2.1.1 วิวัฒนาการของเหมืองข้อมูล

ใน ปี ค.ศ. 1960 Data Collection มีการนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือเพื่อป้องกันการสูญหายได้เป็นอย่างดี ต่อมาในปี ค.ศ. 1980 Data Access มีการนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์ระหว่างกันเพื่อนำไปวิเคราะห์และตัดสินใจอย่างมีประสิทธิภาพ ต่อมาในปี ค.ศ. 1990 Data Warehouse and Decision Support มีการนำข้อมูลมาเก็บลงในฐานข้อมูลขนาดใหญ่ครอบคลุมการใช้งานทั้งหมดขององค์กรเพื่อ ช่วยสนับสนุนการตัดสินใจ และในปี ค.ศ. 2000 เหมืองข้อมูล มีการนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผลโดยสร้างแบบจำลองและความสัมพันธ์ทางสถิติ

2.1.2 วัตถุประสงค์ในการใช้เหมืองข้อมูล

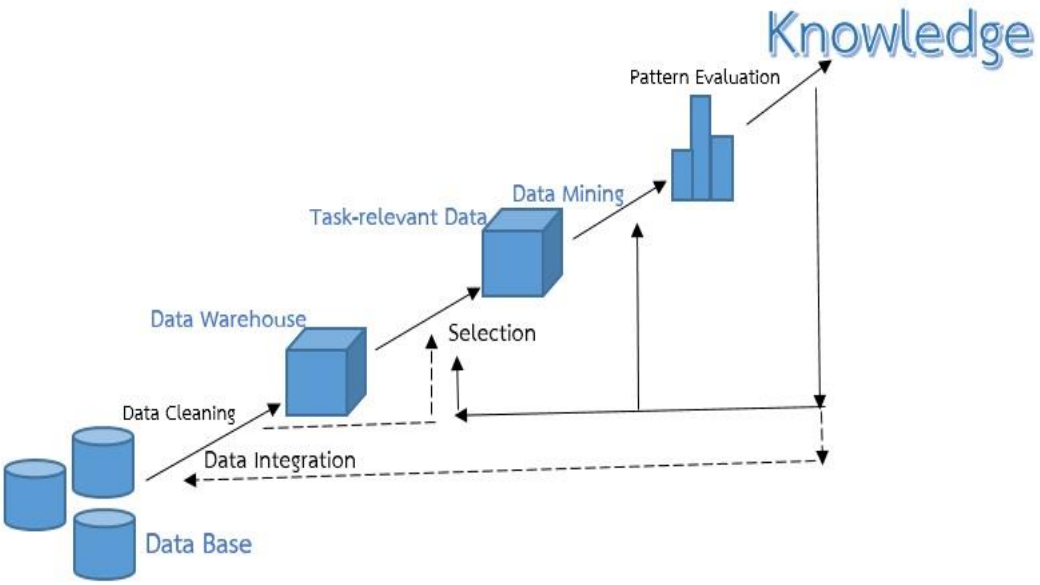
1. เพื่อการค้นหาองค์ความรู้ใหม่ในฐานข้อมูล (Knowledge Discovery in Databases)
2. เพื่อการสกัดองค์ความรู้ที่ซ่อนเร้นอยู่ (Knowledge Extraction)
3. เพื่อจัดการกับข้อมูลในอดีต (Data Archeology)
4. เพื่อสำรวจข้อมูล (Data Exploration)
5. เพื่อค้นหา Pattern ของข้อมูลที่ซ่อนอยู่ (Data Pattern Processing)
6. เพื่อใช้ขุดเจาะข้อมูล (Data Dredging)
7. เพื่อให้ได้มาซึ่งสารสนเทศที่มีประโยชน์

2.1.3 เป้าหมายหลักของเหมืองข้อมูล

คุณลักษณะและเป้าหมายหลักของเหมืองข้อมูล คือ ใช้ค้นหา Pattern ของข้อมูลที่ฝังลึกและซ่อนเร้นอยู่ภายในฐานข้อมูลขนาดใหญ่โดยใช้สถาปัตยกรรม Client-Server ใช้เครื่องมือสมัยใหม่ที่สามารถแสดงผลแบบกราฟิกผู้ใช้สามารถดูข้อมูลแบบเจาะลึกและสามารถใช้เครื่องมือในการสอบถามข้อมูลได้อย่างง่ายดายโดยไม่ต้องอาศัยความชำนาญของผู้พัฒนาโปรแกรมเพราะเครื่องมือถูกออกแบบมาให้ใช้งานได้ง่ายบ่อยครั้งอาจค้นพบผลลัพธ์ที่ไม่คาดหวังมาก่อน

2.1.4 กระบวนการทำเหมืองข้อมูล

เป็นกระบวนการในการค้นหาลักษณะแฝงของข้อมูล (Pattern) ที่ซ่อนอยู่ในฐานข้อมูลโดยมีขั้นตอนดังภาพประกอบที่ 2.1

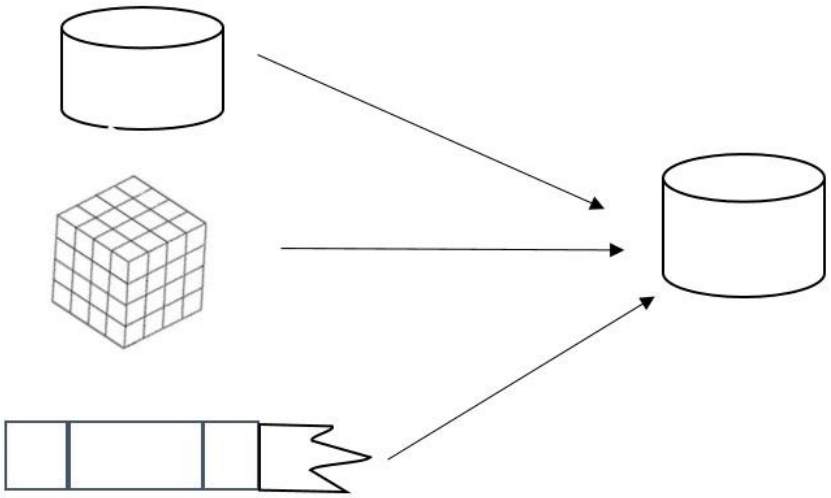


ภาพประกอบที่ 2.1 กระบวนการของเหมืองข้อมูล

ขั้นตอนของการทำเหมืองข้อมูลมี 5 ขั้นตอนดังนี้

1. Data Cleaning เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไปโดยทั่วไปข้อมูลที่จัดเก็บอาจมีความผิดปกติต่างๆได้เช่นข้อมูลบางแอทริบิวต์ขาดหายไป (missing value) ขาดแอทริบิวต์ที่น่าสนใจหรือขาดรายละเอียดของข้อมูลเป็นข้อมูลรบกวน (noisy data) เช่น ข้อมูลมีค่าผิดพลาด (error) หรือมีค่าผิดปกติ (Outliers)

2. Data Integration เป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน เช่นมีข้อมูลในคลังข้อมูล (Data Warehouse) ในรูปแบบของดาต้า คิวบ์ (Data Cube) และมีข้อมูลในรูปแบบฐานข้อมูลเชิงสัมพันธ์ (Relational Database) จำเป็นต้องทำการรวมข้อมูลให้เป็นข้อมูลชุดเดียวกันดังภาพประกอบที่ 2.2



ภาพประกอบที่ 2.2 กระบวนการของ Data Integration

3. Data Selection เป็นขั้นตอนระบุถึงแหล่งข้อมูลที่จะนำมาทำ Mining รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูลเพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น

4. Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่โดยใช้เทคนิคและอัลกอริทึมที่เหมาะสมกับข้อมูลที่มีอยู่

5. Pattern Evaluation เป็นขั้นตอนประเมินผล Pattern และนำเสนอองค์ความรู้ในขั้นตอนนี้จะเป็นการวิเคราะห์ผลลัพธ์ที่ได้แปลความหมายและประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่และนำเสนอ

2.1.5 ประเภทของข้อมูลที่ใช้ในการทำเหมืองข้อมูล

1. ข้อมูลที่มาจากฐานข้อมูลเชิงสัมพันธ์ (Relational databases)
2. ข้อมูลจากคลังข้อมูล (Data warehouses)
3. ข้อมูลจากฐานข้อมูลรายการปรับปรุง (Transactional databases)
4. จากฐานข้อมูลพิเศษหรือที่เก็บข่าวสารพิเศษซึ่งได้แก่
   - ฐานข้อมูลเชิงวัตถุ
   - ฐานข้อมูลเกี่ยวกับเวลา
   - ฐานข้อมูลข้อความ (Text databases) และฐานข้อมูลมัลติมีเดีย
   - ฐานข้อมูลแบบเก่าในอดีตหรือข้อมูลที่มาจากต่างฐานข้อมูลกัน
   - ข้อมูลจากแหล่ง World Wide Web

2.1.6 ชนิดขององค์ความรู้ที่ค้นพบมีดังนี้

1. องค์ความรู้เกี่ยวกับคุณลักษณะของข้อมูล (Characterization) เช่นรู้ว่าคนที่สามารถเรียนต่อในระดับปริญญาเอกได้จะพิจารณาได้จากคุณลักษณะใด

2. องค์ความรู้เกี่ยวกับการจำแนกข้อมูล (Discrimination)

