การพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ Association Rules

Association Rules Discovery for Export Trading

โดย

นางสาวรุจิรา ใหม่จันทร์

รหัส 42067176

อาจารย์ที่ปรึกษา

ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 1 ปีการศึกษา 2544

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หัวข้อ	การพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ Association Rules
นักศึกษา	นางสาวรุจิรา ใหม่จันทร์
อาจารย์ที่ปรึกษา	ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2544

บทคัดย่อ

เนื่องมาจากภาคธุรกิจปัจจุบันมีการแข่งขันกันสูง จึงมีความพยายามที่จะคิดค้นและพัฒนาเทคนิคต่าง ๆ ขึ้นเพื่อให้สามารถแข่งขันได้ในตลาด และเป็นที่ยอมรับกันโดยทั่วไปว่า ข้อมูล คือหัวใจสำคัญในการทำธุรกิจ การที่เรารู้ข้อมูลมากและสามารถนำไปใช้ได้ถูกต้อง ก็จะสร้างโอกาสให้กับธุรกิจมากขึ้น ข้อมูลที่จัดเก็บส่วนใหญ่มักอยู่ในรูปของคลังข้อมูล (Data Warehousing) อาจนำมาใช้ประโยชน์ได้ในระดับหนึ่ง แต่จะทำอย่างไรให้ข้อมูลเหล่านั้นบ่งบอกถึงสิ่งที่ซ่อนอยู่ในข้อมูลโดยที่ธุรกิจอื่นยังไม่รู้และสามารถนำความรู้นั้นไปใช้ให้เกิดความได้เปรียบได้ โครงการนี้จะนำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าเพื่อช่วยหาความสัมพันธ์ของการซื้อสินค้ากับผู้บริโภค โดยใช้ Apriori Algorithm ซึ่งเป็นอัลกอริทึมขั้นพื้นฐานในการหากฎความสัมพันธ์ของข้อมูลประยุกต์ใช้ใน Association Rules ซึ่งเป็นเทคนิคหนึ่งของ Link Analysis ในดาต้าไมนิ่ง (Data Mining) ในการแก้ปัญหา

Title	Association Rules Discovery for Export Trading
Student	Miss Rujira Maichan
Advisor	Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2001

ABSTRACT

Since there is the high competition among Export Trading in the present time, we should try to research and develop new technology, for we can get advantages over the competitors. As we known, information is one of important keys in business. Therefore if we use information effectively, we will gain more business opportunities. Nowadays most data was managed by Data Warehousing, one of the concepts to keep data, which may be used to analyze information in basic level. However, with Data Mining, we will analyze information in advance level that we can use its result to added value in our business.

The propose of this project is to show the step and development of Data Mining for analyzing association between customer and their buying behavior on Export Trading. By using Apriori Algorithm, which is a basic algorithm to generate association rule and being one part of Link Analysis in Data Mining to solve the problem.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ดร. วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงานที่ได้กรุณาให้ความรู้ คำปรึกษาและคำแนะนำต่าง ๆ อันเป็นประโยชน์ต่อการพัฒนาระบบ และได้สละเวลาในการตรวจสอบแก้ไขข้อบกพร่อง

นอกจากนี้ข้าพเจ้าขอกราบขอบพระคุณบุพการี และบุคคลในครอบครัว ที่ได้ให้การสนับสนุนส่งเสริมเป็นกำลังใจในการเรียนตลอดมา ตลอดจนขอขอบคุณเพื่อน ๆ IS8 และเพื่อนๆ จากมหาวิทยาลัยเชียงใหม่ ภาควิชาวิทยาการคอมพิวเตอร์ รุ่น 11 ที่มีส่วนให้ความช่วยเหลือ เป็นกำลังใจ และสนับสนุนให้ผลงานนี้สำเร็จลุล่วงด้วยดี

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าโครงการพัฒนาระบบงานนี้ จะเป็นประโยชน์แก่ผู้ที่สนใจ สำหรับข้อบกพร่องของระบบนี้ ข้าพเจ้าขอรับไว้ เพื่อนำไปปรับปรุงแก้ไขในคราวต่อไป สำหรับความดีที่ได้รับจากโครงการพัฒนาระบบงานนี้ ข้าพเจ้าขอมอบให้แก่บุพการี

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

บทที่

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูปภาพ

หน้า

ภาพที่

สารบัญรูปภาพ(ต่อ)

หน้า

ภาพที่

สารบัญรูปภาพ(ต่อ)

หน้า

ภาพที่

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

จากสภาพการดำเนินธุรกิจท่ามกลางการแข่งขันอย่างรุนแรงในปัจจุบัน ไม่ว่าจะเป็นการแข่งขันทั้งจากภายในและภายนอกประเทศ ทำให้องค์กรต้องปรับตัวเพื่อให้มีความได้เปรียบในเชิงการแข่งขันมากที่สุด เพื่อที่จะสามารถครองส่วนแบ่งทางการตลาดและทำกำไรสูงสุด รวมถึงการสร้างความพึงพอใจสูงสุดให้กับผู้บริโภค การวางแผนการตลาดในการตั้งเป้าหมายของยอดขายโดยดูจากข้อมูลที่จัดเก็บในคลังข้อมูล (Data Warehousing) อย่างเดียวอาจไม่เพียงพอ เนื่องจากการวิเคราะห์ข้อมูลจำนวนมากไม่ใช่เรื่องง่ายและยากที่จะวิเคราะห์ถึงความสัมพันธ์และแนวโน้มต่าง ๆ ของข้อมูลจากหลายฐานข้อมูลได้อย่างครบถ้วน จึงต้องหาวิธีที่จะวิเคราะห์ข้อมูลเพื่อที่จะให้ทราบถึงความสัมพันธ์ในรูปแบบต่ง ๆ ที่ซ่อนอยู่ในคลังข้อมูลเพื่อดึงสารสนเทศที่ได้มาช่วยตั้งเป้าหมายทางธุรกิจให้ดี รวมทั้งยังช่วยในการทำนายแนวโน้มและพฤติกรรมของข้อมูลในอนาคต จึงได้นำเอาเทคนิคของดาต้าไมนิ่ง (Data Mining) เข้ามาช่วยในการวิเคราะห์ข้อมูลเพื่อให้ทราบถึงความสัมพันธ์ในรูปแบบต่าง ที่ซ่อนอยู่ในคลังข้อมูล

1.2 วัตถุประสงค์

เพื่อนำเอาเทคนิคของดาต้าไมนิ่งมาใช้ในการวิเคราะห์ความสัมพันธ์ของข้อมูลต่าง ๆ ในการขายสินค้า วัตถุประสงค์เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้ไปใช้วางแผนกลยุทธทางการตลาดได้อย่างมีประสิทธิภาพ รวมทั้งเป็นแนวทางในการนำไปประยุกต์ใช้ในการสนับสนุนการตัดสินใจของผู้บริหารสำหรับวางแผนเพื่อทำรายการส่งเสริมการขาย เพื่อให้เหมาะสมกับความต้องการของลูกค้า

1.3 ขอบเขตการดำเนินงาน

โครงการนี้เป็นการศึกษาถึงการนำเอาเทคนิคของดาต้าไมนิ่งมาประยุกต์ใช้ โดยอาศัยหลักการของ Link Analysis ในการวิเคราะห์ความสัมพันธ์ของข้อมูลในฐานข้อมูลการส่งออกสินค้า

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ภายใต้ขอบเขตของการศึกษา จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1) ศึกษาและเก็บรวบรวมข้อมูลที่มีอยู่ในองค์กร
2) ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องของดาต้าไมนิ่งเพื่อนำมาประยุกต์ใช้
3) ศึกษาทฤษฎี Association Rules เพื่อนำมาประยุกต์ใช้กับระบบ
4) ออกแบบและพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล
5) สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้า คาดว่าจะให้ประโยชน์แก่เจ้าของงานและผู้ค้นคว้าดังนี้

1) เพื่อให้ขบวนการของการทำดาต้าไมนิ่งหารูปแบบความสัมพันธ์ของข้อมูลด้วยเงื่อนไขต่าง ๆ ที่ไม่สามารถคาดการณ์ได้จากข้อมูลเก่า ๆ
2) เป็นแนวทางในการนำดาต้าไมนิ่งมาประยุกต์ใช้กับข้อมูลทางธุรกิจ
3) เข้าใจหลักการและขั้นตอนของการทำดาต้าไมนิ่ง

ในบทนี้เป็นการกล่าวถึงวัตถุประสงค์และขอบเขตของการทำงานในเบื้องต้นของระบบ ในบทต่อไปจะกล่าวถึงรายละเอียดของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

บทที่ 2
ดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

ดาต้าไมนิ่งเป็นเครื่องมือที่มีประสิทธิภาพในการค้นหาสารสนเทศที่มีประโยชน์จากฐานข้อมูลซึ่งเป็นที่รู้จักกันในปัจจุบัน แต่ยังมีคนอีกมากที่ยังไม่รู้ว่าดาต้าไมนิ่งคืออะไรและมีประโยชน์ต่อพวกเขาอย่างไร ในบทนี้จะกล่าวถึงคำจำกัดความของดาต้าไมนิ่งและทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง

ดาต้าไมนิ่งเป็นขบวนการที่สำคัญที่จะดึงส่วนที่เป็นนัยของข้อมูลที่เราไม่ทราบและทำให้เกิดศักยภาพในการใช้ข้อมูลในฐานข้อมูล กระบวนการค้นหาสารสนเทศจากคลังข้อมูลนี้ต้องผ่านกระบวนการจัดเตรียมข้อมูล (Preprocess Data) การค้นหาและจัดรูปแบบ (Search for pattern) จนกระทั่งได้ข้อมูลตามต้องการ การค้นหานี้ อาจทำได้โดย

- ผู้ใช้เป็นผู้กำหนดคำถาม และระบบจะเป็นผู้ตอบคำถามเหล่านั้น เช่น อาจใช้การซักถาม (Query) และการรายงาน (Reporting) ซึ่งข้อบกพร่องจากการค้นหาแบบนี้คือ ผู้ใช้มักจะไม่ได้คิดถึงสิ่งที่สัมพันธ์กันหรือสิ่งที่ต้องการถามได้อย่างครอบคลุมทั้งหมด ทำให้ข้อมูลส่วนที่สำคัญหลายส่วนอาจไม่ได้ถูกคัดเลือก
- โปรแกรมทางด้านดาต้าไมนิ่ง จะค้นหาข้อมูลอย่างอัตโนมัติโดยโปรแกรมจะคิดคำถามที่น่าสนใจด้วยตัวเอง เมื่อพบข่าวสารแล้วจะแสดงในรูปแบบที่เหมาะสม เช่น กราฟ รายงาน หรือตัวอักษร

เดิมทีแม้ว่าข้อมูลในคลังข้อมูลจะผ่านกระบวนการในการจัดเก็บอย่างเป็นระบบ และมีประสิทธิภาพสูง แต่ถ้าขาดซึ่งกระบวนการในการทำสารสนเทศจากคลังข้อมูลมาใช้อย่างมีประสิทธิภาพและถูกวิธีแล้ว ข้อมูลต่าง ๆ ที่ถูกจัดเก็บไว้จะไม่มีประโยชน์เลย ปัจจุบันเราจึงเริ่มนำดาต้าไมนิ่งมาใช้ในการค้นหาข้อมูลควบคู่ไปกับการพัฒนาเครื่องมือเครื่องใช้ในการที่จะอำนวยความสะดวกต่าง ๆ เนื่องจากมองเห็นความสำคัญของดาต้าไมนิ่งในการค้นหาความรู้ในฐานข้อมูลเพื่อให้เข้าใจถึงความสัมพันธ์ต่าง ๆ ในฐานข้อมูลได้เป็นอย่างดี และสามารถนำความรู้ที่ได้ไปประยุกต์ใช้ในธุรกิจสาขาต่าง ๆ ตลอดจนใช้ในชีวิตประจำวัน เทคโนโลยีดาต้าไมนิ่งจึงเป็นเทคโนโลยีในการค้นหาความรู้ในฐานข้อมูลโดยไม่ต้องตั้งสมมติฐานไว้ล่วงหน้า แต่เป็นการนำความรู้ที่ได้มาทดสอบสมมติฐานภายหลัง สารสนเทศที่ได้มาจากการทำดาต้าไมนิ่งต้องมีลักษณะไม่รู้มาก่อน

ล่วงหน้า(Unknown) เป็นข้อมูลที่มีความถูกต้อง(Valid) และสามารถนำไปใช้ประโยชน์ได้จริง (Actionable) กล่าวคือ

● ข้อมูลที่ไม่รู้มาก่อนล่วงหน้า(Unknown) เป็นข้อมูลที่ผู้ใช้งานไม่รู้มาก่อนและไม่ชัดเจน ไม่สามารถตั้งสมมติฐานล่วงหน้าว่าควรเป็นแบบใด เช่น เจ้าของห้างสรรพสินค้าแห่งหนึ่งเพิ่งค้นพบพฤติกรรมของผู้บริโภคใหม่ว่าผู้บริโภคที่เป็นพ่อบ้านมักจะซื้อสินค้าเบียร์และผ้าอ้อมในวันศุกร์ตอนเย็น จากข้อมูลที่ได้เป็นสัญญาณให้เจ้าของกิจการเตรียมสินค้าไว้เพื่อจำหน่าย ขณะเดียวกันห้างสรรพสินค้าคู่แข่งอาจไม่รู้เรื่องนี้เลยก็ได้

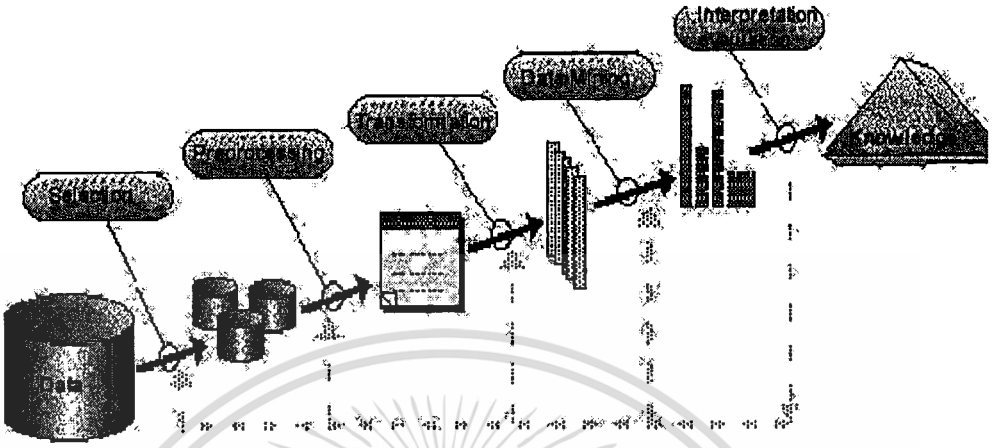
● ข้อมูลที่มีความถูกต้อง(Valid) เป็นข้อมูลที่มีความถูกต้อง เนื่องจากเมื่อผู้ใช้ใช้เทคนิคดาต้าไมนิ่งจะค้นพบสิ่งที่น่าสนใจตลอดเวลา แต่ต้องพิจารณาด้วยว่าสิ่งนั้นถูกต้องหรือไม่ เช่น ผู้ใช้มักพบว่าเมื่อจำนวนความหลากหลายของสินค้ามากขึ้นจะมีความสัมพันธ์ของการซื้อของ 2 สิ่งเสมอ แต่ไม่ได้หมายความว่าจะต้องให้ห้างสรรพสินค้าเก็บสินค้ามากขึ้น เพราะข้อมูลที่ได้อาจเกิดจากความคลาดเคลื่อน

● ข้อมูลที่สามารถนำไปใช้ประโยชน์ได้จริง(Actionable) คือ ข้อมูลจะต้องถูกแปลงออกมาและนำมาตัดสินใจให้เป็นความได้เปรียบเชิงธุรกิจ บางครั้งข้อมูลที่เราค้นพบเป็นสิ่งที่คู่แข่งได้ทำไปแล้วหรือเป็นสิ่งผิดกฎหมาย ข้อมูลดังกล่าวจะไม่มีประโยชน์อะไร ดังนั้น จำเป็นต้องใช้วิจารณญานในการเลือกใช้ข้อมูลด้วย

2.2 กระบวนการทำงานของดาต้าไมนิ่ง

กระบวนการของดาต้าไมนิ่งเป็นกระบวนการของการสร้างแบบจำลอง (Model) โดยสร้างแบบจำลองของกลุ่มข้อมูลเพื่อสร้างความเข้าใจในแนวโน้ม รูปแบบ และความเกี่ยวข้องกันของกลุ่มข้อมูลเพื่อใช้ในการทำนายบนข้อมูลนั้น ๆ โดยสรุปแล้วกระบวนการของดาต้าไมนิ่งประกอบด้วย 5 ขั้นตอนดังแสดงในภาพที่ 2.1

1. กำหนดจุดประสงค์ทางธุรกิจ (Business Objective Determination)
2. การเตรียมข้อมูล (Data Preparation)
3. การทำดาต้าไมนิ่ง (Data Mining)
4. การทำความเข้าใจกับแบบจำลอง (Analysis of Result)
5. การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of knowledge)



ภาพที่ 2.1 ขั้นตอนการทำงานของดาต้าไมนิ่ง [1]

ขั้นตอนที่ 1 : กำหนดจุดประสงค์ทางธุรกิจ

การกำหนดจุดประสงค์ทางธุรกิจจะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะการกำหนดปัญหาให้ชัดเจนจะเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่ง นอกจากนี้แล้วการกำหนดปัญหายังต้องดูถึงความเป็นไปได้ด้วยเพื่อดูว่ามีความจำเป็นที่ต้องใช้ดาต้าไมนิ่งหรือไม่ เพราะไม่ได้หมายความว่าทุกปัญหาสามารถแก้ไขด้วยเทคนิคดาต้าไมนิ่ง ดังนั้น การกำหนดปัญหาที่ไม่ถูกต้องย่อมนำไปสู่ความสำเร็จในการแก้ปัญหาได้ยาก

ขั้นตอนที่ 2 : การเตรียมข้อมูล

เป็นหัวใจของขั้นตอนในการทำและเป็นช่วงที่ใช้เวลามากที่สุด โดยปกติแล้วต้องการเวลาประมาณ 60 เปอร์เซ็นต์ของเวลาทั้งหมดในการเตรียมข้อมูล เนื่องจากบางครั้งอาจมีการนำข้อมูลจากหลายแหล่งมารวมกันเพื่อดูความสัมพันธ์ของข้อมูล ข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความชัดเจน ถูกต้อง ขั้นตอนการเตรียมข้อมูลแบ่งออกเป็น 3 ขั้นตอนย่อยคือ

1) การเลือกข้อมูล (Data selection)

จุดประสงค์หลักคือการระบุลักษณะและเลือกข้อมูลที่ต้องการและนำข้อมูลที่ไม่ต้องการออกไปซึ่งเป็นการเริ่มต้นของการเตรียมการไมนิ่ง การเลือกข้อมูลนั้นจะแตกต่างไปตามจุดประสงค์ของแต่ละธุรกิจที่ได้กำหนดไว้แต่ต้น

การเลือกข้อมูลจำเป็นต้องมีความเข้าใจกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูลและลักษณะอื่น ๆ ตัวแปรข้อมูลมี 2 ลักษณะ

- ตัวแปรแบบ Categorical
  - Nominal : เป็นตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น เพศ(ชาย, หญิง)
  - Ordinal : เป็นตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น ลำดับของสินค้า(ดี, ปานกลาง,เลว)
- ตัวแปรแบบ Quantitative
  - Continuous : ค่าที่เก็บเป็นเลขจำนวนจริง (Real number) หรือเป็นค่าที่ต่อเนื่อง เช่น รายได้
  - Discrete : ค่าที่เก็บเป็นเลขจำนวนเต็ม (Integer) เช่น ข้อมูลจำนวนพนักงาน

นอกจากนี้ ยังมีหลักเกณฑ์ที่ต้องพิจารณาเพิ่มเติมเกี่ยวกับข้อมูลที่จะนำมาใช้อยู่ 4 ประเด็นคือ

1. ระดับของข้อมูลที่พิจารณา

สิ่งที่นำมาช่วยตัดสินใจว่าข้อมูลที่นำมาใช้ควรเป็นข้อมูลระดับรายการ (Item) หรือ ข้อมูลที่สรุปแล้ว คือวัตถุประสงค์ในการทำดาต้าไมนิ่ง เช่น

- การทำไมนิ่งเกี่ยวกับการโทรศัพท์ ถ้าจุดประสงค์ของเราต้องการเน้นไปที่พฤติกรรมการใช้โทรศัพท์ของลูกค้า ข้อมูลที่จัดเก็บโดยปกติแล้วจะมีการจัดเก็บเป็นลักษณะรายละเอียดของแต่ละชุมสาย การเคลื่อนย้ายของอิเล็กตรอนไปยังสวิตชิง ข้อมูลเหล่านี้จะไม่มีประโยชน์เลย เพราะจุดประสงค์ของเราสนใจสิ่งที่อยู่ภายใต้การควบคุมของลูกค้าและมีผลต่อการตลาด ดังนั้น ข้อมูลที่เราสนใจจะเป็น เบอร์โทรศัพท์ของผู้โทร, เวลาเริ่มต้นที่ใช้โทร และเวลาที่ใช้ในการโทรศัพท์แต่ละครั้ง
- ข้อมูลที่ยังไม่สรุป ทำให้จัดการได้ยาก รวมทั้งเกิดจำนวนการคอมไบเนชั่น(Combination) สูง เมื่อใช้เทคนิคของ Association Discovery เพราะข้อมูลของร้านค้าปลีกย่อมมีรายการสินค้าเยอะ ดังนั้น การนำเอาหน่วยวัดในการจัดเก็บสินค้าในคลัง หรือ SKU (Stock Keeping Unit) เข้ามาช่วยจะสามารถลดจำนวนการคอมไบเนชั่นลงได้

2. ลักษณะของข้อมูลที่จัดเก็บ

การจัดเก็บข้อมูลด้วยภาษาคอมพิวเตอร์ที่แต่ละระบบปฏิบัติการเลือกใช้แตกต่างกัน ทำให้ข้อมูลที่นำมาวิเคราะห์มีผลกระทบ เช่น ข้อมูลที่นำมาวิเคราะห์ส่วนมากจัดเก็บด้วยภาษา COBOL และ RPG ข้อมูลที่เป็น Text จะถูกเก็บเป็น EBCDIC และข้อมูลตัวเลขจะเก็บเป็น Packed Decimal ขณะที่ภาษาที่เลือกใช้ในการสร้างระบบ ดาต้าไมนิ่งใช้ภาษา C และ C++ ซึ่งข้อมูลชนิด Text จะมีรูปแบบเป็น ASCII และข้อมูลตัวเลขเก็บเป็น Integer หรือ Floating Point

3. ความแตกต่างของข้อมูลแต่ละแหล่ง

เมื่อข้อมูลที่นำมาวิเคราะห์มาจากหลายแหล่ง ซึ่งแต่ละแหล่งมีรูปแบบการจัดเก็บข้อมูลที่ต่างกัน เช่น การวิเคราะห์ข้อมูลการโทรศัพท์ (Call Detail) เพื่อหาเบอร์โทรศัพท์ที่ใช้ฝากข้อความเข้า Voice Mailbox ในแต่ละเมือง จะมีวิธีการจัดเก็บข้อมูลที่ต่างกัน เช่น เมือง ๆ หนึ่งอาจเก็บเบอร์

โทรศัพท์ที่ใช้โทรเข้า Voice Mailbox ด้วยเบอร์ต้นทางและปลายทาง แต่อีกเมืองหนึ่ง อาจเก็บเบอร์โทรศัพท์ที่ไม่รู้ด้วยเบอร์ปลายทาง อีกเมืองหนึ่งอาจเก็บเบอร์โทรศัพท์ที่โทรเข้า Voice Mailbox จริง ๆ ดังนั้น จึงจำเป็นต้องทำข้อมูลเหล่านี้ให้ออกมาในรูปแบบมาตรฐานเดียวกันก่อน เพื่อที่จะได้เข้าใจถึงความแตกต่างในการเก็บข้อมูลของแต่ละแหล่งได้

4. ข้อมูลที่เป็นข้อความ (Textual Data)

ข้อมูลที่จัดเก็บเป็นแบบ Text อาจก่อให้เกิดความสับสน เช่น '_no' *กับ* 'no_' หรือ 'VOR2J0' กับ 'VOR 2J0' กับ 'VOR-2J0' ซอฟท์แวร์ที่ใช้ในการทำดาต้าไมนิ่งย่อมมองข้อมูลเหล่านี้ไม่เหมือนกัน ในทางแก้ไขคือสร้างตารางเก็บค่าที่ถูกต้อง และแทนที่ข้อมูลที่นำมาวิเคราะห์ด้วย Index ตัวอย่างที่เห็นได้ชัดเจน คือ ฐานข้อมูลแบบสัมพันธ์ (Relational Database) มีการแทนที่ข้อมูลที่เป็น Product_Name ด้วย Product_Code ซึ่งมีการ Unique มากกว่า

2) การกลั่นกรองข้อมูล (Data Preprocessing)

จุดประสงค์เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นถูกต้องและ เหมาะสมที่จะนำไปทำดาต้าไมนิ่ง เนื่องจากข้อมูลที่ถูกเลือกมาจากกระบวนการเลือกข้อมูลข้างต้นอาจมีข้อมูลไม่ถูกต้อง ดังนั้นในขั้นตอนนี้มีประเด็นที่จะต้องพิจารณาเพิ่มเติม 2 ประเด็นคือ

- Noisy Data : เป็นข้อมูลที่มีลักษณะต่างจากข้อมูลที่คาดการไว้ ซึ่งอาจมีความหมายได้ทั้งแง่ดีและร้าย ในแง่ดีคือมันจะแสดงชัดเจนถึงสิ่งที่เรากำลังมองหาอยู่ ในแง่ร้ายคือมันอาจเป็นข้อมูลที่ไม่สมบูรณ์ สาเหตุที่เกิดขึ้นอาจมาจากความเลินเล่อในการบันทึกข้อมูล เช่น บันทึกอายุพนักงานเป็น 200 ปี หรือบันทึกรายได้ติดลบ ค่าเหล่านี้ควรถูกแก้ไขหรือไม่นำมาวิเคราะห์ ควรมีขั้นตอนของการตรวจสอบข้อมูลก่อนนำไปใช้

- Missing Value : ข้อมูลที่ไม่ได้ถูกเลือกมาจากขั้นตอนที่ 1 ถ้าข้อมูลที่ขาดมีจำนวนน้อยอาจตัดทิ้งได้ แต่ถ้าข้อมูลที่ขาดมีมากอาจต้องแทนด้วยค่าเฉลี่ย ในบางกรณี Missing Value นี้อาจชี้ให้เห็นถึงสิ่งที่ผิดปกติก็ได้ เช่น ผู้สมัครไม่ให้ข้อมูลเบอร์โทรศัพท์ที่ทำงาน อาจชี้ให้เราเห็นว่าผู้สมัครรายนี้ไม่ได้ถูกจ้างงานอยู่ ณ ปัจจุบัน หรือ ไม่ต้องการให้เราโทรศัพท์ไปสอบถามเพื่อตรวจสอบเงินเดือนของเขา กับที่เขียนไว้ในใบสมัคร การแทนค่าเบอร์โทรศัพท์ด้วยหลักการ Missing Value อาจใช้ไม่ได้ในกรณีแบบนี้

3) การแปลงข้อมูล (Data Transformation)

เป็นการแปลงข้อมูลให้อยู่ในรูปแบบของข้อมูลที่พร้อมที่จะนำไปวิเคราะห์ตามอัลกอริทึมของดาต้าไมนิ่งที่ใช้ เช่น การแปลงตัวแปรแบบ Quantitative ให้เป็นแบบ Categorical โดยแบ่งค่าของตัวแปรให้เป็นช่วง ๆ เช่น การแปลงข้อมูลเงินเดือน นอกจากนี้ยังมีเทคนิคของการแปลงตัว

แปรแบบ Categorical ให้เป็น Numeric เช่น ยี่ห้อรถ HONDA, TOYOTA และ NISSAN ให้เป็น 001, 010 และ 011

ขั้นตอนที่ 3 : การทำดาต้าไมนิ่ง

เป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ ในขั้นตอนนี้จะมีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนที่ผ่านมา โดยเมื่อทำในส่วนของดาต้าไมนิ่งแล้วอาจต้องย้อนกลับไปทำในขั้นตอนของการเตรียมข้อมูลใหม่ ในการพัฒนาในส่วนของดาต้าไมนิ่ง จะเกี่ยวข้องกับการใช้ อัลกอริทึมหลาย ๆ แบบ ซึ่งแต่ละแบบมีข้อดีและข้อเสียที่ต่างกัน

ขั้นตอนที่ 4 : การทำความเข้าใจกับแบบจำลอง

เป็นการวิเคราะห์ผลของการประมวลผล ซึ่งจะทำการแปลความหมายและประเมินผลที่ได้จากขั้นตอนการทำไมนิ่ง การทำงานในส่วนนี้จำเป็นต้องใช้ทักษะในการวิเคราะห์ข้อมูลและการวิเคราะห์ทางธุรกิจเข้ามาช่วย เครื่องมือทางด้าน Graphical Visualization จะช่วยวิเคราะห์ข้อมูลได้อย่างสะดวกและรวดเร็วขึ้น

