**ระบบแนะนำผลลัพธ์การค้นหาโดยใช้กราฟคำสำคัญสำหรับฐานข้อมูลวิทยานิพนธ์สำนักหอสมุดมหาวิทยาลัยเชียงใหม่**

**Search Result Recommendation System based on Keyword Graph for Thesis Database, Chiang Mai University Library**

**กวินวิชญ์ บุบผาชื่น  
560510598**

**การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของกระบวนวิชา 204491  
สาขาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่  
ปีการศึกษา 2559**

**ระบบแนะนำผลลัพธ์การค้นหาโดยใช้กราฟคำสำคัญสำหรับฐานข้อมูลวิทยานิพนธ์สำนักหอสมุดมหาวิทยาลัยเชียงใหม่**

**Search Result Recommendation System based on Keyword Graph for Thesis Database, Chiang Mai University Library**

กวินวิชญ์ บุบผาชื่น  
560510598

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของกระบวนวิชา 204491  
สาขาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่  
ปีการศึกษา 2559

คณะกรรมการสอบการค้นคว้าอิสระ

.................................................................. ประธานกรรมการ

อาจารย์ ดร.รัศมีทิพย์ วิตา

......................................................................กรรมการ  
อาจารย์ ดร.วัฒนา จินดาหลวง  
วันที่...........เดือน....................................พ.ศ.................

**บทที่ 1**

**บทนำ**

ในบทนี้จะกล่าวถึงภาพรวมของระบบแนะนำผลลัพธ์การค้นหาโดยใช้กราฟคำสำคัญสำหรับฐานข้อมูลวิทยานิพนธ์สำนักหอสมุดมหาวิทยาเชียงใหม่ ในส่วนของหลักการและเหตุผล วัตถุประสงค์ของโครงการ ขอบเขตการพัฒนา ประโยชน์ที่คาดว่าจะได้รับจากการศึกษาค้นคว้าและแผนการดำเนินงานและระยะเวลาดำเนินงาน

**1.1 หลักการและเหตุผล**

ในปัจจุบันการค้นหาเอกสารวิทยานิพนธ์ของสำนักหอสมุด มหาวิทยาลัยเชียงใหม่ยังใช้เป็นระบบค้นหาตาม ชื่อเรื่อง ผู้แต่ง หัวเรื่อง และ ปี ของงานวิทยานิพนธ์ซึ่งเป็นข้อจำกัดในการค้นข้อมูลงานวิทยานิพนธ์ที่ยังไม่มีประสิทธิภาพเท่าที่ควรเนื่องจากจะได้ข้อมูลด้านเดียว หรือ ได้ข้อมูลที่ไม่เพียงพอดังนั้นการเพิ่มประสิทธิภาพการค้นหาข้อมูลเชิงความหมายในฐานข้อมูลงานวิทยานิพนธ์ในการค้นคว้าอิสระนี้ ได้นำเสนอวิธีการค้นหาและเรียงลำดับผลลัพธ์ตามการปรากฏของคำสำคัญในเอกสารโดยการค้นหาที่นำเสนอเป็นการค้นหาผ่านคลังคำศัพท์ในรูปแบบกราฟสองส่วนที่ได้จัดทำขึ้นจากงานวิจัยของปิยบุตร ปัญญาดี โดยในการค้นคว้าอิสระฉบับนี้ได้เพิ่มเติมความครอบคลุมของคำศัพท์โดยใช้คำที่มีความหมายใกล้เคียงจากคลังคำศัพท์เวิร์ดเน็ตเพื่อขยายขอบเขตของการค้นหาและแนะนำเอกสารที่เหมาะสม

**1.2 วัตถุประสงค์ของโครงการ**

1) เพื่อนำเสนอวิธีการเรียงลำดับผลลัพธ์ของการค้นหาจากความสัมพันธ์ของกลุ่มคำสำคัญ

2) แนะนำเอกสารที่มีเนื้อหาใกล้เคียงกับเอกสารที่ผู้ใช้งานเลือก

**1.3 ประโยชน์ที่จะได้รับจากการศึกษาเชิงประยุกต์**

1) สามารถเรียงลำดับผลลัพธ์การค้นหาเอกสารวิทยานิพนธ์ตามความสัมพันธ์ของคำค้นกับเอกสาร

2) สามารถแนะนำเอกสารโดยใช้ความใกล้เคียงของกลุ่มคำศัพท์

3) ประยุกต์ทฤษฏีของกราฟสองส่วนในแอปพลิเคชันการค้นหาได้

**1.4 ขอบเขตของโครงงาน**

**1.4.1 ขอบเขตทางสถาปัตยกรรม**

1) ระบบที่ออกแบบ

เว็บเซอร์วิสเพื่อใช้ในการทำงานร่วมกับระบบประมวลผลฐานข้อมูลงานวิทยานิพนธ์ของสำนักหอสมุด

2) ฮาร์ดแวร์ (Hardware) ที่ใช้ในการพัฒนาระบบ ประกอบด้วย

หน่วยประมวลผล (CPU) อินเทลคอร์ ไอ 7-33เอ็มคิว 2.40 กิกะเฮิรตซ์ (Intel Core i7-4700MQ 2.40 GHz) ฮาร์ดดิสก์ (Hard disk) ขนาดความจุ 750กิกะไบต์ (750 GB) หน่วยความจำหลัก ดีดีอาร์สาม 4 กิกะไบต์ (DDRIII 4GB)

3) ซอฟแวร์ (Software) ที่ใช้ในการพัฒนาระบบ ประกอบด้วย

- ระบบปฏิบัติการไมโครซอฟท์วินโดว์สิบ (Microsoft Windows 10)

- โปรแกรมไพทอนเวอร์ชัน 3.5 (Python 3.5)

4) ภาษา ที่ใช้ในการพัฒนาระบบ

- ภาษาไพทอน (Python)

