การพัฒนาระบบดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining

SYSTEM DEVELOPMENT OF SEQUENTIAL PATTERN MINING

นันทวัน นาคศิริ

NANTAWAN NAKSIRI

อาจารย์ที่ปรึกษา

รศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2551

SYSTEM DEVELOPMENT OF SEQUENTIAL PATTERN MINING

NANTAWAN NAKSIRI

A SYSTEM DEVELOPMENT PROJECT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE PROGRAM IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATION TECNOLOGY

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2/ 2008

หัวข้อ	การพัฒนาระบบดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining
นักศึกษา	นางสาวนันทวัน นาคศิริ
รหัสนักศึกษา	47066227
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2551
อาจารย์ที่ปรึกษา	รศ. ดร. วรพจน์ กรีสุระเดช

บทคัดย่อ

การขุดค้นหาความสัมพันธ์ของข้อมูลที่ใช้วิเคราะห์ข้อมูลการขายสินค้านั้น ยังมีเทคนิควิธีที่น่าสนใจ คือ การหารูปแบบของลำดับข้อมูลที่เกิดขึ้นต่อเนื่องกัน (Sequential Pattern Mining) เป็นการวิเคราะห์เพื่อหาแนวโน้ม หรือพฤติกรรมการซื้อสินค้าแบบต่อเนื่อง โดยใช้อัลกอริทึม CloSpan (Closed Sequential pattern mining) ซึ่งเป็นเทคนิควิธีหารูปแบบของความสัมพันธ์ที่ใหญ่ที่สุดก่อน ซึ่งทำให้ใช้เวลาในการประมวลผลน้อยลง โดยระบบที่พัฒนาขึ้นนี้จะใช้โปรแกรมของ University of Illinois at Urbana-Champaign เป็นเครื่องมือในการขุดค้นข้อมูล ซึ่งระบบที่พัฒนาขึ้นนี้จะทำการจัดเตรียมข้อมูล และนำข้อมูลที่ได้มานั้นสร้างกลยุทธ์เพื่อวางแผนการขายสินค้าต่อไป

Title	System Development of Sequential Pattern Mining
Student	Miss. Nantawan Naksiri
Student ID.	47066227
Degree	Master of Science
Program	Information Science
Academic Year	2008
Advisor	Assoc.Prof. Dr.Worapoj Kreesuradej

ABSTRACT

Sequential Pattern Mining is an active research theme that uses to analyze behavior of customer shopping sequences. CloSpan (Closed Sequential pattern mining) is algorithm to mine frequent closed sequences efficiently in large database. This project present step and prepared database for data mining. Thence the system used CloSpan that developed by University of Illinois at Urbana-Champaign and finally present knowledge data for business.

กิตติกรรมประกาศ

โครงการพัฒนาระบบนี้สำเร็จได้อย่างดี ด้วยการสนับสนุนและความช่วยเหลือจากบุคคลสำคัญที่อยู่เบื้องหลัง ซึ่งข้าพเจ้ารู้สึกซาบซึ้งในความอนุเคราะห์และขอขอบคุณท่านทั้งหลาย ดังนี้

รศ.ดร. วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ที่ปรึกษาโครงการพัฒนาระบบ ที่ให้คำแนะนำและให้คำปรึกษาที่เป็นประโยชน์อย่างยิ่ง โดยเริ่มต้นศึกษาตั้งแต่วิชาสัมมนาจนกระทั่งโครงการนี้สำเร็จลุล่วงด้วยดี

คุณวิศาล เมฆสมณะศักดิ์ ที่ให้คำแนะนำและแนวทางในการวิเคราะห์ข้อมูลทางธุรกิจ รวมถึง บริษัท ดีเอชซี(ประเทศไทย) จำกัด ที่สนับสนุนเครื่องมือต่างๆ และข้อมูลอันสำคัญที่ใช้ในการพัฒนาระบบนี้

คณาจารย์คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยพระจอมเกล้าเจ้าคุณทหารลาดกระบังทุกๆ ท่าน ที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

เพื่อนๆ ทุกคนที่คอยให้กำลังใจ และให้ความช่วยเหลือทุกสิ่งทุกอย่าง

ท้ายที่สุดขอขอบคุณ บิดา มารดา และครอบครัวที่สนับสนุนข้าพเจ้าในทุกๆ เรื่องตลอดมา คุณค่าและประโยชน์จากโครงการพัฒนาระบบนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

นันทวัน นาคศิริ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การทำธุรกิจด้วยระบบเมล์ออเดอร์เป็นช่องการซื้อขายสินค้าที่ได้รับความนิยมในปัจจุบัน เนื่องจากพฤติกรรมการบริโภคสินค้าของคนรุ่นใหม่ได้เปลี่ยนแปลงไป จากการเปิดกว้างของข้อมูลสินค้าผ่านเครือข่ายอินเตอร์เนต ความสะดวกในการชำระเงินผ่านช่องทางต่างๆ และการได้รับสินค้าส่งตรงถึงมือ โดยเฉพาะผลิตภัณฑ์ด้านความงาม ประเทศในแถบเอเชีย ได้แก่ ญี่ปุ่น เกาหลี ไต้หวัน ฮ่องกง และไทย มีการการเติบโตของระบบเมล์ออเดอร์สูงขึ้นเรื่อยๆ มีผลให้หลายแบรนด์ต้องปรับกลยุทธ์แผนการตลาดและเพิ่มช่องทางการขายสินค้าใหม่ๆ เพื่อตอบสนองความต้องการของผู้บริโภคที่ไม่หยุดนิ่ง

ทั้งนี้ระบบเมล์ออเดอร์ได้จัดเก็บประวัติต่างๆ ของลูกค้า รวมถึงใช้รหัสสมาชิกอ้างอิงในข้อมูลใบสั่งซื้อทุกๆ ครั้ง เมื่อธุรกิจดำเนินไปได้สักระยะหนึ่ง ข้อมูลสมาชิกและข้อมูลการสั่งซื้อสินค้าย่อมมีจำนวนเพิ่มมากขึ้น ซึ่งข้อมูลต่างๆ เหล่านี้เป็นข้อมูลที่มีความสำคัญกับธุรกิจเป็นอย่างมาก จึงมุ่งหวังที่จะขุดค้นหาความสัมพันธ์ที่ซ่อนอยู่ในฐานข้อมูล เพื่อศึกษาพฤติกรรมการซื้อสินค้าในระยะยาว วิเคราะห์แนวโน้มในการเลือกซื้อสินค้าแบบต่อเนื่องที่จะเกิดขึ้นในอนาคต โดยใช้คาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining เป็นเครื่องมือหารูปแบบการสั่งซื้อสินค้าที่เกิดขึ้นอย่างต่อเนื่อง ซึ่งจะเป็นข้อมูลสำคัญในการวางแผนกลยุทธ์ทางการตลาดเพื่อให้การเสนอขายสินค้าตรงกับความต้องการของลูกค้ามากขึ้น การวิเคราะห์เพื่อวางแผนกลยุทธ์ทางการตลาดที่แม่นยำนั้น เป็นหัวใจหลักในการส่งเสริมสมาชิกเก่าให้ซื้อสินค้าอย่างต่อเนื่องและเสาะหากลุ่มสมาชิกใหม่ๆ ส่งผลให้ธุรกิจเติบโตขึ้นและสามารถแข่งขันกับคู่แข่งต่อไปได้

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

การศึกษานี้มุ่งหวังที่จะวิเคราะห์ข้อมูล เพื่อให้ได้รูปแบบของการซื้อสินค้าที่เกิดขึ้นต่อเนื่อง เพื่อศึกษาพฤติกรรมของลูกค้าในระยะยาว ดังนั้น จึงได้จัดทำโปรแกรมเพื่อใช้เป็นเครื่องมือในการขุดค้นข้อมูล (Data Mining) ในฐานข้อมูลด้วยเทคนิค Sequential Pattern Mining เพื่อให้ได้รูปแบบความสัมพันธ์ของข้อมูลแบบต่อเนื่องดังที่กล่าวไว้ข้างต้น ซึ่งจะเป็นข้อมูลสำคัญที่ผู้บริหารจะนำมาใช้ในการวางแผนทางธุรกิจได้อย่างมีประสิทธิภาพ และเป็นข้อมูลเบื้องต้นสำหรับวิเคราะห์ข้อมูลในขั้นสูงต่อไป

1.3 ขอบเขตของโครงการ

ระบบที่จะศึกษาและได้ทำการพัฒนาขึ้นนี้ จะเป็นระบบที่ทำงานตามขั้นตอนวิธีของดาต้าไมน์นิ่ง คือ การคัดเลือกข้อมูล การเตรียมข้อมูล การแปลงข้อมูลให้เหมาะสมหลังจากนั้นจะใช้เครื่องมือในการขุดค้นที่ชื่อว่า "CloSpan" ของ "illumine" และสรุปผลลัพท์ที่ได้มา ซึ่งขอบเขตของการศึกษานั้น มีดังนี้

1. สามารถขุดค้นข้อมูลด้วยเทคนิควิธี Sequential Mining Pattern จากฐานข้อมูล MS SQL Server เท่านั้น
2. ระบบเปิดกว้างให้ผู้ใช้สามารถเลือกข้อมูลและกำหนดความสัมพันธ์ของข้อมูลได้เอง
3. สามารถวิเคราะห์ผลและนำผลที่ได้จากการขุดค้นข้อมูลไปวางแผนการขายได้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาและพัฒนาระบบงานนี้ คาดว่าจะได้ประโยชน์ดังต่อไปนี้

1. ได้ศึกษาขั้นตอนการทำดาต้าไมน์นิ่งและสามารถวิเคราะห์ผลลัพธ์ได้
2. ได้ศึกษาค้นคว้า และเลือกเครื่องมือที่มีประสิทธิภาพมาใช้ในการไมน์นิ่ง
3. นำข้อมูลจากฐานข้อมูลมาสร้างมูลค่าให้กับองค์กร
4. นำความรู้และผลที่ได้จากการวิเคราะห์ข้อมูลนี้ไปศึกษาและพัฒนาต่อไป เพื่อขยายผลการวิเคราะห์ข้อมูลที่จะได้ประโยชน์สูงมากยิ่งขึ้น

1.5 ขั้นตอนในการพัฒนาระบบงาน

ขั้นตอนในการพัฒนาระบบงานมีดังต่อไปนี้

1. กำหนดคำถาม ข้อมูลที่ต้องการจากการทำดาต้าไมน์นิ่ง
2. ศึกษาเทคนิคและวิธีการในการทำดาต้าไมน์นิ่ง และอัลกอริทึมที่จะนำมาใช้ในการขุดค้นข้อมูล เพื่อให้ได้ผลตามที่ตั้งไว้
3. เลือกเครื่องมือที่จะใช้ในการทำดาต้าไมน์นิ่ง และศึกษาการทำงานของเครื่องมือนั้น
4. รวบรวมและจัดเตรียมข้อมูลที่จะใช้ในการทำดาต้าไมน์นิ่ง
5. ศึกษารายละเอียดและออกแบบระบบที่จะพัฒนาขึ้นให้รองรับกับเครื่องมือที่ได้เลือกใช้
6. พัฒนาระบบงาน เพื่อเตรียมข้อมูลและเรียกใช้เครื่องมือการทำดาต้าไมน์นิ่ง
7. นำผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่งเฉพาะรูปแบบที่น่าสนใจ มาแสดงผลในรูปแบบที่อ่านง่าย
8. ทดสอบและพัฒนาปรับปรุงระบบให้มีความถูกต้อง
9. สรุปผลการทดสอบจากการใช้งานและจัดทำเอกสารคู่มือของระบบ

1.6 ข้อจำกัดของการศึกษา

ข้อจำกัดของการศึกษาและพัฒนาระบบงานนี้ คือ ข้อมูลที่จะนำใช้ในการทำดาต้าไมน์นิ่งนั้นเป็นข้อมูลสำคัญของการทำธุรกิจและเป็นความลับที่ไม่สามารถเปิดเผยได้ ทำให้ต้องใช้เวลาค่อนข้างมากเพื่อจะจำกัดและเปลี่ยนแปลงข้อมูลที่เป็นกลยุทธ์ทางธุรกิจ ก่อนที่จะนำข้อมูลมาใช้ในการศึกษาและพัฒนาระบบงานนี้ได้

1.7 รายละเอียดในบทต่างๆ

รายละเอียดของเนื้อหาในการพัฒนาระบบงาน แบ่งออกเป็น 6 บท ดังต่อไปนี้

บทที่ 1 กล่าวถึงความเป็นมาของการพัฒนาระบบงาน ความมุ่งหมายและวัตถุประสงค์ ขอบเขตของโครงการ ประโยชน์ที่คาดว่าจะได้รับ ขั้นตอนในการพัฒนาระบบงาน และข้อจำกัดของการศึกษา

บทที่ 2 กล่าวถึงกระบวนการทำดาต้าไมน์นิ่ง คือ ขั้นตอนและวิธีการ โอเปอร์เรชั่นและอัลกอริทึมในการวิเคราะห์ข้อมูล รวมไปถึงรายละเอียดของเทคนิค Sequential Pattern Mining ที่จะนำมาใช้ในการขุดค้นข้อมูล

บทที่ 3 กล่าวถึงเครื่องมือและโปรแกรมที่จะใช้ในการพัฒนาระบบ

บทที่ 4 กล่าวถึงการวิเคราะห์และออกแบบระบบที่จะสร้างขึ้น

บทที่ 5 กล่าวถึงการสร้างและทดสอบระบบ และอธิบายการทำงานของโปรแกรมที่ได้พัฒนาขึ้นแยกตามแต่ละหน้าจอที่ใช้งาน

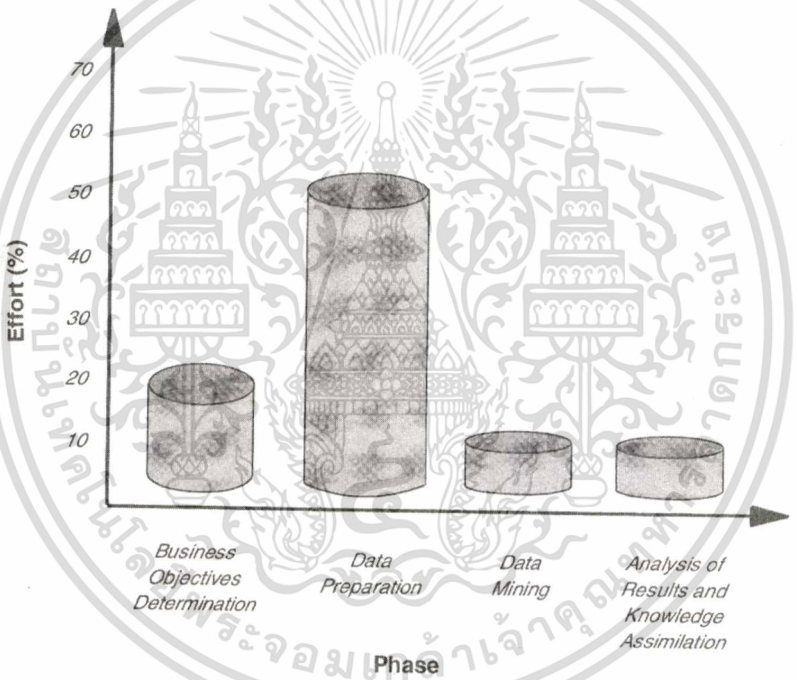
บทที่ 6 เป็นบทสรุปผลการศึกษา ประโยชน์ของระบบที่พัฒนาขึ้น ปัญหาที่พบ และข้อเสนอแนะที่จะพัฒนาต่อไป

บทที่ 2

Data Mining และ Sequential Pattern Mining

ดาต้าไมน์นิ่ง (Data Mining) คือ กระบวนการค้นหาลักษณะที่น่าสนใจของข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่ หรือที่เรียกว่า Knowledge Discovery in Database (KDD) ซึ่งมีวิธีการ (Operation) ต่างๆ มากมาย ที่จะใช้ขุดค้นข้อมูลเพื่อให้ได้ผลลัพธ์ตามเป้าหมายที่ต้องการ