ขั้นตอนที่ 5 : การปรับความรู้ที่ได้เข้ากับธุรกิจ

เป็นการรวบรวมความเข้าใจทางธุรกิจที่เป็นผลมาจากขั้นตอนที่ 4 มารวมเข้ากับส่วนความรู้เพื่อนำไปใช้ในโอกาสต่อไป ในขั้นตอนนี้มีหลักอยู่ 2 ประการคือ การนำเสนอแนวคิดทางธุรกิจที่ค้นพบใหม่ และหาแนวทางที่จะใช้กฎเกณฑ์ใหม่ที่ค้นพบเพื่อให้เกิดประโยชน์สูงสุด

2.3 โอเปอเรชั่นของดาต้าไมนิ่ง (Data Mining Operation)

ดาต้าไมนิ่งประกอบด้วย 4 โมเดลหลักที่ใช้สำหรับประยุกต์ใช้งานทางธุรกิจ ได้แก่ Predictive Modeling, Database Modeling, Database Segmentation และ Link Analysis

1. Database Segmentation เป็นกระบวนการแบ่งกลุ่มของฐานข้อมูลซึ่งสมาชิกในกลุ่มมีค่าความเหมือนกันอยู่เพื่อให้ง่ายต่อการวิเคราะห์ เช่น การแบ่งกลุ่มลูกค้าออกตามอายุ, เพศ, และรายได้ เป็นต้น
2. Predictive Modeling เป็นโมเดลที่ใช้ในการสร้างแบบจำลองพยากรณ์ แบ่งออกเป็น 2 เทคนิคคือ
   - Classification เป็นการทำนายกลุ่มของรายการจากข้อมูลนำเข้า เช่น การทำนายว่าเป็นลูกค้าที่อยู่ในกลุ่มที่ควรส่งจดหมายแนะนำสินค้าและบริการใหม่ไปให้หรือไม่ โดยดูจากประวัติและพฤติกรรมการบริโภค เป็นต้น
   - Value Prediction เป็นการทำนายค่าที่เป็นตัวเลข ใช้เพื่อทำนายค่าของเหตุการณ์ในอนาคต เช่น การทำนายราคาหุ้น หรือการทำนายภาษีที่จะเก็บได้ในปีหน้า เป็นต้น

3. Link Analysis เป็นการวิเคราะห์หาความสัมพันธ์ระหว่างข้อมูลว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันหรือไม่ อย่างไร เช่น เก็บข้อมูลการซื้อสินค้าแต่ละครั้งของลูกค้าเพื่อศึกษาพฤติกรรมการซื้อสินค้า เพื่อนำมาทำนายการส่งเสริมการขายและการจัดชั้นวางสินค้าให้เหมาะสม
4. Deviation Detection เป็นความพยายามหาสิ่งที่แปลกปลอมออกจากกลุ่มของมัน ส่วนมากอาศัยการวาดกราฟ แล้วดูว่าจุดมันมีการกระจายออกไปจากกลุ่มหรือไม่ มักใช้ในการตรวจจับสิ่งผิดปกติต่าง ๆ เช่น การจับการโกง เป็นต้น

โมเดลเหล่านี้นำไปประยุกต์ใช้ในงานทางธุรกิจได้ดังภาพที่ 2.2 แต่จะไม่สามารถเจาะจงได้ว่าธุรกิจ ประเภทใด ต้องใช้โมเดลไหน เพียงแต่บอกว่าลักษณะงานทางธุรกิจใดมีความเกี่ยวข้องกันและลักษณะงานแบบไหน ควรใช้โมเดลแบบใด

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*

ภาพที่ 2.2 โมเดลของดาต้าไมนิ่งกับการประยุกต์ใช้งาน [2]

จากที่กล่าวมาแล้วข้างต้นว่าโอเปอเรชั่นของดาต้าไมนิ่งมีมากมาย สำหรับโครงการที่นำเสนอนี้ จะนำเสนอโอเปอเรชั่นของลิงค์อนาไลซิส (Link Analysis) โดยการใช้แอสโซซิเอชั่นรูล (Association Rules) มาใช้เพื่อหาความสัมพันธ์ของข้อมูลการส่งออกสินค้า โดยจะกล่าวถึงรายละเอียดในบทถัดไป

บทที่ 3

ลิงค์อนาไลซิส(Link Analysis)

หลักการทำงานของลิงค์อนาไลซิส มุ่งเน้นการทำงานบนเรคอร์ดที่สนใจเพื่อหาความสัมพันธ์หรือความเกี่ยวโยงกันระหว่างเรคอร์ด หรือกลุ่มของเรคอร์ด เช่น หาความสัมพันธ์ระหว่างผลิตภัณฑ์ หรือบริการที่ลูกค้าสนใจในเวลาหนึ่ง ๆ หลักการสำคัญของลิงค์อนาไลซิสมี 3 ประการ คือ

1. แอสโซซิเอชั่นดิสคอฟเวอรี่ (Association Discovery)
2. ซีเควนเชียลแพทเทิร์นดิสคอฟเวอรี่ (Sequential Pattern Discovery)
3. ซิมิล่าไทม์ซีเควนซ์ดิสคอฟเวอรี่ (Similar Time Sequence Discovery)

3.1 Association Discovery

ใช้วิเคราะห์หาความสัมพันธ์ของข้อมูลที่เกิดขึ้นในรายการเดียวกัน ศึกษาถึงความสัมพันธ์ที่ถูกปิดซ่อนอยู่ของสินค้า ซึ่งสินค้าเหล่านั้นมักมีแนวโน้มที่จะถูกซื้อควบคู่กันไป การวิเคราะห์แบบนี้บางครั้งเรียกว่า "Market Basket Analysis" (MBA) หรือ "Product Affinity Analysis" คือ การกำหนดว่าสิ่งของอะไรที่จะไปด้วยกัน แนวคิดดังกล่าวนำไปใช้ในร้านซุปเปอร์มาร์เก็ต เพื่อกำหนดว่าสินค้าประเภทใดมักจะถูกซื้อควบคู่กันในการซื้อแต่ละครั้ง ทำให้ทางร้านสามารถกำหนดได้ว่าควรจัดเรียงสินค้าอย่างไร หรือควรเตรียมแคตตาล็อกเพื่อขายสินค้าอย่างไร รวมถึงการวางแผนเพื่อจัดโปรโมชันสนับสนุนการขาย

3.1.1 อัลกอริทึมในการทำงานของ Association Rule มี 3 ขั้นตอน

1. เลือกชุดข้อมูลที่ถูกต้อง (Choosing the Right Set of Items)

2. นำรายการที่เกิดขึ้นมาทำการ Combination หาค่า Support และ Confidence นำมารวมกันสร้างเป็นกฎ (Generating Rule From All This Data)

3.จำกัดจำนวนที่เกิดขึ้น โดยเลือกเฉพาะชุดข้อมูลที่เป็นไปได้ (Overcoming Practical Limit)

1) การเลือกชุดข้อมูลที่ถูกต้อง

โดยปกติแล้ว ข้อมูลที่นำมาใช้มักจะเป็นข้อมูลระดับรายละเอียด หรือ ระดับรายการที่เก็บได้จาก Transaction ณ จุดขาย ซึ่งในทางปฏิบัติร้านค้าเหล่านี้จะมีรายการสินค้าเป็นจำนวนมาก ดังนั้น ในการพิจารณาว่าจะนำข้อมูลระดับไหนมาใช้ ขึ้นกับวัตถุประสงค์ เช่น วิเคราะห์การขายพิซซ่า ถ้าไม่พิจารณาถึง หน้าพิซซ่า (ชีส, หัวหอม, เห็ด) และความหนาของแป้ง (แป้งบาง, แป้งหนา) ไม่ว่าลูกค้าจะซื้อพิซซ่าที่มีรายละเอียดแตกต่างกัน ถือว่าเป็นรายการซื้อพิซซ่า ข้อมูลที่ใช้จะออกเป็นดังภาพที่ 3.1 แต่ถ้าร้านขายพิซซ่าต้องการวิเคราะห์ข้อมูลการขายพิซซ่าโดยพิจารณาการขายรวมถึงประเภทพิซซ่าที่ลูกค้าสั่งซื้อ ข้อมูลที่ใช้ควรเป็นลักษณะดังภาพที่ 3.2

Customer	Pizza	Milk	Sugar	Apple	Coffee
1	√				
2		√	√		
3	√			√	√
4		√			
5	√		√	√	√

ภาพที่ 3.1 การวิเคราะห์การขายพิซซ่าแบบสรุป [3]

Customer	Extra Cheese	Onions	Paper	Mushrooms
1	√	√		
2			√	
3	√	√		√
4		√		
5	√		√	√

ภาพที่ 3.2 การวิเคราะห์การขายพิซซ่าแบบละเอียด [3]

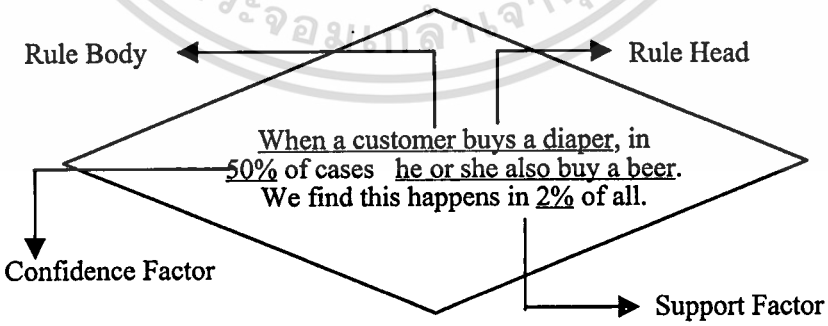
อย่างไรก็ตามระดับของข้อมูลที่สนใจ อาจเปลี่ยนแปลงได้ตลอด เช่น เมื่อเจ้าของร้านค้าพบว่าข้อมูลระดับสรุปไม่เพียงพอกับความต้องการ อาจระบุข้อมูลลงลึกถึงระดับรายการ หรือขณะ

เดียวกันร้านขายพิซซ่านำสินค้าชนิดอื่นมาขายด้วยโดยไม่สนใจการขายพิซซ่าที่มีรูปแบบต่างกัน การวิเคราะห์ ข้อมูลการขายโดยดูที่ระดับรายการย่อมไม่ตรงกับวัตถุประสงค์อีกต่อไป จะเห็นได้ว่าระดับของข้อมูลที่นำมาวิเคราะห์มีความหลากหลาย จึงมีการนำ Taxonomy เข้ามาช่วย คือการจัดกลุ่มของสินค้าให้เป็นตัวใหญ่ขึ้น โดยใช้ กลุ่มของสินค้า หรือ หน่วยที่ใช้จัดเก็บสินค้าในคลังเข้ามาช่วย เนื่องจากข้อมูลระดับ Item จะทำให้มีจำนวนของการ คอมไบเนชั่นสูง และใช้เวลานานในการคำนวณ และบางครั้ง จำนวน Transaction ที่เกิดขึ้นอาจจะมีน้อยเกินจนนำมาพิจารณาไม่ได้ เช่นลูกค้ามี Transaction ซื้อ "Ice Cream" เพียง 2-3 ครั้ง แต่เมื่อวิเคราะห์ด้วย "Frozen Dessert" จะมีจำนวนรายการมากขึ้น ดังนั้น ควรวิเคราะห์ให้ได้ความสัมพันธ์ระดับบนก่อน แล้วค่อยลงลึกในระดับ Item

ในทางปฏิบัติ อาจมีการผสมระหว่าง Item กับ Taxonomy ได้เมื่อวิเคราะห์สินค้าที่ต่างกัน Taxonomy ช่วยในการจัดกลุ่มสินค้าที่ไม่ค่อยเกิดบ่อย ทำให้มีรายการขายสินค้าที่มีความถี่มากขึ้น ขณะที่สินค้าที่มีรายการเกิดบ่อย ๆ ไม่จำเป็นต้องนำมาจัดกลุ่มขึ้นเป็นอีกระดับหนึ่งก็ได้

2)นำรายการที่เกิดขึ้นมาสร้างเป็นกฎ

นำข้อมูลที่สนใจมาทำการคอมไบเนชั่น รูปแบบของกฎที่ได้จากการทำ Association Discovery อยู่ในลักษณะ "IF condition1 THEN condition2" หรือ "WHEN condition1 THEN condition2" โดยที่ condition1 และ condition2 เกิดขึ้นพร้อมกันใน transaction เดียวกัน เรียก condition1 ว่า Rule Body (หรือ Left-hand side, Antecedent) และเรียก condition2 ว่า Rule Head (หรือ Right-hand side, Consequent) ในบทความนี้จะเรียก condition1 ว่า "เหตุ" และ condition2 ว่า "ผล" ตัวอย่างของกฎแสดงได้ดังภาพที่ 3.3



ภาพที่ 3.3 รูปแบบความสัมพันธ์ของ Association Discovery [2]

ในทางปฏิบัติกฎที่นำไปใช้งานได้มักมีผล (Condition2) เพียง 1 รายการ เช่น "IF *Diapers* and *Thursday* , then *Beer* " จะมีประโยชน์กว่า "IF *Thursday*, then *Diapers* and *Beer*"

กฎที่ได้จาก Association Discovery มีตัววัดหลักๆ 2 ตัวคือ Support Factor และ Confidence Factor บางผลิตภัณฑ์ของคาต้าไมนิ่งจะเรียกตัววัดนี้ แตกต่างกันออกไป เช่น MineSet ของ SGI จะเรียก Support Factor และ Confidence Factor ว่า Predictability และ Prevalence ตามลำดับ

Combination	จำนวนเหตุการณ์ที่เกิด
ผ้าอ้อม	20,000
เบียร์	30,000
ผ้าอ้อมและเบียร์	10,000

ภาพที่ 3.4 ข้อมูลจำนวน Transaction ที่เกิดจากการขายเบียร์และผ้าอ้อม จากชุดข้อมูลทั้งหมด 500,000 รายการ

- Support (Prevalence) : คือค่าที่แสดงสัดส่วนระหว่างจำนวนชุดของข้อมูลที่มีทั้งข้อมูลที่เป็นทั้ง "เหตุ" และ "ผล" ของเหตุการณ์เทียบกับจำนวนเหตุการณ์ภายในชุดข้อมูลทั้งหมด

ค่า Support คือ 2 % ซึ่งได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการผ้าอ้อมและเบียร์คู่กัน}}{\text{จำนวนชุดข้อมูลทั้งหมด}}$$

- Confidence (Predictability) : คือค่าที่แสดงสัดส่วนระหว่างจำนวนชุดข้อมูลที่มีทั้งข้อมูลที่เป็น "เหตุ" และ "ผล" เทียบกับจำนวนชุดข้อมูลที่มีเฉพาะเหตุการณ์ที่เป็น "เหตุ"

ค่า Confidence คือ 50 % ซึ่งได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการผ้าอ้อมและเบียร์คู่กัน}}{\text{จำนวนชุดข้อมูลที่มีรายการซื้อผ้าอ้อม}}$$

นั่นคือที่มาของกฎ "50 % ของผู้ที่ซื้อผ้าอ้อม มักจะซื้อเบียร์ไปด้วย" ถ้าในทางกลับกันเปลี่ยนกฎเป็น "เมื่อคนซื้อเบียร์ มักจะซื้อผ้าอ้อมไปด้วย" ค่า Confidence ในทีนี้คือ 33.33 %

กฎที่เราสนใจ คือกฎที่มีค่า Confidence มากที่สุด นอกจากตัววัด 2 ตัวคือ Support และ Confidence แล้ว จากข้อมูลที่เกิดขึ้นพบว่ายังมีตัววัดค่าของเหตุการณ์ที่เกิดขึ้นของแต่ละรายการสินค้า เรียกว่า "Expected Confidence" และ "Lift"

- Expected Confidence : คือค่าที่แสดงสัดส่วนระหว่างชุดข้อมูลที่เป็นตัวแทนของเหตุการณ์ที่สนใจ ซึ่งอาจเป็น "เหตุ" หรือ "ผล" อย่างใดอย่างหนึ่ง เทียบกับจำนวนเหตุการณ์ทั้งหมดภายในชุดข้อมูล

จากข้อมูลของภาพที่ 3.4 Expected Confidence ของ "การซื้อผ้าอ้อม" คือ 4 % และ Expected Confidence ของ "การซื้อเบียร์" คือ 6 % ซึ่งเรียกได้ว่า "4 % ของเหตุการณ์มีรายการซื้อผ้าอ้อม" และ "6 % ของรายการมีการซื้อเบียร์"

- Lift : เป็นปัจจัยหลักที่ใช้ในการเปรียบเทียบคัดเลือกเหตุการณ์ที่ค้นพบจากการทำงานของซอฟท์แวร์และ ใช้สรุปผล ตีความลักษณะของความสัมพันธ์ โดยหาได้จากสัดส่วนระหว่างค่า Confidence กับ Expected Confidence ของจำนวนชุดข้อมูลที่เป็น "ผล"

จากข้อมูลของภาพที่ 3.4 Confidence มีค่า 50 % ค่า Expected Confidence ของเหตุการณ์ที่เป็น "ผล" คือ 6% ค่า Lift คือ 8.33

ค่า Lift ที่ได้จะบ่งบอกถึงความสำคัญของความสัมพันธ์หรือเหตุการณ์ที่ได้ว่ามีมากน้อยแค่ไหน จากข้อมูลที่ได้ ทำให้ได้กฎ "คนซื้อผ้าอ้อมมักจะซื้อเบียร์ในอัตรา 8%" แต่ถ้าค่า Confidence มีค่ามากกว่า 50 % จะทำให้ค่า Lift มีมากกว่า 8.33 % ความสัมพันธ์ระหว่างเหตุการณ์จะน่าเชื่อถือมากยิ่งขึ้น

ข้อควรระวังในการวิเคราะห์ข้อมูลคือ ค่าของ Lift ที่ติดลบ หรือน้อยกว่า 1 หมายถึงเหตุการณ์เหล่านั้นไม่มีทางเกิดขึ้นพร้อมกันได้เลย และกฎที่มีค่า Lift มากหรือน้อยเกินไป บางครั้งอาจเป็นกฎที่ไม่เป็นความจริง เช่น กฎที่ได้คือ "ทุกครั้งที่คนซื้ออาหารสุนัข จะต้องซื้ออาหารแมวไปด้วย" เหตุการณ์เหล่านี้อาจเกิดขึ้นเพียงแค่ 1 วัน เช่น มีรายการโปรโมชัน ณ วันนั้น ว่า ซื้ออาหารสุนัขและอาหารแมว จะแถมอาหารสุนัขอีก 1 ชิ้น

โดยทั่วไป นักวิเคราะห์ข้อมูลมักจะสนใจในกฎที่มีค่า Lift สูงกับต่ำ ซึ่งจะช่วยให้หาความสัมพันธ์ได้ง่าย กฎที่มีค่า Support น้อยอาจแสดงถึงค่าที่ได้มาจากการคำนวณทางสถิติที่ผิดพลาดก็ได้ ถึงแม้จะเป็นกฎที่ถูกต้องก็ตาม การนำกฎที่มีค่า Support ต่ำไปใช้อาจจะได้ผลลัพธ์ที่ไม่คุ้มค่าก็ได้ เพราะค่า Support เป็นตัวบอกว่าเหตุการณ์ดังกล่าวเกิดขึ้นบ่อยแค่ไหน เช่น การตั้งเป้าหมายทางธุรกิจลงทุนเพิ่ม 100 ล้านบาท บนเหตุการณ์ที่เกิดขึ้นปีละ 3 ครั้งย่อมไม่ใช่สิ่งที่คุ้มค่าต่อการลงทุน

3) จำกัดจำนวนที่เกิดขึ้นโดยเลือกเฉพาะชุดข้อมูลที่เป็นไปได้

เนื่องจากการทำงานของ Association Rule มีแนวคิดมาจากการนับจำนวนครั้งของรายการที่เกิดขึ้น และรวมเหตุการณ์เหล่านั้นเข้าด้วยกันในทุกวิถีทางที่เป็นไปได้ ทำให้กฎที่ได้มีจำนวนมาก เทคนิค "Pruning" จะช่วยจำนวนรายการที่นำมาทำการคอมไบเนชั่น ในแต่ละขั้นเพื่อลดจำนวนรายการคอมไบเนชั่น ที่ไม่ตรงกับเงื่อนไขได้ อัลกอริทึมทั่วไปมักจะให้ผู้ใช้ระบุค่า "Minimum Support" และ "Minimum Confidence" ได้ เพื่อให้เวลาที่ใช้ในการคำนวณหากฎไม่มากจนเกินไป

เทคนิค "Pruning" ที่ใช้กันมากที่สุดคือ "Minimum Support Pruning" จะเป็นตัวกำหนดว่ากฎที่ได้จะต้องมาจากรายการที่มีจำนวนการเกิดอย่างน้อยเท่ากับ Minimum Support เช่น มี 1 ล้านรายการ กำหนดค่า Minimum Support 1% ดังนั้น กฎที่สนใจจะต้องมีค่า Support อย่างน้อย 10,000 รายการ Minimum Support จะเป็นตัวจำกัดข้อมูลเป็นทอด ๆ เช่น เมื่อพิจารณากฎที่มี 4 Item

IF *A*, *B*, and *C* , then *D*

เมื่อกำหนดค่า Minimum Support เป็น 1% กฎที่ได้จะเป็นจริงก็ต่อเมื่อ ทุก ๆ Item ต้องมีรายการอย่างน้อย 10,000 ครั้งดังนี้

A ต้องเกิดขึ้นอย่างน้อย 10,000 ครั้ง และ

B ต้องเกิดขึ้นอย่างน้อย 10,000 ครั้ง และ

C ต้องเกิดขึ้นอย่างน้อย 10,000 ครั้ง และ

D ต้องเกิดขึ้นอย่างน้อย 10,000 ครั้ง

หลังจากได้กฎออกมาแล้ว ต้องพิจารณาต่อว่า

A และ B ต้องเกิดพร้อมกันอย่างน้อย 10,000 ครั้งและ

A และ C ต้องเกิดพร้อมกันอย่างน้อย 10,000 ครั้งและ

A และ D ต้องเกิดพร้อมกันอย่างน้อย 10,000 ครั้ง

จะเห็นว่า Minimum Support ช่วยลดจำนวน Item ที่ปรากฏในการทำคอมไบเนชั่นลงได้ ซึ่งมี 2 หนทางในการทำคือ

1. กำจัด Item นั้น ๆ ออกจากการพิจารณา

2. ใช้ Taxonomy ช่วยจัดกลุ่ม Item ให้เป็นสินค้าที่ใหญ่ขึ้น

อย่างไรก็ตาม ในการทำงานจริง ค่า Minimum Support อาจขึ้นกับข้อมูลและสถานการณ์ และสามารถเปลี่ยนแปลงได้ในแต่ละระดับของการทำงาน เช่น ระบุค่า Minimum Support ให้ต่ำเพื่อจะได้ดึงรายการที่เกิดขึ้นไม่บ่อยครั้ง ออกจากรายการที่เกิดขึ้นเป็นประจำได้ หรือ ระบุค่า

Minimum Support ให้มากขึ้น เพื่อดึงรายการที่เกิดขึ้นเป็นประจำออกจากรายการที่เกิดขึ้นไม่บ่อยครั้งได้ อัลกอริทึมการทำงานของ Association Rule ทำงานกับข้อมูลชนิด Categorical Data ดังนั้น ถ้ามีข้อมูลที่ไม่เป็น Categorical เช่น รายได้ของลูกค้า ข้อมูลรายได้นี้จะต้องแปลงมาเป็นช่วงข้อมูล เช่น (0 – 20,000), (20,001 – 40,000) และ (40,001 – 50,000)

3.1.2 กฎที่ได้จากอัลกอริทึม Association Discovery มี 3 ลักษณะ

1) Useful : กฎที่ได้เป็นสารสนเทศที่มีคุณภาพสูง สามารถนำไปตัดสินใจในการดำเนินการทางธุรกิจได้

2) Trival : กฎที่ได้เป็นข้อมูลที่ไม่เป็นสารสำคัญ หรือเป็นข้อมูลที่รู้อยู่แล้ว

3) Inexplicable : กฎที่ได้ไม่สามารถอธิบายได้ ไม่ได้สนับสนุนการตัดสินใจ

3.1.3 ข้อดีของ Association Rule

1) ทำงานได้ดีกับข้อมูลขนาดใหญ่ ขณะที่เทคนิคอื่น ๆ จะมีปัญหากับการทำงานกับข้อมูลปริมาณมาก ๆ นอกจากนี้ในปัจจุบันยังมีงานวิจัยเพื่อเพิ่มประสิทธิภาพของ Association Discovery โดยลดจำนวนของตัวแทนข้อมูล หรือสุ่มตัวอย่างข้อมูลมาทำไมนิ่ง

2) ผู้ใช้สามารถระบุค่า Minimum Support และ Minimum Confidence ได้ ทำให้สามารถควบคุมจำนวนผลลัพธ์ได้

3) สามารถทำการไมนิ่งกับข้อมูลบางส่วนได้ทำให้ลดปัญหากรณีที่มีข้อมูลที่ไม่สมบูรณ์ได้

4) เทคนิคอื่น ๆ เช่น Neural Networks, Decision Trees จะระบุขอบเขตของกลุ่มข้อมูล ทำให้มีการจำกัดข้อมูล มีผลทำให้ข้อมูลที่ถูกเลือกมาทำดาต้าไมนิ่งอาจไม่ใช่ตัวแทนที่แท้จริงของกลุ่มข้อมูล

5) สามารถจัดการกับข้อมูลที่รูปแบบแตกต่างกันได้โดยไม่สูญเสียสารสนเทศ ขณะที่เทคนิคอื่นจะจำกัดรูปแบบและความยาวของข้อมูล

6) เนื่องจากมีการแสดงผลด้วยสัญลักษณ์ ทำให้ข้อมูลที่ได้จาก Association Rule ง่ายต่อการทำความเข้าใจกว่าผลลัพท์ที่ได้จากอัลกอริทึมของ Neural Networks หรือ Classification

3.1.4 ข้อเสียของ Association Rule

1) ถ้าใช้กับข้อมูลที่เกิดขึ้นไม่บ่อยใน Transaction จะทำให้ข้อมูลนี้แยกออกมาจากกลุ่มข้อมูลอื่นอย่างชัดเจน ทำให้ประสิทธิภาพของสารสนเทศที่ได้จากอัลกอริทึมนี้ลดลง

2) กฎที่ได้จากอัลกอริทึมนี้อาจมีปริมาณมากเกินไป ถึงแม้ว่าผู้ใช้สามารถกำหนดค่า Minimum Support และ Minimum Confidence ได้ เพื่อจำกัดจำนวนกฎที่จะสร้างขึ้น แต่อาจทำให้กฎที่ได้ผิดเพี้ยนไป เนื่องจากผู้ใช้กำหนดค่า Minimum Support และ Confidence สูงหรือต่ำเกินไป

4) บอกความแตกต่างของกฎที่ได้มายาก ว่าเป็นกฎจริง หรือกฎที่ได้มาจากการบังเอิญที่

ข้อมูลมาพ้องกัน

4) กฎที่ได้ไม่ได้ให้สารสนเทศถึงความเป็นเหตุเป็นผล ผู้ใช้ทราบว่าอะไรที่เป็นผลกระทบจากเหตุการณ์และหาความสัมพันธ์จาก Association Rule แต่กฎบอกได้เพียงอะไรมีแนวโน้มที่จะเกิดขึ้นด้วยกัน ไม่ได้ให้สารสนเทศเรื่องความเป็นเหตุเป็นผล

3.2 Sequencial Pattern Discovery

ใช้ระบุความเกี่ยวเนื่องกันของการซื้อสินค้าอย่างหนึ่ง และจะซื้อสินค้าอีกอย่างหนึ่งในเวลาต่อมา ดูลำดับการเลือกซื้อสินค้าของลูกค้าเนื่องจากมีจุดมุ่งหมายที่จะเข้าใจพฤติกรรมการซื้อสินค้าของลูกค้าในระยะยาว เช่น ผู้ขายอาจพบว่า ลูกค้าที่ซื้อทีวี มีแนวโน้มที่จะซื้อวีดีโอในเวลาต่อมา

ตัววัดความสัมพันธ์ในอัลกอริทึม Sequential Pattern Discovery เหมือนกับ Association Discovery แต่ Sequential Pattern Discovery มีวิธีการคำนวณหาค่า Support ที่ต่างออกไป โดยค่า Support คำนวณจาก อัตราส่วนจำนวนลูกค้าที่มีข้อมูลการซื้อสินค้าเป็นลำดับต่อจำนวนลูกค้าทั้งหมด แสดงได้ดังตารางที่ 3.1 รูปแสดงข้อมูลของร้านขายขนมปัง ข้อมูลจะเรียงลำดับตามรหัสลูกค้าและ รหัสรายการ เช่น B. Moor มาซื้อสินค้าที่ร้านเป็นเวลา 3 วันต่อกัน ซื้อเบียร์ในวันแรกตามด้วยไวน์และน้ำไซเดอร์ในวันถัดมา และซื้อขนมปังในวันที่ 3 นำข้อมูลเหล่านี้มาจัดเรียงตามลูกค้าได้ดังตารางที่ 3.2 โดยใส่ข้อมูลในวงเล็บตามลำดับของรายการที่เกิดขึ้น

Customer	Transaction Time	Items Bought
B. Adams	*June 21, 1994 5:37 p.m.*	Beer
B. Adams	June 20, 1994 10:30 a.m.	Brandy
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.	WineWine, Cider
B. Moore	June 22, 1994 5:03 p.m.	Brandy
F. Zappa	June 20, 1994 11:02 a.m.	Brandy

ตารางที่ 3.1 ข้อมูล Transaction เรียงลำดับตามลูกค้า[2]

