การพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการใช้บริการธนาคารโดยใช้ Association Rule

Association Rule Discovery for Banking Product and Service

โดย

นางสาว พัชรินทร์ อุทัยจรัสรัศมี

รหัส 44067068

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
ภาคเรียนที่ 2 ปีการศึกษา 2546
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลในการใช้บริการธนาคาร โดยใช้ Association Rule
นักศึกษา	น.ส พัชรินทร์ อุทัยจรัสรัศมี
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

สภาพธุรกิจในปัจจุบันมีการแข่งขันกันสูงมาก ดังนั้นแต่ละธุรกิจต้องพยายามหาวิธีการเพื่อเพิ่มส่วนแบ่งตลาดของตนเอง โดยต้องรักษาฐานลูกค้าเดิมไว้ การกำหนดกลยุทธ์ทางการตลาดเพื่อวางแนวทางการดำเนินธุรกิจจึงเป็นสิ่งที่มีความสำคัญมาก การทำ Cross selling เป็นอีกกลยุทธ์หนึ่งในขยายฐานลูกค้าจากเดิมออกไปโดยการนำเสนอขายผลิตภัณฑ์ต่างๆให้กับลูกค้าในกลุ่มที่เหมาะสม แต่การได้มาซึ่งกลยุทธ์นี้ต้องอาศัยข้อมูลที่มีอยู่ในองค์กรนำมาใช้ให้ประโยชน์ ซึ่งข้อมูลที่มีเก็บอยู่ในองค์กรมีอย่างมากมาย จึงได้มีการนำเอา เทคนิค Data mining เข้ามาช่วยในการวิเคราะห์หารูปแบบ หรือ information ที่มีประโยชน์ที่ซ่อนอยู่ภายในข้อมูลเหล่านั้นนำมาใช้ให้เกิดประโยชน์ได้ โครงการนี้จะนำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการใช้ผลิตภัณฑ์และบริการของธนาคาร เพื่อช่วยหาความสัมพันธ์ของการใช้ผลิตภัณฑ์และบริการของธนาคารกับลูกค้า โดยใช้ Apriori Algorithm ซึ่งเป็นอัลกอริทึมพื้นฐานในการกฎความสัมพันธ์ของข้อมูลใช้ใน Association Rules ซึ่งเป็นเทคนิคหนึ่งของ Link Analysis ของ Data mining

Title	Association Rule Discovery for Banking Products and Services
Student	Miss Patcharin Uthaicharatratsame
Advisor	Asst.Prof. Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

ABSTRACT

In today's business world , there have been a number of trends and cultural changes, which have led to intensified competition. Thus each business sector must aim at developing revenue generating strategies to increase customer 'share of wallet' through cross-selling method (finding the right products for the right customers by creating offerings that match their needs and activities) to existing customer base. Indeed , to do this , it calls for a complete sets of data existed within an organization together with the data mining techniques to help analyze the marketing opportunities within such data.

This project will propose the process and methodology of system enhancement in order to analyze the correlation of data between product usage and banking services. In analyzing this relationship more effectively, 'Apriori Algorithm' , the basic logarithm used in Association Rules,is implemented as one of the reliable techniques of Data Mining Link Analysis.

กิตติกรรมประกาศ

โครงการพัฒนาระบบฉบับนี้ สำเร็จลงได้ด้วยความช่วยเหลือและสนับสนุนจากท่านอาจารย์และบุคคลต่างๆดังต่อไปนี้

ขอกราบขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาที่คอยแนะนำ ให้คำปรึกษาความช่วยเหลือ และหาแนวทางการแก้ไข เมื่อเกิดปัญหาในการจัดทำโครงการ เป็นอย่างดีเสมอมา

ผศ.ดร.อาริต ธรรมโน และ ผศ.ดร.โชติพัชร์ ภรณวลัย คณะกรรมการสอบ ที่สละเวลามาเป็นกรรมการให้กับการสอบโครงการพัฒนาระบบนี้ รวมทั้งขอขอบคุณ คณาจารย์คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ได้ประสิทธิประสาทความรู้ต่างๆให้ผู้จัดทำมีความเข้าใจ และสามารถเรียบเรียงจัดทำโครงการนี้ได้สำเร็จลุล่วงไปด้วยดี

ขอบคุณ คุณสมโภชน์ กิ่งกุ่มกลาง ที่ให้คำแนะนำ และความช่วยเหลือในด้านการพัฒนาระบบ

ขอกราบขอบพระคุณ คุณแม่และ พี่ชาย ที่คอยห่วงใยดูแล คอยให้กำลังใจแก่ผู้จัดทำและการสนับสนุนที่ดีโดยตลอดมา ตลอดจนเพื่อนๆพี่ๆทุกๆคน ที่ให้คำแนะนำ และความช่วยเหลือในด้านต่างๆด้วยดีเสมอมา

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญภาพ

หน้า

สารบัญภาพ (ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

ในสภาวะธุรกิจปัจจุบันมีการแข่งขันกันสูงมาก ทำให้แต่ละองค์กรต้องหากลยุทธ์ต่างๆในการตลาดเพื่อที่จะวางแผนแนวทางการดำเนินธุรกิจเพื่อให้ธุรกิจมีความได้เปรียบในเชิงการแข่งขันและสามารถได้รับส่วนแบ่งตลาดที่มากขึ้น สร้างความเป็นปึกแผ่นให้กับองค์กร และนำไปสู่การทำกำไรที่มากขึ้น ตลอดจนสร้างความพึงพอใจให้กับลูกค้าให้ได้มากที่สุด ซึ่งการวางแผนกลยุทธ์ทางการตลาดนั้นต้องอาศัยข้อมูลที่มีอยู่ในองค์กร ซึ่งข้อมูลเหล่านั้นมีอยู่อย่างมากมายโดยจัดเก็บไว้ในคลังข้อมูล (Data Warehouse) ขององค์กร อย่างไรก็ตามข้อมูลที่มีอยู่ในคลังข้อมูลนั้นอาจจะไม่เพียงพอ และยากต่อการทำการวิเคราะห์เนื่องจากมีข้อมูลจำนวนมาก ดังนั้นจึงได้นำเทคนิค Data mining เข้ามาใช้ในการวิเคราะห์หาสารสนเทศที่ซ่อนอยู่ภายในข้อมูลจำนวนมากนั้น รวมถึงวิเคราะห์หาความสัมพันธ์ของข้อมูล เพื่อศึกษาแนวโน้มและพฤติกรรมของข้อมูลในอนาคตได้ เพื่อใช้เป็นแนวทางในการกำหนดเป้าหมายในการดำเนินงานขององค์กร

1.2 วัตถุประสงค์

เพื่อนำเอาเทคนิคของ Data mining ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลต่างๆในการใช้บริการของลูกค้า วัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้นี้มาใช้ในการวางแผนกลยุทธ์ทางการตลาดได้อย่างมีประสิทธิภาพ รวมทั้งเป็นเป็นแนวทางให้กับผู้บริหารเพื่อใช้ในการสนับสนุนการตัดสินใจ และวางแผนการทำรายการส่งเสริมการตลาดได้อย่างถูกต้องและเหมาะสมกับความต้องการของลูกค้าให้มากที่สุด

1.3 ขอบเขตการดำเนินงาน

โครงงานนี้เป็นการศึกษาการนำเอาเทคนิค Data mining มาประยุกต์ใช้ โดยอาศัยหลักการของ Link Analysis ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลในฐานข้อมูลการใช้บริการของธนาคารพาณิชย์ โดยมีขอบเขตการหาความสัมพันธ์ของผลิตภัณฑ์ด้านเงินฝาก เช่น ออมทรัพย์ เดินสะพัด หรือ เงินฝากระยะยาว เป็นต้น .บัตรประเภทต่างๆ เช่น ATM . บัตร Credit และ บัตร Debit รวมถึงการใช้บริการผ่านทางช่องทางต่างๆเช่น ผ่านทาง Internet เป็นต้น

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1.) ศึกษาแนวคิด และทฤษฎีที่เกี่ยวข้องการ Data mining เพื่อนำมาประยุกต์ใช้ในการพัฒนาระบบงาน

2.) ศึกษาทฤษฎี Association Rule และ Algorithm ที่เกี่ยวข้องเพื่อนำมาประยุกต์ใช้ในระบบงาน

3.) ศึกษา และเก็บรวบรวมข้อมูลที่เกี่ยวข้องที่มีอยู่ภายในองค์กร

4.) วิเคราะห์ ออกแบบ และพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล

5.) สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการที่ได้ศึกษาแนวคิด ทฤษฎีของ Data mining และพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการใช้บริการของธนาคาร คาดว่าจะได้ให้ประโยชน์แกผู้ค้นคว้า และผู้ที่สนใจ ดังนี้

1.) เพื่อให้เข้าใจถึงแนวคิด กระบวนการการทำ Data mining ตลอดจนเทคนิคต่างๆของ Data mining ที่เหมาะสมกับลักษณะงานหรือข้อมูลแต่ละประเภทได้

2.) เพื่อให้เข้าใจถึงแนวคิด และขั้นตอนการนำ Data mining มาใช้ในการวิเคราะห์หาความสัมพันธ์ของข้อมูล

3.) เพื่อเป็นแนวทางในการนำเทคนิค Data mining ไปใช้กับองค์กรธุรกิจต่างๆได้

4.) เพื่อให้สามารถนำผลที่ได้จากการทำ Data mining ไปใช้กับการกำหนดแนวทางการดำเนินธุรกิจได้

ในบทนี้จะกล่าวถึงความเป็นมาของปัญหาในการนำเอา Data mining เข้ามาช่วยในการหาสารสนเทศที่ซ่อนอยู่ในข้อมูล วัตถุประสงค์ ขอบเขตและขั้นตอนการดำเนินงาน รวมไปถึงประโยชน์ที่จะได้รับจากการพัฒนาระบบโครงงานนี้

บทที่ 2

Data Mining

ในปัจจุบันธุรกิจต่างๆมีการเพิ่มขึ้นของข้อมูลในปริมาณที่มากแต่ไม่สามารถที่จะสร้างประโยชน์ได้จากข้อมูลเหล่านั้น ถึงแม้จะมีการนำหลักการ Datawarehouse เข้ามาใช้ในการเก็บข้อมูลสำหรับการวิเคราะห์ และมีเครื่องมือ OLAP(On –Line Analyst Processing) สำหรับการนำเสนอข้อมูลให้กับผู้บริหารในลักษณะมุมมองต่างๆแล้วก็ตาม การใช้ประโยชน์จากข้อมูลยังไม่เต็มที่ ข้อมูลที่ได้เป็นเพียงข้อมูลพื้นฐานเท่านั้น จึงได้มีการนำหลักการ Data Mining เข้ามาใช้ในการหา Information ที่อยู่ภายในข้อมูลเหล่านั้น โดยในบทนี้จะกล่าวถึงนิยาม ขั้นตอนการทำงาน และ เทคนิคต่างๆของ Data Mining

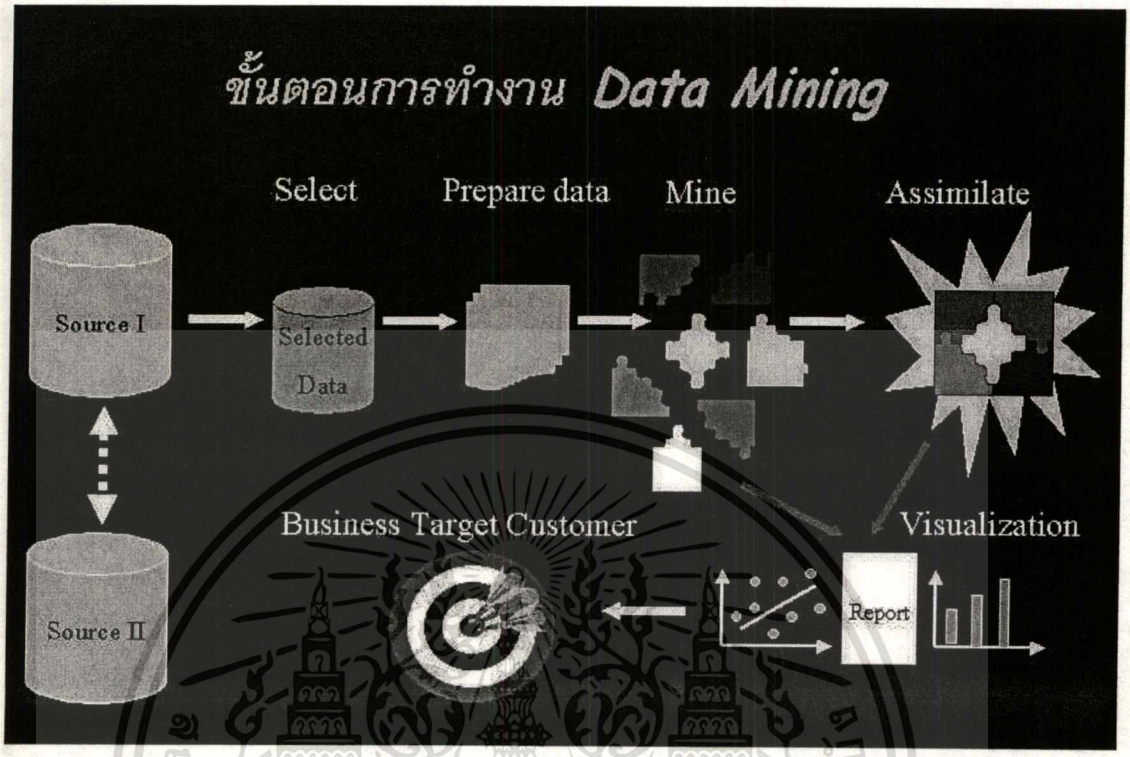
2.1 นิยามของ Data Mining

Data mining เป็นขั้นตอนในการดึง Information ออกมาจากข้อมูล ไม่ว่าจะเก็บในฐานข้อมูล Datawarehouseฯลฯ โดย information ที่ได้นี้ จะต้องเป็น สิ่งที่ไม่รู้มาก่อน (Unknown) , มีความถูกต้อง(Valid) และ สามารถนำ information ไปใช้งานในทางปฏิบัติได้ (Actionable) ซึ่งสิ่งเหล่านี้จะเป็นสิ่งที่ใช้วัดความสำเร็จของการทำ Data mining นอกจากนี้ในปัจจุบันความก้าวหน้าทางเทคโนโลยี ทำให้การทำ Data mining ทำได้เร็วยิ่งขึ้นด้วย

2.2 ขั้นตอนการทำ Data Mining

ขั้นตอนการทำ Data mining เป็นการสร้างแบบจำลองของกลุ่มข้อมูล เพื่อสร้างความเข้าใจในรูปแบบ ความเกี่ยวข้องกันของกลุ่มเข้ามูลเพื่อใช้ในการวิเคราะห์ต่อไป ซึ่งสามารถสรุปขั้นตอนการทำ Data mining ได้ 6 ขั้นตอนดังนี้

1. การกำหนดวัตถุประสงค์ทางธุรกิจ(Business Objective Determination)
2. การเตรียมข้อมูล (Data Preparation)
3. การทำ Data mining (Data Mining)
4. การทำการวิเคราะห์ผล (Analysis of Result)
5. การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of Knowledge)



รูปที่ 2.1 ขั้นตอนการทำ Data Mining

ขั้นตอนที่ 1 : กำหนดวัตถุประสงค์ทางธุรกิจ

การกำหนดวัตถุประสงค์ของการทำ Data mining ว่าจะทำเพื่ออะไร โดยวัตถุประสงค์ที่กำหนดจะต้องมีความชัดเจน มิฉะนั้นแล้วจะทำให้เกิดการตีความหมายผิดพลาด ซึ่งอาจจะทำให้ผลที่ได้ไม่ตรงกับความต้องการทางธุรกิจ นอกจากนี้จะต้องพิจารณาว่า ปัญหาที่กำหนดวัตถุประสงค์ขึ้นมามีสามารถใช้ Data mining ในการแก้ไขได้หรือไม่ เพราะทุกปัญหาไม่สามารถแก้ได้ด้วย Data mining ดังนั้นการกำหนดวัตถุประสงค์ของการทำ Data mining จึงเป็นสิ่งที่จำเป็นต้องพิจารณาอย่างรอบคอบ เพื่อนำไปสู่การแก้ปัญหาได้อย่างตรงจุดได้

ขั้นตอนที่ 2 : การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลเป็นขั้นตอนที่มีความสำคัญมาก จะเป็นขั้นตอนที่ใช้เวลามากที่สุด โดยวัตถุประสงค์หลักคือ เพื่อให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้น มีความถูกต้องเหมาะสม โดยในขั้นตอนนี้จะประกอบด้วย ขั้นตอนย่อย 3 ขั้นตอน ดังนี้

1.) การเลือกข้อมูล(Data Selection)

เลือกข้อมูลที่จะทำการ Mining โดยจะต้องมีการระบุข้อมูลว่าต้องนำข้อมูลจากแหล่งใดภายในหรือภายนอกองค์กร เช่น Data warehouse , Point of Sale , Call center record เป็นต้น และ

ต้องเลือกให้ตรงกับวัตถุประสงค์ที่ตั้งไว้ การเลือกข้อมูลจะต้องมีความเข้าใจในชนิดของข้อมูล ค่าที่เป็นไปได้ รูปแบบและลักษณะต่างๆของข้อมูล โดยข้อมูลจะแบ่งออกเป็น 2 ลักษณะ ได้แก่

- ข้อมูลแบบ Categorical
  - Nominal : ตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น เพศ (ชาย หญิง)
  - Ordinal : ตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น ลำดับของสินค้า(ดี.ปานกลาง, ไม่ดี)
- ข้อมูลแบบ Quantitative
  - Continuous : ค่าที่เก็บเป็นเลขจำนวนจริง หรือเป็นค่าต่อเนื่อง เช่น จำนวนเงิน
  - Discrete : ค่าที่เก็บเป็นเลขจำนวนเต็ม เช่น ข้อมูลจำนวนบุตร

นอกจากการทำความเข้าใจกับชนิดของข้อมูลแล้ว สิ่งสำคัญที่ต้องพิจารณาในการจัดเตรียมข้อมูล ได้แก่

1. ระดับข้อมูลที่ต้องการ : ข้อมูลที่เก็บอยู่ในฐานข้อมูลมีหลายระดับ ตั้งแต่ระดับรายละเอียด จนมาถึงระดับผลสรุป (Summary) ทั้งนี้การจะใช้ข้อมูลระดับใดขึ้นกับวัตถุประสงค์ในการวิเคราะห์นั้นๆ เช่น

- การวิเคราะห์การใช้โทรศัพท์ ของบริษัทแห่งหนึ่ง หากกำหนดวัตถุประสงค์เพื่อศึกษาความหนาแน่นของเครือข่าย เน้นไปที่รูปแบบพฤติกรรมการใช้โทรศัพท์ของลูกค้า ดังนั้นสิ่งที่สนใจคือสิ่งที่เกี่ยวกับสิ่งที่ควบคุมโดยลูกค้า เช่น เบอร์โทรศัพท์ต้นทาง เบอร์โทรศัพท์ปลายทาง เวลา เวลาที่ใช้โทรแต่ละครั้ง เป็นต้น ไม่ใช่ข้อมูลการเคลื่อนย้ายของอิเล็กตรอนไปยังสายภายในวงจรโทรศัพท์
- ข้อมูลที่ไม่ใช่ข้อมูลสรุปบางครั้งอาจจะมีปริมาณที่มากเกินไป ดังเช่นในกรณีของ Market basket analysis ที่ซึ่งจำนวนที่เกิดจากการรวมกันของสินค้าภายในร้านค้าขายปลีกในระดับ stack-keeping unit (SKU)ที่มีปริมาณมากนั้น ดังนั้นการรวมกลุ่มโดยให้สิ่งที่เป็นสินค้าประเภทเดียวกันไว้ในกลุ่มเดียวกัน ทำให้การ combination กัน เพื่อหาความสัมพันธ์ระหว่างข้อมูลมีปริมาณข้อมูลที่น้อยลง