3. องค์ความรู้เกี่ยวกับความสัมพันธ์ของข้อมูล (Association) เช่นมีความสัมพันธ์ของการซื้อสินค้าพบว่าถ้าลูกค้าซื้อป๊อบคอร์น จะต้องซื้อเป๊ปซี่ตามมา

4.องค์ความรู้เกี่ยวกับการแยกประเภทข้อมูล และการพยากรณ์ (Classification/prediction)

5. องค์ความรู้เกี่ยวกับการจัดกลุ่มข้อมูล (Clustering)

6. องค์ความรู้เกี่ยวกับการวิเคราะห์ข้อมูลที่ผิดปกติ (Outlier analysis)

7. องค์ความรู้เกี่ยวกับข้อมูลอื่นๆในงานที่ค้นพบ (Other data mining tasks)

2.1.7 การทำเหมืองข้อมูล สามารถมีขั้นตอนในการขุดค้นข้อมูลดังนี้

1. การวิเคราะห์คุณสมบัติและการแยกแยะข้อมูล (Characterization and Discrimination) การวิเคราะห์คุณสมบัติเช่นการพิจารณารับสมัครพนักงานของฝ่ายทรัพยากรบุคคลต้องวิเคราะห์คุณสมบัติจากใบสมัครและการสัมภาษณ์ การแยกแยะข้อมูล เช่น การสูญเสียการได้ยินเป็นปัญหาสำคัญของคนงานที่ทำงานในโรงงานที่มีเสียงดังซึ่งการทราบถึงสมรรถภาพการได้ยินของตนเองจะช่วยให้คนงานสามารถป้องกันตนเอง จากการสูญเสียการได้ยินได้

2. การหาความสัมพันธ์ของข้อมูล (Association) นิยมใช้ในการวิเคราะห์ตะกร้าสินค้า เพราะสามารถช่วยให้ค้นพบความสัมพันธ์ที่เป็นไปได้ทั้งหมดของการผสมกันของสินค้าที่น่าสนใจเพื่อนำผลที่ได้ไปใช้ทำการตลาด

3. การจัดหมวดหมู่และการวิเคราะห์การถดถอย (Classification/ Regression) การจัดหมวดหมู่ (Classification) ตัวอย่างของการจัดหมวดหมู่ที่นำมาใช้กับงานด้านธุรกิจ เช่น การวิเคราะห์ลูกค้าที่ซื่อสัตย์จงรักภักดีต่อยี่ห้อสินค้า (Brand Loyalty) ขององค์กรเทคนิคของเหมืองข้อมูลที่ใช้ในการแก้ปัญหาแบบ Classification ได้แก่ Decision Tree, Neural Networks, Naïve-Bayes และ K-nearest neighbor (K-NN) ปัญหาแบบ Regression จะเหมือนกับแบบ Classification ต่างกันตรงที่ผลลัพธ์ที่ได้จาก Regression เป็นค่าแน่นอนที่ไม่จำกัดจะเป็นค่าอะไรก็ได้ เช่น แบบจำลองทำนายว่านาย B จะตอบรับข้อเสนอของบริษัท ถ้านาย B ได้รับผลกำไร 1,000 บาท (1,000 เป็นคำตอบเฉพาะที่แน่นอนแต่ไม่จำกัดซึ่งตัวเลขอาจจะเป็นค่าอื่นไปได้เรื่อยๆ ต่างจากคำตอบแบบ Yes, No)

4. การวิเคราะห์การรวมกลุ่มหรือการแบ่งแยกข้อมูล (Cluster analysis / Segmentation) การวิเคราะห์การรวมกลุ่ม (Clustering) เป็นการรวมกลุ่มข้อมูลที่มีลักษณะเหมือนกันรูปแบบหรือแนวโน้มที่จะเหมือนกันการใช้เทคนิค Clustering จะไม่มีผลลัพธ์ (Output) ไม่มีตัวแปรอิสระ (Independent Variable) ไม่มีการจัดโครงร่างของวัตถุเราจะเรียกเทคนิคของ Clusteringว่าเป็นแบบเรียนรู้ข้อมูลโดยไม่ต้องอาศัยครูสอน (Unsupervied Learning) การทำ Clustering จะทำบนพื้นฐานของข้อมูลในอดีต ตัวอย่างเช่นองค์กรต้องการทราบความเหมือนที่มีในกลุ่มของลูกค้าของตนเพื่อที่จะให้เข้าใจลักษณะเฉพาะของลูกค้ากลุ่มเป้าหมายและสร้างกลุ่มของลูกค้าเพื่อที่องค์กรจะได้สามารถขายสินค้าได้ในอนาคตองค์กรจะทำการแยกกลุ่มของข้อมูลลูกค้าออกเป็นกลุ่มๆ (หาส่วนที่เป็นIntersection และ Union) เทคนิคของเหมืองข้อมูล เพื่อแก้ปัญหาแบบ Clustering คือวิธี Demographic Clustering กับ Neural Clustering

5. การประเมินและการพยากรณ์ (Estimation/Prediction) การประเมิน (Estimation) เป็นการประเมินที่ไม่สามารถกำหนดค่าหรือคุณสมบัติที่ชัดเจนได้ใช้จัดการกับค่าที่มีผลแบบต่อเนื่องเช่นใช้ประเมินรายได้ของครอบครัวประเมินความสูงของบุคคลในครอบครัวประเมินจำนวนเด็กๆในครอบครัวการพยากรณ์ (Prediction) จะเหมือนกับ Classification และ Estimation ต่างกันตรงที่ Record ถูกแยกจัดลำดับในการทำนาย ค่าในอนาคตและนำข้อมูลในอดีตมาสร้างเป็นแบบจำลองใช้ทำนายสิ่งที่จะเกิดขึ้นในอนาคตเช่นการทำนายว่าลูกค้ากลุ่มใดที่องค์กรจะสูญเสียไปในอีก 6 เดือนข้างหน้าหรือการทำนายยอดซื้อของลูกค้าจะเป็นเท่าใด ถ้าบริษัทลดราคาสินค้าลง 10%

6. การบรรยายและการแสดงภาพของข้อมูล (Description / Visualization) การบรรยาย (Description) เป็นการหาคำอธิบายถึงสิ่งที่จะเกิดขึ้นโดยอาศัยข้อมูลจากฐานข้อมูล เช่น กลุ่มคนที่มีการศึกษาหรือรายได้น้อยจะเลือกนักการเมืองที่มีนโยบายทุนนิยมมากกว่ากลุ่มคนชั้นกลาง การแสดงภาพของข้อมูล (Visualization) เป็นการนำเสนอข้อมูลในรูปแบบกราฟิกหรืออาจนำเสนอในแบบ 2มิติสร้างรายละเอียดในการนำเสนอให้เข้าใจมากยิ่งขึ้น เช่นองค์กรต้องการหาสถานที่ในขยายสาขาใหม่ที่อยู่ในเขตพื้นที่ภาคเหนือของประเทศ ดังนั้นองค์กรจึงใช้แผนที่ Plot ที่ตั้งขององค์กรคู่แข่งที่มีสาขาอยู่ในเขตนั้นเพื่อพิจารณาสถานที่ตั้งที่เหมาะสมที่สุด

2.1.8 การประยุกต์ใช้เหมืองข้อมูล กับงานด้านธุรกิจสามารถนำเทคนิคของเหมืองข้อมูลไปวิเคราะห์ข้อมูลในฐานข้อมูลเพื่อนาข้อมูลที่ได้ไปใช้ประโยชน์ในงานด้านต่างๆ ดังต่อไปนี้

1. งานด้านการตลาด (Marketing) เช่นการทำ Promotion ส่งเสริมการขาย

2. งานด้านธนาคาร และการเงิน (Banking / Financial Analysis) เช่นใช้ในการวิเคราะห์การให้สินเชื่อแก่ลูกค้าการจัดทำ Package ในการกู้ยืมการทำนายอัตราการจ่ายเงินกู้การแบ่งกลุ่มลูกค้าเพื่อหาเป้าหมายทางการตลาด (ลูกค้าชั้นดี)

3. งานด้านการขายปลีก (Retailing and sales) เป็นงานที่มีการเก็บข้อมูลจำนวนมากประยุกต์ใช้เพื่อหากลยุทธ์ทำให้เกิดการได้เปรียบคู่แข่งทางการค้าในการหาลักษณะการซื้อของลูกค้าความสัมพันธ์ของการซื้อกับช่วงเวลาความสัมพันธ์ระหว่างตัวสินค้าและการวิเคราะห์ประสิทธิภาพของการโฆษณาเป็นต้นช่วยให้สามารถหาวิธีการตอบสนองความต้องการของลูกค้าได้มากที่สุดและอาจหมายถึงส่วนแบ่งทางการตลาดที่เพิ่มขึ้นนั่นเอง

4. งานด้านการวางแผนในการผลิตสินค้า (Manufacturing and production) เช่นการพยากรณ์ยอดจำนวนการผลิตสินค้าเพื่อให้ได้กำไรมากสุด

5. งานด้านนายหน้าและความปลอดภัยด้านการค้า (Brokerage and securities trading) เช่นการพัฒนาวิธีการเพื่อสร้างความเชื่อมั่นในเรื่องความปลอดภัยของข้อมูลในขณะที่มีการพัฒนาวิธีการเข้าถึงข้อมูลและการ Mining ให้สะดวกต่อการใช้งานมากขึ้น

