralf Schenkel</a>, <a class="author"
href="/user/noo_au9/author/Crecelius:T">Tom Crecelius</a>, <a
class="author" href="/user/noo_au9/author/Kacimi:M">Mouna Kacimi</a>, <span
```

```
class='etal'>et al.</span><span class='etallist'><a class="author"
href="/user/noo_au9/author/Michel:S">Sebastian Michel</a>, <a
class="author" href="/user/noo_au9/author/Neumann:T">Thomas Neumann</a>, <a
class="author" href="/user/noo_au9/author/Parreira:JX">Josiane X.
Parreira</a>, <a class="author"
href="/user/noo_au9/author/Weikum:G">Gerhard Weikum</a></span></div>
<span class='vague'>
posted to <span class="taglist"><a class="tag" href="/user/noo_au9/tag/no-
tag">no-tag</a></span>
            by <a href='/user/noo_au9' >noo_au9</a>
on 2010-06-18 14:19:20
<img
                  class="tipsy-hint-s edit_to_read {priority: 3,
user_article_id: 5745732, article_id:3063696}"
                  title="I will read it" alt="***"
                  src="/static/img/star3.png" />/<img class="rating
{article_id:3063696}" style="display:none" src="/static/img/red-
star1.png"></img>
<a href="javascript:toggleInlineLayer('andOthers_5745732')">along with 14
people and 4 groups</a>
<span class="andOthers" id="andOthers_5745732">
<a class="othrusr " href="/user/dungtctin4">dungtctin4</a>
<a class="othrusr " href="/user/thanh">thanh</a>
<a class="othrusr " href="/user/zhaishuang">zhaishuang</a>
<a class="othrusr " href="/user/jarodwen">jarodwen</a>
<a class="othrusr " href="/user/adamsi">adamsi</a>
<a class="othrusr " href="/user/alhoori">alhoori</a>
<a class="othrusr " href="/user/macle">macle</a>
<a class="othrusr " href="/user/eddymier">eddymier</a>
<a class="othrusr " href="/user/andreacapocci">andreacapocci</a>
<a class="othrusr " href="/user/brusilovsky">brusilovsky</a>
<a class="othrusr " href="/user/dtd">dtd</a>
<a class="othrusr " href="/user/wyvern0903">wyvern0903</a>
<a class="othrusr " href="/user/flaviovdf">flaviovdf</a>
<a class="othrusr " href="/user/elsantosneto">elsantosneto</a>
<a class="othrgrp" href="/group/8668">pijitra</a>
<a class="othrgrp" href="/group/3764">Social Web</a>
<a class="othrgrp" href="/group/1252">social_navigation</a>
<a class="othrgrp" href="/group/2118">Adaptive-Web</a>
</span>
</span>
<span class="vague tipsy-hint-s" title="This copy of the article hasn't
been liked by anyone yet" ><img alt="This copy of the article hasn't been
liked by anyone yet"
                  src="/static/img/like14x12.gif" />(0)</span>
<div class="item-icons">
<span class="item-icons-clickable">

                  <a class="articleitem-button articleitem-button-summary
{username:'noo_au9', group_id: '', article_id: '3063696'}">Abstract</a>

</span>
<a id="copy_this_3063696"
      class="articleitem-button articleitem-button-copy tipsy-hint-s
{poster:'noo_au9', article_id:'3063696'}"
      title="Copy to my library"
```

```
        onclick="fnListItems.itemCopyThis(event,this); return
false;">Copy</a>

<a class="articleitem-button articleitem-button-myattachments tipsy-hint-s
{article_id:'3063696'}"
        title="I have attachments on my copy of this article"
        id='mypdf_3063696'>My Attachments</a>
<a id="mycopy_3063696" class="articleitem-button articleitem-button-mycopy
tipsy-hint-s"
        title="View my copy of this article">My Copy</a>
</div>

        <div class="item-abstract">


                <h3>Abstract</h3>
                <p>Online communities have become popular for publishing and
searching content, as well as for finding and connecting to other users.
User-generated content includes, for example, personal blogs, bookmarks,
and digital photos. These items can be annotated and rated by different
users, and these social tags and derived user-specific scores can be
leveraged for searching relevant content and discovering subjectively
interesting items. Moreover, the relationships among users can also be
taken into consideration for ranking search results, the intuition being
that you ...</p>

</div>

</div></span>
```

**ภาพที่ 4.1** แสดงตัวอย่างภาษา HTML จาก www.CiteULike.org

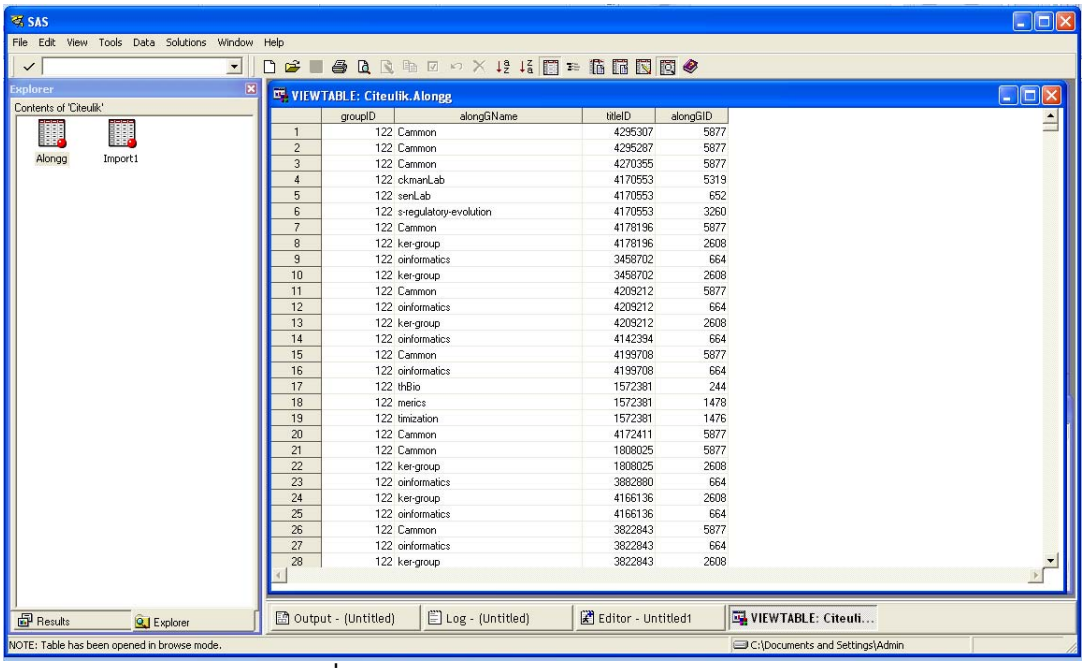



**ภาพที่ 4.2** ตัวอย่างฐานข้อมูล

การจัดเก็บข้อมูลมูลลงโปรแกรม SAS ผู้วิจัยได้ดำเนินการเขียนโปรแกรมเพื่อเตรียมข้อมูลลงในโปรแกรม SAS ปรากฏดังภาพที่ 4.3 และปรากฏผลลัพธ์ดังภาพที่ 4.4

```
PROC IMPORT OUT= CITEULIK.alongG
            DATAFILE= "D:\My Documents\ \test_data\
along_G.xls"
            DBMS=EXCEL REPLACE;
     SHEET="Table1$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
```

**ภาพที่ 4.3** ตัวอย่างโปรแกรมการนำเข้าข้อมูลด้วยโปรแกรม SAS




**ภาพที่ 4.4** ตัวอย่างผลการนำเข้าข้อมูลด้วยโปรแกรม SAS

### 4.2 การศึกษาตัวแบบเพื่อใช้ในการวิเคราะห์ความสัมพันธ์

ในการศึกษาตัวแบบนั้นผู้วิจัยได้เลือกเทคนิคการค้นหากฎความสัมพันธ์ และศึกษาทฤษฎีการค้นหากฎความสัมพันธ์ ดังนี้

ศึกษาการค้นหากฎความสัมพันธ์ รูปแบบของการค้นหากฎความสัมพันธ์สามารถเขียนได้ดังนี้

$$A \rightarrow B$$

โดยที่ A เป็นเงื่อนไข และ B เป็นผลลัพธ์ที่เกิดขึ้น

ในงานวิจัยนี้ผู้วิจัยต้องการค้นหาความสัมพันธ์ระหว่างกลุ่มที่ผู้ใช้งานบนระบบเครือข่ายสังคมออนไลน์ด้านงานวิจัย ซึ่งกำหนดให้เงื่อนไข คือ กลุ่มผู้ใช้ก่อนหน้า หรือ กลุ่มผู้ใช้ X ผลลัพธ์ที่เกิดขึ้น คือ กลุ่มผู้ใช้ที่สัมพันธ์กับกลุ่มก่อนหน้าหรือเรียกว่า กลุ่มผู้ใช้ Y ซึ่งสามารถเขียนกฎความสัมพันธ์ได้ดังนี้

กลุ่มผู้ใช้ X → กลุ่มผู้ใช้ Y

เช่น กลุ่ม Adaptive web → กลุ่ม Social Network

จากความสัมพันธ์นี้สามารถอธิบายได้ว่า ผู้ใช้ที่เป็นสมาชิกกลุ่ม Adaptive web มีแนวโน้มที่จะเป็นสมาชิก กลุ่ม Social Network

การประเมินค่าของกฎจะใช้ค่าสนับสนุน(Support) และค่าความเชื่อมั่น (Confidence) โดยที่

ค่าสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

$$\frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลทั้งหมด}} \quad (4.1)$$

ดังนั้นงานวิจัยนี้สามารถหาค่าสนับสนุนจากสมการดังนี้

ค่าสนับสนุน(กลุ่มผู้ใช้ X → กลุ่มผู้ใช้ Y) เท่ากับ

$$\frac{\text{จำนวนรายการข้อมูลที่มีกลุ่มผู้ใช้ X และกลุ่มผู้ใช้ Y ตามกฎ}}{\text{จำนวนรายการข้อมูลทั้งหมด}} \quad (4.2)$$

ค่าความเชื่อมั่น คือร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนรายการข้อมูลที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

$$\frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลที่เป็นเงื่อนไข}} \quad (4.3)$$

ดังนั้นงานวิจัยนี้สามารถหาค่าสนับสนุนจากสมการดังนี้

ค่าความเชื่อมั่น (กลุ่มผู้ใช้ X → กลุ่มผู้ใช้ Y) เท่ากับ

$$\frac{\text{จำนวนรายการข้อมูลที่มีกลุ่มผู้ใช้ X และกลุ่มผู้ใช้ Y ตามกฎ}}{\text{จำนวนรายการข้อมูลที่มีกลุ่มผู้ใช้ X ตามกฎ}} \quad (4.4)$$

```
%let libname=Au.;
%let datasetname=testmodel;
*ProcessBody;
proc sort data = au.MergetestModel  out=MergetestModel nodupkey;
  by newURL;
run;