2. ความไม่สอดคล้องของข้อมูลที่มาจากหลายแหล่งข้อมูล : ในบางครั้งในการทำ Data mining ต้องอาศัยข้อมูลจากหลายแหล่ง ซึ่งแต่ละแหล่งอาจจะเก็บข้อมูลเดียวกันในรูปแบบที่แตกต่างกันไป เช่น การวิเคราะห์ข้อมูลทางโทรศัพท์ เพื่อหาเบอร์โทรศัพท์ที่ใช้ฝากข้อความเข้า Voice mail ในแต่ละเมือง โดยแต่ละเมืองมีการจัดเก็บที่แตกต่างกัน เช่น เมืองหนึ่งเก็บเบอร์โทรศัพท์ที่ใช้โทรเข้า Vocie mail ด้วยเพียงเบอร์ต้นทาง และเบอร์ปลายทาง ในขณะที่อีกเมืองอาจจะเก็บเบอร์ที่ไม่รู้ด้วยเบอร์ปลายทาง ส่วนอีกเมืองอาจจะเก็บด้วยเบอร์ที่โทรเข้า Voice mail จริงๆ ซึ่งก่อนที่จะทำการวิเคราะห์จะต้องทำการแก้ไขข้อมูลในส่วนนี้ก่อน

3.) ลักษณะการจัดเก็บข้อมูลที่แตกต่างกัน : การจัดเก็บข้อมูลในคอมพิวเตอร์แตกต่างกัน ส่วนใหญ่แล้วข้อมูลที่ใช้ในการวิเคราะห์มักจะเก็บด้วยภาษา COBOL หรือ RPG เก็บ text ในลักษณะ EBCDIC และตัวเลขในลักษณะของ decimal ในขณะที่ ระบบที่ใช้งานทั้งในส่วน data mining หรือ ระบบสนับสนุนการตัดสินใจ จะเก็บข้อมูลด้วย C หรือ C++ เก็บ text ในลักษณะของ ASCII และตัวเลขในลักษณะของ Integer หรือ Floating Point เป็นต้น ดังนั้นจะต้องมีการแปลงข้อมูลให้อยู่ในลักษณะที่สอดคล้องกัน

4.) ข้อมูลที่จัดเป็นข้อความ : ข้อมูลที่จัดเก็บเป็น Text ในบางครั้งนอกจากจะประกอบได้ด้วยข้อมูลที่ไม่จำเป็นสำหรับการวิเคราะห์แล้ว ในบางครั้งก่อให้เกิดความสับสนอีกด้วย ไม่ว่าจะเป็นรหัสไปรษณีย์ ทะเบียนรถ เช่น " V0R2J0" , "V0R 2J0" หรือ "V0R-2J0" ทั้ง 3 ค่านี้คือค่าเดียวกัน แต่ในการวิเคราะห์ข้อมูลโปรแกรมจะมองเป็นค่าที่แตกต่างกัน วิธีการแก้ไข เช่นการสร้างตารางสำหรับเก็บข้อมูลที่ถูกต้อง และแทนที่ข้อมูลที่นำมาวิเคราะห์ด้วยรหัส เช่น ในฐานข้อมูล Relation Database แทนข้อมูล product_name ด้วย product_code ซึ่งเป็นค่า unique ในตารางนั้นๆ

2.) การเตรียมข้อมูลเบื้องต้น (Data Preprocessing)

วัตถุประสงค์เพื่อทำข้อมูลให้มีคุณภาพมากขึ้นแก้ไขปัญหาต่างๆที่พบในข้อมูล เช่นมีข้อมูลผิดพลาดแตกต่างไปจากข้อมูลปกติ หรือข้อมูลหาย ข้อมูลต่างรูปแบบกันเพราะนำมาจากหลายแหล่งข้อมูลเป็นต้น โดยในขั้นตอนนี้จะประกอบไปด้วยขั้นตอนย่อย ดังนี้

2.1) Data Cleaning : ในขั้นตอนนี้จะทำการแก้ไขปัญหาที่พบในข้อมูล ได้แก่

- Missing Value : ข้อมูลที่ขาดหายไป อาจจะเกิดจากการบันทึกผิดพลาด หรือกรอกไม่ครบถ้วน การแก้ไขข้อผิดพลาดเหล่านี้ ถ้าข้อมูลมีปริมาณน้อยก็สามารถตัดรายการนั้นได้เลย หรือ ถ้าข้อมูลมีปริมาณน้อย แต่ข้อมูลนั้นสำคัญมากอาจจะทำการแก้ไขด้วยผู้บันทึกเอง ซึ่งจะต้องพิจารณาเป็นกรณีไป ในทางปฏิบัติทำได้ยาก หรือ แทนค่า missing value ด้วยค่าบางอย่าง เช่น unknown แต่ถ้ามีข้อมูลลักษณะนี้มากก็อาจจะทำให้การ mining ผิดพลาดได้ นอกจากนี้ ถ้าข้อมูลเป็นตัวเลขอาจจะแทนด้วยค่าเฉลี่ยของข้อมูลนั้นๆได้ เป็นต้น

- Nosiy Value : ข้อมูลที่แตกต่างไปจากข้อมูลที่คาดการไว้ อาจจะเกิดจากการบันทึกข้อมูล การป้อนข้อมูลผิดพลาด ปัญหาการส่งข้อมูล เป็นต้น ซึ่งค่าเหล่านี้หากนำไปพิจารณาแล้วอาจจะทำให้ model ที่ได้มีความเบี่ยงเบนไปจากสิ่งที่ควรจะเป็นได้

2.2) Data Integration : เนื่องจากข้อมูลมาจากหลายแหล่งซึ่งอาจจะเก็บในโครงสร้างที่ต่างกัน จึงต้องมีการตรวจสอบ เช่น เลขที่บัญชี ในแหล่งข้อมูลแรก แทนด้วย ACCT_ID แต่สำหรับ

อีกแหล่งข้อมูลแทนด้วย IACCT ซึ่งทั้งสองก็แทนเลขที่บัญชีเช่นเดียวกัน ซึ่งอาจจะมีการป้องกันโดยดูจาก meta data ของแต่ละแหล่งข้อมูล เป็นต้น

2.3) Data Reduction : การลดขนาดของข้อมูล ในบางครั้งข้อมูลที่นำมาพิจารณานั้นอาจจะมีขนาดใหญ่มาก และมีตัวแปรเยอะ ทำให้ต้องมีการลดขนาดลง ซึ่งมีการลดได้ 2 ลักษณะคือ การลดขนาดของข้อมูล (Data size reduction) อาจจะใช้การ sampling ในการเลือกกลุ่มข้อมูล และการลดขนาดของความสัมพันธ์ของตัวแปร (Dimension reduction) เพราะการที่มีตัวแปรมากเกินไปจะทำให้การทำงานช้าลงได้

3. การแปลงข้อมูล(Data Transformation)

ทำการสร้างข้อมูลชุดใหม่ จากข้อมูลเดิม โดยมีวัตถุประสงค์เพื่อทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ และเหมาะสมกับ Algorithm ของ Data mining ที่เลือก เช่น บาง model ใช้ได้สำหรับข้อมูลประมาณต่อเนื่องเท่านั้น ดังนั้น ข้อมูลประเภท categorical เช่น เพศ ต้องแปลงให้เป็นตัวเลข เช่น เพศชาย แทนด้วย 1 และ เพศหญิง แทนด้วย 2 เป็นต้น หรือบางครั้งต้องแปลงข้อมูลแบบ quantitative ให้เป็น categorical โดยอาจจะแบ่งข้อมูลเป็นช่วงๆ เป็นต้น

ขั้นตอนที่ 3 : การทำ Data Mining

ขั้นตอนนี้เป็นการประมวลผลข้อมูลตาม algorithm ที่กำหนดไว้ เพื่อทำการกลั่นกรองรูปแบบข้อมูลจากข้อมูลที่ได้มีการเตรียมมาในขั้นตอนก่อนหน้านี้ ซึ่งในขั้นตอนนี้จะมีการนำเทคนิคต่างๆมาใช้ ซึ่งแต่ละวิธีจะมีข้อดี ข้อเสียต่างกัน ซึ่งผลที่ได้คือการดึงเอารูปแบบหรือ information ที่ซ่อนอยู่ในข้อมูลออกมา

ขั้นตอนที่ 4 : การทำการวิเคราะห์ผล

เป็นขั้นตอนการวิเคราะห์ และแปลความหมายจากผลที่ได้ ซึ่งผลที่ได้จากการทำ data mining นี้เป็นเพียง information ส่วนย่อยๆ โดยขึ้นอยู่กับผู้วิเคราะห์ในการนำเอา informaiton เหล่านี้มาวิเคราะห์ประกอบร่วมกับความรู้ทางธุรกิจเพื่อนำไปใช้ประโยชน์ต่อไป ซึ่งเครื่องมือทางด้าน Graphical Visualization จะช่วยวิเคราะห์ข้อมูลได้สะดวกเพราะสามารถนำเสนอในรูปแบบที่ง่ายต่อความเข้าใจ

ขั้นตอนที่ 5 : การปรับความรู้ที่ได้เข้ากับธุรกิจ

การนำเอา information ที่ได้จากการวิเคราะห์ไปใช้งานต่อไป โดย information ที่ได้นี้เป็นแนวคิดที่ค้นพบใหม่ มีความถูกต้อง และสามารถนำไปใช้ทางธุรกิจให้เกิดประโยชน์สูงสุด

2.3 Data Mining Operation

Data mining ประกอบไปด้วย 4 operation หลัก ได้แก่ Predictive Modeling , Database Segmentation , Link Analysis และ Data Visualization

1. Predictive Modeling เป็นการสร้างแบบจำลอง เพื่อทำการทำนาย จะแบ่งออกเป็น 2 ลักษณะคือ

- Classification เป็นการทำนายกลุ่มของข้อมูลที่นำเข้า ซึ่งผลที่ได้จะเป็นกลุ่ม ซึ่งจะมีการกำหนดกลุ่มไว้ล่วงหน้าแล้ว เช่น การทำนายว่านักเรียนคนนี้จะจบด้วยเกรดใด (A,B,C,D,F)
- Forcasting เป็นการทำนายค่าที่เป็นตัวเลขต่อเนื่อง ใช้เพื่อทำนายค่าของเหตุการณ์ในอนาคต เช่นการทำนายอุณหภูมิ หรือ การทำนายค่าหุ้น

2. Database Segmentation หรือ Clustering เป็นการจัดแบ่งกลุ่มของฐานข้อมูล โดยเริ่มต้นจากการที่ไม่รู้ข้อมูลว่ามีกี่ลักษณะ นำ Data mining เพื่อจัดข้อมูลที่มีลักษณะใกล้เคียงกันไว้ในกลุ่มเดียวกัน

3. Link Analysis เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างของข้อมูล เช่น การเก็บข้อมูลการซื้อสิ้นค้าของลูกค้า ว่าลูกค้ามีการซื้อสินค้าอะไรบ้าง ซื้อวันไหน เพื่อศึกษาพฤติกรรมการซื้อสินค้าของลูกค้า เพื่อมานำมาวางแนวทางส่งเสริมการขาย หรือการจัดชั้นวางสินค้า เทคนิคที่ใช้ได้แก่

- Association Rule Discovery เป็นการค้นหาสิ่งที่มีความสัมพันธ์กัน
- Sequential Pattern Analysis เป็นการค้นหารูปแบบว่า เหตุการณ์ใดเกิด แล้วมักจะมีเหตุการณ์ใดเกิดตามมา

4. Data Visualization เป็นเทคนิคในการนำข้อมูลมาแสดงเป็น graph เพื่อนำมาวิเคราะห์ผล หรือเพื่อเป็นการหาข้อมูลที่แปลกปลอมออกจากกลุ่ม

2.4 การประยุกต์ใช้งาน Data Mining

ในปัจจุบันมีการนำ Data mining มาใช้ประโยชน์ในทางธุรกิจได้หลายทาง ได้แก่

1.) การจัดการทางการตลาด (Market analysis and management) เช่น

- Cross selling เป็นการหาความสัมพันธ์ของสินค้าที่ขายไปด้วยกัน เพื่อนำเสนอขายสินค้าให้กับลูกค้า

- Customer Profiling เป็นการทำ segment ของลูกค้า โดยศึกษาตาม profile ของลูกค้า เพื่อแบ่งลูกค้าเป็นกลุ่ม เพื่อนำเสนอขายโครงการต่างๆให้เหมาะสมกับลูกค้ากลุ่มนั้นๆต่อไป

- Direct mail campaign เป็นการเลือกกลุ่มลูกค้าที่จะทำการส่ง direct mail ไปแล้ว ลูกค้ามีการตอบสนองกลับ เนื่องจากหากทำการสุ่มส่ง mail ออกไปจะมีการตอบสนองกลับมาน้อยมาก ซึ่งทำให้สูญเสียต้นทุนสูงด้วย

- Customer Relation Management (CRM) เป็นวิธีการรักษาความสัมพันธ์ของลูกค้าให้ดีขึ้น ปัจจุบันนิยมใช้กับ call center เพื่อบันทึกการโทรเข้ามาขอรับบริการจากลูกค้า เพื่อนำข้อมูลนั้นมาวิเคราะห์ต่อไป

2.) การวิเคราะห์ความเสี่ยง (Risk Analysis and management) เช่น

- Credit Scoring เป็นการประเมินเบื้องต้นในการปล่อยกู้ให้กับลูกค้า

- Churn Management เป็นการศึกษาถึงสาเหตุใดที่ทำให้ลูกค้าเปลี่ยนไปใช้บริการจากบริษัทอื่น

3. การตรวจจับการโกง (Fraud detection and management) เช่น

- Money laundering การตรวจจับการฟอกเงิน โดยพิจารณาถึงการเชื่อมโยง transaction ที่น่าสงสัย

4. Web Mining เช่น

- การศึกษาการใช้ web site (Web usage mining) เพื่อมีผลต่อการออกแบบหน้าจอต่างๆ หรือการจัด category ของ web site

5. Text Mining เป็นการพัฒนาจากอดีตที่มีแนวคิดว่า data mining มักจะทำกับข้อมูลที่เป็นตัวเลข แต่ข้อมูลบางอย่างทำไม่ได้ จึงเกิดเป็นแนวคิด Text mining ซึ่งอาจจะทำการแบ่งเอกสารเป็นกลุ่ม หรือ แย่งตามคำสำคัญเป็นต้น

จากที่กล่าวมาข้างต้นแล้วนั้น การจะนำ operation ของ data mining ไปใช้กับงานลักษณะใดนั้นไม่สามารถที่จะระบุได้แน่นอนทั้งนี้ ขึ้นกับการพิจารณาของผู้วิเคราะห์ กล่าวโดยสรุปสามารถสรุป operation ของ data mining นำมาประยุกต์ใช้กับงานทางธุรกิจได้ดังตาราง 2.1

Market Management	Risk Management	Fraud Management
Target Marketing *Customer Relationship* *Market basket analysis* *Cross selling* *Market segmentation*	*Forecasting* *Customer retention* *Improved underwriting* *Quality control* *Competitive analysis*	*Fraud detection*

Predictive Modeling	Database Segmentation	Link Analysis	Deviation Detection
Classification *Value prediction*	*Demographic clustering* *Neural clustering*	*Association discovery* *Sequential pattern discovery* *Similar time sequence discovery*	*Visualization* *Statistics*

รูปที่2.2 แสดงการนำ Data mining มาประยุกต์ใช้กับงานทางธุรกิจ[1]

2.5 สรุป

Data mining เป็นกระบวนการที่ใช้ในการหา information ที่ซ่อนอยู่ภายในข้อมูลที่มีอยู่ออกมาเพื่อใช้ประโยชน์ในการดำเนินธุรกิจ ซึ่งขั้นตอนในการทำ data mining แต่ละขั้นตอนมีความสำคัญเพราะแต่ละขั้นเป็นการกำหนด information ที่ต้องการ เพราะหากกำหนดวัตถุประสงค์ไม่ชัดเจน เลือกข้อมูลไม่ถูก ก็จะทำให้ผลลัพธ์ที่ได้ออกมาไม่ได้เป็นไปอย่างที่ต้องการได้ สุดท้ายแล้วผลลัพธ์ที่ได้จากการทำ data mining เป็นเพียง information ย่อยๆไม่ได้กลยุทธ์ทางการตลาด ทั้งนี้ขึ้นอยู่กับผู้วิเคราะห์ในการนำ information ที่ได้เหล่านี้ไปปรับใช้ให้เข้ากับธุรกิจให้ได้มากที่สุด

บทที่ 3

Link Analysis

การทำงานของ Link Analysis เพื่อหาความสัมพันธ์ของรายการที่มีความสนใจ หรือหาความเกี่ยวข้องกันระหว่างรายการ หรือกลุ่มรายการ เช่น หาความสัมพันธ์ระหว่างผลิตภัณฑ์ หรือบริการที่ลูกค้ามีความสนใจ ณ เวลาหนึ่งๆ เทคนิคที่สำคัญของ Link analysis ได้แก่

1. Association Discovery
2. Sequential Pattern Discovery
3. Similar Time Sequence Discovery

3.1 Association Discovery

เทคนิค association rule (กฎความสัมพันธ์) ใช้สำหรับการวิเคราะห์การบริการทางการเงิน และในธุรกิจค้าปลีก ซึ่งเรียกอีกชื่อหนึ่งว่า "Market basket analysis" หรือ Product Affinity Analysis โดยเป็นการกล่าวถึงความคิดที่ว่า คุณสามารถสำรวจสินค้าที่อยู่ในตะกร้า และค้นหารูปแบบที่ซึ่งสามารถถูกใช้สำหรับการลักษณะการวางสินค้าบนชั้นวางได้ดีที่สุด ดังเช่นเรื่อง "beer and diaper" ที่ซึ่ง การวิเคราะห์ association rule ได้แสดงให้เห็นว่า beer และผ้าอ้อม สามารถถูกซื้อไปด้วยกันบ่อยกว่าที่คิด ซึ่งผู้วิเคราะห์ได้กล่าวว่า ผู้ชายที่ธุระที่จะต้องไปร้านขายของเพื่อซื้อผ้าอ้อมจะทำการซื้อเบียร์มาด้วย ซึ่งบทสรุปที่ได้มานี้ทำให้ผ้าอ้อมถูกวางใกล้ๆกับเบียร์ใน supermarket กรณี beer กับผ้าอ้อมนี้เป็นตัวอย่างที่แสดงให้ เห็นถึงว่า association ruleสามารถถูกใช้ในการกำหนดกลยุทธ์ได้อย่างไร รวมถึงเป็นตัวอย่างของความเสี่ยงในการวางหลักการทั่วไป ให้อยู่บนพื้นฐานของกฎที่ไม่ได้เชื่อถือได้เสมอไป

ในทางธุรกิจสามารถนำ Market basket analysis เพื่อกำหนดกลยุทธ์ต่างการตลาดต่างๆเช่น

: การลดราคา หรือคูปองลดราคา (Couponing and discounting)เช่น การที่จะลดราคาทั้งเบียร์ และผ้าอ้อม เพียงลดราคาสินค้าเพียงชนิดเดียวก็ช่วยเพิ่มยอดขายสินค้าอีกชนิดด้วย เนื่องจากสินค้าทั้งสองชนิดมีแนวโน้มที่จะถูกซื้อไปด้วยกัน

: การจัดวางสินค้า(Product placement) การวางสินค้าที่มีความสัมพันธ์กันไว้ใกล้ๆทำให้เกิดข้อดีมากกว่าการวางสินค้าที่มีลักษณะประเภทเดียวกันไว้ด้วยกัน หรือในทางตรงกันข้ามการ

วางสินค้าที่มีความสัมพันธ์กันไว้ไกลๆทำให้ลูกค้าต้องเดินหา ซึ่งจะทำให้มองเห็นสินค้าชนิดอื่นๆ ได้ด้วย