6. งานด้านชีวการแพทย์และวิเคราะห์ DNA (Biomedical an DNA Analysis) เช่น การวิเคราะห์รูปแบบการจัดเรียงตัวของหน่วยพันธุกรรมเพื่อหาสาเหตุความผิดปกติที่ทำให้เกิดโรครวมไปถึงด้านการวินิจฉัยโรคการป้องกันและการรักษานอกจากที่กล่าวมายังนำไปประยุกต์ใช้กับธุรกิจทางด้านประกันภัย (Insurance),Computer hardware และ software, หน่วยงานรัฐบาลและกระทรวงกลาโหม (Government and defense), สายการบิน (Airlines), งานด้านสุขภาพ (Health care), งานด้านการข่าว (Broadcasting) และงานด้านกฎหมาย (Law enforcement) ได้อีกด้วย

2.2 ทฤษฎีการหากฎความสัมพันธ์ (Association Rules Discovery)

การหากฎความสัมพันธ์ (Karthikeyan, T., & Ravikumar, N.A., 2014) เป็นกระบวนการหนึ่งในการทำเหมืองข้อมูลที่ได้รับความนิยมมาก โดยจะใช้กฎความสัมพันธ์ในการหาความสัมพันธ์ของข้อมูลสองชุดหรือมากกว่าสองชุดขึ้นไปภายในกลุ่มข้อมูลที่มีขนาดใหญ่ ในการหากฎความสัมพันธ์นั้นจะมีขั้นตอนวิธีการหาหลายวิธีด้วยกัน ตัวอย่างหนึ่งของกฎความสัมพันธ์ที่ใช้กันก็คือ การวิเคราะห์ตะกร้าสินค้า (Market Basket Analysis) ที่ใช้ในการหาความสัมพันธ์ของสินค้าลูกค้ามักจะซื้อพร้อมกันเพื่อใช้ในการจัดรายการส่งเสริมการขายโดยดูจากกฎความสัมพันธ์ร้อยละของค่าความเชื่อมั่นและค่าสนับสนุนที่เกิดขึ้นในการศึกษาตัวแบบนั้นผู้วิจัยได้เลือกเทคนิคการค้นหากฎความสัมพันธ์และศึกษาทฤษฎีการค้นหากฎความสัมพันธ์ ดังนี้

2.2.1. ศึกษาการค้นหากฎความสัมพันธ์ โดยมีรูปแบบของการค้นหากฎความสัมพันธ์สามารถเขียนแสดงดังสมการที่ 2-1

$$A \Rightarrow B \tag{2-1}$$

โดยที่ A เป็นเงื่อนไข และ B เป็นผลลัพธ์ที่เกิดขึ้น การหากฎความสัมพันธ์ทั้งหมดจะต้องมีค่าสนับสนุนมากกว่าค่าสนับสนุนต่ำสุดที่กำหนดไว้และมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นต่ำสุดที่กำหนดไว้ในงานวิจัยนี้ผู้วิจัยต้องการค้นหาความสัมพันธ์

การประเมินค่าของกฎจะใช้ค่าสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) โดยที่ค่าสนับสนุนคือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดแสดงดังสมการที่ 2-2

$$\text{ค่าสนับสนุน} = \frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลทั้งหมด}} \tag{2-2}$$

ค่าความเชื่อมั่นคือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนรายการข้อมูลที่เป็นเงื่อนไข สามารถเขียนเป็นสมการแสดงดังสมการที่ 2-3

$$\text{ค่าความเชื่อมั่น} = \frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลที่เป็นเงื่อนไข}} \qquad (2\text{-}3)$$

ในการหากฎความสัมพันธ์กฎจะถูกยอมรับก็ต่อเมื่อกฎนั้นค่า Support ของ (X U Y) มากกว่าหรือเท่ากับค่าสนับสนุนที่กำหนดให้เรียกว่า Minimum Support และค่า Confidence ($X \rightarrow Y$) มากกว่าหรือเท่ากับค่าความเชื่อมั่นที่กำหนดให้เรียกว่า Minimum Confidence ถึงจะถือว่ากฎนั้นๆ เป็นกฎที่น่าสนใจ

การสร้าง Association Rules จะพิจารณาสร้างกฎจาก Frequent Itemsets ทั้งหมดที่ได้มา โดยจะนำแต่ละ Frequent Itemsets มาแยกออกเป็นกฎ เช่น ให้ k = Chicken, f = French Fries, p =Pepsi Itemset {k, f, p} สามารถแยกกฎได้เป็นดังนี้

$$\{ k, f \} \rightarrow \{ p \}$$
$$\{ k, p \} \rightarrow \{ f \}$$
$$\{ f, p \} \rightarrow \{ k \}$$
$$\{ k \} \rightarrow \{ f, p \}$$
$$\{ f \} \rightarrow \{ k, p \}$$
$$\{ p \} \rightarrow \{ k, f \}$$

จากกฎข้างต้น จะเห็นว่า Items ทางด้านซ้าย (Antecedent) เมื่อนำมารวมกับ Items ทางด้านขวา (Consequent) แล้วจะต้องได้เท่ากับขนาดของ Frequent Itemsets ที่พิจารณาโดยขนาดของทั้ง Antecedent และ Consequent สามารถเพิ่มหรือลดได้ แต่เมื่อจำนวนข้อมูลของตัวใดตัวหนึ่งเพิ่มข้อมูลอีกตัวที่เหลือจะต้องลดลงและที่สำคัญจำนวน Items ของ Antecedent และ Consequent จะต้องไม่เท่า 0 แล้วนำมาหากฎที่แยกได้มาหาค่า Confidence ของกฎแล้วนำมาเปรียบเทียบกับค่า Minimum Confidence เพื่อหากฎที่สามารถยอมรับได้

2.3 ขั้นตอนวิธีอพริโอริ (Apriori Algorithm)

ขั้นตอนวิธีอพริโอริ (R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, 1996) เป็นขั้นตอนวิธีพื้นฐานในการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยและสร้างกฎความสัมพันธ์ เป็นขั้นตอนวิธีที่ได้รับการยอมรับและได้รับความนิยมเป็นอย่างมาก อีกทั้งขั้นตอนวิธีอพริโอริยังเป็นขั้นตอนวิธีที่มีอิทธิพลต่อการศึกษา และพัฒนาขั้นตอนวิธีอื่นๆ

2.3.1 หลักการทำงานของขั้นตอนวิธีอพริโอริ

หลักการทำงานของขั้นตอนวิธีอพริโอริ โดยหลักๆ แล้วจะประกอบด้วยขั้นตอนการทำงานทั้งหมด 2 ขั้นตอนด้วยกันขั้นตอนแรกคือ การสร้างกลุ่มข้อมูลท้าชิง และขั้นตอนที่สองคือการทดสอบกลุ่มข้อมูลท้าชิงเหล่านั้นว่าเป็นกลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยหรือไม่แสดงดังภาพประกอบที่ 2.4 โดยการทำงานของขั้นตอนวิธีอพริโอริ สามารถอธิบายได้ดังนี้

1) อ่านชิ้นข้อมูลจากฐานข้อมูลครั้งแรกเพื่อนับค่าความถี่ของแต่ละชิ้นข้อมูลที่ปรากฏทั้งหมดในฐานข้อมูล

2) ตรวจสอบค่าความถี่ของแต่ละชิ้นข้อมูล เพื่อคำนวณค่าสนับสนุนโดยหากชิ้นข้อมูลนั้นๆ มีค่าสนับสนุนมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำก็จะถือว่าเป็นกลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยที่มีขนาดของชิ้นข้อมูล 1 ชิ้นข้อมูล ($L1$: Frequent 1-Itemsets)

3) นำ $L1$ ที่ได้มาสร้างกลุ่มข้อมูลท้าชิงที่มีขนาดของชิ้นข้อมูล 2 ชิ้นข้อมูล ($C2$: Candidate 2-Itemsets)

4) อ่านชิ้นข้อมูลจากฐานข้อมูลอีกครั้งเพื่อนับค่าความถี่ของ $C2$ และตัด $C2$ ที่มีค่าสนับสนุนน้อยกว่า ค่าสนับสนุนขั้นต่ำ โดยหาก $C2$ มีค่าสนับสนุนมากกว่า หรือเท่ากับค่าสนับสนุนขั้นต่ำก็จะกลายเป็น $L2$

5) ทำในหัวข้อที่ 3) และ 4) ซ้ำจนกว่าไม่สามารถสร้าง Ck จาก Lk-1 ได้ เมื่อ k คือขนาดของชิ้นข้อมูล จึงสิ้นสุดการสร้างกลุ่มข้อมูลท้าชิงและจบการทำงานของขั้นตอนวิธีอพริโอริ ทำให้ได้กลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยทั้งหมด

```
1   L1 = {large 1-itemsets};
2   For (k=2; Lk-1≠∅; k++) do begin
3     Ck = apriori-gen(Lk-1);    // New candidates
4     Forall transactions t ∈ D do begin
5       Ct = subset(Ck , t);    // Cadidates contained in t
6       Forall candidates c ∈ Ct do
7         c.count++;
8       End
9     L1 = { c ∈ Ck | c.count ≥ minsup};
10  End
11  Answer = Uk Lk;
```

