ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

ระบบค้นหาความสัมพันธ์โดยใช้อพริออริอัลกอริทึม

Association Rules Discovery System by Apriori Algorithms

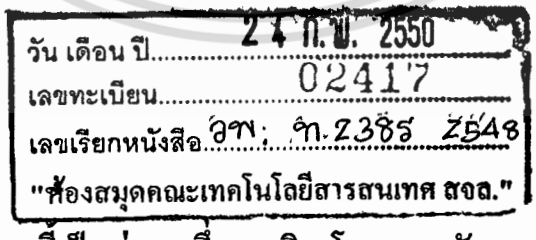
โดย

ทศพล คงน้อย

รหัสประจำตัว 46066231

อาจารย์ที่ปรึกษา

ผศ.ดร. วรพจน์ กรีสุระเดช



b11711206
i1285799x

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบ

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2548

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อหัวข้อ	ระบบค้นหาความสัมพันธ์โดยใช้อพริออริอัลกอริทึม
นักศึกษา	นายทศพล คงน้อย
อาจารย์ที่ปรึกษา	ผศ.ดร. วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2548

บทคัดย่อ

สภาพธุรกิจในปัจจุบันมีการแข่งขันกันสูงมาก ดังนั้นแต่ละธุรกิจต้องพยายามหาวิธีการเพื่อเพิ่มส่วนแบ่งตลาดของตนเองโดยต้องรักษาฐานลูกค้าเดิมเอาไว้ การกำหนดกลยุทธ์ทางการตลาดเพื่อวางแนวทางการดำเนินธุรกิจจึงเป็นสิ่งที่มีความสำคัญมาก จึงได้มีการนำเอาเทคนิค Data mining เข้ามาช่วยในการวิเคราะห์หารูปแบบหรือ informationที่มีประโยชน์ที่ซ่อนอยู่ในฐานข้อมูลแล้วนำมาใช้ให้เกิดประโยชน์ โครงการนี้จะนำเสนอถึงขั้นตอนและวิธีการพัฒนาระบบงานเพื่อวิเคราะห์ความสัมพันธ์ของผลิตภัณฑ์ โดยใช้ Apiori Algorithm ซึ่งเป็น Algorithm พื้นฐานในการหาความสัมพันธ์ของข้อมูลที่ใช้ใน Association Rules ซึ่งเป็นเทคนิคหนึ่งของ Link Analysis ของ Data mining

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Title	Association Rule Discovery by Apriori Algorithms
Student	Mr. Tossapon Kongnoi
Advisor	Asst.Prof.Dr. Worapoj Kreesuradej
Level of Study	Master of Science in Information Technology
Major	Information Science
Academic Year	2005

ABSTRACT

In today's business world have many competition . Thus each business sector must aim at developing revenue generating strategies to increase customer and save the old customer that the organization have. Indeed , to do this , it call for a complete sets of data existed within an organization together with the data mining techniques to help analyze the marketing opportunities within such data . This project will propose the process and methodology of system enhancement in order to analyze the correlation of data between product usage and banking services. In analyzing . This relationship more effectively, Apriori Algorithm the basic logarithm used in AssociationRules, is implemented as one of the reliable techniques of Data Mining Link Analysis

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย.....I
บทคัดย่อภาษาอังกฤษ.....II
สารบัญ.....III
สารบัญตาราง.....VI
สารบัญรูป.....VII
บทที่
1. บทนำ.....1
1.1 ความเป็นมาของปัญหา.....1
1.2 วัตถุประสงค์.....1
1.3 ขอบเขตการดำเนินงาน.....1
1.4 ขั้นตอนและวิธีการดำเนินงาน.....2
1.5 ประโยชน์ที่คาดว่าจะได้รับ.....2
2. Association Rules Discovery.....3
2.1 Data mining.....3
2.1.1 สิ่งที่เกี่ยวข้องกับData mining.....3
2.1.2 สิ่งที่ดาต้ามายนิ่งสามารถทำได้.....4
2.1.3 การใช้งานดาต้ามายนิ่ง กับ การวิเคราะห์ข้อมูล.....5
2.1.4 การทำงานของดาต้ามายนิ่ง.....5
2.1.5 ขั้นตอนการทำดาต้ามายนิ่ง.....6
2.1.6 การประยุกต์ใช้งานดาต้ามายนิ่ง.....8
2.2 กฎความสัมพันธ์.....9
2.3 อพริออริอัลกอริทึม.....14
2.3.1 Apriori_gen()function.....15

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

2.3.2 Algorithm Apriori.....17

2.3.3 ตัวอย่างเครื่องมือและโปรแกรมที่ใช้อพริออริอัลกอริทึม.....19

2.4 ข้อดีและข้อเสียของAssociation Rules Discovery.....19

2.5 สรุป.....20

3 การวิเคราะห์และออกแบบ.....21

3.1 ลักษณะของโปรแกรม.....21

3.2 Use Case Diagram.....21

3.3 Class Diagram.....22

3.3.1 class Associationruleapiori.....23

3.3.2 class Miningwindow.....23

3.3.3 class Genbutton.....23

3.3.4 class Datamining.....23

3.3.5 class Apiori.....23

3.3.6 class table.....24

3.4 Sequence Diagram.....24

3.5 State Diagram.....25

4 การพัฒนาระบบ.....26

4.1 รูปแบบของข้อมูล.....26

4.2 การติดต่อกับฐานข้อมูล.....27

4.3 เครื่องมือที่ใช้ในการพัฒนาโปรแกรม.....27

4.4 ระบบที่ใช้ในการทำงาน.....28

4.5 การเรียกใช้โปรแกรม.....29

4.6 Interface.....29

4.7 กรณีการทดสอบการทำงาน.....31

5 สรุปผลการดำเนินงาน.....37

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ(ต่อ)

หน้า

5.1 ผลการดำเนินงาน.....37
5.2 สรุปผลการทดลอง.....37
5.3 การประยุกต์ใช้งาน.....37
5.4 ปัญหาและอุปสรรค.....37
5.5 ขอเสนอแนะ.....38
บรรณานุกรม.....39

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

หน้า

ตารางที่

2.1 รายการฐานข้อมูลสำหรับตัวอย่าง Algorithm Basic.....12

2.2 แสดงค่า support สำหรับต่างๆในตารางที่2.1โดน large itemset จะดูค่า support ที่ 40%.....12

2.3 confidence ของบาง association rule ที่มีค่า confidence = 60%.....13

2.4 หา Candidate Sets โดยใช้ Apriori_gen().....16

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

หน้า

รูปที่

2.1 ขั้นตอนต่างๆในการทำ Data mining โดย CRISP-DM.....6
2.2 การค้นหาLarge itemset ต่างๆโดยใช้ Apriori Algorithm.....18
3.1 Use Case Diagram ของโปรแกรมที่ใช้ในการหา Association rules.....21
3.2 Class Diagram ของโปรแกรมที่ใช้ในการหา Association rules.....22
3.3 Sequence Diagram แสดงขั้นตอนการสร้าง Association rules.....24
3.4 State Diagramของการสร้าง Association rules.....25
4.1 Design view of database.....26
4.2 Datasheet view of database.....26
4.3 Interface ของ Eclipse.....28
4.4 ภายใน Folder DataminingApiori.....29
4.5 หน้าต่างการทำงานของโปรแกรม.....30
4.6 Warning Dialog messageแสดงถึงว่าไม่สามารถดำเนินการกับฐานข้อมูลได้.....30
4.7 Combo Box List ในที่ใช้ในการกำหนดค่าSupport.....31
4.8 ข้อมูลในฐานข้อมูล1.....32
4.9 Association rules ที่ได้จากข้อมูลในรูป4.7โดยมีค่าMinsupและMincon=0%.....32
4.10 Association rules ที่ได้จากข้อมูลในรูป4.7โดยมีค่าMinsupและMincon=60%.....33
4.11 ข้อมูลในฐานข้อมูล2.....33
4.12 Association rules ที่ได้จากข้อมูลในรูป4.10โดยมีค่าMinsupและMincon=0%.....34
4.13 Association rules ที่ได้จากข้อมูลในรูป4.10โดยมีค่าMinsup=25และMincon=60%.....34
4.14 Association rules ที่ได้จากข้อมูลในรูป4.10โดยมีค่าMinsup=50และMincon=60%.....35
4.15 Association rules ที่ได้จากข้อมูลในรูป4.10โดยมีค่าMinsup=30และMincon=50%และ
แสดงผลออกมาแบบเรียงลำดับโดยดูจากค่า confidence จากน้อยไปมาก.....35

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

เนื่องจากสภาพธุรกิจในปัจจุบันมีการแข่งขันกันสูงมาก ทำให้แต่ละองค์กรต้องหากลยุทธ์ต่างๆในการตลาดเพื่อที่จะวางแผนแนวทางการดำเนินธุรกิจเพื่อให้ธุรกิจมีความได้เปรียบในเชิงการแข่งขันและสามารถได้รับส่วนแบ่งตลาดที่มากขึ้น สร้างความเป็นปึกแผ่นให้กับองค์กร และนำไปสู่การทำกำไรที่มากขึ้น ตลอดจนสร้างความพึงพอใจให้กับลูกค้าให้ได้มากที่สุด ซึ่งการวางแผนกลยุทธ์ทางการตลาดนั้น ต้องอาศัยข้อมูลที่มีอยู่ในองค์กร ซึ่งข้อมูลเหล่านั้นมีอยู่อย่างมากมายโดยจัดเก็บไว้ในคลังข้อมูล (Data Warehouse) ขององค์กร อย่างไรก็ตามข้อมูลที่อยู่ในคลังข้อมูลนั้นอาจจะไม่เพียงพอ และยากต่อการวิเคราะห์เนื่องจากข้อมูลมีเป็นจำนวนมาก ดังนั้นจึงได้นำเทคนิคการทำเหมืองข้อมูล (Data mining) เข้ามาใช้ในการวิเคราะห์หาสารสนเทศที่ซ่อนอยู่ภายในข้อมูลจำนวนมากนั้น รวมถึงวิเคราะห์หาความสัมพันธ์ของข้อมูล เพื่อศึกษาแนวโน้มและพฤติกรรมของข้อมูลในอนาคตได้ เพื่อใช้เป็นแนวทางในการกำหนดเป้าหมายในการดำเนินงานขององค์กร

1.2 วัตถุประสงค์

เพื่อนำเอาเทคนิคของ การทำเหมืองข้อมูลมาใช้ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลต่างๆในการใช้บริการของลูกค้า เพื่อให้องค์กรสามารถนำสารสนเทศที่ได้นี้ มาใช้ในการวางแผนกลยุทธ์ทางการตลาดได้อย่างมีประสิทธิภาพรวมทั้งเป็นแนวทางให้กับผู้บริหารเพื่อใช้ในสนับสนุนการตัดสินใจ และ วางแผนทำรายการส่งเสริมการตลาดได้อย่างถูกต้อง และ เหมาะสมกับความต้องการของลูกค้าให้มากที่สุด

1.3 ขอบเขตการดำเนินงาน

โครงงานนี้เป็นการศึกษาการนำเอาเทคนิค Data mining มาประยุกต์ใช้โดยอาศัยหลักการของ Link Analysis ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลในฐานข้อมูล โดยที่สามารถกำหนดระดับความน่าสนใจของข้อมูลที่ต้องการได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการศึกษาไว้ดังนี้

1) ศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้องกับ Data mining

2) ศึกษา Association Rule และ Algorithm ที่เกี่ยวข้อง

3) ศึกษาการโปรแกรมเชิงวัตถุ

4) วิเคราะห์ ออกแบบ และพัฒนาระบบงานเพื่อวิเคราะห์ข้อมูลจากฐานข้อมูล

5) สรุปผลการศึกษา

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการที่ได้ศึกษาแนวคิดของ Data mining และ พัฒนาโปรแกรม เพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลจากฐานข้อมูล คาดว่าได้ประโยชน์จากผู้ค้นคว้า และผู้ที่สนใจดังนี้

1) เพื่อให้เข้าใจถึงแนวคิดของกระบวนการทำ Data mining ตลอดจนเทคนิคต่างๆของ Data mining ที่เหมาะสมกับลักษณะงานหรือข้อมูลแต่ละประเภทได้

2) เพื่อให้เข้าใจถึงแนวคิด และขั้นตอนการนำ Data mining มาใช้ในการวิเคราะห์หาความสัมพันธ์ของข้อมูล

3) เพื่อเป็นแนวทางในการนำเทคนิค Data mining ไปใช้กับระบบงาน

4) เพื่อให้สามารถนำผมที่ได้จากการทำ Data mining ไปใช้กับการกำหนดแนวทางการดำเนินธุรกิจได้

ในบทนี้จะกล่าวถึงความเป็นมาของปัญหาในการนำเอา Data mining เข้ามาช่วยในการหาสารสนเทศที่ซ่อนอยู่ในฐานข้อมูล วัตถุประสงค์ ขอบเขตและขั้นตอนการดำเนินงาน รวมไปถึงประโยชน์ที่จะได้รับจากการพัฒนาระบบโครงงานนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2

Association Rule Discovery

2.1 Data mining

ในปัจจุบันธุรกิจต่างๆมีการเพิ่มขึ้นของข้อมูลในปริมาณที่มาก แต่ก็ไม่สามารถที่จะสร้างประโยชน์ได้จากข้อมูลเหล่านี้ ถึงแม้จะมีการนำหลักการ Data Warehouse เข้ามาใช้ในการเก็บข้อมูลสำหรับการวิเคราห์ และมีเครื่องมือ OLAP (On-Line Analyst Process) สำหรับการนำเสนอข้อมูลที่ได้ให้กับผู้บริหารในลักษณะมุมมองต่างๆแล้วก็ตาม แต่การใช้ประโยชน์ของข้อมูลยังไม่เต็มที่ ข้อมูลที่ได้เป็ยเพียงข้อมูลพื้นฐานเท่านั้น จึงได้มีการนำหลักการ Data mining เข้ามาใช้ในการหา Information ที่อยู่ภายในข้อมูลเหล่านั้น