:Timing และ Cross-marketing เช่น กำหนดกฎว่า "คนที่ซื้อ VCR(เครื่องบันทึกเทป) มักจะซื้อ กล้องวีดีโอ พกพาในเวลา 2-4 เดือนหลังจากที่ซื้อ VCR ไป" ดังนั้น เมื่อมีการ promote กล้องวีดีโอใหม่ ก็สามารถทำการส่งข้อมูลเหล่านี้ไปให้กับผู้ที่ซื้อ VCR ไป ซึ่งมีแนวโน้มว่าจะกลับมาซื้อในอีก 2- 3 เดือนข้างหน้า

3.1.1 ขั้นตอนการทำงานของ Association Discovery มี 3 ขั้นตอน

1. เลือกชุดข้อมูลที่ถูกต้อง (Choose the Right Set of Items)

2. สร้างกฎ (Generating Rules from All This Data) โดยทำการคำนวณหาค่า Support และ Confidence เพื่อวัดประสิทธิภาพของกฎ

3. เลือกเฉพาะชุดข้อมูลที่เป็นไปได้ (Overcoming Practical Limit)

1.) การเลือกชุดข้อมูลให้ถูกต้อง

ข้อมูลส่วนใหญ่ที่ใช้มักจะเป็นข้อมูลระดับรายละเอียด (transaction) ที่เกิดขึ้น ณ จุดขาย ซึ่งการจะนำข้อมูลระดับไหนมาใช้ขึ้นกับวัตถุประสงค์ในการวิเคราะห์ เช่น ในการวิเคราะห์การซื้อสินค้าในร้านขายของชำ ซึ่งมีสินค้าเป็นร้อยที่อยู่บนหิ้งขาย ซึ่งจะรวมวิเคราะห์เป็นสินค้าแต่ละประเภท หรือทำการวิเคราะห์รายละเอียดถึงการซื้อเช่น ซื้อเบียร์ ยี่ห้ออะไร ขนาดขวดกี่ pack เป็นต้น หรือ ในการวิเคราะห์การขายพิซซ่า ถ้าต้องการวิเคราะห์ถึงลักษณะหน้าพิซซ่าที่ลูกค้านิยมทานก็ต้องพิจารณาถึง topping ที่ทำการใส่ไม่ว่าจะเป็น ชีส เห็ด หรือพริกไทย และความหนาของ พิซซ่า (แป้งหนา,แป้งบาง) ลักษณะข้อมูลที่ได้จะเป็นดังรูปที่ 3.1

Customer	Extra Cheese	Onions	Mushrooms	Pepper
1	✓	✓		
2			✓	
3	✓	✓		✓
4		✓		
5	✓		✓	✓

รูปที่3.1 การวิเคราะห์การขาย พิซซ่าแบบละเอียด [2]

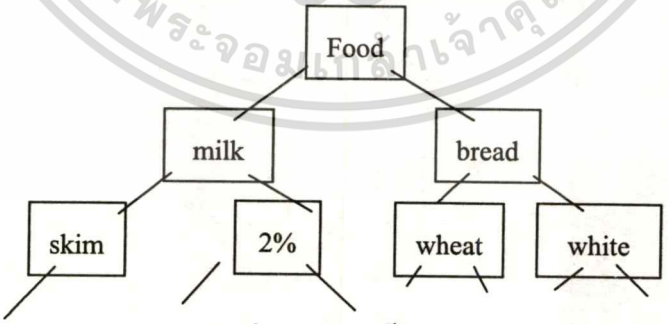
ในขณะที่ในมุมมองของผู้จัดการร้านต้องการทราบข้อมูลสรุป เมื่อลูกค้าเข้าร้านพิซซ่าแล้วมีการสั่งซื้ออะไรบ้าง เช่นสั่งพิซซ่า เครื่องดื่มน้ำอัดลม กาแฟ นม เป็นต้น ระดับข้อมูลที่นำมาวิเคราะห์จะเปลี่ยนไปเป็นดังรูปที่ 3.2

Customer	Pizza	Milk	Sugar	Apples	Coffee
1	✓				
2		✓	✓		
3	✓			✓	✓
4		✓			
5	✓		✓	✓	✓

รูปที่ 3.2 การวิเคราะห์การขายพิซซ่าแบบสรุป [2]

อย่างไรก็ตามการพิจารณาระดับของข้อมูลที่สนใจมีการเปลี่ยนแปลงได้ตลอด เช่นเมื่อพบว่าข้อมูลในระดับที่กำหนดไม่เพียงพอต่อความต้องการ ก็สามารถกำหนดในระดับรายละเอียดที่ลึกลงไปกว่านั้นได้

ในการวิเคราะห์ข้อมูลการพฤติกรรมการซื้อสินค้าของลูกค้าโดยเฉพาะร้านขายปลีก เช่น Supermarket ขนาดใหญ่ที่มีสินค้าอยู่มากมาย ดังนั้นการพิจารณาสินค้าทุกรายการเป็นไปได้ยาก ใช้เวลาในการคำนวณนาน และในบางครั้งทำให้ผลที่ได้ไม่ทำให้เกิดประโยชน์ ดังนั้นถึงมีการรวมกลุ่มสินค้าประเภทเดียวกันไว้ด้วยกัน เรียกว่า **Taxonomy** ดังรูปที่ 3.3



รูปที่ 3.3 ลำดับชั้นของข้อมูล

จาก**รูปที่** 3.3 แสดงลำดับชั้นของข้อมูลประเภทอาหาร เมื่อพิจารณาในระดับล่างสุดนม จะแบ่งเป็น skim(นมพร่องมันเนย) และ นมไขมัน2% ซึ่งหากวิเคราะห์ด้วยข้อมูลระดับล่างนี้ในบางครั้งจำนวน Transaction ที่เกิดมีไม่มาก และต้องเสียเวลาในการคำนวณและ วิเคราะห์นาน ถึงทำการรวมเป็นสินค้าประเภทนม (milk) เป็นต้น

ข้อมูลที่ใช้สำหรับการวิเคราะห์ความสัมพันธ์ ประกอบด้วย 2 ส่วนคือ entity และ attributes โดย entity จะเป็นรายการซื้อ ส่วน attribute เป็นรายการที่ถูกซื้อใน 1 ครั้ง หรือ กำหนด entity คือ คนไข้ ดังนั้น attribute คือ หมอที่รักษา , ยา ,กระบวนการรักษา และค่าใช้จ่าย เป็นต้น ที่ซึ่งมีความสัมพันธ์กับคนไข้

ข้อมูลที่ใช้มี 2 ลักษณะได้แก่ Horizontal และ Vertical

- **Horizontal Format** ประกอบไปด้วย 1 แถวมาจากหลายๆตาราง และแต่ละคอลัมน์คือแต่ละ attribute เช่น การวิเคราะห์ประสิทธิภาพของกระบวนการรักษา ต้องการข้อมูล 1 แถวสำหรับคนไข้แต่ละคน โดยแบ่งตามคอลัมน์ คือ ยา,หมอที่ทำการรักษา และ ค่าใช้จ่าย เป็นต้น สำหรับการวิเคราะห์ market basket (การซื้อสินค้าในแต่ละครั้ง) 1 แถวเป็นรายการสำหรับการซื้อ 1 ครั้ง โดยแต่ละคอลัมน์คือแต่ละสินค้า

ปัญหาที่สำคัญที่พบในข้อมูลลักษณะนี้คือ จำนวนคอลัมน์ที่มีอยู่ในปริมาณที่ใหญ่มาก ซึ่งบางครั้งอาจจะมีถึง 100,000 รายการ ดังนั้นการรวมกลุ่มสินค้าเข้าไว้ด้วยกันเป็นการลดจำนวน column ลงไปได้มาก นอกจากนี้เนื่องจาก schema และข้อมูลมีความสัมพันธ์กัน ดังนั้นเมื่อมีการเพิ่มสินค้าชนิดใหม่เข้ามาก็จะต้องมีการเปลี่ยนแปลง schema ใหม่

ตารางที่ 3.1 ข้อมูลแบบ Horizontal Format

TID	ผ้าอ้อม	ขนมปังกรอบ	เบียร์	ผ้าเช็ดหน้า	ผลิตภัณฑ์นม
111	✓	✓			✓
112		✓	✓		

- **Vertical Format** เพื่อแก้ไขปัญหา ของ Horizontal Format โดยกำหนดให้คอลัมน์เป็น enitiy โดยแต่ละแถวแทนแต่ละ attribute โดยเชื่อมโยงเข้าด้วยกันโดยใช้ common ID ซึ่งวิธีนี้จะมีประโยชน์มากกว่าเมื่อมีจำนวน attribute ที่มาก

ตารางที่ 3.2 ข้อมูลแบบ Vertical Format

ID	Product
111	ผ้าอ้อม
111	ขนมปังกรอบ
112	ขนมปังกรอบ
112	เบียร์

การทำงานของ Association Rule สามารถใช้ได้เพียงกับข้อมูลเชิงคุณภาพ (Categorical Data) สำหรับข้อมูลประเภทต่อเนื่องเช่น รายได้ของลูกค้า จะต้องทำการแปลงเป็นข้อมูลเชิงคุณภาพก่อน โดยอาจจะกำหนดเป็นช่วงของข้อมูลก็ได้ เช่น 0-20,000 , 20,001- 40,0000 เป็นต้น

2.) สร้างกฎความสัมพันธ์

นำข้อมูลที่ได้มาสร้างกฎโดยรูปแบบของกฎที่ได้จะอยู่ในลักษณะ **"IF condition1 THEN condition2"** หรือ **"WHEN condition1 THEN condition2"** โดยที่ condition1 และ condition2 เกิดขึ้นพร้อมกันใน transaction เดียวกันจะเรียกส่วน conditon1 ว่า Rule Body (เหตุ) หรือ Left-Hand side(LHS) และเรียก condition2 ว่า Rule Head (ผล) หรือ Right-Hand side(RHS) ในทางปฏิบัติโดยทั่วไปแล้ว มักจะเกิด condition2 เพียง 1 รายการ เช่น "IF Diapers and Thrusday, then Beer" ซึ่งจะมีความหมายในการวิเคราะห์มากกว่า "IF Thrusday , then Diapers and Beer"

หน่วยวัดหลายๆตัวที่ใช้สำหรับการประเมิน association rule และกำหนดว่า rule ที่ได้นั้นมีประโยชน์หรือไม่ ซึ่งหน่วยวัดเหล่านี้ได้แก่ support , confidence , lift ซึ่งหน่วยวัดเหล่านี้เกี่ยวข้องกับขนาดตัวอย่างของประชากร ซึ่งในการใช้ association rule นั้นจำนวนตัวอย่างที่เหมาะสมนั้นยากต่อการกำหนด ไม่เหมือนกับเทคนิคทางสถิติอื่นๆ

โดยหน่วยวัดประสิทธิภาพของกฎที่ได้ มี 2 ค่าหลัก ได้แก่ Support Factor และ Confidence Factor

● **Support Factor (Prevalence)** แสดงถึงความถี่ในการ เกิดกฎนั้น โดย support ของ A=>B สามารถแทนด้วยสัญลักษณ์ **sup(A=>B)** โดยค่า support คือ ค่าที่แสดงสัดส่วนระหว่างจำนวนชุดข้อมูลที่มีทั้งข้อมูลที่เป็นทั้ง "เหตุ" และ "ผล" ของเหตุการณ์ เทียบกับจำนวนเหตุการณ์ภายในชุดข้อมูลทั้งหมด โดยเมื่อ support มีค่าสูง ซึ่งบ่อยครั้งอาจจะเป็นการบอกสิ่งที่รู้อยู่แล้ว เช่น กฎที่ว่า ATM => Saving account ซึ่งเป็นกฎที่มีค่า support สูง แต่ไม่ได้แสดงความหมายลึกซึ้งอื่นๆ แต่ก็เป็นสิ่งที่แสดงถึงสิ่งที่เรารู้อยู่แล้ว เกี่ยวกับพื้นฐานลูกค้า

$$\text{support} = \frac{\text{จำนวนชุดข้อมูลที่มีรายการผ้าอ้อมและเบียร์}}{\text{จำนวนชุดข้อมูลทั้งหมด}}$$

● **Confidence (Predictability)** เป็นหน่วยวัดถึงการเกิดขึ้นจริงของกฎ หรือ ความบ่อยที่ rule head เป็นจริง เมื่อเกิด rule body สามารถเขียนแทนด้วยสัญลักษณ์ **Conf(A => B)** โดยค่า Confidence คือ ค่าที่แสดงสัดส่วนระหว่างจำนวนชุดข้อมูลที่มีทั้ง "เหตุ" และ "ผล" เทียบกับจำนวนข้อมูลที่มีเฉพาะส่วน "เหตุ"

$$\text{confidence} = \frac{\text{จำนวนชุดข้อมูลที่มีรายการผ้าอ้อม และเบียร์}}{\text{จำนวนชุดข้อมูลที่มีรายการซื้อผ้าอ้อม}}$$

เมื่อกำหนดตัวอย่างรายการซื้อสินค้าของร้านค้าแห่งหนึ่ง ตามตารางที่ 3.3 ว่าเกิดรายการซื้อทั้งหมด 500,000 รายการจะได้ค่า Support ของการซื้อทั้ง ผ้าอ้อม และ beer คือ 10,000 / 500,000 = 2% และ Confidence ของการซื้อ ผ้าอ้อม และ beer จะได้ว่า รายการซื้อทั้ง diaper และ beer เกิด

10,000 รายการ ส่วนการเกิดรายการซื้อ diaper อย่างเดียวมี 20,000 รายการ ดังนั้นได้ Confidence = 10,000 / 20,000 = 50% ในทางตรงกันข้ามหากกำหนดกฎเป็น Buy beer then buy diaper จะได้ confidence = 33.33% แต่ทั้งสองมีค่า support เท่ากัน ทั้งนี้เนื่องจากค่า support ไม่ขึ้นตามทิศทางของกฎ แต่ขึ้นอยู่กับ item ในกฎนั้นๆ

ตารางที่ 3.3 ตัวอย่างข้อมูลการ ซื้อสินค้า [5]

สินค้า	จำนวนรายการ
ผ้าอ้อม	20,000
เบียร์	30,000
ผ้าเช็ดหน้า	10,000
ผ้าอ้อม และเบียร์	10,000
ผ้าเช็ดหน้า และผ้าอ้อม	8,000
ผ้าเช็ดหน้า และเบียร์	220
ผ้าเช็ดหน้า, ผ้าอ้อม และเบียร์	200

ตารางที่ 3.4 Rule และค่าหน่วยวัด [5]

	Left-hand side		Right-hand side	Expected Confidence (%)	Confidence (%)	Support (%)	Lift Ratio
1	ผ้าอ้อม	-->	เบียร์	6	50	2	8.33
2	เบียร์	-->	ผ้าอ้อม	4	33.33	2	8.33
3	ผ้าอ้อม	-->	ผ้าเช็ดหน้า	2	40	1.60	20
4	ผ้าเช็ดหน้า	-->	ผ้าอ้อม	4	80	1.60	20
5	ผ้าเช็ดหน้า	-->	เบียร์	6	2.20	0.04	0.37
6	เบียร์	-->	ผ้าเช็ดหน้า	2	0.73	0.04	0.37
7	ผ้าอ้อมและผ้าเช็ดหน้า	-->	เบียร์	6	2.50	0.04	0.42
8	ผ้าอ้อม และเบียร์	-->	ผ้าเช็ดหน้า	2	2.00	0.04	1
9	ผ้าเช็ดหน้าและเบียร์	-->	ผ้าอ้อม	4	90.91	0.04	22.73
10	ผ้าอ้อม	-->	ผ้าเช็ดหน้าและเบียร์	0.044	1	0.04	22.73
11	ผ้าเช็ดหน้า	-->	ผ้าอ้อมและเบียร์	2	2	0.04	1
12	เบียร์	-->	ผ้าอ้อมและผ้าเช็ดหน้า	1.60	0.67	0.04	0.42

จากตารางที่ 3.4 เมื่อพิจารณา กฎที่ว่า "คนซื้อผ้าอ้อม" 4% (20,000/500,000) และ "คนซื้อเบียร์ " 6% (30,000/500,000) เรียก 4% , 6% นี้ว่า **Expected Confidence** ของการซื้อ diaper หรือ beer โดยไม่สนใจว่าสิ่งใดจะถูกซื้อไปด้วยกัน

● **Expected Confidence** คือค่าที่แสดงสัดส่วนระหว่างชุดข้อมูลที่เป็นตัวแทนเหตุการณ์ที่สนใจ ซึ่งอาจจะเป็น "เหตุ" หรือ "ผล" อย่างใดอย่างหนึ่งเทียบกับจำนวนเหตุการณ์ทั้งหมดภายในชุดข้อมูลนั้น

● **Lift** เป็นค่าที่ใช้ในการเปรียบเทียบคัดเลือกเหตุการณ์ค้นพบจากการทำงานของ software และใช้สรุปผล ประเมินลักษณะของความสัมพันธ์ กล่าวอีกนัยหนึ่งคือ Lift เป็นหน่วยวัดที่แสดงถึงความมีประโยชน์ของกฎนั้นๆ มีค่าเป็นสัดส่วนระหว่างค่า Confidence กับ Expected Confidence

Lift = Confidence / Expected Confidence

ดังนั้นค่า Lift ที่ได้จะบ่งบอกถึงความสำคัญของความสัมพันธ์หรือเหตุการณ์ที่ได้ว่ามีมากน้อยแค่ไหน

ข้อควรระวังในการวิเคราะห์คือ ค่าของ Negative Lift หรือ Lift น้อยกว่า 1 จากตารางที่ 3.4 กฎที่ว่า "คนที่ซื้อผ้าอ้อมและผ้าเช็ดหน้า จะซื้อเบียร์ด้วย" แสดงว่าคนที่ซื้อผ้าอ้อมและผ้าเช็ดหน้า น้อยคนมากที่จะซื้อเบียร์ด้วย หมายถึงเหตุการณ์นี้แทบจะไม่มีทางเกิดขึ้นพร้อมกันได้เลย

สำหรับกฎที่มีค่า Confidence สูงหรือว่าต่ำเกินไป บางครั้งกฎเหล่านั้นอาจจะผิดปกติ เช่น กฎที่กล่าวว่า เมื่อซื้ออาหารสัตว์ จะซื้อที่ใส่อาหารสัตว์ด้วย มี Confidence = 1 (100%) มีความหมายราวกับว่า ไม่มีใครที่ซื้ออาหารสัตว์ แล้วไม่ซื้อที่ใส่อาหารไปด้วย ซึ่งผลที่ได้นี้อาจจะแสดงถึงข้อมูลเพียงวันเดียว และมีการจัดโปรโมชั่นพิเศษ เช่น เมื่อซื้ออาหารสัตว์จะแถมที่ใส่อาหารฟรี เป็นต้น

โดยทั่วไป นักวิเคราะห์จะให้ความสนใจในกฎที่มีค่า Lift ที่สูงหรือต่ำมาก ซึ่งจะช่วยให้หาความสัมพันธ์ได้ง่าย สำหรับกฎที่มีค่า Support ต่ำสามารถแสดงถึงค่าทางสถิติที่ผิดปกติ หรือคำนวณผิดพลาดได้ และถึงแม้กฎที่ได้จะมีความถูกต้อง การทำตามกฎที่มีค่า support ต่ำนั้นอาจจะนำไปสู่ความผิดพลาดที่ใหญ่กว่านี้ เพราะค่า support แสดงถึงความบ่อยของการเกิดกฎนั้น ดังนั้นถ้ามีค่าที่ต่ำแสดงถึงเหตุการณ์ที่เกิดไม่บ่อยนัก เช่น ต้องการลงทุนเพิ่มเพื่อให้เกิดกำไรขึ้นเป็น 100 ล้าน สำหรับเหตุการณ์ที่เกิดขึ้น 3 ครั้งต่อปี จึงไม่คุ้มต่อการลงทุน เป็นต้น

