การพัฒนาระบบค้นหาความสัมพันธ์โดยใช้ Apriori Algorithm

Association Rule Discovery using Apriori Algorithm

โดย

นางสาวดวงกมล มหาวีระ

รหัส 45066091

อาจารย์ที่ปรึกษา

ผศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบค้นหาความสัมพันธ์โดยใช้ Apriori Algorithm
นักศึกษา	นางสาวดวงกมล มหาวีระ
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

ในปัจจุบันการแข่งขันทางธุรกิจและองค์กรมีแนวโน้มเพิ่มขึ้นเรื่อย ๆ และธุรกิจต่าง ๆ มีการใช้ข้อมูลในปริมาณที่สูง และมีข้อมูลมาเกี่ยวข้องด้วยเป็นจำนวนมาก การจัดการด้านข้อมูลของลูกค้าถือว่าเป็นสิ่งสำคัญจึงได้มีการนำเอาหลักการ Data Mining เข้ามาประยุกต์ใช้เพื่อช่วยในการวิเคราะห์ข้อมูลหลาย ๆ รูปแบบและหาความสัมพันธ์ของข้อมูลช่วยในการสนับสนุนการตัดสินใจต่างๆ ที่เป็นประโยชน์ต่อธุรกิจ และยังสามารถคาดการณ์สิ่งที่จะเกิดขึ้นในอนาคตได้ เป็นการเพิ่มประสิทธิภาพให้กับองค์กรนั้น

โครงการนี้นำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลที่เกี่ยวข้องกัน โดยใช้ Apriori Algorithm ซึ่งเป็นอัลกอลิทึมพื้นฐานในการหากฎความสัมพันธ์ของข้อมูล โดยโปรแกรมเริ่มการทำงานโดยการรับข้อมูลเข้ามา ขั้นต่อมาโปรแกรมจะเข้าสู่การเตรียมข้อมูล จากนั้นจะทำการค้นหาค่า Support และค่า Confidence มาแสดง

Title	Association Rule Discovery using Apriori Algorithm
Student	Miss. Doungkamol Mahaveera
Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

At present, 2 trend of competitions in economics and organizations is increasing. These organizations are highly using and having associated data in their business. The customer data management is the most important. So, data mining principle is applied to synthesize many data patterns and to find data associations that support decision making, in which advantage in business. Beside these advantages, it will be effectiveness to organization in future forecasting.

This project presented processes and ways to develop working system, to analyze the association of relative data. The project used to find data association rules. The program start working by data receiving, data preparing, support data finding and confidence value showing, respectively

กิตติกรรมประกาศ

ความสำเร็จของการพัฒนาโครงการศึกษากรณีพิเศษฉบับนี้ สำเร็จขึ้นได้จากความช่วยเหลือของบุคคลหลายๆ ท่าน ข้าพเจ้ามีความรู้สึกขอบคุณทุกท่านที่มีส่วนช่วยเหลือในด้านต่างๆ ด้วยความจริงใจ หากขาดบุคคลที่จะกล่าวถึงดังต่อไปนี้ไป ก็จะส่งผลกระทบต่อความสำเร็จของโครงการศึกษากรณีพิเศษฉบับนี้

ขอขอบคุณบิดา มารดาผู้ให้กำเนิด น้องอาร์ท และพี่แป้นที่คอยให้กำลังใจกันเสมอมา

ขอขอบคุณอาจารย์ ผศ.ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษา ที่ให้ความรู้ และนำหนังสือและให้คำปรึกษา ตลอดจนแนวทางการแก้ไขปัญหา

ขอขอบคุณนางสาวลัดดาวัลย์ อุ่นศิริ ที่ให้คำแนะนำในส่วนของโปรแกรม และ Java Programming

ขอขอบคุณนางสาวดลพร หวังเสริมวงศ์ ที่คอยมาช่วยดูโปรแกรมตอน error ตอนก่อน present

ขอขอบคุณนางสาวกุลธิดา บุญโฉม ที่คอยแนะนำในหลายๆ เรื่อง และเป็นกำลังใจให้ตลอดเวลา

ขอขอบคุณนางสาวชัชพรรณ คล่องพิริยะ ที่ช่วยเหลือข้าพเจ้าในหลายๆ เรื่อง

ขอขอบคุณนายเอกภูมิ อารีรัตน์ ที่คอยให้คำแนะนำ เตือน และช่วยเหลือกันในหลายๆ เรื่อง

ขอขอบคุณเพื่อน IS 13.2 ที่คอยปลุกกันขึ้นมาทำงานตอนดึกๆ และกำลังใจที่มีให้กันเสมอมา

สุดท้าย ขอขอบคุณเพื่อนร่วมงานทุกคน ผู้ที่คอยช่วยเหลือเรื่องงานเวลาที่ลางานบ่อยๆ เพื่อมาทำโปรเจค และเป็นกำลังใจให้เสมอมา

ดวงกมล มหาวีระ

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

ภาพที่

บทที่ 1

บทนำ

1. หลักการและเหตุผล

เนื่องจากการวิเคราะห์ข้อมูลจำนวนมากในปัจจุบันไม่ใช่เรื่องง่าย และยากที่จะวิเคราะห์ถึงความสัมพันธ์และแนวโน้มต่างๆ จากหลายฐานข้อมูลได้อย่างครบถ้วน จึงต้องหาวิธีที่จะวิเคราะห์ข้อมูลเพื่อที่จะให้ทราบถึงความสัมพันธ์ในรูปแบบต่างๆ ที่ซ่อนอยู่ในฐานข้อมูลเพื่อดึงสารสนเทศที่ได้มาช่วยในการทำนายแนวโน้มและพฤติกรรมของข้อมูลในอนาคต จึงได้นำเอาเทคนิคของดาต้าไมนิ่ง (Data Mining) เข้ามาช่วยในการวิเคราะห์ข้อมูลเพื่อให้ทราบถึงความสัมพันธ์ในรูปแบบต่างๆ และช่วยสนับสนุนการตัดสินใจทางธุรกิจ

2. วัตถุประสงค์

เพื่อนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการวิเคราะห์ความสัมพันธ์ของข้อมูลต่างๆ ให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้วางแผนกลยุทธทางการตลาดได้อย่างมีประสิทธิภาพ รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้ในการสนับสนุนการตัดสินใจของผู้บริหารสำหรับวางแผนเพื่อทำรายการส่งเสริมการขาย เพื่อให้เหมาะสมกับความต้องการของลูกค้า และยังทำให้เกิดความเข้าใจความต้องการของลูกค้าได้รวดเร็วยิ่งขึ้น รวมทั้งยังช่วยในการประมวณผลในการวิเคราะห์ข้อมูลลูกค้าใหม่ว่าเป็นลูกค้าที่มีศักยภาพหรือไม่จากกฎที่ค้นหาได้

3. ขอบเขตของการดำเนินงาน

โครงการนี้เป็นการศึกษาถึงการนำเทคโนโลยีดาต้าไมนิ่งมาประยุกต์ใช้ อาศัยหลักการของ Link Analysis มาใช้ในการวิเคราะห์ความสัมพันธ์ของข้อมูล โดยนำโมเดลของ Association rule มาประยุกต์ใช้ และใช้ Apriori Algorithm ในการประมวลผลข้อมูล

4. ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1). ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องของดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้

2). ศึกษาทฤษฎี Association Rules โดยใช้ Apriori Algorithms

3). ออกแบบและพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล

4). สรุปผลการศึกษา

5. เทคโนโลยีที่ใช้ในการพัฒนาระบบ

ระบบได้รับการพัฒนาเป็น Application ซึ่งทำงานในระบบปฏิบัติการ Microsoft Windows 2000 โดยมีรายละเอียดดังนี้

- การพัฒนา Application ใช้โปรแกรม Borland Jbuilder9 Enterprise
- ใช้ Microsoft SQL Server 2000 เป็นฐานข้อมูลหลักในการเก็บข้อมูล

6. ประโยชน์ที่คาดว่าจะได้รับ

1. ผู้บริหารสามารถมองเห็นถึงภาพรวมของธุรกิจได้มากขึ้น เพื่อนำไปเป็นแนวทางในการสนับสนุนการตัดสินใจทางต่อไป
2. สามารถนำผลลัพธ์ที่ได้จากการวิเคราะห์มาเพิ่มประสิทธิภาพให้กับองค์กร
3. ได้รับความรู้เกี่ยวกับเทคโนโลยีดาต้าไมนิ่ง
4. ได้นำทฤษฎีเกี่ยวการการโปรแกรมเชิงวัตถุ มาประยุกต์ใช้ในการออกแบบและพัฒนาระบบงานจริง
5. ได้เครื่องมือที่นำมาประยุกต์ใช้เพื่อหาความสัมพันธ์ของข้อมูลต่างๆ ที่เกิดขึ้นได้ สามารถทำนายแนวโน้มและพฤติกรรมของข้อมูลในอนาคตได้

บทที่ 2

หลักการทำงานของ Data Mining

2.1 Data Mining

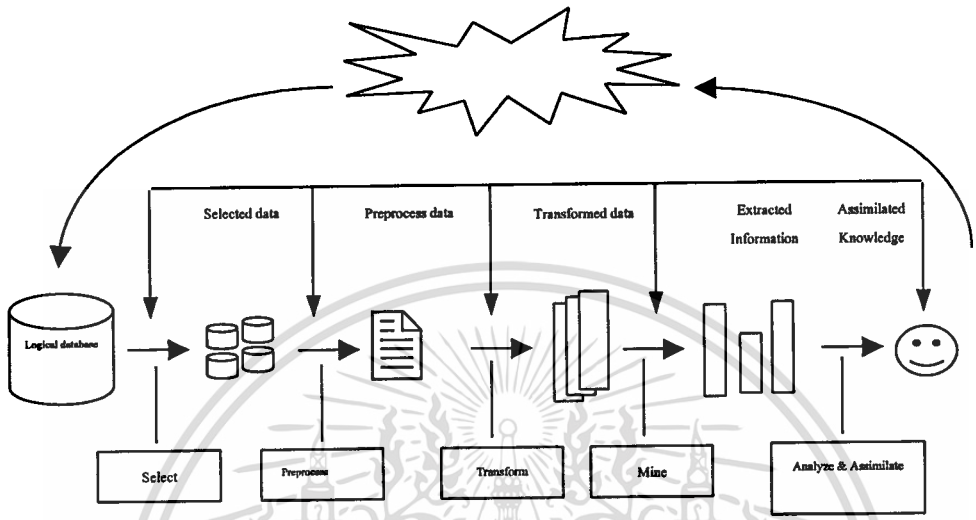
Data Mining เป็นกระบวนการค้นหาสารสนเทศที่เป็นประโยชน์ โดยหาจากความสัมพันธ์และรูปแบบทั่วไปของข้อมูลที่ซ่อนอยู่ในฐานข้อมูลใหญ่ เพื่อใช้เป็นแนวทางในการตัดสินใจ หรือทำนายแนวโน้มและพฤติกรรมโดยอาศัยข้อมูลในอดีต ข้อมูลที่ได้จากการทำ Data Mining ไม่ได้เกิดจากสมมติฐาน หรือการคาดคะเนจากประสบการณ์ แต่เป็นข้อมูลหรือความสัมพันธ์ที่เกิดขึ้นจริงในโลกของธุรกิจปัจจุบันบริษัทต่างๆจะพยายามหาเทคนิคที่สามารถนำความสำเร็จมาสู่บริษัท เช่น ในโลกธุรกิจขนาดย่อมจะสร้างความสัมพันธ์กับลูกค้า โดยสังเกตจากความต้องการ ความชอบและความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากผลสะท้อนในอดีตว่าจะทำอย่างไรให้การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต หรือ บริษัทที่เป็นผู้ออกบัตรเครดิตและธนาคารต่างๆ จะมี ขบวนการที่ใช้ดาต้าไมนิ่งให้เป็นประโยชน์ ในการตัดสินใจว่าลูกค้ากลุ่มใดเป็นกลุ่มที่ดี , ทำความเข้าใจลูกค้า , ช่วยในการแยกประเภทของลูกค้าและจะทำนายกลุ่มของประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต เป็นต้น อย่างไรก็ตามการเรียนรู้นั้นต้องมากกว่าการเก็บสะสมข้อมูลอย่างตรงไปตรงมา จะทำให้การทำงานไม่เป็นประสิทธิภาพ ซึ่งสารสนเทศที่ได้จากการทำ Data Mining จะมีคุณสมบัติ 3 ประการ คือ