/*proc sql;
 create table au.Testmodel as
 select Group_NameX,Group_ID from work.MergetestModel;
 quit;*/
data au.testmodel;
set au.Model;
run;
proc sort data = au.testmodel;
  by Group_NameY;
run;

data au.matchModel au.notmatch;
merge au.testmodel(in=incomp) work.MergetestModel (in=inmerge) ;
by Group_NameY;
 if incomp and inmerge   then do;
        status = "Yes" ;
                    output au.matchModel;
end;
else if incomp and not inmerge then do;
                     status = "No" ;
                    output au.notmatch;
end;
```

```
run;


data au.matchModel1;
  set au.matchModel
      au.notmatch;
run;
```

**ภาพที่ 4.5** ตัวอย่างโปรแกรมการสร้างต้นแบบกฎการวิเคราะห์ความสัมพันธ์ด้วยโปรแกรม SAS

ตารางที่ 4.1 ผลการค้นหากฎความสัมพันธ์

| *Rule* | **ค่าความเชื่อมั่น** *(%)* | **ค่าสนับสนุน** *(%)* |
|---|---|---|
| Genetics-of-Gambling→G4ID | 53.95 | 8.39 |
| G4ID→Genetics-of-Gambling | 100 | 8.39 |
| Philosophy_of_informatic→Blog_and_WikiResearch | 82.46 | 5.16 |
| Blog_and_WikiResearch→Philosophy_of_informatic | 40.22 | 5.16 |
| Statistics and Social Science→Biostatistics | 86.09 | 3.40 |
| Biostatistics→Statistics and Social Science | 88.89 | 3.40 |
| microRNA→Bioinformatics | 72.46 | 3.08 |
| Bioinformatics→microRNA | 17.48 | 3.08 |
| mgh_lcs→Blog_and_WikiResearch | 90.30 | 2.13 |
| Blog_and_WikiResearch→mgh_lcs | 16.60 | 2.13 |
| ReadingLab→Clinical_Psychology | 45.63 | 2.12 |
| Clinical_Psychology→ReadingLab | 85.94 | 2.12 |
| Social navigation→Adaptive-Web | 69.83 | 1.63 |
| Adaptive-Web→Social navigation | 53.76 | 1.63 |
| Automatic sumarization→ASR | 86.64 | 1.31 |
| ASR→Automatic sumarization | 50.50 | 1.31 |
| NLP→ASR | 83.10 | 1.16 |
| ASR→NLP | 44.47 | 1.16 |

จากกฎความสัมพันธ์ที่แสดงในตารางที่ 4.1 สามารถอธิบายได้ดังนี้

กฎที่ 1 สามารถอธิบายได้ว่าผู้ใช้งานที่เป็นสมาชิกในกลุ่ม Genetics-of-Gambling มีแนวโน้มที่จะเข้าไปศึกษาข้อมูลในกลุ่ม G4ID โดยที่คิดเป็น 53.95 % ของจำนวนรายการทั้งหมด

กฎที่ 2 สามารถอธิบายได้ว่าผู้ใช้งานที่เป็นสมาชิกในกลุ่ม G4ID มีแนวโน้มที่จะเข้าไปศึกษาข้อมูลในกลุ่ม Genetics-of-Gambling โดยที่คิดเป็น 100% ของจำนวนรายการทั้งหมด

กฎที่ 3 สามารถอธิบายได้ว่าผู้ใช้งานที่เป็นสมาชิกในกลุ่ม Philosophy_of_informatic มีแนวโน้มที่จะเข้าไปศึกษาข้อมูลในกลุ่ม Blog_and_WikiResearch โดยที่คิดเป็น 82.46 % ของจำนวนรายการทั้งหมด

กฎที่ 4 สามารถอธิบายได้ว่าผู้ใช้งานที่เป็นสมาชิกในกลุ่ม Blog_and_WikiResearch มีแนวโน้มที่จะเข้าไปศึกษาข้อมูลในกลุ่ม Philosophy_of_informatic โดยที่คิดเป็น 40.22 % ของจำนวนรายการทั้งหมด

จากตารางที่ 1 ต้องนำตัวแบบที่ได้มาจำแนกกฎความสัมพันธ์โดยพิจารณาจากหลักเกณฑ์ดังนี้

1. พิจารณาจากค่าความเชื่อมั่นที่สูงสุดของแต่ละเงื่อนไข
2. ถ้าค่าความเชื่อมั่นเท่ากัน ให้พิจารณาค่าสนับสนุนที่สูงสุดของแต่ละเงื่อนไข
3. ถ้าค่าความเชื่อมั่นและค่าสนับสนุนมีค่าเท่ากันให้พิจารณากฎที่มาก่อนให้มีค่าความสำคัญมากกว่า

เมื่อได้ศึกษาขั้นตอนการสร้างตัวแบบโดยการค้นหากฎความสัมพันธ์แล้ว จึงดำเนินการสร้างตัวแบบ และทดสอบตัวแบบในลำดับต่อไป ซึ่งในการทดสอบตัวแบบจะต้องแบ่งข้อมูลออกเป็น 2 ส่วน โดยใช้วิธีการเลือกตัวอย่างแบบมีระบบ

ศึกษาการเลือกตัวอย่างแบบมีระบบ

วีรานันท์ พงศาภักดี (2536 : 59-61) กล่าวว่า การเลือกตัวอย่างแบบมีระบบ เป็นเทคนิคที่ได้รับความนิยมกว้างขวางนอกจากจะใช้ได้กับวิธีการเลือกตัวอย่างด้วยความน่าจะเป็นแบบเท่ากันแล้วยังใช้ได้สะดวกกับวิธีการเลือกตัวอย่างด้วยความน่าจะเป็นที่เป็นสัดส่วนกับขนาด หรือความน่าจะเป็นแบบไม่เท่ากันอีกด้วย

การเลือกตัวอย่างแบบมีระบบจะแบ่งเป็น 2 วิธีคือ

- การเลือกตัวอย่างแบบระบบเส้นตรง

  การเลือกตัวอย่างแบบระบบเส้นตรง มีวิธีการเลือกดังนี้

  1. ให้หมายเลขแก่หน่วยประชากรจาก 1-N
  2. หาช่วงการเลือกตัวอย่าง (Sampling interval) คือ I = N/n โดยที่ n คือ จำนวนตัวอย่างที่ต้องการเลือก
  3. เลือกเลขสุ่ม (Random number) R ใดๆ ขึ้นมาโดยที่ 1< R < I

  ดังนั้นการเลือกตัวอย่างข้อมูล จะได้ผลดังนี้

  หน่วยหมายเลขที่ R คือหน่วยตัวอย่างที่ 1

  หน่วยหมายเลขที่ R+Iคือหน่วยตัวอย่างที่ 2

  ..... .....

  ..... .....

  หน่วยหมายเลขที่ R + (n-1)I คือหน่วยตัวอย่างที่ n

- การเลือกตัวอย่างแบบระบบวงกลม

การเลือกตัวอย่างแบบมีระบบวงกลม มีวิธีการเลือกตัวอย่าง เช่นเดียวกับการเลือกตัวอย่างแบบมีระบบเส้นตรง แต่เป็นวิธีที่เหมาะกับการคำนวณค่า I ซึ่งมีค่าไม่ลงตัวและใช้การปัดเศษให้เป็นตัวเลขจำนวนเต็ม และการเลือกเลขสุ่มจะใช้ R ที่อยู่ระหว่าง 1 – N (0 < R < N)

ตัวอย่าง การเลือกตัวอย่างขนาด 10 หน่วย จากประชากร 500 หน่วย

1. ให้หมายเลขแก่หน่วยประชากรจาก 1 – 500
2. หาช่วงการเลือกตัวอย่าง I = N/n = 500/10 = 50
3. เลือกเลขสุ่ม R = 100 โดยที่ 0 < R < 500

ดังนั้นการเลือกตัวอย่างข้อมูล จะได้ผลดังนี้

หน่วยตัวอย่างที่ 1 คือหน่วยหมายเลขที่ 100

หน่วยตัวอย่างที่ 2 คือหน่วยหมายเลขที่ 150

.....

.....

หน่วยตัวอย่างที่ 9 คือหน่วยหมายเลขที่ 500

หน่วยตัวอย่างที่ 10 คือหน่วยหมายเลขที่ 50 (วนกลับไปเริ่มต้น)

ในงานวิจัยนี้เลือกวิธีการเลือกตัวอย่างแบบมีระบบวงกลม เพราะการสุ่มเลขเริ่มต้นมีช่วงที่กว้างกว่าวิธีการเลือกตัวอย่างแบบมีระบบเส้นตรง ทำให้หน่วยตัวอย่างมีโอกาสที่จะถูกเลือกเท่าๆ กัน

**4.3 ทดสอบตัวแบบเพื่อใช้ในการวิเคราะห์ความสัมพันธ์โดยใช้เทคนิคเหมืองข้อมูล**

ในการสร้างตัวแบบผู้วิจัยจะแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลการเรียนรู้ และข้อมูลตรวจสอบซึ่งการแบ่งข้อมูลจะใช้การเลือกตัวอย่างแบบมีระบบวงกลม

โดยขั้นตอนการเลือกตัวอย่างแบบวงกลมจากนั้นจะทำการแบ่งข้อมูลการเรียกใช้งานของสมาชิกที่ต้องการนำมาทดสอบตัวแบบ ออกเป็นข้อมูลการเรียนรู้ และข้อมูลตรวจสอบ ตามสัดส่วนที่กำหนดโดยใช้วิธีการเลือกตัวอย่างแบบมีระบบวงกลม

สัดส่วนที่ใช้ในการแบ่งข้อมูล เป็นสัดส่วนที่ต้องใช้แบ่งข้อมูลของแต่ละกลุ่มชั้นข้อมูล และการเลือกตัวอย่างแบบมีระบบวงกลมก็จะเลือกตัวอย่างของแต่ละชั้นข้อมูลด้วย โดยมีขั้นตอนดังนี้

4.3.1 กำหนดรหัสลำดับที่ให้กับรายการข้อมูล

กำหนดรหัสลำดับให้กับข้อมูลโดยแบ่งเป็นกลุ่มตามชื่อกลุ่มสมาชิก

4.3.2 คำนวณหาจำนวนรายการข้อมูลทั้งหมดของแต่ละกลุ่ม และคำนวณรายการข้อมูลที่ต้องการตามสัดส่วน

โดยที่

รายการข้อมูลตามสัดส่วน = (จำนวนรายการข้อมูลของแต่ละกลุ่ม * สัดส่วน) / 100

ถ้าตัวเลขที่ได้เป็นทศนิยมให้ปัดเป็นตัวเลขจำนวนเต็ม เช่น 1.66 เมื่อทำการปัดเศษขึ้นเป็นตัวเลขเต็มจะมีค่าเท่ากับ 2

4.3.3 หาช่วงการเลือกตัวอย่าง

4.3.4 เลือกเลขสุ่มโดยการสุ่มตัวเลขเริ่มต้น R ของข้อมูลแต่ละกลุ่ม ซึ่งค่า R จะอยู่ในช่วงตั้งแต่ 1 ถึง N

ดังนั้นผลของการเลือกตัวอย่างจะได้รายการข้อมูลการเรียนรู้ (Train) ที่เลือกได้ในแต่ละกลุ่มจะได้ผลดังตารางและสำหรับข้อมูลตรวจสอบ (Validation) คือรายการข้อมูลที่ไม่ได้ถูกเลือก ซึ่งจะต้องทำการค้นหากฎความสัมพันธ์จากข้อมูลการเรียนรู้และข้อมูลตรวจสอบที่ได้เลือกไว้

**4.4 ทดสอบความถูกต้องของตัวแบบ**

การทดสอบความถูกต้องของตัวแบบสามารถแบ่งเป็นหัวข้อต่างๆดังนี้

4.4.1 การนำข้อมูลตรวจสอบที่เลือกได้มาตรวจสอบความถูกต้องของตัวแบบ

การตรวจสอบความถูกต้องของตัวแบบผู้วิจัยได้ทำการคำนวณหาค่าความเชื่อมั่นข้อมูลเรียนรู้ และค่าความเชื่อมั่นของข้อมูลตรวจสอบ เพื่อทำการตรวจสอบความถูกต้องดังตารางที่4.2

ตารางที่ 4.2 ตัวอย่างผลการทดสอบความถูกต้องตัวแบบ

| กฎที่ | Rule | %ค่าความเชื่อมั่นข้อมูลเรียนรู้ | %ค่าความเชื่อมั่นข้อมูลตรวจสอบ | ความถูกต้อง |
|---|---|---|---|---|
| 1 | Genetics-of-Gambling→G4ID | 53.95 | 45.00 | F |
| 2 | G4ID→Genetics-of-Gambling | 100 | 100 | T |
| 3 | Philosophy_of_informatic→Blog_and_WikiResearch | 82.46 | 90.00 | T |
| 4 | Blog_and_WikiResearch→Philosophy_of_informatic | 40.22 | 18.12 | F |
| 5 | Statistics and Social Science→Biostatistics | 86.09 | 92.35 | T |
| 6 | Biostatistics→Statistics and Social Science | 88.89 | 91.74 | T |
| 7 | microRNA→Bioinformatics | 72.46 | 85.00 | T |
| 8 | Bioinformatics→microRNA | 17.48 | 5.30 | F |
| 9 | mgh_lcs→Blog_and_WikiResearch | 90.30 | 100 | T |
| 10 | Blog_and_WikiResearch→mgh_lcs | 16.60 | 7.87 | F |
| 11 | ReadingLab→Clinical_Psychology | 45.63 | 30.56 | F |
| 12 | Clinical_Psychology→ReadingLab | 85.94 | 95.00 | T |
| 13 | Social navigation→Adaptive-Web | 69.83 | 72.14 | T |
| 14 | Adaptive-Web→Social navigation | 53.76 | 40.12 | F |
| 15 | Automatic sumarization→ASR | 86.64 | 100 | T |
| 16 | ASR→Automatic sumarization | 50.50 | 45.34 | F |