2.1 ขั้นตอนวิธีการทำ Data Mining



ภาพที่ 2.1 แผนภาพแสดงเวลาของขั้นตอนการไมน์นิ่ง

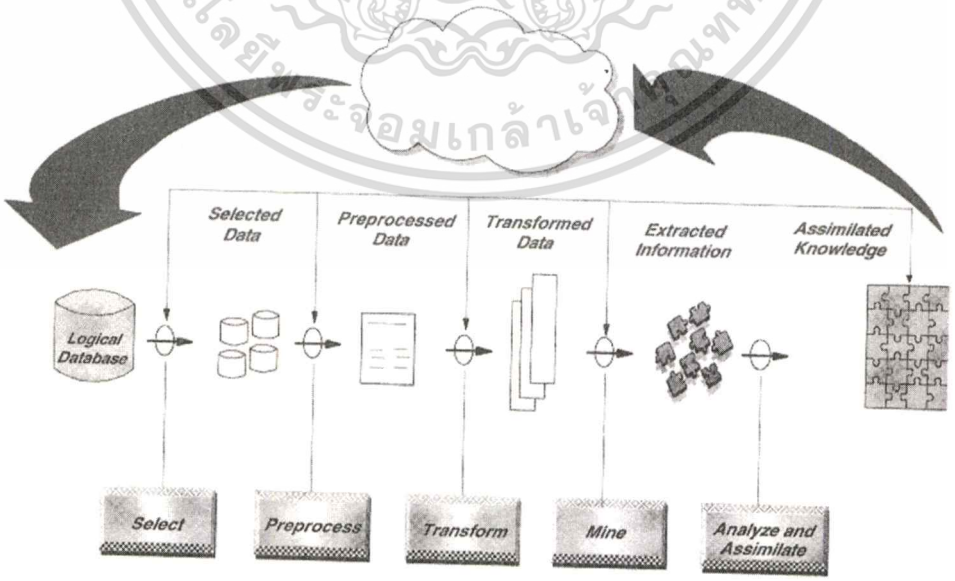
การทำดาต้าไมน์นิ่งแบ่งออกเป็น 4 ขั้นตอนหลักๆ ตามภาพที่ 2.1 อธิบายได้ดังนี้

2.1.1 การสร้างวัตถุประสงค์และกำหนดเป้าหมายในการทำดาต้าไมน์นิ่ง (Business Objectives Determination) ขั้นตอนนี้จะเป็นขั้นตอนที่สำคัญที่สุดที่จะทำให้ผลลัพธ์จากการไมน์นิ่งประสบความสำเร็จ เพราะการสร้างวัตถุประสงค์และกำหนดเป้าหมายในการทำดาต้าไมน์นิ่งให้ชัดเจนนั้น จะเป็นสิ่งที่กำหนดว่าจะต้องเลือกใช้วิธีการ (Operation) และอัลกอริทึมใดในการขุดค้นข้อมูล

2.1.2 การเตรียมข้อมูลที่จะทำการไมน์นิ่ง (Data Preparation) เนื่องจากข้อมูลต่างๆ ที่เก็บรวบรวมไว้ในฐานข้อมูลนั้น บางส่วนอาจจะเป็นข้อมูลที่ผิดพลาด หรือไม่สมบูรณ์ อาจมีสาเหตุมาจากผู้ที่คีย์ข้อมูลอาจใส่ข้อมูลไม่ครบถ้วน หรืออาจเกิดจากมีผู้ทำงานหลายๆ คน ทำให้ข้อมูลที่ใส่อาจมีค่าไม่เหมือนกันทั้งๆ ที่เป็นข้อมูลเดียวกัน เช่น ข้อมูลอาชีพของสมาชิก อาจเป็นได้ทั้ง "โปรแกรมเมอร์" หรือ "Programmer" ดังนั้น จึงต้องมีการจัดเตรียมข้อมูลที่จะทำดาต้าไมน์นิ่งให้ถูกต้องและสมบูรณ์ที่สุดเสียก่อน ขั้นตอนนี้เป็นขั้นตอนที่ใช้เวลามากที่สุดถึง 60 เปอร์เซ็นต์ ซึ่งข้อมูลที่จะนำมาทำการไมน์นิ่งนั้น จะต้องเป็นข้อมูลที่มีความเหมาะสม นั่นคือ ข้อมูลที่จะไมน์นิ่งจะต้องเป็นข้อมูลที่จะใช้เท่านั้น (Data Selection) ส่วนข้อมูลอื่นๆ ที่ไม่เกี่ยวข้องจะต้องนำออกไป ข้อมูลที่จะใช้ไมน์นิ่งจะต้องเป็นข้อมูลที่มีคุณภาพดี (Data Preprocessing) คือ จะต้องแก้ไขให้สมบูรณ์ ต้องไม่มีค่าว่าง และปรับเปลี่ยนรูปแบบข้อมูล (Data Transformation) เพื่อให้ข้อมูลที่ได้มีความเหมาะสมกับการตัดสินใจ

2.1.3 กระบวนการทำดาต้าไมน์นิ่ง (Data Mining) ตามอัลกอริทึมที่เหมาะสม ซึ่งเป็นหัวใจของขั้นตอนทั้งหมด

2.1.4 การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมน์นิ่ง (Analysis of Results and Knowledge Assimilation) ว่าสามารถนำผลลัพธ์ที่ได้นั้นไปใช้ประโยชน์ได้หรือไม่ และถ้าหากผลลัพธ์ที่ได้นั้น ไม่น่าพอใจ ก็ต้องกลับไปตรวจสอบขั้นตอนต่างๆ และทำกระบวนการดาต้าไมน์นิ่งใหม่ เมื่อผลลัพธ์การทำดาต้าไมน์นิ่งเป็นที่น่าพอใจแล้วจึงรวบรวม ประมวลข้อมูล เพื่อเป็นองค์ความรู้ของธุรกิจต่อไป



ภาพที่ 2.2 แสดงขั้นตอนย่อยของขั้นตอนการเตรียมข้อมูลและการวิเคราะห์ผลการทำดาต้าไมน์นิ่ง

2.2 อัลกอริทึมของดาต้าไมน์นิ่ง

จุดประสงค์ของการทำดาต้าไมน์นิ่งนั้นจะวิเคราะห์ข้อมูลออกเป็น 3 ประเภทใหญ่ๆ คือ เพื่อการจัดการทางการตลาด (Market Management), เพื่อการจัดการความเสี่ยง (Risk Management) และ เพื่อการจัดการกลโกง(Fraud Management)

ซึ่งในการทำดาต้าไมน์นิ่งแต่ละวิธีการ (Operation) นั้น มีหลากหลายอัลกอริทึมที่จะใช้ในการประมวลผลข้อมูลซึ่งการจะเลือกใช้อัลกอริทึมใดมาวิเคราะห์ข้อมูลขึ้นอยู่กับผลลัพธ์ที่ต้องการหรือลักษณะของข้อมูลที่มีในระบบฐานข้อมูล ดังแสดงได้ตามตารางด้านล่าง

ตารางที่ 2.1 แสดงความสัมพันธ์ของจุดประสงค์ วิธีการ และอัลกอริทึมในการทำดาต้าไมน์นิ่ง (Data Mining Application and Supporting Operations and Techniques)

	Market Management		Risk Management	Fraud Management
Applications	Target marketing Customer relationship management Market basket analysis Cross selling Market segmentation		Forecasting Customer retention Improved underwriting Quality control Competitive analysis	Fraud detection
Operations	**Predictive Modeling**	**Database Segmentation**	**Link Analysis**	**Deviation Detection**
Techniques	Classification Value prediction	Demographic clustering Neural clustering	Associations discovery Sequential pattern discovery Similar time sequence discovery	Visualization Statistics

2.2.1 Predictive Modeling

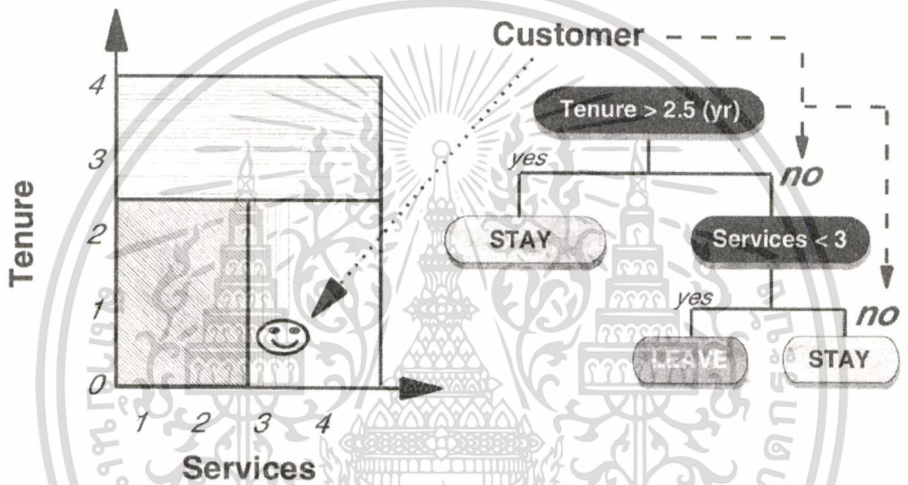
Predictive เป็น Model ในการทำดาต้าไมน์นิ่งเพื่อทำนายหรือหาแนวโน้มของเหตุการณ์โดยอาศัยข้อมูลเดิมที่มีอยู่ จะแยกออกเป็นสองส่วน คือ วิธีการดาต้าไมน์นิ่งแบบทางตรง(Direct Data Mining) และ วิธีการดาต้าไมน์นิ่งแบบทางอ้อม(Undirected Data Mining) โดยวิธีการแบบทางตรงนั้นจะรู้ผลลัพธ์แล้วคำนวณค่าแต่ละผลลัพธ์ ส่วนวิธีการดาต้าไมน์นิ่งแบบทางอ้อมนั้นจะรู้รูปแบบวิธีการในการทำงานแล้วต้องการทราบผลลัพธ์ Predictive Modeling มี 2 เทคนิค คือ

2.2.1.1 Classification เป็นการแบ่งกลุ่มของ Record ในฐานข้อมูล โดยจะสำรวจจุดเด่นข้อมูลที่ปรากฏออกมา แล้วจึงกำหนดจุดเด่นนั้นเป็นตัวแบ่งในการจัดหมวดหมู่

2.2.1.2 Value predictive เป็นการประมาณค่าออกมาเป็นตัวเลขที่ต่อเนื่องที่มีความสัมพันธ์กับ Record ในฐานข้อมูล

ตัวอย่างของ Predictive Modeling คือ อัลกอริทึม Decision Tree เป็นอัลกอริทึ่มที่รู้จักกันดีของเทคนิคคลาสสิฟิเคชัน (Classification) เป็นเทคนิคที่ใช้สำหรับสถานการณ์ที่เจาะจง เป็นการเรียนรู้ฟังก์ชันการจัดกลุ่มข้อมูลจากกลุ่มข้อมูลที่กำหนดให้ ขั้นตอนการเรียนรู้ในการทำดิซิชั่นทรี (Decision Tree)นี้ จะให้ผลลัพธ์ที่สามารถนำมาสร้างเป็นกฎได้

การทำดาต้าไมน์นิ่งด้วย Predictive Modeling ส่วนใหญ่เราจะใช้กับข้อมูลที่แบ่งแยกเป็นประเภทหรือข้อมูลที่ได้ตัดสินใจไว้แล้ว เช่น การอนุมัติการกู้ยืมเงิน เป็นต้น



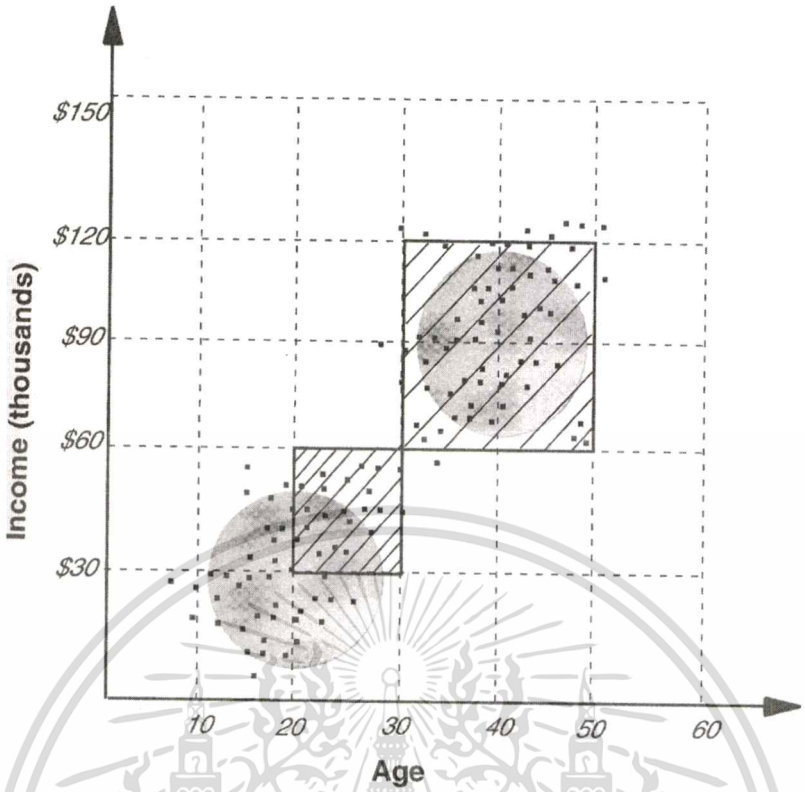
ภาพที่ 2.3 ตัวอย่างการวิเคราะห์ด้วยดิซิชั่นทรี

2.2.2 Database Segmentation

Database Segmentation หรือ Clustering Analysis เป็นเทคนิคหนึ่งในการทำดาต้าไมน์นิ่งเพื่อใช้ในการจำแนกกลุ่ม (Clustering) ของข้อมูลในฐานข้อมูล โดยที่ข้อมูลในแต่ละกลุ่มจะมีความคล้ายคลึงกัน เมื่อแบ่งข้อมูลได้แล้ว จึงมาดูข้อมูลในแต่ละกลุ่มว่ามีความเหมือนหรือคล้ายกันอย่างไร ซึ่งมี 2 เทคนิค คือ

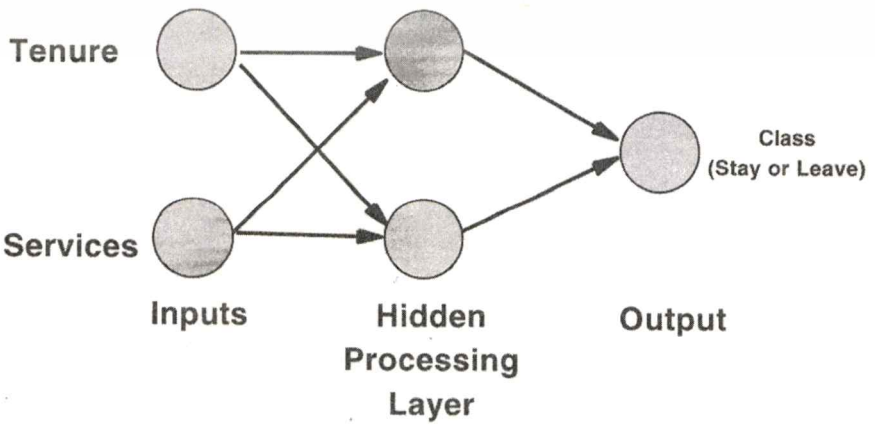
2.2.2.1 Demographic clustering เป็นการแบ่งกลุ่มโดยอาศัยเทคนิคการวัดที่ใช้พื้นฐานของการโหวต ที่เรียกว่า Condorset เช่น การแบ่งกลุ่มของข้อมูลที่เข้ามาใหม่ โดยเทียบกับข้อมูลของกลุ่มเดิม ว่ามีความใกล้เคียงกับกลุ่มไหนมากที่สุด ก็จะถูกจัดเข้าไปอยู่กลุ่มนั้น แต่จะยากตรงที่จะใช้คุณสมบัติอะไรเป็นตัวเปรียบเทียบ

ตัวอย่างของ Segmentation คือ อัลกอริทึม K-mean เป็นอัลกอริทึ่มที่รู้จักกันดีของเทคนิคนี้ เป็นเทคนิคที่ให้ข้อมูลเรียนรู้กันเองว่าอยู่ในกลุ่มไหนแล้วหาค่าเฉลี่ยใหม่ จากนั้นก็ทำการเรียนรู้ต่อไปเรื่อยๆ จนกระทั่งค่าเฉลี่ยที่ได้ไม่เปลี่ยนแปลง