1. ข้อมูลที่ไม่เคยทราบมาก่อน (Unknown) เป็นข้อมูลที่ผู้ใช้งานไม่รู้มาก่อน และไม่ชัดเจนไม่สามารถตั้งสมมติฐานล่วงหน้าว่าควรเป็นแบบใด
2. ข้อมูลที่มีความถูกต้อง มีเหตุผลรับรอง และสามารถพิสูจน์ได้ (Valid)
3. ข้อมูลที่สามารถนำไปใช้ประโยชน์ได้ (Actionable)

Data Mining นั้นประกอบด้วยหลายๆ ขั้นตอน และแต่ละขั้นตอนก็จะมีการทำซ้ำในขั้นตอนนั้นๆ หรือต้องมีการวนกลับมาทำซ้ำใหม่ ในการศึกษาข้อมูลแต่ละครั้งจึงจำเป็นต้องมีการกำหนดวัตถุประสงค์ของการ Mining ที่ชัดเจนก่อน ซึ่งการทำ Mining เป็นเพียงขั้นตอนหนึ่งของกระบวนการทั้งหมด

2.2 กระบวนการทำ Data Mining

กระบวนการของ Data Mining เป็นกระบวนการของการสร้างแบบจำลอง (Model) ประกอบด้วย 5 ขั้นตอน แสดงในภาพที่ 2.1



ภาพที่ 2.1 แสดงกระบวนการทำงานของ Data Mining

1. การกำหนดวัตถุประสงค์ (Business Objective Determination)

การกำหนดวัตถุประสงค์จะต้องกำหนดปัญหาและความต้องการทางธุรกิจให้ชัดเจน ซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจ และการวิเคราะห์ข้อมูลเบื้องต้นว่า เรามีข้อมูลใดบ้างและต้องการอะไรจากข้อมูลบ้าง ซึ่งขั้นตอนนี้ จะสามารถมองถึง อัลกอลิทึม และฐานข้อมูลที่จะใช้งานเบื้องต้น เป็นการนำไปสู่การสร้างแบบจำลองที่เหมาะสม ขึ้นอยู่กับเป้าหมายทางธุรกิจ นอกจากนี้แล้วการกำหนดปัญหายังต้องดูถึงความเป็นไปได้ด้วยเพื่อดูว่ามีความจำเป็นที่ต้องใช้ Data Mining หรือไม่ เพราะไม่ได้หมายความว่าทุกปัญหาสามารถแก้ไขด้วยเทคนิค Data Mining ดังนั้นการกำหนดปัญหาที่ไม่ถูกต้องย่อมนำไปสู่ความสำเร็จในการแก้ปัญหาได้ยาก

2. การจัดเตรียมข้อมูล (Data Preparation)

เป็นขั้นตอนที่สำคัญ ซึ่งต้องใช้เวลาและความพยายามอย่างมากกว่าขั้นตอนอื่น ๆ ทั้งหมด เนื่องจากต้องมีการพิจารณาข้อมูลในแทบทุกเรื่อง โดยจะต้องมีการย้อนกลับมาทำซ้ำในขั้นตอนการเตรียมข้อมูล และการสร้าง Model เนื่องจากการเรียนรู้บางสิ่งจาก Model อาจนำไปสู่การแก้ไขข้อมูล ซึ่งประกอบด้วย 3 ขั้นตอน คือ

- **การคัดเลือกข้อมูล (Data Selection)** คือ การคัดเลือกข้อมูลจากข้อมูลทั้งหมดขององค์กร โดยมีจุดประสงค์หลัก คือ การระบุลักษณะ และเลือกข้อมูลที่ต้องการ และนำข้อมูลที่ไม่

ต้องการออกไป โดยคำนึงถึงวัตถุประสงค์ในการนำข้อมูลมาใช้ นอกจากนี้ยังทำความเข้าใจกับข้อมูล และประเภทของข้อมูลที่จะต้องนำมาใช้ด้วย โดยประเภทของข้อมูลแบ่งเป็น

- Categorical แบ่งเป็น
  - Nominal คือ ตัวแปรที่ลำดับไม่มีความสำคัญ (ลำดับไม่มีผลกับค่า) เช่น สถานะการแต่งงาน (โสด , แต่งงาน , หม้าย) , เพศ (ชาย , หญิง) , ระดับการศึกษา (มหาวิทยาลัย , มัธยม , ประถม) เป็นต้น
  - Ordinal คือ ตัวแปรที่ลำดับมีความสำคัญ (ลำดับมีผลกับค่า) เช่น อัตราการใช้บัตรเครดิตของลูกค้า (good , regular , poor)
- Quantitative แบ่งเป็น
  - Continuous จะเก็บค่าตัวเลขที่เป็นจำนวนจริง (Real number) เช่นค่าเฉลี่ยของการซื้อ , ค่าใช้จ่ายบริษัทเฉลี่ยต่อเดือน
  - Discrete จะเก็บค่าตัวเลขที่เป็นจำนวนเต็ม (integer) เช่น จำนวนพนักงานในบริษัท

ในการเลือกข้อมูลต้องคำนึงถึงอายุของข้อมูลด้วย เช่น ข้อมูลอาชีพของลูกค้า ซึ่งจะมีการเปลี่ยนแปลงบ่อยเมื่อเวลาผ่านไป เพราะฉะนั้นการนำเอาข้อมูลลูกค้ามาใช้ ต้องตรวจสอบให้แน่ชัดว่าเป็นข้อมูลนั้นถูกต้องหรือไม่

- **การตรวจสอบข้อมูล (Data Processing)** ในกระบวนการนี้จะมีปริมาณข้อมูลจำนวนหนึ่งที่ถูกเลือกเข้ามาจากกระบวนการ Data Selection ซึ่งข้อมูลเหล่านี้จะต้องเป็นข้อมูลที่ถูกต้องพร้อมสำหรับการทำ Mining จึงต้องทำการตรวจสอบข้อมูลว่าข้อมูลที่คัดเลือกมาเป็นข้อมูลที่เหมาะสมหรือไม่โดยใช้หลักการทางสถิติ และเทคนิคการนำเสนอข้อมูลที่น่าสนใจของข้อมูล เช่น ข้อมูลประเภท Categorical การจัดการกระจายของข้อมูลจะทำให้เข้าใจข้อมูลที่มีอยู่ได้ดียิ่งขึ้น วิธีการที่ง่ายที่สุด คือ การนำเอาข้อมูลนั้นไปสร้างกราฟ ซึ่งจะช่วยให้เห็นความโน้มเอียงของข้อมูลและข้อมูลที่ผิดปกติได้ ส่วนข้อมูลประเภท Quantitative การวิเคราะห์ข้อมูลจะทำได้โดยการหาค่าสูงสุด (Max) , ค่าต่ำสุด (Min) , ค่าเฉลี่ย (Mean) , ค่าที่ปรากฏบ่อย (Mode) , ค่ามัธยฐาน (Median) เป็นต้น ซึ่งสิ่งที่ผิดปกติที่จะปรากฏให้เห็นในขั้นตอนนี้คือ
  - Noisy Data เป็นข้อมูลที่มีลักษณะแตกต่างจากข้อมูลที่คาดการณ์เอาไว้ หรือค่าของข้อมูลอาจจะผิดไปจากที่ควรจะเป็น ซึ่งอาจเกิดจากการป้อนข้อมูลผิด เช่น บันทึกค่าเงินเดือนพนักงานติดลบ หรือ บันทึกส่วนสูงเป็น 560 เซนติเมตร เป็นต้น ค่าเหล่านี้ควรถูกแก้ไข หรือ ไม่นำมาวิเคราะห์ ดังนั้นควรมีขั้นตอนการตรวจสอบข้อมูลก่อนนำไปใช้

- Missing Values เป็นข้อมูลที่ไม่ได้ถูกเลือกมาจากขั้นตอนการเลือก (Data Selection) หรือเป็นค่าที่ไม่ถูกต้อง (ไม่สมบูรณ์) ซึ่งอาจถูกลบระหว่างการทำ Noise Detectionคือ ข้อมูลบางส่วนอาจหายไป อาจเกิดจากความผิดพลาดของมนุษย์ หรือ ไม่มีข้อมูลส่วนนี้ในขณะที่รับข้อมูล ถ้าข้อมูลที่ขาดมีจำนวนน้อยอาจแก้ไขโดยการตัดข้อมูลนั้นทิ้งทั้งรายการ แต่ถ้าข้อมูลที่ขาดไปมีมากต้องบันทึกส่วนที่หายไปด้วยค่าเฉลี่ย (Mean) หรือ ค่าที่ปรากฏบ่อย (Mode) (สำหรับข้อมูลที่เป็น Categorical อาจบันทึกด้วยค่าฐานนิยมแทน หรือ บันทึกเป็น "Unknown")

- **การปรับเปลี่ยนรูปแบบของข้อมูล (Data Transformation)** ในระหว่างขั้นตอนนี้ข้อมูลที่ได้กลั่นกรองแล้วจะถูกแปลงให้อยู่ในรูปของข้อมูลที่พร้อมจะถูกวิเคราะห์ เพื่อให้อยู่ในรูปแบบของข้อมูลที่ตรงตามอัลกอลิทึม ของ Data Mining ที่จะใช้ เช่นการแปลงข้อมูลตัวเลขให้เป็นช่วง เพื่อใช้กับ Decision tree หรือการปรับอัตราส่วนตัวเลขให้อยู่ในช่วง 0 – 1 เพื่อใช้กับอัลกอลิทึมใน Neural network

3. การทำ Data Mining

เป็นการประมวลผลข้อมูลตาม อัลกอลิทึม ที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนที่ผ่านมา โดยเมื่อทำในส่วนของ Data Mining แล้วอาจต้องย้อนกลับไปทำในขั้นตอนของการเตรียมข้อมูลใหม่ ในการพัฒนา Data Mining นั้นจะเกี่ยวข้องกับการใช้อัลกอลิทึมหลาย ๆ แบบซึ่งแต่ละแบบมีข้อดีข้อเสียแตกต่างกันไป

4. การทำความเข้ากับแบบจำลอง (Analysis of Results)

เป็นการวิเคราะห์ผลการประมวลผล ซึ่งจะทำการแปลความหมายและประเมินผลลัพธ์ที่ได้จากขั้นตอนการทำ mining ว่าสามารถนำไปใช้บรรลุวัตถุประสงค์ที่ต้องการหรือไม่รวมทั้งเป็นการประเมินถึงความถูกต้องของผลที่ได้จากการทำ เพราะบางครั้งผลที่ได้จากการทำ mining อาจมีข้อผิดพลาดได้ เครื่องมือทางด้าน Graphical Visualization จะช่วยวิเคราะห์ข้อมูลได้อย่างสะดวกรวดเร็วยิ่งขึ้น

5. การนำสารสนเทศที่ได้ไปใช้ประโยชน์ (Assimilation of Knowledge)

เป็นการรวบรวมความเข้าใจทางธุรกิจที่เป็นผลมาจาก Analysis of Result มารวมกับส่วนความรู้เพื่อนำไปใช้ต่อไป ซึ่งมีหลักอยู่ 2 ประการ คือ

- การนำเสนอแนวความคิดทางธุรกิจที่ค้นพบใหม่
- หาแนวทางที่จะใช้กฎเกณฑ์ใหม่ที่ค้นพบเพื่อให้เกิดประโยชน์สูงสุด

2.3 Data Mining Operation

Data Mining ประกอบด้วย 4 โมเดลหลัก คือ

1). การสร้างแบบจำลองพยากรณ์ (Predictive Modeling)