โดยทั่วไปการทำ Data mining (บางครั้งก็เรียกกันว่า knowledge discovery) คือกระบวนการของการวิเคราะห์ข้อมูลโดยเทคนิคต่างๆและสรุปมันให้อยู่ในรูปแบบของ information เพื่อนำไปใช้เพิ่มคุณค่าให้แก่ระบบ Data mining software เป็นหนึ่งในหลายๆเครื่องมือที่ใช้สำหรับวิเคราะห์ข้อมูล โดยที่มันมักจะถูกใช้เพื่อการวิเคราะห์จากหลายๆมิติและหลายๆมุมมองแล้วแบ่งหมวดหมู่ข้อมูลพวกนั้น แล้วจึงสรุปความสัมพันธ์ ต่างๆที่ได้ค้นพบ ดังนั้นเทคนิคการทำ Data mining คือกระบวนการของการหาความสัมพันธ์หรือความเกี่ยวพันธ์ของข้อมูลขนาดใหญ่มากๆที่อยู่ในฐานข้อมูลแบบ relational databases ในเรื่องของการเปลี่ยนแปลงอย่างต่อเนื่องกันมานั้นถึงแม้ว่าการทำ Data mining จะเป็นเรื่องใหม่ แต่เทคโนโลยีนั้นกลับไม่ กล่าวคือบริษัทต่างๆได้ใช้คอมพิวเตอร์ที่มีประสิทธิ ภาพมากอยู่แล้วในการประมาณผมการทำงาน เช่นข้อมูลจากเครื่องสแกนตามซุปเปอร์มาเกตที่ลูกค้าจะต้องช่วยตัวเอง และร้านขายสินค้าปลีกนานาชนิด เครื่องคอมพิวเตอร์จะนำข้อมูลต่างๆของสินค้าที่ลูกค้าซื้อไปมาวิเคราะห์เพื่อสร้างรายงานประจำวัน ประจำเดือน หรือประจำปีได้เอง อย่างไรก็ตาม ถึงแม้ว่าประสิทธิภาพของ CPU จะทรงพลังขึ้น, อุปกรณ์ที่ใช้เก็บข้อมูลจะมี ขนาดพื้นที่ความจุมากขึ้น มีความสามารถมากขึ้น ซึ่งสิ่งเหล่านี้เป็นประโยชน์ในการวิเคราะห์โดย data mining แต่ราคากลับถูกลงเรื่อยๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.1.1 สิ่งที่เกี่ยวข้องกับ Data mining

Data (ข้อมูล) คือข้อเท็จจริงใดๆ ตัวเลข หรือ ข้อความ ที่สามารถประมวลผลโดยคอมพิวเตอร์ได้ ในปัจจุบัน องค์กรต่างๆ ได้สะสมข้อมูลเป็นปริมาณมาก และปล่อยให้ข้อมูลนั้นเติบโตขึ้นเรื่อยๆ โดยรูปแบบการเก็บนั้น จะเก็บอยู่ในรูปแบบต่างๆ และฐานข้อมูลแบบต่างๆ เช่น

- operational or transactional data
- nonoperational data
- meta data

information คือความสัมพันธ์ท่ามกลางข้อมูล นั่นคือ data ของเรามีความสัมพันธ์กันแบบไหนนั่นเอง เช่น การวิเคราะห์ ข้อมูลการค้าปลีกของข้อมูลการขาย จะสามารถสร้าง information ที่จะบอกได้ว่าสินค้าอะไรบ้างที่ขายไปแล้ว และเมื่อไหร่ โดยที่ Information สามารถแปลงไปสู่รูปแบบที่เรียกว่า knowledge ได้ โดยจะอยู่ในรูปแบบของประวัติที่บันทึกไว้ และอนาคตที่จะเป็นไปเพื่อประโยชน์ที่จะช่วยในการตัดสินใจ

Data Warehouse คือที่ๆใช้เพื่อทำการเก็บข้อมูลต่างๆ โดยที่ข้อมูลนี้จะมีขนาดใหญ่มาก โดยลักษณะของ Data Warehouse คือ รวมข้อมูลของทั้งองค์กร มีการแยกเก็บเป็นเรื่องๆ เพื่อให้สามารถตอบคำถามในเรื่องที่สนใจได้ ข้อมูลมักจะมีหลายมิติ การดูข้อมูลจะมีการดูเป็นช่วงเวลา และข้อมูลที่เก็บไว้ทั้งหมด จะไม่มีการลบทิ้ง คือจะปล่อยให้ข้อมูลโตต่อไปเรื่อยๆ

2.1.2 สิ่งที่ดาต้ามายนิ่งสามารถทำได้

Data mining มักจะถูกใช้โดยบริษัทต่างๆ เมื่อต้อง การศึกษาลูกค้าโดยเน้นลงไปที่ลักษณะต่างๆ เช่น การค้าปลีก การเงิน การสื่อสาร และองค์กรตลาด โดยจะทำให้บริษัท เหล่านี้ สามารถที่จะทราบความสัมพันธ์ ระหว่างปัจจัยภายใน และ ปัจจัยภายนอก ซึ่งจะช่วยให้สามารถกำหนดผลกระทบต่างๆในการขาย ความพอใจของลูกค้า และ ผลประโยชน์ที่เกี่ยวพันธ์กัน ในท้ายที่สุด Data mining จะสร้างผลสรุปของ information เพื่อที่จะดูรายละเอียดของ transaction ต่างๆ ของข้อมูลได้

ด้วยการทำ Data mining ผู้ค้าปลีกสามารถใช้จุดขายโดยอิงจากความต้องการของลูกค้า เพื่อที่จะได้ทำการส่งเสริมการขายได้ถูกต้อง โดยการวิเคราะห์นี้จะอยู่บนพื้นฐานของประวัติการซื้อของลูกค้า และจากข้อมูลต่าง เรายังสามารถพัฒนาผลิตภัณฑ์และการส่งเสริมการขายแก่ลูกค้าแต่ละราย เป็นแบบเฉพาะเจาะจงไป

ผลลัพธ์ของ Data mining เช่น

- การคาดคะเนว่าอะไรจะเกิดขึ้นในอนาคต

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- จัดประเภท คน หรือ สิ่งของ เป็นหมวดหมู่
- จัดประเภทคนหรือสิ่งของโดยดูคุณสมบัติ
- การเกี่ยวข้องกันของเหตุการณ์ที่เป็นไปได้ที่จะเกิดขึ้นร่วมกัน
- ลำดับต่างๆที่การที่เป็นไปได้ที่จะเกิดขึ้นต่อเนื่องกัน

2.1.3 การใช้งานดาต้ามายนิ่ง กับ การวิเคราะห์ข้อมูล

Data mining เป็นแบบฉบับในการพัฒนาในเรื่องของระบบผู้ชำนาญการ โดยมากจะให้ความสำคัญ ไม่มากนักกับวิธีการทำ คือถ้ามันทำได้ก็ใช้มันไป และยังไม่ต้องการการสันนิษฐานเพื่อสร้างข้อมูล (Data) อีกด้วย Data mining ก็สามารถพบความสัมพันธ์ในฐานข้อมูลขนาดใหญ่ได้ แต่มันต้องการความ เข้าใจในตัวข้อมูล และ ปัญหาทางธุรกิจ

Data analysis เป็นการทดสอบ Model ทางสถิติ แต่ก็ขึ้นอยู่ว่า Model ทางสถิตินั้นจะสันนิษฐานขึ้นมาถูกต้องมากแค่ไหน และยังมีการทดสอบแบบ Hypothesis ซึ่งเป็นการทดสอบความสัมพันธ์อีกด้วย การทดสอบนี้จะทำอยู่บนการ สุ่มข้อมูลขึ้นมา เทคนิคที่ใช้มักจะเกิดข้อผิดพลาดได้เมื่อข้อมูลมีขนาดใหญ่มากๆ และวิธีนี้ยังต้องการความสามารถทางสถิติอย่างมากอีกด้วย

OLAP หรือชื่อเต็มๆว่า On-line Analytical Processing มีหน้าที่สร้างมุมมองที่ดีที่สุด ที่จะแสดงให้ดูได้ว่ากำลังเกิดอะไรขึ้นอยู่ แต่ไม่สามารถบอกได้ว่า ทำไมถึงเกิดเหตุการณ์แบบนี้ และไม่สามารถที่จะทำนายเหตุการณ์ในอนาคตได้อีกด้วย

2.1.4 การทำงานของดาต้ามายนิ่ง

ในขณะที่ระบบสารสนเทศขนาดใหญ่ได้รับการพัฒนาทั้งทาง transaction และ ระบบการวิเคราะห์ อย่างคู่ขนานกันไป Data mining จึงเข้ามาจัดการเชื่อม 2 สิ่งนี้เข้าด้วยกัน software ทาง Data mining จะวิเคราะห์ความสัมพันธ์ และรูปแบบที่อยู่ใน transaction โดยมีพื้นฐานอยู่บน การร้องขอของผู้ใช้งานระบบ ในหลายๆรูปแบบของการวิเคราะห์ ได้ใช้ประโยชน์จากหลักการของสถิติ กลจักร neural network โดยทั่วไป ความสัมพันธ์ที่มักจะค้นหากันมี 4 อย่างคือ Classes, Clusters, Association, Sequential patterns

Data mining ประกอบด้วย 5 ส่วนหลักคือ

- ทำการเก็บข้อมูลลง Data warehouse
- เก็บและจัดการข้อมูลในระบบฐานข้อมูลหลายมิติ
- จัดเตรียมข้อมูลให้แก่ นักวิเคราะห์ธุรกิจและผู้เชี่ยวชาญทางด้านคอมพิวเตอร์
- ทำการวิเคราะห์ข้อมูลโดย Software

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- นำเสนอข้อมูลในรูปแบบที่มีประสิทธิภาพ เช่น กราฟ หรือ ตาราง

และระดับต่างๆของการวิเคราะห์คือ

- Artificial neural networks
- Genetic algorithms
- Decision trees
- Nearest neighbor method
- Rule induction
- Data visualization

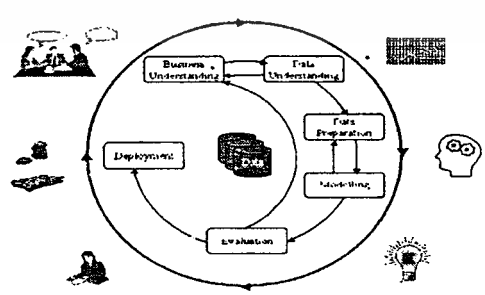
2.1.5 ขั้นตอนการทำดาต้ามายนิ่ง

เราสามารถทำ Data mining ได้ โดยการใช้ประโยชน์จากกรรมวิธี CRISP-DM (Cross Industry Standard Process for Data Mining)

เหตุที่ต้องใช้ CRISP-DM มีดังต่อไปนี้

- กระบวนการData mining ต้องสามารถเชื่อถือได้ และสามารถทำซ้ำโดยคนและความชำนาญเพียงเล็กน้อย
- CRISP-DM มีความยืดหยุ่นที่จะรับผิดชอบความแตกต่างทางด้าน ธุรกิจ ปัญหาตัวแทน และข้อมูล
- CRISP-DM ได้เตรียมรูปแบบของขอบข่ายงานสำหรับ
  - Guidelines
  - เอกสารอ้างอิงประสบการณ์

ในการทำ CRISP-DM จะมีขั้นตอนอยู่ 6 ขั้นตอนตามรูปที่ 2.1



รูปที่ 2.1 ขั้นตอนต่างๆในการทำ Data mining โดย CRISP-DM

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในขั้นตอนที่ 1 Business Understanding คือ การทำความเข้าใจกับธุรกิจ โดยจะต้องหา จุดประสงค์ของธุรกิจ จุดประสงค์ของการทำ Data mining และ กฎเกณฑ์ความสำเร็จ

ขั้นตอนที่ 2 Data Understanding คือการทำความเข้าใจข้อมูล การค้นหาข้อมูลและตรวจสอบคุณภาพของข้อมูล และหา outliers

ขั้นตอนที่ 3 Data preparation เป็นการเตรียมข้อมูลขั้นตอนนี้มีความสำคัญมาก และเป็นขั้นตอนที่ใช้เวลามากที่สุด เพื่อให้มั่นใจได้ว่าข้อมูลที่ถูกเลือกมานั้นมีความถูกต้องเหมาะสม (เวลา90 % ของการทำงานจะอยู่ที่นี่)

- สะสมข้อมูล
- ประเมินข้อมูล
- นำข้อมูลมารวมกันและจัดการให้ดูดี เช่น การทำตารางเชื่อมต่อ ระดับของการรวมเป็นกลุ่ม และค่าที่หายไป
- การเลือกข้อมูล ต้องมีความตื่นตัวในข้อมูลที่ยังไม่ถูกค้นพบ มีการใช้ข้อมูลตัวอย่าง และ ใช้ visualization tools
- มีการเปลี่ยนรูปโดยสร้างตัวแปลใหม่ๆ

ขั้นตอนที่ 4 Model building การสร้างแบบจำลองโดยแบบจำลองมีอยู่ 2 ประเภทคือ แบบการคาดคะเน และ การอธิบาย ซึ่งมีหลักการในการทำดังนี้

- การเลือดเทคนิคในการสร้างแบบจำลองโดยมีพื้นฐานอยู่บนจุดประสงค์ของการทำ Data mining
- การสร้างแบบจำลองเป็นกระบวนการที่ interactive ที่แตกต่างกันสำหรับบิเรียนรู้สิ่งที่ควบคุมดูแลได้และไม่ได้
- บางแบบจำลองไม่จำเป็นที่จะต้องเป็นการอธิบาย หรือ การคาดคะเน ได้หรือไม่

ประเภทของแบบจำลองมีอยู่ 2 แบบคือ

1. การคาดคะเน ซึ่งเป็นแบบจำลองสำหรับการทำนาย และ Classification
    - Regression algorithms จะใช้คาดคะเนสิ่งที่เกี่ยวกับตัวเลขที่จะออกมา รูปแบบของแบบจำลองประเภทนี่ คือ neural networks, rule induction, CART (OLS regression, GLM)
    - Classification algorithms จะใช้คาดคะเนสิ่งที่เกี่ยวกับสัญลักษณ์ที่จะออกมา รูปแบบของแบบจำลองประเภทนี้คือ CHAID, C5.0 (discriminant analysis, logistic regression)
2. การอธิบาย ซึ่งเป็นแบบจำลองสำหรับการแบ่งกลุ่ม และ หาความสัมพันธ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- Clustering/Grouping algorithms: K-means, Kohonen
- Association algorithms: apriori, GRI

ขั้นตอนที่ 5 Model Evaluation คือการตีค่าแบบจำลอง ซึ่งการตีค่าแบบจำลองจะดีแค่ไหนนั้น ก็ต้องขึ้นอยู่กับประสิทธิภาพของข้อมูลที่เอามาทดสอบ วิธี และ กฎเกณฑ์ ที่นำมาใช้จะขึ้นอยู่กับชนิดของแบบจำลอง ในเรื่องของอัตราความผิดพลาดของแบบจำลองนั้นว่า สำคัญหรือไม่ ง่ายหรือยาก ก็ขึ้นอยู่กับ algorithm ที่ใช้