| 17 | NLP→ASR | 83.10 | 99.00 | T |
| 18 | ASR→NLP | 44.47 | 32.12 | F |

จากตารางที่ 4.2 สามารถอธิบายได้ว่า
ลำดับที่ 1 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้ใช้งานในกลุ่ม Genetics-of-Gambling จำนวน 53.95 % ที่จะเข้าไปศึกษาข้อมูลในกลุ่มผู้ใช้ G4ID เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า สมาชิกในกลุ่ม Genetics-of-Gambling 45.00 % ที่เข้าใช้บริการกลุ่ม G4ID ดังนั้นลำดับที่ 1 จึงมีความถูกต้องเท่ากับ F

ลำดับที่ 2 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้ที่ใช้งานในกลุ่ม G4ID ทั้งหมด จะเข้าไปศึกษาข้อมูลในกลุ่ม Genetics-of-Gambling เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่าสมาชิกในกลุ่ม G4ID ทั้งหมด ที่เข้าใช้บริการกลุ่ม Genetics-of-Gambling ดังนั้นลำดับที่ 2 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 3 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้ใช้งานในกลุ่ม Philosophy_of_informatic จำนวน 82.46% ที่จะเข้าไปศึกษาข้อมูลในกลุ่มผู้ใช้ Blog_and_WikiResearch เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า สมาชิกในกลุ่ม Philosophy_of_informatic 90.00 % ที่เข้าใช้บริการกลุ่ม Blog_and_WikiResearch ดังนั้นลำดับที่ 3 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 4 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้ใช้งานในกลุ่ม Blog_and_WikiResearch จำนวน 40.22% ที่จะเข้าไปศึกษาข้อมูลในกลุ่มผู้ใช้ Philosophy_of_informatic เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า สมาชิกในกลุ่ม Blog_and_WikiResearch 18.12 % ที่เข้าใช้บริการกลุ่ม Philosophy_of_informatic ดังนั้นลำดับที่ 4 จึงมีความถูกต้องเท่ากับ F

ลำดับที่ 5 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้ใช้งานในกลุ่ม Statistics and Social Science 86.09% ที่จะเข้าไปศึกษาข้อมูลในกลุ่มผู้ใช้ Bioinformatics เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า สมาชิกในกลุ่ม Statistics and Social Science 92.35 % ที่เข้าใช้บริการกลุ่ม Bioinformatics ดังนั้นลำดับที่ 5 จึงมีความถูกต้องเท่ากับ T เป็นต้น

ตารางที่ 4.3 แสดงร้อยละความถูกต้องของการทดสอบตัวแบบ

| **การสุ่มครั้งที่** | **จำนวนตัวอย่างสุ่ม** | **ร้อยละความถูกต้อง** |
|---|---|---|
| 1 | 50 | 79.13 |
| 2 | 63 | 87.45 |
| 3 | 61 | 85.18 |
| **ค่าเฉลี่ยรวม** | | **83.92** |

จากตารางที่ 4.3 เมื่อนำความถูกต้องของตัวแบบข้อมูลการเรียนรู้จากการสุ่มตัวอย่างครั้ง 3 ครั้ง มาคำนวณร้อยละของความถูกต้องพบว่ามีความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 83.92 ของจำนวนรายการทั้งหมด โดยร้อยละของความถูกต้องครั้งที่ 1 เท่ากับ 79.13, ครั้งที่ 2 เท่ากับ 87.45

และ ครั้งที่ 3 เท่ากับ 85.18 ดังนั้นสามารถนำตัวแบบที่ได้ศึกษานี้ไปใช้ในการทำนายเนื้อหาเว็บได้ต่อไป

**4.4.2 การทดสอบระบบ**

การทดสอบระบบผู้วิจัยได้ทดสอบระบบโดยดูจากผลลัพธ์ที่ได้ในแต่ละกระบวนการ โดยการตรวจสอบจากความถูกต้องของข้อมูล ดังตาราง

ตารางที่ 4.4 แสดงกระบวนการและวิธีทดสอบ

| กระบวนการ | วิธีการทดสอบ |
|---|---|
| การนำเข้าข้อมูล | ตรวจสอบจำนวนรายการข้อมูลที่นำเข้าก่อนและหลังนำเข้า ตรวจสอบชนิด |
| กระบวนการก่อนสร้างตัวแบบ<br>- สร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร็อกซีเซิร์ฟเวอร์ | ตรวจสอบจำนวนรายการทั้งหมด และจำนวนรายการหลังจากประมวลผลต้องเท่ากัน<br>ตรวจสอบจำนวนรายการทั้งหมด และจำนวนรายการหลังจากประมวลผลต้องเท่ากัน |
| การศึกษาตัวแบบ<br>- การเลือกตัวอย่างข้อมูลแบบมีระบบวงกลม | ตรวจสอบการทำงานของโปรแกรม โดยการตรวจสอบรายการข้อมูลที่เลือกได้ต้องถูกต้องตามวิธีการเลือกตัวอย่างแบบมีระบบวงกลม |
| - การสร้างตัวแบบข้อมูลเรียนรู้ | ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวนรายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่าความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้ |
| - การทดสอบความถูกต้องตัวแบบ<br>- การสร้างตัวแบบข้อมูลตรวจสอบ | ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวนรายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่าความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้<br>ตรวจสอบการเปรียบเทียบค่าความเชื่อมั่นมีความถูกต้องตามโปรแกรม และคำนวณร้อยละความถูกต้องของตัวแบบ |
| - คำนวณร้อยละความถูกต้อง | |
| การสร้างตัวแบบ | ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวนรายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่า |

| | |
|---|---|
| | ความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้ |
| การจำแนกกฎความสัมพันธ์ | ตรวจสอบความถูกต้องในแต่ละเงื่อนไขจะต้องมีกฎเดียวและเป็นกฎที่มีค่าความเชื่อมั่นและค่าสนับสนุนสูงกว่าค่าความเชื่อมั่นและค่าสนับสนุนที่ต่ำสุด |

**4.4.3 การประเมินผลระบบ**

ผู้วิจัยได้พัฒนาโปรแกรมและมีการทดสอบการใช้โปรแกรม พบว่าโปรแกรมที่พัฒนาขึ้นสามารถสร้างตัวแบบในการวิเคราะห์ความสัมพันธ์ของการเป็นสมาชิกกลุ่มของผู้ใช้แต่ละคนเพื่อเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัย และสามารถนำตัวแบบที่สร้างไปประยุกต์ใช้ได้ในการทดสอบระบบในแต่ละขั้นตอนมีการสรุปผลดังนี้

**4.4.3.1 สรุปผลการทดสอบความถูกต้องตัวแบบ**

ผู้วิจัยได้สร้างตัวแบบข้อมูลเรียนรู้ และสร้างตัวแบบข้อมูลตรวจสอบ โดยใช้ข้อมูลที่แบ่งได้ตามสัดส่วนต่างๆ และตรวจสอบความสอดคล้องกันของตัวแบบได้โดยผลจากการทดสอบที่ได้สามารถสรุปได้ว่า ในการแบ่งสัดส่วนแต่ละครั้งจะได้ผลลัพธ์ความถูกต้องไม่เท่ากัน เพราะในแต่ละครั้งของการเลือกตัวอย่างข้อมูลจะได้ข้อมูลไม่เหมือนกัน ดังนั้นความถูกต้องของตัวแบบจะขึ้นอยู่กับข้อมูลที่เลือกมาได้ด้วย แต่จะพบว่าการแบ่งสัดส่วนข้อมูลจะมีผลกับความถูกต้องด้วย นั่นคือ ถ้าสัดส่วนการแบ่งข้อมูลมากตัวแบบจะมีร้อยละความถูกต้องมากขึ้นด้วย

**4.4.3.2 สรุปผลการสร้างตัวแบบ**

เมื่อตัวแบบที่ผู้วิจัยศึกษาสามารถนำมาใช้ได้ จึงใช้ข้อมูลการเป็นสมาชิกของผู้ใช้ในแต่ละกลุ่มที่ผ่านกระบวนการสร้างความสัมพันธ์เรียบร้อยแล้วมาหาความสัมพันธ์ และจากการสร้างตัวแบบพบว่ากฎความสัมพันธ์ที่ค้นหาได้มีจำนวนมาก จึงต้องมีการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด เพื่อเลือกเฉพาะกฎความสัมพันธ์ที่มีความเชื่อมั่นสูง

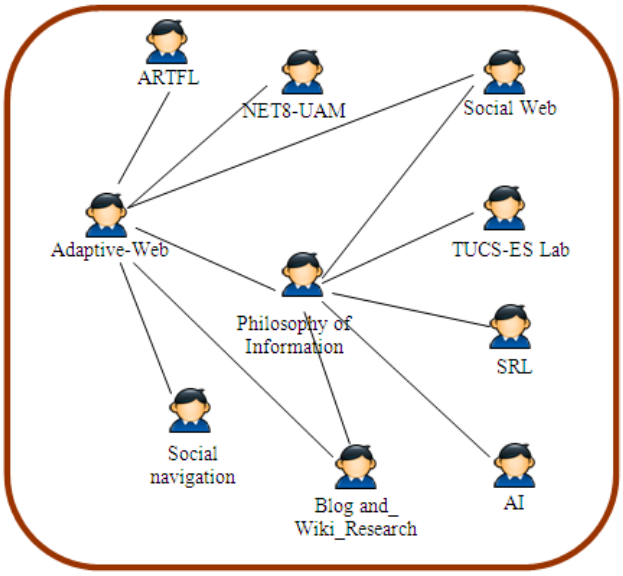
**4.4.3.3 สรุปผลการจำแนกกฎความสัมพันธ์**

การจำแนกกฎความสัมพันธ์เป็นการเลือกกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นและค่าสนับสนุนสูงสุดของแต่ละกฎความสัมพันธ์ จะมีความสอดคล้องกับการกำหนดค่าความเชื่อมั่นและค่าสนับสนุนในการสร้างตัวแบบ นั่นคือเมื่อจำแนกกฎแล้วในแต่ละกฎจะมีค่าความเชื่อมั่นและค่าสนับสนุนน้อยที่สุดจะเท่ากับค่าความเชื่อมั่นและค่าสนับสนุนต่ำสุดตามที่กำหนดตอนสร้างตัวแบบ

**4.4.3.4 สรุปผลการตรวจสอบความถูกต้องกับข้อมูลจริง**

งานวิจัยนี้ได้ทำการตรวจสอบความถูกต้องโดยทำการนำข้อมูลที่ทำการสร้างโมเดลมาเปรียบเทียบกับข้อมูลจริงของข้อมูลซึ่งพบว่าโมเดลที่สร้างขึ้นสามารถหาความสัมพันธ์

ผู้ใช้ที่สนใจเป็นสมาชิกในแต่ละกลุ่ม และสามารถนำมาช่วยในการจัดอันดับผลการค้นหางานวิจัยได้ โดยข้อมูลที่ใช้ในการตรวจสอบเป็นข้อมูลตรวจสอบที่ถูกแบ่งตามสัดส่วนที่นักวิจัยต้องการ ซึ่งหากค่าความเชื่อมั่นของข้อมูลตรวจสอบเพิ่มขึ้น อาจทำให้ประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยเพิ่มขึ้นด้วย นอกจากนี้จากผลการวิเคราะห์ความสัมพันธ์ของกลุ่มผู้ในงานในเครือข่ายสังคมด้านงานวิจัยที่ได้สามารถนำมาสร้างเป็นแผนภาพการเชื่อมโยงความสัมพันธ์ของผู้ใช้ได้ดังภาพที่ 4.6




**ภาพที่ 4.6** ตัวอย่างแผนภาพการเชื่อมโยงความสัมพันธ์จากผลลัพธ์กฎการวิเคราะห์ความสัมพันธ์

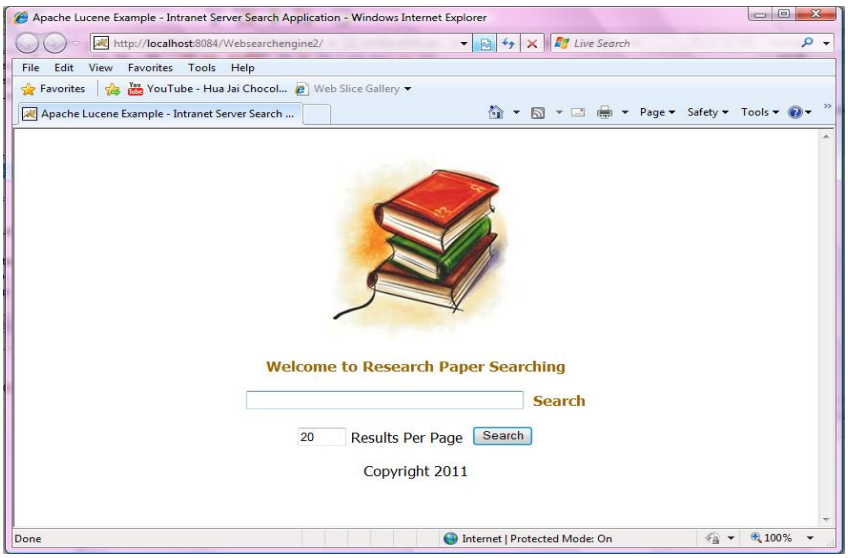
## 4.5 การทดลองและการประเมินผล

### 4.5.1 การทดลองกับผู้ใช้งานระบบ

ผู้วิจัยได้ทำการเชิญผู้ทดสอบระบบ จำนวน 20 คน เพื่อมาทำทดลองและการประเมิน โดยผู้ทดสอบระบบจะต้องทำการค้นหาตามคำค้นหาที่เกี่ยวข้องกับงานวิจัยของผู้ทดสอบระบบแต่ละคน จำนวน 5 คำค้นหา โดยทำการประเมินผลจาก 2 ระบบ คือ

1) ระบบที่พัฒนาจากการจัดอันดับแบบ *Similarity ranking*

2) ระบบที่พัฒนาจากการจัดอันดับแบบ *Similarity ranking* ที่นำกฎการวิเคราห์ความสัมพันธ์มาช่วยในการจัดอันดับ

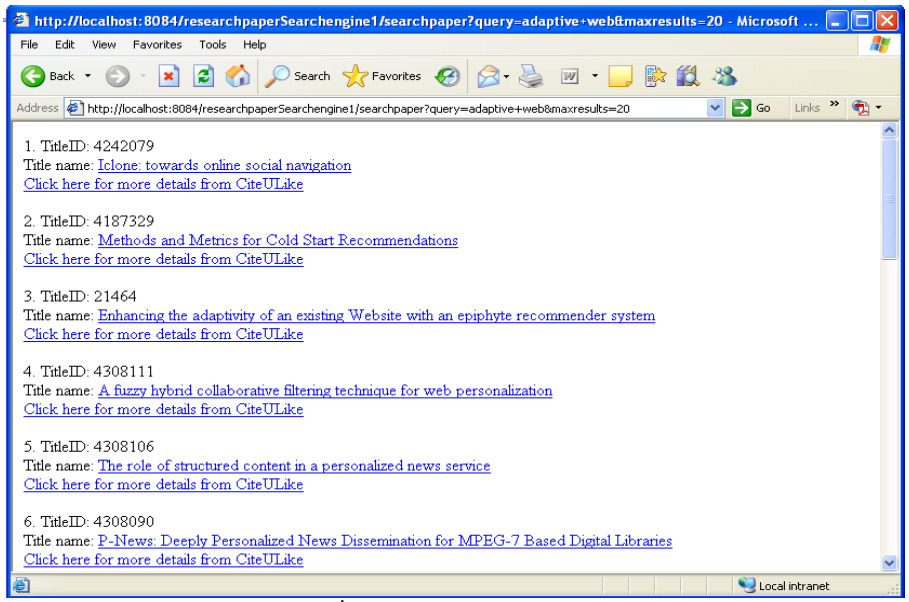



**ภาพที่ 4.7** ตัวอย่างหน้าต่างการทำงานของระบบที่พัฒนาจากการจัดอันดับแบบ *Similarity ranking*

ภาพที่ 4.7 และภาพที่ 4.8 แสดงหน้าต่างการทำงานและผลการค้นหา โดยผู้ทดสอบจะต้องทำการประเมินผลการค้นหาที่ได้จากระบบ ทั้งนี้ใน 1 คำค้นหาจะแสดงผลการค้นหาจำนวน 10 รายการ โดยมีคะแนนประเมินผลการค้นหาถูกกำหนดไว้ 2 ระดับ ดังนี้

ระดับ 1 คือ ผลลัพธ์เกี่ยวข้องกับคำค้น

ระดับ 0 คือ ผลลัพธ์ไม่เกี่ยวข้องกับคำค้นหา