เป็นการคาดคะเน ทำนายถึงความเป็นไปได้ มีลักษณะคล้ายการเรียนรู้ของมนุษย์ คือ จะต้องเข้าใจลักษณะของสิ่งที่ศึกษาอย่างแท้จริง เราจะใช้ Model นี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่ เพื่อกำหนดคุณสมบัติสำคัญของข้อมูล ฉะนั้นข้อมูลที่มีอยู่ต้องเป็นข้อมูลที่สมบูรณ์ จึงจะทำให้แบบจำลองให้คำทำนายที่ถูกต้อง การพัฒนาแบบจำลองพยากรณ์จะนำเอาข้อมูลในอดีตมาสร้างเป็นแบบจำลองโดยแบ่งออกเป็น 2 ขั้นตอนคือ

- ช่วงการเรียนรู้ (Training Phase) เป็นการสร้างโมเดลโดยการใช้ข้อมูลในอดีตและมีจำนวนข้อมูลจำนวนมาก
- ช่วงการทดสอบ (Testing Phase) เป็นการตรวจสอบความน่าเชื่อถือและประสิทธิภาพของโมเดลที่สร้างขึ้นมาว่ามีความเหมาะสมหรือไม่ โดยใช้กับข้อมูลที่ถูกแบ่งเอาไว้สำหรับการทดสอบ ซึ่งเป็นข้อมูลที่มีจำนวนไม่มากนัก

Predictive Modeling แบ่งเป็น 2 ลักษณะคือ

- Classification เป็นแบบจำลองเพื่อทำนายกลุ่มของรายการที่เราสนใจ โดยจะทำการแบ่งกลุ่มของข้อมูลตามชนิดกลุ่มข้อมูลที่ควรจะเป็น เช่น การสร้างแบบจำลองของลูกค้าสินเชื่อ แล้วป้อนจำนวนเงินที่กู้ ศักยภาพ ผลประกอบประกอบการ และฐานะทางการเงิน การปฏิบัติตามเงื่อนไขของธนาคารที่ผ่านมาว่าควรจะให้วงเงินสินเชื่อเพิ่มหรือไม่ เป็นต้น อัลกอลิทึมที่นิยมใช้ คือ Tree Induction และ Neural Induction
- Value Prediction เป็นการทำนายค่าความต่อเนื่องของข้อมูล และค่าที่เป็นตัวเลข เช่น การทำนายราคาหุ้น เป็นต้น โดยมีเทคนิคที่นิยมใช้คือ Linear regression และ Nonlinear regression

2). การแบ่งฐานข้อมูล (Database Segmentation)

Segmentation หรือ Clustering เป็นการทำการแบ่งกลุ่มข้อมูล เพื่อทำการแยกออกให้ทราบว่าข้อมูลชุดนี้มีทั้งหมดกี่กลุ่ม ซึ่งการแบ่งกลุ่มข้อมูลนี้ไม่สามารถกำหนดได้ว่าข้อมูลนี้จะอยู่ในกลุ่มใด โดยการจัดกลุ่มดังกล่าวได้พิจารณาคุณสมบัติหลาย ๆ มิติของข้อมูล ถ้ารายการในข้อมูลมีลักษณะคล้ายคลึงกันเป็นกลุ่มเดียวกันได้ ก็จะรวมเข้าด้วยกัน เพื่อให้ง่ายต่อการวิเคราะห์ เช่น การแบ่งลูกค้าออกตามอายุ , เพศ , รายได้ เป็นต้น เทคนิคที่นิยมใช้ คือ Demographic clustering และ Neural clustering

3). การวิเคราะห์ความสัมพันธ์ (Link Analysis)

เป็นการศึกษาความสัมพันธ์ของข้อมูลว่ามีความสัมพันธ์กันในรูปแบบ ลักษณะใด โดยเรียกความสัมพันธ์นี้ว่าเป็น "Association" เป็นโมเดลที่นิยมมากในการวิเคราะห์หาความสัมพันธ์ระหว่างลูกค้ากับสินค้าหรือบริการ แบ่งได้เป็น 3 ลักษณะ ตามการวิเคราะห์ข้อมูลดิบ

- Association Discovery การวิเคราะห์ข้อมูลที่เกิดขึ้นพร้อมกันภายในกลุ่มข้อมูลเดียวกัน เช่น การวิเคราะห์พฤติกรรมการซื้อของผู้บริโภคเพื่อศึกษาแนวโน้มการซื้อสินค้าและบริการ เพื่อนำกลับมาวางกลยุทธ์ ส่งเสริมการขาย เป็นต้น
- Sequential Pattern Discovery การศึกษาความสัมพันธ์ระหว่างข้อมูลโดยเทียบกับเวลา ซึ่งเป็นการศึกษาพฤติกรรมในระยะยาว
- Similar Time Sequence Discovery การศึกษาพฤติกรรมของข้อมูลที่เกิดขึ้นทั้งหมดหรือเกิดขึ้นในช่วงเวลาเดียวกัน เพื่อหาความสัมพันธ์ระหว่างกลุ่มข้อมูลเหล่านี้

4). การตรวจสอบค่าเบี่ยงเบน (Deviation Detection)

เป็นโมเดลที่จะใช้เทคนิคทางสถิติและการทำให้เห็นภาพ ซึ่งเป็นการสรุปข้อมูลให้ออกมาในรูปแบบการแสดงผลกราฟฟิก operation นี้สามารถใช้ในการตรวจสอบลายเซ็นปลอม หรือ บัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

2.4 ประโยชน์ของ ดาต้าไมนิ่ง

2.4.1 ตัวแบบที่ได้ง่ายต่อการเข้าใจ ผู้ที่ไม่มีความรู้พื้นฐานทางสถิติก็สามารถแปรความจากตัวแบบได้สามารถนำสารสนเทศที่ได้ไปใช้ในกระบวนการทางธุรกิจได้

2.4.2 สามารถวิเคราะห์ข้อมูลจำนวนมากได้ สามารถวิเคราะห์ได้ถึงหน่วยกิกะไบต์

2.4.3 ดาต้าไมนิ่ง จะค้นพบในสิ่งที่เราคาดไม่ถึง เนื่องจากการที่มีตัวแบบและการตรวจสอบที่หลากหลาย จะพบว่าการค้นหาจากตัวแปรที่นำมารวมกัน ทำให้ได้สารสนเทศที่เกี่ยวข้องกับธุรกิจ

2.4.4 ตัวแปรไม่จำเป็นต้องถูกบันทึก ดาต้าไมนิ่ง สามารถรองรับได้ทั้งตัวแปรที่เป็นตัวเลขหรือเป็นกลุ่ม ตัวแปรเหล่านี้จะแสดงให้เห็นในตัวแบบโดยมีรูปแบบเช่นเดียวกับที่เก็บในดาต้าเบส

2.4.5 ตัวแบบมีความถูกต้อง เนื่องจากมีการทดสอบด้วยเทคนิคทางสถิติ

2.4.6 ตัวแบบถูกสร้างได้อย่างรวดเร็ว ตัวแบบต่างๆ จะถูกทดสอบและเลือกตัวแบบที่ดีที่สุดสำหรับผู้ใช้

2.5 ตัวอย่างการประยุกต์ใช้ ดาต้าไมนิ่ง

จากประโยชน์ที่เห็นได้ชัดของ ดาต้าไมนิ่งในการควบคุมค่าใช้จ่าย และเพิ่มรายได้ให้กับองค์กร ทำให้ ดาต้าไมนิ่ง ได้รับความนิยมมากขึ้น นอกจากนี้ยังมีการนำ ดาต้าไมนิ่ง ไปใช้กับฐานข้อมูลทางการตลาดในด้านการหากลุ่มลูกค้าที่น่าจะส่งmail หรือนำเสนอรายการสินค้า เพื่อจะสามารถลดค่าใช้จ่าย และเพิ่มรายได้ที่จะเกิดจากการขายสินค้านั้นๆ ได้ การโฆษณาประชาสัมพันธ์สินค้ากับกลุ่มลูกค้าเป้าหมายจะได้ผลกลับมาดีกว่า ตัวอย่างการนำ Data Mining มาประยุกต์ใช้

Direct Marketing

ดาต้าไมนิ่ง ได้เข้ามามีบทบาทในการทำตลาดขายตรงเพราะความสามารถในการทำนายว่าลูกค้ารายใดน่าจะมีการตอบรับที่ดีในการประชาสัมพันธ์สินค้าด้วยไปรษณีย์ หรือทำนายกลุ่มลูกค้าที่มีแนวโน้มจะซื้อสินค้าชนิดหนึ่ง ซึ่งจะทำให้การประชาสัมพันธ์สินค้าประเภทหนึ่งๆ ต่อกลุ่มลูกค้าเป้าหมายได้ผลมากขึ้น นอกจากนั้นยังช่วยลดค่าใช้จ่ายจากการไม่ต้องส่งใบโฆษณาไปยังลูกค้าทุกราย แต่จะส่งเฉพาะลูกค้าที่มีแนวโน้มจะตอบรับเท่านั้น

Cross Selling

การเพิ่มมูลค่าของลูกค้าก็เป็นงานการตลาดที่สำคัญประการหนึ่ง การเพิ่มมูลค่าของลูกค้าแตกต่างจากการเพิ่มส่วนแบ่งทางการตลาดตรงที่การเพิ่มส่วนแบ่งตลาดได้จากการเพิ่มจำนวนลูกค้า แต่การเพิ่มมูลค่าของลูกค้าคือการเพิ่มปริมาณการซื้อสินค้าของลูกค้าแต่ละราย ซึ่งในอดีตใช้วิธีการเดาว่าในขณะที่ลูกค้าซื้อสินค้าชนิดหนึ่งๆ ควรจะนำเสนอสินค้าใดเพิ่ม แต่เมื่อมีการนำ ดาต้าไมนิ่งมาใช้ทำให้นักการตลาดทราบว่าลูกค้ารายใดน่าจะซื้อสินค้าชนิดใหม่ของบริษัทด้วย หรือควรจะนำเสนอสินค้าชนิดใดไปพร้อมๆ กัน

Trend Analysis

การทำความเข้าใจแนวโน้มของตลาด มีความสำคัญในการลดต้นทุน และการนำสินค้าออกสู่ตลาดในช่วงเวลาที่เหมาะสม สถาบันการเงินต่างๆ ต้องการวิธีการที่จะรู้ถึงรูปแบบของการเปลี่ยนแปลงปริมาณเงินฝากและถอน อย่างรวดเร็ว ส่วนผู้ค้าปลีกก็ต้องการที่จะรู้ว่าสินค้าชนิดใดจะขายได้เมื่อวางขายด้วยกัน

บทที่ 3

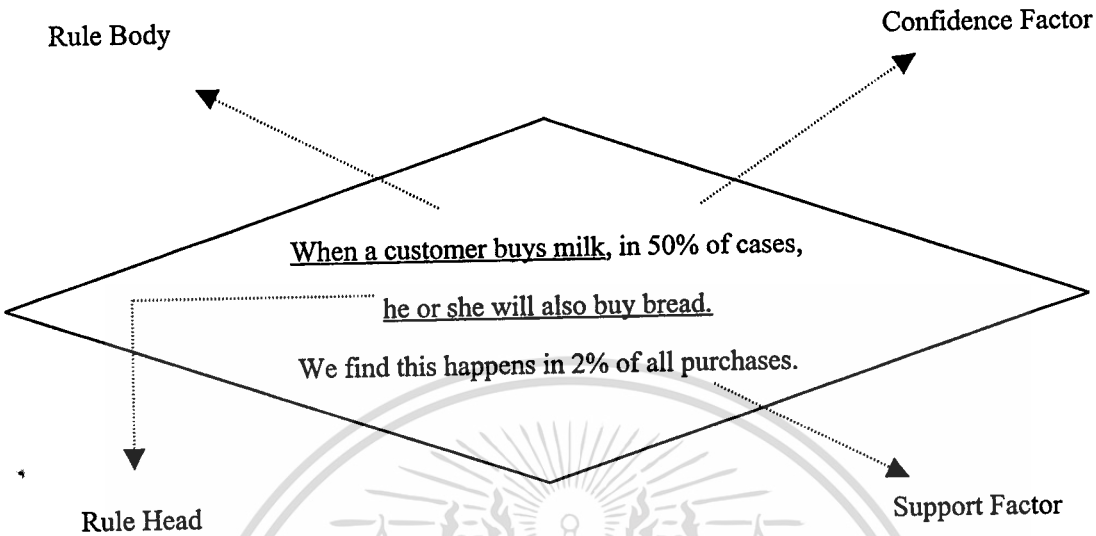
Association Rule และ Apriori Algorithms

3.1 Association Rule

Association Rule เป็นเทคนิคในการค้นหาความสัมพันธ์ของข้อมูล เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่างๆ โดยเทคนิคนี้จะใช้กันอย่างแพร่หลายในการขายสินค้า หรือการวิเคราะห์ข้อมูลที่เป็นทรานแซกชั่น