ขั้นตอนที่ 6 Deployment คือการจัดรูป โดยเราต้องระบุว่าผล ลัพธ์ที่ต้องการเป็นอย่างไร ใครมีความต้องการที่จะใช้มัน และมีความถี่ที่ใช้เป็นเท่าไหร่ เราสามารถจัดผลลัพธ์ของ Data mining ได้โดยการเรียบเรียงฐานข้อมูล และ นำผลลัพธ์มาใช้ประโยชน์เหมือนกฎทางธุรกิจ

2.1.6 การประยุกต์ใช้งานดาต้ามายนิ่ง

ในปัจจุบันมีการนำ Data mining มาใช้ประโยชน์ในทางธุรกิจได้หลายทางซางแล้วแต่ใครจะนำไปประยุกต์ใช้ ได้แก่

1) การจัดการทางการตลาด (Market analysis and management) เช่น

- Cross selling เป็นการหาความสัมพันธ์ของสินค้าที่ขายไปด้วยกัน เพื่อนำเสนอขายสินค้าให้กับลูกค้า

- Customer Profiling เป็นการทำ segment ของลูกค้า โดยศึกษาตาม profile ของลูกค้าเพื่อแบ่งเพื่อแบ่งลูกค้าเป็นกลุ่ม สำหรับนำเสนอขายโครงการต่างๆให้เหมาะสมกับลูกค้ากลุ่มนั้นๆต่อไป

- Direct mail campaign เป็นการเลือกกลุ่มลูกค้าที่จะทำการส่ง direct mail ไปแล้วลูกค้ามีการตอบสนองกลับ :เนื่องจากหากทำการสุ่มส่ง mail ออกไปจะมีการตอบสนองกลับมาน้อยมาก ซึ่งทำให้สูญเสียต้นทุนสูงด้วย

-Customer Relation Management (CRM) เป็นวิธีการรักษาและบริหารความสัมพันธ์กับลูกค้าให้ดีขึ้น ปัจจุบันนิยมใช้ call center เพื่อบันทึกการโทรเข้ามาขอรับบริการจากลูกค้า เพื่อนำข้อมูลนั้นมาวิเคราะห์ต่อไป

2) การวิเคราะห์ความเสี่ยง (Risk Analysis and management) เช่น

- Credit Scoring เป็นการประเมินเบื้องต้นในการปล่อยกู้ให้กับลูกค้า

- Churn Management เป็นการศึกษาว่าสาเหตุใดที่ทำให้ลูกค้าเปลี่ยนไปใช้บริการจากบริษัทอื่น

3) การตรวจจับการโกง (Fraud detection and management) เช่น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

- Money laundering การตรวจจับการฟอกเงิน โดยพิจารณาถึงการเชื่อมโยง transaction ที่น่าสงสัย

4) Web Mining เช่น

- การศึกษาการใช้ web site (Web usage mining) เพื่อมีผมต่อการออกแบบหน้าจอต่างๆ หรือจัด category ของ web site

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

2.2 กฎความสัมพันธ์

กฎความสัมพันธ์ (Association rules) เป็นส่วนหนึ่งของงานวิจัยที่เกี่ยวของกับการทำ Data mining และกำลังได้รับความสนใจ จาก กลุ่มคนที่เกี่ยวข้องกับ ฐานข้อมูลอยู่เพราะว่าสามารถนำมาประยุกต์ใช้ในงานทางการตลาด และ งานค้าปลีกต่างๆ ได้อย่างดีเยี่ยม

เทคนิค association rules ใช้สำหรับการวิเคราะห์การบริหารทางการเงิน และในธุรกิจค้าปลีก ซึ้อาจเรียกอีกชื่อนึงได้ว่า "Market basket analysis" หรือ "Product Affinity Analysis" โดยเป็นการกล่าวถึงความคิดที่ว่า คุณสามารถสำรวจสินค้าที่อยู่ในตระกร้า และค้นหารูปแบบที่ซึ่งสามารถถูกใช้สำหรับลักษณะการวางสินค้าบนชั้นวางได้ดีที่สุด ดังเช่นเรื่อง"beer and diaper" ซึ้งการวิเคราะห์ association rules ได้แสดงให้เห็นว่า beer และ diaper สามารถถูกซื้อไปด้วยกันบ่อยกว่าที่คิด ซึ้งผู้วิเคราะห์ได้กล่าวว่า ผู้ชายที่จะจะต้องไปร้านขายของเพื่อซื้อผ้าอ้อมจะทำการซื้อ beer มาด้วย ซึ่งบทสรุปที่ได้มานี้ทำให้ผ้าอ้อมถูกวางใกล้ๆกับ beer ใน supermarket กรณี beer กับ diaper นี้เป็นตัวอย่างที่แสดงให้เห็นถึงว่า association rules สามารถถูกใช้ในการกำหนดกลยุทธ์ได้อย่างไร รวมถึงเป็นตัวอย่างของความเสี่ยงในการวางหลักการทั่วไปให้อยู่บนพื้นฐานของกฎที่ไม่ได้เชื่อถือได้เสมอไป

ในทางธุรกิจสามารถนำ Market basket analysis มากำหนดกลยุทธ์ทางการตลาดต่างๆเช่น

: การลดราคา หรือคูปองลดราคา (Couponing and discounting) เช่น การที่จะลดราคาทั้ง beer และ diaper เพียงลดราคาสินค้าเพียงชนิดเดียวก็ช่วยเพิ่มยอดขายของสินค้าอีกชนิดได้ด้วย เนื่องจากสินค้าทั้งสองชนิดมีแนวโน้มที่จะถูกซื้อไปด้วยกัน

: การจัดวางสินค้า (Product placement) การวางสินค้าที่มีความสัมพันธ์กันไว้ใกล้ๆกันทำให้เกิดข้อดีมากกว่าการวางสินค้าที่มีลักษณะประเภทเดียวกันไว้ด้วยกัน หรือในทางตรงกันข้ามการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

วางสินค้าที่มีความสัมพันธ์กันไว้ไกลๆทำให้ลูกค้าต้องเดินหา ซึ่งจะทำให้มองเห็นสินค้าชนิดอื่นๆ ได้ด้วย

:Timing และ Cross-marketing เช่นกำหนดกฎว่าคนที่ซื้อ VCR (เครื่องบันทึกเทป) มักจะซื้อ กล้องวีดีโอพกพาในเวลา 2-4 เดือนหลังจากที่ซื้อ VCR ดังนั้น เมื่อมีการ promote กล้องวีดีโอใหม่ ก็สามารถทำการส่งข้อมูลเหล่านี้ไปให้กับผู้ที่ซื้อ VCR ไป ซึ่งมีแนวโน้มว่าจะกลับมาซื้อในอีก 2-4 เดือนข้างหน้า

รูปแบบปัญหาของ Association rules จะมีดังนี้

- ให้ Let I = $\{I_1, I_2, \ldots, I_m\}$ เป็น set ของ m attributes ที่เหมาะสม ซึ่งเราจะเรียกมันว่า literals ให้ D เป็น Dฐานข้อมูล ที่แต่ละ record (tuple) T ถูกระบุได้โดยไม่ซ้ำกัน และ บรรจุ ชุด ของ items ต่างในรูปแบบที่ $T \subseteq I$ แล้ว Association rule หนึ่งๆ คือการ implement ในรูปแบบ $X \Rightarrow Y$ โดยที่ $X, Y \subset I$ เป็นชุดของ item ต่างๆ และเราจะเรียกมันว่า itemsets , และ $X \bigcap Y = \phi$. ในตอนนี้เราจะเรียก X ว่า antecedent และเรียก Y ว่า consequent.

- Support (s) ของ Association rule คือ อัตราส่วนที่อยู่ในรูปของเปอร์เซ็นต์ของ record ต่างๆ ที่บรรจุ X Y Y เพื่อเป็นผลรวมจำนวนของ record ในฐานข้อมูล เพราะฉะนั้น ถ้าเราพูดถึง support ของ rule คือ 5% จะตีความได้ว่า 5% ของ record ทั้งหมด บรรจุ X Y Y โดยทั่วๆไปถ้า support ต่ำกว่า 5 % ผู้จัดการร้านนั้นๆมักจะไม่ค่อยสนใจพึ่งความสัมพันธ์ของรายการสินค้าที่ถูกรวมเข้าด้วยกัน ถึงแม้ว่า support ของ Association rules ที่มีค่าสูงจะเป็นที่ต้องการ แต่บางทีมันก็อาจจะไม่ใช่ทางเลือกที่เป็นไปได้ก็ได้ ตัวอย่างของเหตุการณ์ เช่น ถ้าเราใช้ Association rule เพื่อทำนายความล้มเหลวของระบบการสื่อสาร ที่มีพื้นฐานแบบ switching nodes บน ชุดของเหตุการณ์ ที่จะทำให้ระบบล้มเหลวได้ ในความเป็นจริงถ้าเหตุการณ์นี้ไม่เกิดขึ้นบ่อยนัก ความสัมพันธ์ที่ Association rule ต่างๆแสดงให้เห็นจะมีความสำคัญมาก

- ในเรื่องของจำนวน record นั้น จะมีค่า confidence (α) เป็นอัตราส่วนในรูปของเปอร์เซ็นต์ของจำนวน record ที่บรรจุ X Y Y กับจำนวน record ที่บรรจุ X ด้วยเหตุนี้ ถ้าเราบอกว่ากฎนั้น มี confidence 85% จะหมายความว่า 85% นั้นคือ 85% ของ record ที่บรรจุ X และ บรรจุ Y เช่นเดียวกัน ซึ่งค่านี้จะบอกถึงความเชื้อถือได้ใน Data set ของ X และ Y โดยส่วน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

มาก confidence จะมีผลต่อความน่าเชื่อถือของ Association rule ถ้าชุดของเหตุการณ์ต่างๆ ให้ค่าสัดส่วนที่ไม่มากนัก ความสัมพันธ์นี้ก็จะไม่ค่อยได้รับความสนใจซักเท่าไหร่ Mining ของ Association จากฐานข้อมูล จะรวมทั้ง support และ confidence ดังนั้นปัญหาของ mining association rule ต่างๆ สามารถเป็นได้ทั้ง support และ confidence เช่นขั้นตอนของAlgorithm ดังต่อไปนี้

Algorithm 1 Basic:

Input:

I. D, s. c

Output:

Association rules satisfying s and c

Algorithm:

1) Find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, s.

2) Generate the desired rules using the large itemsets, which have user-specified threshold confidence,

ในขั้นแรกของ Algorithm นี้จะเป็นการหา large itemset ต่างๆ และ itemset ที่เหลือ ที่ไม่ได้ระบุว่าเป็น large itemset ก็จะถูกระบุว่าเป็น small itemset ณ.จุดนี้ itemset หนึ่งๆจะเป็น subset ของชุด itemset ต่างๆ ที่น่าสนใจจากฐานข้อมูล การสังเกตสิ่งที่น่าสนใจ หรือมีประโยชน์จาก large itemset ต่างๆจะดูที่ว่า ถ้า itemset X เป็น small itemset subset ของ X ก็จะเป็น small itemset ด้วย แต่ก็ไม่ได้หมายความว่าถ้า X เป็น large itemset แล้ว subset ของ X ก็จะเป็น large itemset ด้วย ในบทความนี้เราจะให้ L แทนชุดของ large itemset ต่างๆ

ในขั้นที่ 2 ของ Algorithm นี้จะเป็นการหา association rule ต่างๆที่จะใช้กับ large itemset ที่ได้มาจากขั้นตอนแรก

ตัวอย่างต่อไปนี้จะช่วยให้เข้าใจกระบวนการพื้นฐานในการหา association rule จาก large itemset ต่างๆ

ตัวอย่าง พิจารณาถึงฐานข้อมูลขนาดเล็กที่มี 4 item I = {Bread, Butter, Eggs, Milk} และ 4 รายการที่แสดงตาม ตารางที่ 2.1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.1 รายการฐานข้อมูลสำหรับตัวอย่าง Algorithm Basic

Transaction ID	Items
T1	Bread, Butter, Eggs
T2	Butter, Eggs, Milk
T3	Butter
T4	Bread, Butter

และมี support สำหรับ itemset ต่างๆ ในตารางที่ 2.1 โดน large itemset จะดูค่า support ที่ 40% ตามตารางที่ 2.2

ตารางที่ 2.2 แสดงค่า support สำหรับต่างๆในตารางที่ 2.1 โดน large itemset จะดูค่า supportที่ 40%

Itemset	Support, s	Large/Small
Bread	50%	Large
Butter	100%	Large
Eggs	50%	Large
Milk	25%	Small
Bread, Butter	50%	Large
Bread, Eggs	25%	Small
Bread, Milk	0%	Small
Butter, Eggs	50%	Large
Butter, Milk	25%	Small
Eggs, Milk	25%	Small
Bread, Butter, Eggs	25%	Small
Bread, Butter, Milk	0%	Small
Bread, Eggs, Milk	0%	Small
Butter, Eggs, Milk	25%	Small
Bread, Butter Eggs, Milk	0%	Small

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ซึ่งจะทำให้เข้าใจได้ว่า ค่า support ต่ำสุด และค่า confidence ต่ำสุดคือของ association rule คือ 40 % และ 60 % ตามลำดับ อันจะแสดงให้เห็นได้ว่ามีหลาย association ที่เป็นไปได้ แต่จุดหลักๆที่เราจะสนใจจะอยู่ในตารางที่ 2.3

ตารางที่ 2.3 confidence ของบาง association rule ที่มีค่า confidence = 60%

Rule	Confidence	Rule Hold
Bread $\Longrightarrow$ Butter	100%	Yes
Butter $\Longrightarrow$ Bread	50%	No
Butter $\Longrightarrow$ Eggs	50%	No
Eggs $\Longrightarrow$ Butter	100%	Yes