-ระบบฐานข้อมูลเชิงกราฟโอเรียนดีบี (OrientDB)

**1.4.2 ขอบเขตของระบบงาน**

1) พัฒนาระบบเว็บเซอร์วิสเพื่อใช้สืบค้นข้อมูลเอกสารวิทยานิพนธ์ของสำนักหอสมุด

มหาวิทยาลัยเชียงใหม่ ผ่านฐานข้อมูลคำสำคัญในรูปแบบกราฟ

**1.4.3 ขอบเขตของข้อมูล**

1) ข้อมูลเข้า

-ข้อความค้นหาข้อมูลงานวิทยานิพนธ์จากระบบการสืบค้นเอกสารวิทยานิพนธ์สำนักหอสมุด

2) ข้อมูลออก

-ผลลัพธ์การค้นหาข้อมูลงานวิทยานิพนธ์ในรูปแบบแนะนำข้อมูลที่เกี่ยวข้อง

**1.4.4) ขอบเขตการศึกษา**

1) ศึกษารูปแบบการทำงานของการค้นหาข้อมูลงานวิทยานิพนธ์สำนักหอสมุดมหาวิทยาลัยเชียงใหม่

2) ศึกษาวิธีการเลือกคำที่มีความหมายคล้ายกันด้วยเวิร์ดเน็ต **(WordNet)**

3) ศึกษาวิธีการ สืบค้นข้อมูลโดยอาศัยความสัมพันธ์ในเชิงความหมายของผลลัพธ์

**1.5 แผนการดำเนินงานและระยะเวลาดำเนินงาน**

เริ่มดำเนินการตั้งแต่ เดือน สิงหาคม พ.ศ. 2559 สิ้นสุดเดือน ธันวาคม พ.ศ. 2559 แสดงรายละเอียดการดำเนินงานในตาราง ที่ 1.1 โดยมีขั้นตอนการดำเนินงานดังนี้

1) ศึกษาทฤษฏีและเอกสารที่เกี่ยวข้อง

- ศึกษาทฤษฏีของการสืบค้นเอกสารและสืบค้นในเชิงความหมาย

- ศึกษาลักษณะของระบบการค้นหาข้อมูลงานวิทยานิพนธ์สำนักหอสมุดมหาวิทยาลัยเชียงใหม่

- ศึกษารูปแบบลักษณะของระบบ แนะนำผลลัพธ์ และ การค้นหาข้อมูลเชิงความหมายในฐานข้อมูลงานวิทยานิพนธ์

- ศึกษาการเชื่อมต่อและการสืบค้นข้อมูลจากฐานข้อมูลเชิงกราฟ

2) ออกแบบการทดสอบระบบ

3) ออกแบบระบบและจัดทำระบบ

4) ทำการทดสอบ วิเคราะห์ และ ปรับแต่งระบบให้ถูกต้องตามการความต้องการของระบบ

5) จัดทำเอกสาร

**ตาราง 1.1** แสดงระยะเวลาของการดำเนินงานทั้ง 5 ขั้นตอน

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ระยะเวลาปี พ.ศ. 2559**  **ขั้นตอนการดำเนินงาน** | ส.ค. | ก.ย. | ต.ค. | พ.ย. | ธ.ค. |
| 1. ศึกษาทฤษฏีและเอกสารที่เกี่ยวข้อง |  |  |  |  |  |
| 2. ออกแบบการทดสอบระบบ |  |  |  |  |  |
| 3.ออกแบบระบบและจัดทำระบบ |  |  |  |  |  |
| 4. ทำการทดสอบ วิเคราะห์ และ ปรับแต่งระบบให้ถูกต้องตามการความต้องการของระบบ |  |  |  |  |  |
| 5. จัดทำเอกสาร |  |  |  |  |  |

**บทที่ 2**

**ทฤษฎีและหลักการที่เกี่ยวข้อง**

การค้นหางานข้อมูลงานวิทยานิพนธ์ของสำนักหอสมุดมหาวิทยาลัยเชียงใหม่นั้นข้อมูลที่มีอยู่มีขนาดที่ใหญ่การเลือกแนวทางการค้นหาจึงเลือกฐานข้อมูลที่เหมาะสมกับข้อมูลและมีความสามารถที่จะจัดการกับข้อมูลในรูปแบบของกราฟได้อย่างมีประสิทธิภาพ จึงศึกษาทฤษฎีที่เกี่ยวข้องเกี่ยวกับกราฟที่นำมาใช้คือกราฟสองส่วน (Bipartite graph) ที่นำมาแก้ปัญหาเกี่ยวกับการจับคู่คำค้นหา และฐานข้อมูลโอเรียนดีบีซึ่งเป็นโนเอสคิวแอล ที่สร้างฐานข้อมูลที่มีความสัมพันธ์กันในรูปแบบของกราฟได้

**2.1. กราฟสองส่วน (Bipartite graph)**

เป็นกราฟที่เซตจุดยอดสามารถแบ่งได้เป็น 2 เซตที่ไม่มีส่วนร่วมกัน และ จุด 2 จุดใดๆในเซตเดียวกันจะไม่มีเส้นเชื่อม ระหว่างกัน เช่น กราฟ V = {a, b ,c ,d ,e}

a**. .** d

b**.**

c**. .** e

รูปที่ 2.1 ตัวอย่างแสดงกราฟสองส่วน

แบ่งเป็น 2 ส่วนได้ดังนี้

V1 = {a , b , c} , V2= {d , e } โดยที่ V1 https://web.ku.ac.th/schoolnet/snet2/knowledge_math/set/symbol9.gif V2 = Ø และ V1 https://web.ku.ac.th/schoolnet/snet2/knowledge_math/set/symbol8.gif V2 = v