3.) เลือกเฉพาะข้อมูลที่เป็นไปได้

การทำงานของ Association Rule มีแนวคิดมากจากการนับจำนวนครั้งของรายการที่เกิดขึ้นและรวมเหตุการณ์เข้าด้วยกันในทุกทางที่เป็นไปได้ ทำให้กฎที่ได้มีปริมาณมาก จึงมีการนำเทคนิค "Pruning" ช่วยลดจำนวนรายการที่นำมารวมกัน โดยลดจำนวนรายการที่ไม่ตรงกับเงื่อนไขออก โดยผู้วิเคราะห์จะต้องทำการระบุค่าทางสถิติ 2 ค่าได้แก่

1. **Minimum Support** หรือ ***"Minimum Support Threshold"*** หรือ ***"min_sup"*** เป็นการระบุค่าขีดจำกัดล่างของค่า Support ซึ่งค่านี้จะใช้ระบุจำนวนข้อมูลอย่างต่ำที่ยอมรับได้ในการเลือกข้อมูลนั้นมาใช้

2. **Minimum Confidence** หรือ ***"Minimum Confidence Threshold"*** หรือ ***"min_conf"*** เป็นการระบุค่าขีดจำกัดล่างของค่า Confidence ซึ่งค่านี้จะใช้ระบุสัดส่วนอย่างต่ำของจำนวนการเกิดของกลุ่มข้อมูลที่สัมพันธ์กันกับจำนวนการเกิดของข้อมูลบางส่วนในกลุ่มข้อมูลนั้น

เทคนิค "Pruning" ที่นิยมใช้กันมากคือ "Minimum Support Pruning" นั่นคือ กำหนดว่ากฎที่ได้จะต้องมาจากรายการที่มีจำนวนการเกิดอย่างน้อยเท่ากับ Minimum Support เช่น มีจำนวนรายการอยู่ 1,000,000 รายการ และกำหนด minimum support = 1% ดังนั้นจะมีกฎที่มีค่า support มากกว่าเท่ากับ 10,000 รายการเท่านั้นที่ได้รับพิจารณา ซึ่งค่า minimum support นี้จะเป็นตัวจำกัดข้อมูลเป็นทอดๆ เช่น เมื่อพิจารณากฎ 4 item

IF A , B and C then D กำหนด minimum support 10,000 รายการ จะได้ว่า

A , B , C และ D จะต้องเกิดขึ้น item ละอย่างน้อย 10,000 รายการ นั่นคือ minimum support จะช่วยกรอง item ที่เกิดรายการน้อยออกไป จากนั้นพิจารณาวิธีเพื่อทำให้การรวมกันของ item(Combination) ลดลง ซึ่งทำได้ 2 วิธี ได้แก่

1. นำ item นั้นๆออกจากการพิจารณา หรือ

2. ใช้การรวมกลุ่ม สินค้าประเภทเดียวกันไว้ด้วยกัน (Taxonomy)

หลังจากการรวมกันของ item จะได้ว่า

A และ B ต้องเกิดขึ้นอย่างน้อย 10,000 รายการ และ

A และ C ต้องเกิดขึ้นอย่างน้อย 10,000 รายการ และ

A และ D ต้องเกิดขึ้นอย่างน้อย 10,000 รายการ ตามค่า minimum support ที่กำหนด

อย่างไรก็ตาม ในทางปฏิบัติจริง ค่า Minimum Support จะขึ้นกับข้อมูลและสถานการณ์ และสามารถเปลี่ยนแปลงได้ในแต่ละระดับของการทำงาน

3.1.2 ลักษณะกฎ Association Discovery มี 3 ลักษณะ ได้แก่

1.) กฎที่ได้เป็นสารสนเทศที่มีคุณภาพสูง สามารถนำไปตัดสินใจในการดำเนินการทางธุรกิจได้

2.) กฎที่ได้เป็นข้อมูลที่รู้อยู่แล้ว

3.) กฎที่ได้ไม่สามารถอธิบายได้ ไม่ได้สนับสนุนการตัดสินใจ

หากพิจารณาประเภทของข้อมูลที่นำมาใช้ใน Association Discovery สามารถแบ่งกฎที่ได้เป็น 2 ประเภท คือ

1.) **Boolean Association Rule** เช่น ผ้าอ้อม => เบียร์ [support= 6% , confidence=70%]

2.) **Quantitative Association Rule** ข้อมูลเชิงปริมาณ เช่น (อายุ= 26...30) => (รถยนต์=1,2) [support = 3% , confidence=36%]

3.1.3 ข้อดี และข้อเสียของ Association Discovery

ข้อดีของ Association Discovery

1.) ทำงานได้ดีกับข้อมูลขนาดใหญ่ ขณะที่เทคนิคอื่นๆจะมีปัญหาในการทำงานกับข้อมูลปริมาณมากๆ ซึ่งในปัจจุบันมีการวิจัยเพื่อเพิ่มประสิทธิภาพของ Association Discovery โดยลดจำนวนของตัวแทนข้อมูล และสุ่มตัวอย่างขึ้นมาทำ Data mining อีกด้วย

2.) สามารถระบุค่า Minimum Support และ Minimum Confidence ได้ ทำให้สามารถควบคุมจำนวนผลลัพธ์ได้

3.) สามารถทำการ mining กับข้อมูลบางส่วนได้ ทำให้ลดปัญหากรณีข้อมูลไม่สมบูรณ์ได้

4.) เทคนิคอื่นๆ เช่น Decision Trees จะระบุขอบเขตของกลุ่มข้อมูล ทำให้มีการจำกัดข้อมูล มีผลทำให้ข้อมูลที่ถูกเลือกมาอาจจะไม่ใช่ตัวแทนที่แท้จริงของข้อมูล

5.) สามารถจัดการกับข้อมูลที่รูปแบบแตกต่างกันได้ โดยไม่สูญเสียสารสนเทศ ในขณะที่เทคนิคอื่นๆจะจำกัดรูปแบบ และความยาวของข้อมูล

6.) การคำนวณง่าย และ แสดงผลด้วยสัญลักษณ์ที่ชัดเจนและเข้าใจง่าย

7.) สนับสนุนการวิเคราะห์เบื้องต้นเพื่อสร้างแนวทางการทำ Data mining ต่อไป เนื่องจากไม่ต้องกำหนด factor ต่างๆมากนัก

ข้อเสียของ Association Discovery

1.) ถ้าใช้กับข้อมูลที่เกิดขึ้นไม่บ่อย ใน Transaction ทำให้ข้อมูลแยกออกจากกลุ่มข้อมูลอื่นชัดเจน ทำให้ประสิทธิภาพของสารสนเทศที่ได้ลดลง

2.) กฎที่ได้มีปริมาณมากเกินไป ถึงแม้จะมีการกำหนดค่า minimum support และ minimum confidence เพื่อจำกัดจำนวนกฎที่สร้างขึ้น แต่ก็อาจจะทำให้กฎที่ได้ผิดเพี้ยน หากผู้วิเคราะห์ทำการกำหนดค่าเหล่านี้สูงหรือต่ำเกินไป

3.) บอกความแตกต่างของกฎที่ได้มายาก ว่าเป็นกฎจริง หรือว่าได้มาจากการข้อมูลพ้องกัน

4.) กฎที่ได้ไม่ได้ให้ถึงความเป็นเหตุเป็นผล บอกเพียงแนวโน้มที่จะเกิดขึ้นด้วยกัน

5.) การ scan ข้อมูลทั้งฐานข้อมูลค่อนข้างซับซ้อน และคำนวณนาน

6.) การกำหนดจำนวนสินค้าที่จะเข้าวิเคราะห์ทำได้ยาก เพราะหากกำหนดมากเกินไปจะทำให้ได้กฎที่มากเกินไป

7.) เนื่องจากมีการกำหนด minimum support เพื่อลดจำนวนกฎที่ได้ ในบางครั้งไม่พบข้อมูลที่เป็นแนวทางใหม่ๆที่ยังเกิดขึ้นน้อย

3.2) Sequencial Pattern Discovery

ใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าอย่างหนึ่ง และจะซื้อสินค้าอีกอย่างหนึ่งในเวลาต่อมา ดูลำดับการเลือกซื้อสินค้าของลูกค้า เนื่องจากเพื่อให้เข้าใจพฤติกรรมการซื้อสินค้าของลูกค้าในระยะยาว หน่วยวัดความสัมพันธ์ใน Sequential Pttern Discovery เป็นเช่นเดียวกับ Association Discovery แต่ มีการคำนวณค่า Support ต่างไป โดยค่า support เป็นอัตราส่วนจำนวนลูกค้าที่มีข้อมูลการซื้อสินค้าเป็นลำดับต่อจำนวนลูกค้าทั้งหมด จากตารางที่ 3.5 แสดงข้อมูลของร้านขายขนมปัง เรียงตามรหัสลูกค้า และรหัสรายการ เช่น B.Moor มาซื้อสินค้าที่ร้านเป็นเวลา 3 วันติดต่อกัน โดยทำการซื้อเบียร์ในวันแรก และซื้อไวน์ , Cider ในวันที่สอง และในวันที่สามจึงซื้อขนมปัง ดังแสดงในตารางที่ 3.6 แสดงการซื้อสินค้าของลูกค้าแต่ละราย

ตารางที่ 3.5 แสดงข้อมูลการซื้อสินค้าของลูกค้า [1]

Customer	Transaction Time	Items Bought
B.Adams	*June 20, 1994 5:37 p.m*	Beer
B.Adams	*June 21, 1994 10:30 a.m*	Bread
J.Brown	*June 20, 1994 10:13 a.m.*	Juice , Coke
J.Brown	*June 20,1994 11:47 a.m.*	Beer
J.Brown	*June 21,1994 9:22 a.m.*	Wine,Water, Cider
J.Mitchell	*June 21,1994 3:19 a.m.*	Beer,Gin,Cider
B.Moore	*June 20,1994 2:32 p.m.*	Beer
B.Moore	*June 21,1994 6:17 p.m.*	Wine,Cider
B.Moore	*June 22,1994 5:03 p.m.*	Bread
F.Zappa	*June 20,1994 11:02 a.m.*	Bread

Sequential Pattern จะนับจำนวนความถี่ที่เกิดขึ้นของ Transaction ของลูกค้าซึ่งเรียงลำดับเหตุการณ์ไว้แล้ว โดยแสดงเฉพาะคู่เหตุการณ์ที่มีค่ามากกว่า Minimum Support ดังแสดงในตารางที่ 3.7 ได้ข้อสารสนเทศที่ว่า "เมื่อลูกค้าซื้อเบียร์แล้วจะมีการซื้อขนมปังตามในภายหลัง" เหตุการณ์นี้เกิดขึ้นกับลูกค้า 2 ใน 5 คน

ตารางที่ 3.6 การซื้อสินค้าของลูกค้าแต่ละราย

Customer	Customer Sequence
B.Adams	(Beer)(Bread)
J.Brown	(Juice,Coke)(Beer)(Wine,Water,Cider)
J.Mitchell	(Beer,Gin,Cider)
B.Moore	(Beer)(Wine,Cider)(Beer)
F.Zappa	(Bread)

ตารางที่ 3.7 ผลลัพธ์ที่ได้จาก Sequential Pattern Discovery

Sequential Pattern with support >40%	Supporting Customers
(Beer)(Bread)	B.Adams,B.Moore
(Beer)(Wine,Cider)	J.Brown,B.Moore

ข้อควรระวังของเทคนิค Sequencial Pattern คือ

1.) Support Factor ระบุค่า parameter ได้เพียง 1 ค่า

2.) จำนวนข้อมูลถ้ามีมาก จำเป็นต้องมีการตรวจสอบให้มั่นใจว่ามีรหัสที่เป็นตัวแทนของข้อมูล Transaction ของลูกค้าแต่ละราย

3.) ต้องมี Field พิเศษเพื่อเป็นตัวแทนของลูกค้า โดยทั่วไปฐานข้อมูลการขายสินค้ามักจะไม่เก็บรหัสลูกค้าไว้ใน Transaction

4.) เพื่อให้การทำงานดีขึ้น ข้อมูลจะต้องเก็บเรียงตามลำดับเหตุการณ์ที่เกิดขึ้นของลูกค้าแต่ละรายๆไป

3.3) Similar Time Sequence

ใช้ค้นหาความเกี่ยวเนื่องกันระหว่างกลุ่มข้อมูล 2 กลุ่ม ซึ่งขึ้นต่อกันทางด้านเวลา โดยมีรูปแบบการเคลื่อนไหวเหมือนกัน แทนข้อมูลในแนวแกน X ด้วยค่าของเวลา เช่น วัน เดือน ส่วน แกน Y แทนด้วยค่าของตัวแปรที่สนใจ จึงนิยมใช้เทคนิคนี้สำหรับการดูแนวโน้มยอดขาย เพื่อเตรียมดูแล

เรื่อง stock สินค้า และสามารถดูได้ว่า ณ ช่วงเวลาแต่ละวันหรือ แต่ละสัปดาห์ยอดขายสิตค้าใดมีค่าใกล้เคียงกัน ซึ่งแสดงผลด้วยกราฟ

ข้อดีของเทคนิคนี้ คือ แสดงถึงการเคลื่อนไหวของข้อมูล เช่น ยอดขายที่เปลี่ยนแปลง การเคลื่อนไหวของราคาสินค้า และ การเคลื่อนไหวของ stock สินค้า

จากที่กล่าวมาข้างต้น ทั้ง 3 เป็นเทคนิคหลักของ Link Analysis หัวข้อต่อไปนี้กล่าวถึง algorithm พื้นฐานของ Association Rule เพื่อนำไปใช้ในการหาความสัมพันธ์ของข้อมูลการใช้บริการธนาคาร

3.4 Apriori Algorithm

Association Rule เป็นเทคนิคของ data miningในการหาความสัมพันธ์ระหว่างสินค้าในฐานข้อมูลรายการขาย กฎที่ได้จะอยู่ในรูปกฎความสัมพันธ์ เช่น A-> B เช่นลูกค้าที่ทำการซื้อ keyboard มีแนวโน้มในการซื้อ mouse ไปด้วยโดยกฎที่ได้จะแสดงในรูปแบบ *"Keyboard -> Mouse [support=6% , confidence=70%]* " กล่าวคือ ค่า confidence ของกฎคือ 70% แสดงจำนวนเหตุการณ์ที่ซื้อทั้ง keyboard และ mouse เทียบกับจำนวนเหตุการณ์ซื้อ mouse ทั้งหมด ส่วนค่า support ของกฎคือ6% คือจำนวนเหตุการณ์ที่ซื้อทั้ง keyboard และ mouse เทียบกับจำนวนเหตุการณ์ขายทั้งหมดในฐานข้อมูล

Apriori Algorithm เป็นอัลกอริทึมหนึ่งที่เป็นที่ยอมรับในการหาความสัมพันธ์ของข้อมูล โดยอาศัยหลักการทำงานเป็นรอบๆซ้ำๆ นั่นคือใช้รายการใน k-itemset (เซตของข้อมูลที่มีสมาชิก k ตัว ซึ่งเรียกสมาชิกแต่ละตัวว่า item) ในการหารายการใน (k+1)-itemset ต่อไป โดยมีการกำหนดค่า ***"Minimum Support"*** เป็นเงื่อนไขในการเลือกข้อมูลเพื่อทำในรอบต่อๆไปว่าข้อมูลนั้นๆจะต้องมีค่า Support มากกว่าหรือเท่ากับ Minimum Support ที่กำหนด โดยจะทำการวนซ้ำๆไปเรื่อยๆ จนกระทั่งไม่สามารถหาชุดรายการที่มากกว่า k รายการ(k-itemset) ได้อีกแล้วก็จะนำชุดรายการจากรอบนั้นๆ มาสร้างกฎความสัมพันธ์

สำหรับ itemset ที่มีความถี่ในการเกิดรายการมากกว่าหรือเท่ากับ ค่า minimum support จะเรียก itemset นั้นว่า ***"Frequent Itemset"*** ซึ่งมีข้อตกลงว่า **"Subset ของ Frequent Itemset จะต้องเป็น Frequent Itemset ด้วย"** เช่น ถ้า ABCD เป็น Frequent itemset แล้ว AB ซึ่งเป็น subset ของ ABCD ก็จะเป็น Frequent Itemset ด้วย

การทำงานหลักของ Apriori Algorithm ได้แก่การค้นหาชุดรายการ(k-itemset)ที่เป็น Frequent Itemset เพื่อสำหรับนำไปสร้างเป็นกฎความสัมพันธ์ โดยขั้นตอนในการหาชุดรายการ ประกอบไปด้วย 2 ขั้นตอนหลัก ได้แก่ **Join Step** และ **Prune Step**

3.4.1 การรวมรายการ (Join Step)

ในขั้นตอนการ Join นี้จะเป็นการสร้าง C_k ซึ่งเป็นเซตของข้อมูลที่ถูกเลือกมา k รายการ โดย เกิดจากการรวมกันของ L_{k-1} ด้วยตัวเอง(Self – joining L_{k-1})โดยสัญลักษณ์ต่างๆที่ใช้มีดังนี้

ตารางที่ 3.8 สัญลักษณ์ใน Apriori Algothrim

D	แทน Database แต่ละ Transaction เก็บ <TID,items>
TID	แทน ตัวเลขระบุรายการ Transaction
size	แทน จำนวน Item ในแต่ละ set ข้อมูล
k-Itemset	แทน set ของข้อมูลที่แต่ละ set ประกอบด้วยสมาชิกจำนวน k ตัว
L_k	Set ของ Frequent itemset ซึ่งมีสมาชิก k ตัว ซึ่งทุก subset มีค่า support มากกว่าเท่ากับ minimum support โดยสมาชิกใน set ประกอบด้วย {itemset , support count}
C_k	Set ของ Frequent itemset ที่เกิดจากการ join กันของ L_{k-1} กับ L_{k-1} โดยสมาชิกใน set ประกอบด้วย {itemset , support count}
$L_{k-1} \infty L_{k-1}$	แทน การ Self-join ของ L_{k-1} เพื่อทำการหา C_k

ในการ Join กัน ($L_{k-1} \infty L_{k-1}$) เพื่อทำการหา Candidate Itemset (C_k) โดยจะมีการกำหนดค่า minimum support ไว้ด้วย เมื่อทำการพิจารณาสมาชิกใน L_{k-1} จะทำการ join ได้นั้นสมาชิกตำแหน่งที่ 1 ถึง k-2 จะต้องเหมือนกัน ทั้งนี้เพื่อเป็นการป้องกันไม่ให้เกิดรายการซ้ำกันได้ โดยที่ข้อมูลภายใน L_{k-1}นั้น มีการเรียงลำดับจากน้อยไปมาก กล่าวคือ

$$L_{k-1} = \{ item[1], item[2], item[3], \ldots, item[k-1] \}$$

$$where\ item[1] < item[2] < item[3] \ldots < item[k-1]$$

ดังนั้นการ join กันจะอยู่ภายใต้เงื่อนไขดังนี้

If ($item_1[1] = item_2[1]$) and ($item_1[2] = item_2[2]$) and ... ($item_1[k-2] = item_2[k-2]$) and ($item_1[k-1] < item_2[k-1]$) then

$$L_{k-1} \infty L_{k-1} = \{ item_1[1], item_1[2], \ldots, item_1[k-2], item\,2[k-2] \}$$

ซึ่งแต่ละรอบจะทำการนับค่า support สำหรับแต่ละ item โดยเลือกเฉพาะ item ที่มีค่า support มากกว่าหรือเท่ากับ minimum support และในตอนจบแต่ละลูปจะได้ C_k ไปเป็นตัวตั้งต้นในการหา L_kต่อไป

3.4.2 การตัดรายการ (Prune Step)

ในขั้นตอนของการ Pruning ทำเพื่อลดกฎที่ได้ให้มีจำนวนลดลง โดยจะทำการเลือกเฉพาะ C_k ที่ subset แต่ละตัวของ C_k ที่มีสมาชิก k-1 ตัว เป็นสมาชิกในลำดับก่อนหน้านี้ (L_{k-1})