ขั้นตอนแรกจะต้องหา itemset ทั้งหมดที่เป็นไปตามกฎ large itemset ในขั้นที่ 2 ต้องพิสูจน์ว่า กฎนั้นมีค่า confidence อย่างน้อย 60% ถ้าเงื่อนไขเป็นไปตามกฎ เราก็สามารถบอกได้ว่ามีเครื่องแสดงความเชื่อมั่นได้ว่ากฎนั้นถูกยึดถือได้ด้วยค่า confidence 60% itemset ต่างๆที่มีความสัมพันธ์กับ กฎที่กล่าวมาแล้วคือ {Bread, Butter}, และ {Butter, Eggs}. ค่าsupport ของแต่ละ itemset ที่น้อยที่สุดคือ 40% (ตามตารางที่ 2.2) ดังนั้นจะได้ว่า itemset ที่มีค่า support มากกว่า 40% จะเป็น large itemset ค่า confidence ของแต่ละกฎที่แสดงในตารางที่ 2.3 จะแสดงถึงกฎที่ได้ โดยกฎแรกจะถือตาม (Bread $\Longrightarrow$ Butter) แต่อย่างไรก็ตามกฎที่สองจะไม่ยึดถือตาม (Butter $\Longrightarrow$ Eggs) เพราะค่า confidence ต่ำกว่า 60%

Algorithm 2 Find Association Rules Given Large Itemsets:

Input:

I, D, s, c ,L

Output:

Association rules satisfying s and c

Algorithm:

1) Find all nonempty subsets, x, of each large itemset, $l \in L$

2) For every subset,obtain a rule of the form $x \Longrightarrow (l\text{-}x)$ if the ratio of the frequency of occurrence of l to that of x is greater than or equal to the threshold confidence.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากตัวอย่างนี้ เราสามารถคาดเดาได้ว่า กฎข้อแรก{Bread $\Longrightarrow$ Butter} ได้ยึดถือตาม Algorithm ที่ 2 นี้ คือ l = {Bread, Butter}, และ x = {Bread}. ดังนั้น (l-x) = {Butter} ตอนอัตราส่วนของ support(Bread, Butter) เพื่อ support(Bread) คือ 100% ซึ่งมากกว่าค่า minimum confidencc ดังนั้นจะบอกได้ว่า ยึดถือตามกฎ เพื่อความเข้าใจที่ดีขึ้น ให้เราพิจารณากฎข้อที่ 2 Butter $\Longrightarrow$ Eggs ที่ซึ่ง x = {Butter}. และ (l-x) = {Eggs} อัตราส่วน support(Butter, Eggs) หารด้วย support(Butter) คือ 50% ซึ่งน้อยกว่า 60% ดังนั้น เราสามารถบอกได้ว่า {Butter} $\Longrightarrow$ {Eggs} มีค่า confidence ไม่เพียงพอ เมื่อเทียบกับเกณฑ์ที่ต้องมีค่า confidence 60%

2.3 อพริออริอัลกอริทึม

Apriori algorithm ถูกพัฒนาโดย Agrawal ในปี 1994 ซึ่ง Apriori algorithm นับว่าเป็นความสำเร็จในประวัติศาสตร์ในการทำ mining association rules เลยก็ว่า เพราะมันได้พัฒนาความรู้ในเรื่องของ association rule algorithm ไปได้อย่างรวดเร็วมาก โดยเทคนิคนี้จะใช้คุณลักษณะของแต่ละ subset ใน large itemset ดังนั้นเราจะสมมุติได้เลยว่า item ต่างๆนั้นได้ถูกเก็บไว้ใน lexicographic order แล้ว Apriori จะสร้าง candidate item ต่างๆ โดยการ join large item ต่างๆ ในกระบวนการ ผ่าน และ ลบ subset เหล่านั้น ที่ซึ่ง itemset ที่เป็น small itemset จะผ่านไปโดยไม่ถูกพิจารณาเกี่ยวกับรายการต่างๆในฐานข้อมูล เพราะเราจะพิจารณาแต่ large itemset ที่ผ่านออกไป ทำให้จำนวนของ candidate large itemset ต่างๆ จะลดลงอย่างเห็นได้ชัด

ในการผ่านครั้งแรกของ itemset ต่างๆ จะมีเพียงแค่ item เดียวเท่านั้นที่ถูกนับ การค้นพบ large itemset ต่างๆในการผ่านครั้งแรกจะถูกใช้เพื่อสร้าง candidate set ต่างๆของการผ่าน ครั้งที่2ที่จะใช้ใน apriori_gen() function แต่ละครั้งที่ candidate itemset ต่างๆถูกพบ support ต่างๆของพวกมันจะถูกนับ เพื่อใช้ในการค้นพบ large itemset ต่างๆของการผ่านครั้งที่ 2 โดยการแสกนฐานข้อมูล ในการผ่านครั้งที่ 3 large itemset ต่างๆที่ได้จากการผ่านครั้งที่2จะถูกพิจารณาเหมือนว่าเป็น candidate set ต่างๆ เพื่อใช้ในการค้นพบ large itemset ต่างๆ ในการผ่านครั้งนี้ กระบวนการทำซ้ำๆนี้จะยกเลิกเมื่อไม่มี large itemset ตัวใหม่ถูกค้นพบขึ้นมาอีก แต่ละการผ่าน i ของ algorithm จะแสกนฐานข้อมูลครั้งหนึ่ง และจะค้นหา large itemset ต่างๆ Li จะใช้แทน large itemset ต่างของ size I, Ci จะใช้แทน candidate ของ size I

ค่า Minimum Support และ Minimum Confident จะเป็นที่ใช้ในการกำหนดขอบเขตของกฎที่เราต้องการเช่น ให้ Minimum Support = 6% และค่า Minimum Confident = 70% คือให้แสดงเฉพาะกฎที่มีค่า Minimum Support มากกว่า หรือ เท่ากับ 6% และค่า Minimum Confident มากกว่าหรือ เท่ากับ 70% หรือ ในรูปของกฎความสัมพันธ์ที่ว่า "Keyboard $\rightarrow$ Mouse (support = 6%,

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

confidence = 70%)" หมายความว่าค่า 70% ของ confidence นั้นจะแสดงจำนวนเหตุการณ์ที่ซื้อทั้ง keyboard และ mouse เทียบกับจำนวนเหตุการณ์ที่ซื้อ mouse ทั้งหมด ส่วนค่า 6% ของ support แสดงถึงเหตุการณ์ที่ซื้อทั้ง keyboard และ mouse เทียบกับจำนวนเหตุการณ์ขายทั้งหมดในฐานข้อมูล

2.3.1 Apriori_gen()function

The apriori_gen() function ได้ถูกอธิบายขึ้นมาโดย Agrawall ในปี 1994 ว่ามีอยู่ 2 ขั้นตอน โดย

- ช่วงขั้นตอนแรก Lk-1 จะถูก join กันกับตัวของมันเองเพื่อใส่ Ck ลงไป
- ในขั้นตอนที่สอง The apriori_gen() จะลบ itemset ทั้งหมดจากผลลัพธ์ของการ join ซึ่งทำให้มีบาง (k-1)-subset ไม่อยู่ใน Lk-1 แล้วจากนั้นมันจะคืนค่า large k-itemset ต่างๆ

Method: apriori_gen() [Agrawal1994]

Input: set of all large (k-1)-itemsets L_{k-1}

Output: A superset of the set of all large k-itemsets

//Join step

I_i = Items i

insert into C_k

Select $p.I_1, p.I_2, \ldots\ldots p.I_{k-1}, q.I_{k-1}$

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

Where $p.I_1 = q.I_1$ and …… and $p.I_{k-2} = q.I_{k-2}$ and $p.I_{k-1} < q.I_{k-1}$.

//pruning step

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

ตัวอย่างในตารางที่ 2.4 จะช่วยให้เข้าใจ apriori_gen() ได้ดีขึ้น large item ต่างๆหลังจากผ่านครั้งที่ 3 ไปแล้วจะแสดงอยู่ในคอลัมน์แรก ทำให้เข้าใจได้ว่ารายการทั้งหมดคือ {Apple, Bagel, Chicken, Eggs, DietCoke} และหลังจากผ่านการ join L3 กับตัวของมันเองแล้ว C4 จะเป็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

{{Apple, Bagel, Chicken, DietCoke}, {Apple, Chicken, DietCoke, Eggs} ซึ่งจะลดขั้นตอนในการลบ itemset {Apple, Chicken, DietCoke, Eggs} เพราะว่า subset ของมันกับ 3 item {Apple, DietCoke, Eggs} ไม่ได้อยู่ใน L3

ตารางที่ 2.4 หา Candidate Sets โดยใช้ Apriori_gen()

Large Itemsets in the third pass (L_3)	Join (L_3, L_3)	Candidate sets of the fourth pass (C_4 after pruning)
{{Apple, Bagel, Chicken}, {Apple, Bagel, DietCoke}, {Apple, Chicken, DietCoke}. {Apple, Chicken, Eggs}, {Bagel, Chicken, DietCoke}}	{{Apple, Bagel, Chicken, DietCoke}, {Apple, Chicken, DietCoke Eggs}}	{{Apple, Bagel, Chicken, DietCoke}}

subset() function จะคืนค่า subset ต่างๆของ candidate set ต่างๆที่แสดงอยู่ในรายการ การนับ support ของ candidate ต่างๆคือการใช้เวลาให้เสียเปล่าไปในแต่ละขั้นตอนของ algorithm เพื่อที่จะลดจำนวนของ candidate ต่างๆที่ต้องการตรวจสอบสำหรับให้รายการลงนั้น candidate itemset Ck จะถูกใส่ลงใน hash tree แต่ละ node ของ hash tree จะบรรจุ leaf node หรือ ตาราง hash(an internal node) leaf node ต่างๆจะบรรจุ candidate itemset ต่างๆใน sorted order internal node ต่างๆของ hash tree จะมีตารางเชื่อมต่อกับ child node ต่างๆ itemset ต่างๆจะถูกแทรกเข้าไปใน hash tree

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ห้องสมุดคณะเทคโนโลยีสารสนเทศ สจล.

โดยใช้ hash function เมื่อ itemset ถูกแทรกเข้าไปแล้วมันจะถูกร้องขอให้เริ่มตั้งแต่ root แล้วไล่ลงมาเรื่อยๆตาม hash tree จนถึง leaf

2.3.2 Algorithm Apiori

algorithm 3 จะแสดงถึงเทคนิค apriori ที่จะบอกว่าในระยะแรก algorithm จะทำงานซ้ำๆ

```
Function count(C: a set of itemsets, D: database)
begin
  for each transaction T ∈ D=⋃D^i do begin
  forall subsets x ⊆ T do
   if x ∈ C then
    x.count++;
  end
 end
```

Algorithm 3. Apriori [Agrawal1994]

Input:

I, D, s

Output:

L

Algorithm:

//Apriori Algorithm proposed by Agrawal R., Srikant, R. [Agrawal1994]

//procedure LargeItemsets

1) C_1: = I: //Candidate 1-itemsets

2) Generate L_1 by traversing database and counting each occurrence of an attribute in a transaction;

3) for (k = 2; $L_{k-1} \neq \phi$; k++) do begin

//Candidate Itemset generation

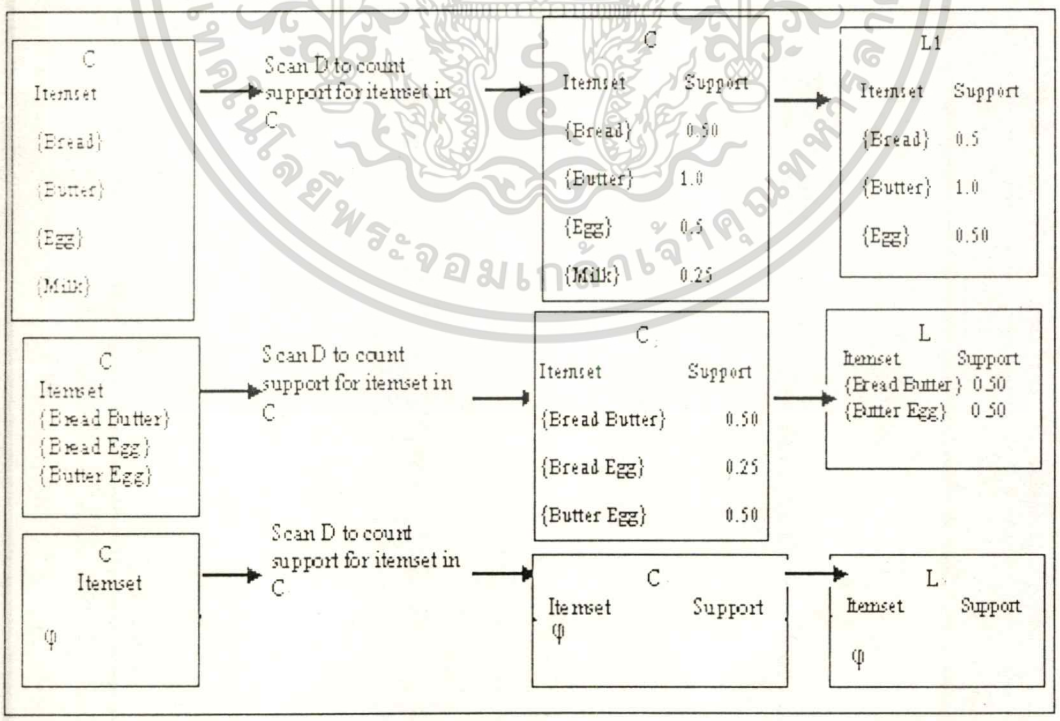
//New k-candidate itemsets are generated from (k-1)-large itemsets

4) C_k = apriori-gen(L_{k-1});