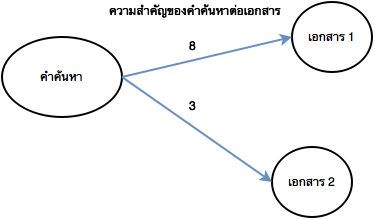
กราฟสองส่วนมีประโยชน์โดยการนำมาแก้ไขปัญหา เช่น การจับคู่ การค้นหา ในส่วนของการค้นหาข้อมูลงานวิทยานิพนธ์ของสำนักหอสมุดมหาวิทยาลัยเชียงใหม่นั้น จะทำการแบ่งฐานข้อมูลออกเป็น 2 ส่วนในระหว่างส่วนของสำคัญที่คัดกรองจากเนื้อหาของเอกสารจากระบบ และ ข้อมูลเอกสารงานวิทยานิพนธ์ที่เก็บในฐานข้อมูลหลักในรูปแบบของกราฟ และใช้หลักการของการ จับคู่ของกราฟสองส่วนในการแก้ปัญหาซึ่งจะให้ผลลัพธ์ที่ตรงตามความต้องการได้มากที่สุดซึ่งเรียงลำดับตามความสำคัญของข้อมูล [1]

**2.2. ฐานข้อมูลเชิงกราฟ (Graph Database)**

เป็นฐานข้อมูลชนิดหนึ่งที่มีลักษณะการเชื่อมต่อกันแบบกราฟถูกนำมาใช้กันมากขึ้น ในด้านโซเชียลเน็ตเวิร์คหรือหากต้องการที่จะเก็บข้อมูลที่มีขนาดใหญ่มากๆ การเชื่อมต่อข้อมูลจะ มีลักษณะ เป็นกลุ่มโหนด เชื่อมต่อกันเป็นจำนวนมากฐานข้อมูลเชิงกราฟสร้างขึ้นโดยอาศัยพื้นฐานความสัมพันธ์จากทฤษฎีเชิงกราฟซึ่งประกอบไปด้วย   
 -Nodes แทนที่ด้วยเอนทิตี้ เช่น จากตัวอย่างคือคำค้นหาและเอกสารเทียบเท่ากับเรคคอร์ดในฐานข้อมูลแบบ Relational Database

-Edges เป็นเส้นที่บ่งบอกถึงความสัมพันธ์ของแต่ละโหนด เช่น แต่ละโหนดติดต่อกันแบบไหน ซึ่งเป็น key concept ของฐานข้อมูลเชิงกราฟจากตัวอย่างคือ โหนดคำค้นหามีเส้นความสัมพันธ์ต่อโหนดเอกสาร 1 และ เอกสาร 2

-Properties เป็นการอธิบายถึงข้อมูลที่มีความเกี่ยวข้องกันของแต่ละโหนด เช่น ในโหนดคำค้นหามี properties คือ คำว่า Data และ ใน Edge คือ 10 เป็นต้น [2]



รูปที่ 2.2 ตัวอย่างการเก็บข้อมูลเชิงกราฟ

**2.3 โนเอสคิวแอล (NoSQL)**

โนเอสคิวแอลนั้นเป็นแนวทางหนึ่งสำหรับจัดการข้อมูลและการออกแบบฐานข้อมูลสำหรับข้อมูลขนาดใหญ่ ซึ่งอยู่อย่างกระจัดกระจาย หลากหลายรูปแบบ จะถูกนำไปใช้กับข้อมูลที่มีขนาดใหญ่ ข้อมูลไม่มีโครงสร้างที่ชัดเจน หรือ ระบบที่ต้องการการทำงานแบบตลอดเวลา

**2.3.1 คุณสมบัติของฐานข้อมูลโนเอสคิวแอล**

1) Dynamic Schemas ไม่จำเป็นต้องมีตารางที่ตายตัวหรือไม่ต้องมีตาราง ก่อนที่จะจัดเก็บข้อมูล ข้อมูลแต่ละแถว สามารถจัดเก็บได้ตามต้องการ สามารถเพิ่มหรือลด จะไม่ส่งผลต่อระบบ ทำให้สามารถจัดเก็บข้อมูลได้ตามต้องการ

**2)** Auto-Sharding เมื่อข้อมูลมีขนาดใหญ่ หรือเราต้องการเพิ่มประสิทธิภาพการ อ่าน-เขียนข้อมูลปริมาณมากๆ การทำ Sharding ในระบบโนเอสคิวแอลดาต้าเบสจะทำการกระจายข้อมูลไปยัง Server ต่างๆอัตโนมัติ (Auto-Sharding) ผู้พัฒนา (Developer) ไม่ต้องเขียนโปรแกรมในการกระจายข้อมูลเอง เหมือน Relational Database

**3)** Replicationการสำเนาข้อมูลจากเครื่องหนึ่งไปยังอีกเครื่องหนึ่งเมื่อ Server หนึ่งเสียหาย อีกเครื่องหนึ่งจะขึ้นมาทำงานแทนทันที โดยข้อมูลของแต่ละเครื่องจะมีข้อมูลเหมือนกัน ดังนั้น Replication เป็นหนึ่งคุณสมบัติที่ตอบสนองต่อการใช้งานที่ต้องการความต่อเนื่องได้ตลอดเวลา (High Availability)

**4)** Integrated Cachingการจัดเก็บข้อมูลที่ใช้งานบ่อยๆ เข้าไว้ใน Memory (RAM) ซึ่งเป็นคุณสมบัติเด่นของโนเอสคิวแอลที่รวม Caching ไว้ในตัวเองอยู่แล้วไม่จำเป็นต้องทำ Cache Layer เหมือน Relational Database ที่ต้องทำ Cache Layer แยก [3]