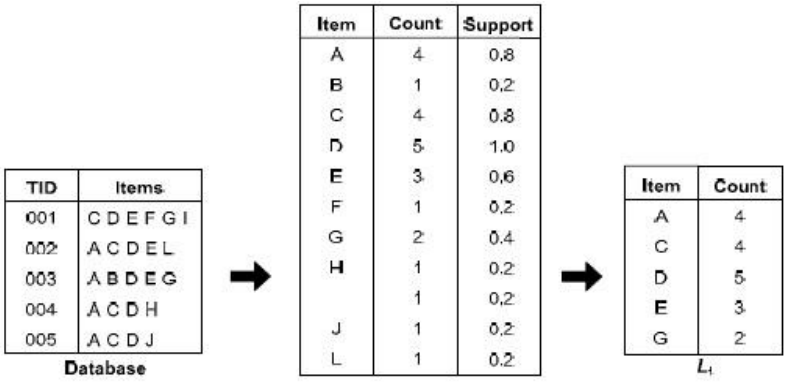
ภาพประกอบที่ 2.3 ขั้นตอนการทำงานของขั้นตอนวิธีอพริโอริ

(R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, 1996)

2.3.2 ตัวอย่างการทำงานของขั้นตอนวิธีอพริโอริ

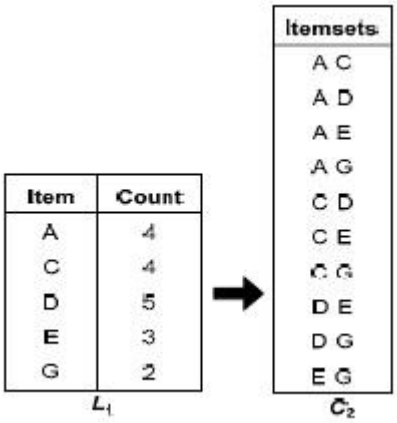
เพื่อให้เข้าใจถึงกระบวนการทำงานของขั้นตอนวิธีอพริโอริ ได้อย่างชัดเจนขึ้น จึงขอยกตัวอย่าง การค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของขั้นตอนวิธีนี้โดยใช้ฐานข้อมูลการซื้อสินค้าในตารางที่ 2.1 ประกอบการอธิบาย พร้อมทั้งกำหนดค่าสนับสนุนขั้นต่ำเป็น 0.4 (มีค่าความถี่อย่างน้อยเท่ากับ 2 รายการข้อมูลปรากฏในฐานข้อมูล) โดยตัวอย่างการทำงานของขั้นตอนวิธีอพริโอริ สามารถอธิบายได้ดังนี้

1) อ่านชิ้นข้อมูลจากฐานข้อมูลเพื่อนับค่าความถี่ของแต่ละชิ้นข้อมูล และคำนวณค่าสนับสนุนจากค่าความถี่ที่ได้ พร้อมทั้งตัดชิ้นข้อมูลที่ไม่ผ่านค่าสนับสนุนขั้นต่ำที่กำหนดไว้ ซึ่งผลลัพธ์ที่ได้จะเรียกว่า $L1$ แสดงดังภาพประกอบที่ 2.4



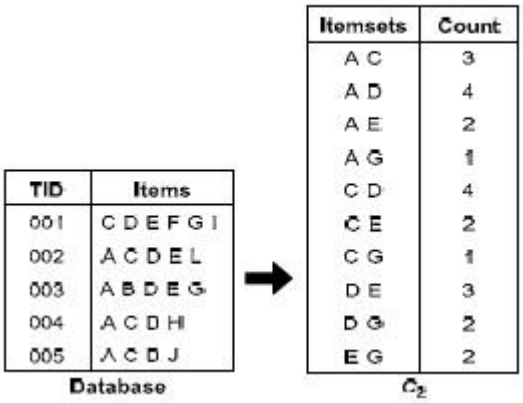
ภาพประกอบที่ 2.4 การอ่านข้อมูลจากฐานข้อมูลเพื่อค้นหา $L1$ ของขั้นตอนวิธีอพริโอริ

2) นำ $L1$ ที่ได้ไปสร้างกลุ่มข้อมูลท้าชิง $C2$ โดยวิธีการสร้าง$C2$ นั้นจะจับคู่ร่วมกัน (Join) ของชิ้นข้อมูลทุกตัวที่เป็น $L1$ ทุกกรณีที่เป็นไปได้ที่ประกอบด้วยชิ้นข้อมูล 2 ชิ้นข้อมูล ดังภาพประกอบที่ 2.5



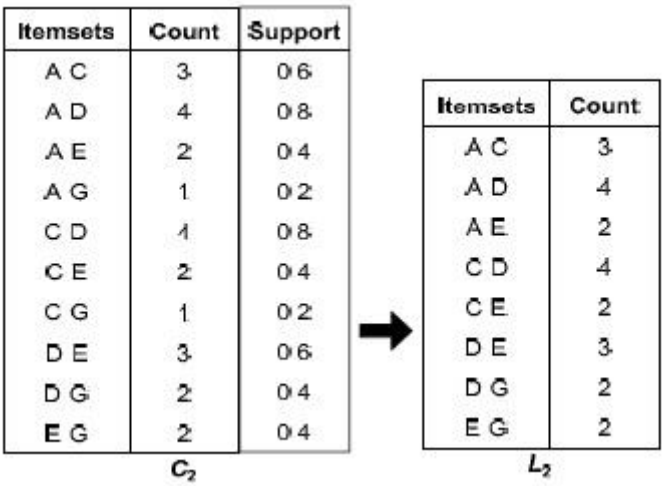
ภาพประกอบที่ 2.5 การสร้างกลุ่มข้อมูลท้าชิง $C2$ ของขั้นตอนวิธีอพริโอริ

3) หลังจากสร้างกลุ่มข้อมูลท้าชิง $C2$ แล้วนั้น ขั้นตอนต่อไปคืออ่านชิ้นข้อมูลจากฐานข้อมูลเพื่อนับค่าความถี่ของ $C2$ แสดงดังภาพประกอบที่ 2.6



ภาพประกอบที่ 2.6 การนับค่าความถี่ของข้อมูลท้าชิง $C2$ ของขั้นตอนวิธีอพริโอริ

4) นำค่าความถี่ของ *C*2 ที่ได้ไปคำนวณหาค่าสนับสนุนเพื่อใช้ในการตัดสินใจให้เป็นกลุ่มข้อมูล *L*2 ผลลัพธ์ที่ได้แสดงดังภาพประกอบที่ 2.7



Itemsets	Count	Support
A C	3	0 6
A D	4	0 8
A E	2	0 4
A G	1	0 2
C D	4	0 8
C E	2	0 4
C G	1	0 2
D E	3	0 6
D G	2	0 4
E G	2	0 4

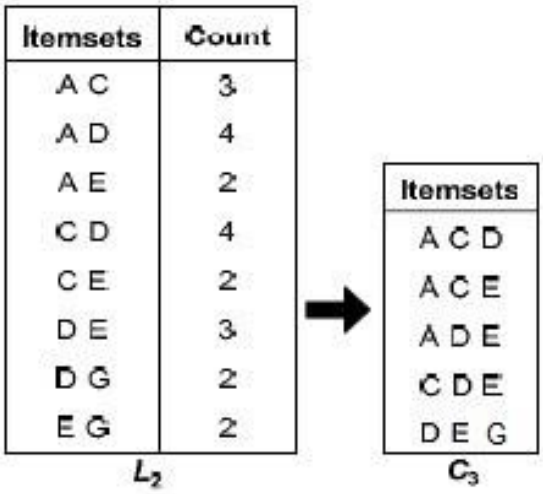
C_2

Itemsets	Count
A C	3
A D	4
A E	2
C D	4
C E	2
D E	3
D G	2
E G	2

L_2

ภาพประกอบที่ 2.7 ผลลัพธ์การค้นหากลุ่มข้อมูล *L*2 ของขั้นตอนวิธีอพริโอริ

5) นำกลุ่มข้อมูล *L*2 ที่ได้ไปสร้างกลุ่มข้อมูลท้าชิง *C*3 โดยวิธีการสร้างจะจับคู่ร่วมกันของกลุ่มข้อมูล ทุกตัวที่เป็น *L*2 ที่มีชิ้นข้อมูลตัวแรกเหมือนกัน ซึ่งผลลัพธ์ของการจับคู่จะประกอบด้วยชิ้นข้อมูล 3 ชิ้นข้อมูล ดังภาพประกอบที่ 2.8