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
//Counting support of C_k
5)          Count (C_k, D)
6)      L_k = {c ∈ C_k | c.count ≥ minsup}
7)  end
9)  L:= ⋃_k L_k
```

รูปที่ 2.2 จะช่วยให้เข้าใจได้ว่า apriori algorithm ทำงานยังไงในตัวอย่าง algorithm 2 เริ่มแรก แต่ละ item ของ itemset จะถูกพิจารณาเหมือนเป็น 1-item candidate itemset ดังนั้น C1 มี 4 1-item candidate sets ซึ่งคือ {Bread}, {Butter}, {Eggs}, และ {Milk} L1 ประกอบด้วย 1-item จาก C1 ที่มี support มากกว่าหรือเท่ากับ 0.4 C2 จะถูกสร้างโดยการ join L1 กับตัวของมันเอง และ ลบ itemset ต่างๆที่มี subset ที่ไม่ได้อยู่ใน L1 วิธีนี้เราจะได้ C2 คือ {{Bread Butter}, {Bread Eggs}, {Butter Eggs}} การนับ support ของ C1,C2 จะพบว่าเป็น {{Bread Butter}, {Butter Eggs}} ด้วยการใช้ apriori_gen() เราจะไม่ได้ candidate itemset ต่างๆจากรอบที่ 3 เพราะว่า เงื่อนไขสำหรับการ join L2 กับตัวของมันเองไม่สามารถจะหาเจอได้ในรอบที่ 3 นี้



รูปที่ 2.2 การค้นหาLarge itemset ต่างๆโดยใช้ Apriori Algorithm

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3.3 ตัวอย่างเครื่องมือและโปรแกรมที่ใช้อพริออริอัลกอริทึม

"Weka" คือ software ทางด้าน data mining ที่เป็นแบบ open source software ตัวนี้ไม่ได้ support แต่ในส่วนการเรียนรู้ algorithm เท่านั้น แต่ยัง support ในส่วนของการเตรียมการข้อมูล และ meta-learners software ตัวนี้ถูกสร้างขึ้นมาโดย JAVA ดังนั้นมันจึงสามารถทำงานบนระบบไหนก็ได้ ข้อดีคือ ข้อมูลทั้งหมดจะถูกเตรียม จัดรูปแบบ และถูกนำไปใช้กับ algorithm ของ data mining ซึ่งหมายความว่า มันต้องการแค่รูปแบบของข้อมูลเท่านั้น แล้วจะพยายาม เปรียบเทียบวิธีต่างๆ เพื่อให้ได้วิธีที่ง่ายที่สุด ข้อเสียของ software นี้คือ มันไม่ได้ใช้เทคนิคที่ใหม่ที่สุดในการทำงาน เทคนิคในการเตรียมข้อมูล และสำรวจที่เสนอมานั้นมักจะไม่ดีพอ

AMADA version 1.2 เป็น tool ที่ถูกออกแบบโดย James Malone เพื่อใช้ในงานวิจัยของ Dr. Ken McGarry ที่มหาวิทยาลัย Sunderland ในเรื่องเกี่ยวกับ data mining โดย tool นี้จะทำงานได้ต้องใช้ MATLAB v 5.2 หรือสูงกว่า

2.4 ข้อดีและข้อเสียของAssociation Rules Discovery

ข้อดีของAssociation Rules Discovery

1) ทำงานได้ดีกับข้อมูลขนาดใหญ่ ขณะที่เทคนิคอื่นจะมีปัญหาในการทำงานกับข้อมูลปริมาณมากๆซึ่งในปัจจุบันมีการจัยเพื่อเพิ่มประสิทธิภาพของ Association Rules Discovery โดยลดจำนวนของตัวแทนข้อมูลและสุ่มตัวอย่างขึ้นมาทำ Data mining อีกด้วย

2) สามารถระบุค่า Minimum Support และ Minimum Confidence ได้ ทำให้สามารถควบคุมจำนวนผลลัพธ์

3) สามารถทำการ mining กับข้อมูลบางส่วนได้ ทำให้ลดปัญหากรณีข้อมูลไม่สมบูรณ์ได้

4) เทคนิคอื่นๆเช่น Decision Trees จะระบุขอบเขตของกลุ่มข้อมูล ทำให้มีการจำกัดข้อมูล มีผลทำให้ข้อมูลที่ถูกเลือกมาอาจจะไม่ใช่ตัวแทนที่แท้จริงของข้อมูล

5) สามารถจัดการกับข้อมูลที่รูปแบบแตกต่างกันได้ โดยไม่สูญเสียสารสนเทศ ในขณะที่เทคนิคอื่นๆจะจำกัดรูปแบบ และความยาวของข้อมูล

6) การคำนวณง่าย และ แสดงผลด้วยสัญลักษณ์ที่ชัดเจนและเข้าใจง่าย

7) สนับสนุนการวิเคราะห์เบื้องต้นเพื่อสร้างแนวทางการทำ Data mining ต่อไป เนื่องจากไม่ต้องกำหนด facto rต่างมากนัก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อเสียของ Association Rules Discovery

1) ถ้าใช้กับข้อมูลที่เกิดขึ้นไม่บ่อย ใน Transaction ทำให้ข้อมูลแยกออกจากกลุ่มข้อมูลอื่นชัดเจน ทำให้ประสิทธิภาพของสารสนเทศที่ได้ลดลง

2) กฎที่ได้มีปริมาณมากเกินไป ถึงแม้จะมีการกำหนดค่า minimum support และ minimum confidence เพื่อจำกัดจำนวนกฎที่สร้างขึ้น แต่ก็อาจจะทำให้กฎที่ได้เพี้ยน หากผู้วิเคราะห์ทำการกำหนดค่าเหล่านี้สูงหรือต่ำเกินไป

3) บอกความแตกต่างของกฎที่ได้มายาก ว่าเป็นกฎจริง หรือได้มาจากการข้อมูลข้องกัน

4) กฎที่ได้ไม่ได้ให้ถึงความเป็นเหตุเป็นผล บอกเพียงแนวโน้มที่จะเกิดขึ้นด้วยกัน

5) การแสกนข้อมูลทั้งฐานข้อมูลล่อนข้างซับซ้อน และคำนวนนาน

6) การกำหนดจำนวนสินค้าที่จะเข้าวิเคราะห์ทำได้ยาก เพราะหากกำหนดมากเกินไปจะทำให้ได้กฎที่มากเกินไป

7) เนื่องจากมีการกำหนด minimum support เพื่อลดจำนวนกฎที่ได้ ในบางครั้งไม่พบข้อมูลเป็นแนวทางใหม่ๆที่ยังเกิดขึ้นน้อย

2.5 สรุป

Association Rule เป็นเทคนิคของ Data mining ในการหาความสัมพันธ์ระหว่างสินค้าในฐานข้อมูลรายการขายกฎที่ได้จะอยู่ในรูปความสัมพันธ์ เช่น A➔B แล้วแสดงค่า support กับ confidence

Apriori Algorithm เป็น Algorithm ที่ใช้ในการหาความสัมพันธ์ของข้อมูลในฐานข้อมูล โดยอาศัยหลักการทำงานเป็นรอบๆซ้ำๆนั่นคือใช้รายการใน k-itemset (set ของข้อมูลที่มีสมาชิก k ตัวซึ่งเรียกสมาชิกแต่ละตัวว่า item) ในการหารายการใน (k+1)-itemset ต่อไป โดยมีการกำหนดค่า minimum support เป็นเงื่อนไขในการเลือกข้อมูลเพื่อทำในรอบต่อๆไปว่าข้อมูลนั้นจะต้องมีค่า support มากกว่าหรือเท่ากับค่า minimum support ที่กำหนด โดยจะทำการวนซ้ำๆไปเรื่อยๆจนกระทั่งไม่สามารถหาชุดรายการที่มากกว่า k รายการ (k-itemset) ได้อีกแล้วก็จะนำชุดรายการจากรอบนั้นๆมาสร้างกฎความสัมพันธ์

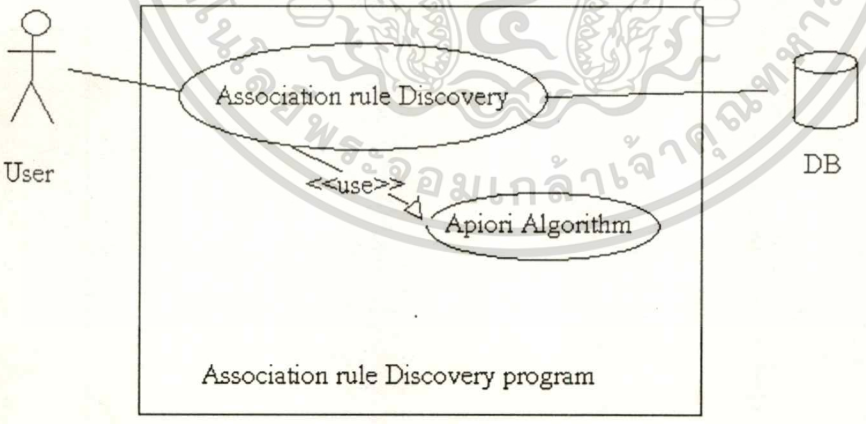
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
การวิเคราะห์และออกแบบระบบ

3.1 ลักษณะของโปรแกรม

โปรแกรมมีลักษณะการทำงานที่จะต้องไปดึงข้อมูลจากฐานข้อมูลที่ผู้ใช้กำหนด แล้วนำข้อมูลที่ได้มาทำการหาความสัมพันธ์ตามแบบ Apriori Algorithm แล้วแสดงผลออกมาเฉพาะความสัมพันธ์ที่อยู่ในขอบเขตที่ผู้ใช้กำหนดไว้ โดยที่สามารถเลือกได้ว่าจะให้ขอมูลในตารางมีการเรียกลำดับการแสดงผลแบบไหนใน 5 แบบที่มีคือ แสดงผลโดยไม่ต้องมีการเรียงลำดับ, แสดงผลโดยเรียงลำดับจากค่า support น้อยสุดไปมากสุด, แสดงผลโดยเรียงลำดับจากค่า support มากสุดไปน้อยสุด, แสดงผลโดยเรียงลำดับจากค่า confidence น้อยสุดไปมากสุด และ แสดงผลโดยเรียงลำดับจากค่า confidence มากสุดไปน้อยสุด

3.2 Use Case Diagram

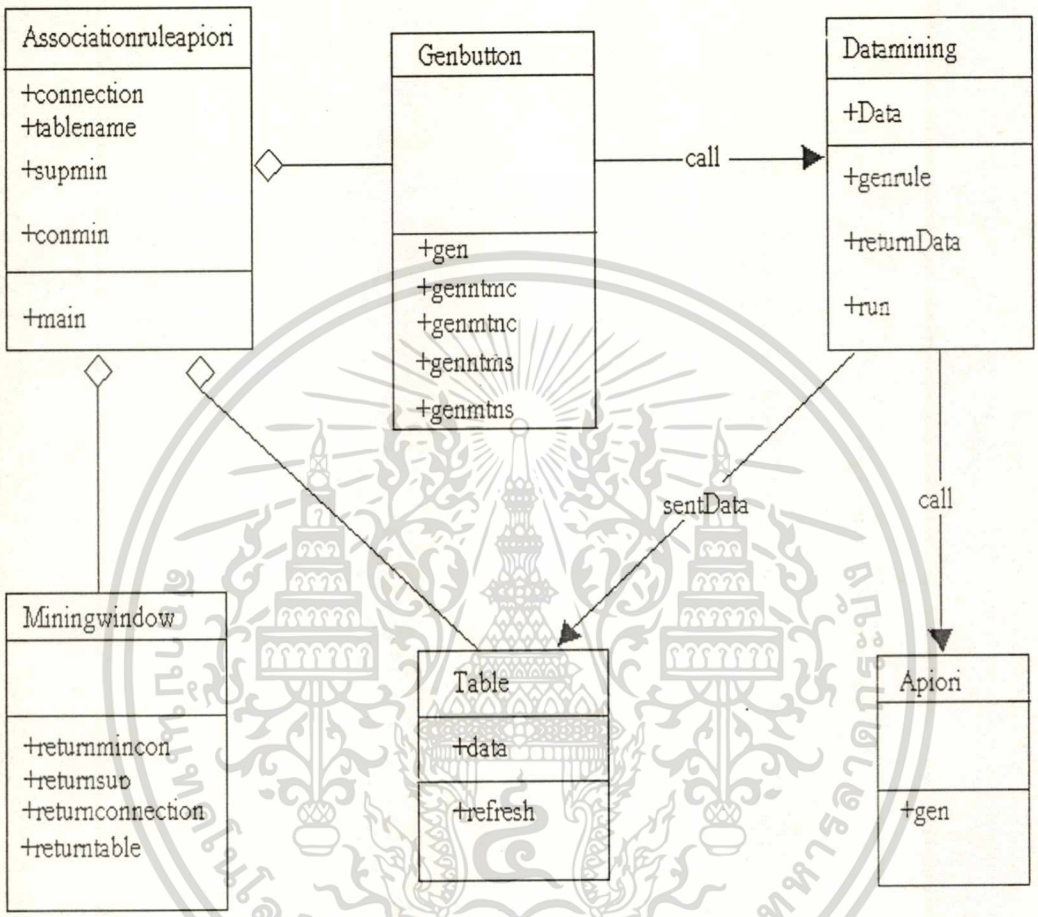


รูปที่ 3.1 Use Case Diagram ของโปรแกรมที่ใช้ในการหา Association rules

ผู้ใช้จะเข้ามาร้องขอระบบให้ทำการสร้าง Association rules ให้ แล้วระบบก้จะไปทำการติดต่อกับฐานข้อมูลที่ผู้ใช้งานบอกแล้วทำการสร้าง Association rules ให้ผู้ที่ร้องขอมา

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3 Class Diagram



รูปที่ 3.2 Class Diagram ของโปรแกรมที่ใช้ในการหา Association rules

จาก Class Diagram จะเห็นว่ามี class อยู่ทั้งหมด 6 class โดยclassที่เป็นclassหลักคือclass Associationruleapiori ซึ่ง classนี้จะเป็น class ที่ใช้ในการเริ่มต้นการทำงานของโปรแกรม class Miningwindow จะทำหน้าที่รับค่า connection, ชื่อ Table, minimum confidence และ minimum support มาจากผู้ใช้งาน แล้วคืนค่ากับมายัง class Associationruleapiori class Genbutton เป็น class ที่ทำหน้าที่ สร้างปุ่มสำหรับเริ่มต้นทำการประมวณผลเพื่อทำ data mining โดย class นี้จะไปเรียกใช้class Datamining ให้ทำการประมวณผล แล้ว class Datamining จะเรียกใช้ class Apiori ในการทำงานบางขั้นตอน หลักจากประมวณผลเสร็จก็จะได้ Association rule ออกมา แล้ว class Dataming ก็จะส่ง Association rule ที่ได้ไปยัง class Table ให้แสดงผลออกมาในรูปแบบตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.1 class Associationruleapiori