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

ตารางที่ 3.2 การจัดกลุ่มTransaction ของลูกค้าแต่ละราย

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

ตารางที่ 3.3 ผลลัพธ์ที่ได้จาก Sequential Pattern Discovery

อัลกอริทึมในการทำงานของเทคนิค Sequential Pattern จะนับจำนวนความถี่ที่เกิดขึ้นของ Transaction ของลูกค้าซึ่งเรียงลำดับเหตุการณ์ไว้แล้ว โดยแสดงผลเฉพาะคู่เหตุการณ์ที่มีค่ามากกว่า Minimum Support

จากตารางที่ 3.3 ได้สารสนเทศออกมาเป็น "เมื่อลูกค้าซื้อเบียร์แล้วจะมีการซื้อขนมปังตามมาในภายหลัง" เหตุการณ์นี้เกิดขึ้นกับลูกค้า 2 คนจาก 5 คน

ข้อดีและข้อเสียของเทคนิค Sequential Pattern จะเหมือนกับ Association Discovery นอกจากนี้ยังมีจุดที่ควรระวัง 4 ประการคือ

1. Support Factor ระบุค่า parameter ได้เพียง 1 ค่า
2. จำนวนข้อมูลยิ่งมาก ยิ่งจำเป็นต้องมีการตรวจสอบให้มั่นใจว่ามีรหัสที่เป็นตัวแทนของข้อมูล Transaction ของลูกค้าแต่ละราย
3. ต้องมีฟิลด์พิเศษเพื่อเป็นตัวแทนของลูกค้า ซึ่งปกติในฐานข้อมูลการขายสินค้าทั่ว ๆ ไป มักจะไม่เก็บรหัสลูกค้าไว้ใน Transaction
4. เพื่อให้เทคนิคนี้ทำงานได้ดี ข้อมูลจะต้องเก็บเรียงตามลำดับเหตุการณ์ที่เกิดขึ้นของลูกค้าแต่ละราย

3.3 Similar Time Sequence

ใช้ค้นหาความเกี่ยวเนื่องกันระหว่างกลุ่มข้อมูล 2 กลุ่มซึ่งมีการขึ้นต่อกันทางด้านเวลา โดยมีรูปแบบการเคลื่อนไหวเหมือนกัน แทนข้อมูลในแนวแกน X ด้วยค่าของเวลา เช่น วันหรือเดือน แกน Y ด้วยค่าของตัวแปรที่สนใจ เทคนิคนี้มักใช้สำหรับการดูแนวโน้มยอดขาย เพื่อเตรียมตัดสต๊อคสินค้า สามารถดูข้อมูลได้ว่า ณ ช่วงเวลาแต่ละวันหรือแต่ละสัปดาห์ยอดขายของสินค้าใดมีค่าใกล้เคียงกัน อัลกอริทึมนี้จะแสดงผลด้วยกราฟ

ข้อดีของอัลกอริทึมนี้คือ เห็นการเคลื่อนไหวของข้อมูล เช่น ยอดขายที่เปลี่ยนแปลง การเคลื่อนไหวของราคาสินค้า การเคลื่อนไหวของสต๊อค

3.4 เทคนิคและอัลกอริทึมของ Apriori

Association Rules เป็นอัลกอริทึมหนึ่งของดาต้าไมนิ่งในการหาความสัมพันธ์ระหว่างสินค้าหรือบริการ กับลูกค้าในการซื้อสินค้าแต่ละครั้ง กฎที่ได้มักอยู่ในรูป $X \rightarrow Y$ โดยที่ X และ Y คือชุดของข้อมูล ตัวอย่างของกฎเช่น "40% ของลูกค้าที่ซื้อโค้กมักจะซื้อข้าวโพดคั่วด้วย, มีลูกค้า 15% ที่ซื้อทั้ง 2 อย่างพร้อมกัน" สามารถเขียนให้อยู่ในรูปของชุดข้อมูลได้ โดยที่ X = {BuyCoke}, Y = {BuyPopCorn} ค่า Confidence ของกฎ คือ 40 % ซึ่งคำนวณได้จากจำนวนเหตุการณ์ที่ซื้อโค้กและข้าวโพดคั่ว เทียบกับจำนวนเหตุการณ์ที่ซื้อโค้ก ค่า Support ของกฎ คือ 15% ซึ่งคำนวณได้จากจำนวนเหตุการณ์ที่ซื้อโค้กและข้าวโพดคั่วเทียบกับจำนวนเหตุการณ์ทั้งหมดในฐานข้อมูล

ขบวนการจัดการกับข้อมูลโดยใช้เทคนิค Association Rules โดยอาศัยอัลกอริทึมการทำงานของ Apriori มีหลักการทำงาน 2 ประการคือ

1. การหาชุดของข้อมูล (ItemSet) ในรายการขายที่มีค่าความถี่ในการเกิดมากกว่าหรือเท่ากับค่า Support ที่น้อยที่สุด (Minimum Support) โดยจะเรียก ItemSet นี้ว่า "Frequent ItemSet" หรือ "Large ItemSet"

2. การนำ "Frequent ItemSet" มาสร้างเป็นกฎ บนพื้นฐาน ถ้า ABCD และ AB เป็น "Frequent ItemSet" สามารถสร้างกฎ $AB \rightarrow CD$ โดยคำนวณค่า Confidence จาก Support(ABCD) / Support(AB) กฎที่ได้จะถูกต้องเมื่อมีค่าConfidence มากกว่าหรือเท่ากับค่า Confidence ที่น้อยที่สุด (Minimum Confidence)

อธิบายขั้นตอนการทำงานของอัลกอริทึมได้ในหัวข้อ 3.4.1 และ 3.4.2 ตามลำดับ โดยกำหนดให้ :

D คือ Database แต่ละ Transaction เก็บ <TID, Items>

TID คือ ตัวเลขระบุรายการ Transaction

Size คือ จำนวน Item ในเซ็ตของข้อมูล

k-ItemSet คือ เซ็ตของข้อมูลที่แต่ละเซ็ตประกอบด้วยสมาชิกจำนวน k ตัว

สรุปสัญลักษณ์ที่ใช้ในอัลกอริทึมได้ดังตารางที่ 3.4

k-ItemSet	เซ็ตของข้อมูลที่มีสมาชิกจำนวน k ตัว
L_k	Set ของ Frequent k-1 ItemSet ซึ่งทุกเซ็ตมีความถี่ในการเกิดมากกว่าหรือเท่ากับค่า Minimum Support สมาชิกในเซ็ตประกอบด้วย 2 ฟิลด์ คือ 1) ItemSet และ 2) Support Count
C_k	Set ของ Candidate k-1 ItemSet ที่ถูกเลือกมาจาก L_k สมาชิกในเซ็ตประกอบด้วย 2 ฟิลด์ คือ 1) ItemSet และ 2) Support Count

ตารางที่ 3.4 สัญลักษณ์ที่ใช้ใน Apriori Algorithm

3.4.1 การหา Frequent ItemSet

อัลกอริทึมในการหา "Frequent ItemSet" เริ่มจากการนับค่า Support สำหรับแต่ละ Item ในแต่ละรอบ โดยเลือกเฉพาะ Item ที่มีค่ามากกว่า Minimum Support จากนั้นวนลูปนับค่า Support ของ Subset ของ Item ที่ได้จากลูปก่อนหน้า เพื่อหาค่า "Candidate ItemSet" ในตอนจบของแต่ละลูปจะเลือก "Candidate ItemSet" ที่มีค่า Support มากกว่า Minimum Support ไปเป็นตัวตั้งต้นในการหา "Frequent ItemSet" ต่อไป ทำไปเรื่อย ๆ จนกว่าจะไม่สามารถหา "Frequent ItemSet" ได้ แสดงได้ดังภาพที่ 3.5

L_1= {frequent 1 – ItemSets};
for (k = 2; L_{k-1} <> ϕ}; k++ **do begin**
C_k= apriori-gen(L_{k-1}); //New Candidates
forall transactions $t \in D$ **do begin**
C_t = subset(C_k, t) ; //Candidates Contained in t
forall candidates $c \in C_t$ **do**
c.count++;
end
$L_k = \{c \in C_k | c.\text{count} \geq \text{minsup}\}$
end
Set of all Frequent ItemSets = $\bigcup_k L_k$;

ภาพที่ 3.5 อัลกอริทึมการหา Frequent ItemSet [4]

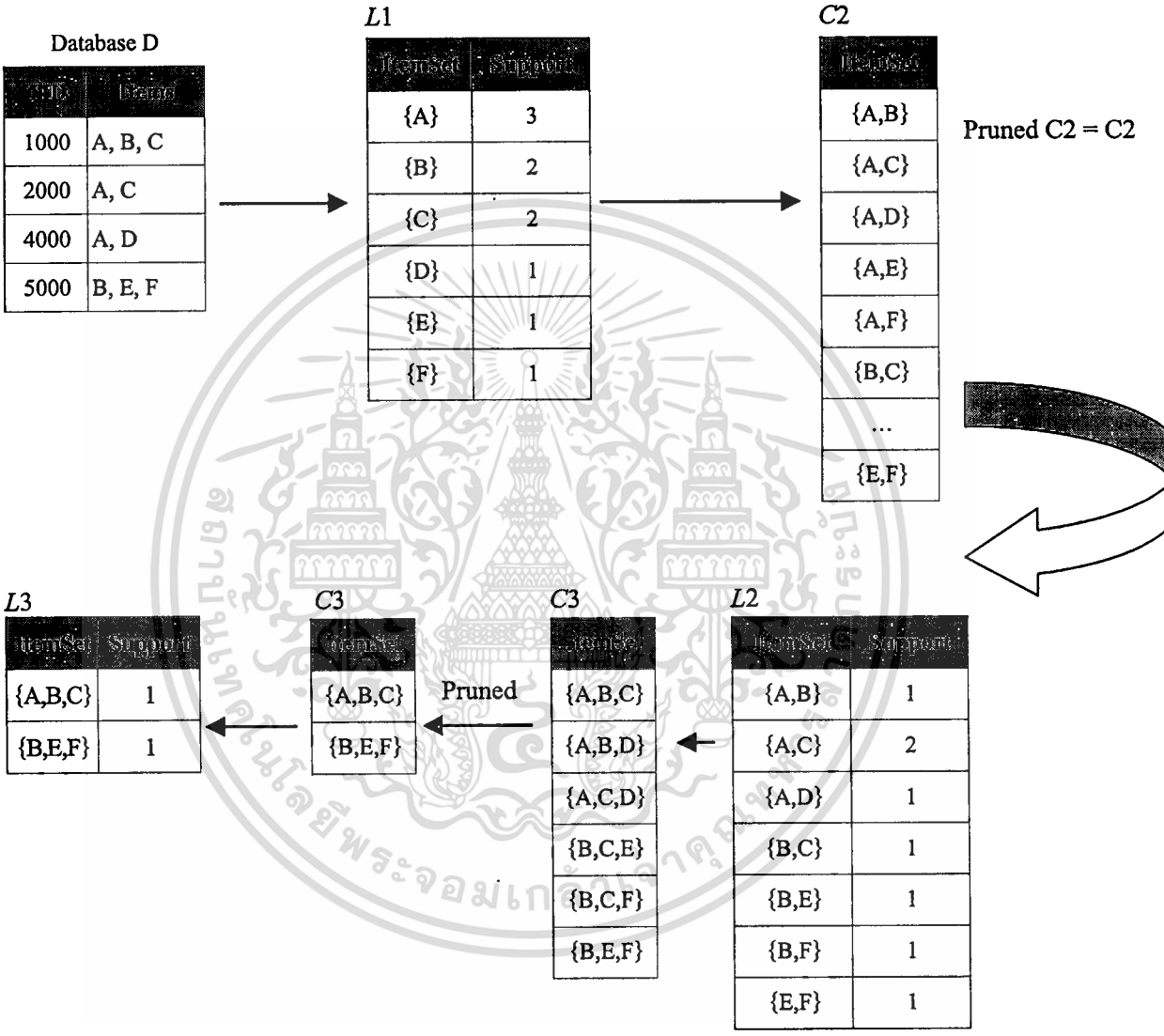
insert into C_k
select $p.\text{item}_1, p.\text{item}_2, \ldots, p.\text{item}_{k-1}$, $q.\text{item}_{k-1}$
from $L_{k-1} p, L_{k-1} q$
where $p.\text{item}_1 = q.\text{item}_1, \ldots, p.\text{item}_{k-2} = q.\text{item}_{k-2}, p.\text{item}_{k-1} < q.\text{item}_{k-1}$;

ภาพที่ 3.6 อัลกอริทึม Apriori-Gen [4]

forall itemsets $c \in C_k$ **do**
forall $(k-1)$ -- subsets s of c **do**
if ($s \notin L_{k-1}$) **then**
delete c from C_k;

ภาพที่ 3.7 ขั้นตอนการ Prunning [4]

ในขั้นตอนของการ Prune จะทำการลบ "Candidate ItemSet" ที่ทุก ๆ k-1 subset ของมันไม่ได้อยู่ใน Frequent ItemSets L_{k-1} ตัวอย่างของข้อมูลที่ผ่านขั้นตอนของการหา Frequent ItemSets โดยกำหนดให้ Minimum Support = 0.25 และ Minimum Confidence = 0.5 แสดงได้ดังภาพที่ 3.8



ภาพที่ 3.8 ตัวอย่างข้อมูลที่ผ่านการคำนวณจากอัลกอริทึมหา Frequent ItemSets [5]

3.4.2 การนำ Frequent ItemSet มาสร้างเป็นกฎ

นำ Frequent ItemSet ที่ได้จากขั้นตอน 2.4.1 มาสร้างเป็นกฎ โดยนำค่า Frequent ItemSet ตั้งแต่ L2 เป็นต้นไป มาคำนวณหา Subset และนำแต่ละ Subset ที่ได้มาสร้างเป็นกฎ อัลกอริทึมการทำงานแสดงได้ดังภาพที่ 3.9 และแสดงตัวอย่างข้อมูลที่ได้หลังจากการสร้างกฎได้ ดังภาพที่ 3.10

```
//Faster Algorithm
forall frequent k-itemsets l_k, k ≥ 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, H_m : set of m – item consequence )
        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 begin
                                output the rule (l_k - h_{m+11} ⇒ h_{m+1}), with confidence = conf
                                                and support = support(l_k);
                        else
                                delete h_{m+11} from H_{m+1};
                end
                call ap-genrules(l_k, H_{m+1} );
        end
```

ภาพที่ 3.9 อัลกอริทึมของ Genrules [5]

L2

ItemSet	Sup
{A,B}	1
{A,C}	2
{A,D}	1
{B,C}	1
{B,E}	1
{B,F}	1
{E,F}	1

Possible Rules →

Rules	Confidence	Support
B->A	0.5	1
A->C	0.67	2
C->A	1	2
D->A	1	1
B->C	0.5	1
C->B	0.5	1
B->E	0.5	1
E->B	1	1
B->F	0.5	1
F->B	1	1
E->F	1	1
F->E	1	1

L3

ItemSet	Sup
{A,B,C}	1
{B,E,F}	1

Possible Rules →

Rules	Confidence	Support
B->A&C	0.5	1
C->A&B	0.5	1
A&B->C	1	1
A&C->B	0.5	1
B&C->A	1	1
B->E&F	0.5	1
E->B&F	1	1
F->B&E	1	1
B&E->F	1	1
B&F->E	1	1
E&F->B	1	1

ภาพที่ 3.10 ผลลัพธ์จากการสร้างกฎ [5]

บทที่ 4
การประยุกต์ใช้ดาต้าไมนิ่งเพื่อวิคราะห์หาความสัมพันธ์ของข้อมูลการส่งออกสินค้า

เพื่อให้การศึกษาบรรลุวัตถุประสงค์ตามที่กำหนดไว้ จึงต้องมีการกำหนดวัตถุประสงค์และกระบวนการต่าง ๆ สำหรับจัดเตรียมข้อมูลให้อยู่ในรูปที่เหมาะสมก่อนที่จะนำมาใช้งาน ซึ่งมีขั้นตอนดังนี้

4.1 กำหนดวัตถุประสงค์

ปัญหาที่เป็นอยู่ของบริษัทในการดำเนินงานของฝ่ายการตลาดคือการกำหนดยอดขายและกลยุทธการขายของผลิตภัณฑ์ต่าง ๆ ในแต่ละตลาดไม่ประสบผลสำเร็จเท่าที่ควร โดยฝ่ายการตลาดมักจะใช้ข้อมูลจากยอดขายในอดีตมากำหนดเป้าหมายการขายในแต่ละปี จากนั้นให้แต่ละหน่วยงานที่รับผิดชอบผลิตภัณฑ์นั้น ๆ ทำการกำหนดกลยุทธการขายของแต่ละหน่วยงานเองอย่างอิสระโดยไม่มีความสัมพันธ์กับอีกหน่วยงานที่รับผิดชอบผลิตภัณฑ์อื่น ทำให้หน่วยงานการขายไม่สามารถทำยอดขายได้ตามเป้าที่กำหนดไว้ หรือบางครั้งสามารถทำยอดขายได้เกินเป้าหมายที่ตั้งไว้

จึงมีความคิดที่จะนำดาต้าไมนิ่งมาประยุกต์ใช้กับงานด้านการขายซึ่งเป็นธุรกิจส่งออกสินค้า โดยมีวัตถุประสงค์เพื่อวิเคราะห์หาความสัมพันธ์ต่าง ๆ ที่เกิดขึ้นในการขายสินค้า เพื่อที่ฝ่ายการตลาดสามารถนำผลการวิเคราะห์ไปเป็นประโยชน์ในการตั้งเป้าหมายยอดขายสินค้าในแต่ละปี ให้ตรงกับประเภทของสินค้าและตลาดสินค้า เพื่อที่จะเพิ่มยอดจำหน่ายสินค้าได้และกำหนดกลยุทธการขายของแต่ละผลิตภัณฑ์ได้

4.2 การคัดเลือกข้อมูล

ข้อมูลที่ต้องนำมาใช้ในการดำเนินการทั้งหมด รวบรวมมาจากฐานข้อมูลการส่งออกสินค้าของเครือซิเมนต์ไทย ซึ่งจัดเก็บอยู่บน Mainframe ด้วย DB2 และมีการโอนข้อมูลลงมาที่ฐานข้อมูลของ Informix ข้อมูลที่นำมาศึกษาจัดเก็บตั้งแต่ปี พ.ศ. 2537 ถึง พ.ศ. 2543

ทำการติดต่อกับฐานข้อมูลของ Informix โดยเชื่อมต่อข้อมูลมาใช้ที่ฐานข้อมูลของไมโครซอฟท์แอคเซส (Microsoft Access) โดยทำการเลือกมาเฉพาะข้อมูลจากตารางการขายระดับ Header ตารางการขายระดับ Detail ตารางข้อมูลสินค้า ตารางข้อมูลลูกค้า ตารางประเทศ และตารางเขตการขาย ทำการคัดเลือกเฉพาะ Attribute ที่สนใจจากตารางดังกล่าวได้ดังนี้

ตารางการขายระดับ Header

ชื่อข้อมูล	ประเภทข้อมูล
รหัสการขาย (So_no)	Text
รหัสลูกค้า (Cust_id)	Number
วันที่ขายสินค้า (Iss_date)	Date/Time
รหัสประเทศของลูกค้า(Cntry_id)	Text
รหัสเขตการขาย (Zone_id)	Text

ตารางที่ 4.1 ตารางการขายระดับ Header

ตารางการขายระดับ Detail

ชื่อข้อมูล	ประเภทข้อมูล
รหัสการขาย (So_no)	Text
ลำดับรายการ (Seq_no)	Number
รหัสสินค้า (Prd_id)	Text
จำนวนเงิน(Amount)	Currency
รหัสสกุลเงิน(Curr_id)	Text

ตารางที่ 4.2 ตารางการขายระดับ Detail

ตารางกลุ่มสินค้า

ชื่อข้อมูล	ประเภทข้อมูล
รหัสแผนกสินค้า (PrdGrp_id)	Text
ชื่อแผนกสินค้า(PrdGrp_longname)	Text

ตารางที่ 4.3 ตารางกลุ่มสินค้า

ตารางข้อมูลสินค้า

ชื่อข้อมูล	ประเภทข้อมูล
รหัสสินค้า(Prd_id)	Text
ชื่อสินค้า (Prd_longname)	Text
รหัสกลุ่มสินค้า (PrdGrp_id)	Text

ตารางที่ 4.4 ตารางข้อมูลสินค้า

ตารางข้อมูลลูกค้า

ชื่อข้อมูล	ประเภทข้อมูล
รหัสลูกค้า (Cust_id)	Number
ชื่อลูกค้า (Cust_name)	Text

ตารางที่ 4.5 ตารางข้อมูลลูกค้า

ตารางประเทศ

ชื่อข้อมูล	ประเภทข้อมูล
รหัสประเทศ (Cntry_id)	Text
ชื่อประเทศ (Cntry_name)	Text

ตารางที่ 4.6 ตารางประเทศ

ตารางเขตการขาย

ชื่อข้อมูล	ประเภทข้อมูล
รหัสเขตการขาย (Zone_id)	Text
ชื่อเขตการขาย (Zone_name)	Text

ตารางที่ 4.7 ตารางเขตการขาย

ตารางสกุลเงิน

ชื่อข้อมูล	ประเภทข้อมูล
รหัสสกุลเงิน (Curr_id)	Text
ชื่อสกุลเงิน (Curr_name)	Text

ตารางที่ 4.8 ตารางสกุลเงิน

จากนั้นทำการปรับข้อมูลและแก้ไขข้อมูลในกรณีที่ค่าของข้อมูลขาดหายไป ดังนี้

1. กรณีที่ในตารางการขายมีรหัสสินค้าที่ไม่มีอยู่ในตารางข้อมูลสินค้า ทำการตัดรายการขายสินค้านั้น ๆ ออกไป

2. กรณีที่มีรหัสลูกค้าในตารางการขาย แต่ไม่มีอยู่ในตารางข้อมูลลูกค้า ทำการตัดรายการขายสินค้านั้น ๆ ออกไป

3. กรณีวันที่ขายสินค้าในตารางการขายหายไป ทำการตัดทิ้ง เนื่องจากไม่สามารถคาดการณ์ได้ว่ารายการขายนั้น ๆ อยู่ในช่วงวันที่เท่าใด

เมื่อทำการตรวจสอบข้อมูลในขั้นต้นเสร็จแล้ว ทำการเตรียมข้อมูลให้อยู่ในรูปของตารางการขาย เพื่อนำเข้าไปวิเคราะห์ผลในระบบโดย ทำการรวบรวมข้อมูลของแต่ละเรคอร์ดในรายการขายที่มีวันที่เดียวกัน เข้าเป็นชุดของรายการขาย (Item Set) เดียวกัน แสดงข้อมูลที่ต้องเรียบเรียงใหม่เพื่อสร้างเป็นตารางการขายที่ใช้ในการวิเคราะห์และนำเข้าโปรแกรมได้ดังตารางที่ 4.9

ชื่อข้อมูล	ประเภทข้อมูล
วันที่ขายสินค้า(Iss_date)	Date/Time
รหัสลูกค้า(Cust_id)	Number
รหัสกลุ่มสินค้า(PrdGrp_id)	Text
จำนวนเงิน(So_amt)	Currency
รหัสสกุลเงิน(Curr_id)	Text
รหัสประเทศ(Cntry_id)	Text
รหัสเขตการขาย(Zone_id)	Text

ตารางที่ 4.9 ตารางการขาย 1

นอกจากการจัดเตรียมข้อมูลให้อยู่ในรูปของตารางที่ 4.9 ซึ่งเป็นการเตรียมข้อมูลในลักษณะตารางที่มีความสัมพันธ์กัน (Relational Table) แล้วยังสามารถจัดเตรียมข้อมูลให้อยู่ในรูปแบบอื่นได้อีก 3 รูปแบบ เพื่อนำข้อมูลไปวิเคราะห์ในระบบ

1. ทำการจัดกลุ่มของข้อมูลตามมุมมองที่สนใจ เช่น จัดกลุ่มของสินค้าตามเขตการขายที่สนใจที่มีวันที่ขายสินค้าเดียวกัน แล้วสร้างเป็นตารางใหม่ดังตารางที่ 4..10 ซึ่งเป็นการเตรียมข้อมูลในลักษณะของความสัมพันธ์ระหว่างรายการขายสินค้า สำหรับที่จะใช้วิเคราะห์หาความสัมพันธ์ในลักษณะที่เป็น Transaction

ชื่อข้อมูล	ประเภทข้อมูล
รหัสการขาย	TID
ชื่อกลุ่มสินค้า(PrdGrp_longname)	Text

ตารางที่ 4.10 ตารางการขาย 2

2. แปลงข้อมูลให้อยู่ในรูปของเท็กซ์ไฟล์ (Text File) เพื่อใช้ในกรณีที่ไม่สามารถใช้งานฐานข้อมูลได้โดยตรง โดยกำหนดให้มีเท็กซ์ไฟล์ที่เก็บชื่อแอททริบิว(Attribute) ของข้อมูล และเท็กซ์ไฟล์ที่เก็บข้อมูล

การจัดเตรียมข้อมูลลักษณะนี้จะใกล้เคียงกับข้อมูลในตารางที่ 4.9 คือ มีเท็กซ์ไฟล์ที่เก็บชื่อแอททริบิวของข้อมูลที่จะนำมาวิเคราะห์ ให้มีนามสกุล .idi ส่วนเท็กซ์ไฟล์ที่เก็บข้อมูลให้เก็บข้อมูลเรียงลำดับตามแอททริบิว บันทึกข้อมูลให้มีนามสกุล .dta

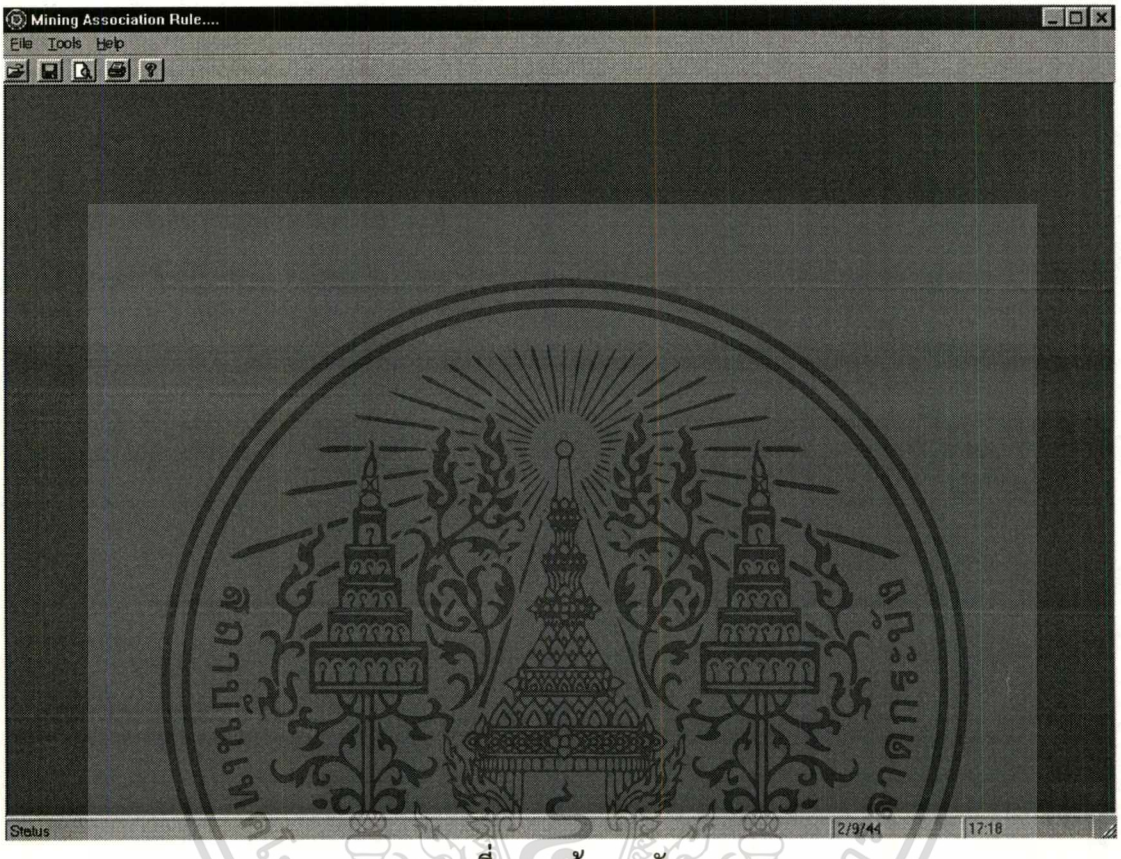
3. แปลงข้อมูลให้อยู่ในรูปของเท็กซ์ไฟล์ที่เก็บข้อมูลที่เป็นรายการขายตามรหัสการขาย การจัดเตรียมข้อมูลลักษณะนี้จะใกล้เคียงกับข้อมูลในตารางที่ 4.10 คือเป็นเท็กซ์ไฟล์ที่เก็บข้อมูลรายการขายที่เกิดขึ้นแต่ละรายการ

สำหรับการนำข้อมูลที่จัดเตรียมในลักษณะที่ 1 ถึง 3 มาใช้วิเคราะห์ในระบบ นี้จะกล่าวถึงโดยละเอียดในภาคผนวก

เมื่อได้ทำการจัดเตรียมข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดไปเป็นการวิเคราะห์ข้อมูลโดยใช้โปรแกรมที่พัฒนาขึ้น ได้เลือกใช้โปรแกรมวิชวลเบสิค เวอร์ชัน 6 (Visual Basic Version 6) ในการเขียนโปรแกรม

