การพัฒนาระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

DEVELOPMENT OF DATA MINING SYSTEM BY QUANTITATIVE ASSOCIATION RULES

โดย

ช่อผกา มงคลสถิตย์พร

CHORPAKA MONGCOLSATITPORN

อาจารย์ที่ปรึกษา

รศ. ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2550

DEVELOPMENT OF DATA MINING SYSTEM BY QUANTITATIVE ASSOCIATION RULES

CHORPAKA MONGCOLSATITPORN

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECHNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2007

หัวข้อ	การพัฒนาระบบคาด้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ
นักศึกษา	นางสาว ช่อผกา มงคลสถิตย์พร
รหัสนักศึกษา	48066808
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2550
อาจารย์ที่ปรึกษา	รศ. ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

การดำเนินธุรกิจในปัจจุบันมีการแข่งขันกันอย่างรุนแรง เพื่อให้เกิดความได้เปรียบทางธุรกิจ จึงได้มีการนำคาด้าไมนิ่งมาช่วยในการวิเคราะห์ข้อมูล เพื่อช่วยในการตัดสินใจทางธุรกิจ โดยข้อมูลที่มีในฐานข้อมูลนั้นมีข้อมูลเชิงปริมาณและข้อมูลเชิงประเภท จึงเหมาะสมที่จะนำเทคนิคการค้นหาความสัมพันธ์ของข้อมูลเชิงปริมาณมาใช้ จะช่วยทำให้เห็นความสัมพันธ์ของข้อมูลที่ซ่อนอยู่ได้ จะส่งผลดีต่องานด้านต่าง ๆ โดยเฉพาะงานด้านการตลาด เป็นเทคนิคหนึ่งที่ช่วยส่งเสริมในด้านการค้นหาว่าสินค้าหรือบริการใดที่มีการซื้อร่วมกัน ซึ่งมีส่วนช่วยเหลือในการพัฒนาแผนการตลาดและการวางกลยุทธ์ขององค์กร ขั้นตอนพัฒนาการทำงานระบบ เริ่มจากการรับข้อมูลเข้าในรูปแบบฐานข้อมูล SQL หรือ ไฟล์ CSV หลังจากนั้นจะทำการเตรียมข้อมูลให้พร้อม และนำอัลกอริทึมที่พัฒนามาจาก Apriori มาใช้ในการหากฎความสัมพันธ์ โดยจะแสดงผลลัพธ์มาในลักษณะรายงาน สิ่งสำคัญของระบบคือความถูกต้องของผลลัพธ์ ซึ่งขึ้นอยู่กับจำนวนข้อมูลที่ใช้สร้างโมเดล การเก็บข้อมูลที่จำเป็นอย่าง ถูกต้อง ครบถ้วน และการเลือกเทคนิคการทำเหมืองข้อมูลที่ตรงกับเป้าหมาย

Title	Development of Data Mining System by Quantitative Association Rules
Student	Ms. Chorpaka Mongcolsatitporn
Student ID.	48066808
Degree	Master of Science
Programme	Information Science
Academic Year	2007
Advisor	Assoc. Prof. Dr. Worapoj Kreesuradej

ABSTRACT

Nowadays in a business word have many competition for get much profit, that is propose to analyze data by data mining technique. Mostly data in database have quantitative and categorical data is profit to using quantitative association rules technique. It very helpful especially marketing to finding buying together association between product or service for increase marketing plan and strategic plan performance. In a system development step start at receive input data in format database or CSV file, next will prepare data are analyzed by extended Apriori algorithm then get association rules and show with report. The principal point is accuracy result depends on amount of training data, data collection and the exact data mining technique.

กิตติกรรมประกาศ

ผู้จัดทำขอขอบพระคุณ รศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาของโครงงานพัฒนาระบบงาน ที่กรุณาให้ความรู้และคำแนะนำอันเป็นประโยชน์อย่างมาก ต่อการพัฒนาโครงงานนี้ตลอดจนตรวจสอบแก้ไขจนกระทั่งโครงงานสำเร็จลุล่วง

ขอขอบคุณเพื่อนๆ และพี่ๆ สาขาวิชาเทคโนโลยีสารสนเทศ ที่ได้ให้คำแนะนำช่วยเหลือผู้จัดทำ ในเรื่องแนวทางการเขียนและแก้ไขโปรแกรมให้สามารถทำงานได้ประสบผลสำเร็จ

ขอขอบคุณเพื่อน ๆ พี่ ๆ ที่ร่วมงานกันในบริษัทจีเอเบิลและ บริษัท แอดวานซ์ อินโฟเซอร์วิส มหาชน จำกัด ที่เป็นกำลังใจ ให้คำปรึกษา และอำนวยความสะดวกตลอดระยะเวลาการศึกษา และการทำงานโครงงานพัฒนาระบบ

สุดท้ายนี้ผู้จัดทำขอกราบขอบพระคุณ บิดา มารดา ญาติพี่น้อง และคุณภาณุวัฒน์ ที่เป็นกำลังใจ และให้การสนับสนุนในทุกเรื่อง ทำให้สามารถทำโครงงานสำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์อันพึงมาจากโครงงานพัฒนาระบบงานฉบับนี้ ผู้จัดทำขอมอบแด่ผู้มีพระคุณทุกท่าน

ช่อผกา มงคลสถิตย์พร

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VI

สารบัญรูป VII

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ในการพัฒนาระบบ 1

1.3 ขอบเขตของการพัฒนาระบบ 2

1.4 ความต้องการในการพัฒนาระบบ 2

1.5 ขั้นตอนการดำเนินงาน 2

1.6 ประโยชน์ที่คาดว่าจะได้รับ 3

บทที่ 2 หลักการและทฤษฎีที่เกี่ยวข้อง 4

2.1 ดาต้าไมนิ่ง 4

2.1.1 นิยาม 4

2.1.2 ปัจจัยที่ทำให้ดาต้าไมนิ่งได้รับการสนใจ 4

2.1.3 ลักษณะของงานดาต้าไมนิ่ง 5

2.1.4 ขั้นตอนการทำดาต้าไมนิ่ง 5

2.2 การค้นหากฎความสัมพันธ์จากข้อมูล 8

2.2.1 กฎความสัมพันธ์พื้นฐาน (Association Rules Discovery) 8

2.2.2 การค้นหากฎความสัมพันธ์ของข้อมูลสำหรับข้อมูลเชิงปริมาณ (Quantitative Association Rules Discovery) 9

2.2.3 ขั้นตอนการค้นหากฎความสัมพันธ์ของข้อมูลเชิงปริมาณ 9

บทที่ 3 การวิเคราะห์และออกแบบระบบ 21

3.1 รายละเอียดที่เกี่ยวข้องกับระบบ 21

สารบัญ (ต่อ)

หน้า

3.2 คอนเท็กซ์ไดอะแกรม..... 21
3.3 แผนภาพกระแสข้อมูลระดับที่ 1..... 22
3.4 แผนภาพกระแสข้อมูลระดับที่ 2..... 24
3.4 อีอาร์ไดอะแกรม..... 27

บทที่ 4 การพัฒนาระบบ.....30
4.1 การจัดเตรียมสภาพแวดล้อมในการพัฒนา..... 30
4.2 ขั้นตอนการพัฒนาระบบ.....30
4.3 การพัฒนาระบบและหน้าจอการทำงาน..... 30

บทที่ 5 บทสรุป..... 40
5.1 ผลการวิเคราะห์และออกแบบระบบการทำคาต้าไมนิ่งแบบ Quantitative Association Rules.....40
5.2 ประโยชน์ที่คาดว่าจะได้รับ.....40
5.3 แนวทางการพัฒนาระบบเพิ่มเติม..... 41

บรรณานุกรม..... 42

ประวัติผู้เขียน..... 43

สารบัญตาราง

ตารางที่ หน้า

2.1 ตัวอย่างข้อมูลรายการ(เปลี่ยนแปลง)....8

2.2 ตัวอย่างตาราง People....9

2.3 ตัวอย่างผลการหากฎความสัมพันธ์ของข้อมูลเชิงปริมาณ....10

2.4 แสดงการแบ่งช่วงอายุ....11

2.5 แสดงตาราง People หลังจากมีการแบ่งช่วงอายุแล้ว....11

2.6 แสดงตารางแทนค่าช่วงอายุด้วยตัวเลขที่เรียงลำดับ....11

2.7 แสดงตารางแทนค่าสถานภาพแต่งงาน....12

2.8 แสดงตาราง People หลังจากมีการแทนค่าแล้ว....12

2.9 แสดงจำนวน Support ทั้งหมด....12

2.10 แสดง Itemsets ทั้งหมดที่ทำการรวมได้....13

2.11 แสดง Itemsets ทั้งหมดที่จะนำไปใช้ในอัลกอริทึม....13

2.12 แสดง Itemsets ที่ทำการ Join รอบที่ 1....16

2.13 แสดง Itemsets ทั้งหมดที่เกิดจากการทำงานของอัลกอริทึมในรอบที่ 1....17

2.14 แสดง Itemsets ที่เกิดจากกการ Join ในรอบที่ 2....17

2.15 แสดง Itemsets ทั้งหมดที่เกิดจากการทำงานของอัลกอริทึมในรอบที่ 2....17

2.16 แสดง Frequent Itemsets ทั้งหมด....17

2.17 แสดงกฎความสัมพันธ์บางส่วนที่สร้างขึ้นมา....19

3.1 โครงสร้างของตารางที่ใช้เก็บข้อมูลที่มีลักษณะเชิงปริมาณ....28

3.2 โครงสร้างของตารางที่ใช้เก็บข้อมูลกฎความสัมพันธ์ที่ได้....28

3.3 โครงสร้างของตารางที่ใช้เก็บรายละเอียดของงาน....29

สารบัญรูป

รูปที่ หน้า

2.1 แสดงขั้นตอนการทำดาต้าไมนิ่ง.....5

2.2 แสดงอัลกอริทึม Apriori..... 14

2.3 แสดงอัลกอริทึมสำหรับสร้างกฎความสัมพันธ์..... 18

3.1 คอมเท็กซ์ไดอะแกรมของระบบ.ดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ..... 22

3.2 แผนภาพกระแสข้อมูลระดับที่ 1 ของระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ.....22

3.3 แผนภาพกระแสข้อมูลระดับที่ 2 ของกระบวนการทำดาต้าไมนิ่งของระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ.....25

3.4 เวิร์กโฟลว์ของของกระบวนการทำดาต้าไมนิ่งในระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ.....26

3.5 แสดงอีอาร์ไดอะแกรมของระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ. 26

4.1 แสดงหน้าจอหลักของโปรแกรม..... 31

4.2 หน้าจอแสดงการสร้างงานใหม่จากฐานข้อมูล SQL..... 32

4.3 หน้าจอแสดงการสร้าง Connection String หรือการติดต่อไปยังฐานข้อมูล..... 33

4.4 หน้าจอแสดงการกำหนดส่วนของ Transaction.....34

4.5 การสร้าง Query String..... 35

4.6 หน้าจอแสดงการกำหนดค่าต่าง ๆ สำหรับ Project.....36

4.7 แสดงผลลัพธ์ความสัมพันธ์ของข้อมูลที่ได้..... 37

4.8 แสดงการสร้างงานใหม่จาก CSV..... 38

4.9 ตัวอย่างข้อมูลในไฟล์ CSV..... 39

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันมีการนำระบบสารสนเทศเข้ามาช่วยในการจัดการข้อมูลมากขึ้น ทำให้ข้อมูลต่างๆถูกรวบรวม และจัดเก็บอย่างมีระบบมากกว่าในอดีต ระบบสารสนเทศไม่ได้เพียงแต่มีบทบาทในการดูแล หรือจัดการในการจัดเก็บข้อมูลเท่านั้น การนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์ ก็เป็นสิ่งสำคัญ ในการนำข้อมูลที่มีอยู่มาวิเคราะห์อย่างจริงจัง จะทำให้ผู้วิเคราะห์ได้รับรู้ถึงข้อมูลที่ซ่อนเร้นอยู่ในข้อมูลเหล่านั้น หรือเรียกว่าการเกิดของข้อมูลที่มีรูปแบบ การค้นหารูปแบบเหล่านี้จะทำให้ผู้วิเคราะห์ทราบถึงข้อมูลซึ่งสามารถนำมาใช้ประโยชน์ได้มากมาย ซึ่งทำให้เกิดการพัฒนาในการค้นหารูปแบบของข้อมูลซึ่งสามารถนำมาใช้ประโยชน์ได้ ส่งผลให้เกิดเทคโนโลยีที่เรียกว่า คาต้าไมนิ่ง ซึ่งคาต้าไมนิ่งเป็นกระบวนการค้นหารูปแบบและความสัมพันธ์ที่มีอยู่ในข้อมูลโดยอาศัยขั้นตอนวิธีทางคอมพิวเตอร์และวิธีทางสถิติ ปัจจุบันการทำเหมืองข้อมูลได้ถูกนำมาประยุกต์ใช้ในการบริหารจัดการด้านต่าง ๆ มากมาย เช่น การจัดการความสัมพันธ์กับลูกค้า (CRM) และการจัดการความเสี่ยง (Risk Management) ตัวอย่างการทำเหมืองข้อมูลที่ใช้แพร่หลาย ได้แก่ การวิเคราะห์แบ่งกลุ่มลูกค้า (Customer Segmentation) การค้นหาว่าสินค้าหรือบริการใดที่มีการซื้อร่วมกัน (Market Basket Analysis) เป็นต้น

อีกทั้งข้อมูลส่วนใหญ่ที่มีในฐานข้อมูลนั้นเป็นข้อมูลเชิงปริมาณ เทคนิคการหากฎความสัมพันธ์ของข้อมูลเชิงปริมาณ จึงเป็นเทคนิคหนึ่งที่ช่วยส่งเสริมในด้าน การค้นหาว่าสินค้าหรือบริการใดที่มีการซื้อร่วมกัน ซึ่งมีส่วนช่วยเหลือในการพัฒนาแผนการตลาด และการวางกลยุทธ์ขององค์กร

1.2 วัตถุประสงค์ในการพัฒนาระบบ

1.2.1 เพื่อศึกษาขั้นตอน และวิธีการในการค้นหาความสัมพันธ์สำหรับข้อมูลเชิงปริมาณโดยใช้อัลกอริทึม Apriori

1.2.2 เพื่อสร้างระบบที่ใช้ในการวิเคราะห์ความสัมพันธ์สำหรับข้อมูลเชิงปริมาณ รูปแบบและแนวโน้มของข้อมูล

1.2.3 เพื่อมองหาแนวทางในการพัฒนา Quantitative Association Rules ที่เหมาะสมในอนาคตต่อไป

1.3 ขอบเขตของการพัฒนาระบบ

ระบบที่พัฒนาขึ้นนี้ มีการทำงานครอบคลุมขั้นตอนทั้งหมดของการทำดาต้าไมนิ่ง มีขอบเขตการศึกษา ดังนี้

1.3.1 การพัฒนาระบบเริ่มตั้งแต่การเตรียมข้อมูล จนกระทั่งได้ผลลัพธ์ ซึ่งก็คือกฎความสัมพันธ์ออกมา

1.3.2 ระบบทำงานแบบ Offline คือทำงานได้ภายในเครื่องคอมพิวเตอร์เครื่องเดียวเท่านั้น

1.3.3 ข้อมูลที่นำมาใช้ในการหากฎความสัมพันธ์ สามารถใช้กับข้อมูลทั่วไป ทั้งข้อมูลเชิงปริมาณหรือข้อมูลเชิงประเภท สามารถรับ Input เป็นฐานข้อมูลหรือไฟล์ CSV

1.3.4 ระบบสามารถจัดกลุ่มข้อมูล Input ได้

1.3.5 ระบบมีการจัดเก็บกฎความสัมพันธ์และค่าเงื่อนไขต่างๆ ไว้

1.3.6 ระบบสามารถบันทึกกฎความสัมพันธ์และสามารถเรียกดูในภายหลังได้

1.3.7 สามารถพัฒนาโปรแกรม ตามทฤษฎี และอัลกอริทึม ได้ถูกต้อง

1.3.8 สามารถนำเสนอผลลัพธ์ให้กับผู้ใช้งานนำไปใช้ประโยชน์ได้อย่างถูกต้อง

1.4 ความต้องการในการพัฒนาระบบ

1.4.1 ฮาร์ดแวร์ (Hardware)

- เครื่องคอมพิวเตอร์แพนเทียม 4 2.0 MHz ขึ้นไป
- ฮาร์ดดิสก์ที่มีหน่วยความจำไม่น้อยกว่า 512 Mb
- เม้าส์และคีย์บอร์ด
- ซีดีรอม

1.4.2 ซอฟต์แวร์ (Software)

- ระบบปฏิบัติการวินโดว์ (Window XP)
- Visual Studio .Net 2005 (Visual Basic)
- คริสตัลรีพอร์ท (Crystal Report)
- ไมโครซอฟต์แอคเซส (Microsoft Access)
- ไมโครซอฟต์เอสคิวแอล เซิฟเวอร์ (Microsoft SQL Server 2003)

1.5 ขั้นตอนการดำเนินงาน

การพัฒนาระบบการจัดเก็บค่าธรรมเนียมผ่านทาง มีขั้นตอนการดำเนินงานในการพัฒนาระบบ ดังต่อไปนี้

1.5.1 กำหนดหัวข้อ เป้าหมาย จุดประสงค์ และขอบเขตของการพัฒนาระบบ

1.5.2 ศึกษาทฤษฎีที่เกี่ยวข้อง ได้แก่ ดาต้าไมนิ่ง, ขั้นตอนในการทำดาต้าไมนิ่ง, การค้นหาความสัมพันธ์สำหรับข้อมูลเชิงปริมาณ, อัลกอริธึมที่พัฒนามาจาก Apriori