จะมีแอตทริบิว connection ทำหน้าที่เก็บค่า Data Source tablename ทำหน้าที่เก็บชื่อตาราง supmin ทำหน้าที่เก็บค่า mininmum support และ conmin เก็บค่า minimum confidence มี method main ใช้เป็น method หลักที่ใช้ในการทำงานของโปรแกรม โดยหลักๆแล้ว class นี้จะเป็น class ที่ใช้บรรจุส่วนต่างๆของโปรแกรม

3.3.2 class Miningwindow

จะทำหน้าที่รับค่า connection โดยใช้ method returnconnection , ชื่อ Table โดยใช้ method return table, minimum confidence โดยใช้ method returnmincon และ minimum support โดยใช้ method returnsup มาจากผู้ใช้งาน แล้วคืนค่ากับมายัง class Associationruleapiori

3.3.3 class Genbutton

เป็น classที่ทำหน้าที่ สร้างปุ่มสำหรับเริ่มต้นทำการประมวณผลเพื่อทำdata mining โดย class นี้จะไปเรียกใช้class Datamining ให้ทำการประมวณผลโดย methodที่อยู่ภายในจะเป็นตัวกำหนดการแสดงผลของข้อมูล โดย method gen จะเป็นการแสดงผลออกมาธรรมดาไม่การการเรียงลำดับของข้อมูลที่ได้ method genmtnc จะเป็นการแสดงผลโดยเรียงลำดับตามค่า confidence จากมากไปน้อย method genntmc จะเป็นการแสดงผลโดยเรียงลำดับตามค่า confidence จากน้อยไปมาก method genmtns จะเป็นการแสดงผลโดยเรียงลำดับตามค่า support จากมากไปน้อย และ method genntms จะเป็นการแสดงผลโดยเรียงลำดับตามค่า support จากน้อยไปมาก

3.3.4 class Datamining

class Datamining เป็น class หลักในการทำ datamining โดยในระหว่างทำงานจะเรียกใช้ class Apiori ในการทำงานบางขั้นตอนเพื่อสร้าง CandidateItemsets หลักจากประมวณผลเสร็จก็จะได้ Association rule ออกมา แล้ว class Dataming ก็จะส่ง Association rule ที่ได้ไปยัง class Table ให้แสดงผลออกมาในรูปแบบตาราง

3.3.5 class Apiori

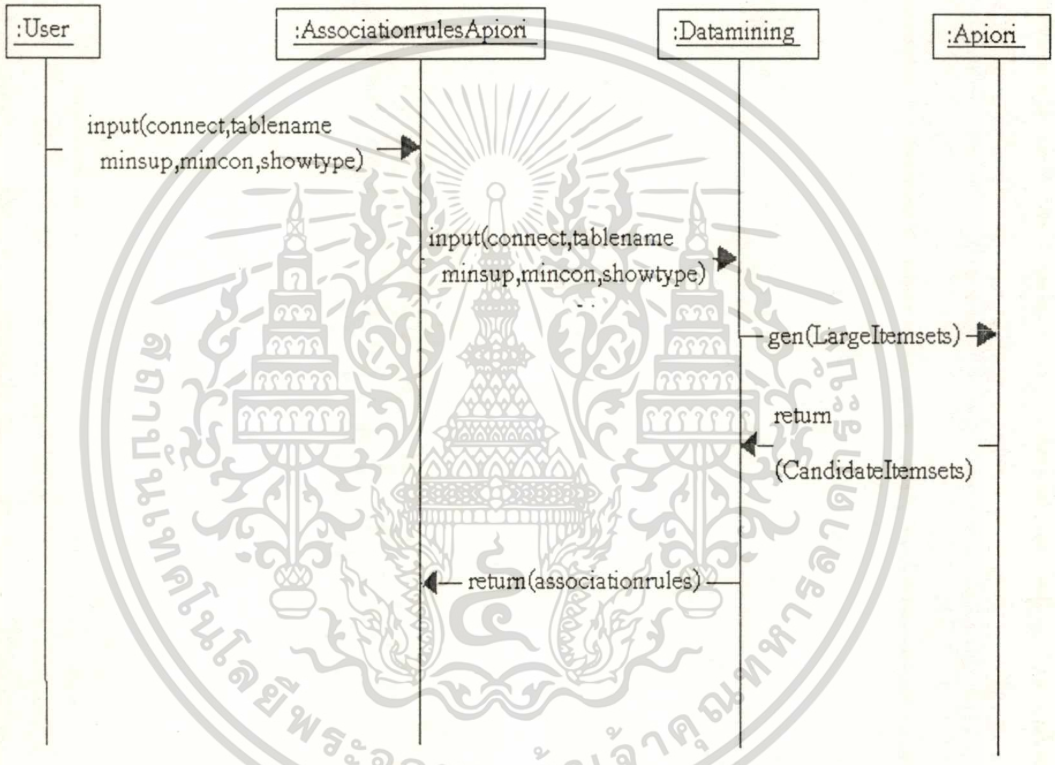
เป็น class ที่ใช้สร้าง CandidateItemset ให้แก่ class Datamining โดยการนำ LargeItemset มาทำการ join กัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.6 class table

Class table จะเป็น class ที่ใช้แสดงผลในรูปแบบของตาราง โดย method ภายในจะแบ่งตามประเภทการแสดงผล

3.4 Sequence Diagram

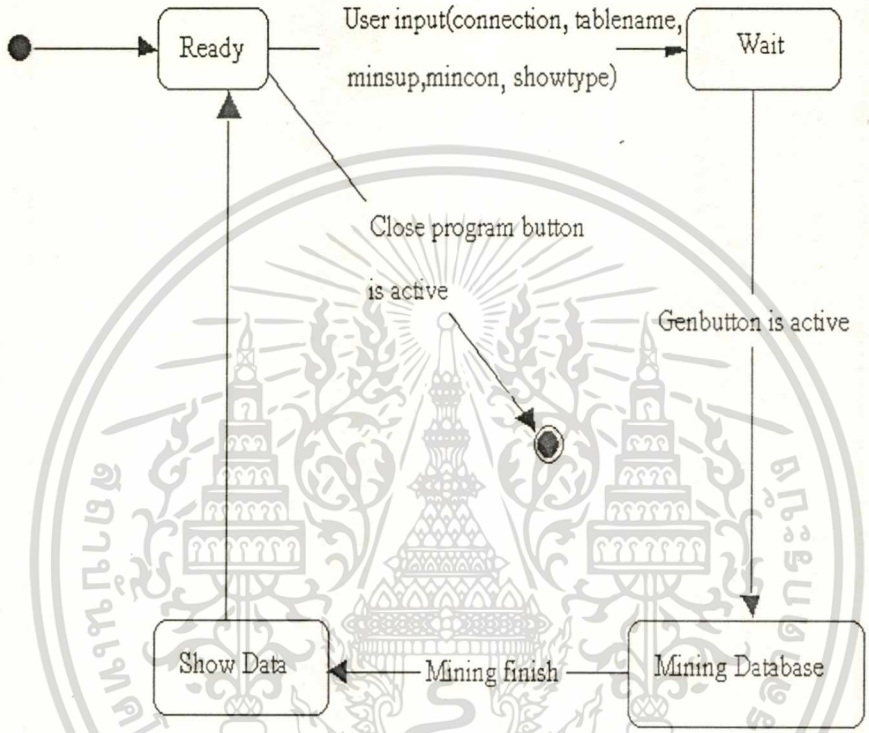


รูปที่ 3.3 Sequence Diagram แสดงขั้นตอนการสร้าง Association rules

เริ่มต้น User จะใส่ค่า connection, ชื่อของ Table, ค่า minimum support, ค่า minimum confidence และรูปแบบในการแสดงผลข้อมูล ให้กับ class AssociationruleApiori แล้ว class นี้ก็จะส่งข้อมูลไปให้ class Datamining ทำการประมวลผล ในระหว่างประมวลผลเมื่อจะทำการสร้าง Candidate Itemsets เพื่อใช้ในการสร้าง Association rule นั้น class Datamining จะส่งค่าของ Large Itemsets ไปให้ class Apiori ทำการสร้าง CandidateItemsetsให้ แล้วส่งกลับมาที่ class Datamining เพื่อทำการประมวลผลสร้างกฎความสัมพันธ์แล้วส่งไปที่ class AssociationruleApiori เพื่อแสดงผลแก่ผู้ใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.5 State Diagram



รูปที่ 3.4 State Diagramของการสร้าง Association rules

เริ่มแรกจะอยู่ในสถานะ Ready เมื่อ User ใส่ข้อมูลของค่า connection, ชื่อของ Table, ค่า minimum support, minimum confidence และรูปแบบในการแสดงผลแล้ว สถานะก็จะเปลี่ยนไปสู่สถานะ wait เพื่อคอย User กดปุ่ม Gen เมื่อUser กดปุ่ม สถานะก็จะเปลี่ยนเป็น Mining Database เพื่อทำการประมวลผลฐานข้อมูลที่ระบุไว้ เมื่อทำการ mining ฐานข้อมูลเสร็จแล้ว ก็จะเปลี่ยนสถานะไปเป็น Show Data คือแสดงผลข้อมูลตามรูปแบบที่ผู้ใช้กำหนดมานั่นเอง แล้วก็จะกลับสู่สถานะ Ready อีกครั้งจากสถานะ Ready นี้ถ้าจะออกจากโปรแกรมก็กดปุ่ม Close program ก็เป็นอันจบการทำงานของโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่4

การพัฒนาระบบงาน

4.1 รูปแบบของข้อมูล

ฐานข้อมูลที่ใช้ได้กับโปรแกรมนี้คือ Microsoft Office Access รูปแบบของฐานข้อมูลจะอยู่ในรูป TABLE (TID,PRODUCE,....) คือ Primary key จะต้องเป็น Composite key โดยที่ตัวแรกเป็น Transaction ID และตัวที่สองเป็นสินค้าที่ซื้อส่วนในรายละเอียดต่อไปจะเป็นอะไรก็ได้ เช่น วันที่ซื้อ เวลาที่ซื้อหรือจำนวนที่ซื้อ เพราะสิ่งที่เราสนใจในการทำ Data mining หรือจำนวนของ Transaction ที่เกิดกับสินค้า ไม่ใช่วันเวลาหรือจำนวนที่ซื้อ

	Field Name	Data Type
🔑	TID	Text
🔑	PRODUCE	Text
	DATE	Date/Time
	TIME	Date/Time
▶	PIECE	Number

รูปที่ 4.1 Design view of database

	TID	PRODUCE	DATE	TIME	PIECE
	0001	beers	1/10/2005	12:01	3
	0001	candy	1/10/2005	12:01	2
	0001	eggs	1/10/2005	12:01	2
	0002	bread	1/10/2005	12:03	3
	0002	candy	1/10/2005	12:03	2
	0002	tissue	1/10/2005	12:04	1
	0003	beers	1/10/2005	12:34	6
	0003	bread	1/10/2005	12:34	1
	0003	candy	1/10/2005	12:34	2
	0003	tissue	1/10/2005	12:35	2
	0004	bread	1/10/2005	13:10	1
✎	0004	tissue	1/10/2005	13:10	2

รูปที่ 4.2 Datasheet view of database

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.2 การติดต่อกับฐานข้อมูล

Database Driver เป็นตัวกลางที่ใช้เชื่อมต่อระหว่างฐานข้อมูลกับโปรแกรมต่างๆ ในระบบปฏิบัติการแบบ windows สามารถเรียกใช้ Database Driver ต่างๆเหล่านี้ได้โดยผ่าน ODBC ในwindowsXp หรือ 2000 จะอยู่ที่ Control Panel -> Administrator Tools แล้วเลือก Data source (ODBC)

ในการติดต่อฐานข้อมูลที่เป็นMicrosoft Accessนั้นจะต้องดำเนินการผ่าน ODBC Driver ซึ่งเราจะต้องสร้าง Data Source ขึ้นมาโดยมีขั้นตอนดังต่อไปนี้

1 เปิด Data Source (ODBC) ขึ้นมาจะได้จอภาพ ODBC Data Source Administrator

2 คลิกปุ่ม Add จะปรากฏจอภาพ Create New Data Source

3 เลือก Microsoft Access Driver

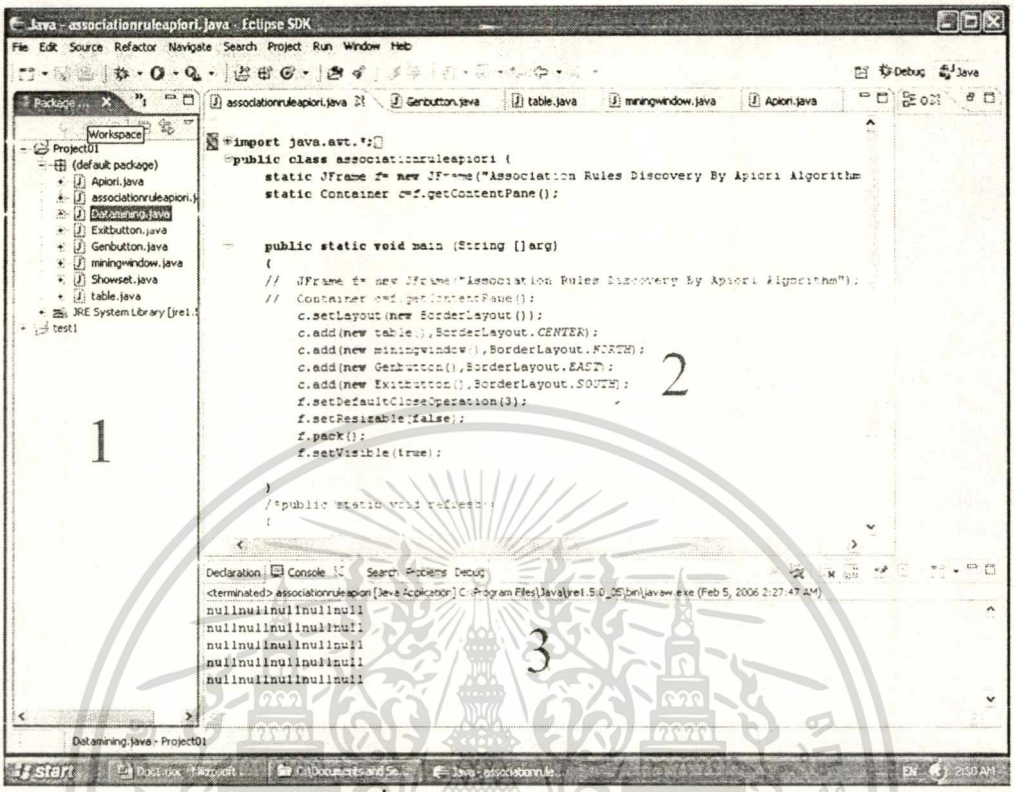
4 คลิกปุ่มFinishจะปรากฏจอภาพ ODBC Microsoft Access Setup โดยที่ ในช่อง Data Source Name ใช้สำหรับตั้งชื่อ Data Source เพื่อใช้เป็นชื่ออ้างอิงในการติดต่อกับฐานข้อมูล ในช่อง Description ใช้แสดงรายละเอียดว่าฐานข้อมูลใช้ทำอะไรบ้าง อาจไม่ระบุก็ได้ หากต้องกำหนด Login และ Password โดยการกดปุ่ม Advanced แล้วก็กำหนดในช่องที่ระบุไว้ คลิกปุ่ม Create สำหรับการสร้างฐานข้อมูลใหม่ หรือ Select ถ้ามีฐานข้อมูลอยู่แล้ว ในค่าในช่องต่างๆ ส่วนตรง Format ให้เลือก Version 4.x เพราะเป็น Version ที่ใช้กับ Microsoft Access Xp และ 2000 เมื่อเรียบร้อยแล้วก็คลิกปุ่ม OK ก็เป็นอันเรียบร้อย

เมื่อเรียบร้อยแล้วที่หน้าจอ ODBC Data Source Adminitrator ในส่วนของ User Data Sources จะมีชื่อ Data Source Name ที่ได้ตั้งชื่อไว้ หมายความว่าเราสามารถจัดการกับฐานข้อมูลผ่านทาง Data Source นี้ได้ทันที

4.3 เครื่องมือที่ใช้ในการพัฒนาโปรแกรม