ตัวอย่างของ association rule ได้แก่ market basket analysis (MBA) ซึ่งเป็นเทคนิคที่นำไปใช้ในด้านการตลาด เพื่อวิเคราะห์พฤติกรรมการซื้อของลูกค้า โดยหาความสัมพันธ์ระหว่างสินค้าต่างๆ ที่ลูกค้าซื้อ การค้นหาความสัมพันธ์สามารถช่วยผู้ขายพัฒนากลยุทธ์ทางการตลาด โดยพิจารณาจากสินค้าที่มักจะถูกซื้อพร้อมๆ กัน เช่น ถ้าลูกค้าซื้อนม เขามักจะซื้อขนมปังพร้อมกัน ข้อมูลนี้สามารถนำไปสู่การเพิ่มยอดขาย โดยการช่วยผู้ขายในการวางแผนการตลาด และการวางแผนการจัดชั้นวางสินค้า เช่น การวางขนมปังและนมไว้ใกล้กัน อาจจะเพิ่มยอดการขายสินค้าทั้งสองนี้ หรืออาจจะวางไว้คนละมุมของร้าน เพื่อที่ว่าลูกค้าซื้อนมแล้ว ต้องการที่จะซื้อขนมปังก็จะต้องเดินผ่านสินค้าตัวอื่นๆ ทำให้มีโอกาสที่ลูกค้าจะซื้อสินค้าตัวอื่นเพิ่มขึ้นด้วย

โดยที่รูปแบบของกฎความสัมพันธ์ที่ได้จะอยู่ในรูปแบบ “If X Then Y” หรือ “If Condition 1 Then Condition 2” โดยที่ X และ Y เกิดขึ้นพร้อมกันในทรานแซกชั่นเดียวกัน หรือเรียกได้ว่า ถ้าเหตุการณ์ X หรือ Condition 1 เกิดขึ้นแล้ว จะเกิดเหตุการณ์ Y หรือ Condition 2 ขึ้นด้วย



ภาพที่ 3.1 Association Rule

กฎที่ได้จาก Association Rule มีตัววัดหลักๆ 2 ตัว คือ ค่า Support Factor (Prevalence) และค่า Confidence Factor (Predictability)

สินค้า	จำนวนทรานแซกชั่น
Bread	4,000
Butter	6,000
Bread, Butter	2,000

ตารางที่ 3.1 แสดงจำนวน Transaction ของสินค้าที่เกิดจากการขายขนมปังและเนยจากชุดข้อมูลทั้งหมด 100,000 รายการ

- ค่า Support คือ ค่าที่แสดงสัดส่วนระหว่างชุดของข้อมูลที่มีทั้งข้อมูล "เหตุ" และ "ผล" เทียบกับจำนวนข้อมูลเหตุการณ์ทั้งหมด โดยทรานแซกชั่นที่จะสนับสนุนกฎ "When X then Y" ถ้ามี Item X และ Y ในกฎเกิดขึ้นในทรานแซคชั่นเดียวกัน จากภาพที่ ค่า Support ที่ได้ คือ 2% ซึ่งได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการนมและขนมปังคู่กัน}}{\text{จำนวนชุดข้อมูลทั้งหมด}}$$

- ค่า Confidence คือ ค่าที่แสดงสัดส่วนระหว่างจำนวนชุดข้อมูลที่มีทั้งข้อมูล "เหตุ" และ "ผล" เทียบกับจำนวนข้อมูลที่มีเฉพาะเหตุการณ์ที่เป็น "เหตุ" ค่า Confidence ที่ได้ คือ 50% ซึ่งได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการนมและขนมปังคู่กัน}}{\text{จำนวนชุดข้อมูลที่มีรายการซื้อขนมปัง}}$$

จากข้อมูลที่เกิดขึ้นพบว่ายังมีตัววัดค่าของเหตุการณ์ที่เกิดขึ้นของแต่ละรายการสินค้า เรียกว่า "Expected Confidence" และ "Life"

- ค่า Expected Confidence คือ ค่าสัดส่วนระหว่างจำนวนทรานแซกชั่นที่สนใจต่อจำนวนทรานแซกชั่นทั้งหมด ซึ่งอาจเป็น "เหตุ" หรือ "ผล" อย่างใดอย่างหนึ่งเทียบกับจำนวนเหตุการณ์ทั้งหมดภายในชุดข้อมูล จากภาพที่ ค่า Expected Confidence ของการซื้อขนมปัง จะเท่ากับ 4% ค่า Expected Confidence ของการซื้อเนย จะเท่ากับ 6%

- ค่าที่แสดงความน่าเชื่อถือของความสัมพันธ์ระหว่างเหตุการณ์ ซึ่งหาได้จากค่าสัดส่วนระหว่างค่า Confidence กับ ค่า Expected Confidence ของจำนวนชุดข้อมูลของ Y จากภาพที่ ค่า Lift จะเท่ากับ 11.1% ซึ่งจะได้กฎว่า "คนที่ซื้อขนมปัง จะซื้อเนยด้วยคิดเป็น 11.1%" ถ้าค่า Confidence ยิ่งมีค่ามากความสัมพันธ์ระหว่างขนมปังกับเนยจะยิ่งมากขึ้นไปด้วย ข้อควรระวังในการวิเคราะห์ข้อมูล คือ ค่าของ Lift ที่ติดลบ หรือน้อยกว่า 1 หมายถึงเหตุการณ์เหล่านั้นไม่มีทางเกิดขึ้นพร้อมกันได้เลย และกฎที่มีค่า Lift มากหรือน้อยเกินไป บางครั้งอาจเป็นกฎที่ไม่เป็นความจริง

3.2 Apriori Algorithm

Apriori เป็นอัลกอลิทึมพื้นฐานในการ mining ชุดข้อมูลของ Association Rule โดย Apriori จะมีหลักการทำงาน 2 ประการ คือ

1. จะทำการหาความถี่ของการเกิดเซตของข้อมูล (Frequent Itemset) โดยที่ความถี่ที่ใช้จะต้องมีค่ามากกว่าหรือเท่ากับค่า Support ที่น้อยที่สุด (Minimum Support) ซึ่งทุกๆ สับเซตของความถี่ของการเกิดเซตข้อมูลจะต้องมีความถี่ของตัวเองด้วย เช่น ถ้า {AB} เป็น frequent itemset ทั้ง {A} และ {B} ควรจะเป็น frequent itemset ด้วย

2. ใช้ความถี่ของการเกิดเซตข้อมูลมาใช้ในการสร้างกฎความสัมพันธ์ โดยกฎที่ได้จะถูกต้องเมื่อมีค่า Confidence มากกว่าหรือเท่ากับค่า Confidence น้อยที่สุด (Minimum Confidence)

ตัวแปรที่จะต้องนำมาพิจารณา คือ

K-Itemset คือ เซตของข้อมูลที่มีจำนวนสมาชิก k ตัว

L_k คือ เซตของ frequent k-itemset ซึ่งทุกเซตจะมีความถี่ในการเกิดมากกว่าหรือเท่ากับค่า Minimum Support แต่ละสมาชิกของเซต จะประกอบด้วย 2 ฟิลด์ คือ itemset และ support count

C_k คือ เซตของ candidate k-itemset เป็นเซตที่ถูกเลือกมาจาก L_k ซึ่งแต่ละสมาชิกของเซตจะประกอบด้วย 2 ฟิลด์ คือ itemset และ support set

D คือ Database ที่แต่ละทรานแซกชั่นเก็บ

t คือ จำนวนทรานแซกชั่นใน Database

3.2.1 การหา Frequent Itemset

ในการหา Frequent Itemset จะเริ่มจากการสแกนข้อมูลในฐานข้อมูลและทำการนับค่า Support สำหรับแต่ละ Item ในแต่ละรอบ โดยเลือกเฉพาะ Item ที่มีค่า Support มากกว่า Minimum Support จากนั้นจะทำการวนลูปนับค่า Support ของ Subset ของ Item ที่ได้จากลูปก่อนหน้า เพื่อหาค่า Candidate Itemset ในตอนจบของแต่ละลูปจะเลือก Candidate Itemset ที่มีค่า Support มากกว่า Minimum Support ไปเป็นตัวตั้งในการหา Frequent Itemset ต่อไป ทำไปเรื่อยๆ จนกว่าจะไม่สามารถหา Frequent Itemset ได้

```
L1 = find_frequent_1-Itemsets(D);
For(k = 2 ; Lk-1 ≠ ϕ ; k++) {
    Ck = apriori_gen (Lk-1 , min_sup);
    For each transaction t ∈ D { // scan D for count
        Ct = subset(Ck , t); //get the subset of t that are candidate
        For each candidate c ∈ Ct
            c.count++ ;
    }
    Lk = {c ∈ Ck | c.count ≥ min_sup}
}
return L = ∪k Lk
```

ภาพที่ 3.2 แสดงอัลกอลิทึมการหา Frequent Itemset

Procedure apriori_gen(L_{k-1} : frequent(k-1) –itemsets ; min_sup : minimum support threshold)

For each itemset $l_1 \in L_{k-1}$

For each itemset $l_2 \in L_{k-1}$

If $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge .. \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] = l_2[k-1])$ then {

$C = l_1 \square l_2$; // join step : generate candidate

If has_infrequebt_subset(c,L_{k-1}) then

Delete c; //prune step : remove unfruitful candidate

Else add c to C_k

}

return C_k

ภาพที่ 3.3 แสดงอัลกอลิทึมการทำงานของ Apriori_gen

การทำงานของ Apriori Algorithm มี 2 ขั้นตอนหลัก คือ

<u>ขั้นที่ 1</u> ขั้นตอนการเชื่อม (Join Step) เป็นขั้นตอนการสร้าง C_k (Candidate itemset) จากการเชื่อมกันของ L_{k-1} กับเซตของตัว L_{k-1} เอง