Itemsets	Count
A C	3
A D	4
A E	2
C D	4
C E	2
D E	3
D G	2
E G	2

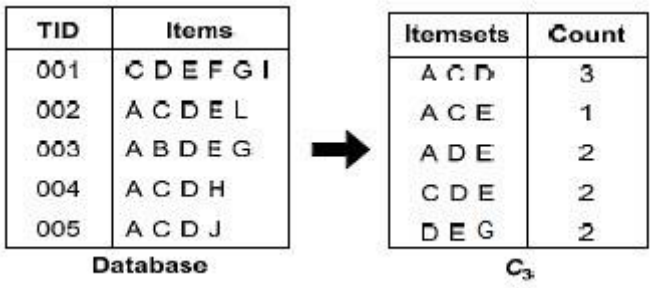
L_2

Itemsets
A C D
A C E
A D E
C D E
D E G

C_3

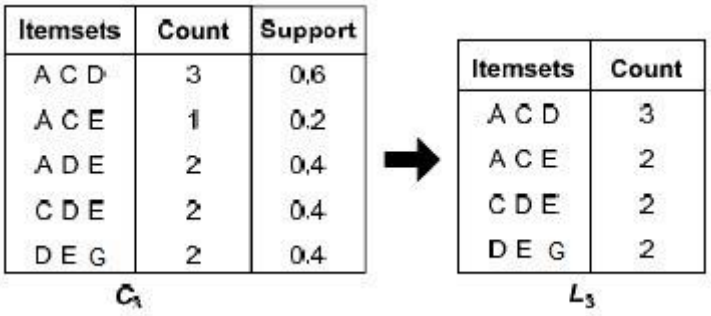
ภาพประกอบที่ 2.8 การสร้างกลุ่มข้อมูลท้าชิง *C*3 ของขั้นตอนวิธีอพริโอริ

6) หลังจากสร้างกลุ่มข้อมูลท้าชิง *C*3 แล้วนั้น ขั้นตอนต่อไปคือ อ่านชิ้นข้อมูลจากฐานข้อมูลเพื่อนับค่าความถี่ของ *C*3 แสดงดังภาพประกอบที่ 2.9



ภาพประกอบที่ 2.9 การนับค่าความถี่ของกลุ่มข้อมูลท้าชิง *C*3 ของขั้นตอนวิธีอพริโอริ

7) คำนวณหาค่าสนับสนุนจากค่าความถี่ที่ได้ พร้อมทั้งตัดกลุ่มข้อมูลท้าชิง*C*3 ที่ไม่ผ่านค่าสนับสนุนขั้นต่ำ ซึ่งผลลัพธ์ที่ได้จะเรียกว่า *L*3 แสดงดังภาพประกอบที่ 2.10



ภาพประกอบที่ 2.10 ผลลัพธ์การค้นหากลุ่มข้อมูล *L2* ของขั้นตอนวิธี อพริโอริ

8) นำ *L*3 ที่ได้ไปสร้าง *C*4 โดยวิธีการสร้าง *C*4 นั้นจะจับคู่ร่วมกันของกลุ่มข้อมูลทุกตัวที่เป็น *L*3 กับตัวมันเองที่มีชิ้นข้อมูลสองตัวแรกเหมือนกัน ซึ่งผลลัพธ์ของการจับคู่จะประกอบด้วยชิ้นข้อมูล 4 ชิ้นข้อมูล แต่จากภาพประกอบที่ 2.10 จะเห็นได้ว่าเมื่อได้ *L*3 แล้วไม่สามารถสร้าง *C*4 จาก *L*3 ได้ เนื่องจากไม่มีกลุ่มข้อมูลใดเลยใน *L*3 ที่มีชิ้นข้อมูล 3 ชิ้นข้อมูลแรกเหมือนกัน จึงหยุดการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อย ดังนั้นผลลัพธ์ที่ได้ทั้งหมด แสดงดังตาราง 2.1โดยตัวเลขที่อยู่หลังเครื่องหมาย “:” ในตารางหมายถึงค่าความถี่ของกลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยๆ

ตารางที่ 2.1 ผลลัพธ์การค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของขั้นตอนวิธีอพริโอริ

Level	Frequent Itemsets
L_1	A:4, C:4, D:5, E:3, G:2
L_2	AC:3, AD:4, AE:2, CD:4, CE:2, DE:3, DG:2, EG:2
L_3	ACD:3, ADE:2, CDE:2, DEG:2

2.3.3 ข้อดีของขั้นตอนวิธี อพริโอริ

ขั้นตอนวิธี Apriori สามารถทำงานได้ดีหากกำหนดค่าสนับสนุนขั้นต่ำมีค่ามากๆมีขนาดของฐานข้อมูลเล็ก และมีจำนวนของชิ้นข้อมูล $L1$ น้อย อีกทั้งกระบวนการทำงานสำหรับการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของขั้นตอนวิธี อพริโอริ นั้นมีลักษณะที่ง่ายและไม่ซับซ้อน

2.3.4 ข้อเสียของขั้นตอนวิธี อพริโอริ

การทำงานของขั้นตอนวิธี อพริโอริ นั้นต้องอ่านข้อมูลจากฐานข้อมูลหลายครั้งเพื่อใช้ตรวจสอบกลุ่มข้อมูลท้าชิงซึ่งอาจก่อให้เกิดปัญหาคอขวดขึ้นได้ และในระหว่างการประมวลผลต้องใช้เนื้อที่ในหน่วยความจำเป็นจำนวนมากสำหรับการสร้างกลุ่มข้อมูลท้าชิง และยังใช้เวลาในการประมวลผลนานหากข้อมูลในฐานข้อมูลมีอัตราส่วนของจำนวนชิ้นข้อมูลปรากฏในรายการข้อมูลมาก และขนาดของฐานข้อมูลมีขนาดใหญ่

2.4 ขั้นตอนวิธี เอฟพี-กโรธ (FP-Growth Algorithm)

เนื่องจากขั้นตอนวิธี อพริโอริ มีการอ่านข้อมูลหลายครั้ง และยังต้องสร้างกลุ่มข้อมูลท้าชิงจำนวนมาก จึงทำให้มีการคิดค้นวิธีการเพื่อแก้ไขข้อบกพร่องนี้ Han และคณะ (Han, J.,Pei, J., and Yin, Y., 2000) ได้พัฒนาขั้นตอนวิธีใหม่ขึ้นมาเพื่อลดจำนวนของการอ่านข้อมูลจากฐานข้อมูล พร้อมทั้งนำเสนอโครงสร้างข้อมูลแบบใหม่ขึ้นมา ที่มีชื่อว่า FP-Tree โดยใช้ชื่อว่าขั้นตอนวิธีเอฟพี-กโรธ เป็นขั้นตอนวิธีที่อ่านข้อมูลจากฐานข้อมูลเพียง 2 ครั้ง และไม่มีการสร้างกลุ่มข้อมูลท้าชิงเพื่อลดระยะเวลาในการประมวลผลให้สามารถทำงานได้เร็วขึ้น

2.4.1 หลักการทำงานของขั้นตอนวิธีเอฟพี-กโรธ

หลักการทำงานของขั้นตอนวิธีเอฟพี-กโรธ แสดงดังภาพประกอบที่ 2.11 ซึ่งขั้นตอนวิธีเอฟพี-กโรธ เป็นขั้นตอนวิธีที่มีลักษณะการค้นหากลุ่มข้อมูลที่ปรากฏบ่อยแบบการเติบโตอย่างเป็นรูปแบบ (Pattern Growth) โดยการทำงานของขั้นตอนวิธีเอฟพี-กโรธ สามารถ อธิบายหลักการทำงานได้ดังนี้

1) อ่านข้อมูลจากฐานข้อมูลครั้งแรกเพื่อนับค่าความถี่ของแต่ละชิ้นข้อมูล แล้วนำชิ้นข้อมูลที่ไม่น้อยกว่าค่าสนับสนุนขั้นต่ำ ($L1$) มาเรียงลำดับตามค่าความถี่ของแต่ละชิ้นข้อมูลจากมากไปหาน้อยแล้วนำมาสร้างตาราง Header (Header Table)

2) อ่านข้อมูลจากฐานข้อมูลครั้งที่สองเพื่อสร้างต้นไม้ FP-Treeโดยอ่านข้อมูลจากฐานข้อมูลทีละรายการข้อมูลจากนั้นตัดชิ้นข้อมูลในรายการข้อมูลนั้นที่ไม่ปรากฏอยู่ในตาราง Header ทิ้งไป แล้วเรียงชิ้นข้อมูลที่เหลือตามลำดับในตาราง Headerแล้วนำชิ้นข้อมูลดังกล่าวไปสร้างโหนด (Node Tree) เพิ่มเข้าไปในต้นไม้ FP-Tree แล้วเชื่อมแต่ละโหนดที่เป็นชิ้นข้อมูลเดียวกันเพิ่มเข้าไปกับตาราง Header

3) สร้าง Conditional pattern base และสร้าง Conditional FP-Tree ของแต่ละชิ้นข้อมูลเพื่อใช้ในขั้นตอนการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อย โดยการพิจารณาจะเริ่มจากชิ้นข้อมูลล่างสุดจนถึงชิ้นข้อมูล ที่อยู่บนสุดในตาราง Header ซึ่ง Conditional pattern base หมายถึงเซตของชิ้นข้อมูลที่เกิดขึ้นพร้อมกับชิ้นข้อมูลที่กำลังพิจารณาในแต่ละเส้นทาง (Path Tree) และกำหนดให้ทุกชิ้นข้อมูลมีค่าความถี่เท่ากับค่าความถี่ ของชิ้นข้อมูลที่กำลังพิจารณาจากต้นไม้ FP-Tree หลังจากนั้นสร้างต้นไม้ FP-Tree บน Conditional pattern base เรียกว่า Conditional FP-Tree ซึ่งเกิดจากการนำค่าความถี่ของแต่ละชิ้นข้อมูลในทุกเส้นทางมารวมกันและ เลือกเฉพาะชิ้นข้อมูลที่ผ่านค่าสนับสนุนขั้นต่ำจาก Conditional FP-Tree เพื่อนำไปสร้างกลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยต่อไป

4) ค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยจากการสร้าง Conditional pattern base และสร้าง Conditional FP-Tree ของแต่ละชิ้นข้อมูล โดยใช้หลักการทำงานแบบแบ่งแยกแล้วเอาชนะ (Divide and Conquer)

Input: FP-tree constructed based on Algorithm 1, using *DB* and a minimum support threshold ξ.