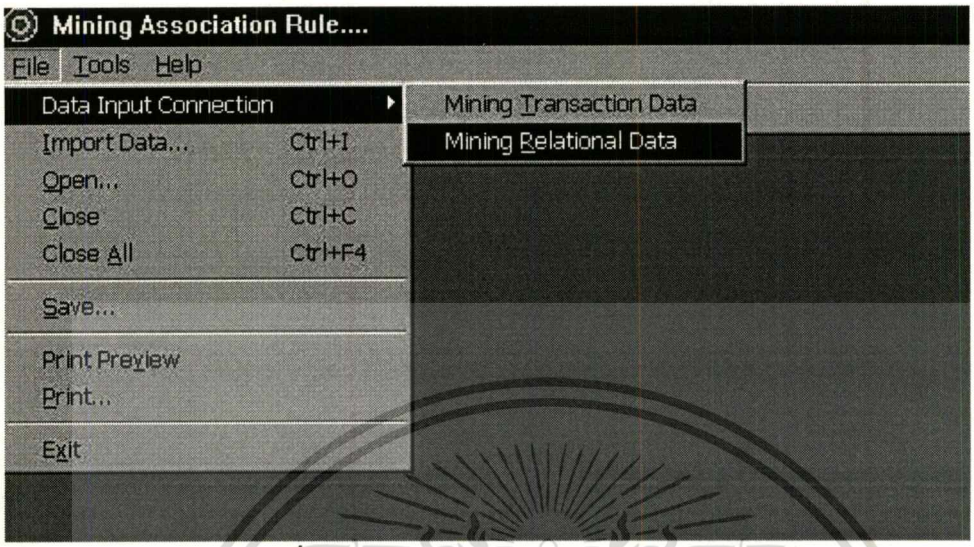
4.3 การติดต่อกับข้อมูลที่นำมาวิเคราะห์

เมื่อเข้าสู่โปรแกรม จะปรากฏเมนูหลัก ดังภาพที่ 4.1



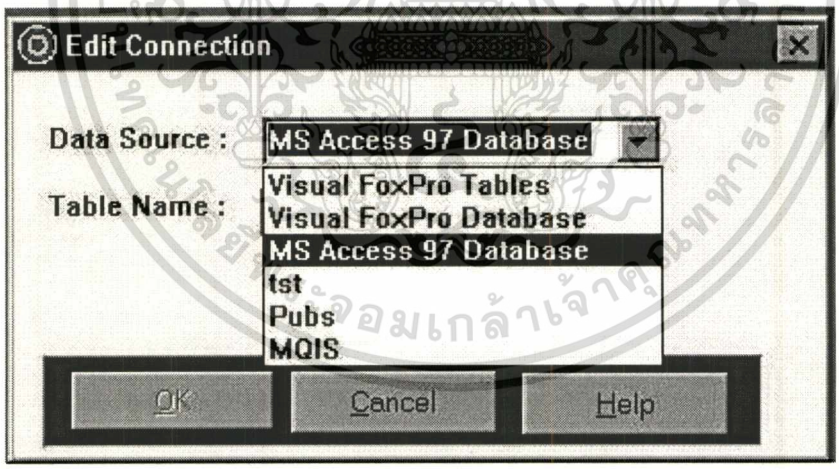
ภาพที่ 4.1 หน้าจอหลักของระบบ

จากเมนูหลักเลือกเมนูย่อย Data Input Connecction จะปรากฏเมนูย่อย 2 เมนู คือ Mining Transaction Data และ Mining Relational Data โดยเมนูแรกเป็นการติดต่อกับฐานข้อมูลที่เป็นรายการขายตามตารางที่ 4.10 และเมนูที่สองเป็นการติดต่อกับฐานข้อมูลที่เป็นรายการขายตามตารางที่ 4.9 ในที่นี้จะทำการวิเคราะห์ข้อมูลที่เป็น Relational Data จึงเลือกคลิกที่ Mining Relational Data ดังหน้าจอตัวอย่างภาพที่ 4.2



ภาพที่ 4.2 หน้าจอแสดงการเลือกวิธีการวิเคราะห์ข้อมูล

จากนั้นจะปรากฏหน้าจอให้เลือกวิธีติดต่อกับฐานข้อมูลดังภาพที่ 4.3



ภาพที่ 4.3 หน้าจอแสดงการเลือกฐานข้อมูลที่ต้องการติดต่อ

4.4 การตรวจสอบคุณภาพของข้อมูลและการจัดกลุ่ม

หลังจากติดต่อกับฐานข้อมูลเสร็จแล้ว จะเข้าสู่กระบวนการของการตรวจสอบคุณภาพของข้อมูล และจัดกลุ่มข้อมูล โดยแบ่งเป็นหัวข้อย่อยได้คือ

4.4.1 การเลือกตารางจากฐานข้อมูลที่ติดต่อมาวิเคราะห์

4.4.2 การกำหนดเงื่อนไขในการนำข้อมูลมาวิเคราะห์

4.4.3 การเลือกแอททริบิวมาวิเคราะห์

4.4.4 การตรวจสอบคุณภาพของข้อมูล

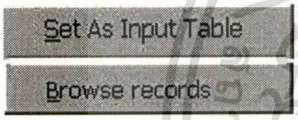
4.4.5 การจัดการกับค่าที่หายไป (Missing Value)

4.4.6 การจัดกลุ่มข้อมูล (Grouping)

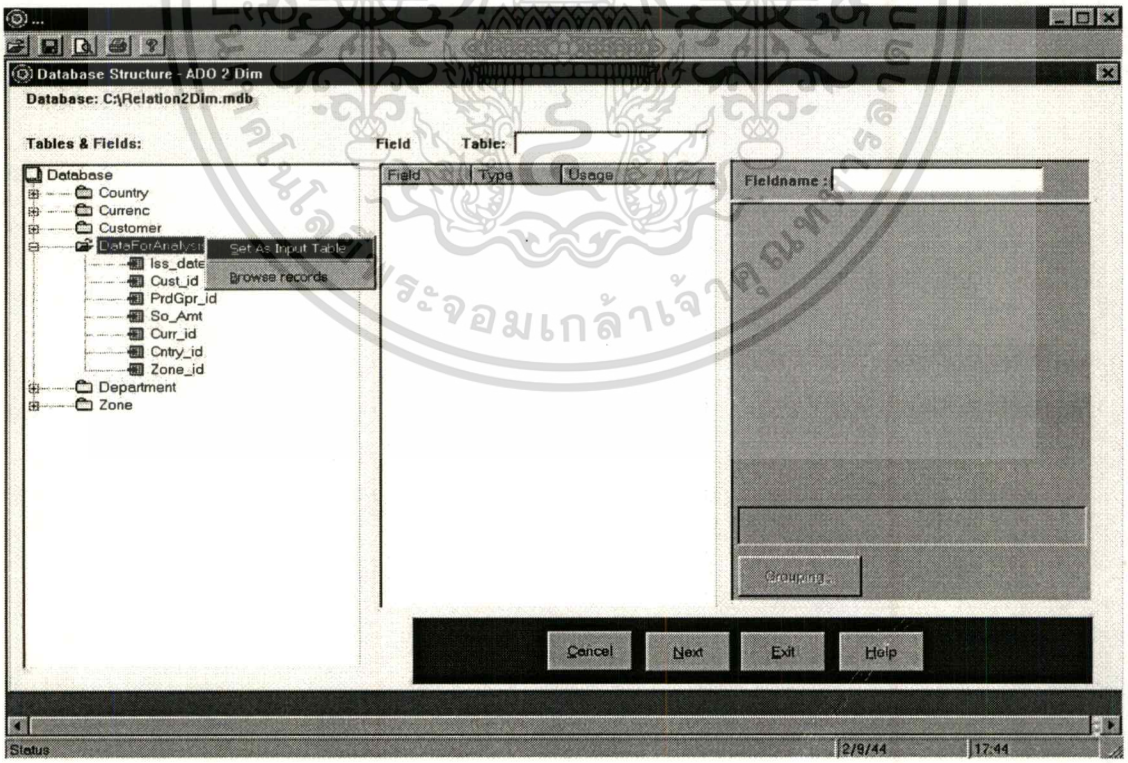
4.4.7 การเชื่อมต่อตาราง (Join)

4.4.1 การเลือกตารางจากฐานข้อมูลที่ติดต่อมาวิเคราะห์

หลังจากที่เลือกฐานข้อมูลมาแล้ว จะปรากฏหน้าจอแสดงรายละเอียดของตารางและแอทริบิวของตารางมาให้เลือก ว่าต้องการใช้ตารางไหนเป็นตารางหลักในการวิเคราะห์ข้อมูลโดยการคลิกเมาส์ขวาที่ชื่อตารางที่จะนำมาวิเคราะห์ จะปรากฏเมนูย่อย 2 เมนู คือ



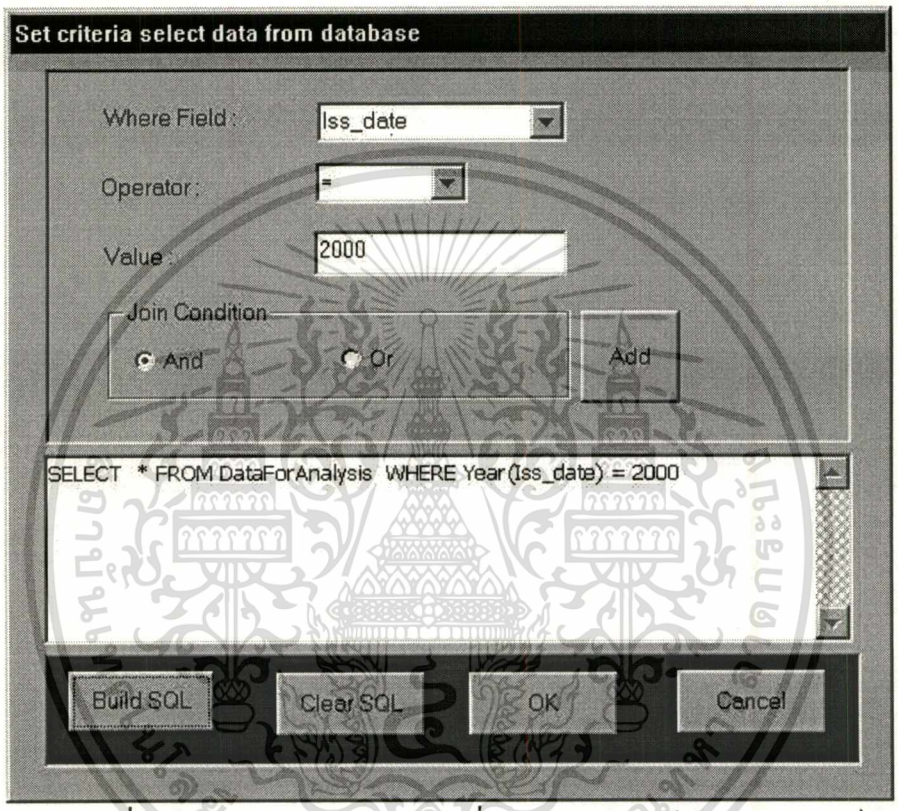
เพื่อกำหนดให้ตารางนั้นเป็นตารางที่ใช้ในการวิเคราะห์ และเมนูเพื่อดูข้อมูลในตารางนั้น ๆ แสดงตัวอย่างหน้าจอได้ดังภาพที่ 4.4



ภาพที่ 4.4 หน้าจอแสดงการเลือกตารางวิเคราะห์

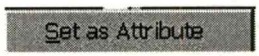
4.4.2 การกำหนดเงื่อนไขในการนำข้อมูลมาวิเคราะห์

หลังจากเลือกตารางที่วิเคราะห์เสร็จแล้ว จะเป็นการกำหนดเงื่อนไขในการเลือกนำข้อมูลจากตารางที่ต้องการติดต่อนั้นว่าต้องการช่วงของข้อมูลอย่างไร หรือมีเงื่อนไขอย่างไรบ้าง ในที่นี้ต้องการข้อมูลการขายเฉพาะปี พ.ศ. 2543 แสดงตัวอย่างของหน้าจอได้ดังภาพที่ 4.5



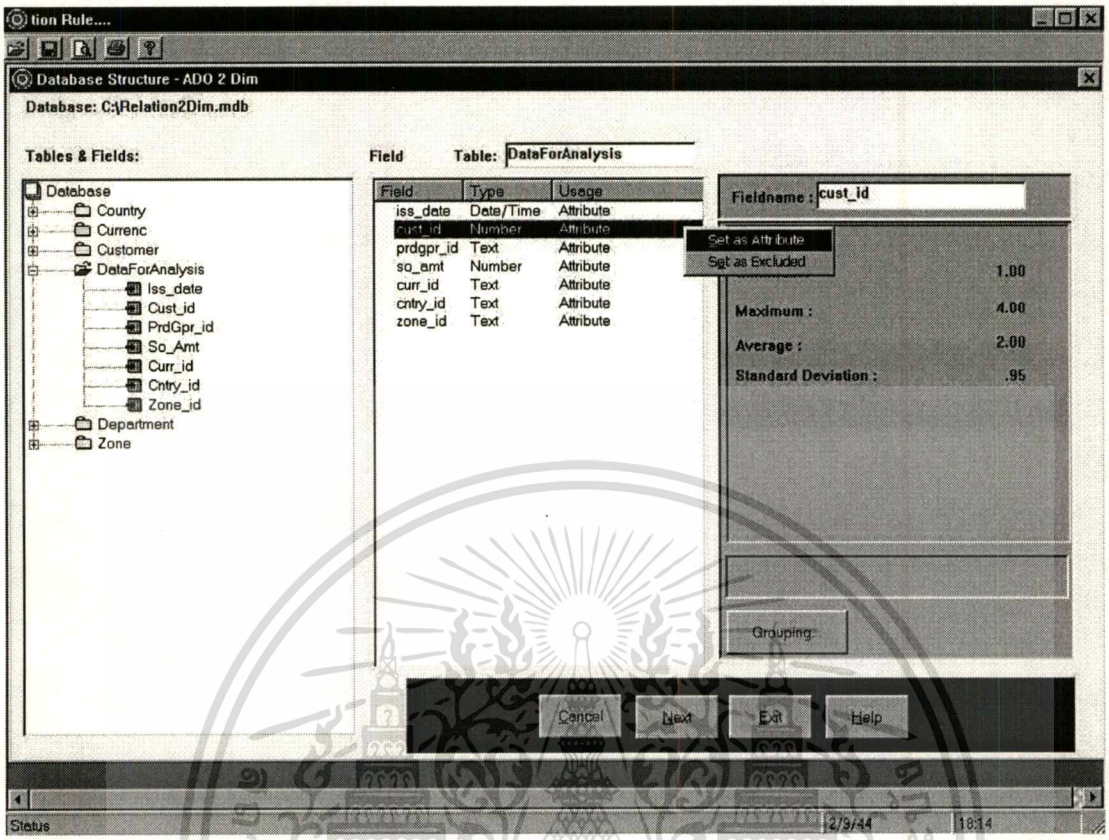
ภาพที่ 4.5 หน้าจอแสดงการกำหนดเงื่อนไขในการนำข้อมูลมาวิเคราะห์

4.4.3 การเลือกแอททริบิวมาวิเคราะห์

ระบบสามารถกำหนดได้ว่าต้องการใช้แอททริบิวไหนบ้างมาใช้วิเคราะห์โดยคลิกขวาที่ชื่อแอททริบิว จะปรากฏเมนูย่อย 2 เมนู คือ 

เพื่อกำหนดให้แอททริบิวนั้นใช้ในการวิเคราะห์ และ 

ให้แอททริบิวนั้นใช้วิเคราะห์ แสดงตัวอย่างหน้าจอได้ดังภาพที่ 4.6



ภาพที่ 4.6 หน้าจอแสดงการเลือกแอททริบิวที่นำมาวิเคราะห์

โดยทำการเลือกเฉพาะแอททริบิวที่สนใจมาวิเคราะห์ ได้แก่

วันที่ขายสินค้า(iss_date) (ซึ่งจะนำไปวิเคราะห์ในมุมมองของไตรมาส)

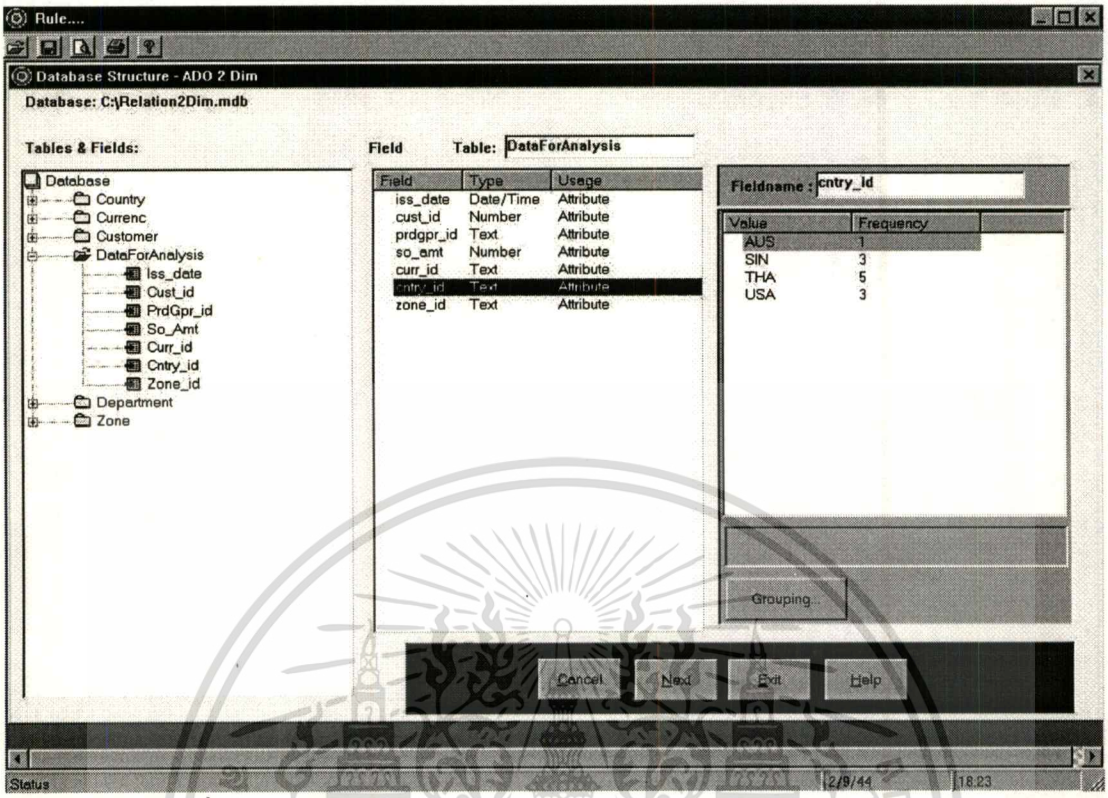
กลุ่มของสินค้า (prdgrp_id)

ประเทศปลายทาง(cntry_id)

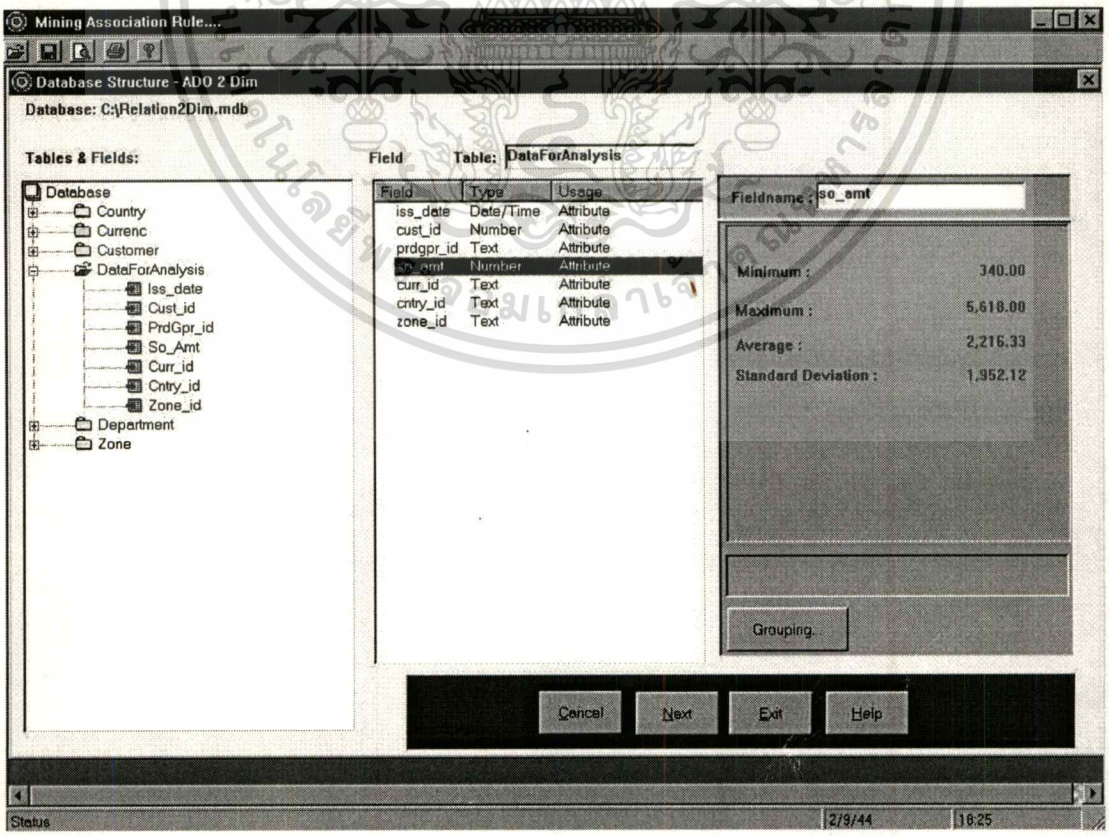
เขตการขาย (zone_id)

4.4.4 การตรวจสอบคุณภาพของข้อมูล

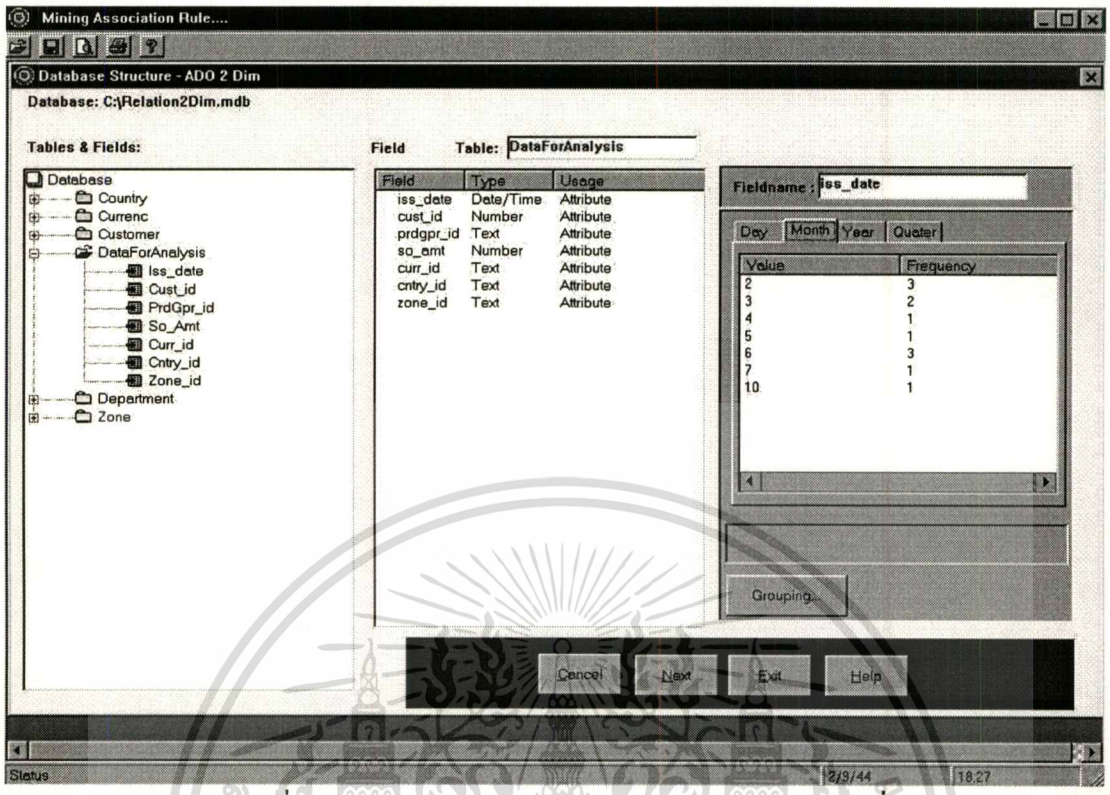
ระบบจะแสดงรายละเอียดของข้อมูลในแต่ละแอททริบิว กรณีที่แอททริบิวใดมีข้อมูลที่หายไป (Missing Value) ระบบจะแสดงข้อความเตือนที่มุมล่างของหน้าจอ พร้อมกับมีปุ่มคำสั่งให้จัดการกับค่าว่างนั้น โดยถ้าเป็นแอททริบิวที่มีชนิดของข้อมูลเป็นข้อความ (Tex) จะแสดงค่าของข้อมูลและค่าความถี่ที่เกิดขึ้นของข้อมูลในแอททริบิวนั้นๆ แสดงตัวอย่างหน้าจอดังภาพที่ 4.7 แอททริบิวที่มีชนิดข้อมูลเป็นตัวเลข(Number) จะแสดงค่าของสูงสุด, ต่ำสุด และค่าเฉลี่ยของข้อมูล แสดงตัวอย่างหน้าจอดังภาพที่ 4.8 และแอททริบิวที่มีชนิดข้อมูลเป็นวันที่(Date/Time) จะแสดงค่าของข้อมูลได้ตามปี, เดือน, วัน และไตรมาส แสดงตัวอย่างหน้าจอดังภาพที่ 4.9



ภาพที่ 4.7 หน้าจอแสดงรายละเอียดของแอททริบิวชนิดข้อความ



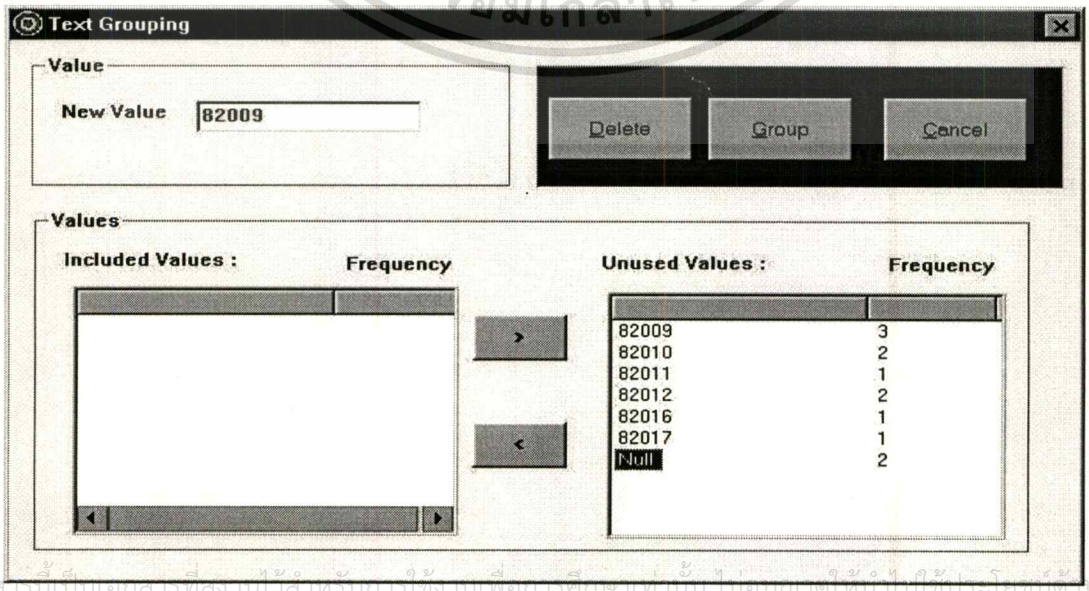
ภาพที่ 4.8 หน้าจอแสดงรายละเอียดของแอททริบิวชนิดตัวเลข



ภาพที่ 4.9หน้าจอแสดงรายละเอียดของแอททริบิวชนิดวันที่

4.4.5 การจัดการกับค่าที่หายไป (Missing Value)