Join Step

Insert into C_k

Select $p.item_1$, $p.item_2$, … , $p.item_{k-1}$, $q.item_{k-1}$

From L_{k-1} p , L_{k-1} q

Where $p.item_1 = q.item_1$, … , $p.item_{k-2} = q.item_{k-2}$,

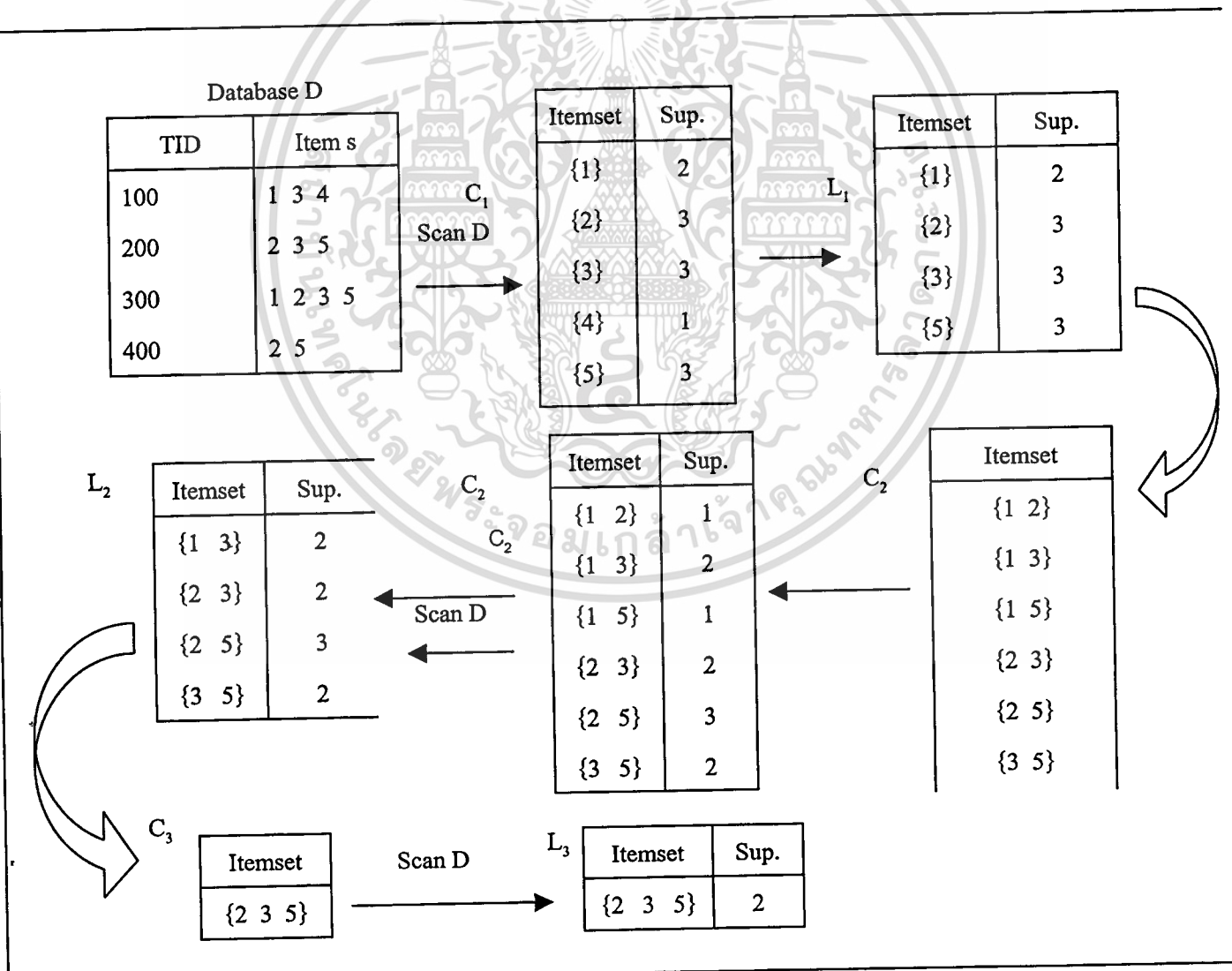
$p.item_{k-1} < q.item_{k-1}$

ภาพที่ 3.4 แสดงอัลกอลิทึมการ Join Step

<u>ขั้นที่ 2</u> เป็นขั้นตอนการตัดทิ้ง (Prune Step) คือ ขั้นตอนของการตัดเซตสมาชิกใน C_k ในขั้นตอนนี้จะทำการลบ Candidate itemset ทุกๆ k-1 subset ของมันที่ไม่ได้อยู่ใน frequent itemset L_{k-1}

```
Pruning Step
Forall itemsets c ∈ C_k do
    Forall (k-1) – subset s of c do
        If (s ∉ L_k-1) then delete c from C_k
```

ภาพที่ 3.5 แสดงอัลกอลิทึมการ Prune Step



ภาพที่ 3.6 ตัวอย่างข้อมูลที่ผ่านการคำนวนจาก Apriori

ตัวอย่างข้อมูล ขั้นตอนการทำงานจะเริ่มจาก

1. นำข้อมูลจาก database มาทำโดยผ่านอัลกอลิทึม ซึ่งอัลกอลิทึมจะนับจำนวนแต่ละรายการจาก transaction ทั้งหมดจะได้เป็นตาราง C1 ออกมา
2. เมื่อได้เซตของ C1 แล้วจะพิจารณาค่าความถี่ในที่นี้คือค่า support ที่มีค่าน้อยกว่าค่า minimum support จะตัดออก จากตัวอย่างจะกำหนดค่า minimum support = 2 จะได้เซต L1
3. นำ L1 มาสร้างเป็น C2 โดย การนำ L1 join L1 จากนั้นนับจำนวนของแต่ละ itemset ใน C2 (หลัง Scan D)
4. หา L2 โดยพิจารณาจากค่า support ที่มากกว่า minimum support ที่กำหนดไว้
5. join L2 เข้าด้วยกัน และจากคุณสมบัติของ Apriori ที่ว่าซับเซตทั้งหมดของ frequent itemset ต้อง frequent ด้วยทำให้ได้ค่าดัง C3
6. สร้าง L3 จาก C3 ที่มีค่า support มากกว่า minimum support
7. จาก L_3 ไม่สามารถหา C4 ได้อีก เพราะเมื่อทำการ join L_3 จะได้ค่าเดิมออกมา ทำให้อัลกอลิทึมสิ้นสุดลง และได้ frequent itemset ทั้งหมดออกมา

3.2.2 การนำ Frequent Itemset มาสร้างเป็นกฎ

การค้นหากฎความสัมพันธ์ มีหลักการว่า การสร้างกฎสำหรับกลุ่มข้อมูล 1 เราจะค้นหา Subset ของ 1 ที่มีค่าไม่ว่างทั้งหมด ซึ่งทุกๆ Subset A ผลลัพธ์ของกฎ a จะอยู่ในรูปแบบ $a \Longrightarrow (1\text{-}a)$ ถ้าสัดส่วนระหว่างค่าสนับสนุน 1 ต่อ ค่าสนับสนุน A ซึ่งค่าน้อยที่สุดเท่ากับค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) จากนั้นเราจะพิจารณา Subset ของ 1 ทั้งหมดเพื่อสร้างกฎร่วมหลายตัวภายหลัง อัลกอลิทึมการทำงานแสดงได้ดังภาพที่ 3.7

forall frequent k-itemset l_k , $k \geq 2$ **do begin**

H_1 = {consequence of rules derived from l_k with one item in the consequent};

call ap-genrules(l_k , H_1);

end

// The genrules generates all valid rules

procedure ap-genrules(l_k : frequent k-itemset , a_m : set of m-item consequence)

A = {(m-1) – itemsets a_{m-1} | $a_{m-1} \subset a_m$

forall $a_{m+1} \in$ A **do begin**

Conf = support(l_k) / support($l_k - a_{m+1}$);

if (conf $\geq$ mincof) **then begin**

output the rule ($a_{m-1} \Rightarrow l_k - a_{m+1}$);

with confidence = conf and support = support(l_k);

if(m-1 > 1) then

call ap-genrules(l_k , a_{m-1}); //to generate rules with subsets of a_{m-1} as the antecedents

end

call ap-genrules(l_k , H_{m+1});

end

ภาพที่ 3.7 แสดงอัลกอลิทึมการ Generate Rule

Itemset	Sup.
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

Possible Rule →

Rule	Con.	Sup.
1 -> 3	1	2
3 -> 1	0.67	2
2 -> 3	0.67	2
3 -> 2	0.67	2
2 -> 5	1	3
5 -> 2	1	3
3 -> 5	0.67	2
5 -> 3	0.67	2

Itemset	Sup
{2 3 5}	2

Possible Rule →

Rule	Con.	Sup.
2 -> 3&5	0.67	2
3 -> 2&5	1	2
5 -> 2&3	0.67	2
3&5 -> 2	1	2
2&5 -> 3	0.67	2
2&3 -> 5	1	2

ภาพที่ 3.9 แสดงผลลัพธ์จากการสร้างกฎ

บทที่ 4

การประยุกต์ใช้ดาต้าไมนิ่งเพื่อหาความสัมพันธ์

เพื่อเป็นการเพิ่มประสิทธิภาพและประสิทธิผลของการพัฒนาโปรแกรม จึงได้มีการศึกษาตัวอย่างของข้อมูลที่จะสามารถนำมาใช้งานกับระบบที่จะทำการศึกษาเกี่ยวกับ Apriori Algorithm ซึ่งในการนำทฤษฎีเกี่ยวกับดาต้าไมนิ่งมาประยุกต์ใช้โดยมีวัตถุประสงค์เพื่อทำการค้นหาความสัมพันธ์ของสิ่งที่เกิดขึ้นและแสดงในรูปแบบของกฎการตัดสินใจ แต่เนื่องจากข้อมูลในฐานข้อมูลเดิมที่มีอยู่ ไม่สามารถนำมาดำเนินการได้ทันที จึงต้องมีกระบวนการต่างๆสำหรับจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมก่อนนำมาใช้งาน ซึ่งจะได้กล่าวดังต่อไป

4.1 แหล่งที่มาของข้อมูล

ข้อมูลที่ต้องนำมาใช้ในการดำเนินการ ได้มาจากฐานข้อมูลการสั่งซื้อสินค้าของลูกค้า ซึ่งเก็บข้อมูลไว้โดยใช้ ฐานข้อมูล Microsoft SQL Server 2000 ดังมีรายละเอียดตารางข้อมูลดังนี้

4.1.1 ตาราง OrderDetails เป็นตารางที่ใช้เก็บข้อมูลทั่วไปของรายการการสั่งซื้อสินค้า

PHYSICAL NAME	DATATYPE	SIZE	DESCRIPTION
OrderID	int	4	เลขที่ใบสั่งซื้อ
ProductID	int	4	รหัสสินค้า
UnitPrice	money	8	ราคาสินค้าต่อหน่วย
Quantity	smallint	2	จำนวนสินค้าที่สั่งซื้อ
Discount	real	4	ส่วนลด

ตารางที่ 4.1 ตารางรายการการสั่งซื้อสินค้า

4.1.2 ตาราง Orders เป็นตารางที่ใช้เก็บข้อมูลการสั่งซื้อ

PHYSICAL NAME	DATATYPE	SIZE	DESCRIPTION
OrderID	int	4	เลขที่ใบสั่งซื้อ
CustomerID	nchar	5	รหัสลูกค้า
EmployeeID	int	4	รหัสพนักงานขาย
OrderDate	datetime	8	วันที่สั่งซื้อ
ShippedDate	datetime	8	วันที่ส่งของ
ShipVia	int	4	รหัสการขนส่ง
ShipName	nvarchar	40	ชื่อผู้ส่ง
ShipAddress	nvarchar	60	ที่อยู่ผู้ส่ง

ตารางที่ 4.2 ตารางข้อมูลการสั่งซื้อ

4.1.3 ตาราง Products เป็นตารางที่ใช้เก็บข้อมูลสินค้า

PHYSICAL NAME	DATATYPE	SIZE	DESCRIPTION
ProductID	int	4	รหัสสินค้า
ProductName	nvarchar	40	ชื่อสินค้า
SupplierID	int	4	รหัสผู้ขายสินค้า
CategoryID	Int	4	รหัสประเภทสินค้า
QuantityPerUnit	nvarchar	20	จำนวนต่อหน่วย
UnitPrice	money	8	ราคาสินค้าต่อหน่วย
UnitsInStock	smallint	2	จำนวนที่เหลือในคลังสินค้า

ตารางที่ 4.3 ตารางข้อมูลสินค้า

4.1.4 ตาราง Categories เป็นตารางที่ใช้เก็บข้อมูลเกี่ยวกับประเภทของสินค้า

PHYSICAL NAME	DATATYPE	SIZE	DESCRIPTION
CategoryID	int	4	รหัสประเภทสินค้า
CategoryName	nvarchar	15	ชื่อประเภทสินค้า
Description	ntext	16	รายละเอียด