ภาพที่ 2.4 ตัวอย่างการวิเคราะห์แบบ Segmentation

2.2.2.2 Neural clustering เป็นการสร้าง Neural Network โดยเทคนิคนี้จะยอมรับเฉพาะข้อมูลที่เป็นตัวเลขเท่านั้น การวิเคราะห์ด้วย Neural Network จะเป็นการให้เรียนรู้จากตัวอย่างต้นแบบเหตุการณ์ที่เกิดขึ้น แล้วได้ผลลัพธ์เป็นคลาสของข้อมูลออกมา การวิเคราะห์ด้วยวิธีนิวรอลเตเวิร์ค ระบบจะสามารถเรียนรู้แล้วแก้ปัญหาที่กว้างขึ้นได้ แต่ปัญหาในการวิเคราะห์แบบ Neural Network คือ จะมีส่วนของ Hidden Processing Layer อยู่ ซึ่งได้ถูกซ่อนอยู่ ส่วนใหญ่แล้วจะไม่รู้ว่ามีการทำงานอย่างไร จะรู้แค่ค่า Input กับ Output เท่านั้น

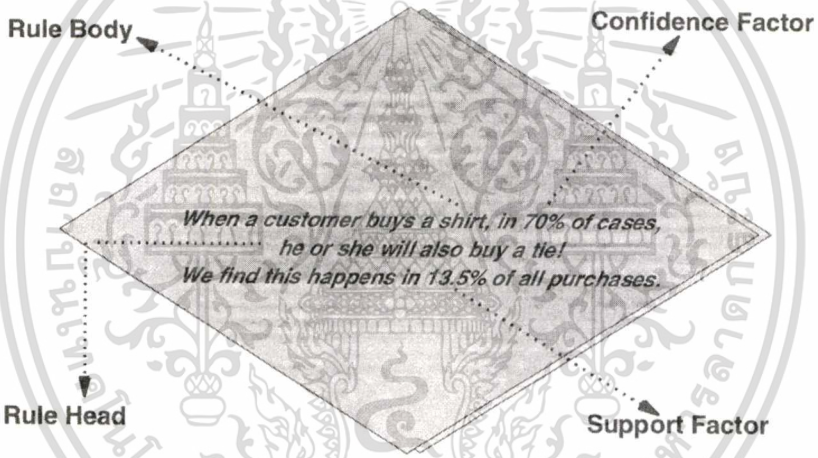


ภาพที่ 2.5 ตัวอย่างการวิเคราะห์ด้วย Neural Network

2.2.3 Link Analysis

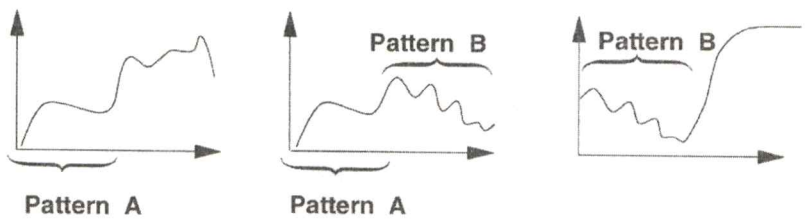
Link Analysis นั้นจะรู้จักกันดีกับอัลกอริทึมเทคนิคกฎความสัมพันธ์(Association Rules) ซึ่งเป็นการวิเคราะห์ความสัมพันธ์ที่น่าสนใจของข้อมูลที่อยู่ในฐานข้อมูล ผลลัพธ์ที่ได้ คือ เซทข้อมูลของเหตุการณ์ที่เกิดขึ้นพร้อมกันบ่อยๆ ซึ่ง มี 3 เทคนิค คือ

2.2.3.1 Association discovery การทำดาต้าไมน์นิ่งด้วยวิธีนี้จะเป็นที่นิยมและถูกใช้ใน Super Market เป็นส่วนใหญ่ ซึ่งอัลกอริทึมนี้จะใช้ในการหาความสัมพันธ์ระหว่างลูกค้ากับสินค้าหรือบริการ เพื่อใช้ประโยชน์ในการจัดโปรโมชั่น หรือวางแผนส่งเสริมการขายสินค้า ตัวอย่างอัลกอริทึมของโอเปอร์เรชั่นนี้ เช่น วิธี Apriori การแก้ปัญหาของอัลกอริทึมจะหาความสัมพันธ์ระหว่างค่า Confidence และ ค่า Support แล้วสร้างกฎขึ้นมาจากการนับจำนวนของข้อมูลทั้งหมดเพื่อเลือกข้อมูลที่มีจำนวนนับสูงกว่าค่า Support



ภาพที่ 2.6 ตัวอย่างการวิเคราะห์แบบ Association Rule

หลังจากที่ได้สร้างกฎของ Association Rule แล้ว จะได้ Pattern ของข้อมูลออกมา แต่เราจะสนใจ Pattern ข้อมูลที่มีลักษณะแปลกๆ เท่านั้น เช่น ถ้าได้ความสัมพันธ์ของขวดนมกับนมผงเด็กนั้น จะเป็น Pattern แบบปกติ เราจะไม่สนใจ แต่ถ้าเป็นข้อมูลของขวดนมกับเบียร์ ซึ่งเป็นความสัมพันธ์ที่ไม่ปกตินั้น เราจะให้ความสนใจ ซึ่งจะใช้วิธีการวิเคราะห์แบบ Pattern Matching



ภาพที่ 2.7 ตัวอย่าง การวิเคราะห์แบบ Pattern Matching

2.2.3.2 Sequential pattern discovery เป็นเทคนิคที่ใช้วิเคราะห์ข้อมูลที่ต่อเนื่องว่าเมื่อเกิดเหตุการณ์หนึ่งขึ้นแล้วจะเกิดอะไรต่อไป เป็นการหาลำดับของเหตุการณ์ที่เกิดขึ้นบ่อยๆ

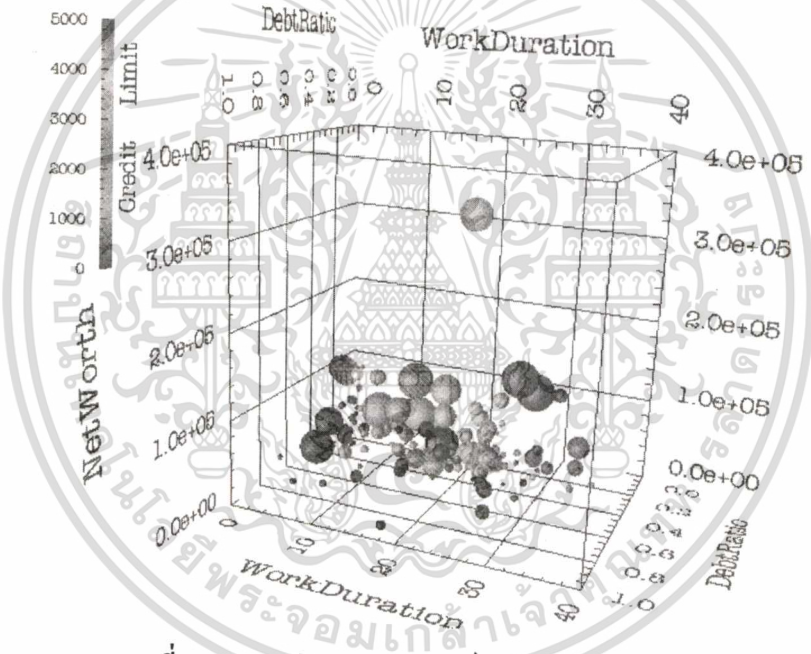
2.2.3.3 Similar time sequence discovery เป็นเทคนิคที่ใช้วิเคราะห์ในตลาดหุ้น

2.2.4 Deviation Detection

Deviation Detection เป็นโมเดลในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน โดยทั่วไปมักใช้วิธีทางสถิติ ส่วนมากอาศัยการ plot graph แล้วการกระจายของจุด เพื่อตรวจสอบกลโกงที่อาจจะเกิดขึ้น โดยดูจากความคลาดเคลื่อน หรือข้อมูลที่เบี่ยงเบนไปจากค่ามาตรฐาน จะมี 2 เทคนิค คือ

2.2.4.1 Visualization

2.2.4.2 Statistics



ภาพที่ 2.8 ตัวอย่างการวิเคราะห์แบบ Fraud Probability

2.3 Sequential Pattern Mining

Sequential Pattern Mining เป็นเทคนิควิธีการทำดาต้าไมน์นิ่งในกลุ่มของ Link Analysis ที่รู้จักกันดี “กฎความสัมพันธ์” ใช้วิเคราะห์ข้อมูลแบบต่อเนื่องโดยมีเรื่องของเวลาเข้ามาเกี่ยวข้องด้วย ว่าเมื่อเกิดเหตุการณ์หนึ่งขึ้นแล้วจะเกิดอะไรต่อไป เป็นการหาลำดับของเหตุการณ์ที่เกิดขึ้นบ่อยๆ เช่น ข้อมูลการซื้อสินค้า, ข้อมูลในการใช้โทรศัพท์, ข้อมูลภัยธรรมชาติ เช่น แผ่นดินไหว หรือพายุเฮอริเคน, การวินิจฉัยโรคและการรักษา, การผันผวนของคลังสินค้า, การวิเคราะห์ Weblog, การวิเคราะห์ DNA sequence เป็นต้น

การทำดาต้าไมน์นิ่งด้วย Sequential pattern นั้น มีการศึกษาและพัฒนาขึ้นหลายอัลกอริทึม เช่น Apriori, GSP, SPADE, PrefixSpan, CloSpan, IncSpan

2.4 ข้อมูลการซื้อสินค้าแบบต่อเนื่อง

ข้อมูลการซื้อสินค้าของลูกค้าที่ใช้รหัสสมาชิกอ้างอิงในข้อมูลใบสั่งซื้อทุกๆ ครั้งนั้น จะสามารถนำมาใช้ในการทำดาต้าไมน์นิ่งด้วยเทคนิค Sequential pattern ได้

ตารางที่ 2.2 แสดงตัวอย่างของข้อมูลการซื้อสินค้าแบบต่อเนื่อง

SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

จากตารางที่ 2.2 อธิบายได้ว่า SID คือ รหัสมาชิก ส่วน sequence คือ เซทของข้อมูลการซื้อสินค้า ถ้าหา Sequence Pattern แล้ว จะได้เป็น

support threshold min_sup =2, <(ab)c> เป็น sequential pattern (2.1)

ในการหา Sequential patterns ในฐานข้อมูลแบบทรานแซคชั่นที่มีขนาดใหญ่ๆ นั้น สิ่งที่สำคัญในการไมน์นิ่งคือ ความเชื่อมั่นและค่าซัพพอร์ต

ให้ $I = \{i_1, i_2, \ldots, i_n\}$ เป็นเซ็ทของข้อมูล(Item)

เรียกซับเซ็ท $X \subseteq I$

$|X|$ เป็นขนาดของ X

ลำดับ $s = (s_1, s_1, \ldots, s_m)$ ลำดับรายการของเซ็ทข้อมูล

$S_i \subseteq I, i \in \{1,\ldots,m\}$; m :sequence number of itemset

Lenge of sequence $s = (s_1, s_1, \ldots, s_m)$ นิยามได้เป็น

$$l \stackrel{\text{def}}{=} \sum_{i=1}^{m} |s_i|. \tag{2.2}$$

$S_a = (a_1, a_2, .., a_n)$ ถูกบรรจุในลำดับ $S_b = (b_1, b_2, .., b_m)$

ถ้า exist integers $1 \leq i_1 < i_2 < \ldots < i_n \leq m$

ดังนั้น $a_1 \subseteq b_{i1}, a_2 \subseteq b_{i2}, \ldots, a_n \subseteq b_{in}$

ถ้า sequence s_a is contained in sequence s_b ดังนั้น

เรียก s_a ว่าเป็น subsequence ของ s_b

และเรียก s_b ว่าเป็น supersequence ของ s_a

2.5 Sequential Pattern Mining Algorithms

ในปัจจุบัน ได้มีผู้คิดค้นและพัฒนาอัลกอริทึมในการ Sequential Pattern Mining ขึ้น โดยมีพื้นฐานมาจากอัลกอริทึม Apiori ซึ่งแจกแจงได้ดังนี้

2.5.1 Apriori (Agrawal & Sirkant'94)

วิธีนี้ถูกออกแบบมาแก้ปัญหาของข้อมูลที่เป็นลักษณะของการซื้อสินค้าในลักษณะเป็นตะกร้ารายการสินค้า วิธีคิด Apriori มี 5 Phase

1) Sort Phase ขั้นการจัดเรียงข้อมูลในฐานข้อมูลโดยใช้ Primary key คือ Customer- id และกำหนด secondary key เป็น Transation- time แล้วเก็บลงฐานข้อมูลโดยจัดกลุ่มตามรหัสลูกค้าเป็นหลัก

2) Litemset Phase ขั้นการกำหนดลิมิตนั้น จะได้ทำการเลือกข้อมูลมา 3 ส่วน คือ Transaction Time, Customer ID, Items Bought จาก Transection Database ซึ่งแสดงดังตารางด้านล่าง

ตารางที่ 2.3 Apriori Algorithm: Litemset Phase

Transaction Time	Customer ID	Items Bought
June 10' 93	2	10,20
June 12' 93	5	90
June 15' 93	2	30
June 20' 93	2	40,60,70
June 25' 93	4	30
June 25' 93	3	30,50,70
June 25' 93	1	30
June 30' 93	1	90
June 30' 93	4	40,70
June 25' 93	4	90

3) Transformation Phase ขั้นการเปลี่ยนสภาพจาก Transection Database มาเป็นข้อมูลในลักษณะของ Set ตามตารางดังนี้

ตารางที่ 2.4 Apriori Algorithm: Transformation Phase

Customer ID Sequence	Original Customer Tramsformed Customer After Mapping
1	[(30)(90)] [{(30)}{(90)}] [{1}{5}]
2	[(10 20)(30)(40 60 70)] [{(30)}{(40), (70), (40, 70)}] [{1}{2, 3, 4}]
3	[(30 50 70)] [{(30)}{(70)}] [{1, 3}]
4	[(30)(40 70)(90)] [{(30)}{(40), (70), (40, 70)}{(90)}] [{1}{2, 3, 4}{5}]
5	[(90)] [{(90)}] [{5}]

4) Sequence Phase ขั้นจัดลำดับ จะทำการเลือกข้อมูลจากเซ็ทมา Candidate กัน ยกตัวอย่างจากตารางด้านล่างเช่น < 1 2 3 > กับ < 1 2 4 > เพราะมี 12 เป็นข้อมูลสองลำดับแรก เหมือนกันคือ < 1 2 > แต่ข้อมูลในเซ็ท < 2 3 4 > จะมีลำดับสองข้อมูลแรกเป็น < 2 3 >

5) Maximal Phase ขั้นหาข้อมูลถูกพบในลำดับถัดไปมากที่สุด เพราะจะเป็นข้อมูลที่มี ความเป็นไปได้สูงที่สุด

ถึงแม้ว่า Apriori จะเป็นอัลกอริทึมที่ได้รับความนิยมมาก แต่ Apriori ก็ยังมีข้อจำกัดคือขาด ในเรื่องของการนำเวลามาเป็นตัวกำหนดเพื่อหาช่วงเวลาที่มีลำดับข้อมูลที่ซ้ำกันบ่อยๆ ได้

2.5.2 GSP : Generalized Sequential Pattern (Agrawal and Sirkant EDBT'96)

GSP เป็นอัลกอริทึมแบบ best-know algorithms ที่มีแนวคิดพื้นฐานจาก Apriori-base โดย จะมีวิธีคิดเป็น 2 Phases

1) Join Phase ข้อมูลที่ได้รับเลือกจะนำมา Candidateตามลำดับที่ถูกสร้าง โดย join ตามลำดับบ่อยตัดสินใจในระดับล่าสุด

จากตารางที่ 2.2 ตัวอย่างข้อมูลนั้น นำมา Join Phase ด้วยค่า support threshold = 2 จะได้ดังตารางที่ 2.5

ตารางที่ 2.5 GPS Algorithm: Join Phase

	<a>	<b>	<c>	<d>	<e>	<f>
<a>	<aa>	<ab>	<ac>	<ad>	<ae>	<af>
<b>	<ba>	<bb>	<bc>	<bd>	<be>	<bf>
<c>	<ca>	<cb>	<cc>	<cd>	<ce>	<cf>
<d>	<da>	<db>	<dc>	<dd>	<de>	<df>
<e>	<ea>	<eb>	<ec>	<ed>	<ee>	<ef>
<f>	<fa>	<fb>	<fc>	<fd>	<fe>	<ff>