**2.3.2 การจัดหมวดหมู่ของโนเอสคิวแอลประกอบไปด้วย 4 ประเภท**

**1)** Column-based databases คือฐานข้อมูลที่เก็บข้อมูลในรูปของตารางคล้ายๆกับฐานข้อมูลเชิงสัมพันธ์แต่จะเก็บข้อมูลในแบบคอลัมน์แทนแบบแถว ซึ่งเหมาะกับการทำเหมืองข้อมูล ตัวอย่างของฐานข้อมูลที่เป็นลักษณะนี้ได้แก่ Hbase, Cassandra, Hypertable เป็นต้น

**2)** Document-based databases คือฐานข้อมูลที่เก็บข้อมูลในรูปแบบกลุ่มของเอกสารซึ่งง่ายต่อการแก้ไขโครงสร้างโดยแต่ละเอกสารจะมีจำนวนฟิลที่แตกต่างกันได้ตัวอย่างของฐานข้อมูลที่เป็นลักษณะนี้ได้แก่ Apache CouchDB , MongoDB เป็นต้น

**3)** Key-Value data Stores คือฐานข้อมูลที่เก็บข้อมูลในรูปแบบ Key และ Value ที่สัมพันธ์กันอยู่ ซึ่ง Value จะถูกค้นเจอได้โดยค่า Key  ตัวอย่างของฐานข้อมูลที่เป็นลักษณะนี้ได้แก่ Redis, Dynomite, Voldemort เป็นต้น

**4)** Graph-based data-stores คือฐานข้อมูลที่เก็บข้อมูลในรูปแบบของกราฟที่มีความสัมพันธ์กันทำให้สามารถใช้ทฤษฏีกราฟได้ในการจัดการกับข้อมูลได้เป็นอย่างดี ระบบที่นำไปใช้เช่น ระบบแนะนำ หรือ ข้อมูลที่เชื่อมโยงกัน ตัวอย่างของฐานข้อมูลที่เป็นลักษณะนี้ได้แก่ nodes. Neo4j เป็นต้น [4]

เนื่องจากระบบแนะนำผลลัพธ์การค้นหาที่จะจัดทำขึ้นนั้นต้องใช้ดาต้นเบสในรูปแบบของกโนเอสคิวแอลซึ่งอยู่ในหมวดหมู่ของฐานข้อมูลเชิงกราฟคือโอเรียนดีบีซึ่งมีคุณสมบัติที่สามารถสร้างฐานข้อมูลในรูปแบบของกราฟและสามารถนำทฤษฏีกราฟมาแก้ปัญหาในการจัดการกับข้อมูลได้

**2.4 โอเรียนดีบี (OrientDB)**

โอเรียนดีบีถูกเขียนขึ้นด้วยภาษาจาวา เป็นฐานข้อมูลแบบผสมซึ่งเป็นการผสมระหว่างความสัมพันธ์และโนเอสคิวแอล สามารถใช้ภาษาเอสคิวแอล ในการบริหารจัดการข้อมูลได้ มีลักษณะการจัดเก็บข้อมูลแบบฐานข้อมูลเชิงกราฟและเป็น Opensource apache 2.0

โอเรียนดีบีมีความสัมพันธ์พื้นฐานระหว่างฐานข้อมูลกับเอกสารจะเป็นการบริหารจัดการฐานข้อมูลในรูปแบบกราฟ จะเป็นการเชื่อมต่อโดยตรงระหว่างกราฟกับเรคคอร์ดนั้นๆ โดยที่เราสามารถเพิ่มไปในทุกๆส่วนหรือในแต่ละส่วนของต้นไม้และกราฟของเรคคอร์ดนั้นและมีการสนับสนุนภาษาเอสคิวแอล ระหว่างการใส่ข้อมูลลงในตารางโอเรียนดีบีนั้นไม่มีความสัมพันธ์กับดีบีเอ็มเอส (DBMS) โดยที่โอเรียนดีบีนั้นเป็นตัวที่ติดอยู่กับการเคลื่อนไหวของโนเอสคิวแอล แม้ว่าจะสนับสนุนส่วนย่อยของการคิวรีของภาษาเอสคิวแอล

การนำโอเรียนดีบีมาใช้เก็บฐานข้อมูลของงานวิทยานิพนธ์ของสำนักหอสมุดมหาวิทยาลัยเชียงใหม่ทำให้สามารถสร้างฐานข้อมูลดาต้าเบสในรูปแบบของกราฟ ซึ่งจะทำให้สามารถใช้หลักของกราฟสองส่วนในการค้นหาข้อมูลงานวิทยานิพนธ์ที่ต้องการจะค้นหาได้ตามจุดประสงค์ [5]

**2.5 ภาษาไพทอนและโอเรียนดีบี (Python and OrientDB)**

การใช้ติดต่อฐานข้อมูลโอเรียนดีบีนั้นสามารถเรียกใช้ได้หลายภาษาซึ่งในงานวิทยานิพนธ์จะใช้ภาษาไพทอน ในการติดต่อฐานข้อมูล โดยทางโอเรียนดีบีจะมีไดเวอร์เฉพาะสำหรับภาษาไพทอนให้อยู่แล้วโดยจะทำงานร่วมกับโอเรียนดีบี version 1.7 ขึ้นไปซึ่งสามารถดาวน์โหลด ได้จาก<https://github.com/mogui/pyorient> [6]

ตัวอย่างการเชื่อมต่อกับโอเรียนดีบีโดยใช้ภาษาไพทอน

client = pyorient.OrientDB(“localhost”,2424)  
session\_id = client.connect(“admin”,”admin”)