(1) L_1 = {large 1-itemsets} ;
(2) **for** ($k = 2$; $L_{k-1} \neq \phi$; k++) **do begin**
(3) C_k = apriori-gen(L_{k-1}) ; // New candidates
(4) **forall** transactions $t \in D$ **do begin**
(5) C_t = subset(C_k , t) ; // Candidates contained in t
(6) **forall** candidates c $\in C_t$ **do**
(7) $c.count$ ++ ;
(8) **End**
(9) L_k = { c $\in C_k$ | $c.count \geq min\_sup$}
(10) **End**
(11) **return** $\bigcup_k L_k$;

รูปที่ 3.4 Apriori Algorithm

apriori-gen(L_{k-1} : frequent (k-1)-itemset) ;

(1) forall itemset $p \in L_{k-1}$
(2) forall itemset $q \in L_{k-1}$
(3) If (p[1] = q[1]) ^ (p[2] = q[2]) ^ … ^ (p[k-2] = q[k-2]) ^ (p[k-1] < q[k-1]) then
(4) c = p ∞ q ; // join step : generate candidates
(5) If **has_infrequent_subset**(c , L_{k-1}) then
(6) delete **c** ;
(7) else add **c** to C_k // prune step
(6) end
(7) End
(8) return C_k ;

รูปที่ 3.5 Algorithm สร้าง Candidate Itemset

has_infrequent_subset(c , L_{k-1}) ;

(1) **forall** ***itemsets c in*** C_k do

(2) **forall** ***(k-1)-subsets of c*** do

(3) If (s $\notin$ L_{k-1}) then delete c from C_k

รูปที่ 3.6 การทำ Prunning

ตัวอย่างการทำงานของ Apriori Algorithm โดยกำหนดให้ฐานข้อมูลแสดงรายการการขายสินค้าทั้งหมด 4 รายการ โดยให้แต่ละรายการมีเลขที่(TID) เป็น 100, 200 , 300 และ 400 และแต่ละรายการประกอบไปด้วยรายการสินค้าที่ขายได้(item) เช่น TIDที่ 100 มีการขายสินค้าที่ 1,3 และ 4 เป็นต้น ดังแสดงในตารางที่ 3.9 โดยกำหนดค่า minimum support = 2

ตารางที่ 3.9 ตัวอย่างรายการขายสินค้า

TID	Items
100	1 , 3 , 4
200	2 , 3, 5
300	1 , 2 , 3 , 5
400	2 , 5

ขั้นตอนการทำงานมีดังนี้

1. ในการทำงานรอบแรก ทำการเรียงลำดับรายการในแต่ละ TID เพื่อง่ายต่อการ join กัน โดยในรอบแรกนี้จะทำการหา C_1 โดยจะทำการแยกสินค้าที่อยู่ในแต่ละ TID ออกมาทั้งหมด และนับจำนวนการเกิดทั้งหมดของแต่ละรายการสินค้าในฐานข้อมูล(support)

2 . ทำการตัดรายการสินค้าที่จำนวนการเกิดน้อยกว่า min_sup ที่กำหนด รายการที่เหลือเป็น Frequent Itemset(L_1)

3. รอบที่สอง ทำการค้นหารายการ(C_2) ที่เกิดจากการ join ของ $L_1 \infty L_1$

4. ทำการนับจำนวนการเกิดแต่ละรายการ(support) โดยทำการเปรียบเทียบกับค่า min_sup โดยตัดรายการที่ค่า support น้อยกว่า min_sup ไปทำให้ได้รายการที่เป็น L_2

5. รอบที่สาม ทำการค้นหารายการ(C_3) ที่เกิดจากการ join กันของ L_2 สามารถแสดงรายละเอียดได้ดังนี้

$C_3 = L_2 \infty L_2$

$= \{ \{1,3\},\{2,3\},\{2,5\},\{3,5\} \} \infty \{ \{1,3\},\{2,3\},\{2,5\},\{3,5\} \}$

$= \{ \{1,2,3\} , \{2,3,5\} \}$

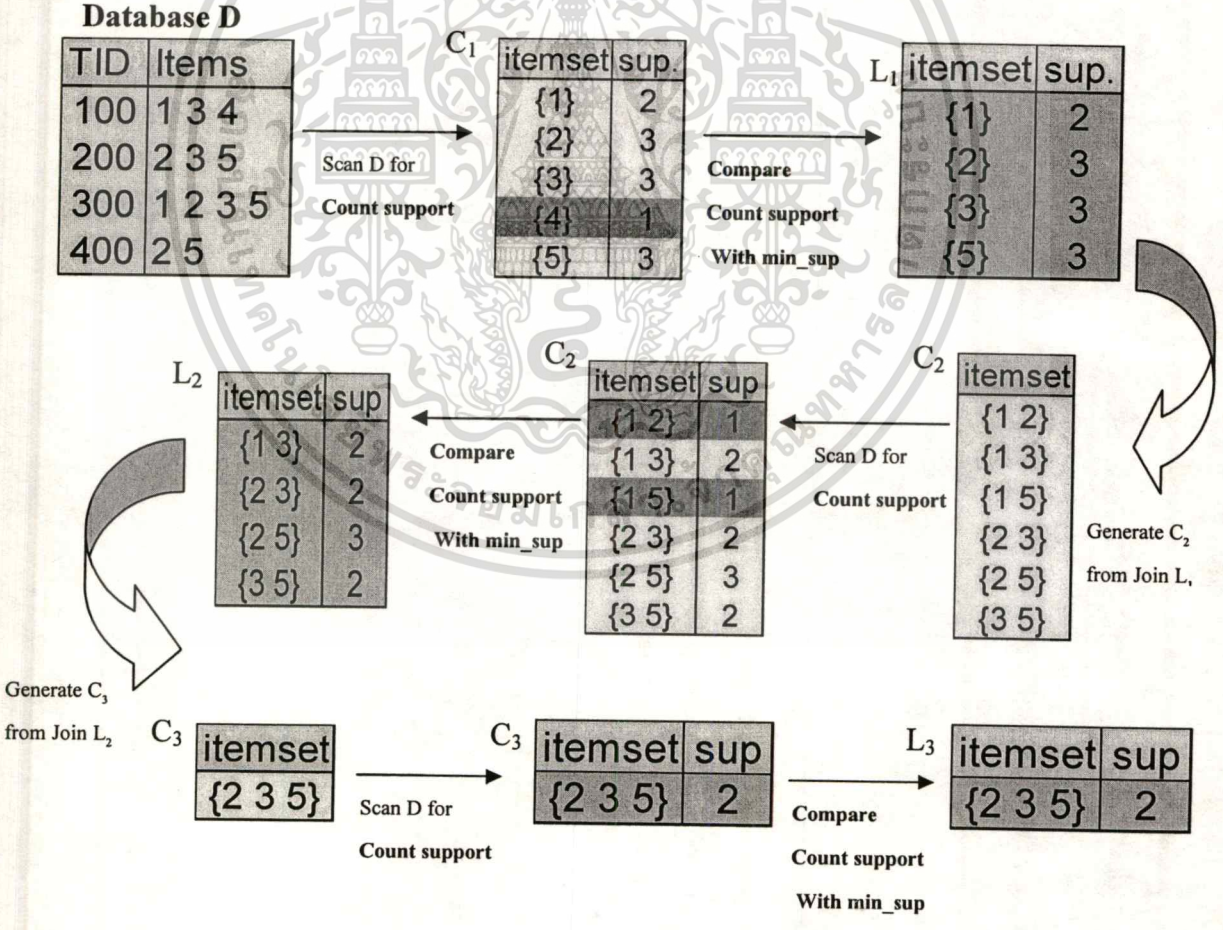
จากการ join ทำให้ได้รายการ 2 ชุดรายการ ซึ่งประกอบไปด้วย {1,2,3} , {2,3,5} ขั้นต่อไปคือการตัดรายการ (prune) โดยจะหา 2- itemsets ของแต่ละชุดรายการต้องเป็นสมาชิกอยู่ใน L_2 รายละเอียดมีดังนี้

- {1,2,3} ประกอบไปด้วย 2-itemsets คือ {1,2} , {1,3} ,{2,3} แต่พบว่า {1,2} ไม่เป็นสมาชิกอยู่ใน L_2 ดังนั้น {1,2,3} ไม่เป็นสมาชิกใน C_3

- {2,3,5} ประกอบด้วย 2-itemsets คือ {2,3} , {2,5} , {3,5} ซึ่งพบว่า 2-itemsets ทุกตัวเป็นสมาชิกอยู่ใน L_2 ดังนั้นมีเพียง {2,3,5} เท่านั้นเป็นสมาชิกใน C_3

6. ทำการนับจำนวนการเกิดรายการของแต่ละสมาชิกใน C_3 และทำการเปรียบเทียบกับค่า min_sup ต้องมีค่ามากกว่าเท่ากับ min_sup จะได้ชุดรายการ L_3

7. ทำการค้นหา C_4 โดยทำการ join L_3 ด้วยกัน พบว่าไม่สามารถทำการ join ได้แล้ว เนื่องจากเหลือสมาชิกเพียงตัวเดียว ทำให้ $C_4 = \varnothing$ ดังนั้นจบการทำงานที่ได้ชุดรายการ L_3



รูปที่ 3.7 ตัวอย่างการหา Frequent Itemset และ Candidate Itemset

3.4.2 การนำ Frequent Itemset มาสร้างเป็นกฎความสัมพันธ์

การสร้างกฎความสัมพันธ์ เป็นการนำเอา subset ที่ไม่ใช่เซตว่าง ที่เป็นสมาชิกของ L_k (Frequent Itemset) มาสร้างเป็นกฎ ให้อยู่ในรูปแบบดังนี้

กำหนดให้ s เป็น subset ที่ไม่ใช่เซตว่างของ l ซึ่งเป็นสมาชิกใน L_k โดย $k \geq 2$ โดยกฎที่ได้จะอยู่ในรูปแบบ "$s \Rightarrow l - s$" และจะมีการคำนวณหาค่า confidence ของแต่ละกฎ เช่น

"AB=>CD" จะได้ค่า ***confidence = support(ABCD) / support(AB)***

ซึ่งกฎที่สนใจจะต้องมี ค่าconfidenceมากกว่าหรือเท่ากับ minimum confidence(***min_conf***)

```
// Faster Algorithm
1) forall large k-itemsets l_k, k ≥ 2 do begin
2)     H_1 = { consequents of rules derived from l_k with one item in the consequent };
3)     call ap-genrules(l_k, H_1);
4) end

procedure ap-genrules(l_k: large k-itemset, H_m: set of m-item consequents)
    if (k > m + 1) then begin
        H_{m+1} = apriori-gen(H_m);
        forall h_{m+1} ∈ H_{m+1} do begin
            conf = support(l_k)/support(l_k − h_{m+1});
            if (conf ≥ minconf) then
                output the rule (l_k − h_{m+1}) ⟹ h_{m+1} with confidence = conf and support = support(l_k);
            else
                delete h_{m+1} from H_{m+1};
        end
        call ap-genrules(l_k, H_{m+1});
    end
```

รูปที่ 3.8 Algorithm ในการ Generate Rule

ตัวอย่างการสร้างกฎความสัมพันธ์ กำหนดให้ min_conf = 70% เมื่อนำFrequent Itemset (L_k) ที่ได้จากขั้นตอนที่ 3.4.1 มาสร้างเป็นกฎ เช่น L_3 = { {2,3,5} } ซึ่งประกอบไปด้วยสมาชิก l = {2,3,5} ซึ่งประกอบไปด้วยเซตย่อยดังนี้ {2,3} , {2,5} , {3,5} , {2} , {3} , {5} นำมาสร้างกฎได้ ดังตาราง

ตารางที่ 3.10 กฎความสัมพันธ์ที่ได้จาก L_k

กฎที่	กฎ	Confidence
1.	{2,3} => {5}	2/2 = 100%
2.	{2,5} => {3}	2/3 = 66.67%
3.	{3,5} => {2}	2/2 = 100%
4.	{2} => { 3,5}	2/3 = 66.67%
5	{3} => { 2,5}	2/3 = 66.67%
6	{5} => { 2,3}	2/3 = 66.67%

จากกฎที่ได้ออกมาเมื่อนำมาเปรียบเทียบกับ min_conf (70%) ที่กำหนด จะพบว่ามีเพียงกฎความสัมพันธ์ที่ 1,3 เท่านั้น เนื่องจากสมาชิกใน L_3 มีเพียงตัวเดียว จึงพิจารณาต่อ L_2ซึ่ง สมาชิกประกอบด้วย {1,3},{2,3},{2,5},{3,5} นำสมาชิกแต่ละตัวมาหา subset และสร้างกฎความสัมพันธ์ตามวิธีข้างต้นต่อไป

	Body	Implies	Head	Supp (%)	Conf (%)	F	G	H	I
1	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00'	28.45	40.4				
2	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00'	20.46	29.05				
3	cost(x) = '0.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	59.17	84.04				
4	cost(x) = '0.00~1000.00'	==>	revenue(x) = '1000.00~1500.00'	10.45	14.84				
5	cost(x) = '0.00~1000.00'	==>	region(x) = 'United States'	22.56	32.04				
6	cost(x) = '1000.00~2000.00'	==>	order_qty(x) = '0.00~100.00'	12.91	69.34				
7	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '0.00~500.00'	28.45	34.54				
8	order_qty(x) = '0.00~100.00'	==>	cost(x) = '1000.00~2000.00'	12.91	15.67				
9	order_qty(x) = '0.00~100.00'	==>	region(x) = 'United States'	25.9	31.45				
10	order_qty(x) = '0.00~100.00'	==>	cost(x) = '0.00~1000.00'	59.17	71.86				
11	order_qty(x) = '0.00~100.00'	==>	product_line(x) = 'Tents'	13.52	16.42				
12	order_qty(x) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	23.88				
13	product_line(x) = 'Tents'	==>	order_qty(x) = '0.00~100.00'	13.52	98.72				
14	region(x) = 'United States'	==>	order_qty(x) = '0.00~100.00'	25.9	81.94				
15	region(x) = 'United States'	==>	cost(x) = '0.00~1000.00'	22.56	71.39				
16	revenue(x) = '0.00~500.00'	==>	cost(x) = '0.00~1000.00'	28.45	100				
17	revenue(x) = '0.00~500.00'	==>	order_qty(x) = '0.00~100.00'	28.45	100				
18	revenue(x) = '1000.00~1500.00'	==>	cost(x) = '0.00~1000.00'	10.45	96.75				
19	revenue(x) = '500.00~1000.00'	==>	cost(x) = '0.00~1000.00'	20.46	100				
20	revenue(x) = '500.00~1000.00'	==>	order_qty(x) = '0.00~100.00'	19.67	96.14				
21									
22									
23	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
24	cost(x) = '0.00~1000.00'	==>	revenue(x) = '0.00~500.00' AND order_qty(x) = '0.00~100.00'	28.45	40.4				
25	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
26	cost(x) = '0.00~1000.00'	==>	revenue(x) = '500.00~1000.00' AND order_qty(x) = '0.00~100.00'	19.67	27.93				
27	cost(x) = '0.00~1000.00' AND order_qty(x) = '0.00~100.00'	==>	revenue(x) = '500.00~1000.00'	19.67	33.23				

Sheet1

รูปที่ 3.9 กฎความสัมพันธ์ และค่า Support , Confidence

บรรณานุกรม

Simoudis,E. 1998. **Discovering Data Mining From Concept to implementation.** New Jersey : Prentice Hall.

Michael J. A. Berry and Gordon, S. Linoff. 1997. **Data Mining Techniques For Marketing , Sales and Customer Support.** , Wiley Computer and Son.

Agrawal, R. and Srikant, R. 1994. **Fast Algorithms for Mining Association Rules.** [Online]. Available: http://citeseer.nj.nec.com/agrawal94fast.html

Estelle, B. and Rob,G. 1998. **Association and Sequencing.** [Online]. Available: http://www.dbmsmag.com/9807m03.html

Han, J. and Kamber, M. 2000. **Data Mining : Concepts and Techniques.** [Online]. Available : http://www.cs.sfu.ca/~han/dmbook

Sander, J. 2002. **Knowledge Discovery in Databases : Association Rules.** [Online]. Available: www.cs.ualberta.ca/~joerg/courses/cmput690/ slides/AssociationRules-s4.pdf

บทที่ 4
การวิเคราะห์และพัฒนาระบบงาน

4.1 ลักษณะการดำเนินธุรกิจ

บริษัททำนำข้อมูลมาวิเคราะห์ในครั้งนี้ ดำเนินธุรกิจเป็นธนาคารพาณิชย์ ให้บริการด้านการเงิน ไม่ว่าจะเป็นเงินฝาก สินเชื่อ การลงทุน ตลอดจนผลิตภัณฑ์เพื่ออำนวยความสะดวกในการใช้จ่าย ไม่ว่าจะเป็นบัตร ATM , บัตรเครดิตประเภทต่างๆ และบัตร Debit เป็นต้น โดยลูกค้าสามารถใช้บริการธนาคารผ่านช่องทางต่างๆได้มากขึ้น นอกจากการเข้ารับบริการที่เคาน์เตอร์ของสาขาแล้ว ได้แก่ การถอนเงินผ่านเครื่อง ATM , การรับฝากเงินสดผ่านเครื่องรับฝากเงินอิเล็กทรอนิกส์ หรือการทำธุรกรรมทางการเงินผ่านทาง Internet เป็นต้น ซึ่งพบว่าในปัจจุบันธนาคารมีผลิตภัณฑ์ต่างๆมากมายหลายประเภท ดังนั้นการนำเสนอผลิตภัณฑ์เหล่านั้นให้กับกลุ่มลูกค้าที่เหมาะสม จึงเป็นสิ่งที่สำคัญอย่างยิ่ง เพื่อเป็นการเพิ่มศักยภาพการบริการลูกค้าแต่ละกลุ่มให้มีประสิทธิภาพ และสร้างความพึงพอใจต่อลูกค้ามากที่สุดด้วย

4.2 วัตถุประสงค์

ในปัจจุบันการแข่งขันเพื่อเพิ่มส่วนแบ่งตลาดสำหรับธุรกิจธนาคารมีสูง แต่ละธนาคารต้องหากลยุทธ์ และวิธีการเพื่อที่จะรักษาลูกค้าเดิมไว้ และสร้างลูกค้าใหม่ให้เข้ามาใช้บริการของธนาคาร โดยอาจจะมีการเพิ่มผลิตภัณฑ์ที่น่าสนใจ หรือการบริการใหม่ๆที่สะดวก รวดเร็วและสร้างความพึงพอใจให้กับลูกค้ามากขึ้น ดังนั้นการที่ธนาคารจะสร้างผลิตภัณฑ์ หรือบริการใหม่ๆได้นั้นจะต้องมีความเข้าใจในพฤติกรรมของลูกค้าว่าเป็นเช่นไร ดังนั้นจึงมีแนวความคิดที่จะหาความสัมพันธ์ของข้อมูลการใช้บริการธนาคารของลูกค้า เพื่อวิเคราะห์ลักษณะผลิตภัณฑ์ และการใช้บริการธนาคารของลูกค้าธนาคารพาณิชย์ สำหรับจุดประสงค์ ในการประเมินประสิทธิภาพของผลิตภัณฑ์ในปัจจุบัน และ กำหนดกลยุทธ์ทางการตลาด และ สร้างโอกาสทางการตลาดในอนาคต

โดยได้นำเอาแนวคิด Association Rule สำหรับวิเคราะห์การใช้บริการของลูกค้า โดยอาศัยหลักการคือการระบุรูปแบบที่สำคัญของผลิตภัณฑ์ที่ถูกซื้อไปด้วยกัน หรือ product basket โดยการวิเคราะห์รูปแบบนี้ ทำให้สามารถระบุแนวโน้มของสินค้าหรือบริการที่ซึ่งบ่อยครั้งจะถูกซื้อไปด้วยกัน หรือ ไม่ถูกซื้อไปด้วยกัน ซึ่งสามารถนำมาประสานเข้ากับวัตถุประสงค์ของธนาคารเพื่อกำหนดนโยบาย สร้างกลยุทธ์เพิ่มยอดขาย สร้างช่องทางการตลาดใหม่ๆในอนาคตต่อไป