จากตารางที่ 2.5 ผลของการ Join จะใช้วิธีการ Join ด้วย Natural Join โดยได้ข้อมูลทั้งหมดออกมาจาก 6 ข้อมูล ได้เป็น 36 Pattern

2) Prune Phase ข้อมูลที่ได้รับเลือกให้ Candidate นั้นจะต้องมีลำดับที่มีข้อมูลเหล่านั้นติดกัน แล้วเหตุการณ์ภายหลังข้อมูลที่ไม่พบบ่อย (เหนือกว่าระดับขีดสุดการสนับสนุนน้อยที่สุด) จะถูกลบ

ตารางที่ 2.6 GPS Algorithm: Prune Phase

	<a>	<b>	<c>	<d>	<e>	<f>
<a>		<ab>	<ac>	<ad>	<ae>	<af>
<b>			<bc>	<bd>	<be>	<bf>
<c>				<cd>	<ce>	<cf>
<d>					<de>	<df>
<e>						<ef>
<f>						
Apriori prunes 44.57% candidates						

2.5.3 SPADE : Sequential PAttern Discovery using Equivalent Class (Zaki 2001)

SPADE เป็นวิธีคิดเพื่อลด I/O โดยลดจำนวนในการสแกนฐานข้อมูล ซึ่งจะต้องทำการสแกนฐานข้อมูลตามลำดับถึง 3 ครั้งจึงจะได้ข้อมูลที่พร้อมในการทำดาต้าไมน์นิ่ง วิธีคิดใช้รายการ

ของข้อมูลในแนว vertical ซึ่งจะเก็บข้อมูลเป็นคู่ๆ ดังตารางที่ 2.7

ตารางที่ 2.7 SPADE Algorithm

FREQUENT SEQUENCES
Frequent 1-Sequences
A – 4
B – 4
D – 2
F – 4
Frequent 2-Sequences
AB – 3
AF – 3
B ⟶ A – 2
BF – 4
D ⟶ A – 2
D ⟶ B – 2
D ⟶ F – 2
Frequent 3-Sequences
ABF – 3
BF ⟶ A- 2
D ⟶ BF – 2
D ⟶ B ⟶ A – 2
D ⟶ F ⟶ A – 2
Frequent 4-Sequences
BF ⟶ ABF ⟶ A – 2

2.5.4 Prefix Span : Prefix-projection Sequential PAttern Mining (Han et al.@KDD'00; Pei, et al.@ICDE'01)

แนวคิดของอัลกอริทึม PrefixSpan

$(\alpha, i, S|\alpha)$ (2.3)

1. Scan $S|\alpha$ once, find the set of frequent items b such that

- b can be assembled to the last element of α to form a sequential pattern; or
- <b> can be appended to α to form a sequential pattern.

2. For each frequent item b, appended it to α to form a sequential pattern α', and output α';

3. For each α', construct α'-projected database S|α', and call PrefixSpan (α', i+1,S|α').

ตารางที่ 2.8 ตัวอย่างข้อมูลในการทำดาต้าไมน์นิ่งด้วยอัลกอริทึม PrefixSpan

SID	Sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

ขั้นตอนในการทำ PrefixSpan คือ

1) หา length-1 sequential patterns จะได้เป็น <a>, <b>, <c>, <d>, <e>, <f> ดังตาราง

ตารางที่ 2.9 PrefixSpan: หา length-1 sequential patterns

Prefix	Sequential Patterns
<a>	<a>,<aa>,<ab><a(bc)>,<a(bc)a>,<aba>,<abc>,<(ab)>,<(ab)c>,<(ab)d>,<(ab)f>,<(ab)dc>,<ac>,<aca>,<acb>,<acc>,<ad>,<adc>,<af>
<b>	<b>, <ba>, <bc>, <(bc)>, <(bc)a>, <bd>, <bdc>,<bf>
<c>	<c>, <ca>, <cb>, <cc>
<d>	<d>,<db>,<dc>, <dcb>
<e>	<e>,<ea>,<eab>,<eac>,<eacb>,<eb>,<ebc>,<ec>,<ecb>,<ef>,<efb>,<efc>,<efcb>
<f>	<f>,<fb>,<fbc>, <fc>, <fcb>

2) เขียนเซ็ทของข้อมูลที่ขึ้นต้นตาม Prefix ด้านบน เช่น <a>-projected database: <(abc)(ac)d(cf)>, <(_d)c(bc)(ae)>, <(_b)(df)cb>, <(_f)cbc>

ตารางที่ 2.10 PrefixSpan: เขียนเซ็ทของข้อมูลที่ขึ้นต้นตาม Prefix ของ length-1

Prefix	Projected(suffix) databases	Sequential Patterns
<a>	<(abc)(ac)d(cf)>, <(_d)c(bc)(ae)>, <(_b)(df)cb>, <(_f)cbc>	<a>,<aa>,<ab><a(bc)>, <a(bc)a>,<aba>,<abc>,<(ab)>, <(ab)c>,<(ab)d>,<(ab)f>, <(ab)dc>,<ac>,<aca>,<acb>, <acc>,<ad>,<adc>,<af>

3) วนลูปหา length-2 sequential patterns เช่น Pattern <a> ได้เป็น <aa>, <ab>, <(ab)>, <ac>, <ad>, <af> แล้วเขียน Pattern ตาม Prefix ที่ได้ จากนั้นก็วนลูปหา Prefix เพิ่มขึ้นเป็น length-3,4,... ไปเรื่อยๆ

2.5.5 CloSpan : Closed Sequential Patterns (Yan,Han & Afshar @SDM'03)

ในการพัฒนาระบบนี้ ใช้อัลกอริทึม CloSpan ในการไมน์นิ่งข้อมูล ซึ่ง CloSpan มีแนวคิดพื้นฐานมาจาก PrefixSpan แต่ CloSpan เป็นอัลกอริทึมที่จะหารูปแบบลำดับเหตุการณ์ที่ใหญ่ที่สุดก่อน ซึ่งจะครอบคลุมรูปแบบของเหตุการณ์ย่อยด้วย ทำให้ใช้เวลาในการอ่านข้อมูลน้อยลงเพราะลดจำนวนครั้งในการอ่านข้อมูล

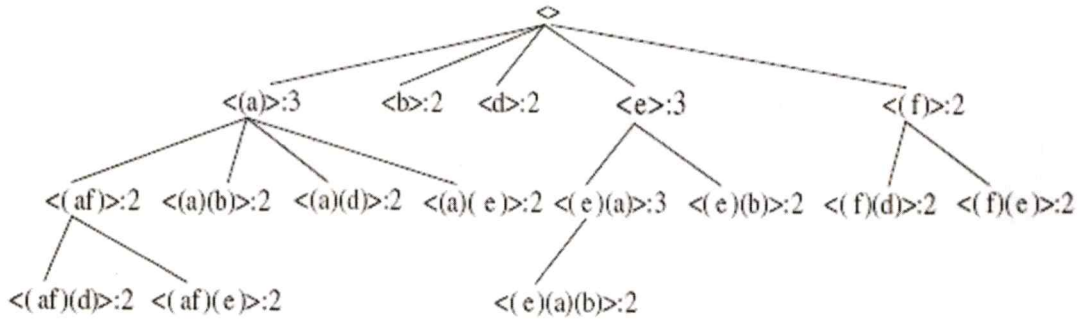
ตารางที่ 2.11 ตัวอย่างข้อมูลสำหรับ CloSpan

Seq ID	Sequence
0	<(a f)(d)(e)(a) >
1	<(e)(a)(b)>
2	<(e)(a b f)(b d e) >

จากข้อมูลข้างต้นจะได้ I(D) หมายถึง จำนวนข้อมูลทั้งหมด เป็น I(D) = 15 เขียนได้เป็น

$$\mathcal{I}(D) = \sum_{i=1}^{n} l(s_i). \qquad (2.4)$$

เมื่อนำข้อมูลมาแตกเป็นแผนภูมิต้นไม้ตามลำดับเหตุการณ์ (Lexicographic Sequence Tree) จะได้ดังภาพที่ 2.9 ซึ่งเป็นเหตุการณ์ที่เกิดขึ้นทั้งหมด 16 รูปแบบ แต่ถ้าเป็น Closed Pattern แล้วจะได้จำนวนรูปแบบที่น้อยลง กำหนดให้ค่า min_sup = 2 จะได้รูปแบบของเหตุการณ์แค่ 4 รูปแบบ คือ <(a f)(d)>: 2 , <(a f)(e)>: 2 , <(e)(a)>: 3 , <(e)(a)(b)>:2



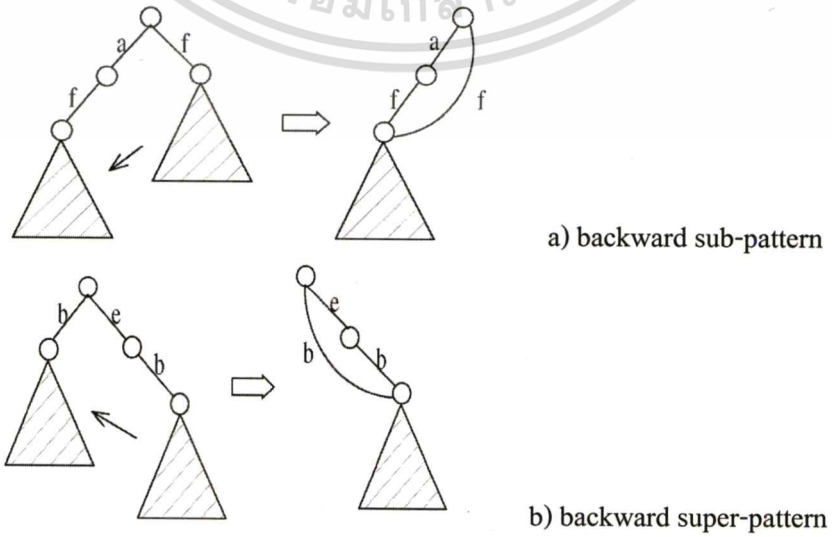
ภาพที่ 2.9 นำข้อมูลมาเขียนเป็น Tree ของ Sequence

CloSpan มีแนวคิดในการไมน์นิ่ง 3 ขั้นตอน ดังนี้

ขั้นแรกเรียกว่า Common Prefix คือ การหา Prefix ที่มีจำนวนเท่ากับค่า min_sup ยกตัวอย่างเช่น มีเซ็ทของข้อมูล {<(d)(e)(a f)>,<(d)(e)(f g)>} ถ้ากำหนดให้ค่า min_sup = 2 ดังนั้นจะได้ <(d)(e)> เป็น Common Prefix ซึ่งเป็นเทคนิคในการ Pruning

ขั้นที่สอง เรียกว่า Partial Order จากตารางที่ 2.11 Seq ID 0 คือเซ็ท <(a f)(d)(e)(a)> ซึ่งขึ้นต้นด้วย <(a f)> แต่เราจะไม่ค้นหาลำดับที่ขึ้นต้นด้วย <(f)>

ขั้นที่สาม เรียกว่า Early Termination by Equivalence ถ้าพบว่ามีเซ็ทของ <(a)(f)(Pattern)> และเซ็ท <(f)(Pattern)> แล้ว จะถือว่า Sequence <(f)> เป็นซัพเซ็ทของ Sequence <(a)(f)> ก็จะไม่สนใจ Sequence <(f)> เรียกว่า backward sub-pattern ในทำนองกลับกัน <(b) (Pattern)> และเซ็ท <(e)(b)(Pattern)> แล้ว จะถือว่า Sequence <(b)> เป็นซัพเซ็ทของ Sequence <(e)(b)> ก็จะไม่สนใจ Sequence <(b)> เรียกว่า backward super-pattern ดังภาพที่ 2.10



ภาพที่ 2.10 CloSpan ใช้แนวคิด Backward ใน Prune Phase

อัลกอริทึมของ CloSpan เขียนได้ดังนี้

(2.5)

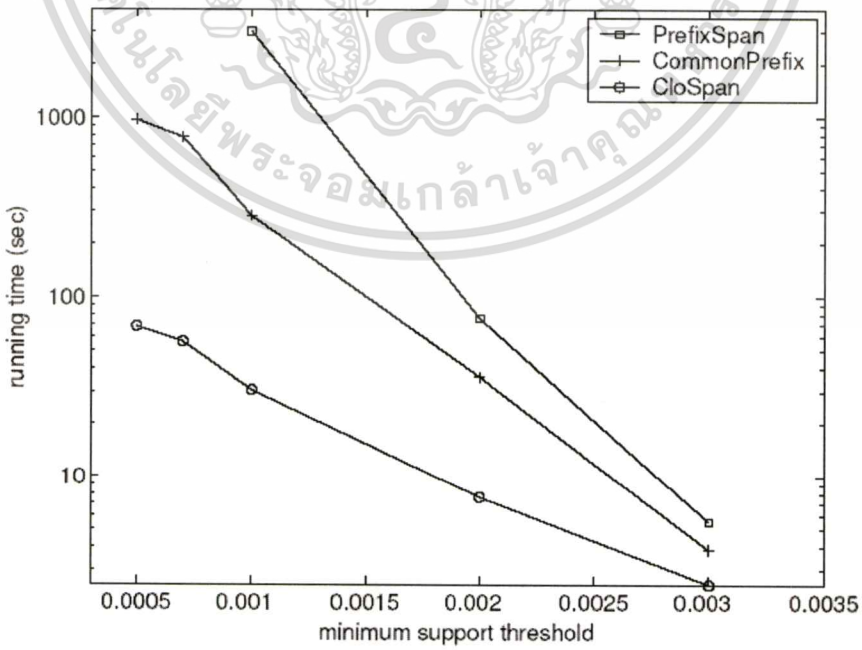
CloSpan(s, D_s, *min_sup*, L)

Input: A sequence s, a projected DB D_s, and *min_sup*.
Output: The prefix search lattice L.

1: Check whether a discovered sequence s' exists s.t. either $s \sqsubseteq s'$ or $s' \sqsubseteq s$, and $\mathcal{I}(D_s) = \mathcal{I}(D_{s'})$;
2: **if** such super-pattern or sub-pattern exists **then**
3: modify the link in L, **return**;
4: **else** insert s into L;
5: Scan D_s once, find every frequent item α such that
(a) s can be extended to $(s \diamond_i \alpha)$, or
(b) s can be extended to $(s \diamond_s \alpha)$;
6: **if** no valid α available **then**
7: **return**;
8: **for each** valid α **do**
9: Call CloSpan($s \diamond_i \alpha$, $D_{s \diamond_i \alpha}$, *min_sup*, L);
10: **for each** valid α **do**
11: Call CloSpan($s \diamond_s \alpha$, $D_{s \diamond_s \alpha}$, *min_sup*, L);
12: **return**;

ภาพที่ 2.11 อัลกอริทึม CloSpan

จากการทดสอบประสิทธิภาพในการไมน์นิ่ง เปรียบเทียบระหว่างอัลกอริทึม PrefixSpan และ CloSpan นั้น CloSpan ใช้เวลาได้น้อยกว่า ดังภาพที่ 2.12



ภาพที่ 2.12 เปรียบเทียบเวลาที่ใช้ในการประมวลผลของอัลกอริทึม PrefixSpan และ CloSpan

บทที่ 3
เครื่องมือและโปรแกรมที่ใช้ในการพัฒนาระบบ

ในบทนี้จะกล่าวถึงเครื่องมือและโปรแกรมที่ใช้ในการพัฒนาระบบ ซึ่งแบ่งออกเป็น 3 ส่วนหลักๆ คือ ภาษาที่ใช้ในการพัฒนาโปรแกรม(MS Visual Basic 2005) ระบบฐานข้อมูล(MS SQL Server) และเครื่องมือที่ใช้ในการทำดาต้าไมน์นิ่ง (Illimine)

3.1 MS Visual Basic 2005

Microsoft Visual Basic 2005 (VB 2005)คือ ภาษาเขียนโปรแกรมซึ่งเป็นหนึ่งในภาษาของ Microsoft Visual Studio 2005 ซึ่งได้ปรับปรุงและพัฒนาให้มีแนวคิดสอดคล้องกับหลักการเขียนโปรแกรมแบบ OOP (Object Oriented Programming)