1.5.3 ออกแบบระบบ

1.5.4 พัฒนาระบบงานเพื่อวิเคราะห์ข้อมูล

1.5.5 ทดสอบระบบ

1.5.6 ปรับปรุง และแก้ไขข้อผิดพลาดที่เกิดขึ้น

1.5.7 สรุปผลการศึกษา

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 เข้าใจขั้นตอนและวิธีในการทำดาต้าไมนิ่ง

1.6.2 ได้ระบบที่ใช้ในการวิเคราะห์ความสัมพันธ์ของข้อมูล เพื่อนำไปประยุกต์ใช้ในธุรกิจต่าง ๆ ได้

บทที่ 2
หลักการและทฤษฎีที่เกี่ยวข้อง

การสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลที่มีขนาดใหญ่ (Knowledge Discovery in Database: KDD) หรือที่เรียกกันว่าดาต้าไมนิ่ง เป็นสาขาหนึ่งที่กำลังได้รับความสนใจอย่างสูงในปัจจุบัน ข้อมูลขนาดใหญ่จะถูกวิเคราะห์และสืบค้นความรู้ที่สำคัญออกมา รวบรวมและจัดเก็บให้อยู่ในรูปฐานความรู้ เพื่อใช้สำหรับการสืบค้นสิ่งที่ต้องการซึ่งไม่สามารถสืบค้นได้จากวิธีการของระบบการจัดการฐานข้อมูลโดยทั่วไป เช่น การวิเคราะห์หาความสัมพันธ์ของข้อมูลหรือการทำนายปรากฏการณ์ต่างๆของข้อมูลที่กำลังจะเกิดขึ้น ตลอดจนนำความรู้ที่ได้ไปช่วยในกระบวนการตัดสินใจ โดยเทคนิคต่าง ๆ เหล่านี้ สามารถนำไปใช้ให้เกิดประโยชน์ได้ในหลาย ๆ สาขา

2.1 ดาต้าไมนิ่ง

2.1.1 นิยาม

ดาต้าไมนิ่ง คือ การค้นหาความสัมพันธ์และรูปแบบทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ในข้อมูลจำนวนมาก ดาต้าไมนิ่งจะทำการสำรวจและวิเคราะห์ข้อมูลดังกล่าว เพื่อให้ได้สารสนเทศที่อยู่ในรูปแบบที่เต็มไปด้วยความหมายและในรูปของกฎ สารสนเทศที่ได้นี้แสดงให้เห็นถึงความรู้ต่าง ๆ ที่มีประโยชน์ มีความถูกต้อง และสามารถนำไปใช้ได้จริง อาจเรียกได้ว่าดาต้าไมนิ่งเป็นการค้นหาความรู้ในฐานข้อมูล

2.1.2 ปัจจัยที่ทำให้ดาต้าไมนิ่งได้รับการสนใจ

ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ได้รับความสนใจอย่างสูงมีดังต่อไปนี้

1. จำนวนและขนาดข้อมูลขนาดใหญ่ถูกผลิตและขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทางอินเตอร์เน็ต, ดาวเทียม และแหล่งผลิตข้อมูลอื่น ๆ เช่น เครื่องอ่านบาร์โค้ด, เครดิตการ์ด, อีคอมเมิร์ส เป็นต้น
2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ เป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์เพื่อการตัดสินใจ ส่วนมากข้อมูลถูกจัดเก็บแยกมาจากระบบปฏิบัติงาน โดยจัดอยู่ในรูปของคลังข้อมูล (Data Warehouse) ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้
3. การทำคาตาไมนิ่งประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนและความต้องการการคำนวณสูง จึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์สมรรถนะสูง ซึ่งในปัจจุบัน

ระบบคอมพิวเตอร์สมรรถนะสูงมีราคาที่ต่ำลง พร้อมด้วยเริ่มมีเทคโนโลยีที่นำเครื่องไมโครคอมพิวเตอร์จำนวนมากมาเชื่อมต่อกันโดยเครือข่ายความเร็ว ทำให้ได้ระบบคอมพิวเตอร์สมรรถนะสูงในราคาถูก

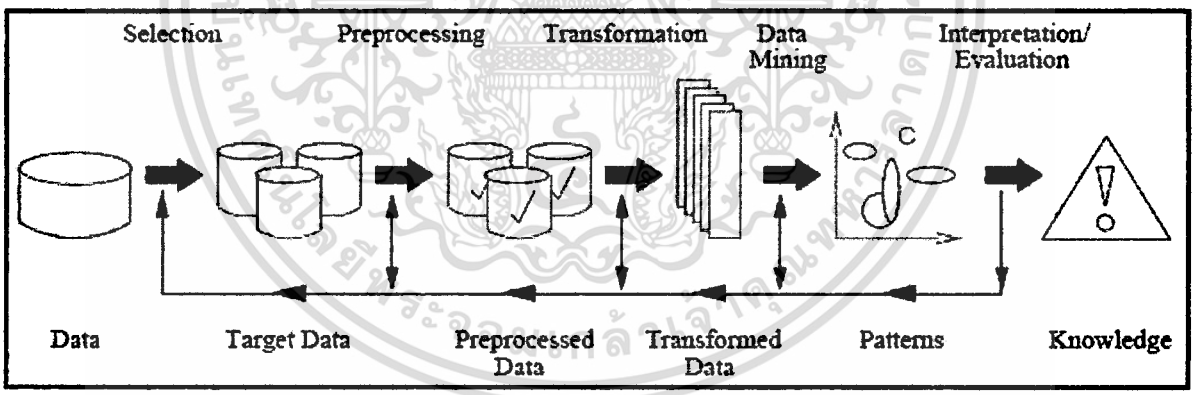
4. การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า เนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า มีการผลิตข้อมูลไว้อย่างมากมายแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์เท่าที่ควร จึงจำเป็นอย่างยิ่งที่จะต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูล เพื่อที่จะนำความรู้ที่ได้รับไปวิเคราะห์ประกอบการตัดสินใจการจัดการในระบบต่าง ๆ ให้เกิดประโยชน์สูงสุด

2.1.3 ลักษณะของงานดาต้าไมนิ่ง

แบ่งออกเป็น 4 ลักษณะ ดังนี้

1. การวิเคราะห์ความสัมพันธ์ (Association Analysis)
2. การจัดหมวดหมู่และการทำนายค่า (Classification and Prediction)
3. การประเมินค่า (Estimation)
4. การแบ่งกลุ่ม (Clustering)

2.1.4 ขั้นตอนการทำดาต้าไมนิ่ง



รูปที่ 2.1 แสดงขั้นตอนการทำดาต้าไมนิ่ง

1. การกำหนดวัตถุประสงค์ (Business Objective Determination)

เป็นการทำความเข้าใจกับปัญหาและความต้องการทางธุรกิจของหน่วยงานหรือองค์กร ซึ่งจะเป็นการพิจารณาว่าจะนำดาต้าไมนิ่งไปแก้ไขปัญหาใด ปัญหานั้นมีความสำคัญพอที่จะทำดาต้าไมนิ่งหรือไม่ ลักษณะและประเภทข้อมูลที่มีอยู่สามารถทำการค้นหาสารสนเทศที่ต้องการได้หรือไม่

2. การเตรียมข้อมูล (Data Preparation)

เป็นการจัดรูปแบบข้อมูลให้อยู่ในรูปแบบมาตรฐาน ที่มีความเหมาะสม โดยสามารถแบ่งเป็นขั้นตอนย่อย ๆ ได้ดังนี้

1) การเลือกข้อมูล (Data Selection)

คือ การระบุลักษณะของข้อมูลที่จะนำมาใช้ประโยชน์ได้ และเลือกข้อมูลเหล่านั้นจากข้อมูลจำนวนมากที่มีอยู่ ในขั้นตอนนี้ถือว่าเป็นหัวใจหลักของกระบวนการทำคาด้าไมนิ่งโดยจะใช้เวลา 60% ของเวลาทั้งหมด ซึ่งเกณฑ์ที่ใช้ในการเลือกข้อมูลนั้นจะแตกต่างไปตามเป้าหมายของการทำคาด้าไมนิ่ง ในการเลือกข้อมูลจำเป็นจะต้องเข้าใจความหมายและคุณลักษณะของข้อมูลนั้น ๆ โดยข้อมูลแบ่งได้เป็น 2 ประเภท ดังนี้

ข้อมูลเชิงประเภท (Categorical) แบ่งเป็น

- Nominal ข้อมูลที่ลำดับไม่มีความสำคัญ เช่น สถานะการแต่งงาน (โสด, แต่งงาน, หย่า, ไม่ทราบ)
- Ordinal ข้อมูลที่ลำดับมีความสำคัญมีลำดับ เช่น ลำดับของลูกค้า (ดี , ปานกลาง, ไม่ดี)

ข้อมูลเชิงปริมาณ (Quantitative) แบ่งเป็น

- ข้อมูลที่มีค่าต่อเนื่อง (Continuous) เช่น ข้อมูลรายได้
- ข้อมูลที่ไม่ต่อเนื่อง (Discrete) เช่น ข้อมูลจำนวนพนักงาน

2) การประมวลผลข้อมูลเบื้องต้น (Data Preprocessing)

เป็นการกลั่นกรองข้อมูล เพื่อทำให้มั่นใจว่าคุณภาพของข้อมูลที่ถูกเลือกนั้นเหมาะสมหรือไม่ แต่บางครั้งข้อมูลที่ได้มายังไม่มีความพร้อมเพียงพอ เนื่องจากมีอุปสรรคอยู่ 2 ประการ คือ

- Noisy Data คือ ข้อมูลที่มีความผิดเพี้ยนจากข้อมูลที่ได้คาดการณ์ไว้ เช่น ข้อมูลอายุคนเป็น 300 ปีหรือค่าของรายได้ติดลบ เป็นต้น ซึ่งค่าเหล่านี้ควรจะถูกแก้ไขหรือเอาออกจากการวิเคราะห์
- Missing Values คือ ค่าที่ถูกเว้นว่างไว้

เทคนิคที่ใช้ในกระบวนการนี้ ได้แก่

- Data Cleaning คือการปรับแต่งข้อมูลที่ยังไม่เหมาะสมให้เหมาะสมมากยิ่งขึ้น เช่น การลด Missing Values โดยการเพิ่มค่าลงในค่าว่างนั้น ๆ การกำจัดNoisy Data ออกจากข้อมูล เป็นต้น
- Data Integration คือการรวบรวมแหล่งที่มาของข้อมูลจากหลาย ๆ แหล่งเข้าด้วยกัน
- Data Reducing คือ การลดขนาดของข้อมูลที่นำมาทำไมนิ่ง โดยอาจจะเป็นการลดจำนวนของข้อมูล คือการลดจำนวนระเบียนลงโดยเลือกใช้เฉพาะข้อมูลตัวอย่าง หรือลดมิติของข้อมูล คือ การลบตัวแปรหรือแอตทริบิวของข้อมูลที่ไม่จำเป็นออก

3) การแปลงข้อมูล (Data Transformation)

คือการจัดรูปแบบของข้อมูลให้สอดคล้องกับแบบจำลองการวิเคราะห์และอัลกอริทึมที่ใช้การแปลงข้อมูลสามารถทำได้โดยวิธีการต่าง ๆ ดังนี้

- Normalization คือ การกำหนดช่วงหรือค่าของข้อมูลที่เป็นไปได้ให้เล็กลง เช่น แปลงแอตทริบิวที่มีค่าเป็นจำนวนจริง ให้มีค่าระหว่าง 0.0 – 1.0 เป็นต้น
- Discretization คือ การเปลี่ยนแปลงข้อมูลที่มีค่าต่อเนื่องให้เป็นค่าไม่ต่อเนื่อง เช่น การแบ่งข้อมูลเป็นช่วง ๆ แล้วใช้ค่าที่เหมาะสมมาเป็นตัวแทนของข้อมูล
- Attribute Construction คือ การรวมแอตทริบิวย่อย ๆ มาเป็นแอตทริบิวเดียว
- Generalization เป็นการแปลงค่าข้อมูลจากระดับต่ำให้มีระดับสูงขึ้น เช่น ค่าของอายุ ถูกแปลงเป็นช่วงอายุ

3. การทำไมนิ่ง (Data Mining)

เป็นขั้นตอนในการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้ให้เหมาะสมกับปัญหาและวัตถุประสงค์ ในบางครั้งเพื่อให้ได้ผลลัพธ์ที่ดี การทำไมนิ่งอาจต้องใช้วิธีการและเทคนิคหลาย ๆ อย่างรวมกัน

4. การวิเคราะห์ผลลัพธ์ (Analysis of Results)

เป็นการวิเคราะห์และแปลความหมายผลที่ได้จากการทำไมนิ่ง โดยอาจเลือกวิเคราะห์เฉพาะผลลัพธ์ที่น่าสนใจ คือเป็นผลลัพธ์ที่ง่ายต่อความเข้าใจ เป็นสารสนเทศที่ไม่เคยรู้มาก่อน และเป็นไปตามวัตถุประสงค์ที่ได้กำหนดไว้

5. การนำความรู้หรือสารสนเทศที่ได้ไปใช้ประโยชน์ (Assimilation of Knowledge)

เป็นขั้นตอนในการเลือก รวบรวมความรู้ที่ได้จากการแปลและวิเคราะห์ผลลัพธ์นำไปประยุกต์ใช้กับองค์กรจริง ๆ เพราะผลลัพธ์ที่ได้จากการไมนิ่งอาจมีรูปแบบจำนวนมากมาย บางผลลัพธ์ที่ได้อาจจะไม่มีประโยชน์ต่อองค์กรเลย ดังนั้นจึงต้องมีการวัดความน่าสนใจของผลลัพธ์ที่ได้จาก

- เป็นสารสนเทศที่ไม่เคยรู้มาก่อน
- ความถูกต้อง คือ สารสนเทศที่ได้ต้องมีความถูกต้องและน่าเชื่อถือ
- สามารถนำไปใช้ให้เกิดประโยชน์กับองค์กรได้จริง

2.2 การค้นหากฎความสัมพันธ์จากข้อมูล

2.2.1 กฎความสัมพันธ์พื้นฐาน (Association Rules Discovery)

การค้นหากฎความสัมพันธ์ เป็นเทคนิคหนึ่งของดาต้าไมนิ่งในการค้นหาความสัมพันธ์ระหว่างรายการ (Item) ในแต่ละเรคคอร์ดหรือกลุ่มของเรคคอร์ด ที่ปรากฏขึ้นในฐานข้อมูล โดยรูปแบบของกฎจะแสดงข้อมูลในลักษณะ "ถ้า X แล้ว Y" ใช้สัญลักษณ์ X => Y

ข้อมูลที่นำมาใช้ในการหากฎความสัมพันธ์พื้นฐาน จะเป็นไปในลักษณะที่แต่รายการ จะประกอบไปด้วยตัวระบุและกลุ่มของรายการ เช่น

ตารางที่ 2.1 ตัวอย่างข้อมูลรายการ (เปลี่ยนแปลง)

ID	Itemsets
1	Item1, Item2
2	Item3
3	Item2, Item5

จากตัวอย่างข้อมูลที่ใช้จะเห็นได้ว่าการหากฎความสัมพันธ์พื้นฐานจัดว่าเป็น Boolean Association Rules คือแต่ละรายการจะมองในลักษณะว่ามีหรือไม่มี

ในการหากฎความสัมพันธ์นี้มีค่า 2 ค่าที่ใช้กันอยู่ก็คือ

- Support Factor เป็นค่าแสดงความสัมพันธ์ ระหว่างจำนวนของเหตุการณ์ X => Y ที่เกิดขึ้น กับจำนวนรายการ (Transaction) ที่เกิดขึ้นทั้งหมด

$$X => Y: \text{support factor} = P(X \cup Y)$$

- Confidence Factor เป็นค่าแสดงความเป็นจริงของกฎ สามารถคำนวณโดยการหารค่า X=> Y ด้วยจำนวนเหตุการณ์ X ที่เกิดขึ้น

$$X => Y: \text{confidence factor} = P(X|Y)$$

$$\text{โดย } P(X|Y) = \frac{\text{จำนวน Support}(X \cup Y)}{\text{จำนวน Support}(X)}$$

จะต้องมีการกำหนด ค่า Support Factor เริ่มต้น (Min_sup) และ ค่า Confidence Factor เริ่มต้น (min_conf) เพื่อใช้ในการเลือกกฎความสัมพันธ์

กระบวนการในการค้นหากฎความสัมพันธ์มีอยู่ 2 ขั้นตอน ดังนี้

1. การหา Frequent Itemsets ทั้งหมด โดยทั่วไปจะนิยมใช้อัลกอริทึม Apriori การหา Frequent Itemsets จะได้ผลลัพธ์เป็น Itemsets (กลุ่มของรายการ) ทั้งหมดที่สัมพันธ์กัน และนับจำนวนครั้งได้มากกว่าหรือเท่ากับค่า Min_sup ที่กำหนดไว้

2. การสร้างกฎความสัมพันธ์จาก Frequent Itemsets คือ จะเป็นการเลือกเซตย่อยทั้งหมดของ Frequent Itemsets เป็นกฎความสัมพันธ์ โดยที่แต่ละเซตย่อยต้องมีค่า Confidence Factor ที่มากกว่าหรือเท่ากับค่า min_conf ที่ได้กำหนดไว้

2.2.2 การค้นหากฎความสัมพันธ์ของข้อมูล สำหรับข้อมูลเชิงปริมาณ (Quantitative Association Rules Discovery)