**2.6 เวิร์ดเน็ต (WordNet)**

(Bird, Klein, and Loper 2009) ได้กล่าวถึง ฐานความรู้เวิร์ดเน็ตคือฐานข้อมูลคำศัพท์ซึ่งมีการจัดหมวดหมู่ของคำที่มีความหมายเหมือนกัน เข้าไว้ด้วยกันเป็นกลุ่มคำ มีการจัดแบ่งกลุ่มคำตามชนิดของคำศัพท์ คือ คำนาม(Nouns) คำกริยา (Verb) คำวิเศษ (Adverb) และคำคุณศัพท์ (Adjective) เวิร์ดเน็ตมีความสัมพันธ์ที่สำคัญ เช่น ความพ้องกันเชิงความหมายของคำ (Synonymy) โดยที่เวิร์ดเน็ตจะนิยามคำพ้องในลักษณะต่างๆ เป็นชุดคำพ้องแต่ละชุดและจัดให้มีหมายเลขกำกับคำพ้องแต่ละชุด (Synset number) แต่ละ Synset จะแทนลำดับชั้นของแนวคิดเพียงคำศัพท์เดียวและจะเชื่อมโยงไปยัง Synset อื่นในรูปแบบ semantic network ด้วยหมายเลขความสัมพันธ์ ความสัมพันธ์ที่แตกต่างกันจะถูกนิยามว่า Synset เช่น เมื่อหาความหมายของคำว่า car จะได้ชุดของคำที่มีความหมายใกล้เคียงกัน car, auto, automobile, machine, motorcar [7]

**2.7 งานวิจัยที่เกี่ยวข้อง**

**2.7.1 งานวิจัยเรื่อง** การเพิ่มประสิทธิภาพการค้นหาเอกสารในฐานข้อมูลเชิงสัมพันธ์โดยการใช้ดัชนีแบบกราฟ: กรณีศึกษาฐานข้อมูลวิทยานิพนธ์ระดับบัณฑิตศึกษา สำนักหอสมุดกลาง มหาวิทยาลัยเชียงใหม่

จากการศึกษาของปิยบุตร ปัญญาดี ได้นำเสนอเกี่ยวกับการเพิ่มประสิทธิภาพการสืบค้นข้อมูลเอกสารวิทยานิพนธ์โดยการพิจารณาเนื้อความเอกสารเพื่อใช้เป็นการสืบค้นซึ่งจะเน้นไปที่เนื้อความในส่วนบทคัดย่อที่เป็นบทสรุปของเอกสารวิทยานิพนธ์เพื่อสร้างดัชนีคำศัพท์ที่เกี่ยวข้องและมีการนำแนวคิดในการจัดเก็บข้อมูลในฐานข้อมูลเชิงกราฟ และทฤษฏีกราฟสองส่วนมาประยุกต์ในการจัดเก็บความสัมพันธ์ระหว่างดัชนีและตัวเอกสารทั้งมีการประยุกต์เวิร์ดเน็ตเพื่อใช้ในการสร้างคำที่มีความหมายเดียวกันหรือใกล้เคียงและทดสอบการสืบค้นเอกสารจากดัชนีที่สร้างขึ้นโดยอาศัยคุณลักษณะในการเชื่อมโยงความสัมพันธ์ของข้อมูลเชิงกราฟจะทำการทดสอบเปรียบเทียบประสิทธิภาพการค้นหาข้อมูลได้จากการวัดค่าการค้นคืนหรือค่าความระลึก (Precision) และการวัดค่าความแม่นยำ (Recall) วัดความสามารถของระบบในการค้นหาเอกสารและวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องกับคำศัพท์ที่เป็นคำค้นโดยใช้ทฤษฏีของ TF-IDF มาใช้ในการหาคำสำคัญในเอกสาร โดยหาค่า TF-IDF ของคำศัพท์ภายในเอกสารและทำการทดสอบการใช้ TF-IDF เมื่อมีค่ามากจะแสดงว่าเป็นคำศัพท์ที่เอกสารนั้นให้ความสำคัญสูง

TF-IDF คือการหาค่าลำดับความสำคัญของคำศัพท์ที่เกิดขึ้นในกลุ่มคำศัพท์ในเอกสารเพื่อเรียงลำดับความสำคัญของคำศัพท์โดยการให้ลำดับความสำคัญว่าคำศัพท์เดียวกันในเอกสารฉบับใดควรอยู่ในลำดับความสำคัญที่เท่าไร โดยการนำค่า TF (Term frequency) ค่าความถี่ของคำศัพท์ในเอกสารคูณกับค่าน้ำหนักของคำศัพท์ IDF (Inverse document frequency) ค่าน้ำหนักของคำสำคัญในเอกสาร จะได้ค่าลำดับความสำคัญของคำศัพท์ที่เกิดขึ้นในกลุ่มคำเอกสารนั้น

**สูตรการคำนวณหาค่า TF**



**กำหนดให้**

tf(t,d) = ค่าความถี่ของคำศัพท์ในเอกสาร

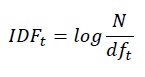
f(t,d) = ความถี่ของคำศัพท์

t = คำศัพท์ภายในเอกสาร

d = เอกสาร

f = ค่าความถี่

**สูตรการคำนวณหาค่า IDF**

****

**กำหนดให้**

IDFt = ค่าการให้ลำดับความสำคัญของคำในเอกสาร

N = จำนวนคำศัพท์ทั้งหมดในเอกสารที่ได้มีการทำการแบ่งข้อความออกเป็นคำ

dft = ค่าความถี่ของคำศัพท์ในเอกสาร TF (Term frequency)

IDF (Inverse document frequency) เป็นการเปรียบเทียบน้ำหนักของคำสำคัญ ในเอกสาร ถ้า