คุณสมบัติของภาษา VB 2005 ที่ถือว่าเป็น OOP คือ

1) Encapsulation เป็นคุณสมบัติที่ผู้เขียนโปรแกรมไม่ต้องสนใจรายละเอียดที่ไม่จำเป็น เช่น การเขียนเมนูนั้น VB2005 จะเป็นผู้จัดการเมนู เช่น การซ่อน การแสดง หรือการใส่รูปภาพนั้นทำอย่างไร เพียงแค่กำหนดคุณสมบัติและใช้งานตามความสามารถเท่านั้น

2) Inheritance เป็นคุณสมบัติที่คลาสต้องสามารถสืบทอดได้ เช่นเดียวกับการกำหนดคอนโทรลของ VB 2005 เป็นออบเจ็กต์ที่สืบทอดได้โดยดีไรฟคลาส รวมทั้งยังสามารถเพิ่มเติมและปรับเปลี่ยนพร็อพเพอร์ตี้ หรือ เมทธอดได้

3) Polymorphism เป็นคุณสมบัติที่คลาสจะต้องเปลี่ยนแปลงความสามารถให้เข้ากับสภาพแวดล้อมได้ เช่น การสร้างคลาสชื่อ Shape ซึ่งจะให้สร้างออบเจ็กต์เป็นรูปต่างๆ เช่น วงกลม สามเหลี่ยม สี่เหลี่ยม และมีเมทธอด Area สำหรับคำนวนพื้นที่ของรูปทรง ซึ่งเมทธอด Area ก็จะต้องมีวิธีการคำนวนที่แตกต่างกันตามการกำหนดพร็อพเพอร์ตี้ชนิดของรูปทรงของคลาสว่าเป็นรูปทรงอะไร

จุดเด่นของ VB 2005 คือ เป็นภาษาที่เหมาะสำหรับการพัฒนาระบบขึ้นเพื่อใช้ในเชิงธุรกิจที่มียังมีการเปลี่ยนแปลงหรือเพิ่มเติมเงื่อนไขในอนาคต เนื่องจาก VB 2005 ได้ออกแบบเครื่องมือหรือคอลโทรลต่างๆ มาให้ ทำให้ประหยัดเวลาในการพัฒนาโปรแกรมหรือปรับปรุงแก้ไขเงื่อนไขการทำงานของโปรแกรม

3.2 MS SQL Server

Microsoft SQL Server คือ ฐานข้อมูลเชิงสัมพันธ์(RDBMS) ของบริษัทไมโครซอฟ ที่พัฒนาความสามารถให้สูงขึ้นเรื่อยๆ และได้รับความนิยมอย่างแพร่หลายในวงการธุรกิจ ซึ่ง MS

SQL Server นั้น รองรับกับภาษาที่เขียนโปรแกรมได้อย่างหลากหลาย เพราะการเชื่อมต่อกับฐานข้อมูลนั้นจะมีตัวกลางในการจัดการให้ หรือที่เรียกว่า ADO หรือ ADODB โดยเฉพาะถ้าภาษาที่ใช้นั้นเป็นของบริษัทไมโครซอฟเอง เช่น MS Visual Basic หรือ ASP.Net

ทั้งนี้ MS SQL Server ยังรองรับการนำข้อมูลเข้ามา และการนำข้อมูลออกไปยังแหล่งข้อมูลหรือฐานข้อมูลอื่นๆ โดยใช้ความสามารถของ Data Transformation Service Import/Export Wizard ทำให้ระบบที่พัฒนาขึ้นเปิดกว้างให้ผู้ใช้สามารถขุดค้นข้อมูลจากแหล่งข้อมูลใดๆ ได้

3.3 ILLIMINE

ILLIMINE คือ โปรแกรมที่ใช้ในการทำดาต้าไมน์นิ่ง พัฒนาขึ้นโดย Department of Computer Science ของ University of Illinois at Urbana-Champaign ในส่วนของ Sequential Pattern Mining นั้น Illimine ได้นำเสนออัลกอริทึมที่ใช้ในการวิจัย 3 แบบด้วยกัน คือ Prefix Span, CloSpan และ IncSpan ซึ่งได้เผยแพร่ผลงานวิจัย และโปรแกรมต่างๆ ให้ผู้ที่สนใจสามารถเข้าไปดูได้ที่ http://illimine.cs.uiuc.edu

ในการพัฒนาระบบงานนี้ ได้เลือก CloSpan มาใช้เป็นเครื่องมือในการทำดาต้าไมน์นิ่ง เนื่องจาก CloSpan มีประสิทธิภาพสูงกว่า ใช้เวลาน้อยกว่า และสิ้นเปลืองทรัพยากรน้อยกว่า นอกจากนี้ CloSpan ยัง Open Source อีกด้วย

บทที่ 4
การวิเคราะห์และออกแบบระบบ

ในบทนี้จะกล่าวถึงการวิเคราะห์และออกแบบระบบ เพื่อศึกษาความต้องการของระบบให้ทำงานได้ตามเป้าหมายที่วางไว้ ซึ่งจะวิเคราะห์และออกแบบระบบเชิงวัตถุ

4.1 ขอบเขตของระบบ

แนวคิดในการออกแบบระบบที่จะพัฒนาขึ้น จะแบ่งฟังก์ชันการทำงานของระบบออกเป็นแต่ละขั้นตอนย่อยๆ ตามกระบวนการทำดาต้าไมน์นิ่ง ดังที่กล่าวไว้ในบทที่ 2 เพื่อให้เห็นภาพการทำงานของระบบชัดเจนและง่ายต่อการพัฒนา จึงแบ่งออกเป็น

4.1.1 การเลือกข้อมูลที่มาทำดาต้าไมน์นิ่ง ซึ่งควรจะเปิดกว้างให้ผู้ใช้สามารถเลือกข้อมูลจากตารางและคอลัมน์ใดๆ ก็ได้ในฐานข้อมูล แต่จะต้องครบถ้วนตามตัวแปรสำคัญที่จะต้องใช้ใน Sequential Pattern Mining และในเมื่อระบบเปิดกว้างให้ผู้ใช้สามารถเลือกแหล่งของข้อมูลได้เอง จึงจำเป็นต้องให้ผู้ใช้ระบุความสัมพันธ์จากตารางที่ได้เลือกเอาไว้ เพื่อให้ระบบสามารถสร้างคำสั่ง SQL เพื่อ Query ข้อมูลได้ถูกต้อง

4.1.2 เรียกดูข้อมูลที่จะไมน์นิ่ง เมื่อได้ข้อมูลตามต้องการแล้ว ก็จะต้องทำความสะอาดข้อมูลให้ข้อมูลสิ่งเดียวกัน มีคำอธิบายเพียงอย่างเดียว และสำหรับข้อมูลที่ไม่มีความต่อเนื่องหรือข้อมูลที่ไม่ปกติจะต้องตัดทิ้งไป

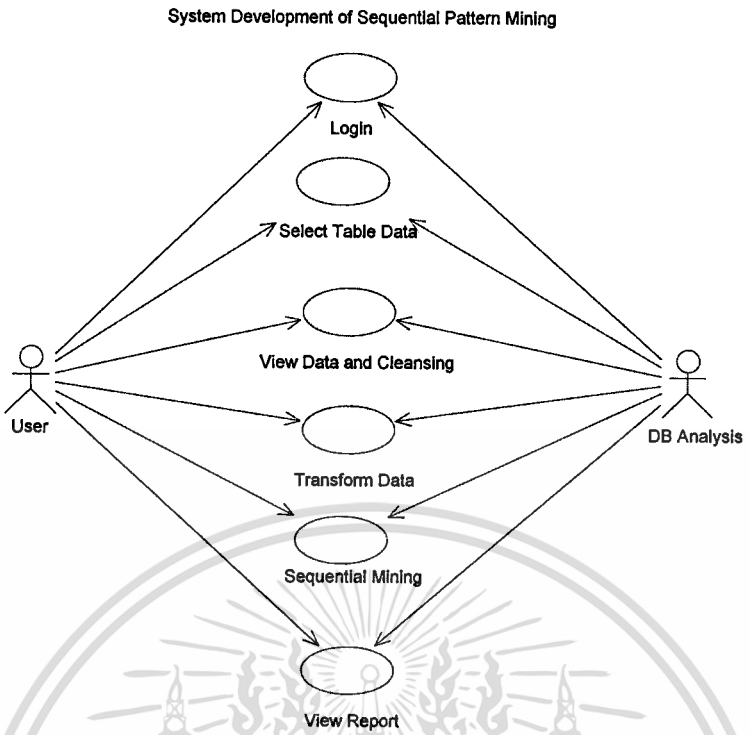
4.1.3 นำข้อมูลมาเตรียมสำหรับการไมน์นิ่ง ขั้นตอนนี้จะเป็นการนำข้อมูลที่ได้เลือกเอาไว้มาเตรียมเพื่อให้อยู่ในรูปแบบที่จะใช้ในการไมน์นิ่ง

4.1.4 ดาต้าไมน์นิ่ง คือ ขั้นตอนการทำดาต้าไมน์นิ่งและอ่านผลลัพธ์ที่ได้จากการไมน์นิ่งมาเก็บไว้ในฐานข้อมูล

4.1.5 รายงานผลลัพธ์ เป็นการแสดงรายงานผลที่ได้จากการไมน์นิ่ง ซึ่งควรจะตัดเอาเฉพาะข้อมูลที่น่าสนใจมาแสดงในรายงาน

4.2 ยูสเคสไดอะแกรม

จากการศึกษาการทำงานของระบบแล้ว อธิบายด้วยยูสเคสไดอะแกรม (Use Case Diagram) ซึ่งเป็นไดอะแกรมที่ใช้ค้นหาความต้องการของระบบ และกำหนดขอบเขตของระบบที่จะพัฒนาขึ้นได้ดังภาพที่ 4.1 ดังนี้



ภาพที่ 4.1 ยูสเคสไดอะแกรม

ยูสเคสไดอะแกรมของการพัฒนาระบบดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining ประกอบด้วย 2 แอ็กเตอร์ และ 6 ยูสเคส ซึ่งอธิบายได้ดังนี้

แอ็กเตอร์เป็นการแสดงถึงบุคคลที่เกี่ยวข้องกับระบบ มี 2 แอ็กเตอร์ดังนี้

1. User คือ ผู้ใช้งานระบบ
2. DB Analysis คือ ผู้ดูแลและวิเคราะห์ดาต้าเบส

ยูสเคสอธิบายฟังก์ชันหลักของระบบ ซึ่งมี 6 ยูสเคส ดังนี้

1. Login เป็นฟังก์ชันที่ตรวจสอบสิทธิ์ในการเข้าใช้ระบบ
2. Select Table Data เป็นฟังก์ชันที่ให้ผู้ใช้ระบบเลือกตารางข้อมูลที่จะนำมาใช้ในการทำดาต้าไมน์นิ่ง
3. View Data and Cleansing เป็นฟังก์ชันที่แสดงข้อมูลทั้งหมดที่ได้เลือกไว้ เพื่อให้ผู้ใช้ทำการตรวจสอบข้อมูลก่อนจะทำการไมน์นิ่ง
4. Transform Data เป็นฟังก์ชันที่จะเตรียมข้อมูลให้อยู่ในรูปแบบที่ระบบต้องการนำไปใช้ในการไมน์นิ่ง
5. Sequential Mining เป็นฟังก์ชันที่เรียกใช้ Illimine เพื่อทำการไมน์นิ่ง และอ่านที่ได้เข้าไปเก็บในฐานข้อมูล
6. View Report เป็นฟังก์ชันที่เรียกดูรายงานผลจากการไมน์นิ่งซึ่งจะแสดงผลเฉพาะข้อมูลที่น่าสนใจเท่านั้น

4.3 ยูสเคสดิสคริพชัน

ยูสเคสดิสคริพชันเป็นส่วนที่จะอธิบายรายละเอียดของการทำงานของแต่ละฟังก์ชัน จากยูสเคสไดอะแกรม สามารถอธิบายรายละเอียดตามตารางที่ 4.1 – 4.6 ได้ดังนี้

ตารางที่ 4.1 ยูสเคสดิสคริพชัน Login

Use case Name: Login	ID: 1	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า วิธีการที่จะเข้าใช้งานระบบจะต้องมีการ Login ป้อนข้อมูล Server Name, Database Name, Username และ Password ทุกครั้งที่เข้าใช้งาน		
Trigger: User ต้องติดต่อผู้ดูแลระบบดาต้าเบส เพื่อขอข้อมูล Server Name, Database Name, Username และ Password Type: External		
Relationship: Association: User, DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ผู้ที่จะเข้ามาใช้ระบบ จะต้องถูกกำหนดสิทธิให้เข้าใช้งานได้จากผู้ดูแลระบบดาต้าเบส โดยจะต้องนำข้อมูล Server Name, Database Name, Username และ Password มาใช้ในการ login มีขั้นตอนดังนี้ 1. ผู้ใช้ระบบป้อนชื่อ Server Name, Database Name, User Name และ Password 2. ผู้ใช้กดปุ่ม OK 3. ระบบนำข้อมูลทั้งหมด ไปตรวจสอบสิทธิในการเข้าในงาน 4. ผู้ใช้เข้าไปใช้งานระบบ		
Subflows:		
Alternate/exceptional flows: 1.1 ถ้าข้อมูลไม่ถูกต้อง หรือไม่พบสิทธิก็ไม่อนุญาตเข้าใช้ระบบได้ ผู้ใช้จะต้องไปติดต่อผู้ดูแลระบบฐานข้อมูลก่อน		

ตารางที่ 4.2 ยูสเคสดิสคริพชัน Select Table Data

Use case Name: Select Table Data	ID: 2	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า ข้อมูลที่จะเลือกเข้ามาใช้ในการทำดาต้าไมน์นิ่งนั้น ผู้ใช้งานจะเป็นผู้เลือกโดยระบุชื่อของตารางและชื่อคอลัมน์ของข้อมูล		
Trigger: Type: External		
Relationship: Association: User, DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ข้อมูลที่จะนำไปใช้ในการทำดาต้าไมน์นิ่งนั้น ผู้ใช้งานจะเป็นผู้ระบุเองจากระบบฐานข้อมูล MS SQL Server มีขั้นตอนดังนี้ 1. ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ของข้อมูล รหัสสมาชิก 2. ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ของข้อมูล วันที่สั่งซื้อ 3. ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ของข้อมูล เลขที่ใบสั่งซื้อ 4. ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ของข้อมูล รหัสสินค้า 5. ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ของข้อมูล ชื่อสินค้า 6. ผู้ใช้ระบุความสัมพันธ์จากตารางที่ระบุไว้ข้างต้น 7. ผู้ใช้กดปุ่ม สิ้นสุดการระบุเงื่อนไข 8. ระบบจะทำการสร้างคำสั่ง SQL Statement จากเงื่อนไขที่ผู้ใช้กำหนดไว้ 9. กดปุ่มเพื่อทดสอบคำสั่ง SQL ระบบจะทำการดึงตัวอย่างของข้อมูลขึ้นมา 10. ผู้ใช้กดปุ่มยืนยัน ระบบจะนำข้อมูลทั้งหมดมาเก็บไว้ในตารางข้อมูลสำหรับการทำดาต้าไมน์นิ่ง		
Subflows:		
Alternate/exceptional flows:		

ตารางที่ 4.3 ยูสเคสดิสคริพชัน View Data and Cleansing

Use case Name: View Data and Cleansing	ID: 3	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า ระบบจะแสดงข้อมูลให้ผู้ใช้ได้ตรวจสอบความถูกต้องก่อนจะนำไปทำดาต้าไมน์นิ่ง และยังเพิ่มฟิลเตอร์ให้ผู้ใช้เลือกข้อมูลจากช่วงเวลาที่กำหนดได้ หรือเลือกเฉพาะสินค้าที่สนใจ		
Trigger: Type: External		
Relationship: Association: User, DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ระบบจะแสดงข้อมูลให้ผู้ใช้ได้ตรวจสอบความถูกต้องก่อนจะนำไปทำดาต้าไมน์นิ่ง และยังเพิ่มฟิลเตอร์ให้ผู้ใช้เลือกข้อมูลจากช่วงเวลาที่กำหนดได้ หรือเลือกเฉพาะสินค้าที่สนใจ มีขั้นตอนดังนี้ 1. ผู้ใช้ระบุช่วงเวลาที่ต้องการนำข้อมูลมาทำดาต้าไมน์นิ่ง 2. ผู้ใช้กดปุ่ม View Data 3. ระบบแสดงข้อมูลทั้งหมดตามช่วงเวลาที่เลือกให้ผู้ใช้ได้ตรวจสอบ 4. ระบบแสดงลิสต์ของสมาชิกโดยเรียงลำดับจากจำนวนใบสั่งซื้อมากที่สุด 5. ระบบแสดงลิสต์ของสินค้าทั้งหมด 6. ผู้ใช้เลือกรหัสสมาชิกและรายชื่อสินค้าที่ต้องการนำมาทำดาต้าไมน์นิ่ง 7. ผู้ใช้กดปุ่ม สิ้นสุดการตรวจสอบข้อมูล		
Subflows:		
Alternate/exceptional flows:		