ในงานทางธุรกิจหรือทางวิทยาศาสตร์ ชนิดของข้อมูลเป็นไปได้หลายชนิด ซึ่งหนึ่งในนั้นก็คือ ข้อมูลเชิงปริมาณ (Quantitative Data) เช่น ข้อมูลอายุ, เงินเดือน เป็นต้น ซึ่งข้อมูลเหล่านี้สามารถนำมาเรียงลำดับหรือใช้เปรียบเทียบได้ บางครั้งชนิดของข้อมูลอาจจะมาในลักษณะของข้อมูลเชิงประเภท (Categorical Data) เช่นข้อมูลรหัสไปรษณีย์, ข้อมูลเพศ เป็นต้น ซึ่งจะไม่สามารถนำไปเรียงลำดับได้ และข้อมูลเชิงประเภทนี้สามารถมองเป็นแบบ Boolean ได้ ซึ่งหลักพื้นฐานในการสร้างกฎความสัมพันธ์ คือเป็นการหามาจากข้อมูลที่มองในลักษณะเป็น Boolean หรือที่เรียกว่า Boolean Association Rules ดังนั้นการที่จะหากฎความสัมพันธ์จากข้อมูลเชิงปริมาณ จึงจำเป็นต้องแปลงข้อมูลให้เป็นลักษณะ Boolean ก่อน แล้วจึงทำการหา Frequent Itemsets และหากฎความสัมพันธ์ต่อไป เช่น ตัวอย่างแสดงผลของกฎความสัมพันธ์ข้อมูลเชิงปริมาณดังตารางที่ 3.2

2.2.3 ขั้นตอนการค้นหากฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

ในที่นี้เราจะใช้วิธีการของ R. Srikant และ R. Agrawal มี 5 ขั้นตอน ดังนี้

1. สร้างช่วงข้อมูลสำหรับ แอตทริบิว ที่เป็นข้อมูลเชิงปริมาณ
2. แทนค่าข้อมูลด้วยตัวเลขที่เรียงลำดับ
3. หา Frequent Itemsets
4. สร้างกฎความสัมพันธ์
5. เลือกกฎความสัมพันธ์ที่เหมาะสมเป็นผลลัพธ์

ตารางที่ 2.2 ตัวอย่างตาราง People

RecordID	Age	Married	Numcars
100	23	No	1
200	25	Yes	1
300	29	No	0
400	34	Yes	2
500	38	Yes	2

(min_sup = 40%, min_conf = 50%)

ตารางที่ 2.3 ตัวอย่างผลการหากฎความสัมพันธ์ข้อมูลเชิงปริมาณ

Rules (Sample)	Support	Confidence
(Age: 30…39) and (Married: Yes) => (NumCar: 2)	40%	100%
(NumCars: 0…1) => (Married: No)	40%	66.6%

โดยแต่ละขั้นตอนของการค้นหาความสัมพันธ์ของข้อมูลเชิงปริมาณมีรายละเอียด ดังนี้

1. สร้างช่วงข้อมูลสำหรับ แอตทริบิว ที่เป็นข้อมูลเชิงปริมาณ

จะเป็นการสร้างช่วง สำหรับแทนค่าข้อมูลเชิงปริมาณ โดยจะไม่เลือกสร้างในทุก แอต ทริบิวก็ได้ ในการแบ่งช่วงข้อมูลนั้นมีความสำคัญมาก เพราะจะมีผลต่อค่า Confidence และค่า Support ดังตัวอย่าง

กฎ "(age: 20…24) => (married: No)" มีค่า Confidence = 100%

แต่ถ้าแบ่งช่วงเปลี่ยนไป

"(age: 20…25) => (married: No)" มีค่า Confidence = 50%

ซึ่งในการแบ่งช่วงข้อมูลแรก ๆ อาจจะไม่ถูกต้องได้ ก็จะสามารถรวมช่วงข้อมูลได้ในภายหลัง วิธีการหาจำนวนช่วงที่จะ ทำได้ด้วยวิธี Equi-Depth Partitioning ซึ่งจะมีสูตร ดังนี้

จำนวนช่วงทั้งหมด = (2 * n) / m * (K - 1)

ซึ่ง n คือ จำนวน แอตทริบิว ทั้งหมดที่เป็นข้อมูลเชิงปริมาณ

m คือ ค่า Min_sup

K คือ ค่า Partial Completeness Level ซึ่งจะกำหนดเอง เป็นค่าที่จัดการเกี่ยวกับจำนวนของข้อมูลที่หายไปในการแบ่งช่วง โดยยิ่งมีค่าน้อยเท่าไหร่ ก็จะทำให้ข้อมูลหายน้อยมากเท่านั้น แต่ถ้าหากใส่ค่าน้อยเกินไปจะทำให้มีจำนวนการแบ่งช่วงมาก ทำให้ต้องใช้เวลาในการทำงานนาน

โดยในที่นี้ จะกำหนดค่า K= 3.5 หาจำนวนช่วงทั้งหมดได้ดังนี้

n = 2 (คือ ในตาราง People มีแอตทริบิวที่เป็นข้อมูลเชิงปริมาณ 2 แอตทริบิว คือ Age และ Numcars) m = 40% = 0.4

จำนวนช่วงทั้งหมด = (2 * 2) / 0.4 *(3.5 - 1) = 4

ในตาราง People แบ่งช่วงข้อมูลเป็น 4 ช่วง โดยจะแบ่งได้เฉพาะ Age แต่ Numcars ไม่สามารถแบ่งได้ เพราะช่วงข้อมูลมีค่า น้อยกว่า 4 สามารถแบ่งได้ดังนี้

ตารางที่ 2.4 แสดงการแบ่งช่วงอายุ

ช่วงอายุ
20...24
25...29
30...34
35...39

เมื่อได้ทำการแบ่งช่วงอายุแล้วตาราง People จะมีลักษณะ ดังนี้

ตารางที่ 2.5 แสดงตาราง People หลังจากมีการแบ่งช่วงอายุแล้ว

RecordID	Age	Married	Numcars
100	20…24	No	1
200	25…29	Yes	1
300	25…29	No	0
400	30…34	Yes	2
500	35…39	Yes	2

2. แทนค่าข้อมูลด้วยตัวเลขที่เรียงลำดับ

เพราะในขั้นตอน การหา Frequent Itemsets นั้นจะใช้อัลกอริทึมที่พัฒนามาจาก Apriori ซึ่งต้องใช้รายการที่เรียงลำดับ

สำหรับแอตทริบิวที่เป็นข้อมูลเชิงปริมาณ และไม่ได้ทำการแบ่งช่วง จะใช้ค่าของข้อมูลจริง เช่น Numcars แต่ถ้าเป็นแอตทริบิวที่เป็นข้อมูลเชิงปริมาณ และทำการแบ่งช่วงแล้ว เช่น Age จะแทนด้วยเลขลำดับของการแบ่งช่วง และถ้าเป็นข้อมูลเชิงประเภท จะแทนค่าด้วยตัวเลขเรียงลำดับที่กำหนดไว้กับค่าข้อมูลจริง ตัวอย่างเช่น

ตารางที่ 2.6 แสดงตารางแทนค่าช่วงอายุด้วยตัวเลขที่เรียงลำดับ

ช่วงอายุ	เลขเรียงลำดับ
20…24	1
25…29	2
30…34	3
35…39	4

ตารางที่ 2.7 แสดงตารางแทนค่าสถานภาพแต่งงาน

Married	เลขเรียงลำดับ
Yes	1
No	2

เมื่อทำการแทนค่าเสร็จเรียบร้อยแล้วจะได้ตารางดังนี้

ตารางที่ 2.8 แสดง ตาราง People หลังจากมีการแทนค่าแล้ว

RecordID	Age	Married	Numcars
100	1	2	1
200	2	1	1
300	2	2	0
400	3	1	2
500	4	1	2

3. หา Frequent Itemsets มีขั้นตอน ดังนี้

1) ในขั้นนี้สามารถที่จะรวมช่วงตัวเลขเข้าด้วยกันได้ โดยการรวมนี้จะต้องรวมเข้ากับช่วงตัวเลขที่ติดกัน และเมื่อรวมแล้ว ช่วงข้อมูลใหม่ที่เกิดขึ้นนั้น ต้องมีค่าจำนวน Support ไม่มากกว่าค่า Max_sup ที่ได้กำหนดไว้ (เป็นค่าที่ต้องระบุเข้าไปเพื่อเป็นการกำหนดขอบเขตในการรวมช่วงข้อมูล) ซึ่งในที่นี้กำหนดค่า Max_sup เท่ากับ 60% ดังนั้น เมื่อรวมช่วงข้อมูลแล้ว แต่ละ Itemsets ที่เกิดขึ้น จะต้องมีจำนวน Support ไม่มากกว่า 3 ดังตัวอย่างในตาราง 2.9 จากนั้นทำการรวมช่วงตัวเลขดังตัวอย่างในตารางที่ 2.10

ตารางที่ 2.9 แสดงจำนวน Support ทั้งหมด

Itemsets	จำนวน Support
{<Age:1>}	1
{<Age:2>}	2
{<Age:3>}	1
{<Age:4>}	1
{<Married:1>}	3
{<Married:2>}	2

ตารางที่ 2.9 (ต่อ)

Itemsets	จำนวน Support
{<NumCars:0>}	1
{<NumCars:1>}	2
{<NumCars:2>}	2

ตารางที่ 2.10 แสดง Itemsets ทั้งหมดที่ทำการรวมได้

Itemsets	จำนวน Support
{<Age:20..24>}	1
{<Age:20..29>}	3
{<Age:25..29>}	2
{<Age:25..34>}	3
{<Age:30..34>}	1
{<Age:30..39>}	2
{<Age:35..39>}	1
{<Married:Yes>}	3
{<Married:No>}	2
{<NumCars:0>}	1
{<NumCars:0..1>}	3
{<NumCars:1>}	2
{<NumCars:2>}	2

2) เลือกเฉพาะ Itemsets ที่มีจำนวน Support มากกว่า Min_sup ซึ่งค่า Min_sup เท่ากับ 40% ก็จะเลือก เฉพาะ Itemsets ที่มีจำนวน Support ตั้งแต่ 2 ขึ้นไป ดังตัวอย่างในตารางที่ 2.11

ตารางที่ 2.11 แสดง Itemsets ทั้งหมดที่จะนำไปใช้ในอัลกอริทึม

Itemsets	จำนวน Support
{<Age:20..29>}	3
{<Age:25..29>}	2
{<Age:25..34>}	3
{<Age:30..39>}	2

ตารางที่ 2.11 (ต่อ)

Itemsets	จำนวน Support
{<Married:Yes>}	3
{<Married:No>}	2
{<NumCars:0..1>}	3
{<NumCars:1>}	2
{<NumCars:2>}	2

3) ใช้อัลกอริทึมที่พัฒนามาจาก Apriori ที่เพิ่มเติมการ Interest Prune โดยอัลกอริทึม Apriori นั้นมีการทำงานดังรูปที่ 2.2 ดังนี้

1) L_1 = {large 1-itemsets};
2) for ($k = 2$; $L_{k-1} \neq \emptyset$; 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) $L_k = \{c \in C_k \mid$ c.count $\geq$ minsup$\}$
10) end
11) Answer = $\bigcup_k L_k$;

รูปที่ 2.2 แสดงอัลกอริทึม Apriori

อธิบายการทำงานของอัลกอริทึมที่พัฒนามาจาก Apriori ได้ดังนี้ คือจะทำงานซ้ำไปเรื่อย ๆ โดย k-Itemsets (จำนวน Itemsets ที่ประกอบด้วย k ค่า) เช่น เซต {computer, financial_software} คือ 2-Itemsets ซึ่ง k-Itemsets จะใช้สร้าง (k+1)-Itemsets ต่อไปเรื่อย ๆ โดยเริ่มต้น จะมี frequent 1-Itemsets โดยแทนค่าด้วย L1 และ L1 จะนำไปใช้หาค่า L2 และได้ค่า frequent 2-Itemsets สำหรับใช้หา L3 ต่อไปเรื่อย ๆ จนกระทั่งไม่สามารถหาค่า frequent k-Itemsets ได้อีกต่อไป

ขั้นตอนที่ Lk-1 ใช้ในการหา Lk ประกอบด้วยขั้นตอน ดังนี้

- Join Phase จะหาเซตของ Candidate k-itemsets ด้วยการ join L_{k-1} ด้วยตัวมันเอง โดยเซตของ Candidate k-itemsets แทนค่าด้วย C_k ให้ l_1 และ l_2 เป็นแต่ละ Itemsets ใน L_{k-1} ให้ค่า $l_1[j]$ จะหมายความว่า j คือลำดับที่ของรายการใน

แต่ละ Itemsets ซึ่งแต่ละ Itemsets จะต้องมีการเรียงลำดับอยู่แล้ว การ Join ระหว่าง L_{k-1} และ L_{k-1} นั้นทำได้โดย ทุก ๆ ค่า $l_1[1...k-2]$ และ $l_2[1...k-2]$ ในแต่ละลำดับต้องเท่ากัน และค่าที่ $l_1[k-1]$ ต้องน้อยกว่า $l_2[k-1]$ เท่านั้นเพื่อไม่ให้เกิดค่าที่ซ้ำกัน ดังตัวอย่าง

$L_2 = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$

จะหา L_3, k = 3 จากการ join กันระหว่าง L_2 สามารถหาได้ได้ดังนี้

$l_1 = \{1, 2\}$ โดยที่ $l_1[1] = 1, l_1[2] = 2$

$l_2 = \{1, 3\}$ โดยที่ $l_2[1] = 1, l_2[2] = 3$

$l_3 = \{2, 3\}$ โดยที่ $l_3[1] = 2, l_3[2] = 3$

ทำการ Join ระหว่าง l_1 และ l_2 เพราะ ค่า $l_1[1] = l_2[1]$ และ $l_1[2] < l_2[2]$ ได้ค่า $\{1, 2, 3\}$

แต่ระหว่าง l_2 และ l_3 จะไม่ทำการ Join เพราะ ค่า $l_2[1]$ ไม่เท่ากับค่า $l_3[1]$

ดังนั้นจะได้ $L_3 = \{\{1, 2, 3\}\}$

- Subset Prune Phase จะเลือกเฉพาะ Candidate k-itemsets ที่ทุกเซตย่อยของ Candidate k-itemsets เป็นสมาชิกของ L_{k-1} เท่านั้น

- Interest Prune Phase โดยจะต้องมีการกำหนด Interest Level(R) ขั้นตอนนี้เป็นการหา Candidate k-itemsets ที่มีค่า Support และ Confidence มากกว่าค่า Expected Value (แสดงไว้ในหัวข้อ 5) ที่ได้กำหนดไว้ โดยจะทำการลบทุก ๆ Candidate k-itemsets ที่มีค่า Support มากกว่า 1/R

- Prune Phase จะเลือกเฉพาะ Candidate k-itemsets ที่มีค่า Support มากกว่าหรือเท่ากับค่า Min_sup

ตัวอย่างการใช้อัลกอริทึม Apriori ที่เพิ่มเติมการ Interest Prune

1) Join Phase(1) ซึ่ง L_1 คือ ตาราง 11 ได้ผลการ Join แสดงในตารางที่ 2.12

2) Subset Prune Phase(1) ได้ผลแสดงในตารางที่ 2.12

3) Interest Prune Phase(1) ได้ผลแสดงในตารางที่ 2.12

4) Prune Phase(1) ซึ่ง Min_sup เท่ากับ 40% หรือ 2 Records ได้ผลแสดงในตารางที่ 2.13

5) Join Phase(2) ซึ่ง L_2 คือตาราง 2.13 ได้ผลการ Join แสดงในตารางที่ 2.14

6) Subset Prune Phase(2) ได้ผลแสดงในตารางที่ 2.15

7) Interest Prune Phase(2) ได้ผลแสดงในตารางที่ 2.15

8) Prune Phase(2) ได้ผลแสดงในตารางที่ 2.15

ตารางที่ 2.12 แสดง Itemsets ที่ทำการ Join รอบที่ 1

Itemsets	จำนวน Support
{<Age: 20..29>,<Married:Yes>}	1
{<Age: 20..29>,<Married:No>}	2
{<Age: 20..29>,<NumCars:0..1>}	3
{<Age: 20..29>,<NumCars:1>}	2
{<Age: 20..29>,<NumCars:2>}	0
{<Age: 25..29>,<Married:Yes>}	1
{<Age: 25..29>,<Married:No>}	1
{<Age: 25..29>,<NumCars:0..1>}	2
{<Age: 25..29>,<NumCars:1>}	1
{<Age: 25..29>,<NumCars:2>}	0
{<Age: 25..34>,<Married:Yes>}	2
{<Age: 25..34>,<Married:No>}	1
{<Age: 25..34>,<NumCars:0..1>}	2
{<Age: 25..34>,<NumCars:1>}	1
{<Age: 25..34>,<NumCars:2>}	1
{<Age: 30..39>,<Married:Yes>}	2
{<Age: 30..39>,<Married:No>}	0
{<Age: 30..39>,<NumCars:0..1>}	0
{<Age: 30..39>,<NumCars:1>}	0
{<Age: 30..39>,<NumCars:2>}	2
{<Married:Yes>,<NumCars:0..1>}	1
{<Married:Yes>,<NumCars:1>}	1
{<Married:Yes>,<NumCars:2>}	2
{<Married:No>,<NumCars:0..1>}	2
{<Married:No>,<NumCars:1>}	1
{<Married:No>,<NumCars:2>}	0

ตารางที่ 2.13 แสดง Itemsets ทั้งหมดที่เกิดจากทำงานของอัลกอริทึมในรอบที่ 1

Itemsets	จำนวน Support
{<Age: 20..29>,<Married:No>}	2
{<Age: 20..29>,<NumCars:0..1>}	3
{<Age: 20..29>,<NumCars:1>}	2
{<Age: 25..29>,<NumCars:0..1>}	2
{<Age: 25..34>,<Married:Yes>}	2
{<Age: 25..34>,<NumCars:0..1>}	2
{<Age: 30..39>,<Married:Yes>}	2
{<Age: 30..39>,<NumCars:2>}	2
{<Married:Yes>,<NumCars:2>}	2
{<Married:No>,<NumCars:0..1>}	2

ตารางที่ 2.14 แสดง Itemsets ที่เกิดจากการ Join ในรอบที่ 2

Itemsets	จำนวน Support
{<Age: 20..29>,<Married:No>,<NumCar:0..1>}	2
{<Age: 20..29>,<Married:No>,<NumCar:1>}	2
{<Age: 30..39>,<Married:Yes>,<NumCar:2>}	2