**ภาพที่ 4.8** ตัวอย่างหน้าต่างผลการค้นหา

**4.5.2 การวัดประสิทธิภาพความแม่นยำในการสืบค้น**

$$recall = \frac{\textit{จำนวนเอกสารที่เกี่ยวข้องทั้งหมดและถูกดึงออกมา}}{\textit{จำนวนเอกสารที่ถูกต้องทั้งหมด}} \quad (4.5)$$

$$\text{Pr}\,ecision = \frac{\textit{จำนวนเอกสารที่ถเกี่ยวข้องและถูกดึงออกมา}}{\textit{จำนวนเอกสารที่ถูกดึงออกมาทั้งหมด}} \quad (4.6)$$

งานวิจัยนี้ได้ทำการประเมินผลโดยใช้ Precision/recall ตามที่ปรากฏดังสมการที่ (4.7), (4.8) ดังรายละเอียดต่อไปนี้

$$recall = (D(u,r)) = \frac{1}{|U|}\sum_{u \in U}\frac{|rp(u,r) \cap D(u,r)|}{|rp(u,r)|} \quad (4.7)$$

$$precision = (D(u,r)) = \frac{1}{|U|}\sum_{u \in U}\frac{|rp(u,r) \cap D(u,r)|}{|D(u,r)|} \quad (4.8)$$

กำหนดให้

$|rp(u,r)|$ คือ จำนวนเอกสารที่เกี่ยวข้องทั้งหมด

$|D(u,r)|$ คือเอกสารที่ถูกดังออกมา

$|rp(u,p) \cap D(u,p)|$ คือเอกสารที่เกี่ยวข้องทั้งหมดและถูกดังออกมา

ในการนำเสนอวิธีการประเมินประสิทธิภาพในระบบค้นหางานวิจัยนั้นนั้นนิยมคำนวนหาค่า Precision และ Recall โดยที่ค่า Recall เท่ากับ 1 หมายถึงระบบสามารถค้นคืนงานวิจัยที่เกี่ยวข้องกับภาพสอบถามทั้งหมดออกมาได้ เช่นเดียวกับค่า Precision ค่า Precision ที่สูงหมายถึงระบบสามารถค้นคืนงานวิจัยออกมาได้โดยมีเอกสารงานวิจัยที่ไม่เกี่ยวข้องปะปนอยู่น้อย ดังนั้นค่า Precision และค่า Recall จึงมักจะนำมาพิจารณาร่วมกัน เช่น พิจารณาค่า Precision ที่ Recall เท่ากับ 0.5 เป็นต้น สามารถกล่าวได้ว่าระบบค้นคืนรูปภาพมีประสิทธิภาพดีถ้ามีค่า Precisionสูงด้วยค่า Recall ที่เท่ากัน

ตารางที่ 4.5 แสดงร้อยละความถูกต้องของการทดสอบตัวแบบ

| *ค่าเฉลี่ย Precision* | *Original Method* | *Our Method* |
|---|---|---|
| P@1 | 55.0% | 73.2% |
| P@2 | 43.1% | 65.7% |
| P@3 | 40.2% | 63.2% |
| P@4 | 38.7% | 61.1% |
| P@5 | 35.2% | 55.5% |
| P@10 | 27.0% | 49.0% |
| P@15 | 25.5% | 44.1% |

เนื่องจากงานวิจัยนี้ได้ทำการกำหนดผลการค้นหาให้กับผู้ประเมินเพียง 10 อันดับจึงสามารถคำนวณได้เพียงค่า Precision และค่าเฉลี่ย Precision เท่านั้น ดังปรากฏในตารางที่ 4.5 ทั้งนี้จากตารางแสดงค่า ให้เห็นว่าค่า *Precision* ที่พัฒนาขึ้นมาใหม่โดยใช้เทคนิคเหมืองข้อมูลมาทำการประยุกต์ใช้นั้นสามารถเพิ่มประสิทธิภาพผลการค้นได้ถึงร้อยละ 20

# บทที่ 5
# สรุปและข้อเสนอแนะ

ในบทนี้ได้อธิบายถึงการสรุปผลการทดลอง และข้อเสนอแนะโดยมีรายละเอียดดังนี้

## 5.1 สรุปผลการวิจัย

ผู้วิจัยได้เสนอขั้นตอนวิธีการศึกษาตัวแบบและพัฒนาตัวแบบ เพื่อนำตัวแบบที่ได้นำไปวิเคราะห์หาความสัมพันธ์ของกลุ่มผู้ใช้เพื่อนำมาช่วยในการจัดอันดับผลการค้นหางานวิจัย โดยขั้นตอนวิธีที่นำเสนอสามารถแบ่งได้ดังนี้

ส่วนแรกเป็นการดำเนินการก่อนสร้างตัวแบบโดยทำการเลือกกลุ่มผู้ใช้ที่ต้องการวิเคราะห์ความสัมพันธ์ และทำการสร้างความสัมพันธ์ของข้อมูล เพื่อเป็นการเตรียมข้อมูลก่อนการศึกษาตัวแบบหรือก่อนการสร้างตัวแบบ

ส่วนที่ 2 เป็นการศึกษาตัวแบบเพื่อใช้ในการวิเคราห์ความสัมพันธ์ โดยใช้เทคนิคการค้นหากฎความสัมพันธ์ ผู้วิจัยได้ศึกษาตัวแปรที่จะนำมาใช้สร้างเป็นเงื่อนไขและผลลัพธ์ ซึ่งในการวิจัยนี้ได้ใช้กลุ่มผู้ใช้ที่มีความสัมพันธ์ก่อนหน้าเป็นเงื่อนไข ส่วนผลลัพธ์คือกลุ่มผู้ใช้ที่มีความสัมพันธ์ถัดมา นั่นคือการสร้างตัวแบบนี้จะต้องหากฎความสัมพันธ์ระหว่าง กลุ่มผู้ใช้ที่ผู้ใช้งานบนระบบเครือข่ายสังคมสนใจ โดยที่กฎความสัมพันธ์ที่ได้จะต้องคำนวณค่าความเชื่อมั่นและค่าสนับสนุน

ส่วนที่ 3 เป็นการพัฒนาตัวแบบโดยนำข้อมูล ที่ผ่านการกระบวนการก่อนสร้างตัวแบบเรียบร้อยแล้วมาสร้างตัวแบบตามที่ได้ศึกษามา โดยแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลเรียนรู้ และข้อมูลตรวจสอบ ในการแบ่งข้อมูลได้ใช้วิธีการเลือกตัวอย่างแบบมีระบบวงกลม และนำข้อมูลแต่ละชุดมาสร้างตัวแบบเป็นตัวแบบข้อมูลเรียนรู้ และตัวแบบข้อมูลตรวจสอบ

ส่วนที่ 4 เป็นการทดสอบตัวแบบโดยจะนำตัวแบบข้อมูลเรียนรู้และตัวแบบข้อมูลตรวจสอบมาเปรียบเทียบเพื่อหาความสอดคล้องกันของตัวแบบ ถ้าผลการเปรียบเทียบได้ร้อยละความสอดคล้องเกินร้อยละ 50 นั่นหมายความว่าตัวแบบที่ศึกษาและพัฒนามานั้นสามารถนำไปใช้ในการวิเคราะห์ความสัมพันธ์ได้

ส่วนที่ 5 เป็นตรวจสอบผลการสร้างความสัมพันธ์จากข้อมูลจริง โดยทำการนำข้อมูลที่ทำการสร้างโมเดลมาเปรียบเทียบกับข้อมูลจริงของข้อมูลซึ่งพบว่าโมเดลที่สร้างขึ้นสามารถนำไปวิเคราะห์หาความสัมพันธ์ของกลุ่มผู้ใช้ได้และสามารถนำไปประยุกต์ใช้ในการเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหางานวิจัยได้

ส่วนที่ 6 งานวิจัยฉบับนี้ได้พัฒนาระบบและนำเทคนิคการจัดอันดับผลการค้นหา ของ 1) การจัดอันดับแบบ *Similarity ranking* และ 2) *Similarity ranking*ที่นำกฎการวิเคราะห์ความสัมพันธ์มาช่วยในการจัดอันดับ ทั้งนี้การพัฒนาอัลกอริทึมดังกล่าวพัฒนามาจากการสร้างดัชนีโดยใช้ “Tag Title และ Abstract” หรือ TTA โดยงานวิจัยนี้ได้ทำการทดสอบระบบจากผู้ทดสอบระบบจำนวน **20** คน โดยมีคะแนนการประเมินผลการค้นหาจำนวน **2** ระดับ

ผู้วิจัยได้ดำเนินการประเมินผลการทดลองโดยใช้ ค่าเฉลี่ย Precision โดยผลการทดลองพบว่า *Similarity ranking*ที่นำกฎการวิเคราะห์ความสัมพันธ์มาช่วยในการจัดอันดับ มีค่าสูงกว่าการจัดอันดับโดยใช้อัลกอริทึมแบบเดิม ซึ่งสามารถที่จะเพิ่มประสิทธิภาพการจัดอันดับผลการค้นหาได้อย่างมีประสิทธิภาพ

งานวิจัยที่จะทำการพัฒนาต่อในอนาคตเป็นการเพิ่มประสิทธิภาพโดยการสร้างโปรไฟล์ผู้ใช้เพื่อทำการเพิ่มประสิทธิภาพผลการค้นหาสำหรับงานวิจัยบนระบบเครือข่ายสังคมและโซเชียลบุ๊กมาร์กต่อไป

**5.2 ข้อเสนอแนะ**

1. การทำงานโดยใช้เทคนิคเหมืองข้อมูลในงานวิจัยนี้ความถูกต้องจะขึ้นอยู่กับการสร้างตัวแบบและข้อมูล ที่นำมาเสร้างตัวแบบ
2. การจัดเก็บข้อมูลจากเว็บจำเป็นจะต้องทราบถึงโครงสร้างของเว็บที่ต้องการดึงข้อมูล
3.. ยังมีอัลกอริทึมอื่น ๆ ที่ใช้ในระบบเครือข่ายสังคมเช่น อัลกอริทึมระบบการแนะนำข้อมูล (Recommender System)

**บรรณานุกรม**


ดร.ศุภชัย ตั้งวงศ์ศานต์ (2551). **ระบบการจัดเก็บและการสืบค้นสารสนเทศด้วยคอมพิวเตอร์.** กรุงเทพฯ: โรงพิมพ์พิทักษ์การพิมพ์, 2551

Berberich, K., Vazirgiannis, M., and Weikum, G.(2004) '**T-Rank: Time-Aware Authority Ranking**', *WAW 2004*, 2004.

Capocci, A. and Caldarelli, G. '**Folksonomies and Clustering in the Collaborative System CiteULike**', Internet: *http://arxiv.org/abs/0710.2835,* [accessed 8/4/2010].

Choochaiwattana, W., and Spring, M.B. (2009)'**Applying Social Annotations to Retrieve and Re-rank Web Resources**'. *Proceedings of 2009 International Conference on Information Management and Engineering (ICIME 2009), April 3 – 5, 2009*,Kuala Lumpur, Malaysia.

CiteULike, Internet: http://www.CiteULike.org, [accessed 8/04/2010].

Farooq, U., Ganoe, C.H., Carroll, J.M., and Giles, C.L. (2007a) '**Supporting distributed scientific collaboration: Implications fordesigning the CiteSeer collaborator**' *Proceedings of the Hawaii Int'l Conference on System Sciences*, IEEE Compute Society, *January 3-6, 2007*,Waikoloa, Hawaii.

Farooq, U., Kannampallil, T.G., Song, Y. , Ganoe, C.H. , Carroll, John M. ,and Giles, C. Lee. (2007b) '**Evalating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics**', *Proceedings of the 2007 international ACM conference on Supporting group work (GROUP'07). November 4-7, 2007*,Sanibel Island, Florida, USA.

Gordon, M., and Pathak, P. (1999) 'Finding information on the World Wide Web: the retrieval effectiveness of search engines'. ***Information Processing and Management: an International Journal***, Vol. 35, No. 2, pp.141–180.

Jomsri, P. , Sanguansintukul, S., and Choochaiwattana, W.(2009a) ' **Improve Research paper Searching with social tagging-A Preliminary Investigation**'. , *The Eight International Symposium on Natural Language Processing (SNLP2009). October 20-21, 2009.* Bangkok.

Jomsri, P. Sanguansintukul, S., and Choochaiwattana , W. (2009b) '**A Comparison of Search Engine Using "Tag Title and Abstract" with CiteULike – An Initial Evaluation**' *The 4th International Conference for Internet Technology and Secured Transactions (ICITST-2009)*. November 9-12 ,2009. London ,UK.

Long, H., Lv, B., Zhao, T., and Liu, Y. (2007)'Evaluate and Compare Chinese Internet Search Engines Based on Users'Experience'. *Wireless*

*Communications, Networking and Mobile Computing, 2007. WiCom 2007. September 21-25, 2007*.

Richardson, M. , Prakash, A. ,and Brill, E. (2006)'**Beyond PageRank: Machine Learning for Static Ranking'**. *Proceedings of the 15th international conference on World Wide Web***,** *WWW 2006.* May 23–26, 2006, Edinburgh, Scotland.