Output: The complete set of frequent patterns.

Method: Call FP-tree (FP-tree, null).

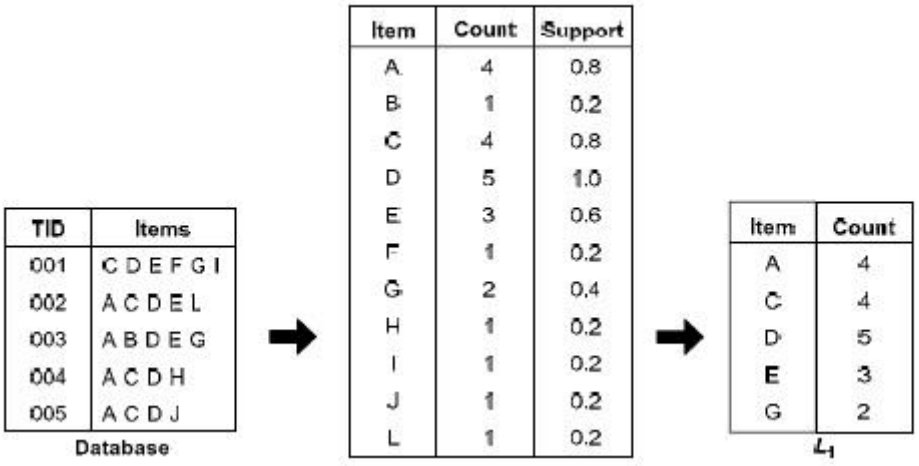
Procedure FPF-tree (FP-tree, α)

```
{
1   IF Tree contains a single path P
2   Then for each combination (denoted as β) of the nodes in the path P do
3       generate pattern β U α with support = minimum support of node in β;
4   Else for each a_i in the header of Tree do    {
5       generate pattern β = a_i U α with support =  a_i.support;
6       construct β's conditional pattern base and then β 's conditional FP-tree
            Tree_β;
7       If Tree_β ≠ ∅
8       Then call FP-Tree (Tree_β, β)              }
}
```

ภาพประกอบที่ 2.11 หลักการทำงานของขั้นตอนวิธีเอฟพี-กโรธ (Han,J.,Pei, J., and Yin,Y., 2000)

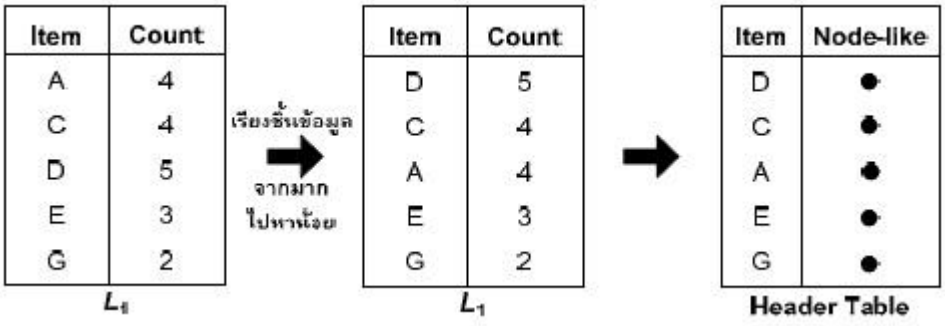
2.4.2 ตัวอย่างการทำงานของขั้นตอนวิธีเอฟพี-กโรธ

1) อ่านข้อมูลจากฐานข้อมูลครั้งแรกเพื่อนับค่าความถี่ของแต่ละชิ้นข้อมูล และคำนวณค่าความถี่ เพื่อกำจัดชิ้นข้อมูลที่ปรากฏไม่บ่อย ซึ่งผลลัพธ์ที่ได้คือ $L1$ ดังภาพประกอบที่ 2.12



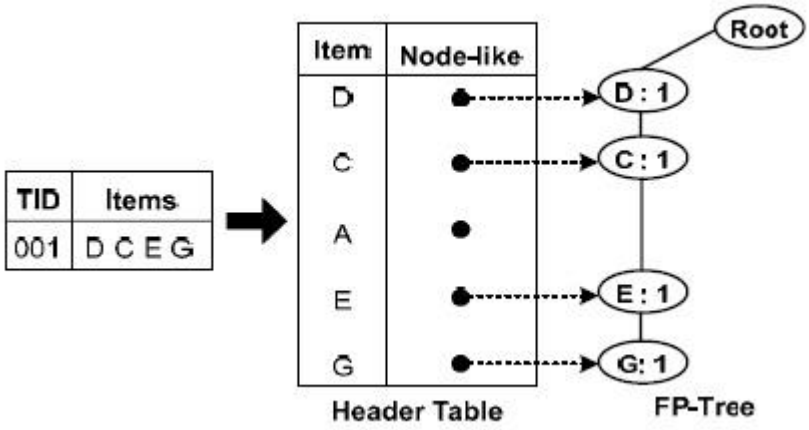
ภาพประกอบที่ 2.12 การอ่านข้อมูลจากฐานข้อมูลเพื่อค้นหา $L1$ ของขั้นตอนวิธีเอฟพี-กโรธ

2) จัดเรียง $L1$ ใหม่ตามค่าความถี่ของแต่ละชิ้นข้อมูลจากมาก

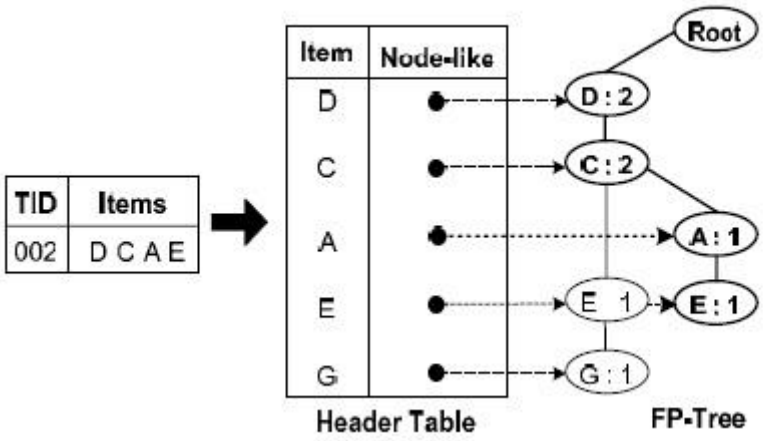


ภาพประกอบที่ 2.13 การสร้างตาราง Header ของขั้นตอนวิธีเอฟพี-กโรธ

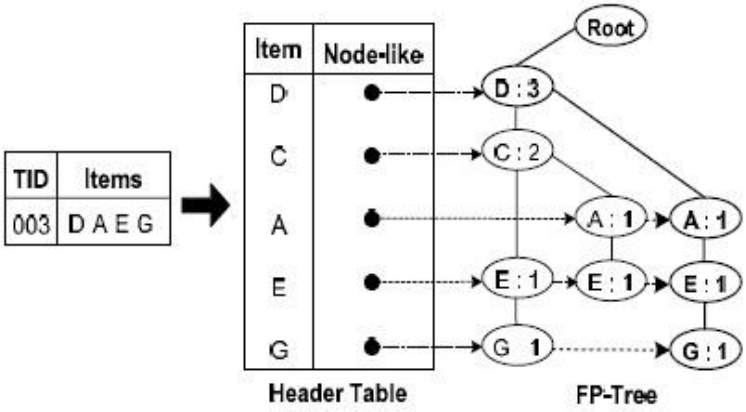
3) อ่านข้อมูลจากฐานข้อมูลครั้งที่สอง โดยเริ่มอ่านชิ้นข้อมูลทั้งหมดจากรายการข้อมูลแรกในฐานข้อมูลซึ่งประกอบด้วยชิ้นข้อมูล C D E F G และ I จากนั้นตัดชิ้นข้อมูลที่ไม่ปรากฏในตาราง Header ออกไป แล้วเรียงลำดับชิ้นข้อมูลที่เหลือใหม่ตามลำดับในตาราง Header จะได้ลำดับของชิ้นข้อมูลของรายการข้อมูลแรกคือ D C E และ G จากนั้นนำชิ้นข้อมูลที่ได้ไปสร้างโหนดของชิ้นข้อมูลเพิ่มเข้าไปในต้นไม้ FP-Tree แล้วเชื่อมแต่ละโหนดที่เพิ่มเข้าไปกับตาราง Header ผลลัพธ์แสดงดังภาพประกอบที่ 2.14 โดยตัวเลขที่อยู่ หลังเครื่องหมาย ":" ในแต่ละโหนดหมายถึงค่าความถี่ของกลุ่มข้อมูล และทำตามขั้นตอนข้างต้นนี้ กับทุกรายการข้อมูลในฐานข้อมูล จะสามารถแสดงผลลัพธ์ของการเพิ่มโหนดชิ้นข้อมูลของแต่ละรายการข้อมูลเข้าไปในต้นไม้ FP-Tree ดังภาพประกอบที่ 2.14 ถึง 2.18