ตารางที่ 2.15 แสดง Itemsets ทั้งหมดที่เกิดจากทำงานของอัลกอริทึมในรอบที่ 2

Itemsets	จำนวน Support
{<Age: 20..29>,<Married:No>,<NumCar:0..1>}	2
{<Age: 30..39>,<Married:Yes>,<NumCar:2>}	2

สรุปแล้วได้ Frequent Itemset ทั้งหมด ดังนี้

ตารางที่ 2.16 แสดง Frequent Itemsets ทั้งหมด

Itemsets	จำนวน Support
{<Age: 20..29>,<Married:No>}	2
{<Age: 20..29>,<NumCars:0..1>}	3
{<Age: 20..29>,<NumCars:1>}	2
{<Age: 25..29>,<NumCars:0..1>}	2

ตารางที่ 2.16 (ต่อ)

Itemsets	จำนวน Support
{<Age: 25..34>,<Married:Yes>}	2
{<Age: 25..34>,<NumCars:0..1>}	2
{<Age: 30..39>,<Married:Yes>}	2
{<Age: 30..39>,<NumCars:2>}	2
{<Married:Yes>,<NumCars:2>}	2
{<Married:No>,<NumCars:0..1>}	2
{<Age: 20..29>,<Married:No>,<NumCar:0..1>}	2
{<Age: 30..39>,<Married:Yes>,<NumCar:2>}	2

4) สร้างกฎความสัมพันธ์จาก Frequent Itemsets ด้วยการใช้อัลกอริทึมสำหรับสร้างกฎความสัมพันธ์ มีการทำงานดังรูปที่ 2.2

- จากแต่ละ Frequent Itemsets l, สร้างเซตย่อย s ที่ไม่ใช่เซตว่างของ l
- แต่ละเซตย่อยที่สร้างขึ้นมา นำมาสร้างกฎ ได้ดังนี้

"s => (l-s)" และทุกกฎที่สร้างขึ้นมาต้องเป็นกฎที่มีค่า Confidence มากกว่าหรือเท่ากับค่า min_conf และค่า Support มากกว่าหรือเท่ากับค่า Min_sup ดังตัวอย่าง โดยจะแสดงตัวอย่างกฎบางส่วน

```
// Simple Algorithm
forall large itemsets l_k, k ≥ 2 do
    call genrules(l_k, l_k);

// The genrules generates all valid rules â ⟹ (l_k − â), for all â ⊂ a_m
procedure genrules(l_k: large k-itemset, a_m: large m-itemset)
1)  A = {(m−1)-itemsets a_{m−1} | a_{m−1} ⊂ a_m};
2)  forall a_{m−1} ∈ A do begin
3)      conf = support(l_k)/support(a_{m−1});
4)      if (conf ≥ minconf) then begin
7)          output the rule a_{m−1} ⟹ (l_k − a_{m−1}), with confidence = conf and support = support(l_k);
8)          if (m − 1 > 1) then
9)              call genrules(l_k, a_{m−1}); // to generate rules with subsets of a_{m−1} as the antecedents
10)     end
11) end
```

รูปที่ 2.3 แสดงอัลกอริทึมสำหรับสร้างกฎความสัมพันธ์

5) เลือกกฎความสัมพันธ์ที่เหมาะสมเป็นผลลัพธ์ จะใช้หลักการที่เรียกว่า "greater-than-expected-value" สำหรับเลือกกฎความสัมพันธ์

ตารางที่ 2.17 แสดงกฎความสัมพันธ์บางส่วนที่สร้างขึ้นมา

Rules	Support	Confidence
<Age: 30..39> and <Married:Yes> => <NumCars: 2>	40%	100.00%
<Age: 20..29> and <Married:No> => <NumCars: 0..1>	40%	100.00%
<Age:25..34> => <NumCars:0..1>	40%	66.67%
<Age:25..29> => <NumCars:0..1>	40%	100.00%
<Age: 20..29> => <NumCars: 0..1>	60%	100.00%

โดยก่อนที่จะหาค่า Expected Value ได้ จะต้องทำความเข้าใจกับความหมายของ Generalization ก่อน อธิบายตามตัวอย่างดังนี้

A = {<Age: 20..24>}

$\hat{A}$ = {<Age: 20..29>}

$\hat{A}$ เป็น Generalization ของ A ซึ่งก็คือ $\hat{A}$ มีช่วงข้อมูลที่คลอบคลุมช่วงข้อมูลของ A

วิธีการหาค่า Expected Values

$E_{pr(\hat{Z})}[Pr(Z)]$ คือค่า Expected Value ของ Pr(Z) (ค่า Support ของ Z) มีพื้นฐานจาก $Pr(\hat{z})$, โดยที่ $\hat{z}$ คือ Generalization ของ Z โดยที่หาค่าได้ดังนี้

$$E_{pr(\hat{Z})}[Pr(Z)] = \frac{Pr(< z_1, l_1, u_1 >)}{Pr(< z_1, l'_1, u'_1 >)} * ... * \frac{Pr(< z_n, l_n, u_n >)}{Pr(< z_n, l'_n, u'_n >)} * Pr(\hat{z})$$

$E_{pr(\hat{Y}|\hat{X})}[Pr(Y|X)]$ คือค่า Expected Confidence ของกฎ X => Y มีพื้นฐานจากกฎ $\hat{X}$ => $\hat{Y}$, ที่ $\hat{X}$ และ $\hat{Y}$ คือ generalization ของ X และ Y

$$E_{pr(\hat{Y}|\hat{X})}[Pr(Y|X)] = \frac{Pr(< y_1, l_1, u_1 >)}{Pr(< y_1, l'_1, u'_1 >)} * ... * \frac{Pr(< y_n, l_n, u_n >)}{Pr(< y_n, l'_n, u'_n >)} * Pr(\hat{Y}|\hat{X})$$

หลักการเลือกกฎความสัมพันธ์ทำได้ดังนี้

1) กฎที่เลือกนั้นจะต้องมีค่า Support มากกว่าหรือเท่ากับค่า R (Interest Level ที่กำหนดให้) คูณกับค่า Expected Value ที่คำนวณได้

2) กฎที่เลือกนั้นจะต้องมีค่า Confidence มากกว่าหรือเท่ากับค่า R (Interest Level ที่กำหนดให้) คูณกับค่า Expected Confidence ที่คำนวณได้

ดังตัวอย่าง เลือกกฎความสัมพันธ์ {<Age:25..34> => <NumCars:0..1>} สมมติว่ากำหนดค่า R = 1.5

X => Y = <Age:25..29> => <NumCars:0..1>

$\hat{X}$ => $\hat{Y}$ = <Age:25..34> => <NumCars:0..1>

หาค่า Expected Value = (2/2)*0.4 = 0.4 เมื่อนำค่า 0.4 *1.5 = 0.6 ดังนั้น จะเห็นได้ว่า กฎ {<Age:25..34> => <NumCars:0..1>} จะไม่ถูกเลือก

โดยสรุปแล้ว การหากฎความสัมพันธ์ของข้อมูลเชิงปริมาณ ในตัวอย่างแสดงข้อมูลเชิงปริมาณจำนวน 2 แอตทริบิว และข้อมูลเชิงประเภทจำนวน 1 แอตทริบิว โดยการหากฎความสัมพันธ์นี้ คือจะทำการแบ่งช่วงข้อมูลของข้อมูลเชิงปริมาณให้เหมาะสม และอาจจะทำการรวมช่วงข้อมูลที่แบ่งแล้วได้ ตามเงื่อนไขที่ระบุไว้เพื่อไม่ให้เกิดความสูญหายของข้อมูล ต่อมาก็ทำการแทนค่าแอตทริบิวทั้งหมดที่แบ่งช่วงหรือไม่แบ่งช่วงด้วยตัวเลขที่เรียงลำดับ สำหรับเตรียมค่า Itemsets และหลังจากนั้นก็นำอัลกอริทึมที่พัฒนามาจาก Apriori ที่เพิ่มเติมในส่วน Interest Prune มาใช้หา Frequent Itemsets เพื่อนำไปสร้างกฎความสัมพันธ์ ในขั้นตอนสุดท้ายก็จะเป็นการเลือกกฎที่เหมาะสมออกมาเป็นผลลัพธ์ โดยสิ่งที่เพิ่มขึ้นมาของการหากฎความความสัมพันธ์ของข้อมูลเชิงปริมาณจากการหากฎความสัมพันธ์พื้นฐาน(ใช้ในข้อมูลที่เป็น Boolean) นั้นจะเพิ่มในส่วนวิธีการแบ่งช่วงข้อมูล เพิ่มส่วน Interest Prune ในการหา Frequent Itemsets และส่วนที่ใช้ในการเลือกกฎความสัมพันธ์ จึงต้องมีการกำหนดค่าคงที่เพิ่มขึ้นคือค่า Max_sup สำหรับกำหนดขอบเขตการรวมข้อมูล, ค่า Partial Completeness Level(K) สำหรับกำหนดปริมาณช่วงข้อมูลที่ต้องการแบ่ง และค่า Interest Level(R) สำหรับกำหนดระดับความน่าสนใจของกฎที่จะเลือก

บทที่ 3
การวิเคราะห์และออกแบบระบบ

การวิเคราะห์และออกแบบระบบคาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณมีรายละเอียดการออกแบบ ดังนี้

3.1 รายละเอียดที่เกี่ยวข้องกับระบบ

จากการวิเคราะห์พบว่ารายละเอียดที่เกี่ยวข้องกับระบบ มีทั้งหมด ดังนี้

- List of External Entities หรือ ผู้ใช้ระบบหรือสิ่งที่อยู่ภายนอกขอบเขตระบบ ประกอบด้วย

1. ผู้ใช้ คือ ผู้ใช้งานระบบที่ต้องการนำข้อมูลมาทำงานไมนิ่ง

- List of Data หรือ ข้อมูลที่ถูกจัดเก็บในระบบ ประกอบด้วย

1. ข้อมูลที่ต้องการทำไมนิ่ง คือ ข้อมูลเบื้องต้นที่ต้องการนำมาทำไมนิ่ง ซึ่งอาจจะนำมาในลักษณะเป็นฐานข้อมูลหรือไฟล์ CSV โดยข้อมูลในส่วนนี้จะเก็บไว้เฉพาะในระหว่างที่ระบบกำลังทำงานเท่านั้น
2. ข้อมูลรายละเอียดงานและเงื่อนไขต่าง ๆ คือ ข้อมูลรายละเอียดต่าง ๆ ของ Project และข้อมูลเงื่อนไขต่าง ๆ ที่ต้องใช้ในอัลกอริทึมสำหรับหากฎความสัมพันธ์
3. ข้อมูลกฎความสัมพันธ์ คือ ข้อมูลผลลัพธ์กฎความสัมพันธ์ที่ระบบประมวลได้

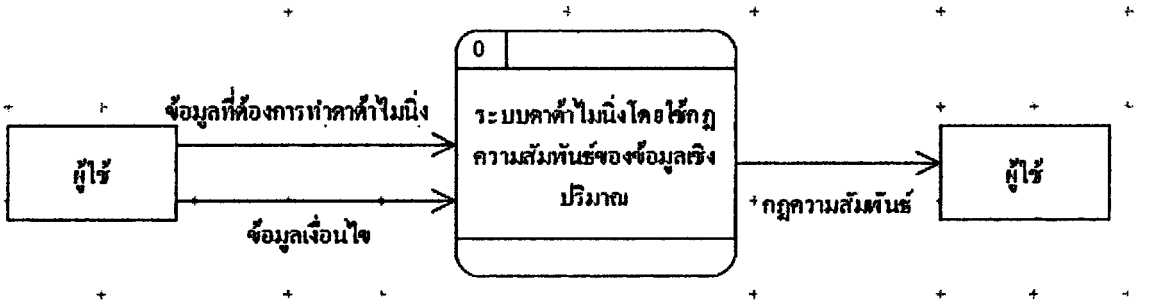
- List of Processes หรือ กระบวนการต่าง ๆ ที่ใช้ในระบบ ประกอบด้วย

1. เรียกดู Project เก่า
2. เพิ่ม Project ใหม่
3. การทำไมนิ่ง

3.2 คอนเท็กซ์ไดอะแกรม

เป็นไดอะแกรมที่แสดงสิ่งแวดล้อมของระบบ เพื่อให้เห็นว่าระบบมีการโต้ตอบกับเอ็กซ์เทอร์นัลเอ็นติตี้ใดบ้าง ในระบบนี้สามารถเขียนคอนเท็กซ์ไดอะแกรมได้ ดังนี้

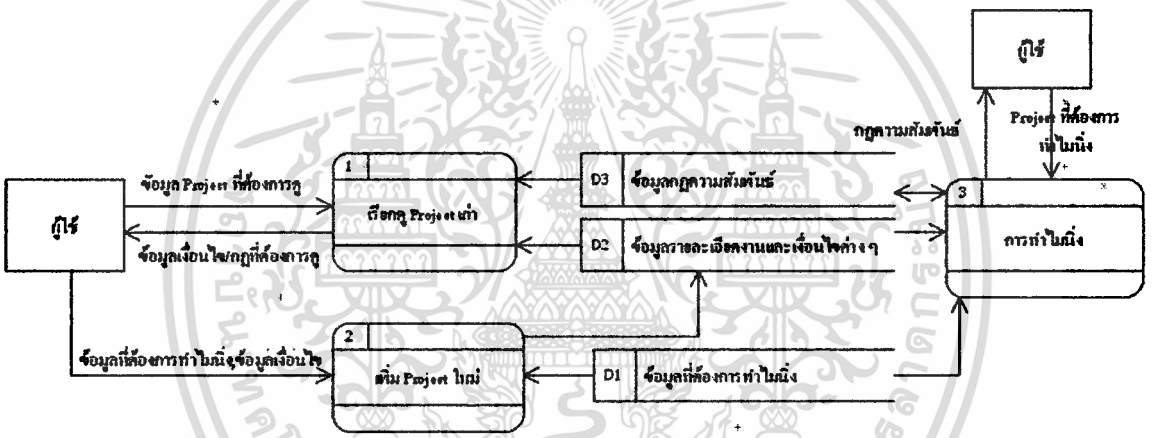
อธิบายได้ว่า ระบบจะมีผู้ใช้งานจากภายนอกเพียงแค่กลุ่มเดียว โดยผู้ใช้จะส่งข้อมูลที่ต้องการทำคาต้าไมนิ่งทั้งหมด เข้าไปในระบบ และระบบจะนำข้อมูลที่ได้ไปประมวลผลได้ผลลัพธ์เป็นกฎความสัมพันธ์ส่งกลับไปให้ผู้ใช้



รูปที่ 3.1 คอนเท็กซ์ไดอะแกรมของระบบคาค้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

3.3 แผนภาพกระแสข้อมูลระดับที่ 1

เป็นไดอะแกรมที่อธิบายคอนเท็กซ์ไดอะแกรมให้ละเอียดมากยิ่งขึ้น



รูปที่ 3.2 แผนภาพกระแสข้อมูลระดับที่ 1 ของระบบคาค้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

อธิบายในแต่ละส่วนงานดังนี้

1. เรียกดู Project เก่า เป็นกระบวนการที่ใช้สำหรับเรียกดูรายละเอียด Project เงื่อนไข Project และกฎความสัมพันธ์ ใน Project เก่าที่เคยประมวลผล โดยผู้ใช้งานจะส่งค่า Project ID ที่ต้องการดูเข้าไปในระบบ และระบบจะแสดงข้อมูลเงื่อนไขและกฎความสัมพันธ์ออกมา มีลำดับการทำงานดังนี้

(1) ระบบแสดงรายการ Project เก่าทั้งหมด

(2) ผู้ใช้เลือก Project เก่าที่ต้องการดู

(3) ระบบจะแสดงรายละเอียดต่าง ๆ ของ Project เก่า คือ

- ชื่อ Project
- รายละเอียด Project

- ถ้าเป็นการไมนิ่งโดยใช้ฐานข้อมูลจะแสดง ค่า Connection String ที่กำหนดไว้ ถ้าเป็นการไม่นิ่งโดยใช้ไฟล์ CSV จะแสดงที่มาของไฟล์ CSV
- แสดงตารางหรือ SQL Statement ที่ระบุที่มาของข้อมูล
- แสดง Transaction ID ที่เลือกไว้
- แสดงแอตทริบิวเชิงปริมาณที่เลือกไว้
- แสดงค่าเงื่อนไขต่าง ๆ ที่กำหนดไว้ คือค่า Minimum Support ค่า Minimun Confidence ค่า Maximum Support ค่า K-Level Completeness และค่า Interest
- กฎความสัมพันธ์

2. เพิ่ม Project ใหม่ เป็นกระบวนการที่ใช้สำหรับเพิ่ม Project ใหม่ โดยจะเพิ่มรายละเอียด Project เงื่อนไข Project และข้อมูลที่ต้องการทำไมนิ่ง ระบบจะทำการบันทึกข้อมูลทั้งหมดไว้ในฐานข้อมูล โดยจะสร้าง Project ID ใหม่สำหรับงานใหม่ มีลำดับการทำงาน ดังนี้

(1) เลือกว่าต้องการทำไมนิ่งด้วยฐานข้อมูลหรือข้อมูลจากไฟล์ CSV
- ถ้าเลือกฐานข้อมูล โดยระบบจะใช้กับฐานข้อมูล Microsoft SQL Server ซึ่งจะต้องกำหนดค่า Connection String ที่ประกอบด้วยค่า Server Name Database Name User และ Password
- ถ้าเลือกไฟล์ CSV จะต้องระบุ ตำแหน่งที่วางไฟล์ CSV นั้นไว้