โปรแกรมนี้มีเครื่องมือในการพัฒนาหลักๆอยู่สองอย่างคือ โดยมี JAVA 1.5.0 เป็นภาษาที่ใช้พัฒนาและ Eclipse 3.1 เป็นเครื่องมือที่ใช้ในการ implement run โปรแกรม และทดสอบโปรแกรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.3 Interface ของ Eclipse

รูปที่ 4.3 เป็นรูปที่แสดงหน้าจอการทำงานของ Eclipse โดยหน้าต่างการทำงานหลักๆจะมีอยู่3ส่วนด้วยกันคือ

1) ส่วนที่แสดงถึง Package Class Attribute และ Method ต่างๆ

2) ส่วนที่ไว้ใช้ในการ coding สำหรับ Class ต่างๆ

3) ส่วนที่เอาไว้ mormitor ค่าต่างที่ให้แสดงผลอออกมา และเป็นส่วนที่แสดง error ต่างๆ อีกด้วย

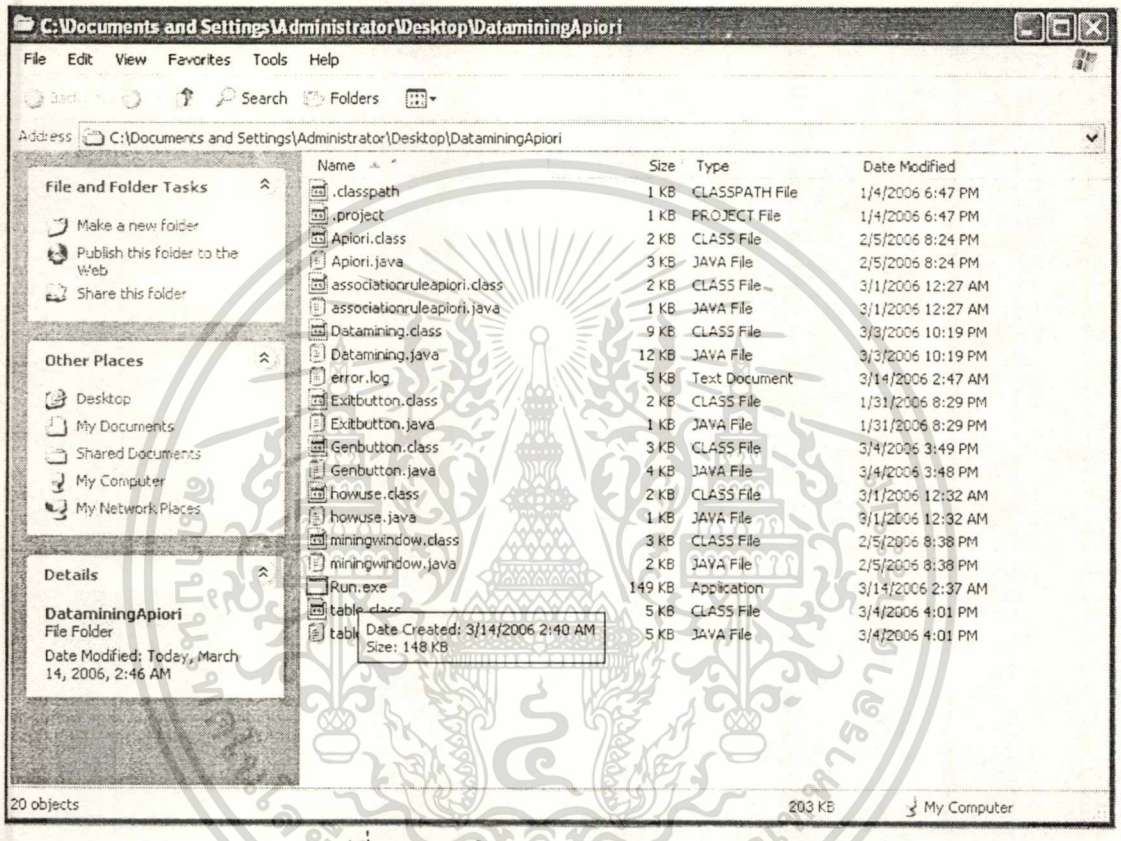
4.4 ระบบที่ใช้ในการทำงาน

ระบบที่เหมาะสมในการทำงานของโปรแกรมนี้ควรจะทำงานบน Window Xp CPU 1.5 GHz RAM 512 MB มีการติดตั้ง JAVA 1.5.0 แต่ในกรณีที่ข้อมูลมีเป็นปริมาณมาก CPU กับ RAM ควรจะมีความสามารถมากกว่านี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.5 การเรียกใช้โปรแกรม

การเรียกใช้โปรแกรมให้ไปที่ Folder Datamining Apriori เมื่อเข้าไปแล้วจะมี file ต่างๆอยู่ ซึ่งจะเป็น class ต่างๆและ Run.exe ให้เลือกที่ Run.exe โปรแกรมก็จะเริ่มทำงาน



รูปที่ 4.4 ภายใน Folder DataminingApiori

4.6 Interface

หน้าจอการทำงานของโปรแกรมนี้มีดังนี้

จากรูปที่ 4.5 ส่วนประกอบต่างๆของ interface มีดังนี้

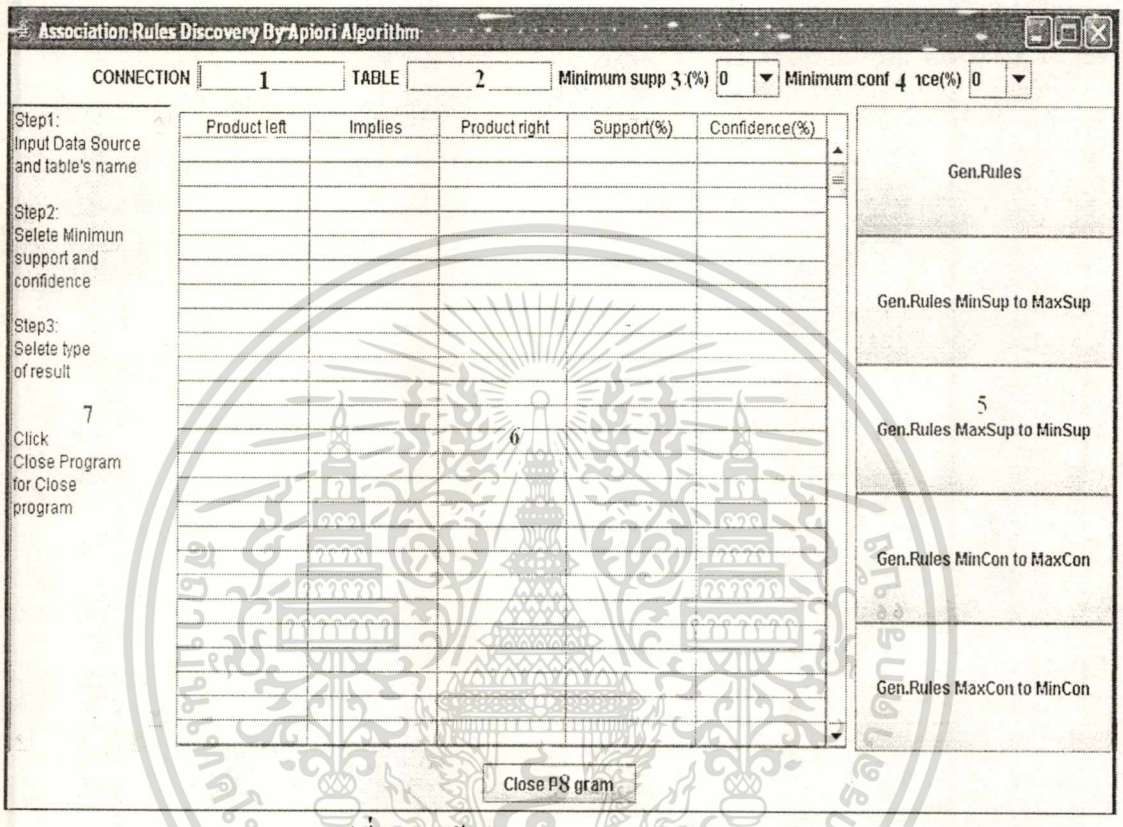
1) เป็นส่วนที่ไว้ใส่ค่า Data source name

2) เป็นส่วนที่ไว้ใส่ชื่อตารางถ้าค่าที่ใส่ไม่ถูกต้องหรือไม่สามารถติดต่อกับฐานข้อมูลได้ ก็จะขึ้นหน้าต่างดังรูปที่ 4.6

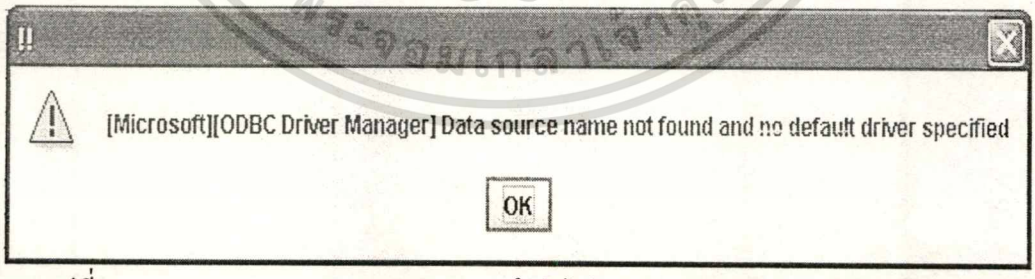
3) เป็นส่วนที่ได้กำหนดค่า Minimum Support โดยค่าที่กำหนดจะอยู่ในรูปแบบเปอร์เซ็น วิธีการกำหนดค่าจะเป็นแบบ Combo Box List ซึ่งจะมีค่าให้เลือกตั้งแต่ 0-100 โดยแต่ระค่าห่างกัน 5 ดังรูปที่ 4.7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4) เป็นส่วนที่ได้กำหนดค่า Minimum Confidence โดยค่าที่กำหนดจะอยู่ในรูปแบบของเปอร์เซ็นวิธีการกำหนดค่าจะเป็นแบบ Combo Box List เช่นเดียวกับในส่วนที่สอง

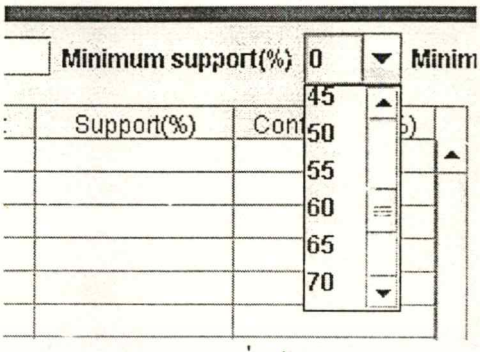


รูปที่ 4.5 หน้าต่างการทำงานของโปรแกรม



รูปที่ 4.6 Warning Dialog message แสดงถึงว่าไม่สามารถดำเนินการกับฐานข้อมูลได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.7 Combo Box List ในที่ใช้ในการกำหนดค่า Support

5) เป็นปุ่มที่ใช้กดเมื่อเราต้องการที่จะให้โปรแกรมทำการหา Association Rules โดยจะมีอยู่ด้วยกัน5ปุ่มซึ่งจะแตกต่างกันในด้านของการแสดงผลของ Association rules ที่ได้ออกมา โดยปุ่มแรก Gen.Rules จะแสดง Association rules ออกมาธรรมดาๆโดยไม่มีการเรียงลำดับใดๆ ,ปุ่มที่สอง Gen.Rules MinSup to MaxSup จะแสดงผลของ Association rules ออกมาโดยเรียงลำดับตามลำดับของค่าsupportจากน้อยไปมาก , ปุ่มที่ 3 Gen.Rules MaxSup to MinSup จะแสดงผลของAssociation rules ออกมาโดยเรียงลำดับตามลำดับของค่า support จากมากไปน้อย , ปุ่มที่ 4 Gen.Rules MinCon to MaxCon จะแสดงผลของ Association rules ออกมาโดยเรียงลำดับตามลำดับของค่า confidence จากน้อยไปมาก และสุดท้ายปุ่มที่ 5 Gen.Rules MaxCon to MinCon จะแสดงผลของ Association rules ออกมาโดยเรียงลำดับตามลำดับของค่า confidence จากมากไปน้อย

6) เป็นส่วนที่แสดง Association rules ที่สร้างขึ้นมาได้ในรูปแบบของตาราง

7) เป็นส่วนที่บอกวิธีการใช้งานโปรแกรมนี้แบบคร่าวๆ

8) เป็นปุ่มปิดการทำการของโปรแกรม

4.7 กรณีการทดสอบการทำงาน

ถ้าข้อมูลในฐานข้อมูลเป็นดังรูปที่ 4.8 แล้วเราทำการหา Association rules โดยกำหมดค่า Minimum support เป็น 0 และค่า Minimum confidence เป็น 0 และให้สร้าง Association rules ออกมาแบบไม่ต้องเรียงลำดับแล้วจะได้ Association rules ออกมาตามรูปที่ 4.9

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	TID	PRODUCE	DATE	TIME	PIECE
▶	0001	beers	1/10/2005	12:01	3
	0001	candy	1/10/2005	12:01	2
	0001	eggs	1/10/2005	12:01	2
	0002	bread	1/10/2005	12:03	3
	0002	candy	1/10/2005	12:03	2
	0002	tissue	1/10/2005	12:04	1
	0003	beers	1/10/2005	12:34	6
	0003	bread	1/10/2005	12:34	1
	0003	candy	1/10/2005	12:34	2
	0003	tissue	1/10/2005	12:35	2
	0004	bread	1/10/2005	13:10	1
	0004	tissue	1/10/2005	13:10	2

รูปที่ 4.8 ข้อมูลในฐานข้อมูล 1

Product left	Implies	Product right	Support(%)	Confidence(%)
[beers]	======>	[candy]	0.5	1.0
[candy]	======>	[beers]	0.5	0.6666666666...
[bread]	======>	[candy]	0.5	0.6666666666...
[candy]	======>	[bread]	0.5	0.6666666666...
[bread]	======>	[tissue]	0.75	1.0
[tissue]	======>	[bread]	0.75	1.0
[candy]	======>	[tissue]	0.5	0.6666666666...
[tissue]	======>	[candy]	0.5	0.6666666666...
[bread]	======>	[candy, tissue]	0.5	0.6666666666...
[bread, candy]	======>	[tissue]	0.5	1.0
[bread, tissue]	======>	[candy]	0.5	0.6666666666...
[candy]	======>	[bread, tissue]	0.5	0.6666666666...
[candy, tissue]	======>	[bread]	0.5	1.0
[tissue]	======>	[bread, candy]	0.5	0.6666666666...

รูปที่ 4.9 Association rules ที่ได้จากข้อมูลในรูป 4.7 โดยมีค่า Minsup และ Mincon = 0%

แต่ถ้าเรากำหนดค่าค่า Minimum support เป็น 60 และค่า Minimum confidence เป็น 60 แล้ว จะได้ Association rules ออกมาตามรูปที่ 4.10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Product left	Implies	Product right	Support(%)	Confidence(%)
[bread]	=======>	[tissue]	0.75	1.0
[tissue]	=======>	[bread]	0.75	1.0

รูปที่ 4.10 Association rules ที่ได้จากข้อมูลในรูป 4.7 โดยมีค่า Minsup และ Mincon = 60%

แต่ถ้าข้อมูลในฐานข้อมูลเป็นดังรูปที่ 4.11 แล้วเราทำการหา Association rules โดยกำหนดค่า Minimum support เป็น 0 และค่า Minimum confidence เป็น 0 แล้วจะได้ Association rules ออกมาตามรูปที่ 4.12

TID	PRODUCE	PRICE
0001	Apple	10
0001	Beer	20
0001	Coke	15
0001	Visky	70
0002	Beer	20
0002	Coke	15
0003	Apple	10
0003	Eeg	8
0003	Fish	50
0004	Beer	20
0004	Coke	15
0004	Fish	50
0004	Visky	70
0005	Coke	15
0006	Apple	10
0007	Apple	10
0007	Coke	15
0007	Eeg	8
0007	Fish	50
0008	Beer	20
0008	Coke	15
0008	Visky	70
0009	Apple	10
0009	Fish	50
0010	Beer	20
0010	Fish	50

รูปที่ 4.11 ข้อมูลในฐานข้อมูล 2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จะสังเกตุได้ว่าค่า Minimum support กับค่า Minimum confidence จะเป็นตัวกำหนดจำนวน Association rules ที่ได้ออกมา ถ้าเรากำหนดค่าไว้ต่ำ Association rules ที่ได้ก็จะมีมากเกินไป แต่ถ้ากำหนดค่าทั้งสองไว้สูงเกินไป Association rules ที่ได้ก็จะน้อยเกินไปหรืออาจไม่มีเลย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการดำเนินงาน

5.1 ผลการดำเนินงาน