**สูตรการคำนวณหาค่า** **TF-TDF**

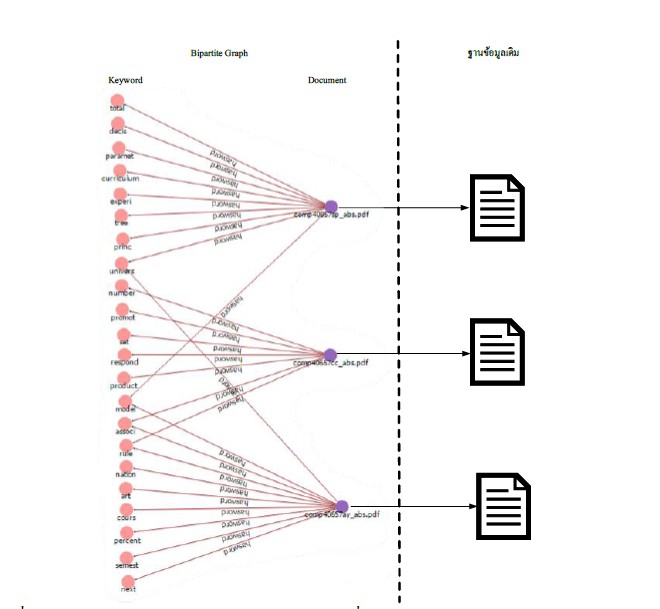


**กำหนดให้**

TF-IDF = ค่าน้ำหนักของคำศัพท์ในกลุ่มเอกสาร

TF=ค่าความถี่ของคำศัพท์ในเอกสาร

IDF= ค่าลำดับความสำคัญของคำศัพท์ในเอกสาร

จากการหาค่าลำดับความสำคัญของคำศัพท์ที่เกิดขึ้นในกลุ่มคำศัพท์ในเอกสารที่กล่าวมาจากนั้นจะบันทึกข้อมูลเข้าสู่ฐานข้อมูลเชิงกราฟเพื่อจัดทำดัชนีเอกสาร ผลลัพธ์ที่จะออกมาจะอยู่ในรูปของฐานข้อมูลเชิงกราฟ

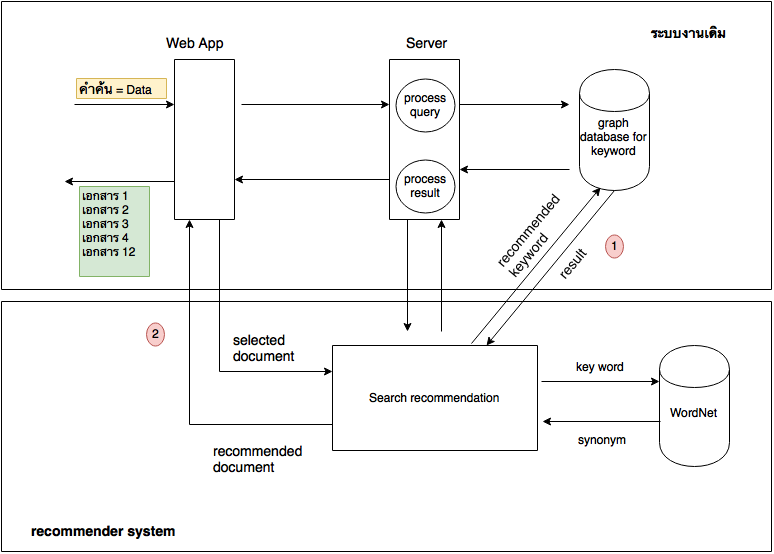
รูป 2.3 แผนผังแสดงความสัมพันธ์ของข้อมูลดัชนีกับตัวเอกสารเมื่อทำการบันทึกข้อมูลลงในฐานข้อมูลเชิงกราฟ

ค่าลำดับความสำคัญของคำศัพท์จะถูกบันทึกลงในฐานข้อมูลเชิงกราฟที่เส้นของความสัมพันธ์ (Edges) ระหว่างคำสำคัญ (Keyword) ต่อ เอกสาร(Document) ซึ่งค่าลำดับความสำคัญของคำศัพท์ที่เกิดขึ้นในกลุ่มคำศัพท์ในเอกสารนั้นจะถูกนำไปใช้ในการสืบค้นข้อมูลเอกสารที่ได้จัดทำขึ้น

**บทที่ 3  
การออกแบบการทดลอง**

**3.1 ขั้นตอนการสืบค้นข้อมูล**

ในบทนี้จะกล่าวถึงวิธีการสืบค้นข้อมูลจากฐานข้อมูลงานวิทยานิพนธ์สำนักหอสมุดมหาวิทยาลัยเชียงใหม่แผนผังโดยรวมของการสืบค้นข้อมูล รูปที่ (3.1)

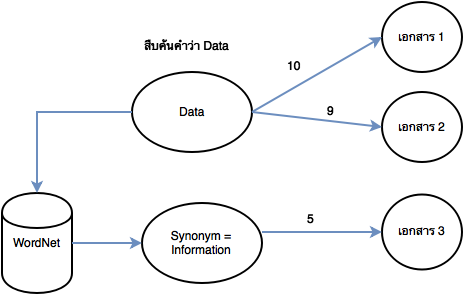


รูปที่ 3.1 แผนผังการทำงานของระบบ

การทำงานของระบบแนะนำผลลัพธ์ (Recommendation system) แบ่งเป็น 2 ขึ้นตอนหลัก

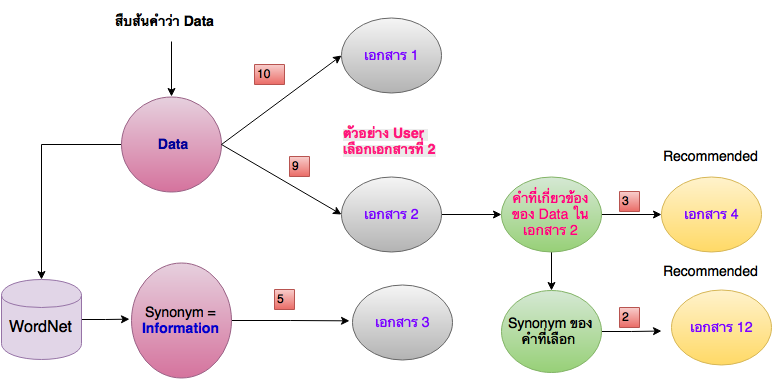
1.แนะนำคำค้น (recommended keyword)  
2.แนะนำเอกสาร (recommended document)