(2) ผู้ใช้ระบุค่ารายละเอียดของงาน คือ
- ชื่อ Project
- รายละเอียด Project
- ถ้าเป็นการไมนิ่งโดยใช้ฐานข้อมูล คือระบุ ค่า Connection String หรือถ้าเป็นการไม่นิ่งโดยใช้ไฟล์ CSV คือ ตำแหน่งที่วางไฟล์ CSV นั้นไว้
- ถ้าเป็นการไมนิ่งโดยใช้ฐานข้อมูล เลือกตารางหรือ SQL Statement เพื่อระบุที่มาของข้อมูล
- เลือก Transaction ID
- เลือก แอตทริบิวเชิงปริมาณ
- กำหนดค่าเงื่อนไขต่าง ๆ คือค่า Minimum Support ค่า Minimun Confidence ค่า Maximum Support ค่า K-Level Completeness และค่า Interest

(3) ระบบทำการบันทึกข้อมูลพร้อมกับบันทึกค่า Project ID ใหม่ที่ระบบสร้างขึ้นมาด้วย

3. การทำไมนิ่ง คือ กระบวนการที่ใช้อัลกอริทึมเพื่อหากฎความสัมพันธ์ และบันทึกกฎความสัมพันธ์ที่ได้ในฐานข้อมูล โดยรายละเอียดและลำดับการทำงานจะแสดงในแผนภาพกระแสข้อมูลระดับที่ 2

3.4 แผนภาพกระแสข้อมูลระดับที่ 2

โดยจะแสดงในกระบวนการ การทำไมนิ่งตามรูปที่ 3.3 ในส่วนงานนี้ได้มีการแบ่งกระบวนการทำงานเป็น 6 ขั้นตอน คือ

1) รวมข้อมูลที่ผู้ใช้เลือกเป็น 1 ตาราง ขั้นตอนนี้ก็คือการเตรียมข้อมูล จะได้รับข้อมูลและเงื่อนไขต่าง ๆ จากผู้ใช้ แล้วทำการรวมข้อมูลต่าง ๆ ที่ผู้ใช้ต้องการนำมาหากฎความสัมพันธ์ โดยข้อมูลที่ผู้ใช้เลือกอาจมาจากหลายตาราง ระบบจึงต้องนำข้อมูลต่าง ๆ มารวมเป็น 1 ตารางก่อน และมีการเก็บรายละเอียดงานพร้อมทั้งเงื่อนไขไว้ในระบบ โดยเงื่อนไขที่ผู้ใช้ระบุจะประกอบด้วยค่า Minimum Support, Minimum Confidence, Maximum Support, K-Level Completeness และค่า Interest

2) ทำการ Group ข้อมูลเป็นกลุ่ม ๆ คือทำการรวมกลุ่มข้อมูลในแต่ละแอตทริบิว เช่น แอตทริบิวที่เป็นข้อมูลเครื่องดื่ม อาจจะมีข้อมูลต่างกันมากมาย เช่น น้ำส้ม น้ำแครอท น้ำโค้ก น้ำมะเขือเทศ ซึ่งอาจจะรวมเป็นกลุ่มได้คือ น้ำอัดลม น้ำผัก น้ำผลไม้ เป็นต้น

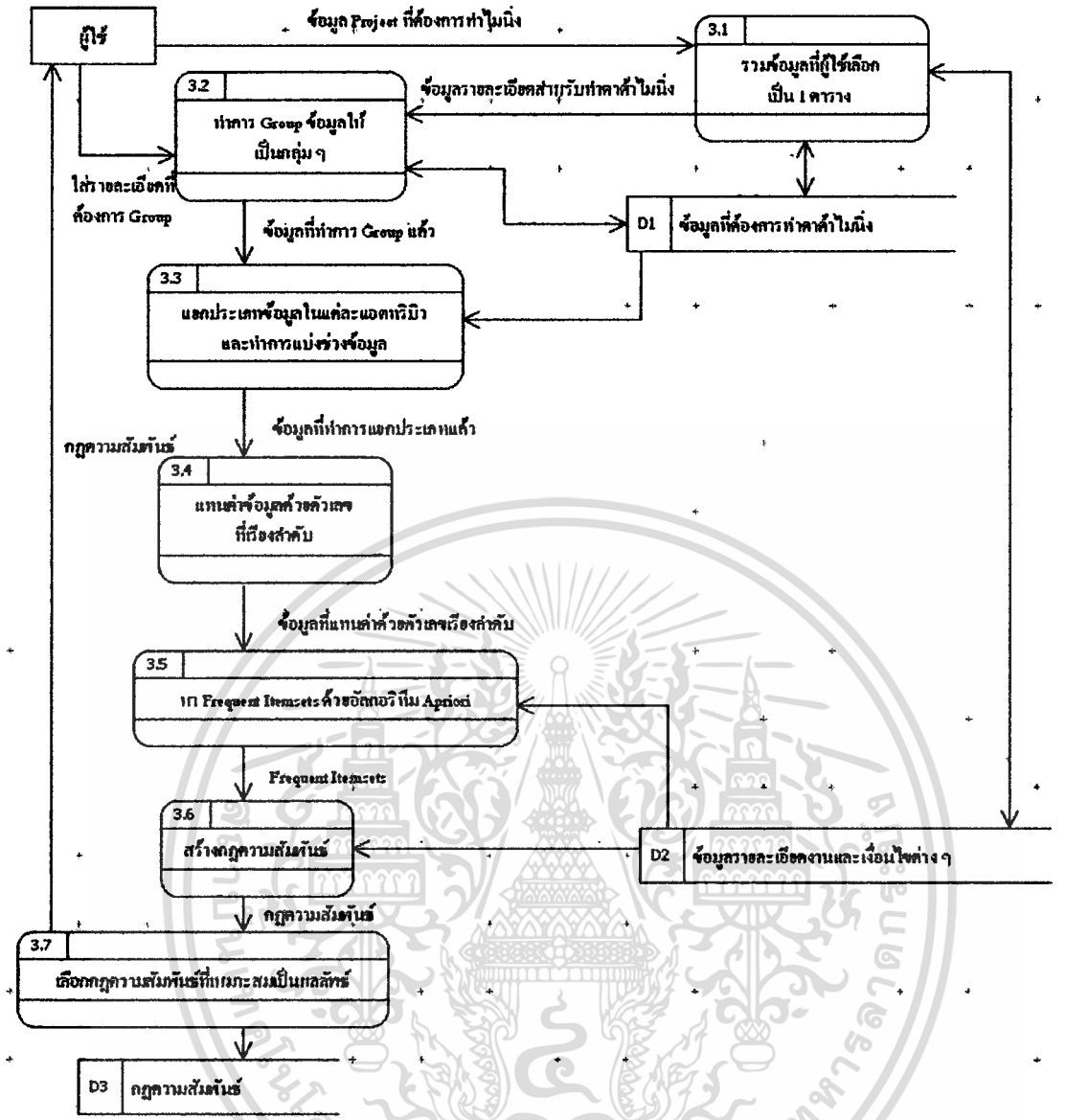
3) แยกประเภทข้อมูลในแต่ละแอตทริบิวและทำการแบ่งช่วงข้อมูล ในขั้นตอนนี้เป็นการเริ่มทำคาต้าไมนิ่ง โดยจะเริ่มแบ่งประเภทของข้อมูล โดยถ้าเป็นข้อมูลเชิงปริมาณก็จะมีการแบ่งช่วงข้อมูล

4) แทนค่าข้อมูลด้วยตัวเลขที่เรียงลำดับ เป็นการแทนค่าข้อมูลให้เรียงกันเพื่อง่ายต่อการนำข้อมูลไปใช้ในอัลกอริทึมที่พัฒนามาจาก Apriori

5) หา Frequent Itemsets ด้วยอัลกอริทึมที่พัฒนามาจาก Apriori เป็นการหากลุ่มข้อมูลที่มีค่ามากกว่าค่า Minimum Support ที่กำหนดไว้

6) สร้างกฎความสัมพันธ์ เป็นกระบวนการนำ Frequent Itemsets มาสร้างกฎความสัมพันธ์

7) เลือกกฎความสัมพันธ์ที่เหมาะสมเป็นผลลัพธ์ โดยมีหลักการเลือกโดยใช้ค่า Interest มาช่วยในการเลือก เมื่อเลือกกฎความสัมพันธ์ได้แล้วก็จะนำกฎความสัมพันธ์ที่ได้เก็บในไว้ระบบและส่งกฎความสัมพันธ์ที่เลือกได้ไปให้ผู้ใช้



รูปที่ 3.3 แสดงแผนภาพกระแสข้อมูลระดับที่ 2 ของกระบวนการการทำไมนิ่งของระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

1) ผู้ใช้เลือกแหล่งที่มาของข้อมูลสำหรับทำดาต้าไมนิ่ง ในส่วนนี้ผู้ใช้จะเลือกฐานข้อมูลที่ต้องการทำดาต้าไมนิ่ง ถ้าเป็นการไมนิ่งโดยใช้ฐานข้อมูล คือระบุ ค่า Connection String หรือถ้าเป็นการไมนิ่งโดยใช้ไฟล์ CSV คือ ตำแหน่งที่วางไฟล์ CSV นั้นไว้

2) เมื่อผู้ใช้เลือกฐานข้อมูลแล้ว โปรแกรมจะเข้าไปอ่านค่าในฐานข้อมูลหรือจากไฟล์ CSV เพื่อนำตารางและแอตทริบิวมาแสดงให้ผู้ใช้เลือก

รูปที่ 3.4 เวิร์กโฟลว์ของกระบวนการการทำไมนิ่งในระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

3) ผู้ใช้เลือกข้อมูลที่ต้องการทำดาต้าไมนิ่ง ในขั้นตอนนี้ผู้ใช้จะทำการเลือกแต่ละแอตทริบิวที่ต้องการนำมาหากฎความสัมพันธ์ โดยหลักการเลือกจะต้องเลือกข้อมูลที่สัมพันธ์กันที่สามารถนำมารวมกันเป็น 1 ตารางได้

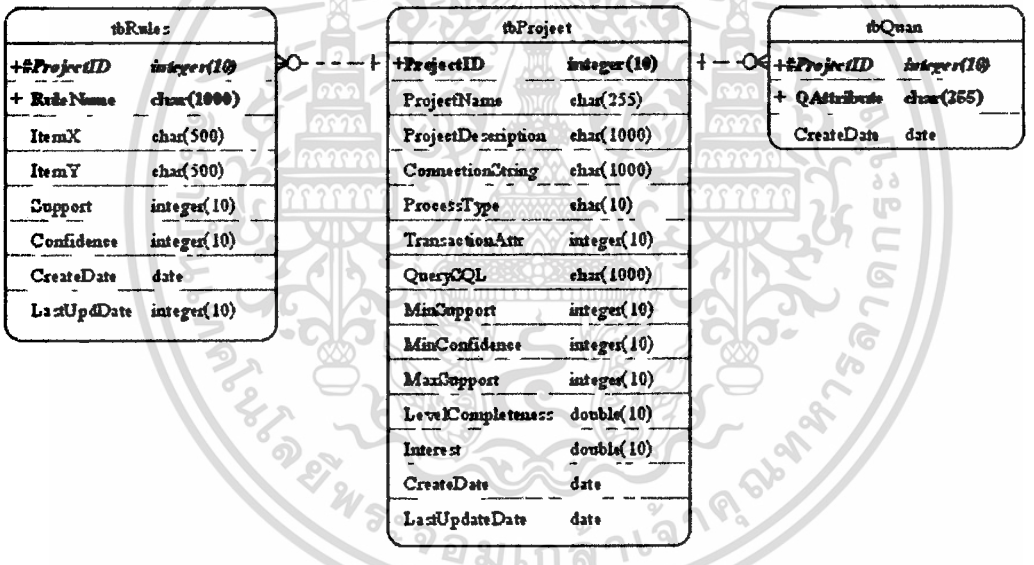
4) ทำการรวมข้อมูลที่ผู้ใช้เลือกทั้งหมดเป็น 1 ตาราง โดยระบบจะนำแต่ละแอตทริบิวที่ผู้ใช้เลือกมารวมกันเป็น 1 ตาราง เพราะในขั้นตอนการทำดาต้าไมนิ่งจะต้องใช้ข้อมูลในลักษณะที่เป็นทรานเซกชั่น อยู่ในตารางเดียว

5) ผู้ใช้กำหนดเงื่อนไขสำหรับทำดาต้าไมนิ่ง กระบวนการนี้ผู้ใช้จะต้องระบุค่าเงื่อนไขต่าง ๆ ที่ต้องนำมาใช้ในการหากฎความสัมพันธ์

6) ระบบค้นหากฎความสัมพันธ์ในขั้นตอนนี้ระบบจะค้นหากฎความสัมพันธ์ ถ้าหาเจอก็จะแสดงและจัดเก็บกฎความสัมพันธ์ที่ได้

7) แต่ถ้าหากฎความสัมพันธ์ไม่ได้ ผู้ใช้ก็จะไปสร้าง Project ใหม่ ทำการกำหนดเงื่อนไขในการหากฎความสัมพันธ์ใหม่อีกครั้ง และค้นหาใหม่อีกครั้ง

3.5 อีอาร์ไดอะแกรม



รูปที่ 3.5 แสดงอีอาร์ไดอะแกรมของระบบดาต้าไมนิ่งโดยใช้กฎความสัมพันธ์ของข้อมูลเชิงปริมาณ

โดยแต่ละตารางสามารถอธิบายรายละเอียดได้ ดังนี้

ตารางที่ 3.1 โครงสร้างของตารางที่ใช้เก็บข้อมูลที่มีลักษณะเชิงปริมาณ

ชื่อตาราง : tbQuan **ใช้สำหรับ :** เก็บข้อมูลที่มีลักษณะเชิงปริมาณ				
ชื่อฟิลด์	**รายละเอียด**	**ประเภท**	**ชนิดของคีย์**	**ตาราง ที่อ้างถึง**
ProjectID	รหัสของงาน	Integer	PK	tbProject
QAttribute	ชื่อของ Attribute ในข้อมูลหลักที่เป็นข้อมูลเชิงปริมาณ	Char(255)	PK	
CreateDate	วันเวลาที่บันทึกข้อมูล	Date		

ตารางที่ 3.2 โครงสร้างของตารางที่ใช้เก็บข้อมูลกฎความสัมพันธ์ที่ได้

ชื่อตาราง : tbRules **ใช้สำหรับ :** เก็บข้อมูลกฎความสัมพันธ์ที่ได้				
ชื่อฟิลด์	**รายละเอียด**	**ประเภท**	**ชนิดของคีย์**	**ตาราง ที่อ้างถึง**
ProjectID	รหัสของงาน	Integer	PK	tbProject
RuleName	ชื่อกฎความสัมพันธ์ที่ได้	Char(1000)	PK	
ItemX	ค่าทางด้านซ้ายของกฎความสัมพันธ์	Char(500)		
ItemY	ค่าทางด้านขวาของกฎความสัมพันธ์	Char(500)		
Support	ค่า Supportของกฎความสัมพันธ์	Integer		
Confidence	ค่า Confidence ของกฎความสัมพันธ์	Integer		
CreateDate	วันเวลาที่บันทึกข้อมูล	Date		
LastUpdateDate	วันเวลาที่แก้ไขข้อมูล	Date		

ตารางที่ 3.3 โครงสร้างของตารางที่ใช้เก็บรายละเอียดของงาน

ชื่อตาราง : tbProject **ใช้สำหรับ :** เก็บข้อมูลงานที่ทำการวิเคราะห์ รวมถึงรายละเอียดของงานที่ทำการวิเคราะห์				
ชื่อฟิลด์	**รายละเอียด**	**ประเภท**	**ชนิดของคีย์**	**ตาราง ที่อ้างถึง**
ProjectID	รหัสของงาน	Integer	PK	-
ProjectName	ชื่อของงานที่ทำการไมนิ่ง	Char(255)		
ProjectDescription	รายละเอียดของงาน	Char(1000)		
ConnectionString	แหล่งที่อยู่ของฐานข้อมูลที่ทำการไมนิ่ง	Char(1000)		
ProcessType	ชนิดของงาน ซึ่งมี 2 ประเภท คือ 1. ข้อมูลจาก SQL 2. ข้อมูลจากไฟล์ CSV	Char(10)		
TransactionAttr	อ้างอิงถึง Attribute ซึ่งเก็บรหัสข้อมูลที่ระบุถึงข้อมูลที่เกิดขึ้นในครั้งเดียวกัน	Integer		
MinSupport	ค่า Minimum Support ของงานไมนิ่ง	Integer		
MinConfidence	ค่า Minimum Confidence ของงานไมนิ่ง	Integer		
MinSupport	ค่า Support ที่มากสุดที่สามารถรวมช่วงได้	Double		
LevelCompleteNess	ค่าที่ใช้ในการแบ่งช่วงข้อมูลเชิงปริมาณ	Double		
Interest	ค่า Interest ของงานไมนิ่ง	Integer		
CreateDate	วันเวลาที่บันทึกข้อมูล	Date		
LastUpdateDate	วันเวลาที่แก้ไขข้อมูล	Date		

บทที่ 4
การพัฒนาระบบ

4.1 การจัดเตรียมสภาพแวดล้อมในการพัฒนา

ในการพัฒนาระบบการจัดเตรียมสภาพแวดล้อมให้พร้อมสำหรับการพัฒนา โดยการติดตั้ง Window XP หลังจากนั้นจึงติดตั้ง Visual Studio .Net 2005 สำหรับทำการพัฒนาโปรแกรม นอกจากนี้เพื่อให้สะดวกในการพัฒนาการติดตั้ง MSDN เพื่อช่วยค้นหาข้อมูลต่างๆในการพัฒนา งานฐานข้อมูลจะเก็บข้อมูลโดยใช้ MS Access 2000 ในการจัดเก็บ และส่วนที่เป็น Input ของระบบจะใช้ฐานข้อมูล MS SQL Server 2005 หรือไฟล์ CSV

4.2 ขั้นตอนการพัฒนาระบบ

ขั้นตอนในส่วนของพัฒนาระบบ มีลำดับการพัฒนา ดังนี้

1. พัฒนาในส่วนของการเพิ่ม Project เข้ามาในระบบ ซึ่ง Project คืองาน 1 งานที่ต้องการทำไมนิ่ง โดยในส่วนนี้จะสามารถใส่ค่าเงื่อนไขต่าง ๆ ที่ต้องการทำไมนิ่ง รวมทั้งข้อมูล Input ของระบบที่เป็นฐานข้อมูลหรือไฟล์ CSV