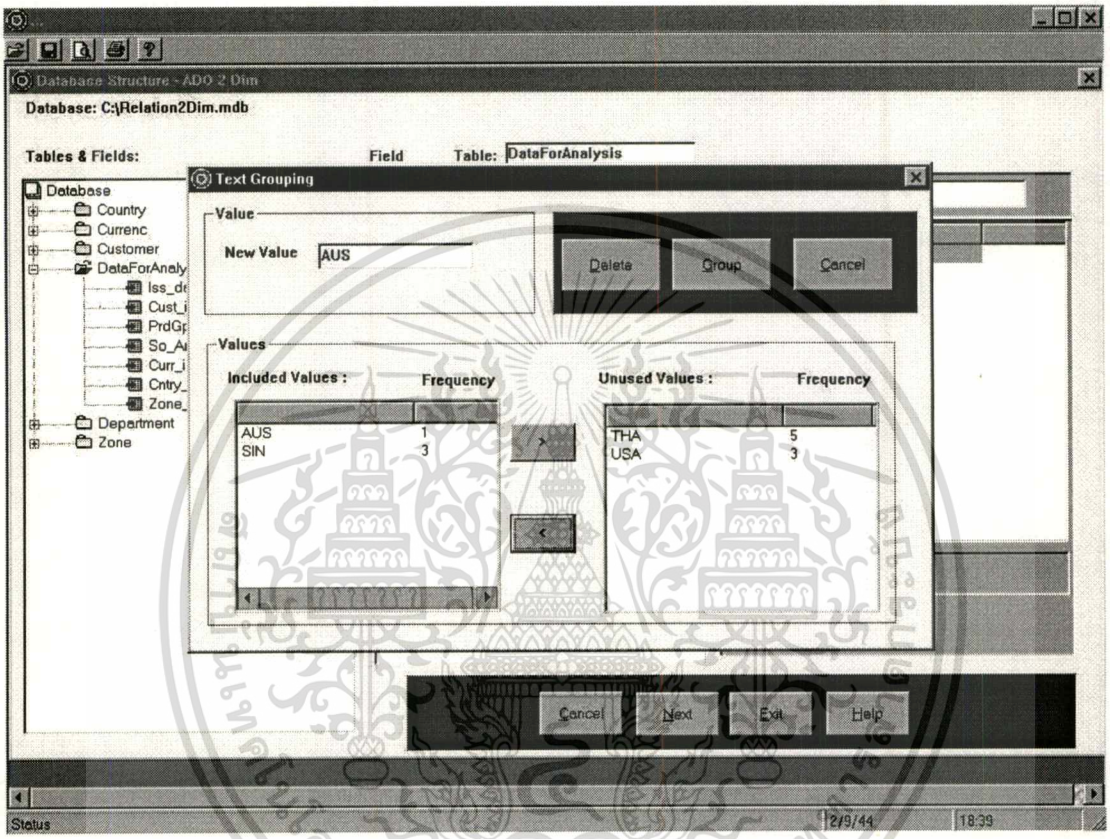
ระบบจะแสดงข้อความเตือนในมุมล่างของหน้าจอในกรณีที่มีค่าของข้อมูลที่ขาดหายไป คลิกที่ปุ่ม Missing Value... เพื่อทำการจัดการกับค่าว่าง ซึ่งสามารถที่จะลบเรคอร์ดที่ประกอบด้วยค่าว่างนี้ หรือจะจัดกลุ่มรวมกับค่าอื่น หรือจะแทนด้วยค่าใหม่ แสดงตัวอย่างของหน้าจอดังภาพที่ 4.10



ภาพที่ 4.10 หน้าจอแสดงการจัดการกับข้อมูลที่มีค่าที่หายไป

4.4.6 การจัดกลุ่มข้อมูล

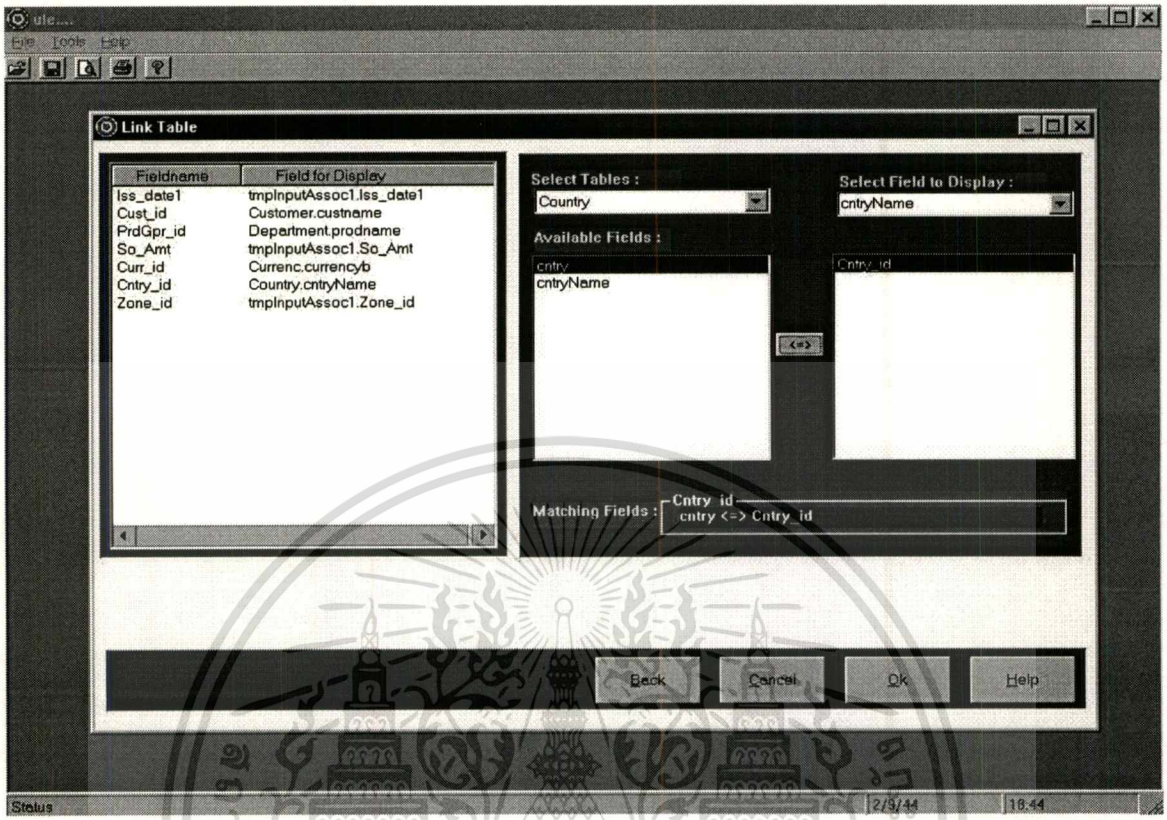
สำหรับข้อมูลที่มีค่าความถี่ที่เกิดขึ้นน้อย สามารถที่จะจัดกลุ่มข้อมูลรวมกับค่าอื่นได้ โดยคลิกที่ปุ่ม Grouping... จะปรากฏหน้าจอดังภาพที่ 4.11



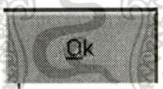
ภาพที่ 4.11 หน้าจอแสดงการจัดกลุ่มข้อมูล

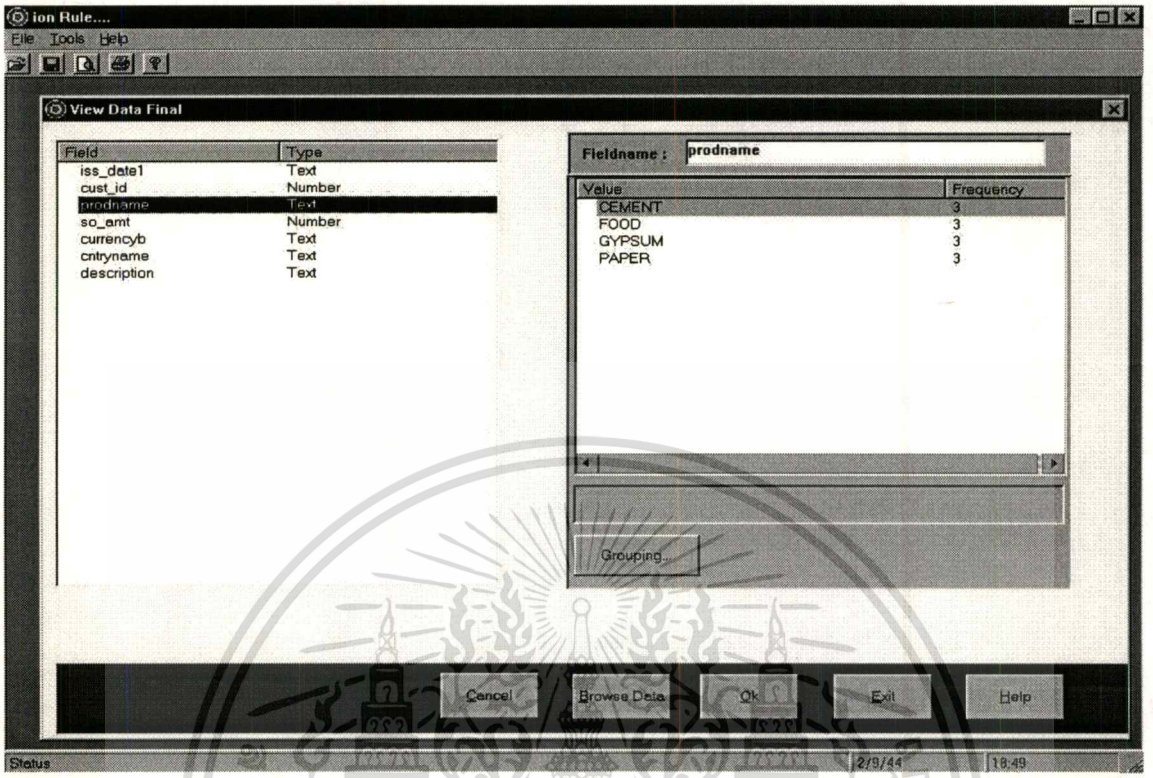
4.4.7 การเชื่อมตารางที่เลือกมาวิเคราะห์กับตารางอื่น

หลังจากที่จัดการกับค่าข้อมูลที่หายไปและจัดกลุ่มข้อมูลเสร็จแล้ว สามารถเชื่อมต่อตาราง (Join) กับตารางอื่นได้ เพื่อเลือกใช้แอททริบิวจากตารางอื่นมาใช้ในการวิเคราะห์แทน แสดงตัวอย่างหน้าจอดังภาพที่ 4.12



ภาพที่ 4.12 หน้าจอแสดงการเชื่อมต่อตาราง

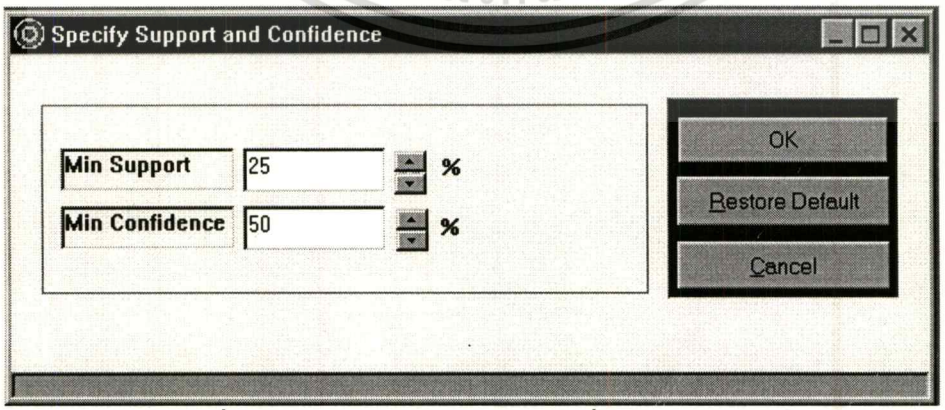
หลังจากเชื่อมตารางแล้วคลิกที่ปุ่ม  จะปรากฏหน้าจอให้ตรวจสอบข้อมูลอีกครั้งหนึ่ง ถ้าการเชื่อมตารางมีค่าของข้อมูลที่หายไปสามารถจัดการกับค่าว่างได้ หรือจัดกลุ่มข้อมูลได้ใหม่ เหมือนกับขั้นตอน 4.4.4 แสดงตัวอย่างของหน้าจอดังภาพที่ 4.13



ภาพที่ 4.13 หน้าจอแสดงการตรวจสอบข้อมูลหลังจากเชื่อมต่อตาราง

4.5 การกำหนดเงื่อนไขให้กับโปรแกรม

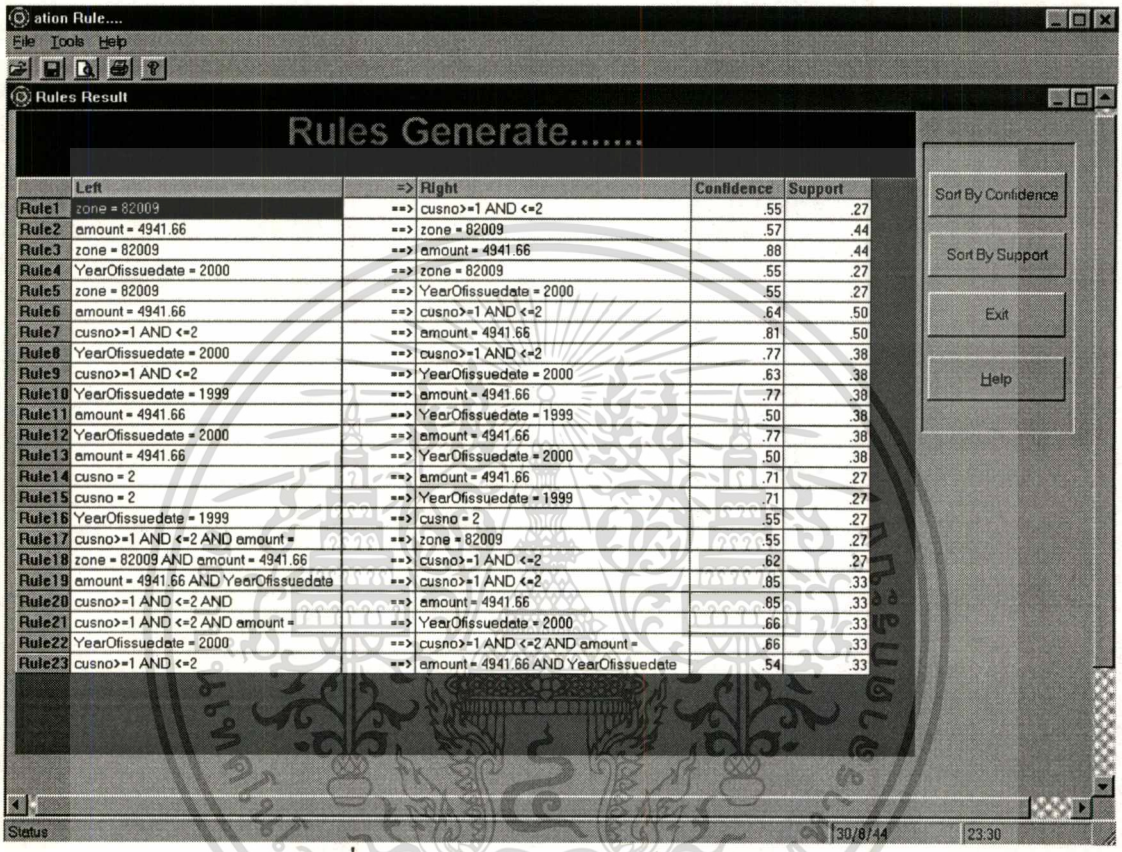
หลังจากที่ทำการการตรวจสอบคุณภาพของข้อมูลเสร็จแล้วจะเข้าสู่ขั้นตอนของการกำหนดค่า Minimum Suppport และ Minimum Confidence เพื่อกำหนดเงื่อนไขในการสร้างกฎ



ภาพที่ 4.14 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

4.6 การแสดงผล

หลังจากโปรแกรมทำการวิเคราะห์ข้อมูลเสร็จแล้ว จะแสดงหน้าจอผลลัพธ์จากการสร้างกฎดังภาพที่ 4.15



	Left	=>	Right	Confidence	Support
Rule1	zone = 82009	==>	cusno>=1 AND <=2	.55	.27
Rule2	amount = 4941.66	==>	zone = 82009	.57	.44
Rule3	zone = 82009	==>	amount = 4941.66	.88	.44
Rule4	YearOfissuedate = 2000	==>	zone = 82009	.55	.27
Rule5	zone = 82009	==>	YearOfissuedate = 2000	.55	.27
Rule6	amount = 4941.66	==>	cusno>=1 AND <=2	.64	.50
Rule7	cusno>=1 AND <=2	==>	amount = 4941.66	.81	.50
Rule8	YearOfissuedate = 2000	==>	cusno>=1 AND <=2	.77	.38
Rule9	cusno>=1 AND <=2	==>	YearOfissuedate = 2000	.63	.38
Rule10	YearOfissuedate = 1999	==>	amount = 4941.66	.77	.38
Rule11	amount = 4941.66	==>	YearOfissuedate = 1999	.50	.38
Rule12	YearOfissuedate = 2000	==>	amount = 4941.66	.77	.38
Rule13	amount = 4941.66	==>	YearOfissuedate = 2000	.50	.38
Rule14	cusno = 2	==>	amount = 4941.66	.71	.27
Rule15	cusno = 2	==>	YearOfissuedate = 1999	.71	.27
Rule16	YearOfissuedate = 1999	==>	cusno = 2	.55	.27
Rule17	cusno>=1 AND <=2 AND amount =	==>	zone = 82009	.55	.27
Rule18	zone = 82009 AND amount = 4941.66	==>	cusno>=1 AND <=2	.62	.27
Rule19	amount = 4941.66 AND YearOfissuedate	==>	cusno>=1 AND <=2	.85	.33
Rule20	cusno>=1 AND <=2 AND	==>	amount = 4941.66	.85	.33
Rule21	cusno>=1 AND <=2 AND amount =	==>	YearOfissuedate = 2000	.66	.33
Rule22	YearOfissuedate = 2000	==>	cusno>=1 AND <=2 AND amount =	.66	.33
Rule23	cusno>=1 AND <=2	==>	amount = 4941.66 AND YearOfissuedate	.54	.33

ภาพที่ 4.15 หน้าจอแสดงผลลัพธ์การสร้างกฎ

4.7 วิเคราะห์ผลการดำเนินงาน

ผลลัพธ์ที่ได้จากการทดสอบ โดยเลือกวิเคราะห์ยอดขายของปี พ.ศ. 2000 จำนวน 100 รายการ พบความสัมพันธ์จากการเลือกวิเคราะห์แอททริบิวต่าง ๆ ดังนี้

- เมื่อกำหนดให้วิเคราะห์ความสัมพันธ์ระหว่างลูกค้าและสินค้า โดยกำหนดให้มี Minimum Support เท่ากับ 20 และ Minimum Confidence เท่ากับ 50 พบความสัมพันธ์ที่น่าสนใจคือ
    1. ถ้าลูกค้าอยู่ถ้าลูกค้าอยู่ในเขตการขายอเมริกาเหนือ แล้วจะซื้อสินค้าเหล็กด้วยค่า Support 50.52 และ Confidence 80.76
    2. ถ้าลูกค้าอยู่ในเขตการขายจีน/ฮ่องกง แล้วจะซื้อสินค้าเซรามิคด้วยค่า Support 25.51 และ Confidence 76.44

- เมื่อกำหนดให้วิเคราะห์ความสัมพันธ์ระหว่างกลุ่มสินค้าและเวลา โดยกำหนดให้มี Minimum Support เท่ากับ 20 และ Minimum Confidence เท่ากับ 50 พบความสัมพันธ์ที่น่าสนใจคือ
    1. ถ้าขายสินค้าเซรามิคแล้วจะขายได้ดีในไตรมาสที่ 1 ด้วยค่า Support 91.66 และ Confidence 91
    2. ถ้าขายสินค้าเหล็กแล้วจะขายได้ดีในไตรมาสที่ 1 ด้วยค่า Support 83 และ Confidence 90
    3. ถ้าขายสินค้าปูนซิเมนต์แล้วจะขายได้ดีในไตรมาสที่ 2 ด้วยค่า Support 75 และ Confidence 80
- เมื่อกำหนดให้วิเคราะห์ความสัมพันธ์ระหว่างเขตการขาย กลุ่มสินค้าและเวลา โดยกำหนดให้มี Minimum Support เท่ากับ 20 และ Minimum Confidence เท่ากับ 50 พบความสัมพันธ์ที่น่าสนใจคือ
    1. ถ้าขายสินค้าเหล็กและขายในไตรมาสที่ 1 แล้วจะขายได้ดีในเขตอเมริกาเหนือด้วยค่า Support 59 และ Confidence 99
    2. ถ้าขายสินค้าเซรามิค แล้วจะขายได้ดีในไตรมาสที่ 1 และในเขตการขายอินโดไชนาด้วยค่า Support 50.21 และ Confidence 72.54
    3. ถ้าขายสินค้าปูนซิเมนต์ แล้วจะขายได้ดีในไตรมาสที่ 2 ในเขตการขายอินโดไชนา ด้วยค่า Support 45.32 และ Confidence 65.71
    4. ถ้าลูกค้าอยู่ในเขตอเมริกาเหนือเมื่อซื้อสินค้าเหล็ก แล้วจะซื้อสินค้าปูนซิเมนต์ เกิดเหตุการณ์เหล่านี้ด้วยค่า Support 25 และ Confidence 85
- เมื่อกำหนดให้วิเคราะห์ความสัมพันธ์ระหว่างเขตการขาย กลุ่มสินค้าและเวลา โดยเลือกวิเคราะห์เฉพาะเขตการขายจีน/ฮ่องกง กำหนดให้มี Minimum Support เท่ากับ 20 และ Minimum Confidence เท่ากับ 50 พบความสัมพันธ์ที่น่าสนใจคือ
    1. ถ้าลูกค้าซื้อสินค้าอลูมิเนียม แล้วจะซื้อสินค้าเซรามิคและยิบซั่ม ในไตรมาสที่ 2 ด้วยค่า Support 16 และ Confidence 80
    2. ถ้าลูกค้าซื้อสินค้าเหล็ก แล้วจะซื้อสินค้าปูนซิเมนต์ในไตรมาสที่ 1 ด้วยค่า Support 34 และ Confidence 75
    3. ถ้าลูกค้าซื้อสินค้าเซรามิค แล้วจะซื้อสินค้ายิปซั่ม ในไตรมาสที่ 3 ด้วยค่า Support 45 และ Confidence 65

จากการทดลองดังกล่าวพบความสัมพันธ์ที่น่าสนใจมากมาย ซึ่งต้องอาศัยกระบวนการวิเคราะห์ข้อมูลเพื่อตัดสินใจว่าความสัมพันธ์ไหนที่จะเป็นประโยชน์กับธุรกิจได้ เช่น จากความสัมพันธ์ของลูกค้า สินค้า และเวลาดังกล่าว ที่พบว่าลูกค้าที่อยู่ในเขตการขายจีน/ฮ่องกง เมื่อซื้อสินค้าเหล็กแล้วจะซื้อสินค้าปูนซิเมนต์ด้วย เหตุการณ์นี้เกิดในไตรมาสที่ 1 ทำให้ฝ่ายการตลาดนำข้อมูลดังกล่าวไปวางแผนการตลาดต่อไปได้ เช่นขยายตลาดการส่งออกสินค้ายิปซั่มไปที่เขตการขายจีน/ฮ่องกง โดยไม่ต้องการให้ยอดขายสินค้าปูนซิเมนต์และเหล็กลดลง บริษัทสามารถจัดทำโปรโมชั่นเมื่อลูกค้าซื้อสินค้าปูนซิเมนต์แล้วจะได้ส่วนลดในการซื้อสินค้ายิปซั่มไปด้วย ทั้งนี้ไม่ทำให้ยอดขายเหล็กลดลง เพราะเราค้นพบความสัมพันธ์แล้วว่า เมื่อลูกค้าซื้อสินค้าปูนซิเมนต์แล้วจะต้องซื้อสินค้าเหล็กตามไปด้วยอยู่แล้ว ดังนั้น บริษัทสามารถขยายตลาดการส่งออกยิปซั่มไปที่เขตการขายจีน/ฮ่องกงได้โดยไม่กระทบกับยอดขายเดิม

บทที่ 5
สรุปผลการศึกษาและข้อเสนอแนะ

โครงการพัฒนาระบบนี้เป็นโครงการที่จัดทำขึ้นมาเพื่อนำเสนอให้เห็นถึงประโยชน์ของการนำทฤษฎีของดาต้าไมนิ่งมาใช้เพิ่มประสิทธิภาพ ในการหาความสัมพันธ์ในรูปแบบต่าง ๆ ของการซื้อสินค้าของลูกค้าในฐานข้อมูลการส่งออกสินค้า เพื่อนำผลลัพธ์ที่ได้ไปใช้วางแผนทางการตลาดต่อไป

5.1 สรุปผลการดำเนินงาน

ดาต้าไมนิ่งเป็นกระบวนการที่ใช้เพื่อค้นหาข้อมูลที่มีประโยชน์ออกจากฐานข้อมูลเพื่อนำมาช่วยในการตัดสินใจ วิธีการแก้ปัญหาด้วยดาต้าไมนิ่งมีอยู่หลายรูปแบบขึ้นอยู่กับวัตถุประสงค์ของการทำงาน การที่จะนำเทคนิคของดาต้าไมนิ่งเข้ามาช่วยในการทำงานนั้น จำเป็นที่จะต้องเข้าใจลักษณะที่แท้จริงของเนื้องานก่อน เพื่อที่จะได้มีความสามารถในการกำหนดปัญหาและขอบเขตปัญหาได้ ถ้ากำหนดปัญหาได้ถูกต้องก็จะนำไปสู่ขั้นตอนการทำดาต้าไมนิ่งที่ถูกต้องและนำไปสู่ผลลัพธ์ที่ต้องการได้จากปัญหาขององค์กร

ในโครงการนี้ได้เสนอเทคนิคการหาความสัมพันธ์ระหว่างข้อมูลที่เกิดขึ้นโดยใช้อัลกอริทึมของ Apriori Algorithm ในการหา Association Rules ซึ่งเป็นอัลกอริทึมหนึ่งของ Link Analysis จากผลการศึกษาพบความสัมพันธ์ที่น่าสนใจหลายรูปแบบ ทำให้องค์กรสามารถกำหนดลูกค้ากลุ่มเป้าหมายสำหรับสินค้าแต่ละประเภทและแต่ละช่วงเวลาได้ ว่าลูกค้ากลุ่มใดคือกลุ่มเป้าหมายหลักของแต่ละสินค้าและสินค้าแต่ละประเภทควรมีกลยุทธการตลาดตามช่วงเวลาอย่างไร ทั้งนี้ข้อมูลจะเป็นประโยชน์ได้ต้องอาศัยกระบวนการขั้นตอนต่าง ๆ ดังที่ได้กล่าวมาคือ การตั้งวัตถุประสงค์ การคัดเลือกและจัดการกับข้อมูลที่จะนำมาวิเคราะห์ ถ้ามีขั้นตอนใดขั้นตอนหนึ่งผิดจะส่งผลให้ผลลัพธ์ที่ได้ผิดพลาดตามไปด้วย และต้องอาศัยกระบวนการวิเคราะห์ผลลัพธ์ที่ได้จากกฎด้วยว่ามีความสมเหตุสมผลกันไหม เพื่อดูถึงความเป็นไปได้ในการนำกฎที่ได้ไปประยุกต์ใช้เพื่อสร้างประโยชน์ให้กับองค์กรต่อไป ซึ่งผลลัพธ์ที่ได้ไม่อาจยืนยันได้ถึงความสำเร็จ 100 เปอร์เซนต์ในการกำหนดทิศทางตลาดแต่ก็เป็นเครื่องยืนยันอย่างหนึ่งถึงโอกาสของความสำเร็จที่จะเกิดขึ้น

5.2 ข้อเสนอแนะ

ระบบนี้สามารถที่จะลดขั้นตอนการทำงานลงได้ ด้วยการใช้ข้อมูลจาก DataWarehouse เพราะว่าจะเป็นข้อมูลที่ได้รับการทำความสะอาดมาเรียบร้อยแล้ว ทำให้สามารถลดขั้นตอนในการทำความสะอาดข้อมูลลงได้ ทั้งยังให้ความมั่นใจในความถูกต้องเพิ่มขึ้นด้วย

ระบบที่พัฒนาขึ้นมานี้สามารถใช้วิเคราะห์กับข้อมูลในธุรกิจอื่น ๆ ได้ เนื่องจากไม่ได้จำกัดขอบเขตไว้เฉพาะกับธุรกิจการส่งออกเท่านั้น ระบบสามารถติดต่อกับฐานข้อมูลที่เลือกได้หลายรูปแบบ และผู้ใช้เป็นคนกำหนดได้ว่าต้องการวิเคราะห์ข้อมูลอะไร จากตารางไหนได้บ้าง ดังนั้นสามารถนำระบบนี้ไปวิเคราะห์หาความสัมพันธ์ระหว่างข้อมูลของระบบอื่นได้

เอกสารอ้างอิง

[1] DBMS Data Mining Solution Supplement., "Association and Sequencing", [Online]. Available: http://www. dbmsmag.com/9807m03.html

[2] Simoudis, E. , 1998 , " Discovering Data Mining From Concept to implementation", Prentice Hall, New Jersey.

[3] Berry, M. J. and Linoff, G., 1997 , "Data Mining Techniques For Marketing, Sales and Customer Support", WILEY COMPUTER PUBLISHING.

[4] Rakesh Agrawal and Ramakrishnam Srikant, 1994 , "Fast Algorithms for Mining Association Rules", In *Proc. Of the 20th Int'l Conference on Very Large databases,* Santiago, Chile.

[5] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques.", [Online]. Available: http://www. cs.sfu.ca/~han/dmbook.