ตารางที่ 4.4 ตารางข้อมูลเกี่ยวกับประเภทสินค้า

4.2 การคัดเลือกข้อมูล

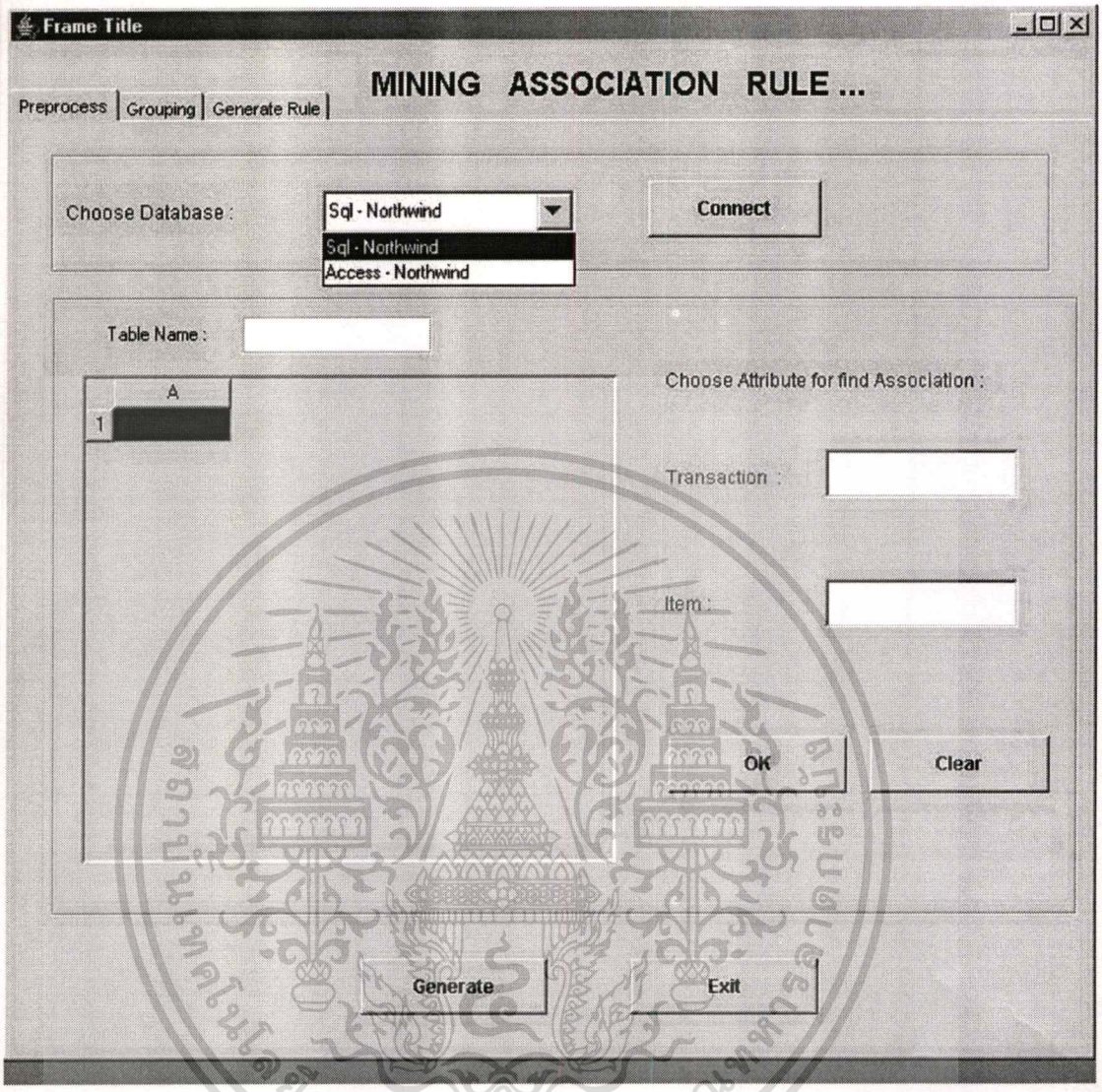
การคัดเลือก Attributes ต่างๆเพื่อนำมาใช้วิเคราะห์เป็นงานส่วนที่มีความสำคัญมาก ทั้งนี้เพราะหากเลือก Attribute ที่ไม่มีความสำคัญหรือไม่เหมาะสมก็จะมีผลต่อกฎที่จะเกิดขึ้นได้ ในการหาความสัมพันธ์โดยใช้หลักการของ Association จะใช้ข้อมูลเพียง 2 column ในตัวอย่างนี้จะทำการหาความสัมพันธ์ระหว่างรายการการขายสินค้า โดยผู้ใช้สามารถจัดกลุ่มของสินค้าสำหรับใช้วิเคราะห์หาความสัมพันธ์ในลักษณ์ที่เป็น Transaction

4.3 ขั้นตอนการทำงานของระบบ

การทำงานของระบบจะเป็นไปตามกระบวนการของ Data Mining ดังนี้

4.3.1 ขั้นตอนการดึงข้อมูล

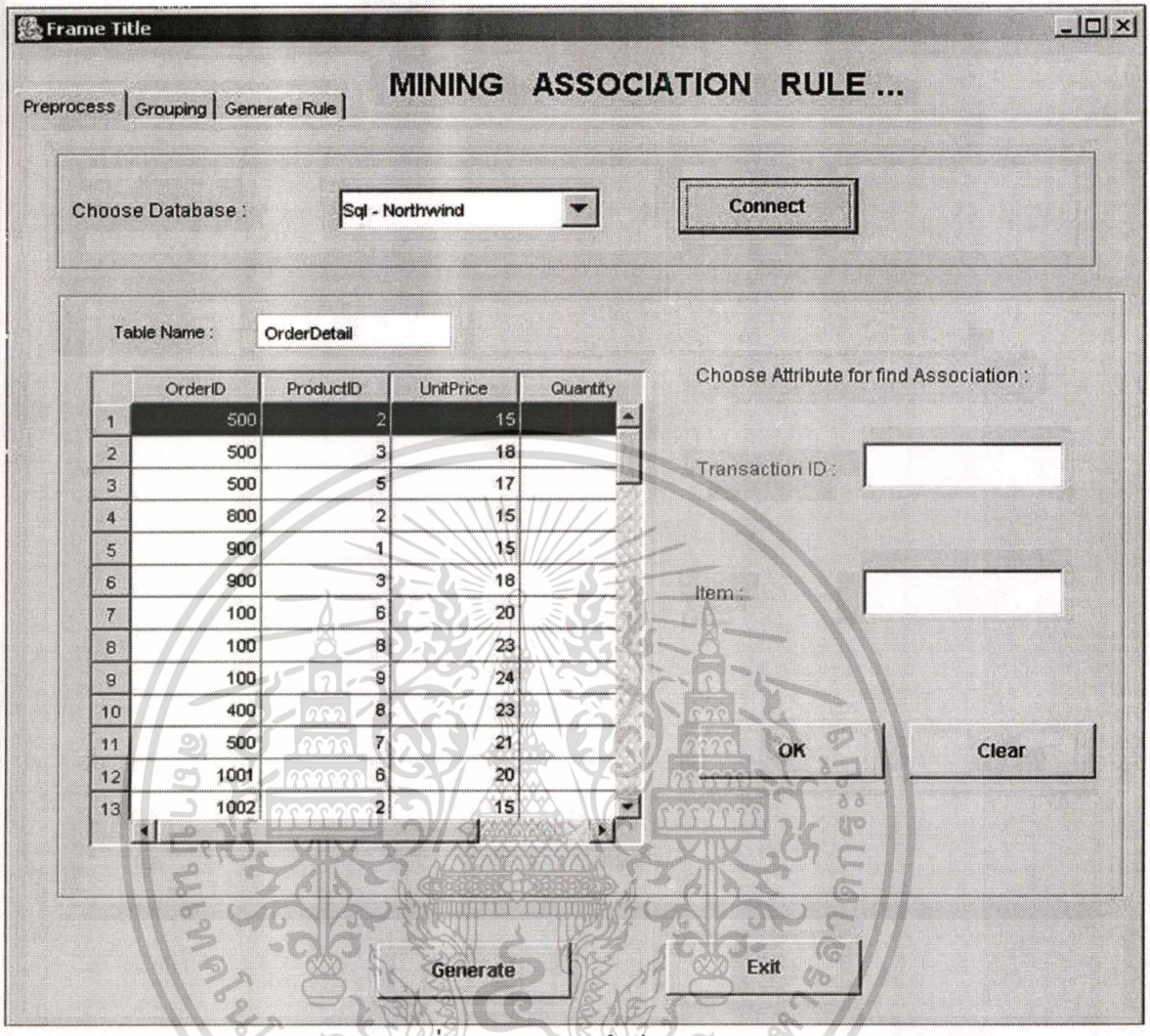
เมื่อเข้าสู่ระบบ จะปรากฏหน้าจอหลัก ดังภาพที่ 4.1