2. พัฒนาในส่วนของการใช้อัลกอริทึมที่ดัดแปลงมาจาก Apriori มาทำการหากฎความสัมพันธ์ของข้อมูลและมีการจัดเก็บกฎความสัมพันธ์ที่ได้ไว้ในระบบ ซึ่งรายละเอียดการทำงานของอัลกอริทึม มีดังนี้

   2.1 สร้างช่วงข้อมูลสำหรับ แอตทริบิว ที่เป็นข้อมูลเชิงปริมาณ

   2.2 แทนค่าข้อมูลด้วยตัวเลขที่เรียงลำดับ

   2.3 หา Frequent Itemsets

   2.4 สร้างกฎความสัมพันธ์

   2.5 เลือกกฎความสัมพันธ์ที่เหมาะสมเป็นผลลัพธ์

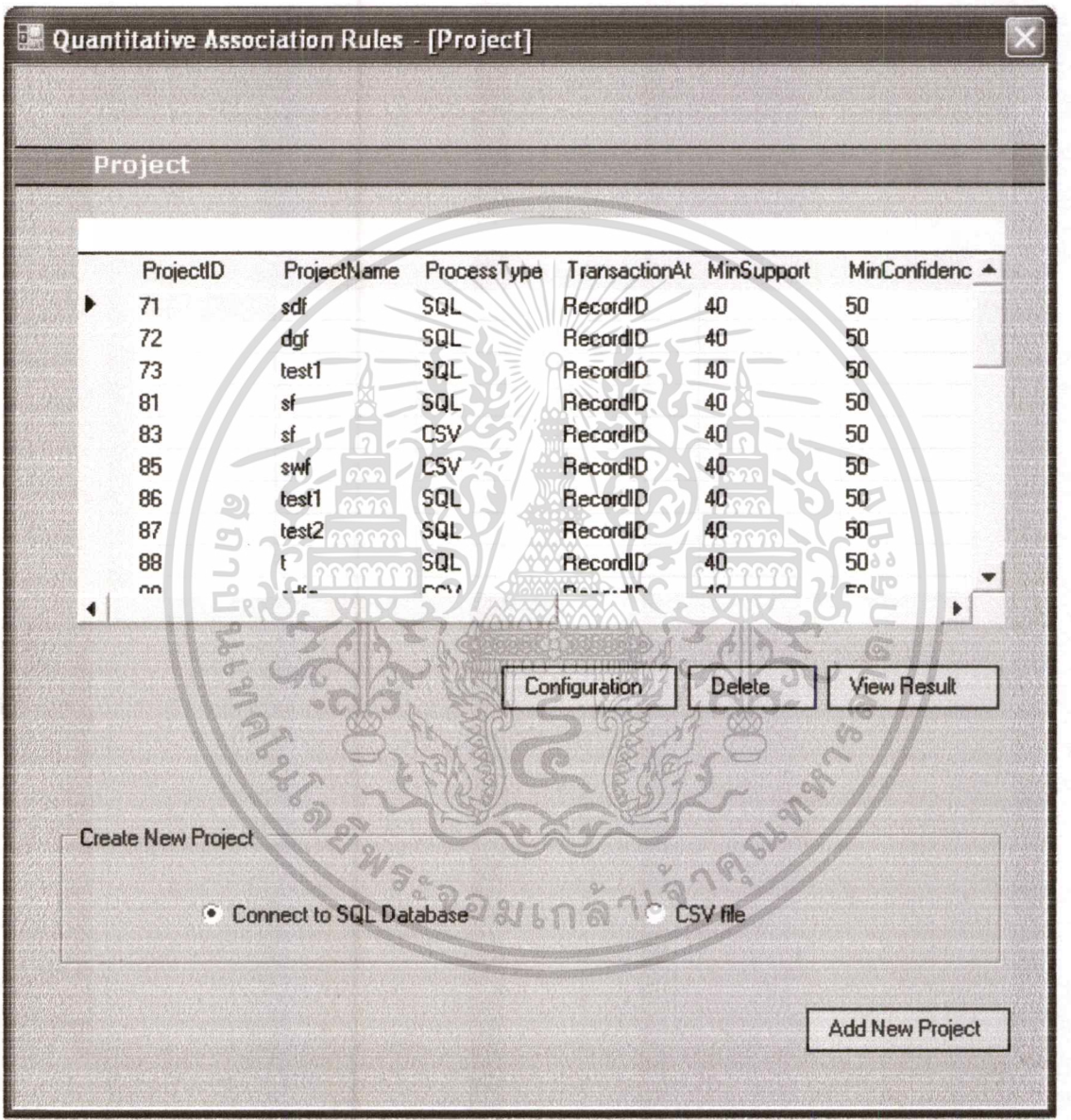
3. พัฒนาในส่วนของการเรียกดูข้อมูล Project ที่ได้เคยค้นหากฎความสัมพันธ์มาแล้ว โดยสามารถดูค่าเงื่อนไขต่าง ๆ ที่กำหนดไว้ ส่วนของการ Input ข้อมูล และกฎความสัมพันธ์ที่ได้

4.3 การพัฒนาระบบและหน้าจอการทำงาน

จากการวิเคราะห์และออกแบบระบบงานจะแบ่งส่วนของหน้าจอการทำงานต่างๆได้ดังนี้

1. หน้าหลักของระบบ ซึ่งจะแสดงรายละเอียดเกี่ยวกับงานที่จะทำการวิเคราะห์ และการกำหนดข้อมูลเริ่มต้น แบ่งออกเป็น 2 ส่วน ดังรูปที่ 4.1 คือ

- Project List คือส่วนของการแสดงงานที่ได้กำหนดรายละเอียดของงานไว้แล้วหรืออาจทำการไมนิ่ง และทราบผลลัพท์แล้ว รายละเอียดของข้อมูลที่แสดงในตารางเกี่ยวกับงานมีดังนี้



รูปที่ 4.1 แสดงหน้าจอหลักของโปรแกรม

o Project ID แสดงรหัสของงานที่ได้ทำการบันทึกไว้

o ProjectName แสดงชื่อของงานที่ได้ทำการบันทึกไว้

o ProjectType แสดงประเภทของงาน ซึ่งมี 2 ประเภท คือ SQL, CSV เพื่อบ่งบอกที่มาของฐานข้อมูลที่นำมาวิเคราะห์ในครั้งนี้ ผู้ใช้งานสามารถเปิดงานที่ต้องการโดยการเลือกที่ชื่องานนั้น และคลิ๊กที่ปุ่มต่างๆ เพื่อทำงานดังนี้

o TransactionAttribute แสดงแอตทริบิวที่เป็น Transaction ID ของงานที่เคยได้ทำการบันทึกไว้

o MinSupport แสดงค่า Minimum Support ที่เคยได้ทำการกำหนดไว้

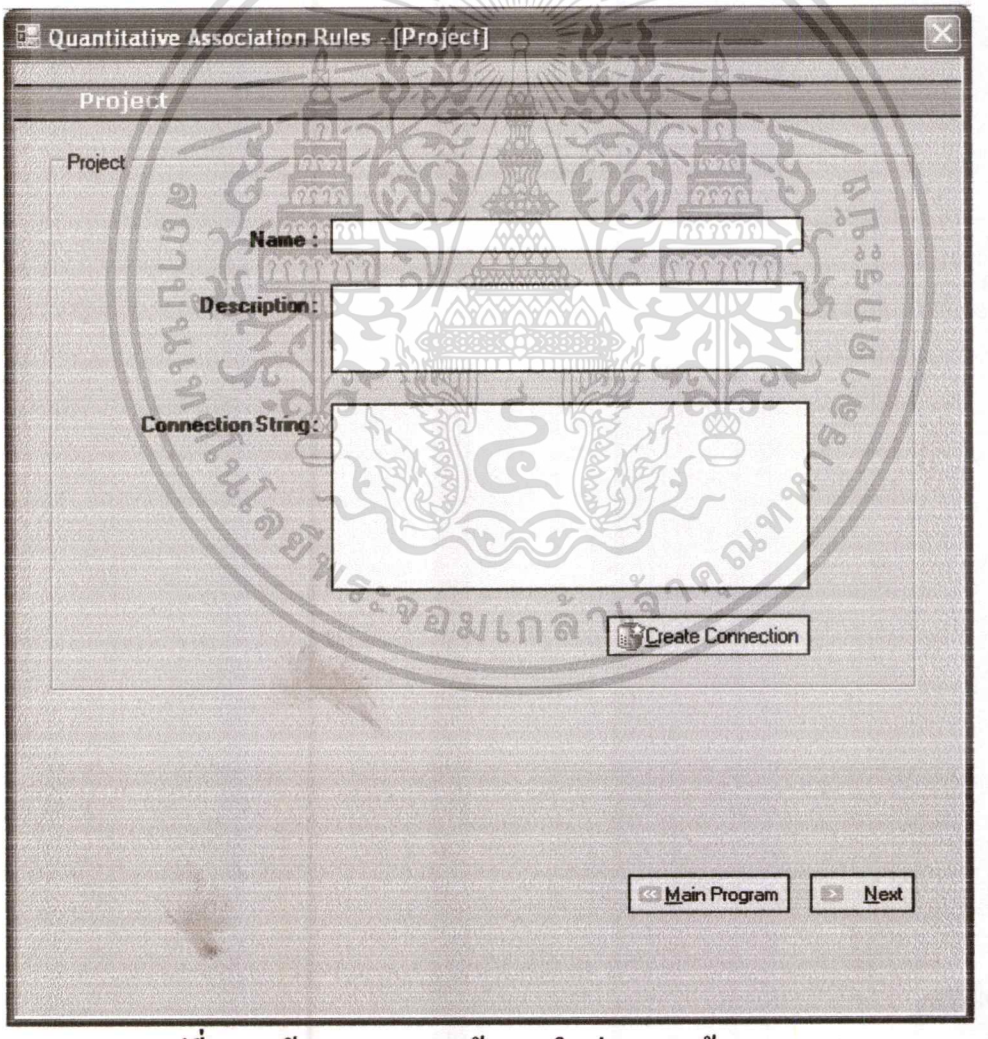
o MinConfidence แสดงค่า Minimum Confidence ที่เคยได้ทำการกำหนดไว้

o ProjectDesc แสดงรายละเอียดของงานที่ผู้ใช้งานสามารถบันทึกเพิ่มเติมเพื่ออธิบายเกี่ยวกับลักษณะของงานได้

o ปุ่ม Configuration เพื่อทำการดูข้อมูลที่เคยได้บันทึกไว้

o ปุ่ม Delete เพื่อทำการลบงาน และผลการทำไมนิ่งออกจากระบบ

o ปุ่ม View Result จะทำการแสดงผลของการทำไมนิ่งของงานที่กำลังเลือกอยู่ โดยระบบจะดึงผลการไมนิ่งจากการทำไมนิ่งในครั้งสุดท้ายของงานชิ้นนั้นขึ้นมาแสดง



รูปที่ 4.2 หน้าจอแสดงการสร้างงานใหม่จากฐานข้อมูล SQL

- Create New Project คือส่วนสำหรับสร้างงานใหม่ ซึ่งผู้ใช้งานจะต้องทำการระบุชนิดของงานที่ต้องการสร้างก่อน ซึ่งมีด้วยกัน 2 ประเภท คือ SQL, CSV ตามแหล่งที่มาของข้อมูล และ

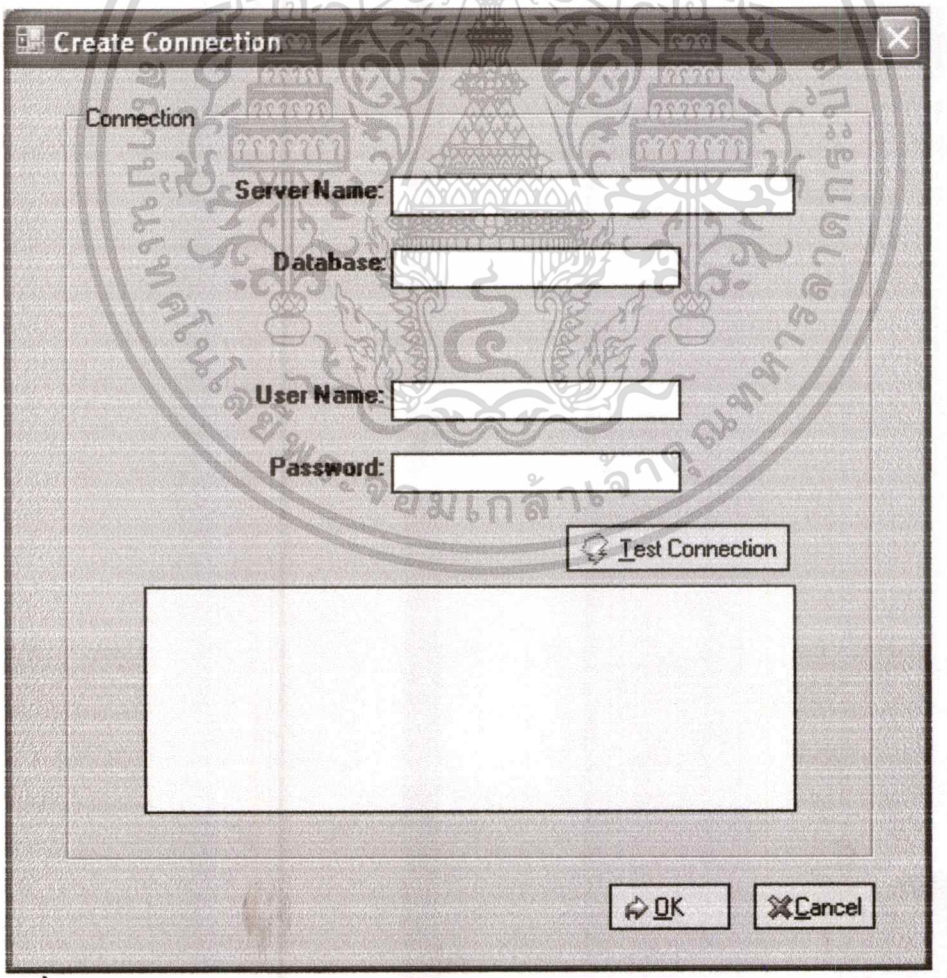
หลังจากนั้นจึงทำการคลิ๊กที่ปุ่ม Add New Project ระบบจะนำเข้าสู้ขั้นตอนในการสร้างงานใหม่ต่อไป

2. การสร้างงานใหม่จากฐานข้อมูล SQL หากผู้ใช้งานต้องการสร้างงานใหม่โดยใช้แหล่งที่มาของข้อมูลจากฐานข้อมูลของ SQL หลังจากผู้ใช้งานคลิ๊กปุ่ม Create New Project แล้ว ระบบจะแสดงหน้าจอให้ผู้ใช้งานกำหนดรายละเอียดเกี่ยวกับงาน ตามรูปประกอบที่ 4.2 ซึ่งมีรายละเอียดดังนี้

o Name ชื่อของงานที่ต้องการระบบ

o Description รายละเอียดของงาน เพื่อช่วยอธิบายรายละเอียดของงานที่จะทำการไมนิ่ง

o Connection String คือส่วนของการกำหนดส่วนเชื่อมต่อกับฐานข้อมูล SQL โดยผู้ใช้งานสามารถกำหนดโดยการคลิ๊กที่ปุ่ม Create Connection ซึ่งจะปรากฏหน้าจอดังรูปประกอบที่ 4.3 ซึ่งมีรายละเอียดดังนี้



รูปที่ 4.3 หน้าจอแสดงการสร้าง Connection String หรือการติดต่อไปยังฐานข้อมูล

o Server Name ชื่อ Server ที่ข้อมูลเก็บอยู่

o Database Name ชื่อของฐานข้อมูลที่ต้องการนำมาใช้ในการทำไมนิ่ง

o User Name ชื่อของผู้ใช้งานที่มีสิทธิเข้าใช้งานในฐานข้อมูลดังกล่าว

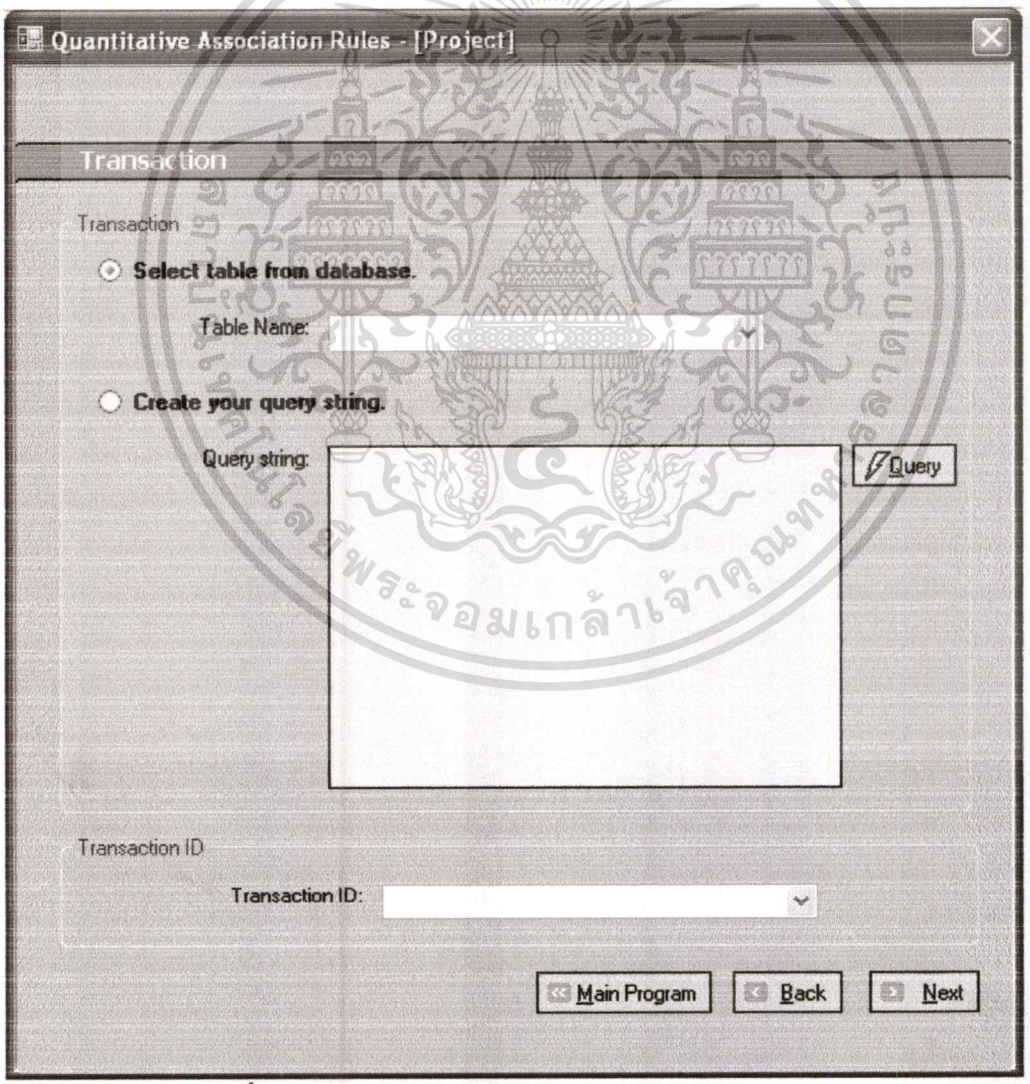
o Password รหัสผ่านที่ใช้สำหรับผู้ใช้งานที่จะเข้าไปยังฐานข้อมูลได้