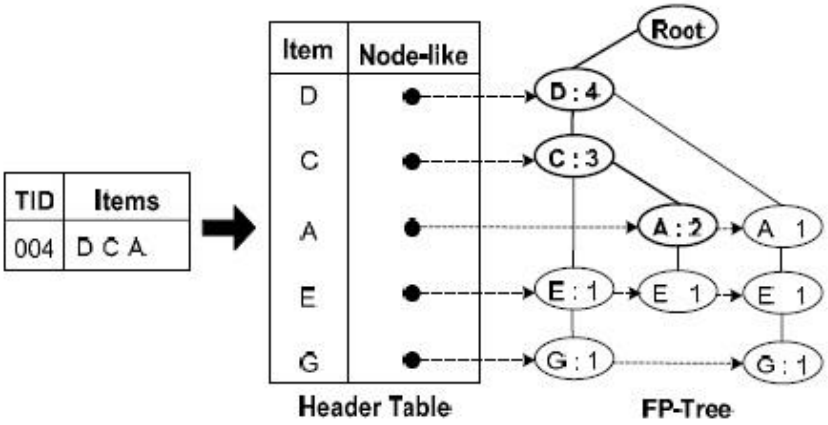
ภาพประกอบที่ 2.14 การอ่านรายการข้อมูลแรกจากฐานข้อมูลของขั้นตอนวิธีเอฟพี-กโรธ



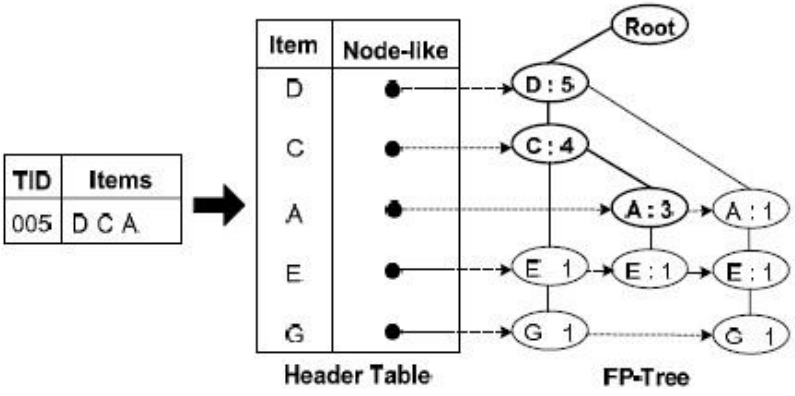
ภาพประกอบที่ 2.15 การอ่านรายการข้อมูลที่สองจากฐานข้อมูลของขั้นตอนวิธีเอฟพี-กโรธ



ภาพประกอบที่ 2.16 การอ่านรายการข้อมูลที่สามจากฐานข้อมูลของขั้นตอนวิธีเอฟพี-กโรธ

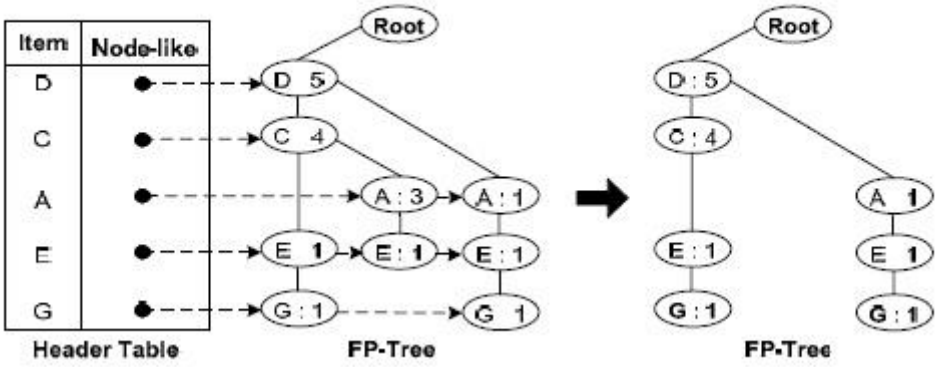


ภาพประกอบที่ 2.17 การอ่านรายการข้อมูลที่สี่จากฐานข้อมูลของขั้นตอนวิธีเอฟพี-กโรธ



ภาพประกอบที่ 2.18 การอ่านรายการข้อมูลที่ห้าจากฐานข้อมูลของขั้นตอนวิธีเอฟพี-กโรธ

4) สร้าง Conditional pattern base และสร้าง Conditional FP-Tree โดยเริ่มพิจารณาจากชิ้นข้อมูลสุดท้ายในตาราง Header Table นั่นคือชิ้นข้อมูล G ซึ่งจากภาพประกอบที่ 2.19 จะเห็นได้ว่าต้นไม้ FP-Tree มีโหนด G ปรากฏอยู่ 2 เส้นทางได้แก่ เส้นทาง D C E มีค่าความถี่เป็น1 (เกิดร่วมกับ ชิ้นข้อมูล G จำนวน 1 ครั้ง) และเส้นทาง D A E มีค่าความถี่เป็น 1 ดังนั้นจะได้ Conditional pattern base ของชิ้นข้อมูล G คือ {(DCE:1), (DAE:1)} ซึ่งจากทั้ง 2 เส้นทางจะเห็นว่ามีชิ้นข้อมูล D และ E ปรากฏร่วมกันกับชิ้นข้อมูล G เหมือนกันทั้ง 2 เส้นทางดังนั้นจะสร้าง Conditional FP-Tree ได้เป็น {(DE:2)} | G และทำเช่นนี้กับทุกชิ้นข้อมูลในตาราง Header จะสามารถหา Conditional pattern base และสร้าง Conditional FP-Tree ของทุกชิ้นข้อมูลได้ดังแสดงในตารางที่ 2.2



ภาพประกอบที่ 2.19 การสร้างตาราง Header ของขั้นตอนวิธีเอฟพี-กโรธ

ตารางที่ 2.2 ค่า Conditional pattern base และค่า Conditional FP-Tree ของขั้นตอนวิธีเอฟพี-กโรธ

Items	Conditional Pattern base	Conditional FP-Tree
G	{(DCE:1), (DAE:1)}	{(DE:2)} | G
E	{(DC:1), (DCA:1), (DA:1)}	{(D:3), (DC:2), (DA:2)} | E
A	{(DC:3), (D:1)}	{(D:4), (DC:3)} | A
C	{(D:4)}	{(D:4)} | C
D	∅	∅

5) ค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยจากการสร้าง Conditional pattern base และสร้าง Conditional FP-Tree โดยเริ่มพิจารณาจากชิ้นข้อมูล G ซึ่งจากตารางที่ 2.2 จะเห็นได้ว่าชิ้นข้อมูล G มี Conditional pattern base เป็น {(DCE:1), (DAE:1)} และ Conditional FP-Tree เป็น {(DE:2)} | G ทำให้การค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของชิ้นข้อมูล G แบ่งการทำงานของ {(DE:2)} | G เป็น 2 ส่วนคือ EG:2 และ DG:2โดยส่วนแรก EG:2 จะแยกได้อีกเป็น {(D:2)} | EG ซึ่งสุดท้ายจะได้เป็น DEG:2 และส่วนที่สอง DG:2 ไม่สามารถแบ่งได้อีกดังนั้นจะได้กลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของชิ้นข้อมูล G เป็น G:2 EG:2 DG:2 และ DEG:2 ผลลัพธ์ของการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยทั้งหมด แสดงดังตารางที่ 2.3

ตารางที่ 2.3 ผลลัพธ์ของการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยของขั้นตอนวิธีเอฟพี-กโรธ

Items	Frequent Itemsets
G	G:2, EG:2, DG:2, DEG:2
E	E:3, DE:3, CE:2, AE:2, DCE:2, DAE:2
A	A:4, DA:4, CA:3, DCA:3
C	C:4, DC:4
D	D:5