4.3 การคัดเลือกข้อมูล

ข้อมูลที่นำมาใช้ในการดำเนินงานทั้งหมด รวบรวมมาจากฐานข้อมูลลูกค้า และการใช้บริการธนาคาร ที่จัดเก็บอยู่ในระบบ Datawarehouse บน Mainframe ทำให้ข้อมูลที่ได้มีความน่าเชื่อถือและมีความถูกต้อง โดยข้อมูลที่นำมาศึกษาเป็นข้อมูลตั้งแต่ มกราคม 2546 ถึง ธันวาคม 2546 โดยได้เลือกข้อมูลบางส่วนมาจัดเก็บใน Microsoft Access โดยเลือกมาเฉพาะข้อมูลบางส่วนที่ใช้ในการทำ data mining ซึ่งสามารถแสดงออกมาเป็นตารางที่เกี่ยวข้องได้ดังนี้

1. ตารางข้อมูลลูกค้า (CUSTOMER) ซึ่งเก็บข้อมูลเกี่ยวกับลูกค้า ได้แก่ รหัสลูกค้า อายุ เพศ รายได้ เป็นต้น ดังแสดงในตารางที่ 4.1

2. ตารางข้อมูลบัญชี (ACCOUNT) ซึ่งเก็บข้อมูลเกี่ยวกับบัญชี ได้แก่ เลขที่บัญชี ประเภท สถานะ เป็นต้น ดังแสดงในตารางที่ 4.2

3. ตาราง Product (PRODUCT) ซึ่งเก็บข้อมูลเกี่ยวกับผลิตภัณฑ์ต่างๆ ได้แก่ รหัส product รหัสกลุ่ม product เป็นต้น ดังแสดงในตารางที่ 4.3

4. ตารางTransaction (TRANSACTION) ซึ่งเก็บข้อมูลของรายการที่เกิดขึ้นกับบัญชี ได้แก่ วันเวลาที่เกิดรายการ จำนวนเงิน เป็นต้น ดังแสดงในตารางที่ 4.4

5. ตาราง Channel (CHANNEL) ซึ่งเก็บข้อมูลช่องทางการให้บริการ ได้แก่ รหัส channel รายละเอียด ดังแสดงในตารางที่ 4.5

6. ตาราง Group Product (GROUP_PRODUCT) ซึ่งเก็บข้อมูลกลุ่มของผลิตภัณฑ์และช่องทางการให้บริการ ดังแสดงในตารางที่ 4.6

7.ตาราง Ref_code (REF_CODE) ซึ่งเก็บข้อมูลของรายละเอียดของรหัสต่างๆที่ใช้ เช่น เงินเดือน , อาชีพ เป็นต้น ดังแสดงในตารางที่ 4.7

รายละเอียดต่างๆแสดงในแต่ละตาราง จะกำหนดว่าข้อมูลใน Field นั้นๆสามารถเป็นค่าว่างได้หรือไม่ และเป็นข้อมูลประเภทใด ดังต่อไปนี้

ตารางที่ 4.1 ตารางข้อมูลลูกค้า

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**CUST_ID**	**CHAR**	**20**	**0**	**N**
2	C_NAME	CHAR	50	0	Y
3	C_SEX	CHAR	1	0	Y
4	C_BTH_DTE	CHAR	10	0	Y
5	C_INCOME	CHAR	2	0	Y
6	C_OCCUP	CHAR	2	0	Y
7	C_EDU	CHAR	2	0	Y

ตารางที่ 4.2 ตารางข้อมูลบัญชี

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**CUST_ID**	**CHAR**	**20**	**0**	**N**
2	**ACCT_ID**	**CHAR**	**20**	**0**	**N**
3	**PRD_CD**	**CHAR**	**4**	**0**	**N**
4	ACCT_STATUS	CHAR	2	0	Y
5	ACCT_OPEN_DTE	CHAR	10	0	Y
6	ACCT_CLOSE_DTE	CHAR	10	0	Y

ตารางที่ 4.3 ตาราง Product

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**GRP_CD**	**CHAR**	**4**	**0**	**N**
2	**PRD_CD**	**CHAR**	**4**	**0**	**N**
3	PRD_DESC	CHAR	100	0	Y

ตารางที่ 4.4 ตาราง Transaction

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**CUST_ID**	**CHAR**	**20**	**0**	**N**
2	**ACCT_ID**	**CHAR**	**20**	**0**	**N**
3	**TRAN_DTE**	**CHAR**	**10**	**0**	**N**
4	**TRAN_TIME**	**CHAR**	**10**	**0**	**N**
5	PRD_CD	CHAR	4	0	Y
6	TRAN_CHN	CHAR	4	0	Y
7	TRAN_AMOUNT	DECIMAL	17	2	Y

ตารางที่ 4.5 ตาราง Channel

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**GRP_CD**	**CHAR**	**4**	**0**	**N**
2	**CHN_CD**	**CHAR**	**4**	**0**	**N**
3	CHN_DESC	CHAR	100	0	N

ตารางที่ 4.6 ตาราง Group Product

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	**GRP_CD**	**CHAR**	**4**	**0**	**N**
2	GRP_DESC	CHAR	50	0	Y

ตารางที่ 4.7 ตาราง Ref_Code

ลำดับ	ชื่อ Column	ประเภทข้อมูล	ความยาว	ทศนิยม	เป็นค่าว่าง (Y/N)
1	TABLE_CD	CHAR	4	0	N
2	VALUE_CD	CHAR	4	0	N
3	VALUE_DESC	CHAR	100	0	Y

4.4 การวิเคราะห์ และออกแบบระบบงาน

การวิเคราะห์ และออกแบบระบบงาน ประกอบด้วยขั้นตอนการทำงานหลักดังนี้

4.4.1 การจัดเตรียมข้อมูล

ในส่วนนี้ผู้ใช้สามารถทำการเลือกส่วนติดต่อกับฐานข้อมูลได้ว่าผู้ใช้ต้องการติดต่อกับฐานข้อมูลใดผ่านทาง ODBC ซึ่งเมื่อผู้ใช้เลือกฐานข้อมูลที่ต้องการจะนำมาใช้ในการวิเคราะห์แล้ว ผู้ใช้สามารถทำการกำหนดเงื่อนไขในการเลือกข้อมูลที่ต้องการนำมาวิเคราะห์ได้

4.4.2 การจัดกลุ่มข้อมูล

ในส่วนนี้เป็นการจัดกลุ่มข้อมูลของผลิตภัณฑ์หลายๆประเภท จัดให้อยู่ในกลุ่มเดียวกัน โดยสามารถเลือกประเภทผลิตภัณฑ์ที่มีอยู่ในระบบ ให้มาอยู่ในกลุ่มเดียวกันได้ หรือจะทำการลบกลุ่มที่ได้จัดไว้แล้วได้ ทั้งนี้ การจัดกลุ่มมีวัตถุประสงค์เพื่อจัดรายการผลิตภัณฑ์ที่มีอยู่หลายประเภทให้มาอยู่กลุ่มเดียวกัน เพื่อใช้สำหรับการหาความสัมพันธ์ระหว่างผลิตภัณฑ์ได้หลากหลายมากขึ้น

4.4.3 การ Mining

ในส่วนนี้เป็นการนำเอาข้อมูลที่ได้ผ่านการจัดเตรียม และจัดกลุ่มเรียบร้อยแล้ว มาผ่านกระบวนการ Mining โดยใช้ Apriori Algorithm ในการค้นหาชุดรายการ และสร้างกฎความสัมพันธ์จากชุดรายการ โดยในการ Mining จะทำการกำหนดเงื่อนไขคือ กำหนดค่า Minimum Support(min_sup) และ Minimum Confidence (min_conf) นอกจากนี้ยังสามารถกำหนดค่าสูงสุดของความสัมพันธ์ของผลิตภัณฑ์ที่ต้องการได้ด้วย

4.4.4 การแสดงผลลัพธ์

ในส่วนนี้เป็นส่วนการแสดงผลลัพธ์ที่ได้จากการทำ Mining โดยผลลัพธ์ที่ได้อยู่ในรูปแบบของกฎความสัมพันธ์ พร้อมแสดงค่า Support และ Confidence ของแต่ละกฎด้วย โดยผู้ใช้สามารถทำการเรียงข้อมูลตามค่า Support หรือ Confidence ได้ อีกทั้งสามารถทำการบันทึกกฎความสัมพันธ์ที่ได้ ในรูปแบบของ Excel file ได้อีกด้วย

4.5 สภาพแวดล้อมของการพัฒนาระบบงาน

สภาพแวดล้อมการพัฒนาระบบ ประกอบด้วย 2 ส่วน ได้แก่

4.5.1 รายละเอียดด้าน Hardware ประกอบด้วย

- เครื่องคอมพิวเตอร์ (Personal Computer) ความเร็ว CPU-PentiumIV 1.80 GHz
- หน่วยความจำ (RAM) 256 MB
- Harddisk ไม่น้อยกว่า 1 GB
- Mouse , Keyboard
- CD-ROM Drive
- Floppy Disk Drive

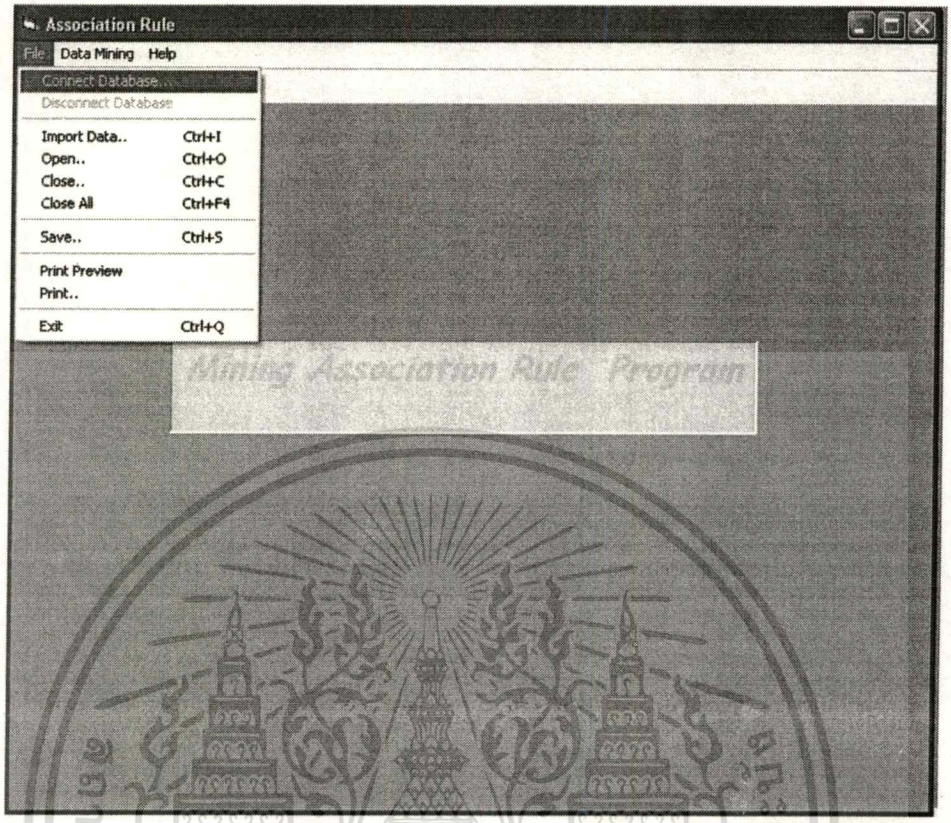
4.5.2 รายละเอียดด้าน Software ประกอบด้วย

- ระบบปฏิบัติการ Windows XP (Service Pack1)
- โปรแกรมMicrosoft Access ใช้เป็นฐานข้อมูล
- โปรแกรม Visual Basic 6.0 (Service Pack 4)
- โปรแกรม Crystal Reports

4.6 การพัฒนาโปรแกรม

การพัฒนาระบบงานในครั้งนี้ใช้ฐานข้อมูลเป็น Microsoft Access ส่วนทางด้านโปรแกรมพัฒนาด้วย Microsoft Visual Basic 6.0(SP 4) อยู่บนระบบปฏิบัติการ Window XP โดยใช้ Crystal Reports ในการสร้างรายการกฎความสัมพันธ์ที่ได้จากการประมวลผล โดยจากการวิเคราะห์และออกแบบระบบงาน ทำให้สามารถพัฒนาโปรแกรมการทำงานแบ่งออกเป็นส่วนๆต่างๆ ดังที่ได้กล่าวแล้วในหัวข้อที่ 4.4

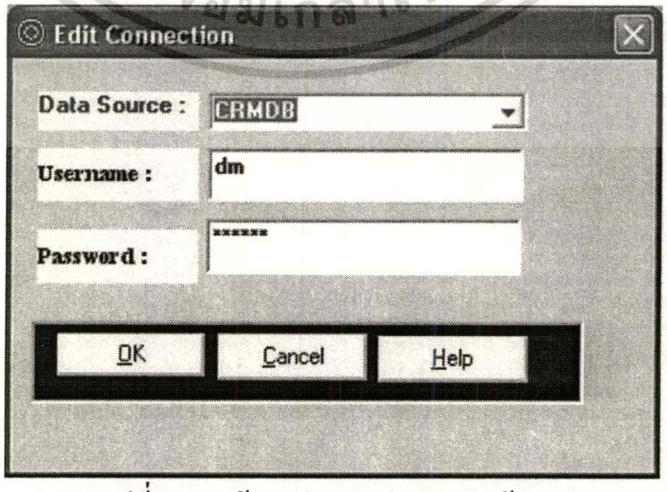
เมื่อทำการ double click โปรแกรม Association_Proj.exe จะปรากฏหน้าจอหลักของระบบดังรูป ที่ 4.1 โดยในส่วนหน้าจอหลักของโปรแกรม เมื่อกด **Menu File** ผู้ใช้สามารถทำการจัดเตรียมข้อมูลโดยการติดต่อกับฐานข้อมูล , Import ข้อมูลจาก Text file , เปิดกฎที่ได้ทำการบันทึกไว้ และเมื่อต้องการออกจากระบบสามารถเลือกที่ **Exit**



รูปที่ 4.1 หน้าจอหลักของระบบ

4.6.1 ส่วนการจัดเตรียมข้อมูล

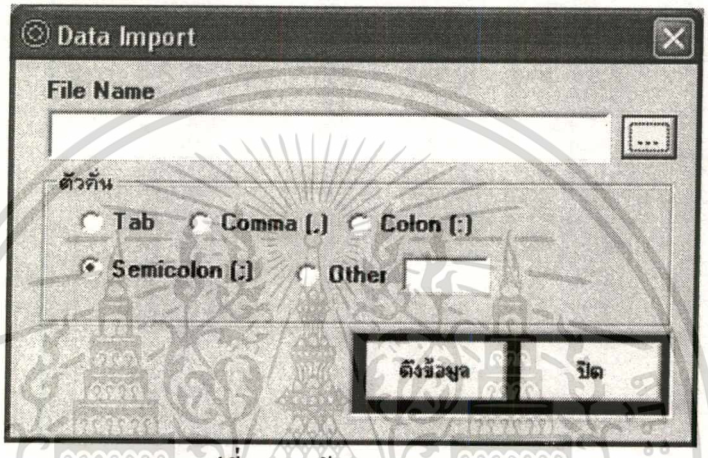
ส่วนการจัดเตรียมข้อมูลโดยผู้ใช้ทำการเลือกฐานข้อมูลที่ต้องการติดต่อผ่านทาง ***"Menu File -> Connect Database"*** จะปรากฏหน้าจอดังรูปที่ 4.2



รูปที่ 4.2 หน้าจอส่วนติดต่อกับฐานข้อมูล

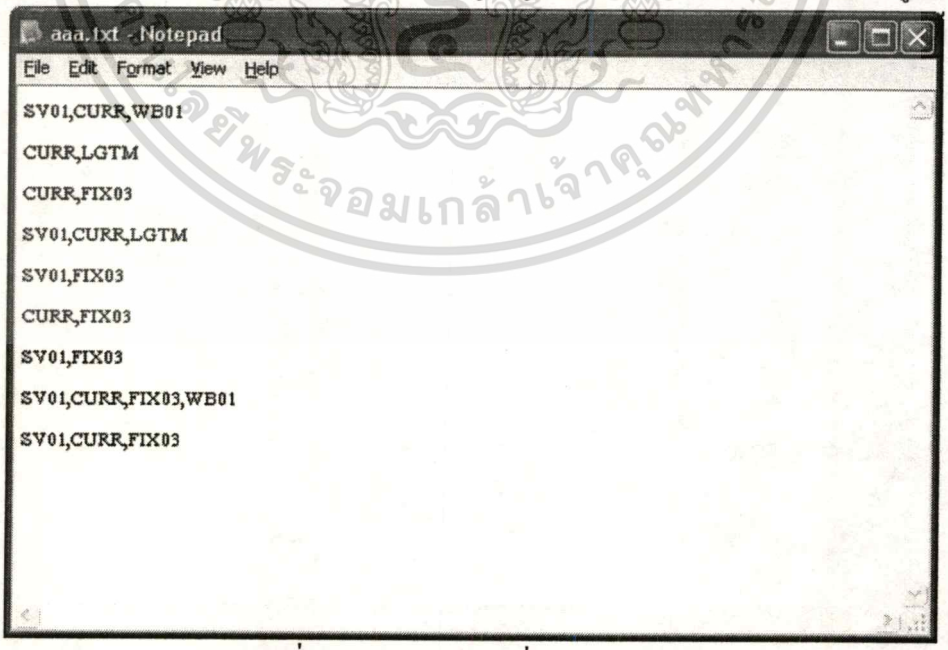
ผู้ใช้สามารถเลือกฐานข้อมูลได้จากในส่วนของ **Data Source** และกำหนด User Name และ Password สำหรับติดต่อกับฐานข้อมูลนั้นๆที่ช่อง **Username** และ **Password** โดยเมื่อกำหนดเรียบร้อยแล้ว กดปุ่ม OK จะปรากฏข้อความแสดงการ Connect สำเร็จ

กรณีที่ผู้ใช้ต้องการทำการ Import ข้อมูลจาก Text file ก็สามารถทำได้โดยการเลือก "***Menu File -> Import Data..***" จะปรากฏหน้าจอดังรูปที่ 4.3 เพื่อให้ผู้ใช้ทำการเลือก File ข้อมูล จากการกดปุ่ม ... และ กำหนดตัวคั่นข้อมูล เช่น ; หรือ tab เป็นต้น