ภาพที่ 4.1 หน้าจอแสดงหน้าจอหลัก

จะมี ComboBox ให้เลือกว่าจะติดต่อกับฐานข้อมูลที่เป็น Sql หรือ Access เมื่อกดปุ่ม

จะทำการติดต่อกับฐานข้อมูลประเภทต่างๆ ซึ่งจะแสดงดังภาพที่ 4.2

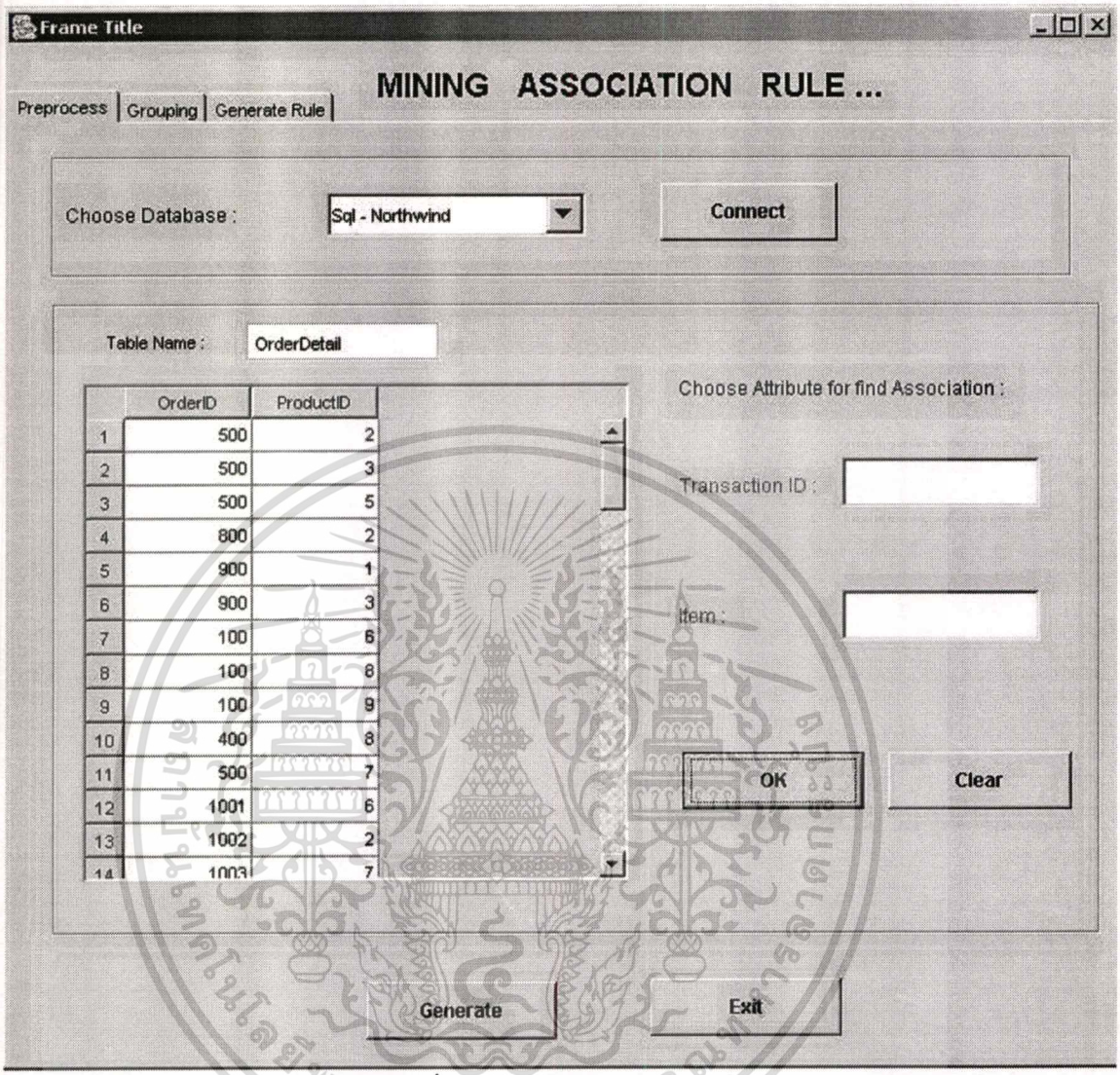


ภาพที่ 4.2 แสดงการดึงข้อมูล

4.3.2 ขั้นตอนการเลือกข้อมูล

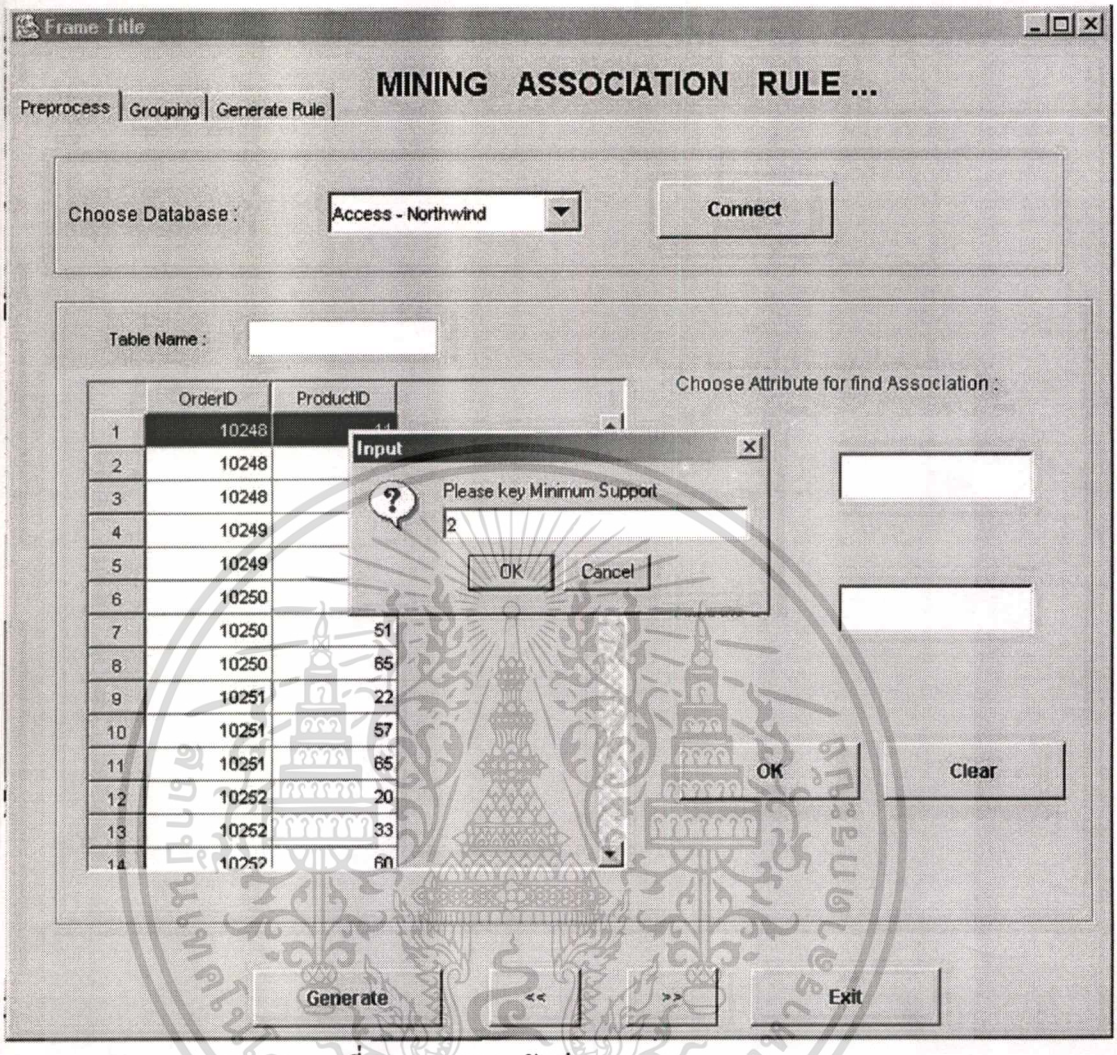
หลังจากนั้นผู้ใช้สามารถใส่ชื่อ Attribute ที่ต้องการหาความสัมพันธ์ซึ่งหลักการของ Association จะให้เลือกแค่ 2 Attribute ที่สัมพันธ์กันในระบบนี้จะเลือก Attribute OrderID และ ProductID เมื่อเลือก Attribute ที่ต้องการแล้วกดปุ่ม 

ตารางจะแสดงข้อมูลที่มีเฉพาะ Attribute ที่เลือก ดังภาพที่ 4.3

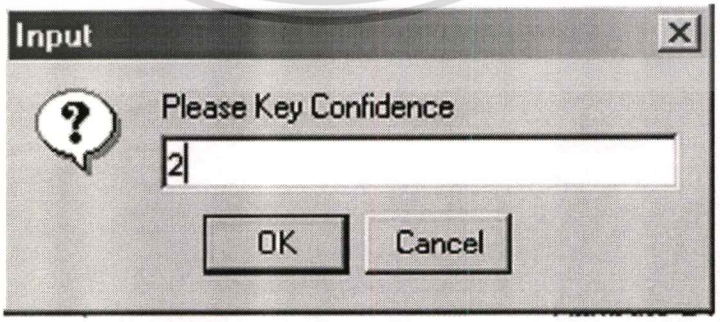


ภาพที่ 4.3 แสดงการเลือกข้อมูล

เมื่อได้ Attribute ที่ต้องการหาความสัมพันธ์เรียบร้อยแล้ว จะสามารถเข้าสู่การทำงานของ Apriori Algorithm โดยทำการกดปุ่ม Generate จะให้ผู้ใช้งานใส่ค่า Minimum support และ Confidence ที่ต้องการ ซึ่งจะรับค่าเป็นเปอร์เซนต์ ดังภาพ



ภาพที่ 4.4 แสดงการรับค่า Minimum Support



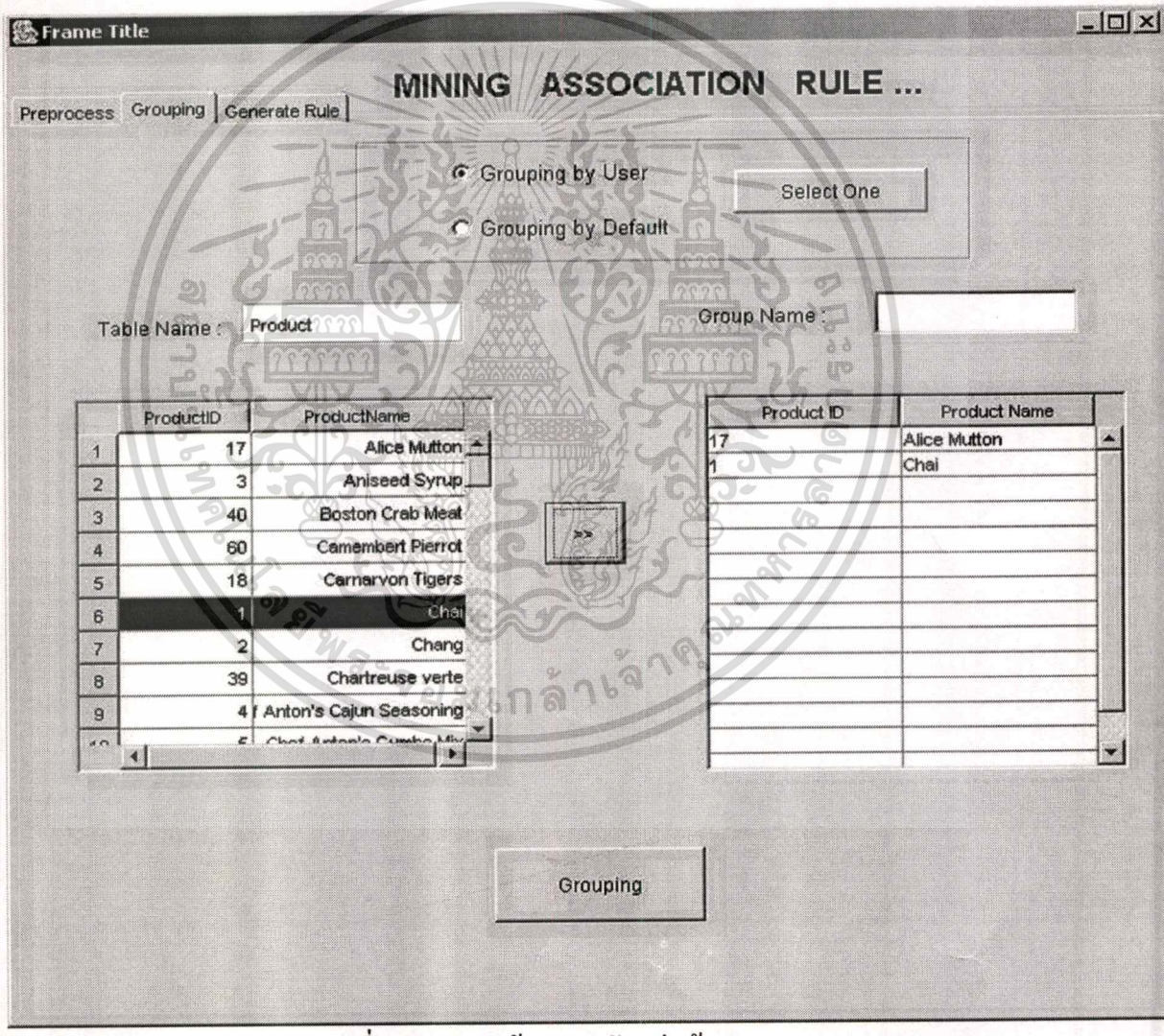
ภาพที่ 4.5 แสดงการรับค่า Confidence

4.3.3 การจัดการกับค่าที่หายไป (Missing Value)

เนื่องจากตัว algorithm ที่เขียนมีความสามารถในการจัดการกับค่า Missing value โดยจะทำการลบเรคคอร์ดที่ประกอบด้วยค่าว่างนี้ทิ้งไป

4.3.4 การจัดกลุ่มข้อมูล

สำหรับข้อมูลที่มีความถี่ที่เกิดขึ้นน้อย สามารถจัดกลุ่มข้อมูลรวมกันกับค่าอื่นได้โดยเลือกที่ Tap ของ Grouping แสดงดังภาพที่ 4.6