R. Forsati, M. R. Meybodi และ A. Ghari Neiat (2009) '***Web Page Personalization based on Weighted Association Rules***' 2009 International Conference on Electronic Computer Technology

Santos-Neto, E. , Ripeanu, M. , and Iamnitchi, A. (2007) '**Tracking usage in collaborative tagging communities'**. Internet: *http://www.csee.usf.edu/~anda/papers/CAMA07_ready_v2.pdf,* [accessed 8/4/2010].

Sun, Y. and Lee Giles, C. (2007) **'Popularity Weighted Ranking for Academic Digital Libraries'**. *ECIR 2007*, LNCS 4425, pp. 605–612, 2007.

Thomas, M.,(2006) **'Implementation and Evaluation of a Quality-Based Search Engine'**. *Conference on Hypertext and Hypermedia*, ACM,2006.

Toine Bogers, T., and van den Bosch, A. (2008) **'Recommending Scientific Articles Using CiteULike'**. *RecSys'08*, *October 23-25, 2008*.Switzerland.

# ภาคผนวก ก
# เอกสารการนำเสนอผลงานวิจัยระดับนานาชาติ

# Analyzing the Relation of Community Group for Research Paper Bookmarking by Using Association Rule

P. Jomsri

***Abstract*—** Currently searching through internet is very popular especially in a field of academic. A huge of educational information such as research papers are overload for user. So community-base web sites have been developed to help user search information more easily from process of customizing a web site to need each specifies user or set of user. In this paper propose to use association rule analyze the community group on research paper bookmarking. A set of design goals for community group frameworks is developed and discussed. Additionally Researcher analyzes the initial relation by using association rule discovery between the antecedent and the consequent of a rule in the groups of user for generate the idea to improve ranking search result and development recommender system.

***Keywords*—** association rule, information retrieval, research paper bookmarking.

## I. Introduction

THE Recently, researching within the information retrieval has considered an alternative approach of retrieving the information based on community of users in the system. Many social bookmarking systems have been designed and implemented for improve systems. Especially, Social resource sharing systems are web-based systems that allow users to upload all kinds of resources.

Furthermore, Search engines are the important tools that people search document on internet. It can return search result by user query. Nowadays, social network has recently received a wide adoption by various web services such as social bookmarking systems. They provide functions that allow users to share content with one another. In a field of academic have a several work of research to regard one which use search engine for searching research paper and investigate the literature reviews such as *CiteULike*[1]. It helps scientists, researchers and academics store, organize, share and discover links to academic research papers. *Connotea*[2] is a free online reference management for all researchers, clinicians and scientists. *BibSonomy*[3] is a system for sharing bookmarks, lists of literature and BIBTEX based publication entries simultaneously. However, the best known in the academic and research paper arena is CiteULike.

As part of social research paper bookmarking system has community group which perhaps each community may concentrate on the same topic. In addition, user in social bookmarking system can join with another groups or communities that user interest. Those communities which users are members may related content or research topic. Therefore, in this paper proposed to analyze the relation of community group for research paper bookmarking by using association rule. The main point is study the relation of user group by using data mining techniques for optimize ranking.

The paper is organized as follows. Section II discusses related works. The framework of this paper is described in Section III. The association rule analysis explained in Section IV, The experimental setting is shown in Section V. Results and discussions from the experiments are presented in Section VI. Finally, the conclusion and future work are given in Section VII.

## II. Related Work

This section contain in to two parts: first was background of community based on social bookmarking and second was related research with CiteULike.

### *A. Community based on social bookmarking*

In recent years, many studies of community-based on search engine have been carried out. The main techniques involved in community-based search engine include *recommendation*, *relevance feedback*, *personalization*, and their combinations. Many research try to measures the similarity or relation between groups for improve the performance of recommender system such as Senot and *et al.* build group profile of TV viewing data by combine with individual user for showing how group interest[4]. Therefore the group personnel relationship exists in social groups of all sorts, which can be researched using the knowledge of the complex social networks system.

There are several specific research projects on community of social network such as Cohen, and Havlin studied the degree distribution of co-author research network in mathematics and the neuroscience domain. These distributions do not strictly follow the power-law distribution [15]. Zhang and Di described the clustering algorithms of co-author research network [16]. Chang, and Daren showed the results of proprietary Chinese medicine network in 2005 [23]. Hong, Wei-dong, and Wen analyze relation of group personnel relationship. By comparing the group personnel relationship model and the empirical model, the simulation results in according with the empirical findings quit well [17].

Some researchers applied association rule mining for improve the web performance such as Heymann, Ramage, and Garcia-Molina, [12] use association rule mining based in combinations with other measures for link prediction on social

tags. Schmitz and *et al.* [13] describe the idea of using association rules to determine hypernymy and hyponymy relations between tags in social tagging data. They have a strong emphasis on formal concept analysis and its usage in context of social tagging data.

Although this paper are following a similar initial thought of utilizing classical data mining techniques for discovering structures in social bookmarking. This paper focus is on structuring groups of user. Aim of this is suggest to the user group bundles for organizing information. The idea is that research paper was assigned by any given users which are a reflection of his interesting and share research paper with other user in the same group. In addition, relationships between group and their user perceived "paper" can be gained.

### *B. Citeulike*

CiteUlike (www.citeulike.org) is a web-based social bookmarking services and traditional bibliographic management tools. It assists researchers and academics in storing, organizing, sharing and discovering links to academic research papers. Like many successful software tools, CiteUlike has a flexible filing system based on the tags. It has been available as a free web service since November 2004. As of September 2011, there are approximately 5,549,945 articles on CiteULike. Their metadata, abstracts, and links to the papers at the publishers' websites. Users can also add reading priorities, personal comments, and tags to their papers. CiteULike also offers the possibility of users setting up and joining groups that connect users sharing academic or topical interests. These group pages report on recent activity. The full text of articles is not accessible from CiteULike, although links to online articles can be added.