ภาคผนวก ก

คู่มือการใช้ระบบวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ AssociationRules

ก.1 การติดตั้งและการทำงานของระบบวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ Association Rules

ก.1.1 การติดตั้งโปรแกรม

ก่อนทำการติดตั้ง ควรทราบถึงความต้องการของระบบ ซึ่งสรุปได้ดังนี้

1) ความต้องการทางด้านซอฟต์แวร์

ซอฟต์แวร์ที่ต้องการใช้ มีดังนี้

(1.1) ระบบปฏิบัติการวินโดวส์ 95 ขึ้นไป

(1.2) โปรแกรมไมโครซอฟต์วิชวลเบสิค 6.0

(1.3) โปรแกรมไมโครซอฟต์แอกเซส เวอร์ชัน 97

2) ความต้องการทางด้านฮาร์ดแวร์

ความต้องการทางด้านฮาร์ดแวร์ สามารถสรุปได้ดังนี้

(2.1) เครื่องคอมพิวเตอร์ที่มีโปรเซสเซอร์เพนเทียมทรี ขึ้นไป

(2.2) เนื้อที่ว่างบนฮาร์ดดิสก์อย่างน้อย 2 เมกะไบต์

(2.3) หน่วยความจำสำรอง อย่างน้อย 128 เมกะไบต์

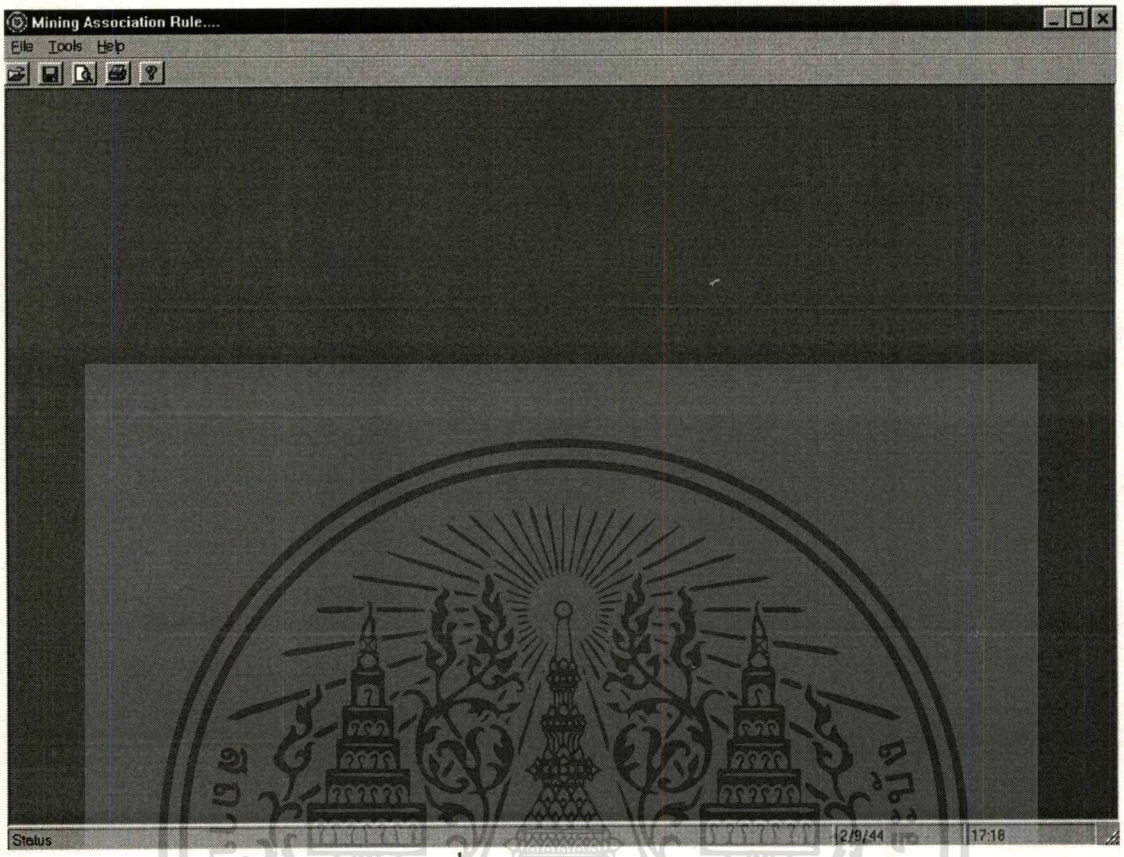
(2.4) จอภาพชนิด Super VGA ความละเอียดของจอภาพอย่างน้อย 256 สี

การติดตั้งโปรแกรม ผู้ค้นคว้าได้ทำแผ่นดิสเกตต์สำหรับติดตั้งโปรแกรม จำนวน 3 แผ่น เมื่อต้องการติดตั้งโปรแกรม ให้ผู้ใช้นำแผ่นดิสก์เกตต์ที่ต้องการติดตั้งใส่ในไดรฟ์ A แล้วดับเบิ้ลคลิกที่ไฟล์ Setup.exe หรือ พิมพ์คำสั่ง run a:\setup.exe ที่หน้าจอรับคำสั่ง จากนั้นโปรแกรมจะทำการติดตั้งไว้ไดรฟ์ C:\Program Files\MiningAsoc

ก.1.2 การทำงานของระบบวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ Association Rules

การเรียกใช้งานระบบ สามารถทำได้โดยดับเบิ้ลคลิกที่ไอคอน MiningAssoc ที่ระบบสร้างไว้ให้ที่เดสท์ทอปของเครื่อง หรือคลิกที่โปรแกรมไฟล์ และเลือกเมนูย่อย Mining Assoc

เมื่อถึงขั้นตอนนี้ ระบบวิเคราะห์ความสัมพันธ์ของข้อมูลการส่งออกสินค้าโดยใช้ Association Rules พร้อมที่จะทำงานแล้ว ซึ่งจะแสดงหน้าจอแรกดังภาพที่ ก.1



ภาพที่ ก.1 หน้าจอแรกของระบบ

ก.1.3 เมนูและปุ่มคำสั่ง

1) เมนู (File)

ประกอบด้วยเมนูย่อย 9 เมนู คือ

1.1) เมนู Data Input Connection ใช้สำหรับเลือกติดต่อกับข้อมูลที่นำมาวิเคราะห์ที่เป็นฐานข้อมูล ซึ่งประกอบด้วยเมนูย่อย 2 เมนูย่อยคือ

1. Mining Transaction Data ใช้เพื่อวิเคราะห์ข้อมูลในลักษณะที่เป็นตารางที่ประกอบด้วย Transaction ID ที่เป็นรหัสตัวแทนของรายการขาย และ Item ID ที่เป็นรายการขาย
2. Mining Relational Data ใช้เพื่อวิเคราะห์ข้อมูลในลักษณะที่เป็นตารางที่มีความสัมพันธ์กัน (Relational Table)

1.2) เมนู Import Data ใช้สำหรับติดต่อกับข้อมูลที่นำมาวิเคราะห์ที่เป็นเท็กซ์ไฟล์ (Text File)

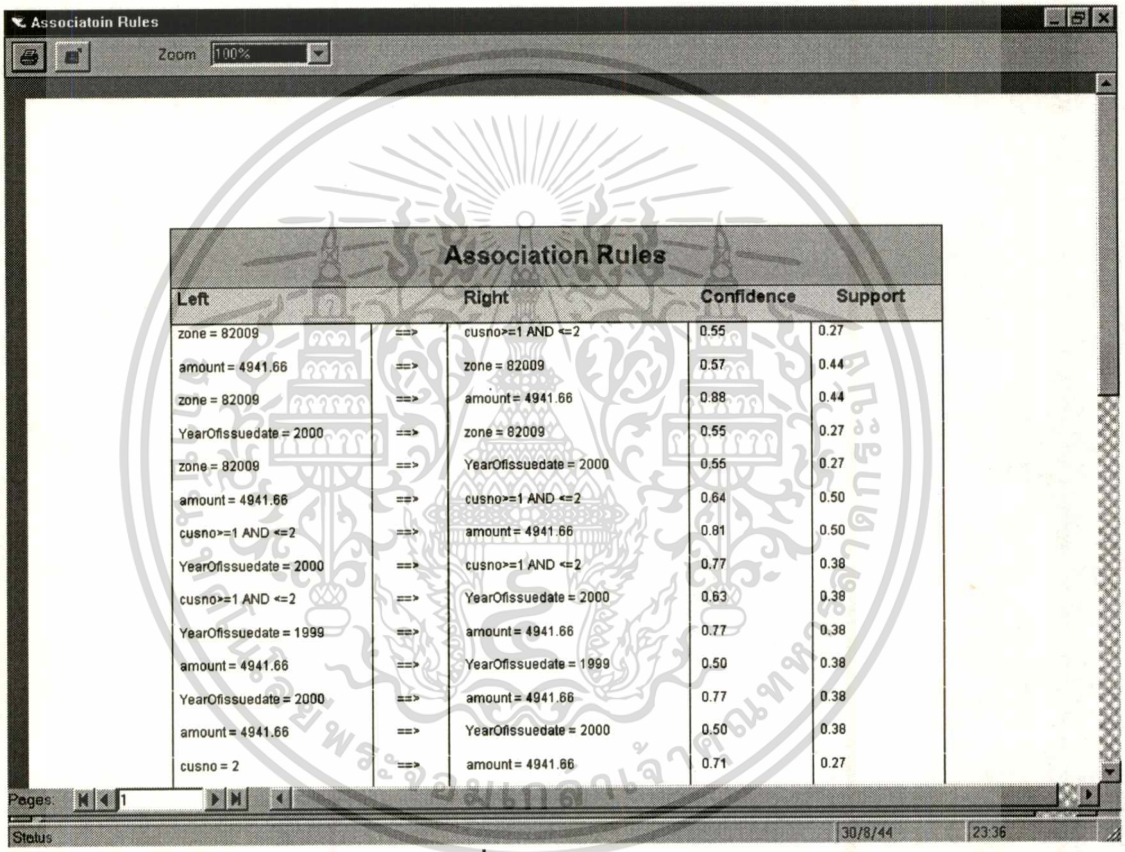
1.3) เมนู Open ใช้เพื่อเรียกดูกฎที่ได้ทำการบันทึกไว้แล้ว นามสกุล .rul

1.4) เมนู Close ใช้เพื่อปิดหน้าต่างที่เปิดอยู่ขณะนั้น

1.5) เมนู Close All ใช้เพื่อปิดหน้าต่างการทำงานทั้งหมดในระบบ

1.6) เมนู Save ใช้เพื่อบันทึกกฎที่ระบบสร้างให้ เพื่อเรียกดูภายหลังด้วยโปรแกรมนี้ หรือเรียกดูด้วยโปรแกรมไมโครซอฟท์เอ็กเซล (Microsoft Excel)

1.7) เมนู Print Preview ใช้เพื่อเรียกดูกฎที่ระบบสร้างให้ก่อนสั่งพิมพ์ แสดงตัวอย่างของหน้าจอ Print Preview ได้ดังภาพที่ ก.2

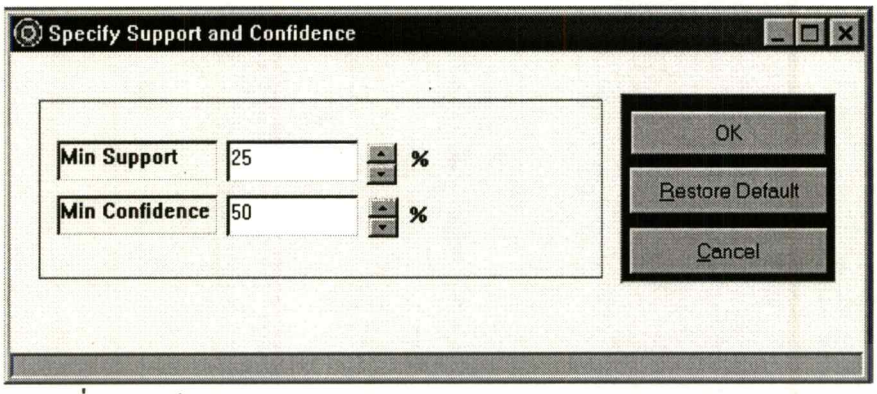


ภาพที่ ก.2 หน้าจอการ Print preview

1.8) เมนู Print ใช้เพื่อสั่งพิมพ์กฎที่เรียกดู

1.9) เมนู Exit ใช้เพื่อออกจากระบบ

2) เมนู Tools ใช้เพื่อกำหนดค่า Minimum Support และ Minimum Confidence แสดงตัวอย่างหน้าจอได้ดังภาพที่ ก.3



ภาพที่ ก.3 หน้าจอการกำหนดค่า Minimum Support และ Minimum Confidence

3) เมนู Help ใช้เพื่อเรียกดูคำสั่งช่วยเหลือการทำงาน
4) ปุ่ม ใช้เพื่อเรียกดูกฎที่ได้ทำการบันทำไว้แล้ว นามสกุล .rul
5) ปุ่ม ใช้เพื่อบันทึกกฎที่ระบบสร้างให้ เพื่อเรียกดูภายหลังด้วยโปรแกรมนี้ หรือเรียกดูด้วยโปรแกรมไมโครซอฟท์เอ็กเซล (Microsoft Excel)
6) ปุ่ม ใช้เพื่อเรียกดูกฎที่ระบบสร้างให้ก่อนสั่งพิมพ์
7) ปุ่ม ใช้เพื่อสั่งพิมพ์กฎที่เรียกดู
8) ปุ่ม ใช้เพื่อเรียกดูคำสั่งช่วยเหลือการทำงาน

ก.2 เริ่มต้นการทำงาน

สำหรับโปรแกรมที่พัฒนาขึ้น สามารถติดต่อกับข้อมูลที่นำมาวิเคราะห์ได้ 2 รูปแบบคือ ข้อมูลที่เป็นฐานข้อมูล จะใช้การติดต่อผ่าน ODBC (Object Database Connectivity) หรือผ่านไดรฟ์เวอร์ของโปรแกรมฐานข้อมูลนั้น เช่น โปรแกรมไมโครซอฟแอคเซส (Microsoft Access) และ ข้อมูลที่เป็นเท็กซ์ไฟล์ โดยโปรแกรมสามารถวิเคราะห์ข้อมูลได้ใน 2 ลักษณะ คือ

1. ข้อมูลที่มีความสัมพันธ์กัน (Relational Data)
2. ข้อมูลที่เป็น Transaction ID และรายการขาย (Transaction Data)

ก.2.1 ข้อมูลที่มีความสัมพันธ์กัน (Relational Data)

เป็นข้อมูลที่มีความสัมพันธ์ระหว่างหลาย ๆแอททริบิว ผลที่ได้จากการสร้างกฎ เช่น

ProductGroup = "CEMENT" AND Zone = "Indochina" ==> Quarter = "1",.50,.75

หมายถึง ถ้าเป็นการขายของกลุ่มสินค้าซิเมนต์และเขตการขายอินโดไชนาแล้วจะเป็นการ

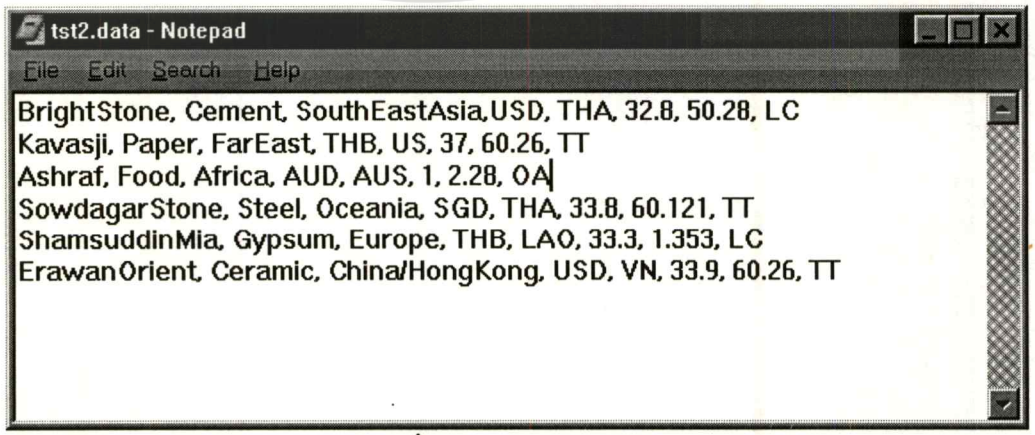
ขายที่กิดขึ้นในไตรมาสที่ 1 โดยเหตุการณ์นี้เกิดขึ้นด้วยค่า Confidence 50% และ Support 75 %

- กรณีที่เป็นการเตรียมข้อมูลด้วยตาราง (Table) จะคล้ายกับฐานข้อมูลของระบบ Relational Database ทั่ว ๆ ไป และนำตาราง Tranaction Data มาใช้วิเคราะห์ โดยระบบสามารถเชื่อมความสัมพันธ์ระหว่าง Transaction Data และ Master Data ใน Relational Database ได้
- กรณีที่เป็นการเตรียมข้อมูลด้วยเท็กซ์ไฟล์ (Text File) จะประกอบด้วยไฟล์นามสกุล .nam และ .dat โดย

  .name : เก็บข้อมูลที่เป็นชื่อแอททริบิวของข้อมูลและประเภทของข้อมูลว่าเป็นข้อมูลที่เป็น Continuous หรือ Dicrete แสดงตัวอย่างของข้อมูลดังภาพที่ ก.4

  .dat : เก็บข้อมูลที่เป็น Transaction ของรายการเรียงลำดับตามชื่อแอททริบิวที่อยู่ในไฟล์ .nam คั่นด้วยเครื่องหมายคอมมา (,) แสดงตัวอย่างของข้อมูลดังภาพที่ ก.5

ภาพที่ ก.4 ตัวอย่าง file .nam



ภาพที่ ก.5 ตัวอย่าง file .dat

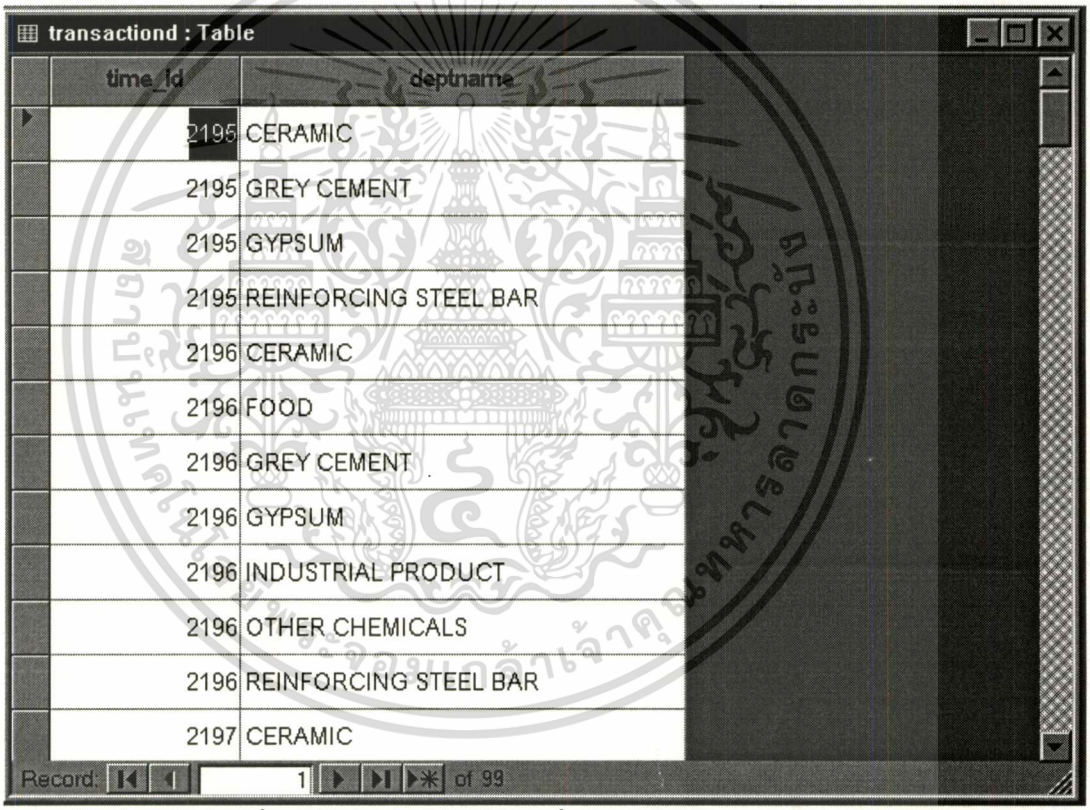
ก.2.2 ข้อมูลที่เป็น Transaction Data

เป็นข้อมูลที่มีความสัมพันธ์ระหว่างรายการขาย (TID) และรายการสินค้า ผลที่ได้จากการสร้างกฎ เช่น

CEMENT ==> STEEL, .50, .75

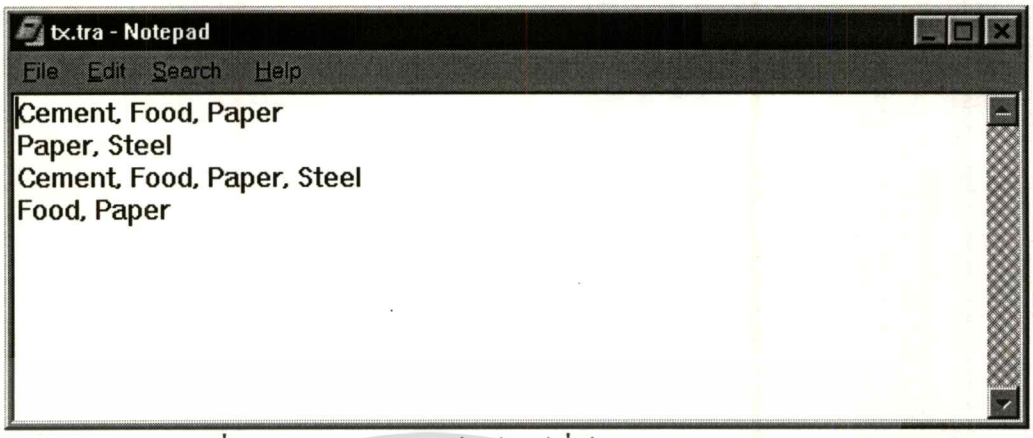
หมายถึง เมื่อลูกค้าซื้อสินค้าซิเมนต์แล้วจะซื้อสินค้าเหล็กตามไปด้วย โดยเหตุการณ์นี้เกิดขึ้นด้วยค่า Confidence 50% และ Support 75 %

- กรณีที่เป็นการเตรียมข้อมูลด้วยตาราง (Table) จะอยู่ในรูปของ TID และ VALUE ดังตัวอย่างในภาพที่ ก.6



ภาพที่ ก.6 ตัวอย่างตารางข้อมูลที่เป็น Transaction Data

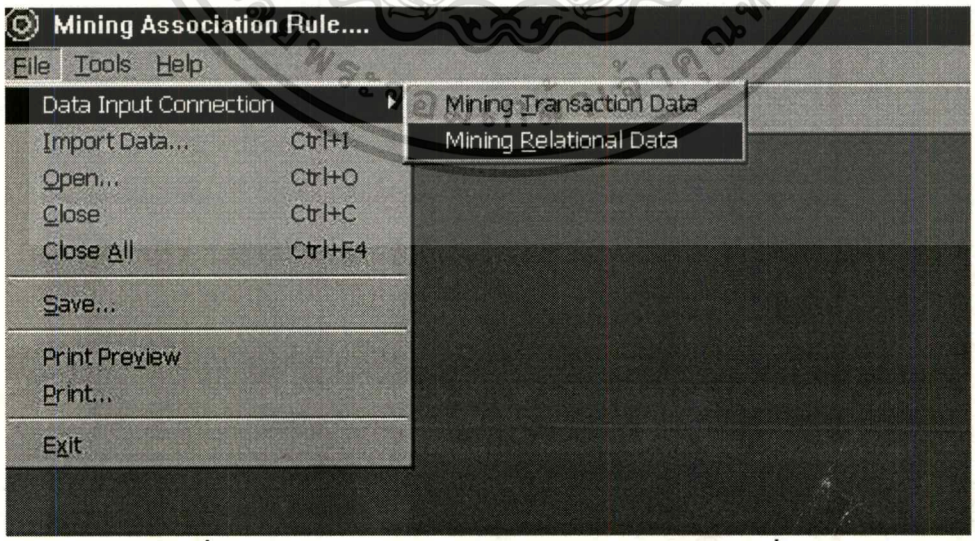
- กรณีที่เป็นการเตรียมข้อมูลด้วยเท็กซ์ไฟล์ จะอยู่ในรูปของ TID และ VALUE โดยคั่นด้วยเครื่องหมายคอมมา (,) บันทึกเป็นไฟล์นามสกุล .tra ดังตัวอย่างในภาพที่ ก.7



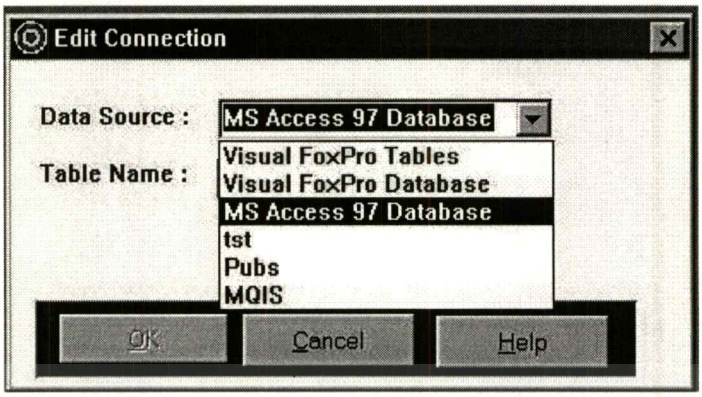
ภาพที่ ก.7 ตัวอย่างข้อมูลเท็กซ์ไฟล์ที่เป็น Transaction Data

ก.3 การวิเคราะห์ข้อมูลจากฐานข้อมูล

ข้อมูลที่เป็นฐานข้อมูล จะใช้การติดต่อผ่าน ODBC (Object Database Connectivity) หรือผ่านไดรฟ์เวอร์ของโปรแกรมฐานข้อมูลนั้น เช่น โปรแกรมไมโครซอฟแอคเซส (Microsoft Access) จากเมนูหลักเมื่อเลือกเมนู Data Input Connecction จะปรากฏหน้าจอดังภาพที่ ก.8 ซึ่งประกอบด้วยเมนูย่อย 2 เมนู คือ Mining Transaction Data และ Mining Relational Data โดยเมนูแรกเป็นการติดต่อกับฐานข้อมูลที่เป็น Relational Data และเมนูที่สองเป็นการติดต่อกับฐานข้อมูลที่เป็น Transaction Data หลังจากเลือกว่าต้องการติดต่อกับฐานข้อมูลลักษณะไหนเสร็จแล้ว จะปรากฏหน้าจอให้เลือกวิธีติดต่อกับฐานข้อมูลดังภาพที่ ก.9

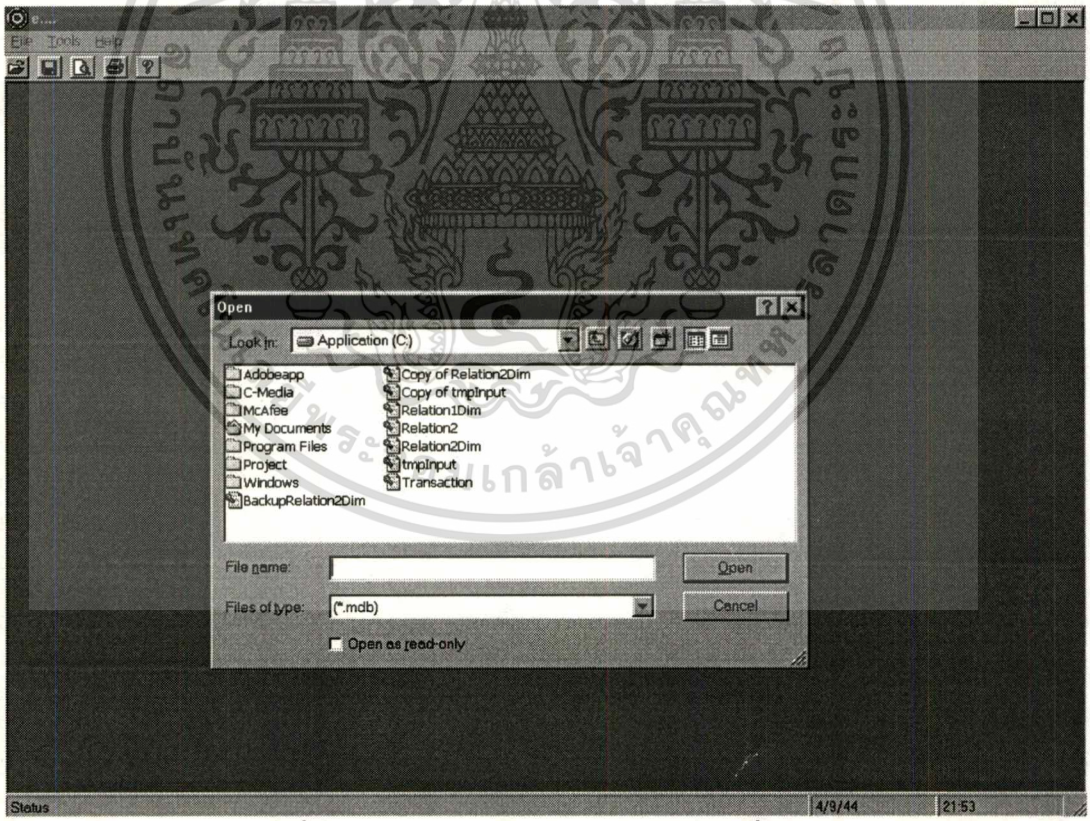


ภาพที่ ก.8 หน้าจอหลักสำหรับเลือกการติดต่อกับข้อมูลที่เป็นฐานข้อมูล



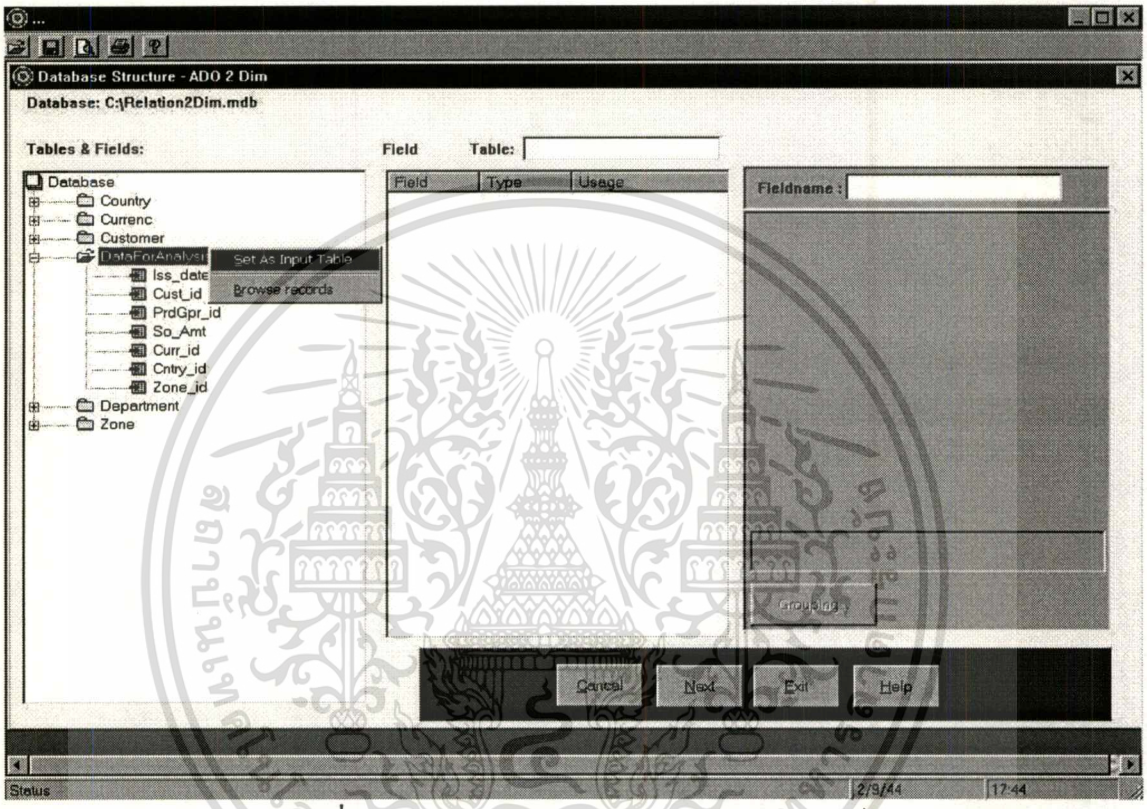
ภาพที่ ก.9 หน้าจอแสดงการเลือกวิธีติดต่อฐานข้อมูล