- ปุ่ม Test Connection ใช้เพื่อทำการทดสอบการใช้งานว่าสามารถติดต่อไปยังฐานข้อมูลดังกล่าวได้หรือไม่

- ปุ่ม Main Program เพื่อให้ผู้ใช้งานกลับไปยังหน้าแรกของระบบ เมื่อผู้ใช้งานไม่ต้องการจะทำส่วนของการสร้างงานใหม่ต่อไป

- ปุ่ม Next เมื่อผู้ใช้งานทำการกำหนดรายละเอียดครบถ้วนแล้ว ก็จะทำการคลิ๊กที่ปุ่มนี้เพื่อเข้าสู่ขั้นตอนในลำดับต่อไป

3. การกำหนดส่วนของ Transaction เมื่อผู้ใช้งานผ่านในส่วนของการกำหนดการติดต่อกับฐานข้อมูลแล้วจะเข้าสู้หน้าจอในส่วนของ Transaction นั้นคือ ผู้ใช้งานจะต้องระบุราย



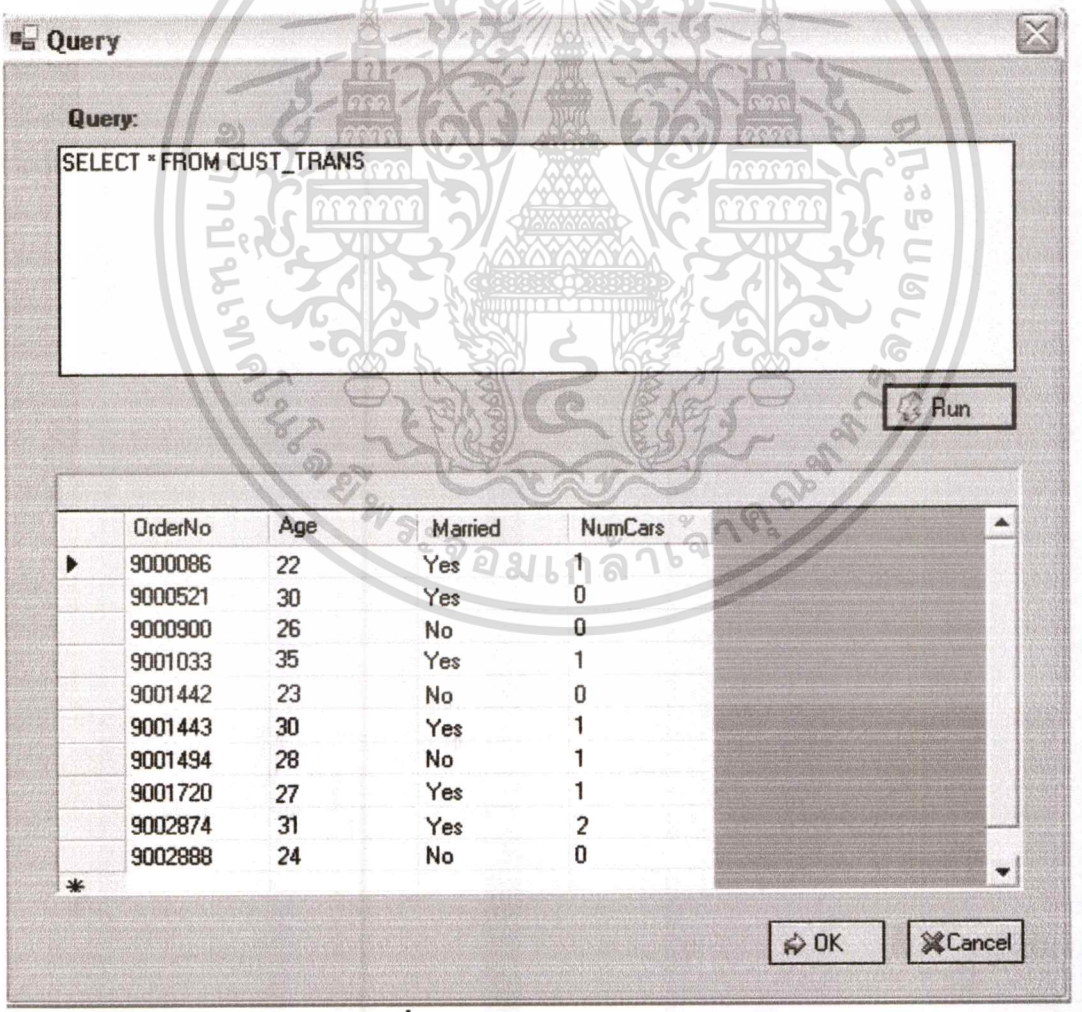
รูปที่ 4.4 หน้าจอแสดงการกำหนดส่วนของ Transaction

ละเอียดของตารางที่จะนำข้อมูลมาวิเคราะห์โดยตารางที่จะนำมาเป็น Transaction คือ ข้อมูลหลักที่จะใช้นั้นเอง เช่นถ้าเป็นการขายสินค้าคือข้อมูลในส่วนของบันทึกการขายสินค้า เป็นต้น ดังรูปประกอบที่ 4.4 ในหน้าจอนี้จะมีรายละเอียด ดังนี้

- Select table from database ถ้าผู้ใช้มีข้อมูลที่สมบูรณ์เก็บไว้ให้ตารางใดตารางหนึ่ง ระบบจะแสดงรายการตารางจากในฐานข้อมูลทั้งหมดมาให้ ผู้ใช้จะต้องเลือกว่าจะเลือกจากตารางที่มีอยู่แล้ว และทำการเลือกชื่อตารางในรายการที่แสดงไว้ให้

- Transaction ID ระบบข้อมูลซึ่งใช้เก็บรหัสของชุดข้อมูล

- Create your query string หากผู้ใช้ต้องการกำหนดรายละเอียดของข้อมูลซึ่งมาจากการผนวกตารางมากกว่าหนึ่งตารางไว้ด้วยกัน ผู้ใช้สามารถเลือกในการกำหนดแบบ Create query string และหลังจากนั้นผู้ใช้งานต้องคลิ๊กที่ปุ่ม Query เพื่อทำการสร้าง Query ที่ต้องการ จะปรากฏหน้าจอเพื่อทำการสร้าง Query ดังรูปประกอบที่ 4.5 ซึ่งมีรายละเอียดการทำงานดังนี้



รูปที่ 4.5 การสร้าง Query String

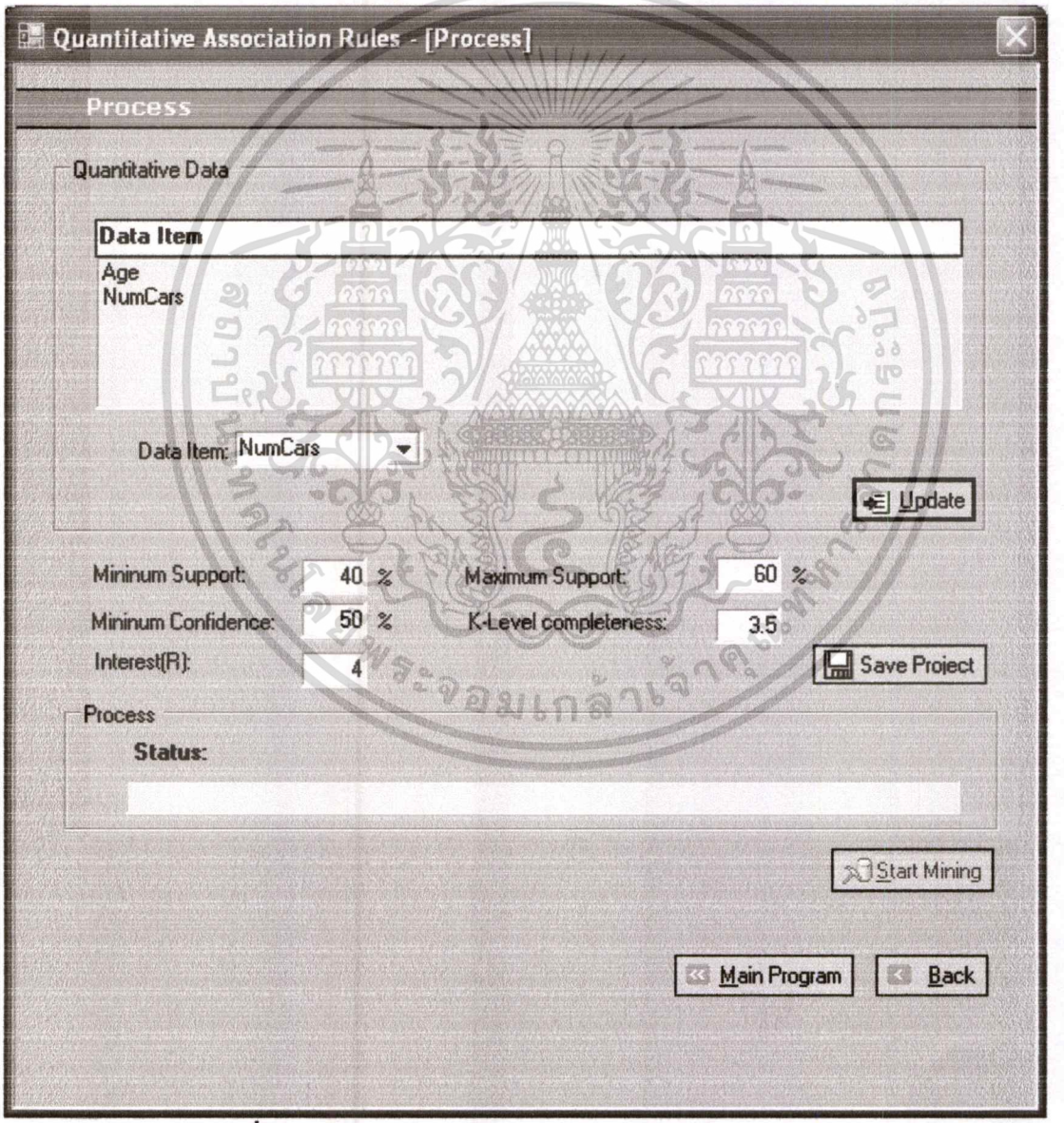
- Query คือส่วนที่ผู้ใช้งานกำหนด Query String ตามรูปแบบของ SQL ลงไป และหลังจากนั้นคลิ๊กยังปุ่ม Run เพื่อทำการประมวลผล และตรวจสอบความถูกต้องของ Query ที่กำหนดเมื่อ Query ที่กำหนดถูกต้อง จึงคลิ๊กปุ่ม OK เพื่อกลับเข้าสู่หน้าจอ Transaction ต่อไป

- ปุ่ม Main Program เพื่อให้ผู้ใช้งานกลับไปยังหน้าแรกของระบบ เมื่อผู้ใช้งานไม่ต้องการจะทำส่วนของการสร้างงานใหม่ต่อไป

- ปุ่ม Back เมื่อผู้ใช้งานต้องการกลับไปยังหน้าจอก่อนหน้านี้

- ปุ่ม Next เมื่อผู้ใช้งานทำการกำหนดรายละเอียดครบถ้วนแล้ว ก็จะทำการคลิ๊กที่ปุ่มนี้เพื่อเข้าสู่ขั้นตอนในลำดับต่อไป

4. เลือก Attribute ที่เป็นข้อมูลเชิงปริมาณ และทำการกำหนดค่าต่าง ๆ ตามรูปที่ 4.6 ดังนี้



รูปที่ 4.6 หน้าจอแสดงการกำหนดค่าต่าง ๆ สำหรับ Project

- Attribute เลือก Attribute ที่ต้องการระบุเป็นข้อมูลเชิงปริมาณ

- K-Level Completeness คือ ค่าตัวเลขที่ใช้ในการแบ่งช่วงข้อมูล ยิ่งมีค่ามากก็จะแบ่งได้ช่วงน้อย

- Maximum support คือค่าที่ใช้ในการรวมข้อมูลที่แบ่งแล้ว

- ปุ่ม Update เมื่อต้องการ กำหนดค่า Atribute ที่ระบุเป็นข้อมูลเชิงปริมาณ

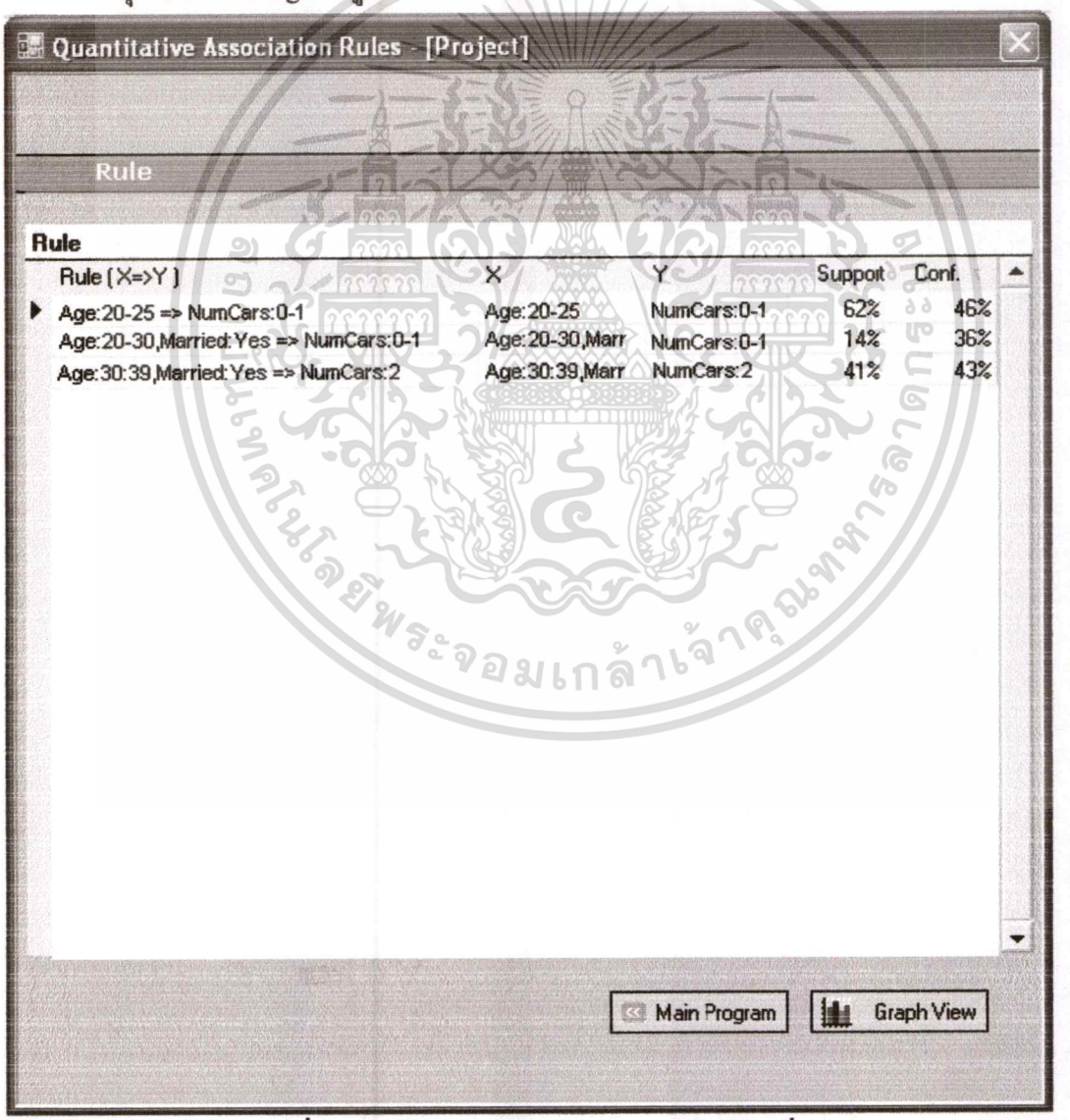
- Minimum support คือค่า Minimum support