Toine Bogers [10] divide a type of metadata from the CiteULike website into five types. First is Topic-related metadata: including all metadata descriptive of the article's topic. Second is Person-related metadata: such as the authors of the article. Third is Temporal metadata: such as the year. Fourth is Miscellaneous metadata: such as the article type. Fifth is User-specific metadata: including the tags assigned by each user, comments by users on an article, and reading priorities.

As CiteULike offers the possibility of users setting up groups that connect users that share similar academic and topical interests for each group we collected the group name, a short textual description, and a list of its members.

Many previous works related to research paper searching focus on improving the efficiency of academic web resource searching. Researchers who studied in research paper searching such as CiteULike: Jomsri, Sanguansintukul, and Choochaiwattana [6], [7] create three heuristic indexers: "tag"(T), "title, abstract"(TA) , "tag, title and abstract"(TTA) and compare with CiteULike. Experiment found that TTA is the best indexer. Furthermore they create a new algorithm for ranking method that is a combination of similarity ranking with paper posted time or *CSTRank* [5]. Capocci and Caldarelli [8] analyzed the small-world properties of the CiteULike folksonomy and the other researcher are [10], [11], [9], and [14].

This paper uses different views to re-ranking search results of research paper bookmarking with focus on the diversity and reliability. This paper extends the method of association rule that is data mining technique to re-ranking search results.

## III. Framework for Community Group of Research Paper Bookmarking

A framework for community group of research paper bookmarking is follows in Fig 1. General community of users who interesting in research paper bookmarking will post papers that they interest to server system of social bookmarking system such as CiteULike. This technique can provide paper with other users for search paper. The framework mechanism was designed in four steps:

1) Historical data of each user groups: After process of user share all their public entries with user community and comment on other papers. Java programming is used to implement a crawler on the research documents. The crawler collects data from research paper bookmarking. The collected documents consist of research papers and each record in the paper corpus contains: article ID, article name, abstract, tag of each paper, link for viewing full text article, groups name, along with group are interest the same paper, book title that published paper, posted date, posted time, paper priority ,and etc.
2) Association rule: This step is preparing and cleaning data for creating association rule model. The relation during users group that interested in the same paper was analyzed. This technique is recommending base on similarity and were describe in section IV.
3) Search Function: Cosine similarity is a similarity measurement between two vectors of *n* dimensions. This involves finding the cosine of the angle between two vectors. This measurement is often used to compare documents in text mining.
4) Re-ranking search result: this step is effect after similarity measurement for improve search result. The ranking of search results are rearranged from the highest similarity score to the lowest similarity score.

## IV. Association Rule Analysis

Association rule discovery is a popular data mining method and well researched method for discovering interesting relations between variables in large databases. Many the research lead data mining and association rule for analyzes and increase efficiency in searching result [18], [19], [20], [21], [22].

This paper analyzed basic data relation by using association rule discovery from personalized function for explore pattern to improve ranking. Researcher explored association of a set title name of paper and set groups of user. We expect that the article were posted more than one group will should significant for create ranking. The data set has over 64,320 rows. Each row of the data set represents a user group that papers were appearing. There for, a single paper can have multiple rows in the data set.

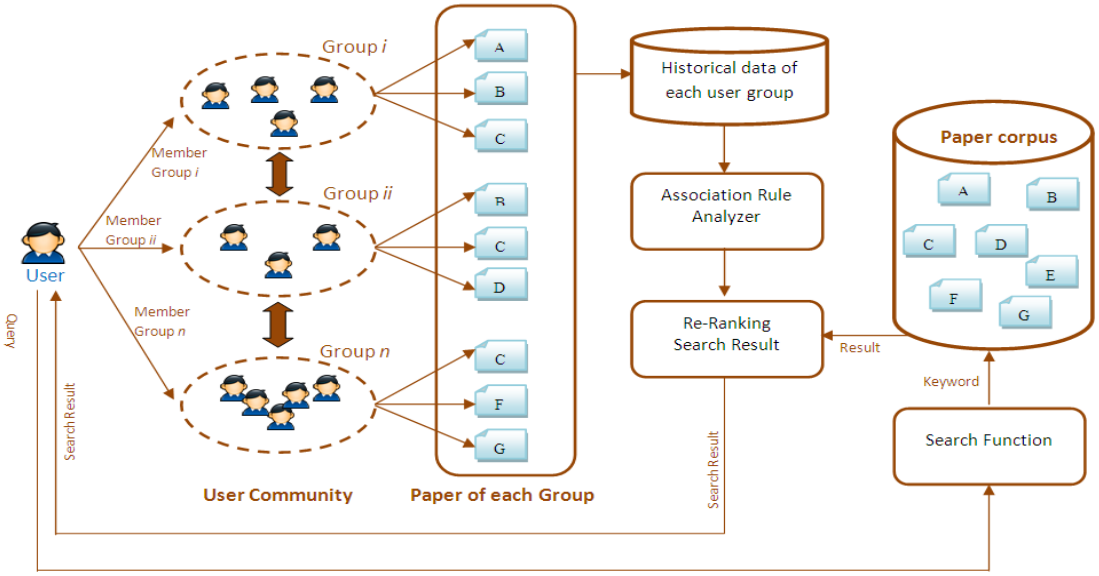



Fig. 1 A Framework for community group of research paper bookmarking

*A. Association rule discovery*

The rules tab in the form of $X \rightarrow Y$ is applied for extracting rules. Where X and Y are disjoint item sets of user group. For each rule of the form $X \rightarrow Y$, researcher defines the *supp* and *conf* as the *support* and *confidence* as follows.

$$conf(X,Y) = \frac{count(X,Y)}{count(X)} \quad (1)$$

such as *conf* (*group X*, *group Y*)

$$conf(group\ X, group\ Y) = \frac{count(group\ X\ , group\ Y)}{count(group\ X\ )} \quad (2)$$

$$sup(X,Y) = \frac{count(X,Y)}{count(All)} \quad (3)$$

such as *sup*(*group X*, *group Y*)

$$sup(group\ X, group\ Y) = \frac{count(group\ X, group\ Y)}{count(All)} \quad (4)$$

TABLE I
EXAMPLES OF RELATION MODELS OF GROUP THAT USER POST WITH CONFIDENCE AND SUPPORT VALUES.

| *Rule* | *Conf (%)* | *Sup (%)* |
|---|---|---|
| microRNA→Bioinformatics | 72.46 | 3.08 |

Table I shows examples of rules for predicting the group. Confidence and support value are used for rule selections. Because plenty of rules are generated, some simple concerns in rule selections include:

1) Select the rule with maximum confidence.
2) Select the rule with maximum support if confidence value is equal.
3) Select the rule that happens first when confidence and support values are equal.

From table I, shows the rule explains:

- Support of $X \rightarrow Y$ is the probability that a paper has both *X* group and *Y* group

Confidence of $X \rightarrow Y$ is probability that a paper appear in *Y* group given that the paper appear in *X*

*B. Result of Association rule discovery*

Table II shows total association prediction model for group of users with confidence and support values.

TABLE II
RELATION MODELS OF GROUP THAT USER POST WITH CONFIDENCE AND SUPPORT VALUES.

| *Rule* | *Conf (%)* | *Sup (%)* |
|---|---|---|
| Genetics-of-Gambling→G4ID | 53.95 | 8.39 |
| G4ID→Genetics-of-Gambling | 100 | 8.39 |
| Philosophy_of_informatic→Blog_and_WikiResearch | 82.46 | 5.16 |
| Blog_and_WikiResearch→Philosophy_of_informatic | 40.22 | 5.16 |
| Statistics and Social Science→Biostatistics | 86.09 | 3.40 |
| Biostatistics→Statistics and Social Science | 88.89 | 3.40 |
| microRNA→Bioinformatics | 72.46 | 3.08 |
| Bioinformatics→microRNA | 17.48 | 3.08 |
| mgh_lcs→Blog_and_WikiResearch | 90.30 | 2.13 |
| Blog_and_WikiResearch→mgh_lcs | 16.60 | 2.13 |
| ReadingLab→Clinical_Psychology | 45.63 | 2.12 |
| Clinical_Psychology→ReadingLab | 85.94 | 2.12 |
| Social navigation→Adaptive-Web | 69.83 | 1.63 |
| Adaptive-Web→Social navigation | 53.76 | 1.63 |
| Automatic sumarization→ASR | 86.64 | 1.31 |
| ASR→Automatic sumarization | 50.50 | 1.31 |
| NLP→ASR | 83.10 | 1.16 |
| ASR→NLP | 44.47 | 1.16 |

## V. Experimental Setting

The experimental setting is divided into two sections. Section *A)* describes the data set, section *B)* discusses describes evaluation metrics.

### *A. The data set*

The crawler collected data from CiteULike during March to May 2010. The collected documents consist of 64,320 research papers. There are groups that are related to the computer science field. Each record in the paper corpus contains: title ID, title name, abstract, tag of each paper, and link for viewing full text article, book title within which the paper was published, posted date, posted time ,paper priority and the along with group.