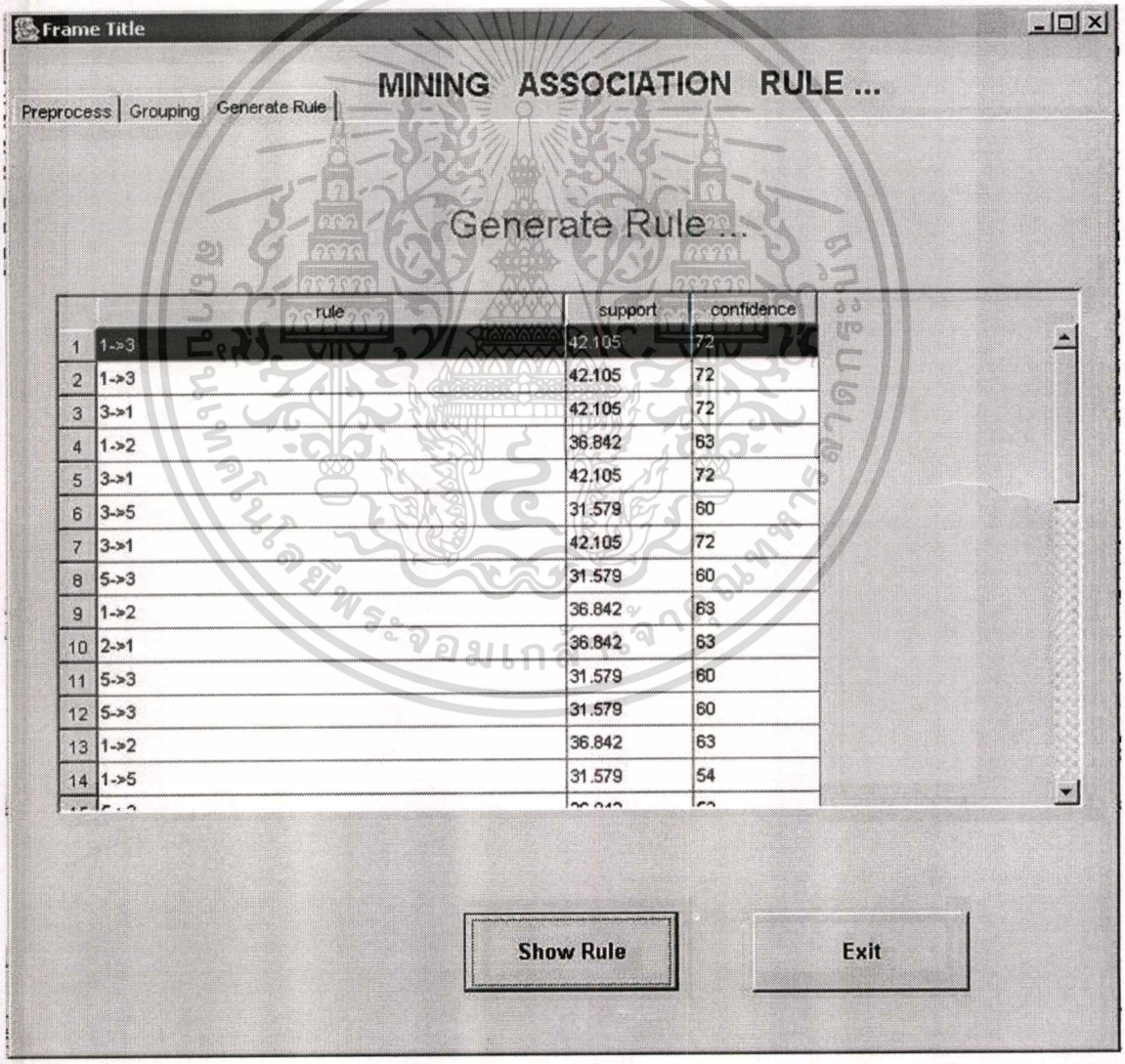
ภาพที่ 4.6 แสดงหน้าจอการจัดกลุ่มข้อมูล

ในหน้าจอการทำการ Grouping ข้อมูลผู้ใช้สามารถเลือกทำได้ 2 แบบ คือ Grouping by user คือ สามารถเลือกข้อมูลจากอีกฝั่งหนึ่งไปยังอีกฝั่งหนึ่งเพื่อทำการ grouping ได้ หรือ แบบที่ 2

จะทำการ Grouping by default คือจะทำการ grouping จากข้อมูลที่ได้จัดกลุ่มไว้แล้วจากฐานข้อมูล เมื่อทำการเลือกประเภทของการ grouping แล้ว ระบบจะทำการดึงข้อมูลขึ้นมาแสดง เมื่อทำการกดปุ่ม Grouping ระบบจะทำการจัดกลุ่มข้อมูลให้ และจะทำการ refresh หน้าจอต้องขั้นตอน preprocess ให้เป็นข้อมูลที่ได้ทำการ grouping

4.3.5 การสร้างกฎ

หลังจากโปรแกรมทำการวิเคราะห์ข้อมูลเรียบร้อยแล้ว จะแสดงผลลัพธ์จากการสร้างกฎ ดังภาพที่ 4.7



ภาพที่ 4.7 แสดงการสร้างกฎของข้อมูล

4.4 วิเคราะห์ผลการดำเนินงาน

ผลลัพธ์ที่ได้จากการทดสอบ โดยเลือกวิเคราะห์รายการขายสินค้า จำนวน 2155 รายการและกำหนดค่า Minimum Support เท่ากับ 0.02 และค่า Minimum Confidence เท่ากับ 0.02 พบความสัมพันธ์จากรายการขายสินค้าดังนี้

1. ลูกค้าที่ซื้อสินค้าชนิดที่ 1 และจะซื้อสินค้าอีกชนิดหนึ่งพร้อมกันด้วยค่า Support ในช่วง 0.09 – 0.18 และ Confidence อยู่ในช่วง 0.05 – 0.08
2. ลูกค้าที่ซื้อสินค้าชนิดที่ 1 และชนิดที่ 2 จะซื้อสินค้าชนิดที่สามพร้อมกันด้วยค่า Support เท่ากับ 0.09 และ Confidence อยู่ในช่วง 0.08 – 1

จากการทดลองดังกล่าวพบความสัมพันธ์ที่น่าสนใจมากมาย ซึ่งต้องอาศัยกระบวนการตัดสินใจว่าความสัมพันธ์ไหนจะเป็นประโยชน์ต่อธุรกิจได้ จากตัวอย่างนี้สามารถที่จะเพิ่มค่า Minimum Support หรือค่า Minimum Confidence เพื่อให้ได้กฎที่น้อยที่สุดได้ ซึ่งถ้าไม่สามารถหากฎความสัมพันธ์ได้อาจจะใช้หลักการจัดกลุ่มของข้อมูลมาช่วยในการหาความสัมพันธ์ต่อไปได้

บทที่ 5
บทสรุป

โครงการพัฒนาระบบนี้เป็นโครงการที่จัดทำขึ้นมาเพื่อนำเสนอให้เห็นถึงประโยชน์ของการนำเอาทฤษฎีดาต้าไมนิ่งมาใช้เพื่อเพิ่มประสิทธิภาพในการหาความสัมพันธ์ของข้อมูลในรูปแบบต่างๆ ของการซื้อขายสินค้าของลูกค้า เพื่อนำผลลัพธ์ที่ได้ไปใช้วางแผนทางการตลาดต่อไป

5.1 สรุปหลักการที่ใช้ในระบบ

ดาต้าไมนิ่งเป็นกระบวนการที่ใช้เพื่อค้นหาข้อมูลที่มีประโยชน์เพื่อนำมาช่วยในการตัดสินใจ เทคนิคที่ใช้ในการหาความสัมพันธ์ของข้อมูลได้นำเอาเทคนิคของ Association มาใช้ในการพัฒนาระบบโดยใช้หลักการของ Apriori Algorithm ซึ่งเป็นอัลกอลิทึมหนึ่งของ Link Analysis ใช้ในการหาความสัมพันธ์ที่เกิดขึ้นโดยอัลกอลิทึมนี้จะทำการนับค่าความสัมพันธ์ที่เกิดขึ้นร่วมกันโดยจะนำมาเปรียบเทียบกับค่า Minimum Support และค่า Minimum Confidence ถ้ามีค่าน้อยกว่าค่าที่กำหนดจะทำการตัดความสัมพันธ์นั้นออกไปเพื่อให้ได้ความสัมพันธ์ที่เป็นไปได้มากที่สุด หลังจากนั้นจะนำความสัมพันธ์ที่ได้มาสร้างเป็นกฎความสัมพันธ์เพื่อนำไปวิเคราะห์ว่ากฎที่ได้มีความสมเหตุสมผลกันหรือไม่ เพื่อดูถึงความเป็นไปได้ในการนำกฎที่ได้ไปประยุกต์ใช้เพื่อสร้างประโยชน์ให้กับองค์กรต่อไป

5.2 สรุปกระบวนการในการทำงาน

โครงการนี้เริ่มต้นจากการเลือก Attribute ที่มีความสัมพันธ์กัน โดยจากการวิเคราะห์ Attributes ที่เกี่ยวข้อง ทำให้การคัดเลือก Attributes ที่คาดว่าจะเป็นปัจจัยที่มีความสำคัญขึ้นมา จากนั้นจึงนำข้อมูลที่ได้ไปเข้าอัลกอลิทึม เพื่อสร้างเป็นกฎความสัมพันธ์ โดยส่วนของการแสดงผลจะแสดงในรูปแบบกฎ IF-THEN จากนั้นจึงนำเอาผลลัพท์ที่ได้ไปทำการทดสอบกับข้อมูลอื่นที่ไม่ใช่ข้อมูลที่ใช้ในการสร้างกฎความสัมพันธ์เพื่อทดสอบความถูกต้องของอัลกอลิทึมที่ได้

5.3 สรุปผลการพัฒนาโปรแกรม

จากการพัฒนาโปรแกรมได้ทำให้ผู้พัฒนาได้เข้าใจถึงหลักการของ Data Mining เพื่อนำมาใช้ประโยชน์ ผลจากการทดลองนั้นมีความถูกต้องในระดับหนึ่ง ซึ่งการทำงานของ Data Mining

นั้นขึ้นอยู่กับการเตรียมข้อมูลที่ดีซึ่งถ้าขั้นตอนนี้กระทำไม่ดีแล้ว ผลการทดสอบก็จะได้ผลไม่ดีเท่าที่ควร ทั้งนี้ข้อมูลจะเป็นประโยชน์ได้ต้องอาศัยกระบวนการต่างๆ ที่ได้กล่าวมาแล้ว ถ้าขั้นตอนใดขั้นตอนหนึ่งผิดจะส่งผลให้ผลลัพธ์ที่ได้ผิดพลาดตามไปด้วย

5.4 ข้อเสนอแนะ

เนื่องจากอัลกอลิทึมที่ใช้ในการพัฒนาต้องทำการนับข้อมูลทุกๆ ครั้งซึ่งทำให้ช้าในการประมวลผล ถ้าข้อมูลมีปริมาณไม่มากอัลกอลิทึมนี้ก็จะสามารถใช้งานได้ แต่ถ้ามีปริมาณข้อมูลที่มากๆ ควรจะใช้อัลกอลิทึมที่มีประสิทธิภาพมากขึ้น

บรรณานุกรม

Cabena and Hadjinian and Sradler .et.al., 1998 **Discovery Data Mining From Concept to Implementation**. Prentice Hall, New Jersey

Jiawei Han, Micheline Kamber. 2000. **Data Mining: Concept and Techniqies**. Morgan Kaufmann Publishers.

Karuna Pande Joshi. **Analysis of Data Mining Algorithms**. [Online]. Available : http://userpages.umbc.edu/~kjoshi/data-mine/proj_rpt.htm

Rakesh Agrawal and Ramakrishnan Srikant. 1994 .**Fast algorithms for mining association rules** In Proc. of the VLDB Conference, Santiago, Chile.

ประวัติผู้เขียน

ชื่อ-นามสกุล : นางสาวดวงกมล มหาวีระ

วันเดือนปีเกิด : 12 เดือนมีนาคม พ.ศ. 2522

สถานที่เกิด : กรุงเทพมหานคร

ประวัติการศึกษา

- สำเร็จการศึกษาระดับประถม โรงเรียนถนอมพิศวิทยา
- เข้ารับการศึกษาระดับมัธยม สายวิทย์-คณิต โรงเรียนสตรีวิทยา 2
- สำเร็จการศึกษามัธยมปลาย จากศูนย์การศึกษานอกโรงเรียน
- สำเร็จการศึกษาระดับปริญญาตรี คณะวิทยาศาสตร์ สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศรีนครินทร์วิโรฒ (ประสานมิตร)
- กำลังศึกษาระดับปริญญาโท คณะเทคโนโลยีสารสนเทศ สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีเจ้าคุณทหารลาดกระบัง

ประวัติการทำงาน

- การประปานครหลวง (ปี พ.ศ. 2544)

ตำแหน่ง วิทยากร 3