ระบบที่สร้างขึ้นมาสามารถสร้าง Association rules จากฐานข้อมูลใดๆที่มีรูปแบบตามที่กำหนดได้ โดยแสดงข้อมูลออกมาตามขอบเขตของค่า Support และ Confidence ที่ตั้งไว้ โดยที่ Association rules จะแสดงออกมาผ่านตาราง และถ้าผู้ใช้งานต้องการก็สามารถเรียงลำดับการแสดงผลโดยเลือกได้ว่าจะเรียงที่ค่า Support จากมากไปน้อยหรือน้อยไปมาก และ เรียงที่ค่า Confidence จากมากไปน้อยหรือน้อยไปมาก

5.2 สรุปผลการทดลอง

จากค่าของ Association rules ที่แสดงผลออกมาตามรายการสินค้าจากหัวข้อ 4.7 ถ้าความสัมพันธ์ไหนมีค่า Support และ Confidence มาก ก็จะมีโอกาสเกิดเหตุการณ์นั้นมากกว่าความสัมพันธ์ที่มีค่าทั้งสองน้อย

5.3 การประยุกต์ใช้งาน

การนำเอาเทคนิค Data mining เข้ามาใช้ในการวิเคราะห์หาสารสนเทศที่ซ่อนอยู่ภายในข้อมูลจำนวนมาก และวิเคราะห์หาความสัมพันธ์ของข้อมูล กฎที่ได้จากการหาความสัมพันธ์นั้นจำเป็นที่จะต้องอาศัยผู้ที่มีประสบการณ์ในด้านนั้นๆ ในการตีความหมายของกฎเหล่านั้น เนื่องจากกฎที่ได้มีเป็นจำนวนมากซึ่งไม่จำเป็นว่าทุกกฎที่ได้จะมีประโยชน์ หรือ สามารถวิเคราะห์หาความสัมพันธ์ได้จริง ดังนั้นกฎนี้จึงเป็นเพียงส่วนหนึ่งที่ช่วยในการตัดสินใจเท่านั้น ซึ่งยังมีอีกหลายปัจจัยที่ใช้ประกอบกับการณ์กำหนดกลยุทธ์ หรือนโยบายทางการตลาด ไม่ว่าจะเป็น สภาพการดำเนินงานของบริษัท ปัจจัยทางเศรษฐกิจ ฯลฯ มาใช้ประกอบเพื่อใช้เป็นแนวทางในการกำหนดเป้าหมายเพื่อดำเนินงานขององค์กรต่อไป

5.4 ปัญหาและอุปสรรค

เนื่องจากว่า Algorithm ที่ใช้มีการคำนวนสูง เพื่อทำการสร้างรายการ Candidate Itemset ซึ่งในแต่ละรอบการทำงานจะใช้ CPU สูง ทำให้บางครั้งเมื่อมีการกำหนดความสัมพันธ์มากๆจะทำให้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เกิดการหยุดทำงานได้ หรือ เครื่อง Hang

5.5 ข้อเสนอแนะ

สำหรับโครงงานนี้ เป็นระบบที่ใช้หาความสัมพันธ์ในฐานข้อมูลใดๆก็ได้ เพียงแต่รูปแบบของฐานข้อมูลต้องเป็นไปตามที่กำหนด ผู้ที่ใช้งานสามรถนำกฎความสัมพันธ์ที่ได้ไปประกอบการตัดสินใจ หรือจะวิเคราะห์ที่มาของกฎความสัมพันธ์ว่าเหตุใดถึงทำให้เกิดกฎนี้ เพื่อเป็นองค์ความรู้ใหม่ขององค์กร

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

Bill Palace.1996. **Data Mining** . UCLA :Anderson Graduate School of Management at Barry

Devlin.1997. **DATA WAREHOUSE From architecture to implementation.** Addison-Wesley

Elisa Bertino, Barbara Catania ,Gian Piero Zarri.2001. **Intelligent Database Syste.** Addison-Wesley

Margaret H. Dunham, Yongqiao Xiao, Le Gruenwald, Zahid Hossain.2000. **A SURVEY OF ASSOCIATION RULES** . Texas: Department of Computer Science and Engineering Southern Methodist University Dallas

Rakesh Agrawal, RMkrishnan Srikant.1999 **Fast Algorithm for Mining Association Rules.** Madison:Department of Computer Science University of Wisconsin

J.Loughry,J.I van Hemert,L.Schoofs.2000 **Efficiently Enumerating the Subset of Set.**Colorado: Department 3740, Mail Stop X-3741,P.O.BOX 179,Denver

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นายทศพล คงน้อย
วันเกิด	16 มีนาคม พ.ศ. 2524
สถานที่เกิด	กรุงเทพมหานครฯ
ประวัติการศึกษา	ระดับมัธยม โรงเรียน เตรียมอุดมศึกษาพัฒนาการ
	ระดับปริญญาตรี สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
	คณะ วิทยาศาสตร์ แขนงวิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์
	สาขา วิทยาการคอมพิวเตอร์
	ปัจจุบัน ระดับวิทยาศาสตรมหาบัณฑิต สถาบันพระจอมเกล้าเจ้าคุณทหาร
	ลาดกระบัง คณะ เทคโนโลยีสารสนเทศ สาขาเทคโนโลยีสารสนเทศ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Product left	Implies	Product right	Support(%)	Confidence(%)
[Apple]	======>	[Coke]	0.2	0.4
[Coke]	======>	[Apple]	0.2	0.3333333333...
[Apple]	======>	[Eeg]	0.2	0.4
[Eeg]	======>	[Apple]	0.2	1.0
[Apple]	======>	[Fish]	0.3	0.6
[Fish]	======>	[Apple]	0.3	0.6
[Beer]	======>	[Coke]	0.4	0.8
[Coke]	======>	[Beer]	0.4	0.6666666666...
[Beer]	======>	[Fish]	0.2	0.4
[Fish]	======>	[Beer]	0.2	0.4
[Beer]	======>	[Visky]	0.3	0.6
[Visky]	======>	[Beer]	0.3	1.0
[Coke]	======>	[Fish]	0.2	0.3333333333...
[Fish]	======>	[Coke]	0.2	0.4
[Coke]	======>	[Visky]	0.3	0.5
[Visky]	======>	[Coke]	0.3	1.0
[Eeg]	======>	[Fish]	0.2	1.0
[Fish]	======>	[Eeg]	0.2	0.4
[Apple]	======>	[Eeg, Fish]	0.2	0.4
[Apple, Eeg]	======>	[Fish]	0.2	1.0
[Apple, Fish]	======>	[Eeg]	0.2	0.6666666666...
[Eeg]	======>	[Apple, Fish]	0.2	1.0
[Eeg, Fish]	======>	[Apple]	0.2	1.0
[Fish]	======>	[Apple, Eeg]	0.2	0.4
[Beer]	======>	[Coke, Visky]	0.3	0.6

รูปที่ 4.12 Association rules ที่ได้จากข้อมูลในรูป 4.10 โดยมีค่า Minsup และ Mincon = 0%

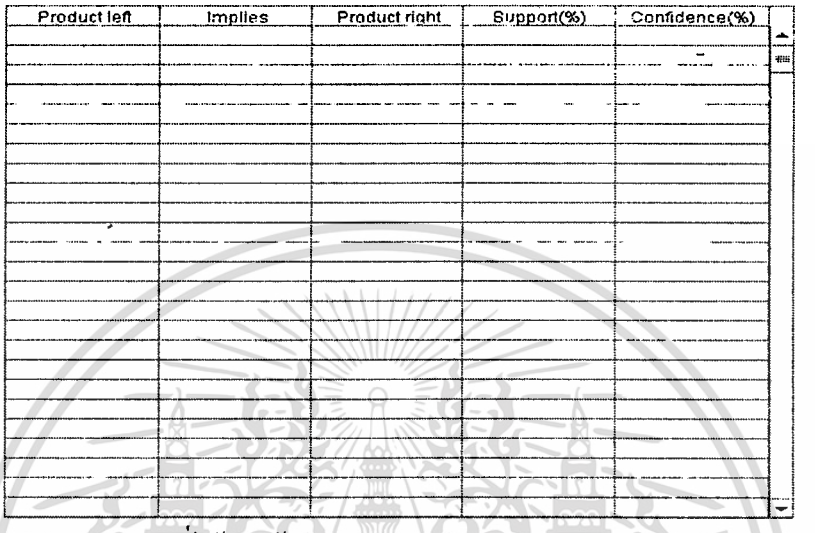
ถ้าเรากำหนดค่าค่า Minimum support เป็น 25 และค่า Minimum confidence เป็น 60 แล้วจะได้ Association rules ออกมาตามรูปที่ 4.13

Product left	Implies	Product right	Support(%)	Confidence(%)
[Apple]	======>	[Fish]	0.3	0.6
[Fish]	======>	[Apple]	0.3	0.6
[Beer]	======>	[Coke]	0.4	0.8
[Coke]	======>	[Beer]	0.4	0.6666666666...
[Beer]	======>	[Visky]	0.3	0.6
[Visky]	======>	[Beer]	0.3	1.0
[Visky]	======>	[Coke]	0.3	1.0
[Beer]	======>	[Coke, Visky]	0.3	0.6
[Beer, Coke]	======>	[Visky]	0.3	0.75
[Beer, Visky]	======>	[Coke]	0.3	1.0
[Coke, Visky]	======>	[Beer]	0.3	1.0
[Visky]	======>	[Beer, Coke]	0.3	1.0

รูปที่ 4.13 Association rules ที่ได้จากข้อมูลในรูป 4.10 โดยมีค่า Minsup = 25 และ Mincon = 60%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แต่ถ้าเรากำหนดค่า Minimum support เป็น 50 และค่าMinimum confidence เป็น 60 แล้วจะได้ Association rules ออกมาตามรูปที่ 4.14



Product left	Implies	Product right	Support(%)	Confidence(%)

รูปที่ 4.14 Association rules ที่ได้จากข้อมูลในรูป 4.10 โดยมีค่า Minsup = 50 และ Mincon = 60%

ถ้าเรากำหนดค่า Minimum support เป็น 30 และค่า Minimum confidence เป็น 50 และให้แสดงผลออกมาแบบเรียงลำดับโดยดูจากค่า confidence จากน้อยไปมาก แล้วจะได้ Association rules ออกมาตามรูปที่ 4.15

Product left	Implies	Product right	Support(%)	Confidence(%)
[Coke]	======>	[Visky]	0.3	0.5
[Coke]	======>	[Beer, Visky]	0.3	0.5
[Beer]	======>	[Coke, Visky]	0.3	0.6
[Beer]	======>	[Visky]	0.3	0.6
[Fish]	======>	[Apple]	0.3	0.6
[Apple]	======>	[Fish]	0.3	0.6
[Coke]	======>	[Beer]	0.4	0.6666666666...
[Beer, Coke]	======>	[Visky]	0.3	0.75
[Beer]	======>	[Coke]	0.4	0.8
[Beer, Visky]	======>	[Coke]	0.3	1.0
[Visky]	======>	[Coke]	0.3	1.0
[Coke, Visky]	======>	[Beer]	0.3	1.0
[Visky]	======>	[Beer, Coke]	0.3	1.0
[Visky]	======>	[Beer]	0.3	1.0

รูปที่ 4.15 Association rules ที่ได้จากข้อมูลในรูป 4.10 โดยมีค่า Minsup = 30 และ Mincon = 50% และแสดงผลออกมาแบบเรียงลำดับ โดยดูจากค่า confidence จากน้อยไปมาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้