### *B. Evaluation Matrix*

The informal was conducted with twenty students that were recruited as experiment participants. In the step of measuring the system accuracy, we need to use information retrieval classification metrics, which evaluate the capability of the system to suggest a short list of interesting items to the user. The precision and recall are the standard measurement for the probability that the system makes a correct or incorrect decision about the user interest. With $r_x$ being the research paper from randomly picked for user *u* and *D(u,r)* is the set of recommended research papers, recall and precision are defined as Equation (5)and (6):

$$recall = (D(u,r)) = \frac{1}{|U|}\sum_{u\in U}\frac{|rp(u,r)\cap D(u,r)|}{|rp(u,r)|} \quad (5)$$

$$precision = (D(u,r)) = \frac{1}{|U|}\sum_{u\in U}\frac{|rp(u,r)\cap D(u,r)|}{|D(u,r)|} \quad (6)$$

Where

$|rp(u,r)|$ is the number of is relevant documents,

$|D(u,r)|$ is the number of retrieved documents.

$|rp(u,p)\cap D(u,p)|$ is the number of relevant documents from the number of retrieved documents.

Recall measures the percentage of interesting items suggested to the users, with respect to the total number of interesting items. Whereas, precision measures the percentage of interesting items suggested to the users, with respect to the total number of suggested items. The values precision and recall are shown in section VI. The twenty subjects were considered as experts in the field participated in the experiment. Therefore, their relevancy ratings are assumed to be perfect. In the study setting, each subject is assigned to investigate the research papers obtained from the $r_x$**.** The 10 documents for relevancy are displayed. Finally, the subjects were asked to rate the relevancy of the search results on a two-point scale: score 0 is not relevant at all and score 1 is relevant.

## VI. Result and Discussion

This section separate in to two parts: first is results from the experiment and the second is the discussion.

### *A. Results*

The results of the paper were described in two section first is result of association rule and second is result of evaluation by using precision and recall

1) Result of association rule

Form table III, We choose the rule that have confidence value more than 60%. The strength rules were hold such as Social navigation with Adaptive-Web has Confidence 69.83%. Article which appears in Social navigation will appear in Adaptive-Web always. Therefore, the relationship of these rule may help to created ranking for optimize search results to user. However, Adaptive-Web with Social navigation has Confidence 53.76%. So article which appears in Adaptive-Web group will not appear in Social navigation always. Therefore, the relationship of these rule may not help to created ranking for optimize search results.

TABLE III
Confidence of Association Rule where $\alpha$ = 60%

| *Rule* | *Conf (%)* | *Rule Hold* |
|---|---|---|
| Genetics-of-Gambling→G4ID | 53.95 | No |
| G4ID→Genetics-of-Gambling | 100 | Yes |
| Philosophy_of_informatic→Blog_and_WikiResearch | 82.46 | Yes |
| Blog_and_WikiResearch→Philosophy_of_informatic | 40.22 | No |
| Statistics and Social Science→Biostatistics | 86.09 | Yes |
| Biostatistics→Statistics and Social Science | 88.89 | Yes |
| microRNA→Bioinformatics | 72.46 | Yes |
| Bioinformatics→microRNA | 17.48 | No |
| mgh_lcs→Blog_and_WikiResearch | 90.30 | Yes |
| Blog_and_WikiResearch→mgh_lcs | 16.60 | No |
| ReadingLab→Clinical_Psychology | 45.63 | No |
| Clinical_Psychology→ReadingLab | 85.94 | Yes |
| Social navigation→Adaptive-Web | 69.83 | Yes |
| Adaptive-Web→Social navigation | 53.76 | No |
| Automatic sumarization→ASR | 86.64 | Yes |
| ASR→Automatic sumarization | 50.50 | No |
| NLP→ASR | 83.10 | Yes |
| ASR→NLP | 44.47 | No |

In addition, we use link analysis to show the relation of group's users interested in the same paper. Fig.2 shows example of similarity measurement from Adaptive-Web – Various kinds of user adaptive web system: hypermedia, IR, filtering — by using Link Analysis. We found that some articles were appear in Adaptive-Web will appear in ARTFL group, NET8-UAM group, Social Web group, Social Navigation group, and Philosophy of Information group. Form result of the similarity we can develop this model into paper recommendation mechanism.

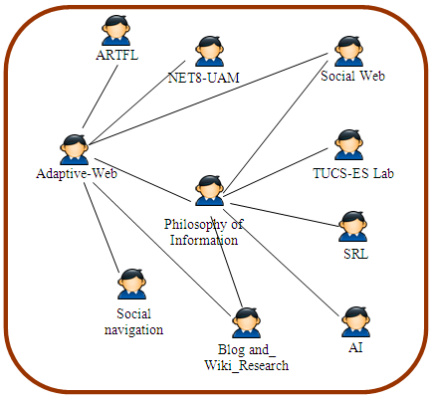



Fig. 2 The result of link analysis

2) Result of evaluation Matrix

Since the subject relevancy ratings, some users only rating one in a group and some users even did not rating any research paper in most groups. The experiment result is depicted in table IV. We use two different correct sets in the experiment. The first is correcting set of original method (not include association rule technique) and the second considers set by include association rule technique before re-ranking method (Our Method). The result is listed in the second and third column.

TABLE IV
PERFORMANCE OF PROPOSED FRAMEWORK

| *Average Precision* | *Original Method* | *Our Method* |
|---|---|---|
| P@1 | 55.0% | 73.2% |
| P@2 | 43.1% | 65.7% |
| P@3 | 40.2% | 63.2% |
| P@4 | 38.7% | 61.1% |
| P@5 | 35.2% | 55.5% |
| P@10 | 27.0% | 49.0% |
| P@15 | 25.5% | 44.1% |

### *B. Discussion*

This paper presents techniques ranking search result for users based on the relation of user group. In the association rule step, the *support* and *confidence value* were used to determine the groups relation. The performance of system by include association rule technique tag based filtering recommendation has accuracy more than the original system**.** Therefore, the relation of user group has a potential and can use this technique for improve ranking search result.

## VII. CONCLUSION AND FUTURE WORK

In this study, a related of community group of research paper bookmarking framework is proposed. This approach studies users' behavior from research paper bookmarking and then use association rule to analysis user's preference and can bring to improve ranking. The experiment has shown some interesting results and it is believed the research direction is promising.

In fact, during our study, it is becoming clear that only relying on one method to predict the preference. Furthermore, in this paper the ranking mechanism only considers one-to-one association rule like *group X→group Y*. This assumption is to simplify the problem.

In addition, This paper preliminary analysis a relation of the group uses that appear the same article by using association rule discovery .Result of preliminary analysis of there some rule is interesting and can bring to improve performance ranking.

In future, researcher plan to use this information to advance analysis for improve web searching.

## ACKNOWLEDGMENT

The authors would like to thank Suan Sunadha Rajabhat University for scholarship support. The study is not possible without the data from CiteULike.

## REFERENCES

[1] CiteULike, http://www.CiteULike.org
[2] Connotea, http://www.connotea.org
[3] BibSonomy, http://www.bibsonomy.org
[4] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, A. Aghasaryan, and C. Bernier ,"Analysis of Strategies for Building Group Profiles,"in *User Modeling, Adaptation, and Personalization 2010*, Lecture Notes in Computer Science, 2010, Volume 6075/2010, pp 40-51.
[5] P. jomsri, A Combination of Similarity Ranking and Time for Social Research Paper Searching, World Academy of Science, Engineering and Technology 78 2011,pp. 638-643