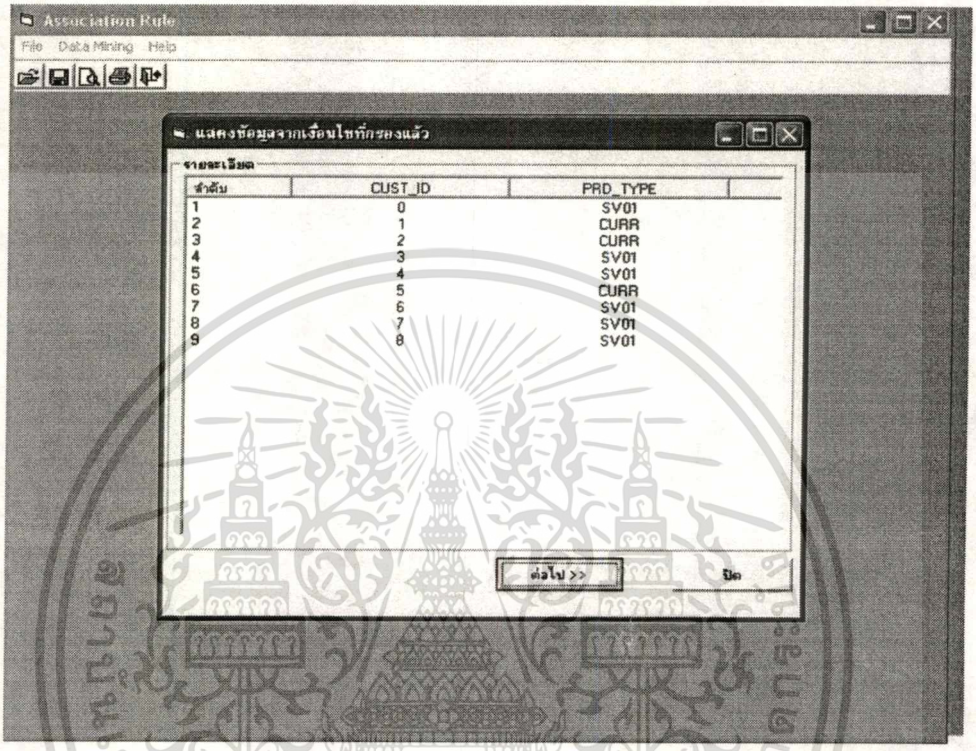
รูปที่ 4.3 หน้าจอImport Data

โดยข้อมูลที่สามารถ Import ได้นั้นจะต้องอยู่ในรูปแบบที่พร้อมจะประมวลผล ดังรูปที่ 4.4



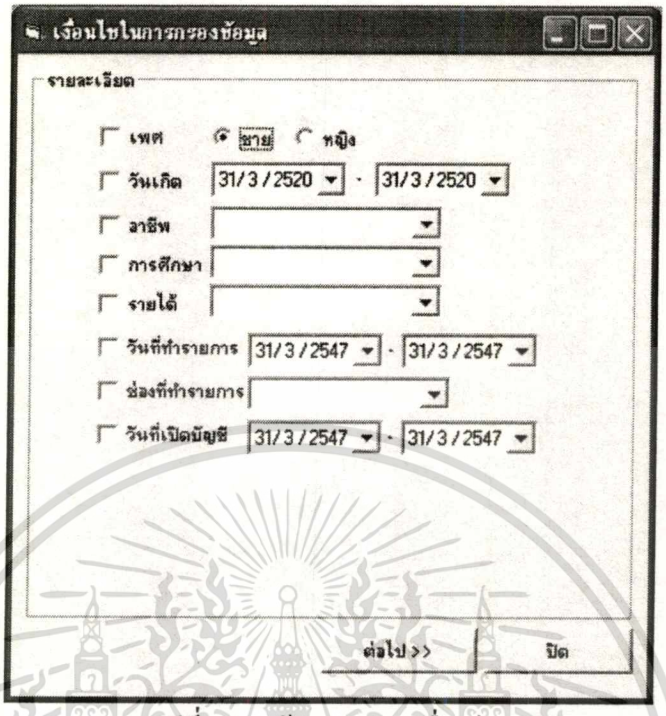
รูปที่ 4.4 ลักษณะข้อมูลที่ทำการ Import

เมื่อทำการกดปุ่ม "***ดึงข้อมูล***" จะปรากฏข้อมูลที่ทำการ Import เข้าเรียบร้อยดังรูปที่ 4.4 เพื่อทำการประมวลผลต่อไป



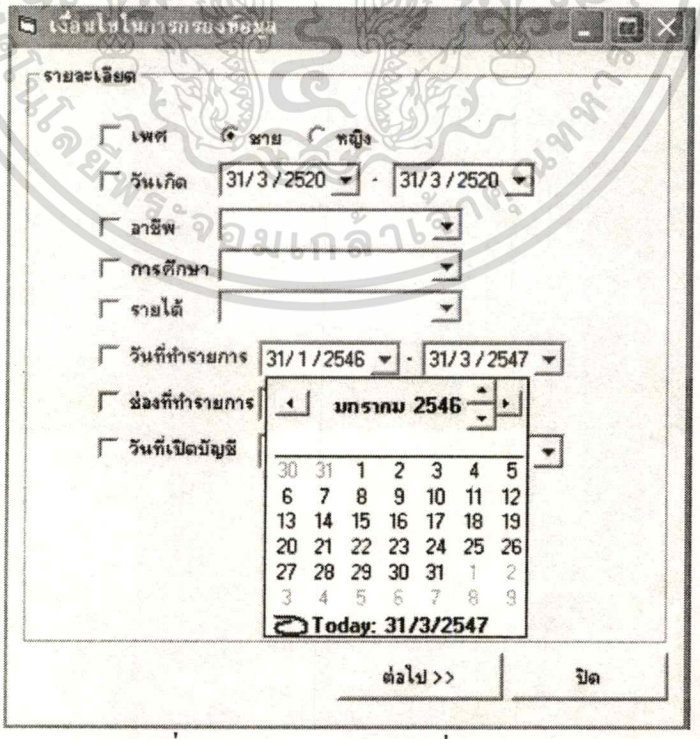
รูปที่ 4.5 แสดงข้อมูลที่ Import

เมื่อผู้ใช้ติดต่อฐานข้อมูลสำเร็จ เริ่มทำการประมวลผลโดย เลือก "***Menu Data Mining -> General Wizard..*** " จะปรากฏหน้าจอให้ทำการกำหนดเงื่อนไขในการเลือกข้อมูลดังรูปที่ 4.6



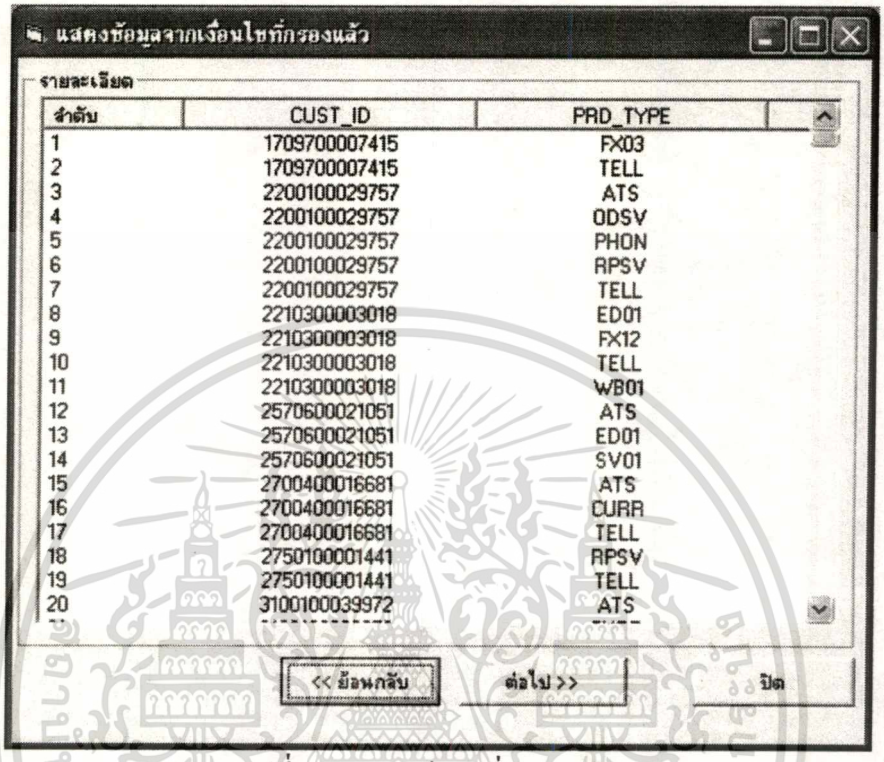
รูปที่ 4.6 หน้าจอกำหนดเงื่อนไข

โดยเงื่อนไขต่างๆ ได้แก่ เพศ , วันเกิด , อาชีพ , การศึกษา, รายได้ , วันที่ทำรายการ , ช่องทางทำรายการ และวันที่เปิดบัญชี เช่น กำหนดเลือกรายการที่ทำรายการตั้งแต่วันที่ 1 มกราคม 2546 ถึง 31 มกราคม 2546 จะทำเลือกวันที่ทำรายการ และกำหนดวันที่ ดังรูปที่ 4.7



รูปที่ 4.7 หน้าจอกำหนดวันที่ทำรายการ

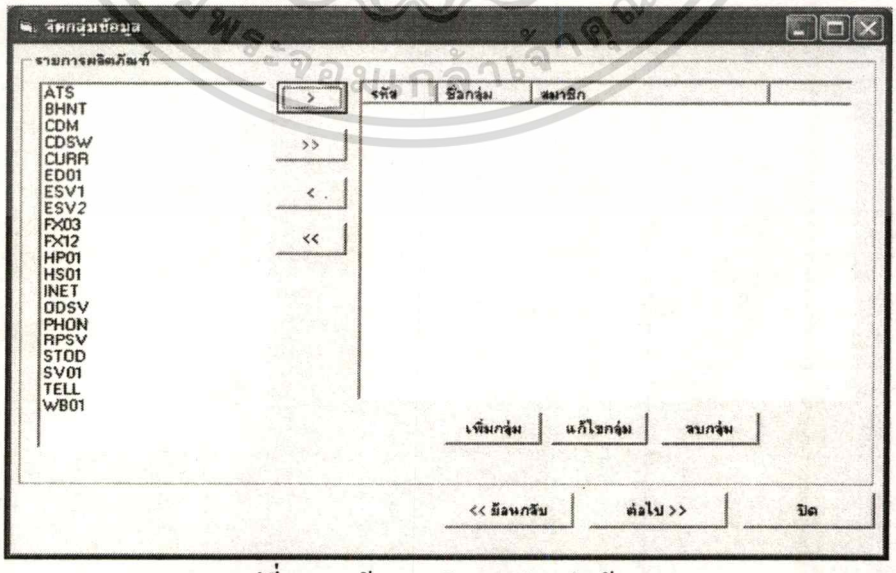
เมื่อทำการกำหนดเงื่อนไขข้อมูลเรียบร้อยแล้ว จะปรากฏข้อมูลทั้งหมดที่ได้ทำการเลือกดังรูปที่ 4.8 จากนั้นกดปุ่ม "*ต่อไป*>>" เพื่อไปยังส่วนการจัดกลุ่มข้อมูล



รูปที่ 4.8 แสดงข้อมูลที่ทำการเลือก

4.6.2 ส่วนการจัดกลุ่มข้อมูล

ส่วนนี้จะเพิ่มความสัมพันธ์ของผลิตภัณฑ์ที่มีอยู่หลากหลาย ให้มีความสัมพันธ์กันมากขึ้น ซึ่งส่วนนี้ผู้ใช้จะทำการกำหนดค่าหรือไม่ทำการกำหนดได้

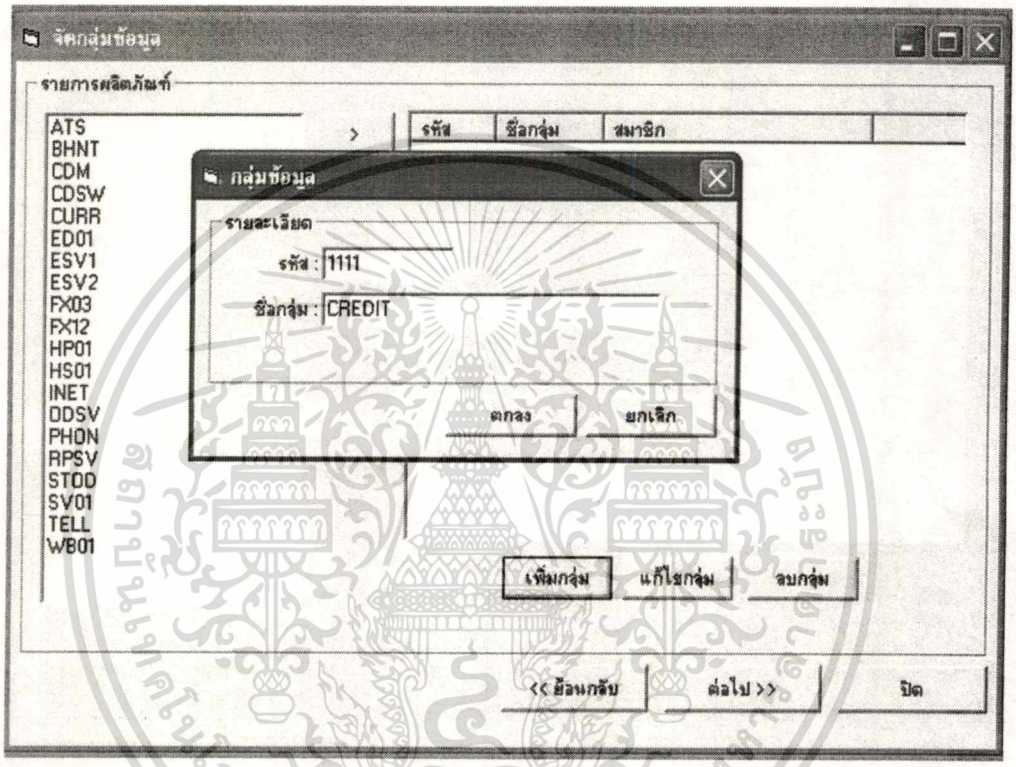


รูปที่ 4.9 หน้าจอหลักการจัดกลุ่มข้อมูล

ในส่วนนี้ส่วนประกอบที่สำคัญได้แก่

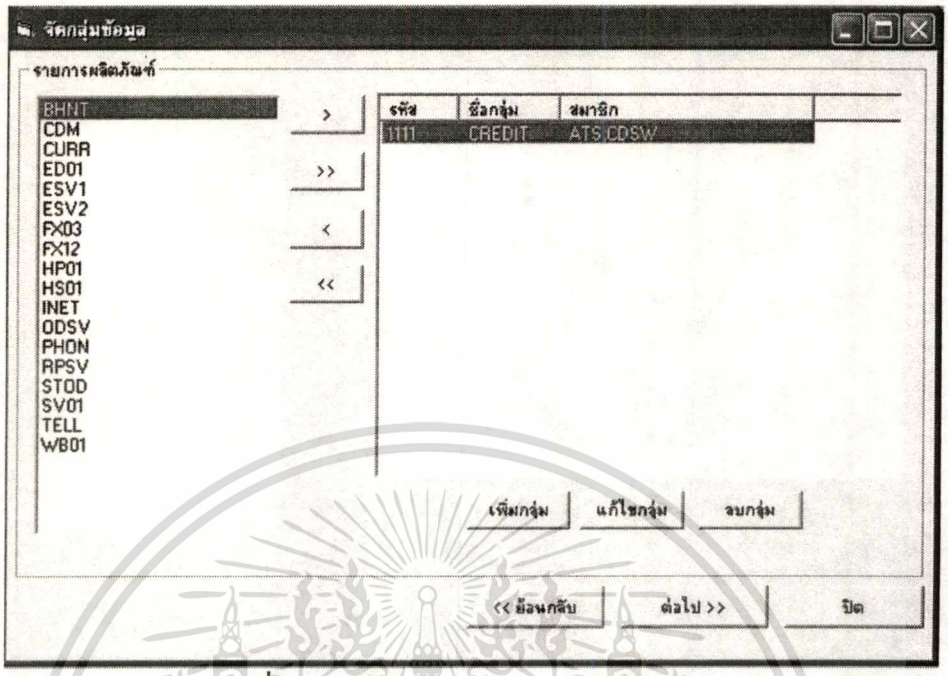
1. รายการผลิตภัณฑ์ ซึ่งจะแสดงรายการผลิตภัณฑ์ทั้งหมดของข้อมูลที่ทำการเลือก

2. การเพิ่มกลุ่ม โดยทำการเลือกปุ่ม “เพิ่มกลุ่ม” จะปรากฏหน้าจอดังรูปที่ 4.10 โดยผู้ใช้จะทำการกำหนดชื่อกลุ่ม และทำการเลือกรายการที่ต้องการให้อยู่ในกลุ่มนั้น



รูปที่ 4.10 หน้าจอการเพิ่มกลุ่ม

3. การแก้ไขกลุ่ม ผู้ใช้ทำการเลือกกลุ่มจากรายการด้านขวา และกดปุ่ม “**แก้ไขกลุ่ม**” จะปรากฏหน้าจอลักษณะเช่นเดียวกับรูปที่ 4.10 เพื่อให้ทำการแก้ไขชื่อกลุ่มได้



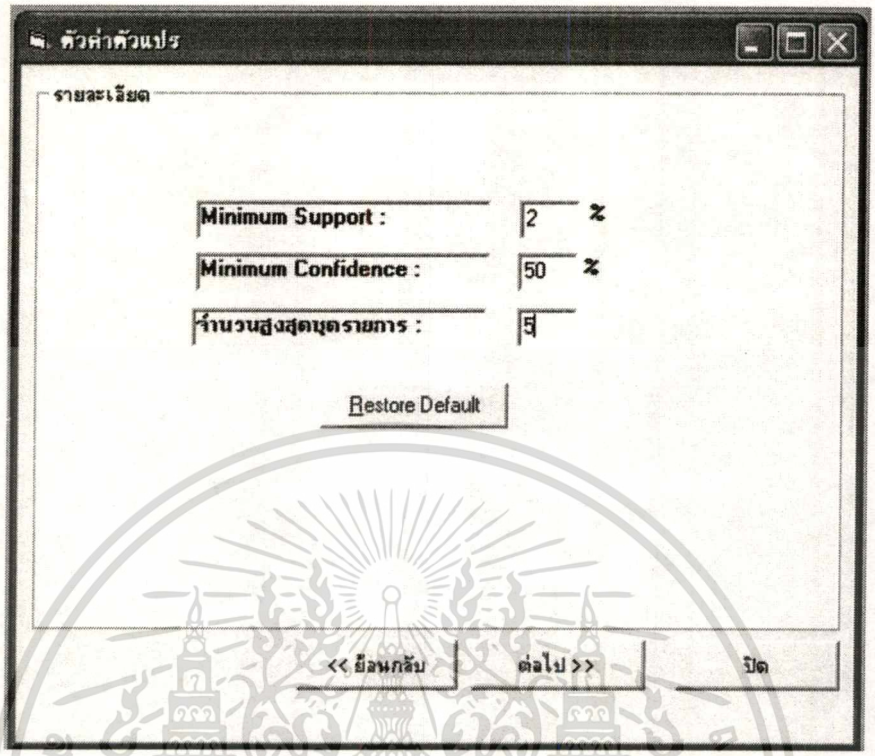
รูปที่ 4.11 หน้าจอการเลือกรายการแก้ไขกลุ่ม

4. การลบกลุ่ม ผู้ใช้ทำการเลือกรายการกลุ่มทางด้านขวามือ และทำการกดปุ่ม "***ลบกลุ่ม***"

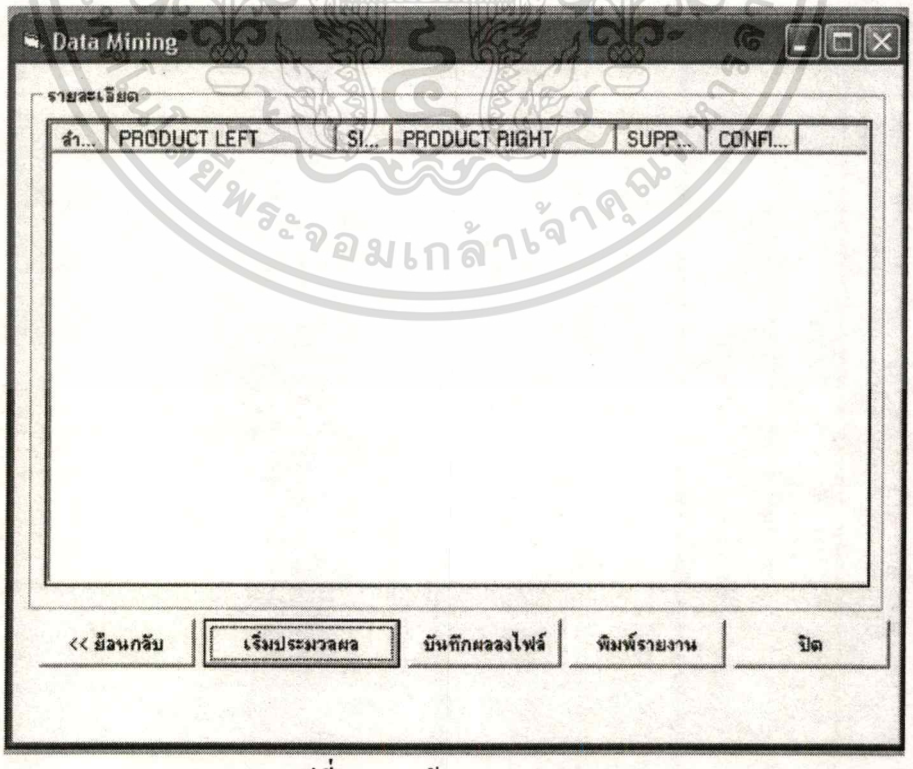
เมื่อทำการกำหนดกลุ่มข้อมูลเรียบร้อยแล้ว กดปุ่ม "***ต่อไป***" และเมื่อต้องการย้อนกลับไปทำการกำหนดเงื่อนไขใหม่ กดปุ่ม "***ย้อนกลับ***"