หลังจากเลือกวิธีติดต่อกับฐานข้อมูลแล้วจะปรากฏหน้าจอให้เลือกชื่อฐานข้อมูลที่ติดต่อ ดังภาพที่ ก.10



ภาพที่ ก.10 หน้าจอแสดงการเลือกฐานข้อมูลที่ติดต่อ

หลังจากนั้นระบบจะแสดงรายละเอียดของตารางและแอทริบิวของตารางมาให้เลือก ว่าต้องการให้ตารางไหนเป็นตารางหลักในการวิเคราะห์ โดยการคลิกเมาส์ขวาที่ชื่อตาราง จะปรากฏเมนูย่อย 2 เมนูคือ Set As Input Table เพื่อกำหนดให้ตารางนั้นเป็นตารางในการวิเคราะห์ และ Browse Records เพื่อดูข้อมูลในตารางนั้น ๆ ดังตัวอย่างหน้าจอภาพที่ ก.11

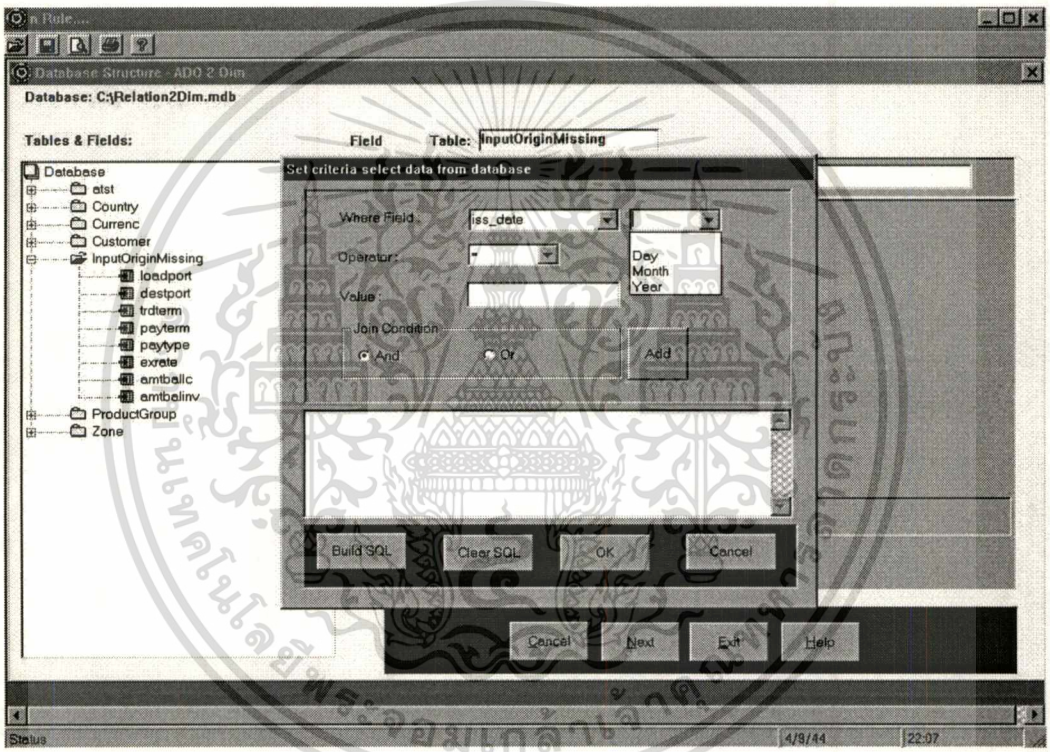


ภาพที่ ก.11 หน้าจอแสดงการเลือกตารางมาวิเคราะห์

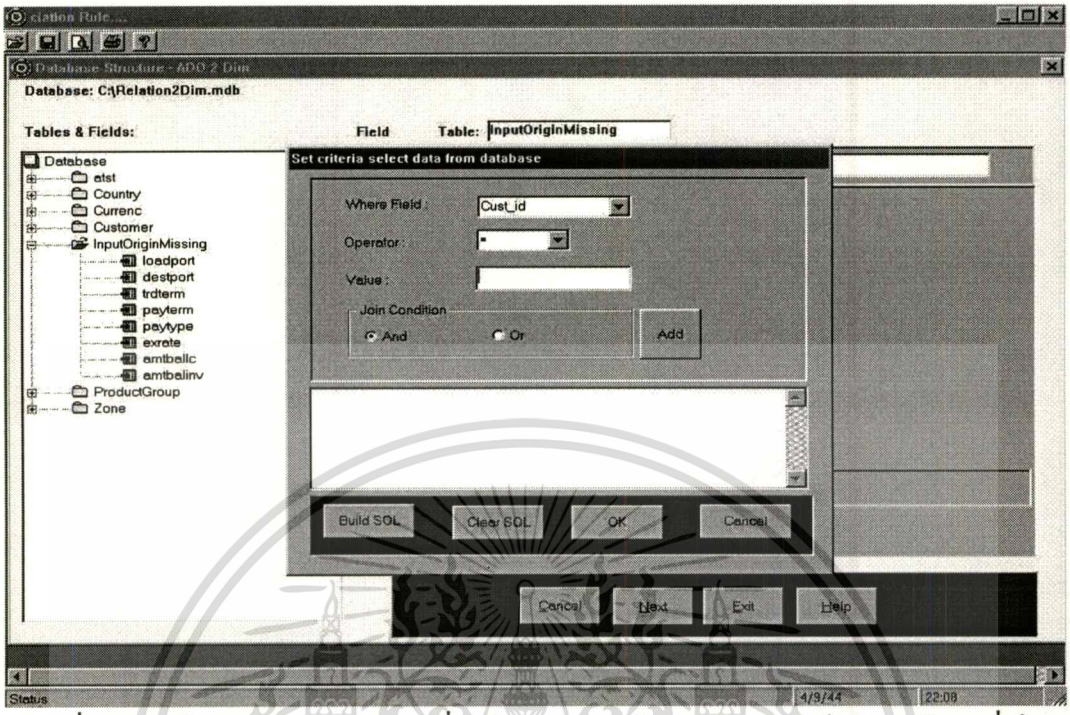
จากนั้นจะปรากฏหน้าจอดังภาพที่ ก.12 เพื่อกำหนดเงื่อนไขในการนำข้อมูลเข้า

- Where Field คือชื่อแอททริบิวที่มีอยู่ในตารางนั้น ๆ ถ้าแอททริบิวนั้นมีชนิดของข้อมูลเป็นวันที่ (Date/Time) จะปรากฏช่องของการเงื่อนไขพิเศษขึ้นมา คือมีการเลือกตาม Day, Month หรือ Year ดังภาพที่ ก.13
- Operator คือ โอเปอเรเตอร์ที่กระทำกับแอททริบิวนั้น ๆ ซึ่งประกอบด้วยค่า =, <=, >=,<, >, <> , LIKE, NULL , NOT NULL
- Value คือค่าของเงื่อนไขที่ต้องการ
- Join Condition ประกอบด้วย 2 ปุ่มที่เป็นออปชั่นให้เลือก คือ AND และ OR
- ปุ่ม ADD คือให้เงื่อนไขที่เลือกนี้เป็นเงื่อนไขหนึ่งในการเลือกข้อมูล

- ปุ่ม Build SQL คือเรียกดูประโยค SQL ที่ระบบสร้างให้
- ปุ่ม Clear SQL คือยกเลิกการสร้างเงื่อนไขที่กำหนดไว้
- ปุ่ม OK คือให้ระบบทำการเลือกข้อมูลจากตารางตามเงื่อนไขที่ระบุไว้ในประโยค SQL
- ปุ่ม Cancel คือการไม่กำหนดเงื่อนไขใด ๆ ในการนำข้อมูลเข้า และกลับไปสู่หน้าจอเดิม



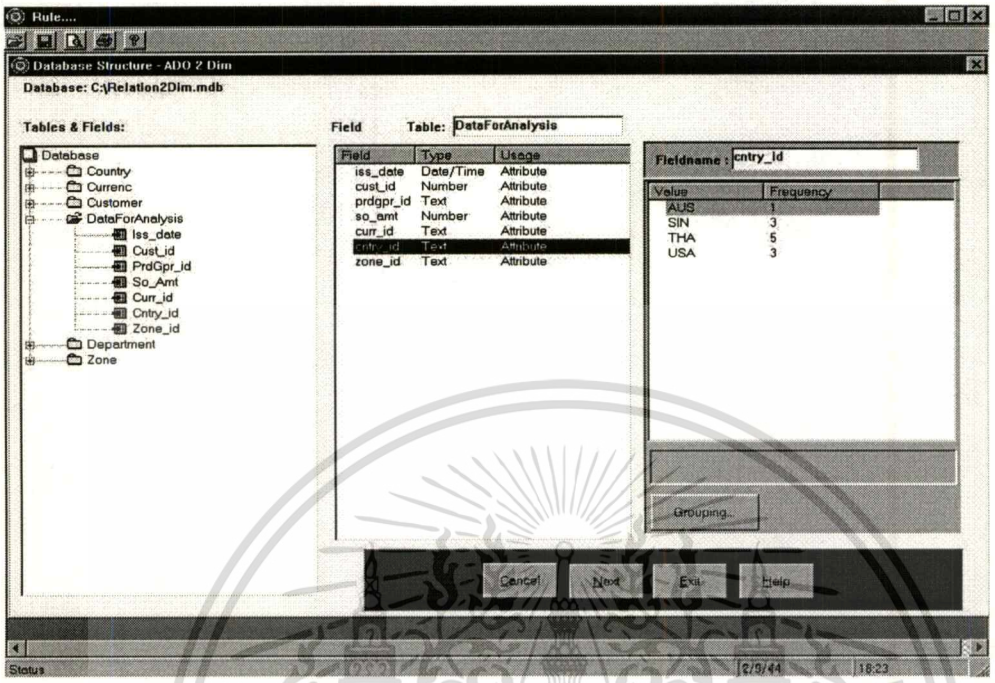
ภาพที่ ก.12 หน้าจอแสดงการกำหนดเงื่อนไขในการนำข้อมูลมาวิเคราะห์



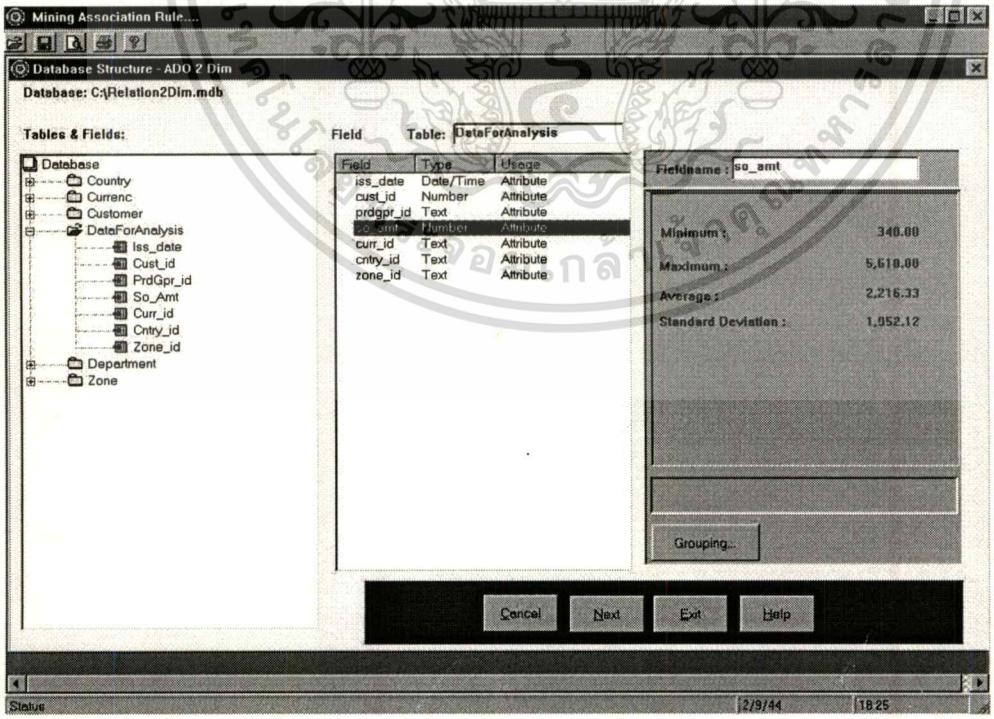
ภาพที่ ก.13 หน้าจอแสดงการกำหนดเงื่อนไขในการนำข้อมูลมาวิเคราะห์สำหรับข้อมูลที่เป็นวันที่

จากนั้น ทำการเลือกแอททริบิวที่จะนำมาวิเคราะห์ โดยการคลิกขวาที่ชื่อแอททริบิว จะปรากฏเมนูย่อย 2 เมนู คือ Set as Attribute และ Set as Excluded ถ้าเป็นการเลือกวิเคราะห์ข้อมูลที่เป็น Relational Data แต่ถ้าเลือกวิเคราะห์ข้อมูลที่เป็น Transaction Data แล้วจะปรากฏเมนูย่อยเป็น Set As Transaction ID และ Set As Item ID

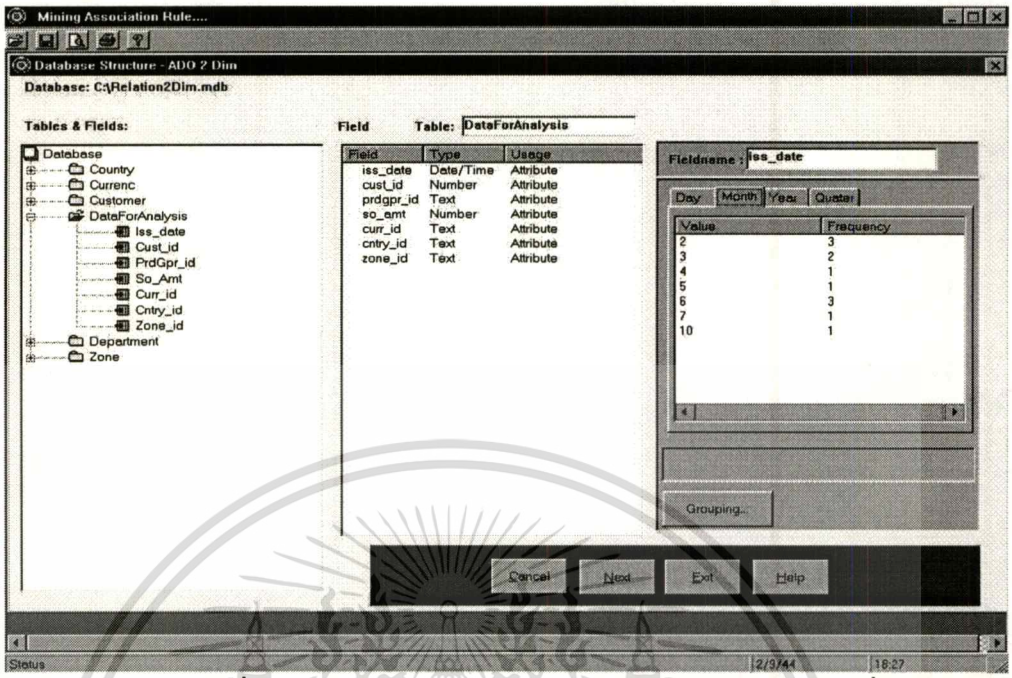
ระบบจะแสดงรายละเอียดของข้อมูลในแต่ละแอททริบิว กรณีที่แอททริบิวใดมีข้อมูลที่หายไป (Missing Value) ระบบจะแสดงข้อความเตือนที่มุมล่างของหน้าจอ พร้อมกับมีปุ่มคำสั่งให้จัดการกับค่าว่างนั้น โดยถ้าเป็นแอททริบิวที่มีชนิดของข้อมูลเป็นข้อความ (Tex) จะแสดงค่าของข้อมูลและค่าความถี่ที่เกิดขึ้นของข้อมูลในแอททริบิวนั้นๆ แสดงตัวอย่างหน้าจอดังภาพที่ ก.14 แอททริบิวที่มีชนิดข้อมูลเป็นตัวเลข(Number) จะแสดงค่าของสูงสุด, ต่ำสุด และค่าเฉลี่ยของข้อมูล แสดงตัวอย่างหน้าจอดังภาพที่ ก.15 และแอททริบิวที่มีชนิดข้อมูลเป็นวันที่(Date/Time) จะแสดงค่าของข้อมูลได้ตามปี, เดือน, วัน และไตรมาส แสดงตัวอย่างหน้าจอดังภาพที่ ก.16



ภาพที่ ก.14 หน้าจอแสดงรายละเอียดของแอททริบิวชนิดข้อความ

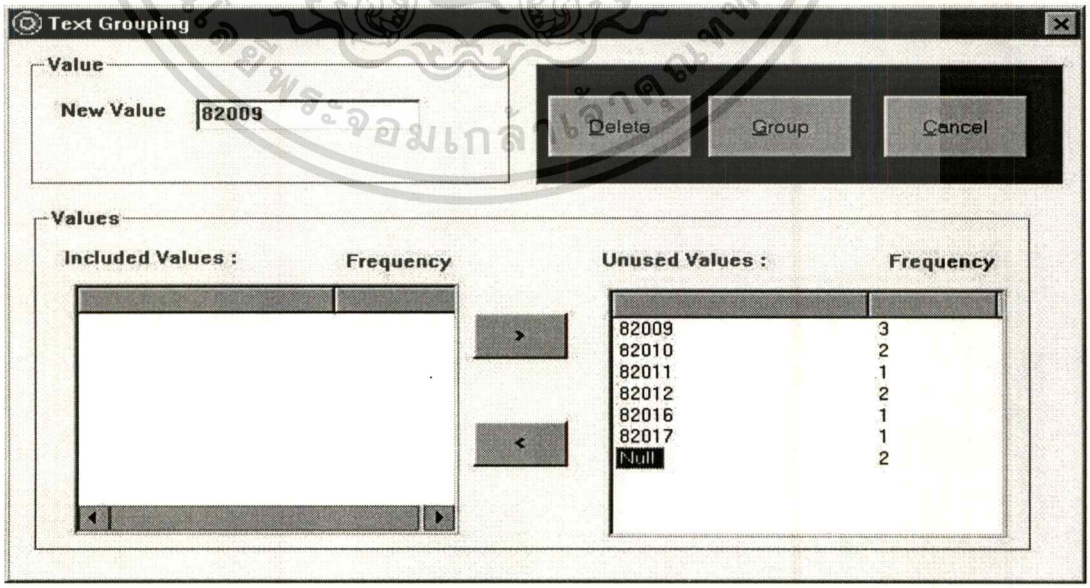


ภาพที่ ก.15 หน้าจอแสดงรายละเอียดของแอททริบิวชนิดตัวเลข



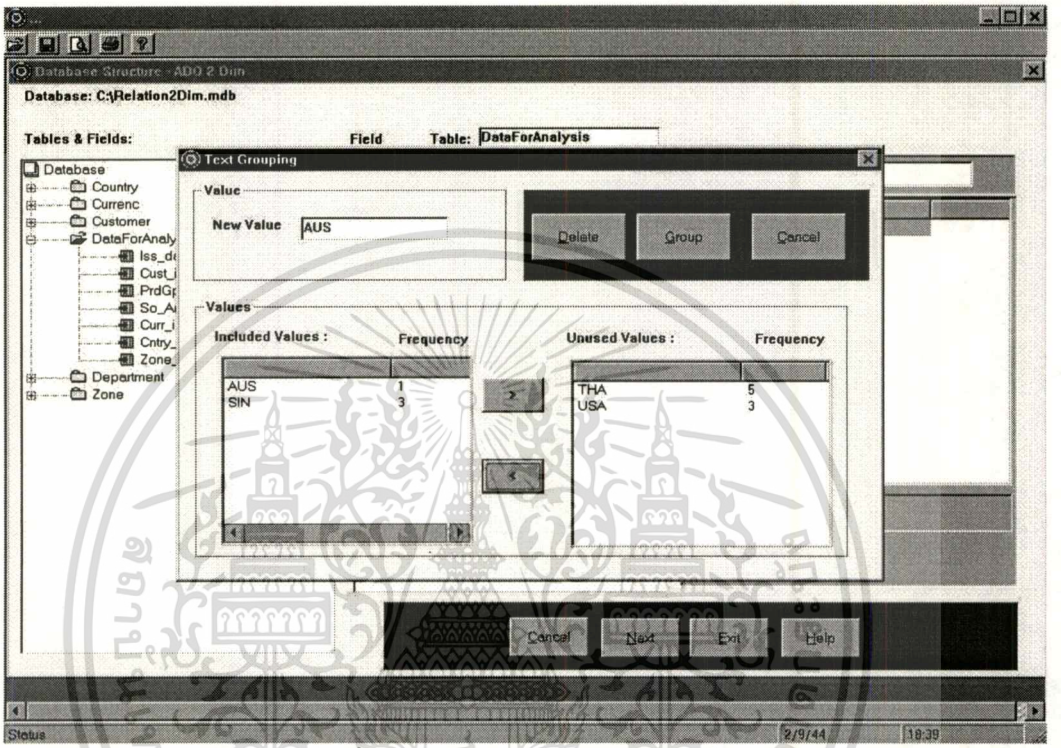
ภาพที่ ก.16 หน้าจอแสดงรายละเอียดของแอททริบิวชนิดวันที่

กรณีที่ต้องการจัดการกับข้อมูลที่ขาดหายไป คลิกที่ปุ่ม Missing Value... จะปรากฏหน้าจอดังภาพที่ ก.17 เพื่อให้จัดการกับค่าของข้อมูลที่หายไป โดยสามารถที่จะลบเรคอร์ดที่ประกอบด้วยค่าว่างนี้ โดยคลิกที่ปุ่ม Delete หรือจะจัดกลุ่มรวมกับค่าอื่น หรือจะแทนด้วยค่าใหม่ โดยคลิกที่ปุ่ม Group



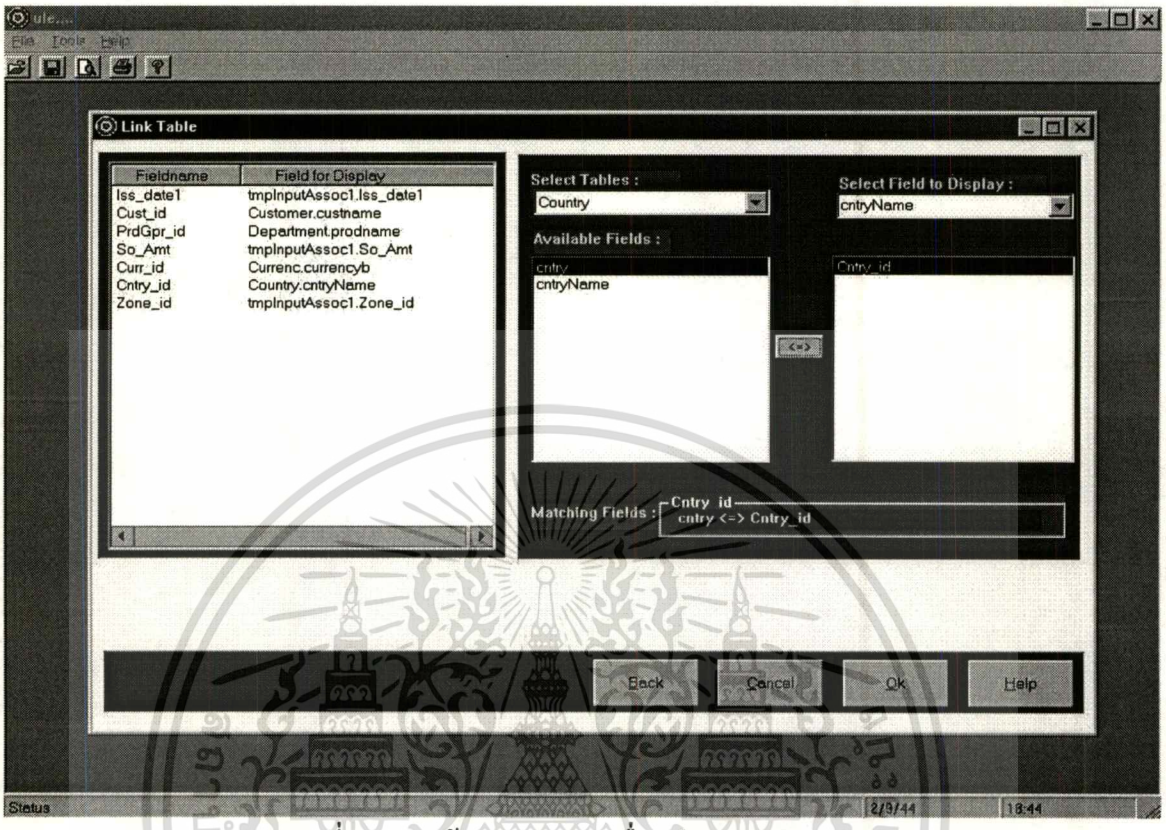
ภาพที่ ก.17 หน้าจอแสดงการจัดการกับข้อมูลที่มีค่าที่หายไป

กรณีที่ต้องการจัดกลุ่มสำหรับข้อมูลที่มีค่าความถี่ที่เกิดขึ้นน้อย สามารถทำได้โดยคลิกที่ปุ่ม Grouping... จะปรากฏหน้าจอดังภาพที่ ก.18

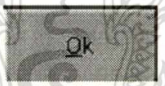


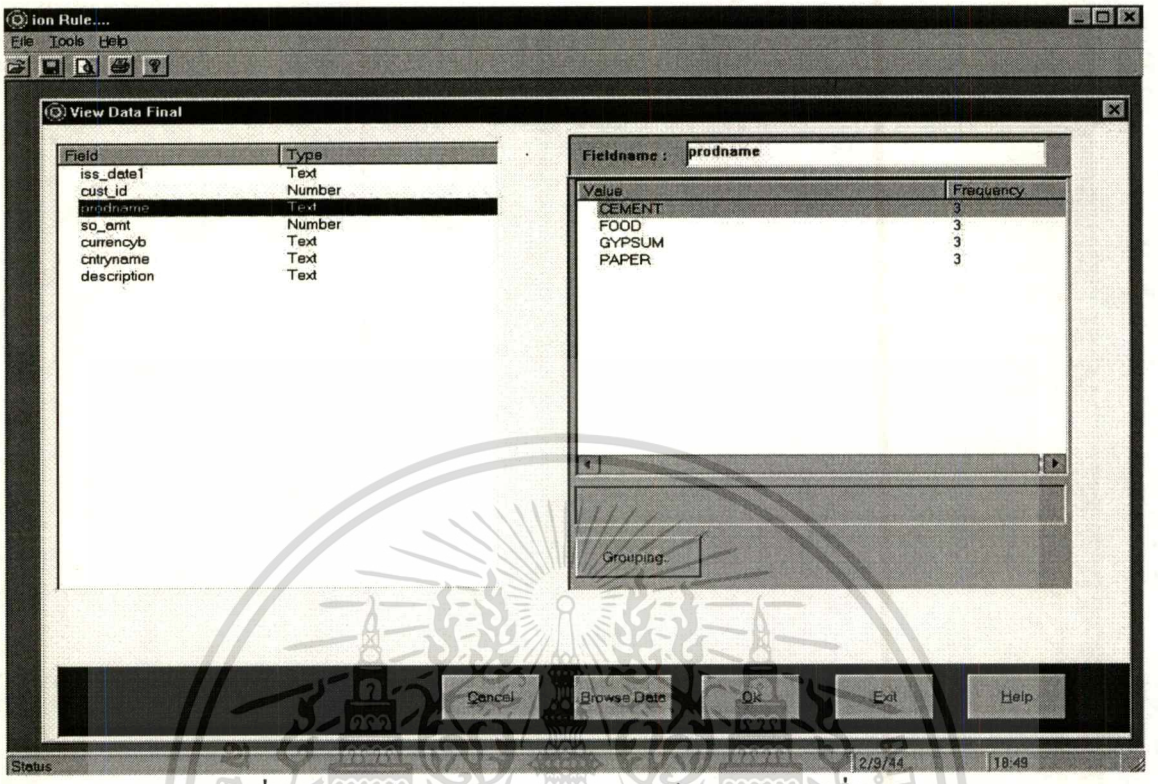
ภาพที่ ก.18 หน้าจอแสดงการจัดกลุ่มข้อมูล

หลังจากที่จัดการกับค่าข้อมูลที่หายไปและจัดกลุ่มข้อมูลเสร็จแล้ว สามารถเชื่อมต่อตาราง (Join) กับตารางอื่นได้ เพื่อเลือกใช้แอททริบิวจากตารางอื่นมาใช้ในการวิเคราะห์แทน แสดงตัวอย่างหน้าจอดังภาพที่ ก.19