จากการใส่คำค้นข้อมูลงานวิทยานิพนธ์ระบบแนะนำผลลัพธ์จะทำการแนะนำคำค้น (recommended keyword) โดยการนำไปหาคำเหมือน (Synonyms) ของคำค้นนั้นแล้วทำการส่งคำค้นเพื่อสืบค้นข้อมูลจากฐานข้อมูลเชิงกราฟในกระบวนการเลือกข้อมูลจากฐานข้อมูลเชิงกราฟนั้นจะมีระบบแนะนำผลลัพธ์ เพื่อเพิ่มข้อมูลในการสืบค้นได้จากนั้นประมวลผลและส่งข้อมูลการสืบค้น

 ในขั้นตอนของการแนะนำคำค้นนั้นระบบจะทำการแนะนำคำศัพท์เพิ่มเติมจากคำที่ผู้ใช้ ค้นโดยหาคำที่มีความหมายใกล้เคียงจากเวิร์ดเน็ต

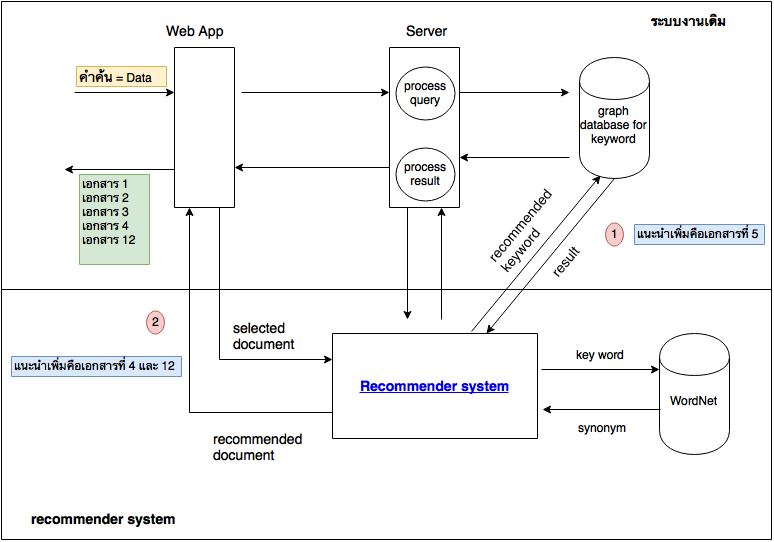
รูปที่ 3.2 ตัวอย่างการแนะนำคำค้น

หลังจากที่ผู้ใช้ได้เลือกเอกสารที่สนใจระบบจะนำคำศัพท์ที่ปรากฏในเอกสารนั้นมาค้นต่อเพื่อแนะนำเอกสารที่มีเนื้อหาเป็นไปในทางเดียวกันกับเอกสารที่ผู้ใช้เลือก



รูปที่ 3.3 ตัวอย่างการแนะนำเอกสาร

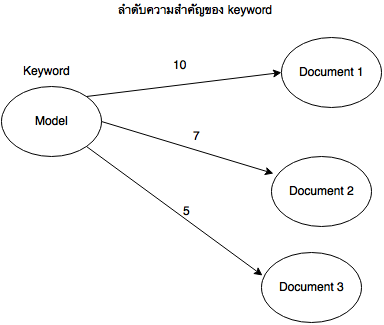
ดั้งนั้น จากตัวอย่างผลลัพธ์ในการค้นหาคำว่า Data จะได้ผลลัพธ์แรกคือ เอกสารที่ 1,2 และ 3 จากนั้นผู้ใช้เลือก  
เอกสารที่ 2 ซึ่งระบบได้แนะนำให้คือเอกสาร 4 และ เอกสาร 12

รูป 3.4 สรุปแผนผังจากการสืบค้นข้อมูล

**บทที่ 4  
การทดลอง**

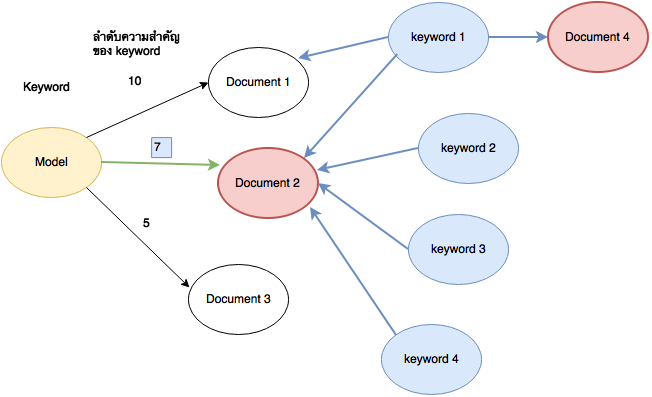
**4.1 ทดสอบขั้นตอนระบบแนะนำผลลัพธ์**

ในขั้นตอนของการค้นหานั้นผู้ใช้จะทำการป้อนข้อมูลคำค้นหาที่ต้องการค้นหาวิทยานิพนธ์ที่ผู้ใช้ต้องการจะค้นหา ในส่วนระบบขั้นตอนการค้นหาของระบบนั้นจะทำการค้นหาตามความสำคัญของคำค้นหาต่อเอกสารงานวิทยานิพนธ์ซึ่งฐานข้อมูลถูกเก็บอยู่ในรูปของฐานข้อมูลเชิงกราฟ ซึ่งจะแบ่งข้อมูลออกเป็น โหนด ซึ่งแบ่งออกเป็นโหนดของคำค้น, เส้นความสัมพันธ์ซึ่งจะระบุความสำคัญของคำค้น ต่อ เอกสารงานวิทยานิพนธ์และโหนดของเอกสาร ซึ่งในขั้นตอนแรกจะให้ความสำคัญกับคำค้นหาก่อน