4.6.3 ส่วนการ Mining

เมื่อทำการจัดเตรียมข้อมูล และทำการจัดกลุ่มเรียบร้อยแล้ว เข้าสู่ส่วนการ Mining โดยมีการกำหนดค่า Parameter ได้แก่ ค่า Minimum Support , ค่า Minimum Confidence และ กำหนด ค่าจำนวนสูงสุดของชุดรายการ ซึ่งระบบได้กำหนดค่าทั้งสามเป็นค่า Default ไว้ที่ 10% , 50% และ 5 ตามลำดับ ซึ่งผู้ใช้สามารถทำการเปลี่ยนแปลงค่าต่างๆได้ ดังรูปที่ 4.12 หรือถ้าต้องการกำหนดค่ากลับเป็นค่า Default โดยการกดปุ่ม "***Restore Default***" หลังจากการกำหนดค่าพารามิเตอร์ต่างๆแล้ว เมื่อทำการกดปุ่ม "***ต่อไป***" จะปรากฏหน้าจอดังรูป 4.13



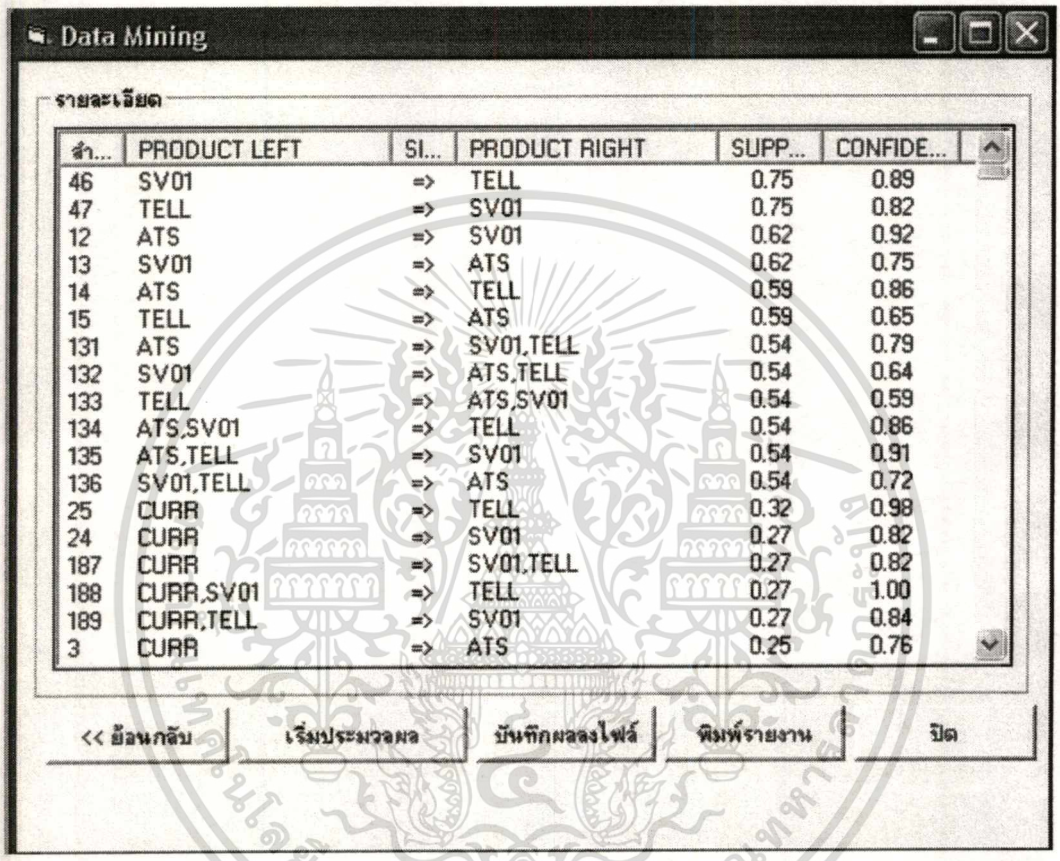
รูปที่ 4.12 หน้าจอกำหนดค่าพารามิเตอร์



รูปที่ 4.13 หน้าจอการทำ Mining

จากรูปที่ 4.13 จะประกอบไปด้วยส่วนสำคัญ ดังนี้

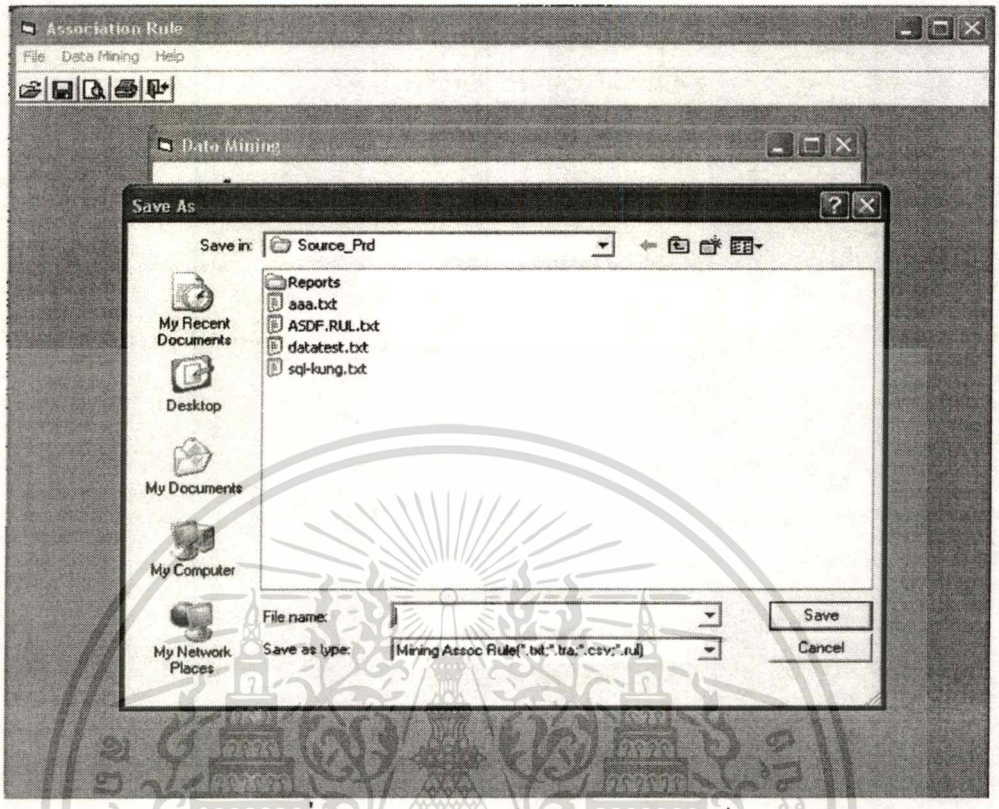
1. Mining โดยทำการกดปุ่ม "***เริ่มประมวลผล***" เพื่อทำงานตาม Apriori Algorithm และสร้างกฎความสัมพันธ์ของสินค้า ตามพารามิเตอร์ที่ได้กำหนด ผลที่ได้ปรากฏดังรูปที่ 4.14



รูปที่ 4.14 หน้าจอตารางผลลัพธ์กฎความสัมพันธ์รายการสินค้า

ส่วนแสดงกฎความสัมพันธ์ของรายการสินค้า จะประกอบด้วยรายการสินค้าตั้งต้น(Product Left) , สินค้าตาม(Product Right) , ค่า Support และค่า Confidence โดยผู้ใช้สามารถทำการเรียงข้อมูลตามต้องการได้โดยทำการกดที่ชื่อ Column ที่ต้องการ เช่นต้องการเรียงข้อมูลตามค่า Support ทำการกดที่ Support เป็นต้น

2. บันทึกกฎความสัมพันธ์ โดยทำการกดปุ่ม "***บันทึกผลลงไฟล์***" จะทำการบันทึกกฎที่ได้ โดยทำการเลือกที่ต้องการเก็บ และกดปุ่ม "***บันทึก***" ดังรูปที่ 4.15



รูปที่ 4.15 หน้าจอบันทึกกฎความสัมพันธ์

3. พิมพ์รายงาน โดยทำการกดปุ่ม "***พิมพ์รายงาน***" แสดงตัวอย่างรายการที่พิมพ์ออกมา โดยในรายงานแสดง ความสัมพันธ์ของรหัสสินค้า , ความสัมพันธ์แบบชื่อสินค้า และค่า Support และค่า Confidence ของแต่ละกฎด้วย ดังรูปที่ 4.16

กฎความสัมพันธ์รายการสินค้า

วันที่ เริ่มต้น 3/1/2546 ถึง 3/1/2546
ค่า Minimum Support 2
ค่า Minimum Confidence 50 %

กฎที่	ความสัมพันธ์แบบรหัสสินค้า	ความสัมพันธ์แบบชื่อสินค้า	Support %	Confidence %
1	CDM==>ATS	ผ่าน Cash Deposit Machine==>ผ่ายการอัตโนมัติ	1.91	75.00
2	CURR==>ATS	เงินฝากเดินสะพัด==>ผ่ายการอัตโนมัติ	24.84	76.47
3	EDU1==>ATS	เงินฝากเพื่อการศึกษา==>ผ่ายการอัตโนมัติ	12.74	60.61
4	FX03==>ATS	เงินฝากประจำ 3 เดือน==>ผ่ายการอัตโนมัติ	6.37	66.67
5	HS01==>ATS	เพื่อการเคหะ==>ผ่ายการอัตโนมัติ	1.27	50.00
6	INET==>ATS	ผ่าน SCB BusinessNet /SIPS / SCB Easy Net==>ผ่ายการอัตโนมัติ	3.82	100.00
7	ODSV==>ATS	เงินฝากกินบัญชี==>ผ่ายการอัตโนมัติ	5.73	69.23
8	PHON==>ATS	ผ่าน SCB Easy Phone==>ผ่ายการอัตโนมัติ	7.01	78.57
9	RPSV==>ATS	ประโยชน์==>ผ่ายการอัตโนมัติ	8.92	66.67
10	STOD==>ATS	โอนอัตโนมัติ (ฝากออก)==>ผ่ายการอัตโนมัติ	3.18	83.33
11	ATS==>SV01	ผ่ายการอัตโนมัติ==>ออมทรัพย์	62.42	91.59
12	SV01==>ATS	ออมทรัพย์==>ผ่ายการอัตโนมัติ	62.42	74.81

รูปที่ 4.16 ตัวอย่างรายงานกฎความสัมพันธ์

เมื่อเสร็จสิ้นการประมวลผล กดปุ่ม "ปิด" และเมื่อต้องการออกจากโปรแกรม "***Menu File -> Exit***"

บทที่ 5
สรุปผลการดำเนินงาน

5.1 ผลการดำเนินงาน

เมื่อกำหนดข้อมูลที่ใช้ทำการทดลอง, ค่า Support และค่า Confidence เพื่อทำการ Mining ตาม Apriori Algorithm ได้กฎความสัมพันธ์ดังนี้

ตารางที่ 6.1 ตารางแสดงกฎความสัมพันธ์ของรายการสินค้า

กฎความสัมพันธ์			ค่า Support (%)	ค่า Confidence(%)
CURR	=>	TELL	32	98
ATS,CURR	=>	TELL	24	97
ATS,ED01	=>	SV01	12	95
ATS,ED01,TELL	=>	SV01	10	94
PHON	=>	TELL	8	93
ATS	=>	SV01	62	92
PHON,SV01	=>	TELL	7	92
ATS,PHON	=>	SV01	6	91
ATS,PHON	=>	TELL	6	91
ATS,TELL	=>	SV01	54	91
SV01	=>	TELL	75	89
ED01	=>	TELL	18	88
ATS,CURR,TELL	=>	SV01	21	87
ATS	=>	TELL	59	86
PHON	=>	SV01	8	86
WB01	=>	SV01	16	86

จากตารางที่ 6.1 นำค่าที่ได้จากกฎที่ 3 และ 4 มาพิจารณาได้ว่า ลูกค้าที่มีบัญชีเงินฝากระยะยาวเพื่อการศึกษา และทำรายการตัดเงินอัตโนมัติ มีความสัมพันธ์กับ บัญชีเงินฝากออมทรัพย์ โดยมีผลิตภัณฑ์และบริการทั้งสามพร้อมกันต่อรายการการใช้บริการทั้งหมด ได้ค่า Support เท่ากับ 12% และมีผลิตภัณฑ์และบริการทั้งสาม ต่อ รายการที่มีบัญชีเงินฝากออมทรัพย์ตามค่า Confidence เท่ากับ 95% ในขณะที่ ลูกค้าที่มีบัญชีเงินฝากระยะยาวเพื่อการศึกษา,ทำรายการตัดเงินอัตโนมัติ และทำรายการผ่าน Counter สาขา มีความสัมพันธ์กับบัญชีเงินฝากออมทรัพย์ โดยมีการใช้ผลิตภัณฑ์และบริการทั้งหมด ต่อรายการการใช้บริการทั้งหมด ตามค่า Support เท่ากับ 10% และ มีการใช้ผลิตภัณฑ์และบริการทั้งหมด ต่อรายการที่มีบัญชีเงินฝากออมทรัพย์ ตามค่า Confidence เท่ากับ 94%

นั่นคือ ถ้าบริษัทใช้บริการและผลิตภัณฑ์ต่างๆ 100 หน่วย จะใช้ทำรายการตัดเงินอัตโนมัติ และมีบัญชีเงินฝากระยะยาวเพื่อการศึกษา และบัญชีเงินฝากออมทรัพย์ 12 หน่วย และถ้าบริษัทมีบัญชีเงินฝากออมทรัพย์จำนวน 100 หน่วย จะมีการใช้บริการและผลิตภัณฑ์สินค้าทั้งสาม 95 หน่วย

5.2 สรุปผลการทดลอง

จากชุดกฎความสัมพันธ์ตามรายการสินค้าที่นำมาพิจารณาจากค่า Support และ ค่า Confidence ที่มีค่ามาก ทั้งนี้การเลือกพิจารณากฎที่ให้ค่าทั้งสองมาก มีโอกาสเกิดขึ้นพร้อมกันมากกว่า กฎที่ให้ค่าทั้งสองน้อย

5.3 การประยุกต์ใช้

การนำเอาเทคนิค Data mining เข้ามาใช้ในการวิเคราะห์หาสารสนเทศที่ซ่อนอยู่ภายในข้อมูลจำนวนมาก วิเคราะห์หาความสัมพันธ์ของข้อมูล กฎที่ได้จากการหาความสัมพันธ์นั้น จำเป็นที่จะต้องอาศัยผู้ที่มีประสบการณ์ในด้านนั้นๆในการตีความหมายของกฎเหล่านั้น เนื่องจากกฎที่ได้มีจำนวนมากซึ่งไม่จำเป็นว่าทุกกฎที่ได้จะมีประโยชน์ หรือสามารถวิเคราะห์หาความสัมพันธ์ได้จริง ดังนั้นกฎนี้จึงเป็นเพียงส่วนหนึ่งที่ช่วยในการตัดสินใจเท่านั้น ซึ่งยังมีอีกหลายปัจจัยที่ใช้ประกอบกับการกำหนดกลยุทธ์ หรือนโยบายทางการตลาด ไม่ว่าจะเป็น สภาพการดำเนินงานของบริษัท ปัจจัยทางเศรษฐกิจ ฯลฯ มาใช้ประกอบเพื่อใช้เป็นแนวทางในการกำหนดเป้าหมายในการดำเนินงานขององค์กรต่อไป

5.4 ปัญหาและอุปสรรค และข้อเสนอแนะ

เนื่องจากว่า Algorithm ที่ใช้มีการคำนวณสูง เพื่อทำการสร้างรายการ Candidate Itemset ซึ่งในแต่ละรอบการทำงานจะใช้ CPU สูง ทำให้บางครั้งเมื่อมีการกำหนดความสัมพันธ์มากๆจะทำให้เกิดการหยุดทำงานได้ หรือ เครื่อง Hang

ข้อเสนอแนะ

สำหรับโครงงานนี้ ทางผู้จัดทำได้ทำการวิเคราะห์หาความสัมพันธ์ของผลิตภัณฑ์ของธนาคาร เพื่อศึกษาถึงรูปแบบลักษณะการซื้อสินค้าและบริการของลูกค้าธนาคาร จากผลลัพธ์ที่ได้สามารถแยกโอกาสทางการตลาดได้เป็น ภายในคือในธนาคาร และภายนอกคือส่วนของลูกค้า โดยอยู่บนพื้นฐานสินค้าที่ลูกค้าซื้อหรือใช้บริการร่วมกัน โดยภายในทำให้สามารถจำแนกได้ถึงสินค้าที่มีความสัมพันธ์ เพื่อกำหนดกลยุทธ์ต่างสำหรับการพัฒนาให้มีประสิทธิภาพยิ่งขึ้น ส่วนภายนอก เพื่อพัฒนาให้ลูกค้าธนาคารมีความหลากหลายมากขึ้น และโอกาสในการทำ cross-selling โดยเฉพาะอย่างยิ่งสำหรับ Credit card และบริการด้านอื่นๆ โดยทางธนาคารอาจจะนำเสนอผลิตภัณฑ์อื่นๆที่ไปในทิศทางเดียวกันได้

บรรณานุกรม

Simoudis,E. 1998. **Discovering Data Mining From Concept to implementation.** New Jersey : Prentice Hall.

Michael J. A. Berry and Gordon, S. Linoff. 1997. **Data Mining Techniques For Marketing , Sales and Customer Support.** , Wiley Computer and Son.

Agrawal, R. and Srikant, R. 1994. **Fast Algorithms for Mining Association Rules.** [Online]. Available: http://citeseer.nj.nec.com/agrawal94fast.html

Estelle, B. and Rob,G. 1998. **Association and Sequencing.** [Online]. Available: http://www.dbmsmag.com/9807m03.html

Han, J. and Kamber, M. 2000. **Data Mining : Concepts and Techniques.** [Online]. Available : http://www.cs.sfu.ca/~han/dmbook

Sander, J. 2002. **Knowledge Discovery in Databases : Association Rules.** [Online]. Available: www.cs.ualberta.ca/~joerg/courses/cmput690/ slides/AssociationRules-s4.pdf

ประวัติผู้เขียน

ชื่อ : นส. พัชรินทร์ อุทัยจรัสรัศมี

วันเดือนปีเกิด : 5 พฤษภาคม 2521

ประวัติการศึกษา:

: มัธยมศึกษา โรงเรียนเตรียมอุดมศึกษา

: ปริญญาตรี คณะพาณิชยศาสตร์และการบัญชี ภาควิชาสถิติ
จุฬาลงกรณ์มหาวิทยาลัย

ประวัติการทำงาน :

: 2543 – ปัจจุบัน เจ้าหน้าที่พัฒนาโปรแกรม ธนาคารไทยพาณิชย์จำกัด(มหาชน)