ตารางที่ 4.4 ยูสเคสดิสคริพชัน Transform Data

Use case Name: Transform Data	ID: 4	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า ระบบจะทำการแปลงข้อมูลให้อยู่ในรูปแบบที่จะนำไปใช้ในการทำดาต้าไมน์นิ่ง		
Trigger: Type: External		
Relationship: Association: User, DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ระบบจะแสดงสถานะการทำงานของโปรแกรมให้ผู้ใช้ได้ทราบจำนวนข้อมูล และลำดับของข้อมูลที่ได้ถูกแปลงเรียบร้อยแล้ว มีขั้นตอนดังนี้ 1. ระบบทำการคำนวนจำนวนสมาชิก, จำนวนข้อมูล ค่าเฉลี่ยของข้อมูล 2. ผู้ใช้กดปุ่มเพื่อสร้าง Text File 3. ระบบแสดงสถานะการทำงานเป็น Progress Bar เพื่ออ่านข้อมูลแต่ละเรคคอร์ดมาเขียนข้อมูลลงไฟล์ โดยจะมีสัญลักษณ์เพิ่มเติม เพื่อคั่นข้อมูลใบสั่งซื้อและรหัสสมาชิก 4. เมื่อระบบทำงานสำเร็จจะขึ้นข้อความ "การทำงานเสร็จสมบูรณ์" 5. ผู้ใช้กดปุ่มเพื่อแปลง Text File เป็น Binary File 6. ระบบจะแปลงข้อมูลเป็น Binary File		
Subflows:		
Alternate/exceptional flows:		

ตารางที่ 4.5 ยูสเคสดิสคริพชัน Sequential Mining

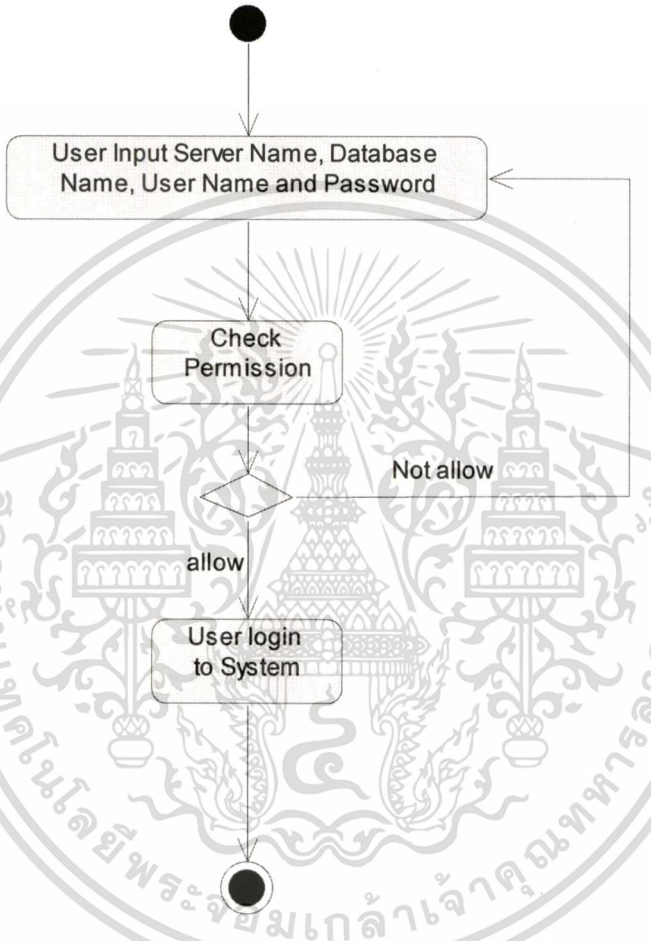
Use case Name: Sequential Mining	ID: 5	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า ระบบจะส่งข้อมูลออกไปให้ CloSpan ทำดาต้าไมน์นิ่ง แล้วระบบจะอ่านผลลัพธ์จากการไมน์นิ่งมาเก็บไว้ในฐานข้อมูล		
Trigger: Type: External		
Relationship: Association: User, DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ระบบจะทำการโหลดโปรแกรม CloSpan ขึ้นมาเพื่อใช้ในการทำดาต้าไมน์นิ่ง มีขั้นตอนดังนี้ 1. ระบบให้ผู้ใช้ระบุค่า Support 2. ผู้ใช้กดปุ่ม Sequential Mining 3. ระบบจะทำการทำดาต้าไมน์นิ่ง 4. เมื่อระบบทำงานเสร็จแล้ว จะขึ้นกล่องข้อความแจ้งผู้ใช้ระบบ 5. ระบบจะอ่านข้อมูลผลลัพธ์ที่ได้จากการไมน์นิ่งเข้าไปเก็บไว้ในฐานข้อมูล 6. ระบบขึ้นกล่องข้อความแจ้งให้ผู้ใช้ทราบว่าอ่านข้อมูลเสร็จแล้ว		
Subflows:		
Alternate/exceptional flows:		

ตารางที่ 4.6 ยูสเคสดิสคริพชัน View Report

Use case Name: View Report	ID: 4	Importance level: High
Primary actor: User, DB Analysis	User case type: Detail, essential	
Stakeholders and interest: User – ผู้ใช้งานระบบ DB Analysis – ผู้ดูแลและวิเคราะห์ดาต้าเบส		
Brief description: use case นี้ใช้เพื่ออธิบายว่า ระบบจะนำเฉพาะข้อมูลที่น่าสนใจ ที่ได้มาจากการทำดาต้าไมน์นิ่งมาแสดงในรูปแบบที่อ่านง่าย		
Trigger: Type: External		
Relationship: Association: User DB Analysis Include: - Extend: - Generalization: -		
Normal flow of events: ระบบจะนำข้อมูลที่ได้มาจากการทำดาต้าไมน์นิ่งมาวิเคราะห์ผล มีขั้นตอนดังนี้ 1. ผู้ใช้กดปุ่ม View Report 2. ระบบแสดงรายงานจากการวิเคราะห์ข้อมูล		
Subflows:		
Alternate/exceptional flows:		

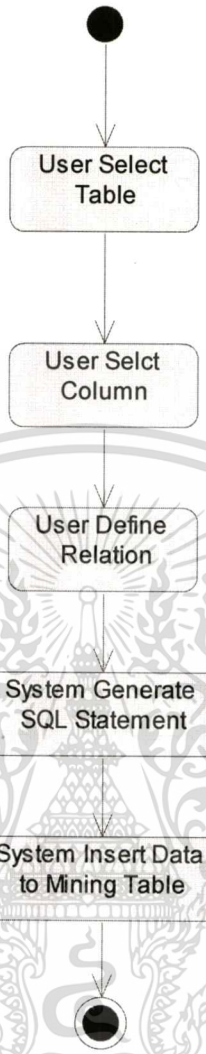
4.4 แอกทิวิตีไดอะแกรม

แอกทิวิตีไดอะแกรมใช้อธิบายรายละเอียดขั้นตอนการทำงานของแต่ละยูสเคส ซึ่งอธิบายรายละเอียดตามภาพที่ 4.2 – 4.7 ได้ดังนี้



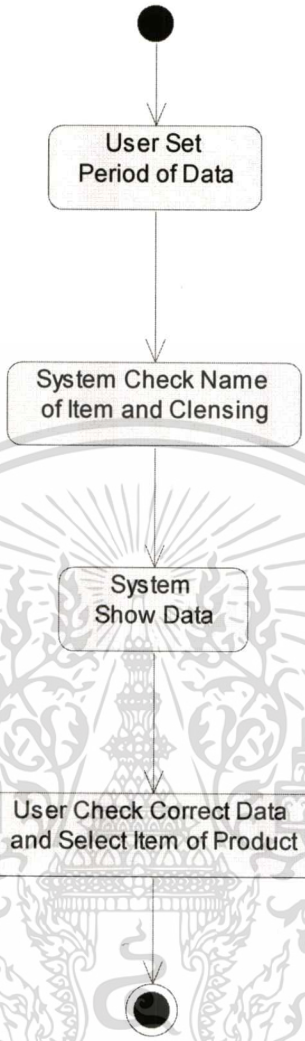
ภาพที่ 4.2 แอกทิวิตีไดอะแกรมของยูสเคส Login

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส Login มีขั้นตอนดังนี้ ผู้ใช้ใส่ข้อมูลที่จะใช้ล็อกอินเข้าระบบ ซึ่งจะเป็นสิทธิของการเข้าถึงระบบฐานข้อมูล ได้แก่ ชื่อเซิร์ฟเวอร์ ชื่อดาต้าเบส ชื่อผู้ใช้ และรหัสผ่าน ระบบจะนำข้อมูลไปตรวจสอบสิทธิว่าสามารถใช้ล็อกอินเข้าใช้ระบบฐานข้อมูลได้หรือไม่ ถ้าเข้าไม่ได้ ระบบจะให้ผู้ใช้กลับไปกรอกข้อมูลใหม่ แต่ถ้าเช็คสิทธิแล้วผ่าน ระบบจะให้ผู้ใช้เข้าไปทำงานต่อไป สำหรับการล็อกอินไม่ผ่าน ระบบจะแนะนำให้ผู้ใช้ติอต่อกับผู้ดูแลระลฐานข้อมูล



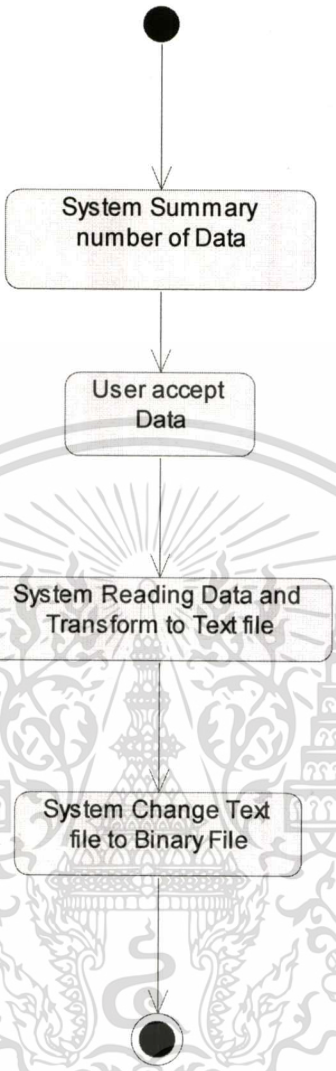
ภาพที่ 4.3 แอกทิวิตีไดอะแกรมของยูสเคส Select Table Data

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส Select Table Data มีขั้นตอนดังนี้ ผู้ใช้ระบุชื่อตารางและชื่อคอลัมน์ ในฐานข้อมูล ตามตัวแปรที่ระบบแจ้งไว้ หลังจากนั้น ผู้ใช้จะต้องกำหนดความสัมพันธ์ของตารางที่ได้เลือกไว้ข้างต้น ว่ามีความสัมพันธ์กันที่คอลัมน์ใด เมื่อกดปุ่มยืนยัน ระบบจะสร้างคำสั่ง SQL ให้ผู้ใช้ทดลองทดสอบคำสั่ง SQL ว่าสามารถไป Query ข้อมูลจากฐานข้อมูลออกมาให้ถูกต้องหรือไม่ ถ้าข้อมูลที่ได้ทดลอง Query ออกมาถูกต้องแล้ว ผู้ใช้กดปุ่มยืนยันและเริ่มบันทึกข้อมูล ระบบจะทำการอ่านข้อมูลจากตารางหลัก มาเก็บไว้ในตารางข้อมูลที่จะใช้ในการทำดาต้าไมน์นิ่ง เพื่อไม่ให้กระทบกับข้อมูลหลัก และเพื่อให้การประมวลผลมีความรวดเร็วอีกด้วย



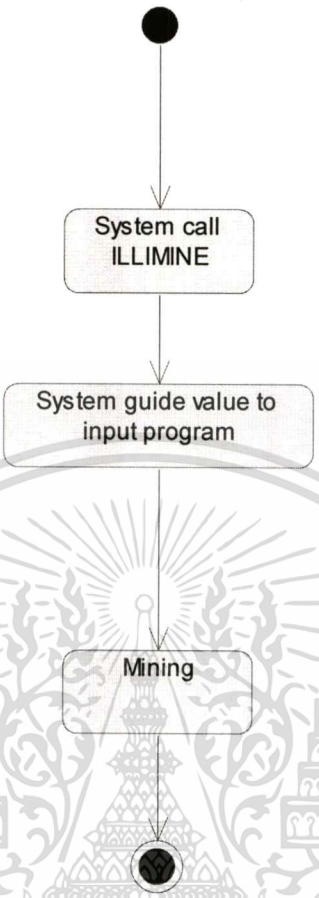
ภาพที่ 4.4 แอกทิวิตีไดอะแกรมของยูสเคส View Data and Cleansing

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส View Data and Cleansing มีขั้นตอนดังนี้ ระบบให้ผู้ใช้กำหนดช่วงเวลาของข้อมูลที่จะนำมาใช้ในการไมน์นิ่ง เพื่อจำกัดข้อมูลให้เล็กลงในกรณีที่มีจำนวนข้อมูลปริมาณมาก หรือผู้ใช้อาจต้องการไมน์นิ่งในช่วงเวลาสั้นๆ หลังนั้นระบบจะไปดึงข้อมูลจากช่วงเวลาดังกล่าว แล้วทำการนับจำนวนใบสั่งซื้อของสมาชิกแต่ละคนว่ามีการซื้อสินค้าต่อเนื่องหรือไม่ สำหรับสมาชิกคนที่ไม่มีการซื้อต่อเนื่อง ระบบจะใส่สถานะไว้ว่าไม่เลือกมาใช้ในการไมน์นิ่ง แต่ถ้าสมาชิกมีการซื้อต่อเนื่องจะใส่สถานะไว้ว่าเลือกมาทำไมน์นิ่ง และทำความสะอาดข้อมูลของสินค้า ในกรณีที่สินค้ารหัสหนึ่งมีชื่อมากกว่า 1 ชื่อ จะแก้ไขให้มีเพียงชื่อเดียว แล้วระบบจะแสดงข้อมูลออกมาให้ผู้ใช้สามารถเลือกได้ตามความต้องการของผู้ใช้อีกครั้งก่อนจะนำไปสู่ขั้นตอนการเตรียมข้อมูล



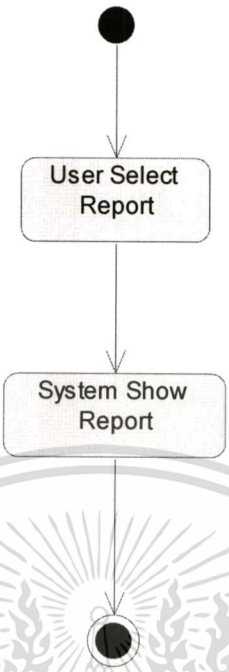
ภาพที่ 4.5 แอกทิวิตีไดอะแกรมของยูสเคส Transform Data

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส Transform Data มีขั้นตอนดังนี้ ระบบจะทำการนับจำนวนข้อมูล แล้วเริ่มวนลูปอ่านข้อมูลโดยเรียงลำดับจากรหัสสมาชิก วันที่ใบสั่งซื้อ เลขที่ใบสั่งซื้อ แล้วนำมาเขียนลง Text File โดยจะมีตัวเลขคั่นระหว่าง ใบสั่งซื้อ และคั่นระหว่างสมาชิกแต่คนไว้ ระบบจะแสดงสถานะการทำงานเป็นแทบความสำเร็จของงานคิดเป็นเปอร์เซนต์เทียบกับจำนวนข้อมูลทั้งหมด เมื่อวนลูปอ่านข้อมูลทั้งหมดมาเขียนลง Text File เสร็จสิ้นแล้ว ระบบจะแจ้งผู้ใช้ เพื่อให้ผู้ใช้ได้ทำงานขั้นต่อไป คือ แปลง Text File เป็น Binary File เป็นอันเสร็จกระบวนการเตรียมข้อมูลสำหรับการทำดาต้าไมน์นิ่ง