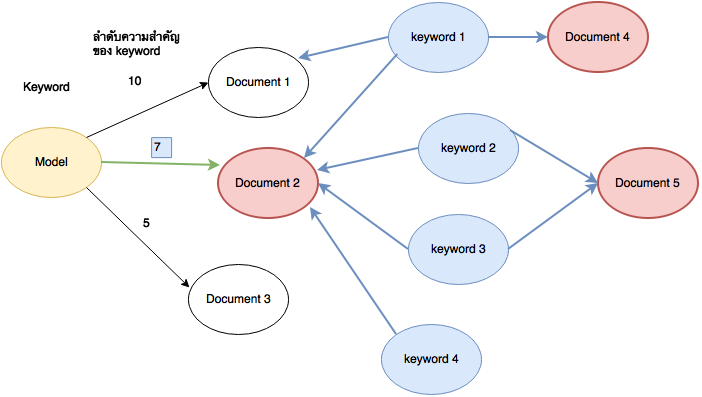


รูป 4.1 แสดงลำดับความสำคัญของแต่ละคำค้นต่อแต่ละเอกสาร

ผลลัพธ์ที่ได้จากการค้นหาคำว่า Model จะปรากฏเอกสารให้เลือก 3 เอกสารซึ่ง จะเรียงตาม ความสำคัญของคำค้นต่อเอกสารผลลัพธ์คือจะค้นหาได้ Document 1 , 2 ,3 ตามลำดับความสำคัญของแต่ละ เอกสารต่อคำค้น

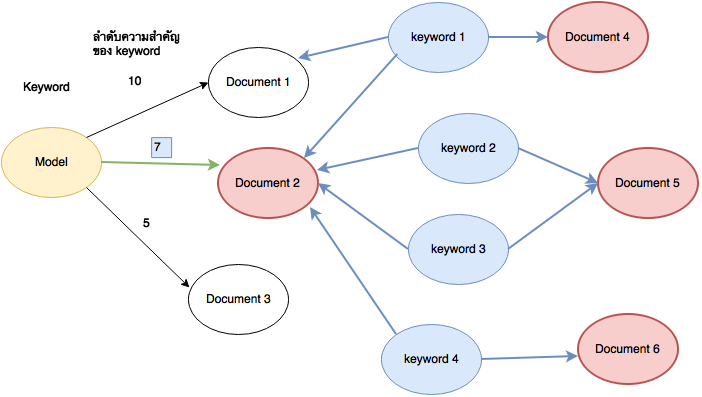
หลังจากนั้นสมมุติว่าผู้ใช้เลือก Document 2 ผลลัพธ์จากการเลือก Document ที่ 2 นั้น จะให้ความสำคัญกับ คำค้นที่มีการแชร์ร่วมกันระหว่างเอกสารมาก่อน จากตัวอย่าง Document 2 มี คำค้นหาที่เกี่ยวข้องกับ Document 1 คือ keyword 1 ซึ่ง Document 4 นั้น มีความสัมพันธ์กับ keyword 1 ที่เกี่ยวข้องระบบจึงแนะนำ Document 4 เป็นลำดับแรก

รูป 4.2 แสดงการแนะนำเอกสารขั้นตอนแรก

จากนั้น keyword 2 และ keyword 3 นั้นมีความสัมพันธ์กับ Document 5 จึงแนะนำ ลำดับที่สองต่อจาก Document 4

รูป 4.3 แสดงการแนะนำเอกสารขั้นตอนที่สอง

ลำดับสุดท้ายคือแนะนำ Document 6 ซึ่งมีความสัมพันธ์กับ keyword 4 เพียง keyword เดียว



รูป 4.4 แสดงการแนะนำเอกสารขั้นตอนสุดท้าย

ดังนั้นการแสดงผลลัพธ์การค้นหาทั้งหมดนั้นที่ได้คือ Document 1 , Document 2, Document 3  
 ในครั้งแรกหลังจากป้อนคำค้นหาและหลังจากที่เลือก Document ที่ 2 จะได้ผลลัพธ์ Document 4 Document 5 และ Document 6 ตามลำดับ

**เอกสารอ้างอิง**

[1] กราฟสองส่วน . (11 กันยายน พศ.2559) เข้าถึงได้จาก <https://th.wikipedia.org/wiki/กราฟสองส่วน>

[2] ฐานข้อมูลเชิงกราฟ (11 กันยายน พศ.2559) เข้าถึงได้จาก<https://en.wikipedia.org/wiki/Graph_database>

[3] คุณสมบัติของฐานข้อมูลโนเอสคิวแอล (13 กันยายน พศ. 2559)  
เข้าถึงได้จาก [www.thaimongo.com/บทความ-mongodb/37-ทำความรู้จัก-nosql-คืออะไร.html](http://www.thaimongo.com/บทความ-mongodb/37-ทำความรู้จัก-nosql-คืออะไร.html)

[4] การจัดหมวดหมู่ของโนเอสคิวแอล (13 กันยายน พศ. 2559)  
เข้าถึงได้จาก <http://cokezaadd.blogspot.com/2014/11/nosql.html>

[5] โอเรียนดีบี (15 กันยายน พศ. 2559) เข้าถึงได้จาก <http://www.poolsawat.com/orientdb/>

[6] ภาษาไพทอนและโอเรียนดีบี (15 กันยายน พศ. 2559) เข้าถึงได้จาก <http://orientdb.com/docs/2.1/Programming-Language-Bindings.html>

[7]เวิร์ดเน็ต (15 กันยายน พศ. 2559) เข้าถึงได้จาก <https://sampaan.wordpress.com/wordnet/>