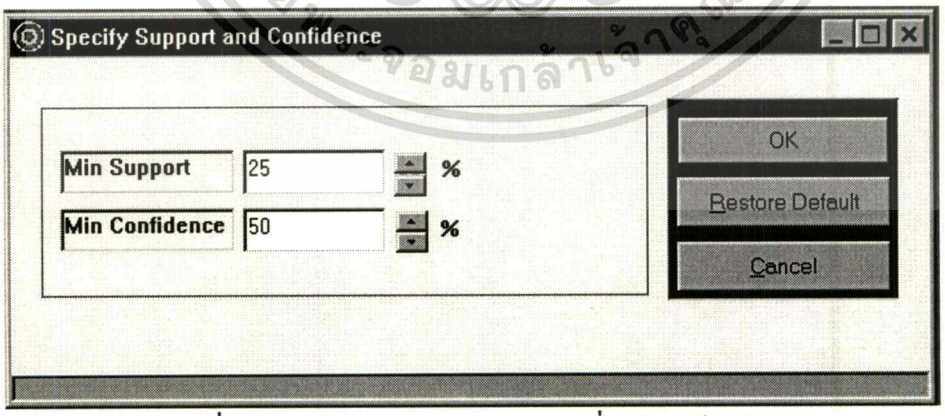
ภาพที่ ก.19 หน้าจอแสดงการเชื่อมต่อตาราง

หลังจากเชื่อมตารางแล้วคลิกที่ปุ่ม  จะปรากฏหน้าจอให้ตรวจสอบข้อมูลอีกครั้งหนึ่ง ถ้าการเชื่อมตารางมีค่าของข้อมูลที่หายไปสามารถจัดการกับค่าว่างได้ หรือจัดกลุ่มข้อมูลได้ใหม่ แสดงตัวอย่างของหน้าจอดังภาพที่ ก.20



ภาพที่ ก.20 หน้าจอแสดงการตรวจสอบข้อมูลหลังจากเชื่อมต่อตาราง

หลังจากที่ทำการการตรวจสอบคุณภาพของข้อมูลเสร็จแล้วจะเข้าสู่ขั้นตอนของการกำหนดค่า Minimum Support และ Minimum Confidence เพื่อกำหนดเงื่อนไขในการสร้างกฎ ดังภาพที่ ก.21



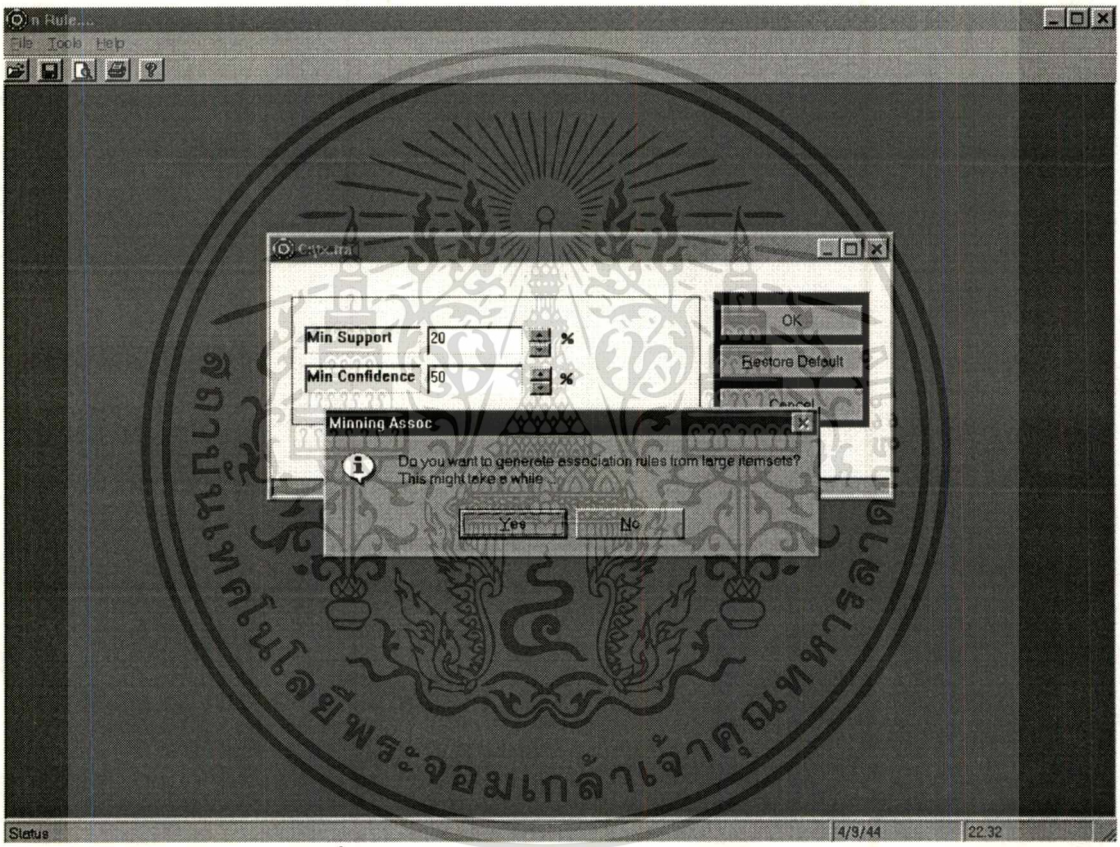
ภาพที่ ก.21 หน้าจอแสดงการกำหนดเงื่อนไขให้กับโปรแกรม

จากภาพที่ ก.21

ปุ่ม Restore Default เป็นการ Reset ค่า Minimum Support และ Minimum Confidence ให้เป็นตามค่าตั้งต้นของโปรแกรม

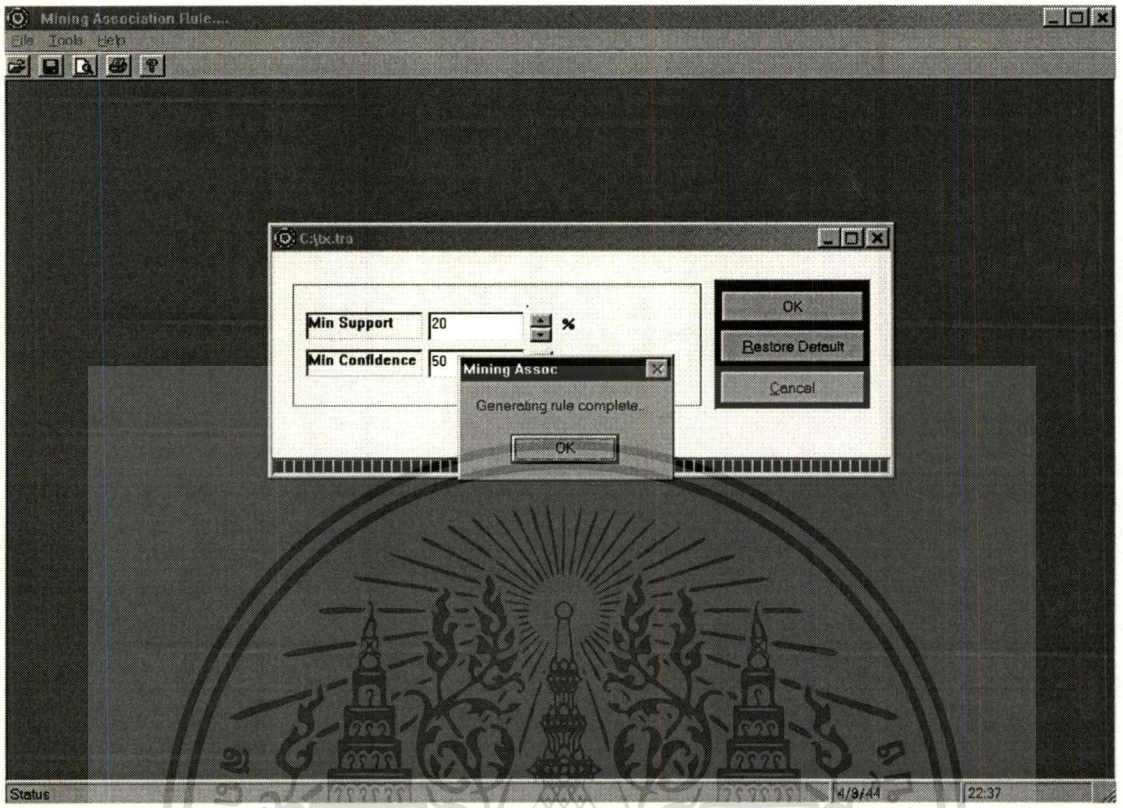
ปุ่ม Cancel คือยกเลิกการสร้างกฎ

ปุ่ม OK คือให้ระบบทำการสร้างกฎให้ ซึ่งจะปรากฏหน้าจอให้ยืนยันการสร้างกฎ ดังภาพที่ ก.22

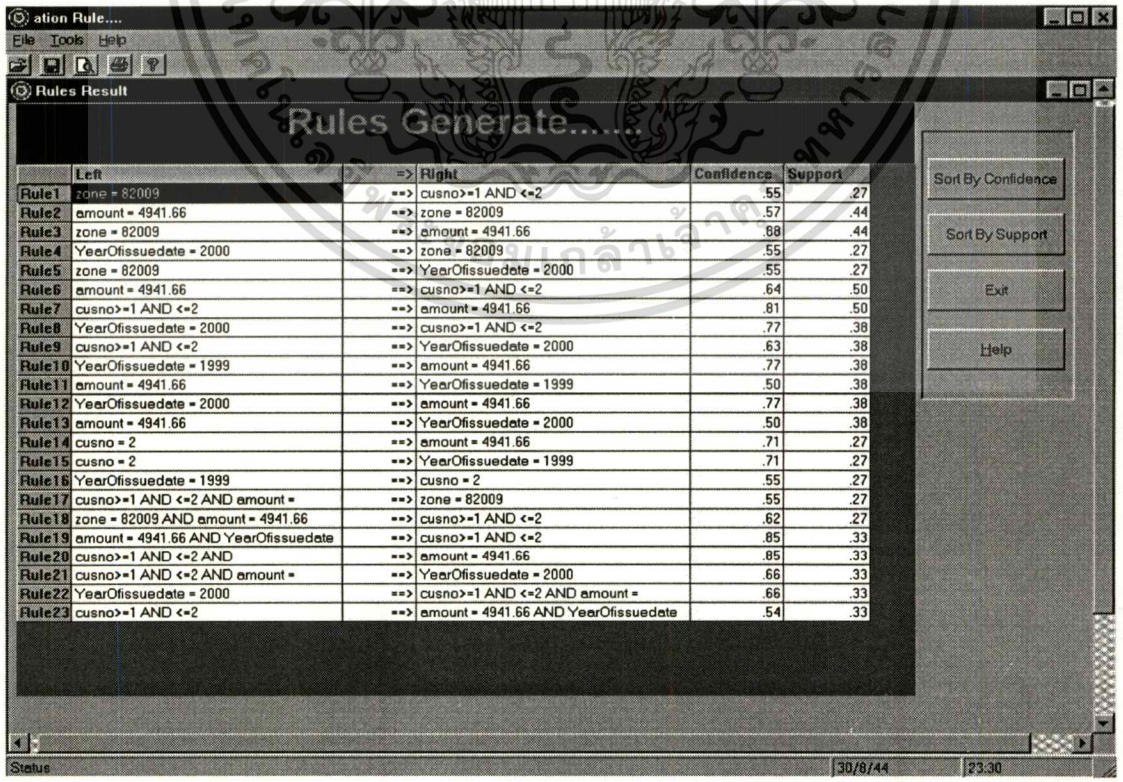


ภาพที่ ก.22 หน้าจอแสดงการยืนยันการสร้างกฎ

จากภาพที่ ก.22 คลิกที่ปุ่ม Yes เพื่อให้ระบบทำการวิเคราะห์ข้อมูล เมื่อระบบทำงานสำเร็จ จะปรากฏข้อความแสดง ดังภาพที่ 23 จากนั้น เมื่อคลิกปุ่ม OK จะปรากฏหน้าจอแสดงผลลัพธ์จากการสร้างกฎดังภาพที่ ก.24



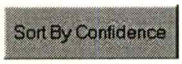
ภาพที่ ก.23 หน้าจอแสดงข้อความแสดงการสร้างกฎสำเร็จ



	Left	=>	Right	Confidence	Support
Rule1	zone = 82009	==>	cusno>=1 AND <=2	.55	.27
Rule2	amount = 4941.66	==>	zone = 82009	.57	.44
Rule3	zone = 82009	==>	amount = 4941.66	.88	.44
Rule4	YearOfissuedate = 2000	==>	zone = 82009	.55	.27
Rule5	zone = 82009	==>	YearOfissuedate = 2000	.55	.27
Rule6	amount = 4941.66	==>	cusno>=1 AND <=2	.64	.50
Rule7	cusno>=1 AND <=2	==>	amount = 4941.66	.81	.50
Rule8	YearOfissuedate = 2000	==>	cusno>=1 AND <=2	.77	.38
Rule9	cusno>=1 AND <=2	==>	YearOfissuedate = 2000	.63	.38
Rule10	YearOfissuedate = 1999	==>	amount = 4941.66	.77	.38
Rule11	amount = 4941.66	==>	YearOfissuedate = 1999	.50	.38
Rule12	YearOfissuedate = 2000	==>	amount = 4941.66	.77	.38
Rule13	amount = 4941.66	==>	YearOfissuedate = 2000	.50	.38
Rule14	cusno = 2	==>	amount = 4941.66	.71	.27
Rule15	cusno = 2	==>	YearOfissuedate = 1999	.71	.27
Rule16	YearOfissuedate = 1999	==>	cusno = 2	.55	.27
Rule17	cusno>=1 AND <=2 AND amount =	==>	zone = 82009	.55	.27
Rule18	zone = 82009 AND amount = 4941.66	==>	cusno>=1 AND <=2	.62	.27
Rule19	amount = 4941.66 AND YearOfissuedate	==>	cusno>=1 AND <=2	.85	.33
Rule20	cusno>=1 AND <=2 AND	==>	amount = 4941.66	.85	.33
Rule21	cusno>=1 AND <=2 AND amount =	==>	YearOfissuedate = 2000	.66	.33
Rule22	YearOfissuedate = 2000	==>	cusno>=1 AND <=2 AND amount =	.66	.33
Rule23	cusno>=1 AND <=2	==>	amount = 4941.66 AND YearOfissuedate	.54	.33

ภาพที่ ก.24 หน้าจอแสดงผลลัพธ์การสร้างกฎ

จากภาพที่ ก.24 มีปุ่มคำสั่ง 4 ปุ่ม คือ



เพื่อให้ระบบเรียงค่าที่แสดงตาม Confidence จากมากไปน้อย

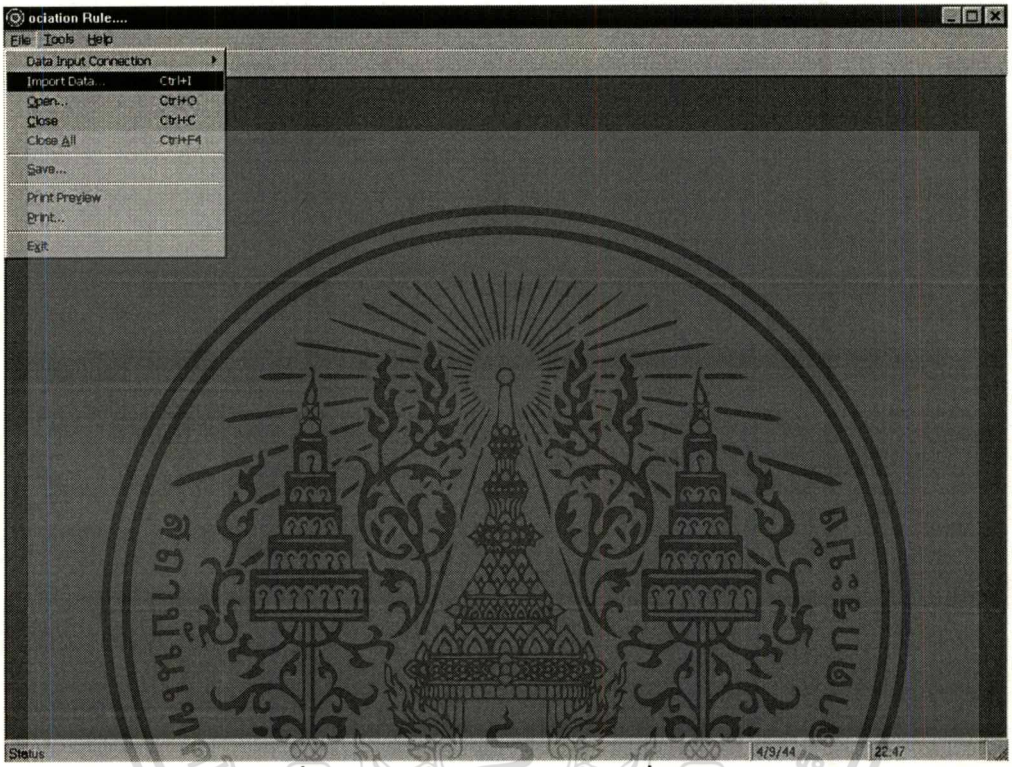
เพื่อให้ระบบเรียงค่าที่แสดงตาม Support จากมากไปน้อย

เพื่อออกจากหน้าจอ

เพื่อเรียกดูความช่วยเหลือจากระบบ

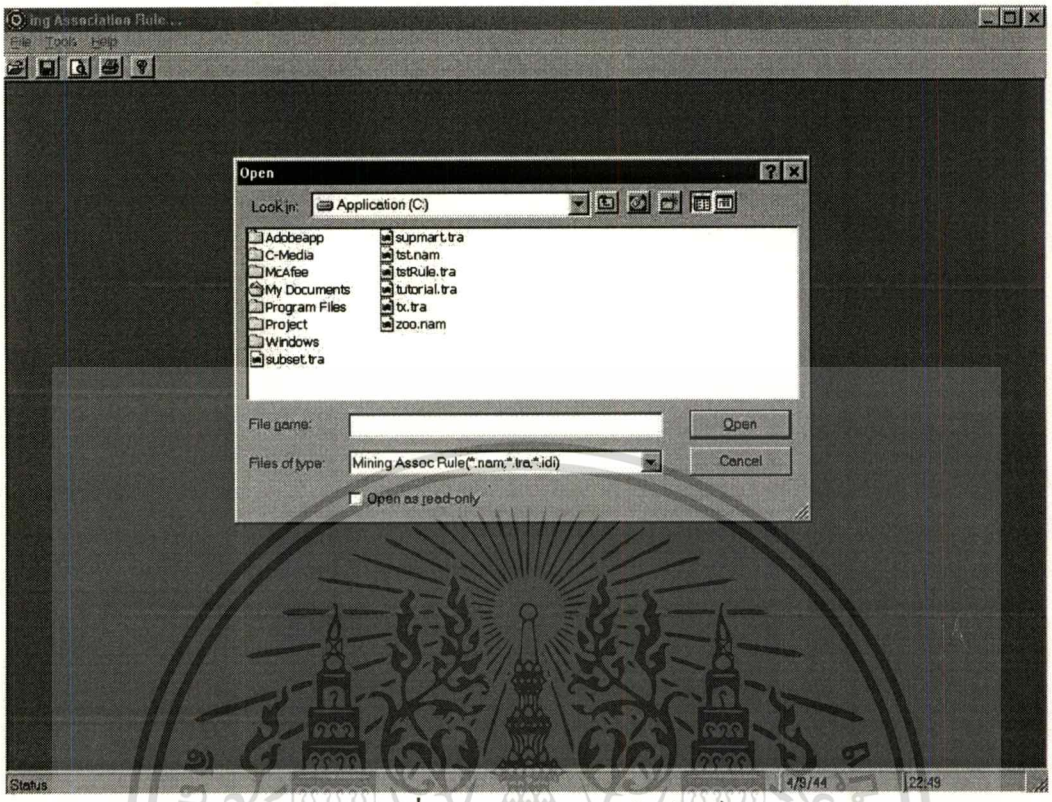
ก.4 การวิเคราะห์ข้อมูลจากเท็กซ์ไฟล์

การติดต่อกับข้อมูลที่เป็นเท็กซ์ไฟล์ ทำได้โดยคลิกที่เมนู Import Data จากเมนูหลัก ดังภาพที่ ก.25



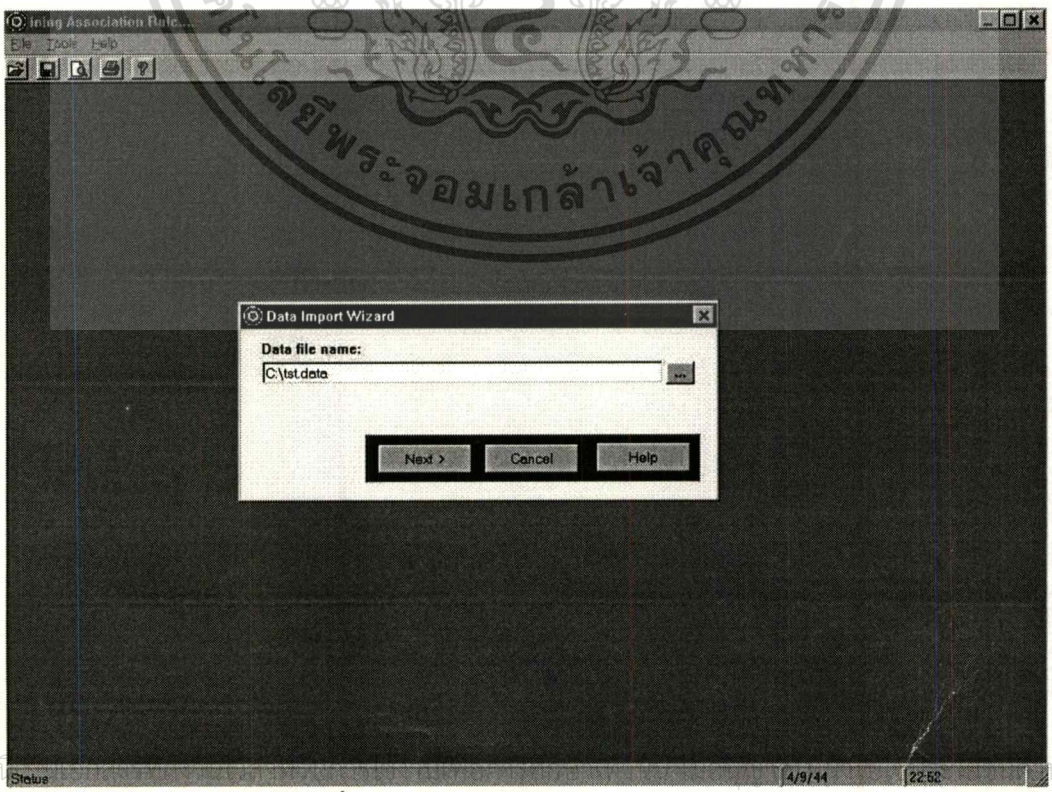
ภาพที่ ก.25 เมนูการติดต่อกับข้อมูลที่เป็นเท็กซ์ไฟล์

จากนั้น จะปรากฏหน้าจอดังภาพที่ ก.26 เพื่อให้เลือกชื่อไฟล์ที่ต้องการติดต่อ โดยระบบจะกำหนดให้เลือกได้เฉพาะไฟล์ที่นามสกุล .tra และ .nam



ภาพที่ ก.26 หน้าจอการเลือกไฟล์

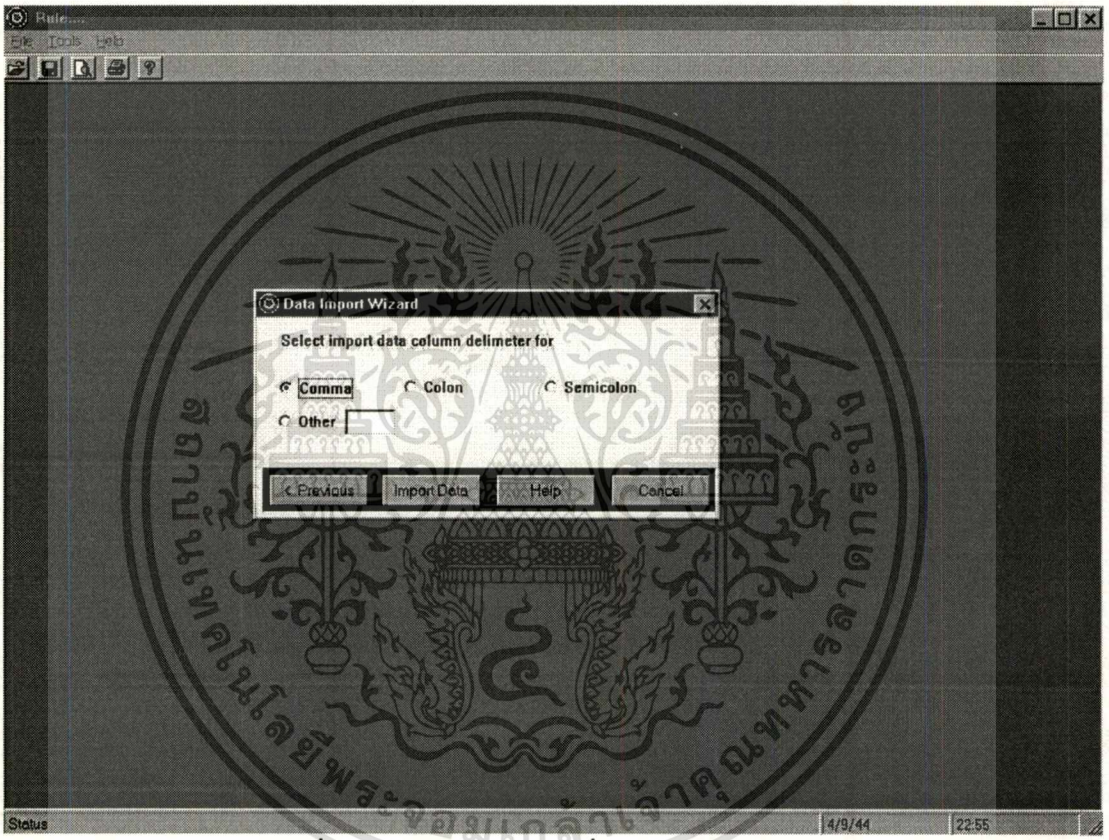
- กรณีที่เลือกวิเคราะห์ข้อมูลที่เป็น Relational Data หลังจากที่เลือกไฟล์นามสกุล .nam แล้ว จะปรากฏหน้าจอดังภาพที่ ก.27 เพื่อให้เลือกชื่อไฟล์ที่เก็บข้อมูลที่มีนามสกุล .dat



ภาพที่ ก.27 หน้าจอการเลือกไฟล์ .dat

จากภาพที่ ก.27 ระบบจะแสดงชื่อไฟล์และโฟลเดอร์ที่จัดเก็บไฟล์เป็นโฟลเดอร์เดียวกับ .nam file ให้โดยอัตโนมัติ ถ้าไฟล์ .dat ไม่ได้เก็บไว้ที่เดียวกัน สามารถคลิกที่ปุ่ม ... เพื่อเลือกหาไฟล์ได้

จากนั้นจะปรากฏหน้าจอดังภาพที่ ก.28 เพื่อให้ผู้ใช้เลือกเงื่อนไขในการดึงข้อมูลเท็กซ์ไฟล์เข้าว่าข้อมูลที่พิมพ์เข้ามาแต่ละค่า คั่นด้วยเครื่องหมายอะไร



ภาพที่ ก.28 หน้าจอการเลือกเงื่อนไขการดึงข้อมูลเข้า

จากนั้น จะปรากฏหน้าจอดังภาพที่ ก.21 เพื่อให้ระบุค่า Minimum Support และ Minimum Confidence คลิกที่ปุ่ม OK เพื่อให้ระบบสร้างกฎ และเมื่อสร้างกฎสำเร็จ จะปรากฏหน้าจอคล้ายกับภาพที่ ก.24

ประวัติผู้เขียน

ชื่อ : นางสาวรุจิรา ใหม่จันทร์

วันเกิด : 2 กรกฎาคม พ.ศ. 2517

มัธยมศึกษา : โรงเรียนศรีสวัสดิ์วิทยาคาร จ. น่าน

ปริญญาตรี : มหาวิทยาลัยเชียงใหม่ คณะวิทยาศาสตร์ สาขาวิทยาการคอมพิวเตอร์ ปีการศึกษา 2539

สถานที่ทำงาน : บริษัท ไอทีวัน จำกัด

ตำแหน่ง : Business/Technical Analyst

E-mail Address: rujiram@hotmail.com