ภาพที่ 4.6 แอกทิวิตีไดอะแกรมของยูสเคส Sequential Mining

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส Sequential Mining มีขั้นตอนดังนี้ ผู้ใช้ใส่ค่า Support ที่ต้องการสำหรับการทำดาต้าไมน์นิ่ง ระบบจะส่งค่า Support และ Binary File ไปให้โปรแกรม CloSpan ประมวลผล จากนั้นระบบจะแจ้งสถานะการทำงานให้ผู้ใช้ทราบว่ากำลังประมวลผลอยู่ ซึ่งอาจใช้เวลาพอสมควร ขึ้นอยู่จำนวนข้อมูลและค่า Support ที่ได้ระบุไป เมื่อทำดาต้าไมน์นิ่งเสร็จแล้ว จะได้ไฟล์ข้อมูลผลลัพธ์ของไมน์นิ่ง และระบบจะแจ้งสถานะว่าไมน์นิ่งเสร็จแล้ว เริ่มเข้าสู่ขั้นตอนการอ่านข้อมูล แล้วระบบก็จะเริ่มต้นอ่านข้อมูลมาทีละบรรทัด แล้วค่อยๆ ตัดคำ จากรูปแบบของเซ็ทข้อมูล ที่มีตัวอักษรคั่นระหว่างเซ็ทและค่า Support มาใส่ในฐานข้อมูล จนกระทั่ง End of File ซึ่งข้อมูลที่ได้อ่านมาเก็บไว้ในฐานข้อมูลนี้ จะนำมาใช้ในการวิเคราะห์และแสดงรายงานผลลัพธ์ของการไมน์นิ่งต่อไป ซึ่งถ้าผลลัพธ์ที่ได้ไม่ค่อยน่าพอใจ ผู้ใช้สามารถกลับไปเลือกข้อมูล หรือ เปลี่ยนค่า Support แล้วทำดาต้าไมน์นิ่งใหม่ เพื่อให้ได้ผลที่พึงพอใจ และเป็นประโยชน์สูงสุดได้



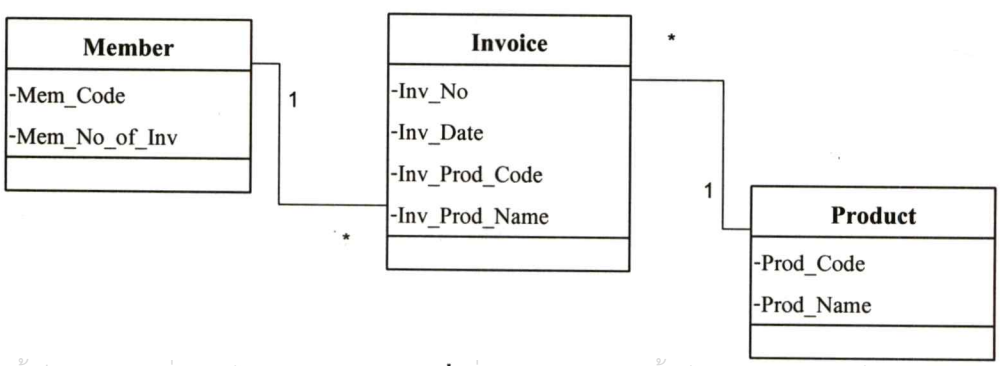
ภาพที่ 4.7 แอกทิวิตีไดอะแกรมของยูสเคส View Report

แอกทิวิตีไดอะแกรมอธิบายการทำงานของยูสเคส View Report มีขั้นตอนดังนี้ ผู้ใช้กดปุ่มแสดงรายงาน ระบบจะแสดงรายงานออกมา

4.5 คลาสไดอะแกรม

จากขั้นตอนการวิเคราะห์การทำงานของระบบ คลาสต่างๆ ที่จำเป็นต่อการพัฒนาระบบมีดังนี้

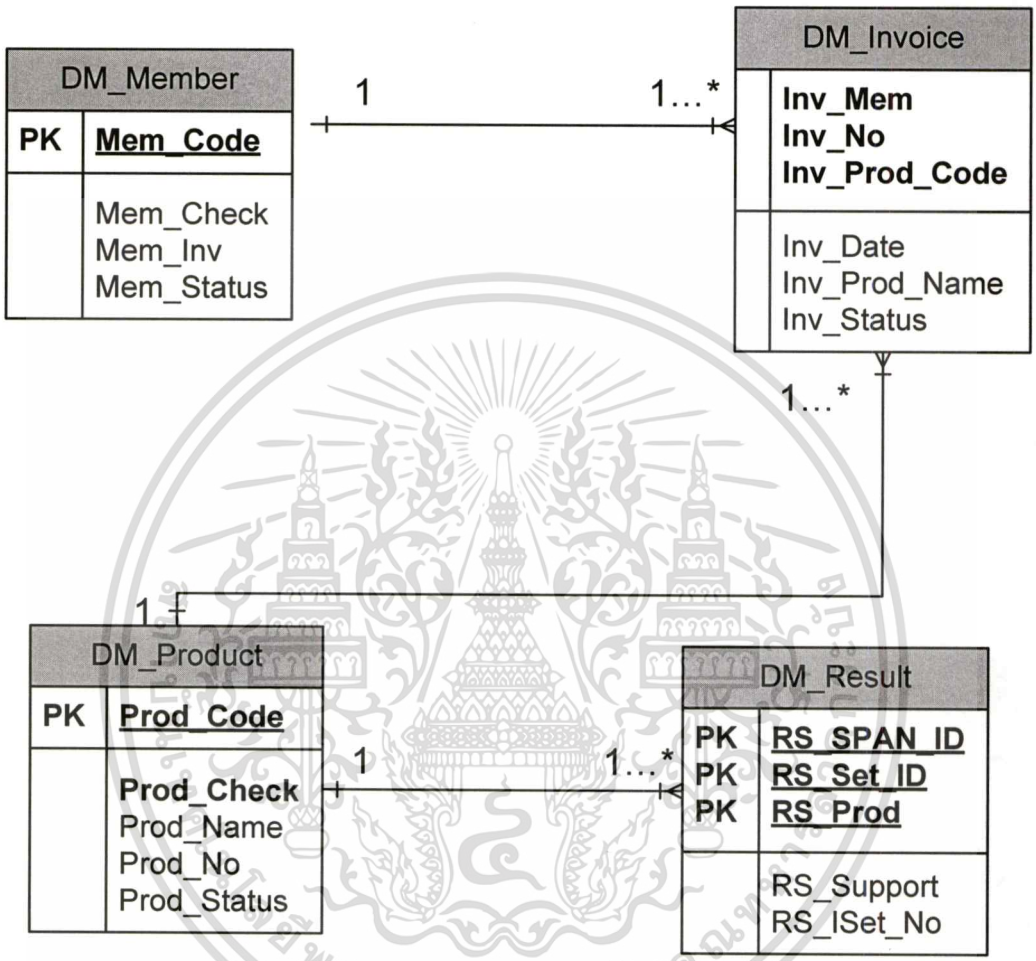
1. Member คือ คลาสสมาชิก
2. Invoice คือ คลาสใบสั่งซื้อสินค้า
3. Product คือ คลาสสินค้า



ภาพที่ 4.8 คลาสไดอะแกรม

4.6 การออกแบบจำลองข้อมูล

ข้อมูลที่จะนำมาใช้ในการทำดาต้าไมน์นิ่ง สามารถออกแบบจำลองของข้อมูลได้ดังนี้



ภาพที่ 4.9 แบบจำลองความสัมพันธ์ของข้อมูล ER Diagram

ตารางที่ 4.7 รายละเอียดของแต่ละเอนทิตี

Table Name	Table Description
DM_Member	เก็บข้อมูลสมาชิก
DM_Product	เก็บข้อมูลของสินค้า
DM_Invoice	เก็บข้อมูลการซื้อสินค้า
DM_Result	เก็บข้อมูลผลลัพธ์จากการทำดาต้าไมน์นิ่ง

4.7 พจนานุกรมข้อมูล

หลังจากที่ได้วิเคราะห์และออกแบบฐานข้อมูลแล้ว ข้อมูลที่จะทำการคัดลอกเพื่อนำไปทำดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern นั้น จะเลือกเฉพาะข้อมูลที่จำเป็นต่อระบบเท่านั้น อธิบายรายละเอียดของแต่ละเอนทิตีได้ดังนี้

ตารางที่ 4.8 พจนานุกรมข้อมูลของ DM_Member

ชื่อแอททริบิวต์	คำอธิบาย	ชนิดข้อมูล	คีย์	ตารางอ้างอิง
Mem_Check	เลือก/ ไม่เลือก	Varchar(3)		
Mem_Code	รหัสสมาชิก	char(13)	PK	DM_Invoice
Mem_Inv	จำนวนใบสั่งซื้อ	integer		
Mem_Status	สถานภาพของสมาชิก	tinyint		

ตารางที่ 4.9 พจนานุกรมข้อมูลของ DM_Product

ชื่อแอททริบิวต์	คำอธิบาย	ชนิดข้อมูล	คีย์	ตารางอ้างอิง
Prod_Check	เลือก/ ไม่เลือก	Varchar(3)		
Prod_Code	รหัสสินค้า	Varchar(10)	PK	DM_Invoice DM_Result
Prod_Name	ชื่อสินค้า	Varchar(100)		
Prod_No	รหัสสินค้าแปลงเป็นตัวเลข	integer		
Prod_Status	สถานะของสินค้า	Char(1)		

ตารางที่ 4.10 พจนานุกรมข้อมูลของ DM_Invoice

ชื่อแอททริบิวต์	คำอธิบาย	ชนิดข้อมูล	คีย์	ตารางอ้างอิง
Inv_Mem	รหัสสมาชิก	char(13)	FK	DM_Member
Inv_Date	วันที่ใบสั่งซื้อ	Datetime		
Inv_no	เลขที่ใบสั่งซื้อ	char(12)	PK	DM_Invoice
Inv_Prod_Code	รหัสสินค้า	varchar(10)	FK	DM_Product
Inv_Prod_Name	ชื่อสินค้า	varchar(100)		
Inv_Status	สถานะของใบสั่งซื้อ	tinyint		

ตารางที่ 4.11 พจนานุกรมข้อมูลของ DM_Result

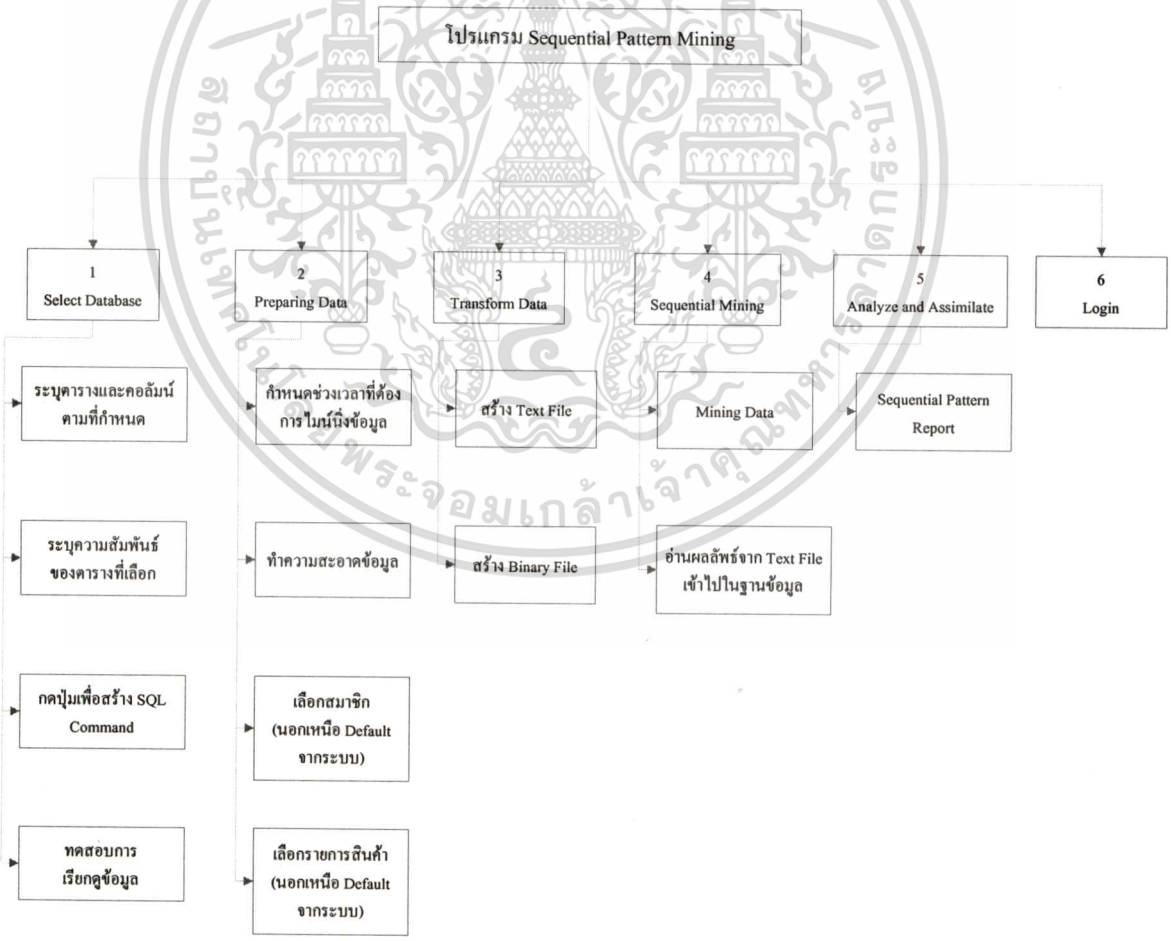
ชื่อแอททริบิวต์	คำอธิบาย	ชนิดข้อมูล	คีย์	ตารางอ้างอิง
RS_SPAN_ID	ลำดับของ Squential Pattern	Integer	PK	
RS_Set_ID	วันที่ใบสั่งซื้อ	Integer	PK	
RS_Prod	เลขที่ใบสั่งซื้อ	char(10)	FK	DM_Product
RS_Support	รหัสสินค้า	Integer	FK	
RS_ISet_No	ชื่อสินค้า	Integer		

บทที่ 5
การสร้างและทดสอบระบบ

ในการสร้างและทดสอบระบบของ การพัฒนาระบบดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining นั้น ได้นำข้อมูลการซื้อสินค้ามาเป็นข้อมูลในการศึกษา เพื่อหารูปแบบการเลือกซื้อสินค้าของลูกค้าที่เกิดขึ้นอย่างต่อเนื่อง ในระบบฐานข้อมูล MS SQL Server และนำผลลัพธ์มาแสดงในรูปแบบที่อ่านง่าย ซึ่งจะนำมาวางแผนทางธุรกิจต่อไป

5.1 รายละเอียดของระบบงาน

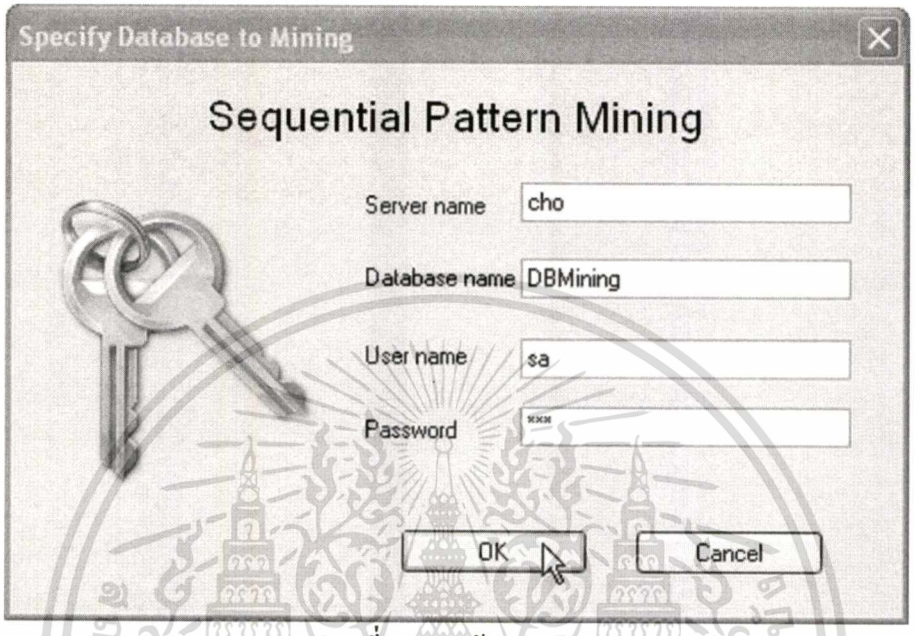
โครงสร้างของโปรแกรมที่ได้พัฒนาขึ้น มีรายละเอียด ดังนี้