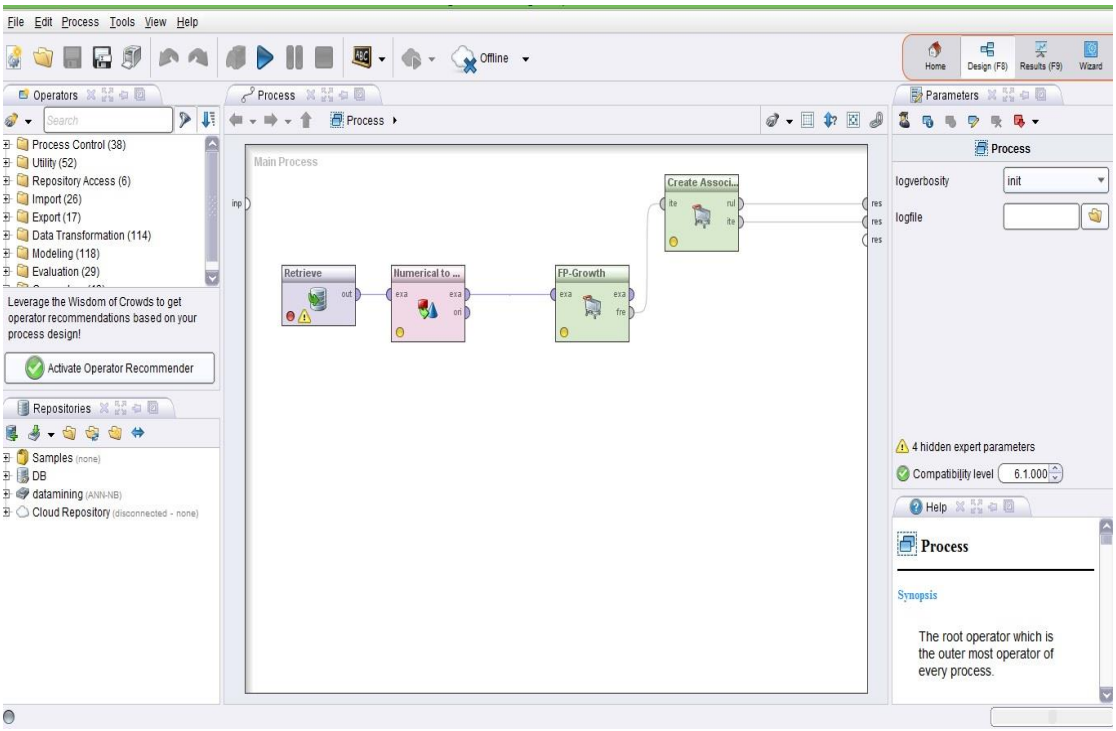
2.4.3 สรุปการทำงานขั้นตอนวิธีเอฟพี-กโรธ

ดังนั้นการทำงานของขั้นตอนวิธีเอฟพี-กโรธ ทำให้ช่วยลดจำนวนการอ่านข้อมูลจากฐานข้อมูลสำหรับการค้นหาข้อมูลที่ปรากฏร่วมกันบ่อยเหลือเพียง 2 ครั้ง และกระบวนการทำงานใช้หลักการทำงานแบบพลวัตร (Dynamic Programming) ทำให้การทำงานมีประสิทธิภาพเหมาะสมกับฐานข้อมูลที่มีขนาดเล็กและขนาดใหญ่ มีจำนวนชิ้นข้อมูล ในฐานข้อมูลน้อย และลักษณะข้อมูลที่เหมาะสมต้องมีความหนาแน่นของข้อมูลสูง คืออัตราส่วนของจำนวนชิ้นข้อมูลที่ปรากฏอยู่ในรายการข้อมูลมีมาก การทำงานสามารถทำงานได้ดีหากกำหนดค่าสนับสนุนขั้นต่ำ

มีค่ามากๆ เพราะจะใช้เวลาในการท่องไปยังแต่ละโหนดสำหรับการค้นหากลุ่มข้อมูลที่ปรากฏร่วมกันบ่อยได้เร็ว และลดการใช้เนื้อที่ในการสร้างต้นไม้ FP-Tree สำหรับจัดเก็บข้อมูล

2.5 โปรแกรม Rapid Miner Studio 6

โปรแกรม Rapid Miner Studio 6 (เอกสิทธิ์ พัชรวงศ์ศักดา, 2558) เป็นเครื่องที่ใช้ในการวิเคราะห์ข้อมูลที่มีขนาดใหญ่หรือเหมืองข้อมูล และสามารถทำการวิเคราะห์ข้อมูลแบบต่างๆ ได้ เช่น การจำแนกชนิดข้อมูล การจัดกลุ่มข้อมูล ซอฟต์แวร์ Rapid Miner Studio 6 แรกเริ่มพัฒนาขึ้นจากบริษัทที่ชื่อว่า Rapid-I ในประเทศเยอรมนี และเมื่อช่วงปลายปี 2013 เปลี่ยนชื่อบริษัทจาก Rapid-I เป็น Rapid Miner สามารถรองรับการใช้งานไฟล์ได้หลายประเภท เช่น ไฟล์ Excel สามารถแสดงข้อมูลได้หลายรูปแบบ เช่น scatter plot 3D สามารถแสดงผลโมเดลที่สวยงามและแก้ไขการแสดงผลให้สามารถอ่านได้ง่ายขึ้น สามารถบันทึกไฟล์โมเดลออกเป็นไฟล์ภาพประเภทต่างๆ เช่น PNG, JPG หรือ PDF มีวิธีการเตรียมข้อมูล (preprocess) และการวิเคราะห์ได้หลากหลายรูปแบบ



ภาพประกอบที่ 2.20 โปรแกรม Rapid Miner Studio 6

2.6 งานวิจัยที่เกี่ยวข้อง

ปัณฑารีย์ ปริยภารัตน์ (2554) พัฒนาระบบสนับสนุนการตัดสินใจเพื่อการวางแผนการจัดโปรโมชั่นทางการตลาด สำหรับอาหารฟาสต์ฟู้ด โดยใช้เทคนิคกฎความสัมพันธ์ (Association Rules) ระบบที่พัฒนาขึ้นนี้อยู่ในรูปแบบเว็บแอพพลิเคชั่นโดยใช้ C#.NET เป็นเครื่องมือในการพัฒนาและใช้ Microsoft SQL Server 2005 เป็นฐานข้อมูล ในการวิเคราะห์ยอดขายสินค้าได้เป็นกฎความสัมพันธ์ทั้งหมด 7 กฎโดยให้ค่าสนับสนุนเท่ากับ 80 และค่าความเชื่อมั่นเท่ากับ 90 โดยมีเพียง 1 กฎที่มีค่าความเชื่อมั่นเท่ากับ 100 คือ { Fried Chicken } { Roasted Wing Herbs, Pepsi } ชุดโปรโมชั่นดังกล่าวสามารถนำไปประยุกต์ใช้จริงและส่งผลให้มียอดขายเพิ่มขึ้น

ศิริพันธ์ เทพมาก (2554) พัฒนาระบบหาความสัมพันธ์ของการซื้อสินค้า โดยใช้เทคนิคกฎความสัมพันธ์กรณีศึกษา สินค้าเครื่องสำอาง ซึ่งระบบจะมีการสร้างกฎความสัมพันธ์ของสินค้าขึ้นมาจากการคำนวณค่าความเชื่อมั่น และค่าสนับสนุนของกฎความสัมพันธ์นั้น จากนั้นระบบจะแสดงกฎความสัมพันธ์ของการซื้อสินค้า เพื่อให้ผู้ประกอบการสามารถนำข้อมูล มาประกอบการตัดสินใจในการจัดทำรายการส่งเสริมการขาย ซึ่งจะทำให้เข้าถึงลูกค้าได้มากขึ้น และจัดทำรายการส่งเสริมการขายได้ตรงตามความต้องการของลูกค้า

เอมอมร ปินปันคง (2554) ออกแบบและพัฒนารูปแบบพยากรณ์ความต้องการสินค้าอุปโภคและบริโภค โดยใช้เทคนิคกฎการวิเคราะห์ความสัมพันธ์ กรณีศึกษาห้างขายปลีกแห่งหนึ่ง โดยใช้ อัลกอรึทึ่ม J48 เพื่อทำการจัดกลุ่มของข้อมูลที่ลูกค้าส่วนใหญ่ที่มีความต้องการสินค้าและไม่เป็นที่ต้องการ และ อพริโอริ เพื่อหาความสัมพันธ์ของสินค้าเพื่อเป็นการเพิ่มโอกาสทางการขายให้มากขึ้น

จากการศึกษาทฤษฎี และงานวิจัยที่เกี่ยวข้องตามที่ได้กล่าวมาในข้างต้นแล้ว ทำให้ผู้วิจัยมีแนวคิดในการสร้างแบบจำลองกฎความสัมพันธ์สำหรับฐานข้อมูลการสั่งซื้อสินค้าโดยใช้เทคนิค เอฟพี-กโรธ ซึ่งจะวิเคราะห์ข้อมูลจากการซื้อสินค้าของลูกค้าในซุปเปอร์มาเก็ตขนาดใหญ่ เพื่อหาความสัมพันธ์ของการซื้อสินค้าของลูกค้า จะมีหลักการทำงานคือ ใช้โครงสร้างข้อมูลที่เรียกว่า เอฟพี-ทรี ในการเก็บข้อมูลก่อนหน้าของรายการที่มีการเปลี่ยนแปลงโดยเฉพาะรายการที่พิจารณาค่าสนับสนุนต้องมากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำเท่านั้น ขั้นตอนการทำงานของอัลกอริทึมมี 2 ขั้นตอนคือ การสร้าง เอฟพี-ทรี จากข้อมูล และการหาเซตรายการความถี่จากเอฟพี-ทรี ที่สร้างขึ้น จะทำให้ลดระยะเวลาในการประมวลผลจากฐานข้อมูลการซื้อสินค้าของลูกค้า จึงทำให้เทคนิคดังกล่าวทำงานได้รวดเร็วกว่าเทคนิค อพริโอริ ซึ่งจะต้องอ่านข้อมูลจากฐานข้อมูลหลายครั้งเพื่อใช้ตรวจสอบกลุ่มข้อมูลท้าชิงซึ่งอาจก่อให้เกิดปัญหาคอขวดขึ้นได้

และในระหว่างการประมวลผลต้องใช้เนื้อที่ในหน่วยความจำเป็นจำนวนมากสำหรับการสร้างกลุ่มข้อมูลท้าชิง และยังใช้เวลาในการประมวลผลนานหากข้อมูลในฐานข้อมูลมีอัตราส่วนของจำนวนชิ้นข้อมูลปรากฏในรายการข้อมูลมาก และขนาดของฐานข้อมูลมีขนาดใหญ่