[6] P.Jomsri, S. Sanguansintukul, W. Choochaiwattana, "Improve Research paper Searching with social tagging-A Preliminary Investigation," in *the Eight International Symposium on Natural Language Processing*, Thailand, 2009,pp.152-156.
[7] P.Jomsri, S. Sanguansintukul, W. Choochaiwattana, "A Comparison of Search Engine Using "Tag Title and Abstract" with CiteULike – An Initial Evaluation," *in the 4th IEEE Int. Conf. for Internet Technology and Secured Transactions (ICITST-2009),*United Kingdom,2009.
[8] A. Capocci, and G.Caldarelli, "Folksonomies and Clustering in the Collaborative System CiteULike," *arXiv Press, eprint No. 0710.2835*, 2007.
[9] U. Farooq, T.G. Kannampallil, Y. Song, C.H. Ganoe, M.C., John, L. Giles, "Evalating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics," in *Proc. of the 2007 international ACM conference on Supporting group work (GROUP'07)*, Sanibel Island, Florida, USA, 2007,pp.351-360.
[10] T. Bogers, and A. van den Bosch, "Recommending Scientific Articles Using CiteULike," in *Proc. of the 2008 ACM conference on Recommender systems(RecSys'08)*, Switzerland,2008 ,pp.287-290.
[11] E. Santos-Neto, M. Ripeanu, and A. Iamnitchi, " Tracking usage in collaborative tagging communities".
[12] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008, pp. 531–538.
[13] C. Schmitz, A. Hotho, R. J\lschke, and G. S. and,"Mining association rules in folksonomies," in DataScience and Classification, ser. Studies in Classification,Data Analysis, and Knowledge Organization. SpringerBerlin Heidelberg, 2006, pp. 261–270. [Online].Available:http://www.springerlink.com/content/gmv832553g0x3673/
[14] U. Farooq, C.H. Ganoe, , J.M. Carroll, and C.L. Giles, "Supporting distributed scientific collaboration: Implications fordesigning the CiteSeer collaborator," in *IEEE Proc. of the Hawaii Int'l Conference on System Sciences*, Waikoloa, Hawaii,2007.

[15] R. Cohen, and S. Havlin, “Scale-free network are ultrasmall Physica”, A311, p590
[16] Peng Zhang, Zengru Di, Complex System and Complexity Science, 2(3), pp.30-34
[17] W. Hong, W. Wei-dong , X. Na, H. Wen, “Group Personnel Relationship Analysis Based on Social Networks”, in *IEEE International Symposium on IT in Medicine & Education, 2009*. (ITIME '09),pp.1003 – 1008
[18] X. Chen, and Y. Wu. Personalized Knowledge Discovery: Mining Novel Association Rules from Tex t. Available: http://www.siam.org/meetings/sdm06/proceedings/067chenx.pdf
[19] C. Schmitz, A. Hotho, R. Jäschke, and G. Stumme. (2008, Oct). Mining Association rule in Folksonomies. *Journal of Information Science (JIS)* [Online].Available:*http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.93.9741&rep=rep1&type=pdf*.
[20] C. Haruechaiyasak, M. Shyu, and S. Chen, “A Data mining Framework for Building A Web-Page Recommender System”, Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration, IRI - 2004, November 8-10, 2004, Las Vegas Hilton, Las Vegas, NV, USA. pp. 357-362
[21] R. Forsati, M.R. Meybodi, A. Ghari Neiat, “Web Page Personalization based on Weighted Association Rules”, *International Conference on Electronic Computer technology 2009*, pp. 130-135
[22] S. niwa , T. Doi, and S. Honiden, “ Web Page Recommender System based on Folksonomy Mining for ITNG’06 Submissions”, *Proceedings of the Third International Conference on Information Technology:New Generations (ITNG’06)*
[23] Hui Chang, Daren He, *Science and Technology Review* , 24(9),pp. 84-87

# INVITATION LETTER

July 24, 2012

Ms. Pijitra Jomsri
Suan Sunandha Rajabhat University
Thailand

To Whom It May Concern,

This invitation letter is to confirm that your peer-reviewed & refereed full paper entitled "Improved Image-retrieval Performance Base on a Combination of Social Tagging and Image Descriptions for Create Indexing" is accepted for oral presentation at the ICCESSE 2012 : International Conference on Computer, Electrical, and Systems Sciences, and Engineering to be held in Singapore, SG during September 12-13, 2012.

This invitation letter serves as confirmation of your conference attendance.

Sincerely Yours,


Conference Council
ICCESSE 2012 Singapore
SG

Conference Venue
River View Hotel Singapore
382 Havelock Road, Singapore 169629
Tel : +65-6732 9922
Fax : +65-6732 1034
web: www.riverview.com.sg

# ประวัติผู้เขียน

นางสาวพิจิตรา จอมศรี

**วันเกิด:** 29 มกราคม พ.ศ. 2524
**หมายเลขบัตรประชาชน:** 3-10020-2196-79-1
**ที่อยู่** 283/16 ซ.ศรีเผื่อน ถ.ริมคลองประฝั่งซ้าย บางซื่อ กทม. 10800
**E-mail:**pijitra_jom@hotmail.com
**โทรศัพท์มือถือ:** 081-8662946 **โทรศัพท์บ้าน** 02-910-4642

**ประวัติการศึกษา**

- วิทยาการคอมพิวเตอร์ (วท.ม.) คณะวิทยาศาสตร์
  สาขาวิทยาการคอมพิวเตอร์
  มหาวิทยาลัยศิลปากร สำเร็จการศึกษา 29 มี.ค. 2550
- วิทยาศาสตรบัณฑิต (วท.บ.) คณะวิทยาศาสตร์และเทคโนโลยี
  สาขาสถิติ
  มหาวิทยาลัยธรรมศาสตร์ สำเร็จการศึกษา 24 ก.พ. 2545

**ประสบการณ์ทำงาน**

2551-ปัจจุบัน อาจารย์ประจำ สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา

2552 อาจารย์พิเศษ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏนครปฐม

2551 วิทยากรอบรมทักษะมาตรฐานการใช้เทคโนโลยีสารสนเทศและการสื่อสาร ศูนย์เทคโนโลยีสารสนเทศสำนักวิทยบริการและเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏสวนสุนันทา

2546-2547 อาจารย์พิเศษ ศูนย์การศึกษาโรงเรียนเทคนิคพาณิชการธนบุรี มหาวิทยาลัยราชภัฏสวนสุนันทา

**ประสบการณ์สอนระดับอุดมศึกษา วิชา :**

1. การสื่อสารข้อมูลและเครือข่าย
2. ความปลอดภัยของระบบสารสนเทศ
3. ข่ายงานบริเวณเฉพาะที่และข่ายงานแฟ้มบริการ
5. การออกแบบและพัฒนายูสเซอร์อินเตอร์เฟส
6. ปฏิสัมพันธ์ระหว่างมนุษย์กับคอมพิวเตอร์
7. คณิตศาสตร์ดิสครีตสำหรับเทคโนโลยีสารสนเทศ
8. พาณิชอิเล็กทรอนิกส์
9. คลังและเหมืองข้อมูล
10. ประกันคุณภาพซอฟต์แวร์

**ผลงานวิจัยและบทความที่ได้รับการตีพิมพ์ระดับนานาชาติและในประเทศ**

**ระดับนานาชาติ**

2553 P. Jomsri, S. Sanguansintukul**,** and W. Choochaiwattana, "A Framework for Tag-Based Research Paper Recommender System: An IR Approach", 24th IEEE International Conference on Advanced Information Networking and Applications (AINA 2010), Perth, Australia, 20-23 April 2010.

2553 P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "A Comparison of Search Engine Using "Tag Title and Abstract" with CiteULike – An Initial Evaluation", The 4th International Conference for Internet Technology and Secured Transactions (ICITST-2009) , *London, UK, November 9-12, 2009*.

2553 P. Jomsri, S. Sanguansintukul, and W. Choochaiwattana, "Improve Research Paper Searching with Social Tagging – A Preliminary Investigation", The Eight International Symposium on natural Language Processing SNLP2009), *October, Bangkok, 20 - 22, 2009***.**

2554 P. Jomsri, "A Combination of Similarity Ranking and Time for Social Research Paper Searching", ICCIT 2011 : International Conference on Computer and Information Technology, *July , Amsterdam, 13-15 ,2011.*

2554 P. Jomsri, S. Sanguansintukul**,** and W. Choochaiwattana, "A Combination Ranking Model for Research Paper Social Bookmarking Systems", AMT2011, 2011 International Conferences on Active Media Technology (AMT 2011), *September,7-9 ,2011*

**ในประเทศ**

2549 พิจิตรา จอมศรี และปานใจธารทัศนวงศ์, "การเพิ่มอัตราการพบในระบบพร๊อกซี่โดยใช้เทคนิคเหมืองข้อมูล", การประชุมวิชาการทางด้านเทคโนโลยีสารสนเทศ NCIT2006/1/2549.

**ประสบการณ์ทำงานพิเศษด้านการฝึกอบรม**

2553 วิทยากรบรรยาย การสร้างและปรับปรุงสินค้าเพื่อนำไปจัดทำร้านค้าออนไลน์ มหาวิทยาลัยราชภัฏสวนสุนันทา

2553 วิทยากรบรรยายอบรมความรู้จัดทำเว็บไซต์ ของดีชุมชน มหาวิทยาลัยราชภัฏสวนสุนันทา

2551-2553 ฝึกอบรมนักศึกษา ด้านทักษะมาตรฐานการใช้เทคโนโลยีสารสนเทศและการสื่อสาร มหาวิทยาลัยราชภัฏสวนสุนันทา

2554 -วิทยากรอบรมความรู้ทางคอมพิวเตอร์สำหรับนักเรียน ICT รุ่นเยาว์ มหาวิทยาลัยราชภัฏสวนสุนันทา

-วิทยากรอบรมความรู้ด้านกฎหมายที่เกี่ยวข้องกับการใช้งานคอมพิวเตอร์ มหาวิทยาลัยราชภัฏสวนสุนันทา