- Minimum confidence คือค่า Minimum confidence

- Interert(R) คือค่าที่ใช้ในการเลือกกฎความสัมพันธ์มาเป็นผลลัพธ์

- ปุ่ม Save Project เมื่อทำการเปลี่ยนแปลงเรียบร้อยแล้ว และต้องการทำการไมนิ่ง ให้ผู้ใช้งานคลิ๊กปุ่ม Save Project เพื่อทำการบันทึก

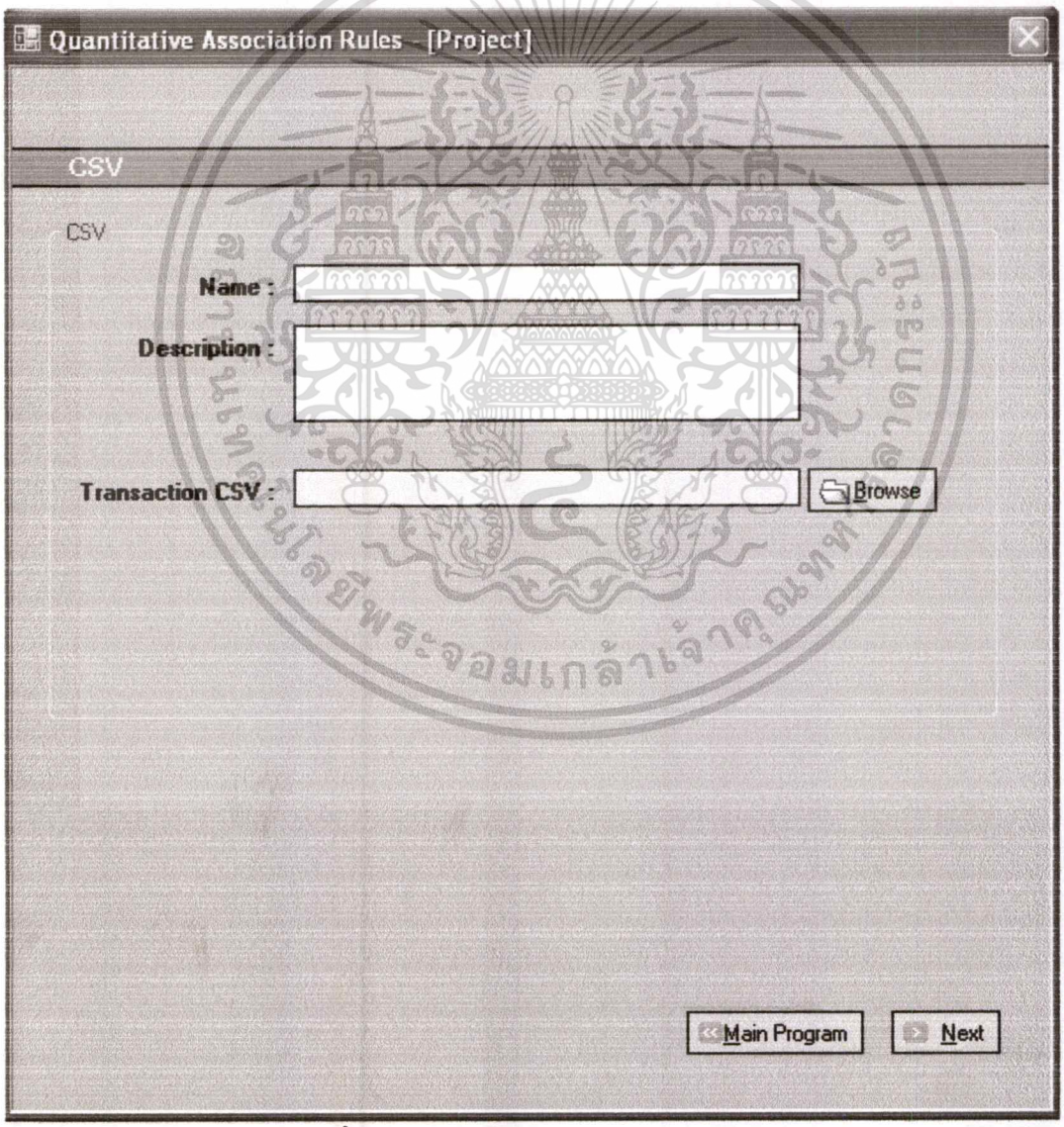
- ปุ่ม Start Mining เมื่อผู้ใช้ต้องการหาผลลัพธ์จากการไมนิ่ง



รูปที่ 4.7 แสดงผลลัพธ์ความสัมพันธ์ของข้อมูลที่ได้

5. แสดงผลการทำไมนิ่ง เมื่อทำการไมนิ่งแล้วระบบจะแสดงผลลัพท์หรือกฎที่ถูกพิจารณาว่าอยู่ในเกณฑ์ที่ผู้ใช้งานสนใจโดยเทียบจาก Minimum Confidence ดังรูปที่ 4.7 โดยข้อมูลต่างๆที่แสดงมีรายละเอียดดังนี้

- Rule (X=>Y) แสดงผลในรูปแบบของความสัมพันธ์ของ Association Rule ในรูปแบบของ Item X ที่มีผลต่อ Item Y เช่น การซื้อสินค้า X แล้วมีโอกาสในการซื้อสินค้า Y เป็นต้น
- X คือ Item X ในกฎเช่นการซื้อขนมปังมีผลในการซื้อนม X จะแทนขนมปัง
- Y คือ Item Y ในกฎเช่นการซื้อขนมปังมีผลในการซื้อนม Y จะแทนนม
- Support คือ ค่า Support หรือความน่าจะเป็นของการเกิด Item X ร่วมกับ Item Y
- Conf. คือ ค่า Confidence หรือค่าความมั่นใจของกฎนี้ ซึ่งหากมีค่าสูงหมายถึงกฎนี้มีความน่าเชื่อถือมาก



รูปที่ 4.8 แสดงการสร้างงานใหม่จาก CSV

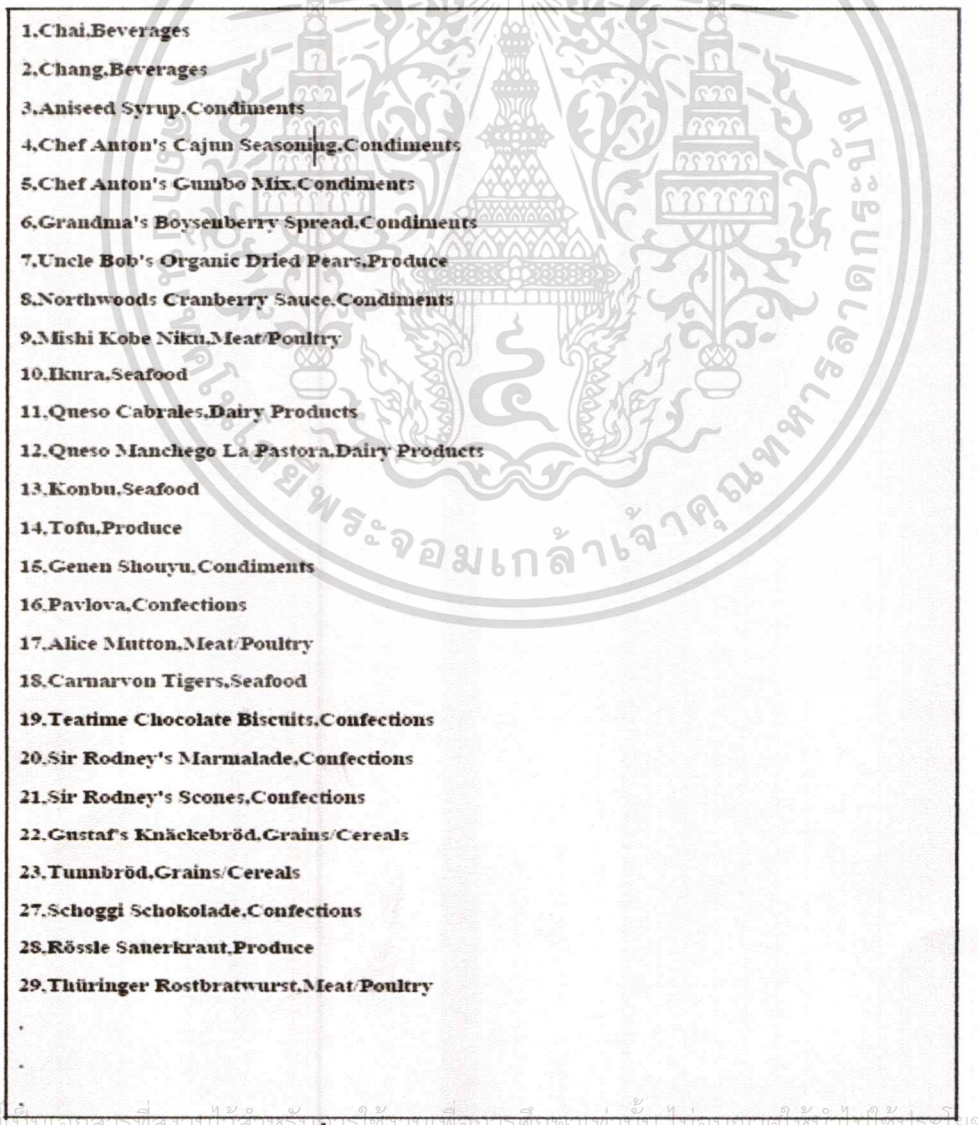
6. การสร้างงานใหม่จากเอกสาร CSV หากผู้ใช้งานต้องการสร้างงานใหม่โดยใช้แหล่งที่มาจากเอกสาร CSV หลังจากผู้ใช้งานคลิ๊กปุ่ม Create New Project จากหน้าจอหลักแล้ว ระบบจะแสดงหน้าจอให้ผู้ใช้งานกำหนดรายละเอียดเกี่ยวกับงาน ตามรูปประกอบที่ 4.9 ซึ่งมีรายละเอียดดังนี้

- Name ชื่อของงานที่ต้องการระบบ

- Description รายละเอียดของงาน เพื่อช่วยอธิบายรายละเอียดของงานที่จะทำการไมนิ่ง

- Transaction CSV คือส่วนของเอกสาร CSV ที่ต้องการใช้ในการทำไมนิ่ง โดยผู้ใช้งานคลิ๊กที่ปุ่ม Browse เพื่อทำการเลือกไฟล์เอกสารหลักที่ต้องการทำไมนิ่ง

โดยในบรรทัดแรกของไฟล์ CSV จะต้องกำหนดชื่อของข้อมูล(ซึ่งเหมือนกับ Attribute ใน Relation) ซึ่งค่าแรกจะหมายถึง TrasactionID ในไฟล์จะต้องมีตัวคั่นเป็นคอมม่า(,) รูปแบบของเอกสาร CSV ต้องมีลักษณะดังรูปที่ 4.9


1,Chai,Beverages
2,Chang,Beverages
3,Aniseed Syrup,Condiments
4,Chef Anton's Cajun Seasoning,Condiments
5,Chef Anton's Gumbo Mix,Condiments
6,Grandma's Boysenberry Spread,Condiments
7,Uncle Bob's Organic Dried Pears,Produce
8,Northwoods Cranberry Sauce,Condiments
9,Mishi Kobe Niku,Meat/Poultry
10,Ikura,Seafood
11,Queso Cabrales,Dairy Products
12,Queso Manchego La Pastora,Dairy Products
13,Konbu,Seafood
14,Tofu,Produce
15,Genen Shouyu,Condiments
16,Pavlova,Confections
17,Alice Mutton,Meat/Poultry
18,Carnarvon Tigers,Seafood
19,Teatime Chocolate Biscuits,Confections
20,Sir Rodney's Marmalade,Confections
21,Sir Rodney's Scones,Confections
22,Gustaf's Knäckebröd,Grains/Cereals
23,Tunnbröd,Grains/Cereals
27,Schoggi Schokolade,Confections
28,Rössle Sauerkraut,Produce
29,Thüringer Rostbratwurst,Meat/Poultry
.
.
.

รูปที่ 4.9 ตัวอย่างข้อมูลในไฟล์ CSV

บทที่ 5
บทสรุป

บทนี้เป็นการสรุปภาพโดยรวมของการศึกษาและพัฒนาระบบจัดเก็บและวิเคราะห์ข้อมูล รวมทั้งข้อเสนอแนะในการพัฒนาระบบเพิ่มเติม โดยสรุปได้เป็นหัวข้อ ดังนี้

5.1 ผลการวิเคราะห์และออกแบบระบบการทำคาต้าไมนิ่งแบบ Quantitative Association Rules

การพัฒนาระบบสำหรับการทำคาต้าไมนิ่งในแบบ Quantitative Association Rules ช่วยให้เพิ่มประสิทธิภาพการหาความสัมพันธ์ข้อมูลสำหรับข้อมูลเชิงปริมาณไม่ให้หายไปด้วยการสร้างเป็นช่วง ซึ่งจะทำให้มีค่า Support เพิ่มมากขึ้น แต่ทั้งนี้การที่จะสร้างข้อมูลเป็นช่วงนั้นจะต้องมีการหาว่าจะแบ่งข้อมูลเป็นทั้งหมดกี่ช่วง ก็ขึ้นอยู่กับค่า K-Level Completeness โดยยิ่งมีค่าน้อยเท่าไหร่ ก็จะทำให้ข้อมูลหายน้อยมากเท่านั้น แต่ถ้าหากใส่ค่าน้อยเกินไปจะทำให้มีจำนวนการแบ่งช่วงมาก ทำให้ต้องใช้เวลาในการทำงานนาน โดยโปรแกรมมีการรองรับการทำงานร่วมกับฐานข้อมูลต่างๆ โดยเฉพาะฐานข้อมูลซึ่งมีการใช้กันมากเช่น ฐานข้อมูล SQL และสนับสนุนการใช้งานร่วมกับเอกสาร CSV ทำให้ผู้ใช้งานสามารถนำเข้าข้อมูลจากระบบอื่นๆ นอกเหนือจากฐานข้อมูลแบบ SQL เพียงอย่างเดียว ซึ่งการนำเข้าจากเอกสาร CSV มีการใช้กันโดยทั่วไปในหลายๆระบบ ซึ่งเพิ่มความยืดหยุ่นสำหรับผู้ใช้งานมากยิ่งขึ้น

5.2 ประโยชน์ที่คาดว่าจะได้รับ

- ผู้ใช้งานสามารถทำการวิเคราะห์ข้อมูลจากการทำคาต้าไมนิ่งในแบบ Association Rule ได้จากฐานข้อมูลในระบบต่างๆที่ผู้ใช้จัดเก็บข้อมูลอยู่ โดยไม่จำเป็นต้องทำการส่งออกข้อมูลหรือเปลี่ยนแปลงรูปแบบ ทำได้การทำคาต้าไมนิ่งได้สะดวกมากขึ้น
- สามารถวิเคราะห์ข้อมูลจากการคาต้าไมนิ่งได้ในข้อมูลเชิงปริมาณได้มีคุณภาพมากกว่าการหาความสัมพันธ์ของข้อมูลแบบธรรมดา
- เป็นแนวทางในการนำข้อมูลที่มีอยู่ในระบบมาใช้งานให้เกิดประสิทธิภาพในด้านการวิเคราะห์และนำผลลัพธ์เพื่อใช้ในการปรับปรุงเปลี่ยนแปลง หรือสร้างสิ่งใหม่ๆซึ่งก่อให้เกิดประโยชน์แก่องค์กรต่อไป

5.3 แนวทางการพัฒนาระบบเพิ่มเติม

• พัฒนาเพิ่มเติมในส่วนของการสามารถนำผลลัพธ์ที่ได้มารวมกับข้อมูลใหม่ให้สามารถแสดงความสัมพันธ์ขึ้นมาใหม่ได้

• พัฒนาเพิ่มเติมในส่วนของรับข้อมูล Input จากฐานข้อมูลเชิงสัมพันธ์ชนิดอื่นนอกเหนือจาก SQL Server ได้

• พัฒนาเพิ่มเติมในส่วนของการหากฎความสัมพันธ์โดยให้สามารถหากฎความสัมพันธ์จากข้อมูล Input ได้มากกว่า 3 แอตทริบิว

• เพิ่มประสิทธิภาพในการทำงานเพื่อให้การหาผลลัพธ์ทำได้รวดเร็วยิ่งขึ้น โดยการวิเคราะห์จากอัลกอริทึมอื่นๆที่เกี่ยวข้องกับ Quantitative Association Rules ในการพัฒนาเปรียบเทียบต่อไป

บรรณานุกรม

กิติ ภักดีวัฒนะกุลและพนิดา พานิชกุล. 2548. **คัมภีร์วิเคราะห์และออกแบบระบบ.** พิมพ์ครั้งที่ 4. กรุงเทพฯ : สำนักพิมพ์ เคทีพี.

สุรสิทธิ์ คิวประสพศักดิ์และนันทนี แขวงโสภา. 2546. **อินไซท์ Visual Basic .NET.** พิมพ์ครั้งที่ 1. กรุงเทพฯ:โปรวิชั่น.

โอภาส เอี่ยมสิริวงศ์. 2548. **การวิเคราะห์และออกแบบระบบ.** พิมพ์ครั้งที่ 1. กรุงเทพฯ:ซีเอ็ดยูเคชั่น.

Dubois, D. and H. Prade. 2003. "A note on quality measures for fuzzy association rules" 21-32. in LNAI, **10th International Fuzzy Systems Association World Congress.** Istambul: Springer-Verlag.

Ramakrishnan Srikant and Rakesh Agrawal. 1996. "Mining Quantitative Association Rules in Large Relational Tables" **In Proceedings of the ACM SIGMOD Conference on Management of Data.** Montreal:ACM.

Rantzau, Ralf. 1997. "Extended Concepts for Association Rule Discovery" **Diploma Thesis of University of Stuttgart.**

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวช่อผกา มงคลสถิตย์พร
วุฒิการศึกษาระดับปริญญาตรี	วท.บ. (วิทยาการคอมพิวเตอร์)
	คณะวิทยาศาสตร์
	มหาวิทยาลัยสงขลานครินทร์
การทำงาน	บริษัท จีเอเบิล จำกัด