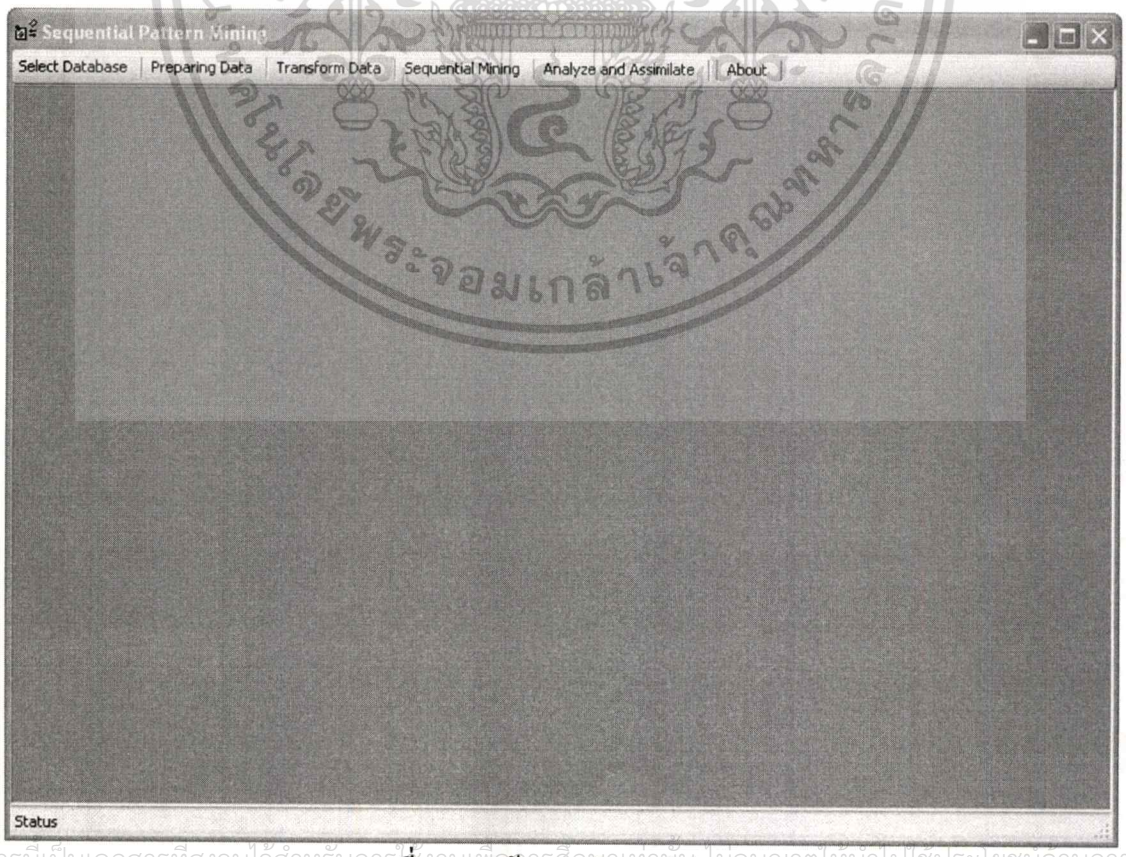
ภาพที่ 5.1 แสดงโครงสร้างของโปรแกรม

5.2 หน้าจอของระบบงาน

ก่อนผู้ใช้งานจะเข้าใช้งานระบบ จะต้องล็อกอินเพื่อตรวจสอบสิทธิในการเข้าถึงฐานข้อมูล ได้แก่ ชื่อเซิร์ฟเวอร์ (Server name), ชื่อฐานข้อมูล (Database name), ชื่อผู้ใช้งาน (User name) และรหัสผ่าน (Password)



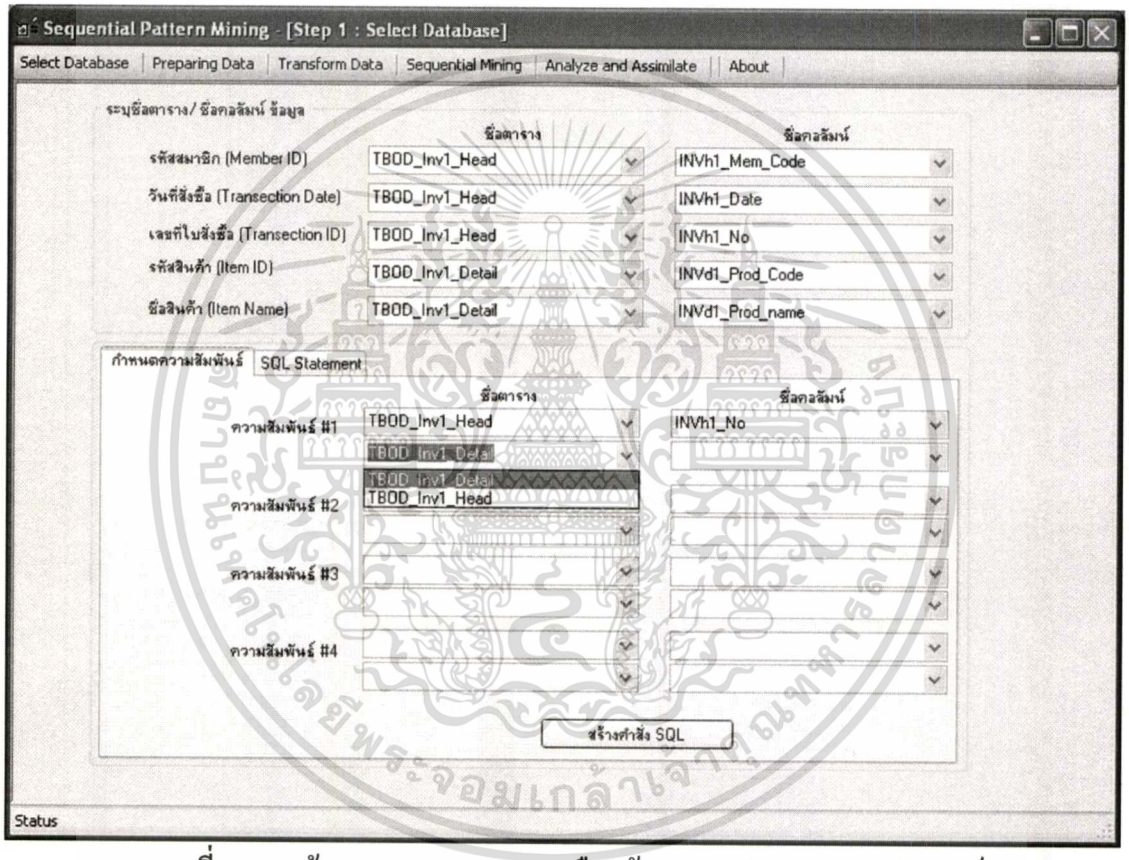
ภาพที่ 5.2 หน้าจอ Login



ภาพที่ 5.3 หน้าจอหลักของโปรแกรม

ในกรณีผู้ใช้ล็อกอินไม่ผ่าน ซึ่งอาจเกิดจากระบุข้อมูลข้างต้นไม่ถูกต้อง หรือไม่มีสิทธิในการเข้าถึงฐานข้อมูลนั้น ระบบจะแจ้งข้อความเตือนและไม่ให้ผ่านให้เข้าไปใช้งาน คำแนะนำเมื่อล็อกอินไม่ผ่าน คือ นำข้อมูลนี้ไปตรวจสอบกับผู้ดูแลระบบฐานข้อมูลว่าข้อมูลต่างๆ นั้นถูกต้อง

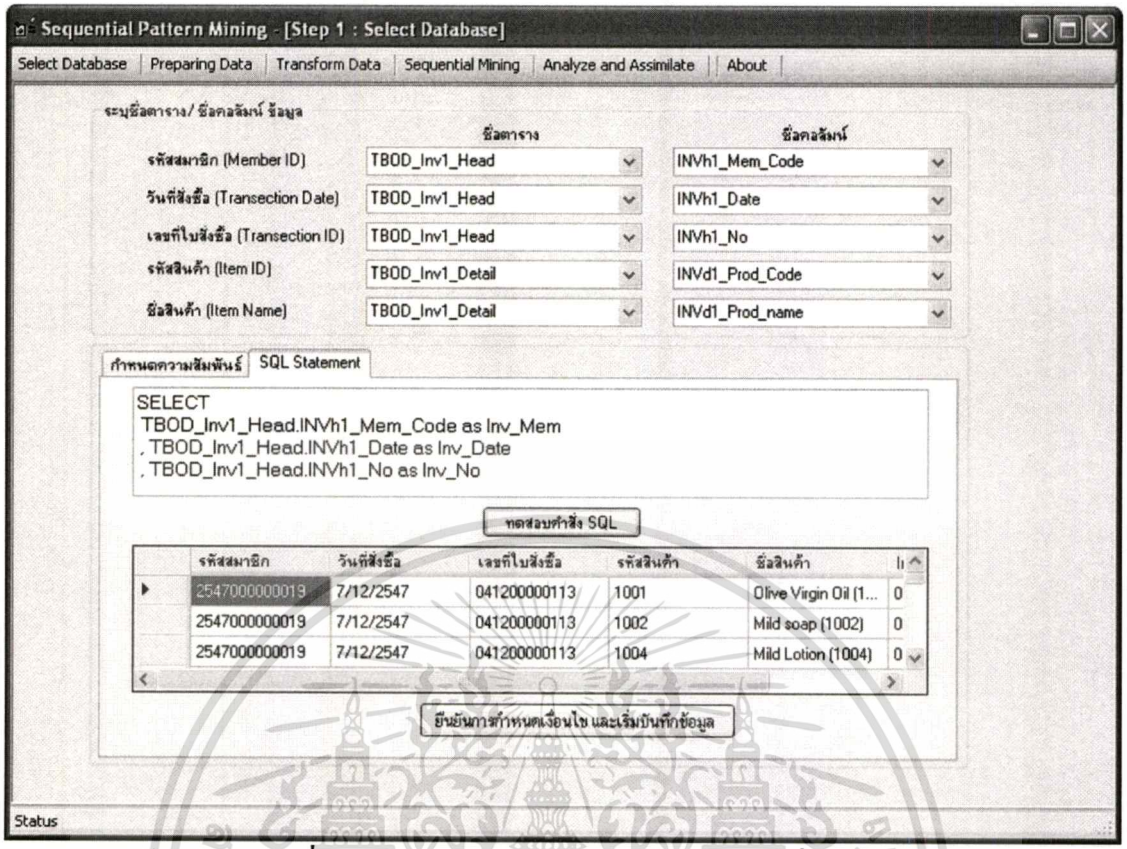
เมื่อผู้ใช้ล็อกอินผ่านเรียบร้อยแล้ว จะพบหน้าจอหลักของโปรแกรมตามภาพที่ 5.3 ซึ่งได้เรียงตามลำดับขั้นตอนของกระบวนการทำดาต้าไมน์นิ่ง คือ Select Database, Preparing Data, Transform Data, Sequential Mining และ Analyze and Assimilate ซึ่งจะอธิบายการทำงานของแต่ละหน้าจอตามลำดับต่อไป



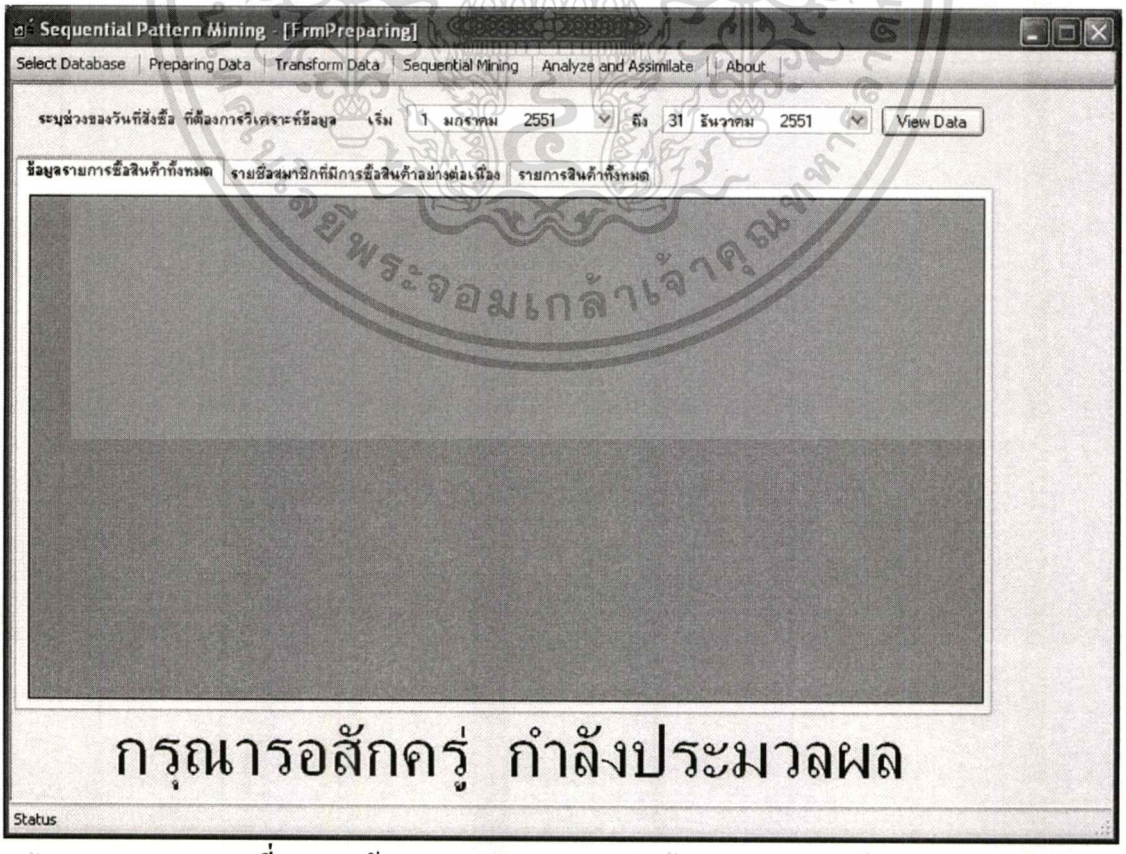
ภาพที่ 5.4 หน้าจอ Select Database เลือกข้อมูลและกำหนดความสัมพันธ์

จากภาพที่ 5.4 คือ หน้าจอ Select Database ระบบจะให้ผู้ใช้ได้ระบุชื่อตารางและชื่อคอลัมน์จากตัวแปรทั้ง 5 ตัวที่นำมาใช้ในการไมน์นิ่ง ได้แก่ รหัสสมาชิก, วันที่สั่งซื้อ, เลขที่ใบสั่งซื้อ, รหัสสินค้า และ ชื่อสินค้า เสร็จเรียบร้อยแล้วลำดับต่อไปจึงกำหนดความสัมพันธ์ของตารางว่ามีความสัมพันธ์กันที่คอลัมน์ใด เสร็จเรียบร้อยแล้วจึงกดปุ่ม “สร้างคำสั่ง SQL” ระบบจะนำข้อมูลที่ระบุไว้ข้างต้น มาแปลงเป็นคำสั่ง SQL ที่จะใช้ Query ข้อมูลจากระบบฐานข้อมูล ดังภาพที่ 5.5

เมื่อกดปุ่ม “ทดสอบคำสั่ง SQL” ระบบจะทดลอง Query ข้อมูลมาเป็นตัวอย่างจำนวน 50 เรคคอร์ด เมื่อตรวจสอบว่าถูกต้องแล้ว ให้กดปุ่ม “ยืนยันการกำหนดเงื่อนไขและเริ่มบันทึกข้อมูล” ระบบจะนำข้อมูลจากตารางหลักเข้าไปในตารางที่เตรียมไว้สำหรับทำดาต้าไมน์นิ่ง



ภาพที่ 5.5 หน้าจอ Select Database ทดสอบคำสั่ง SQL



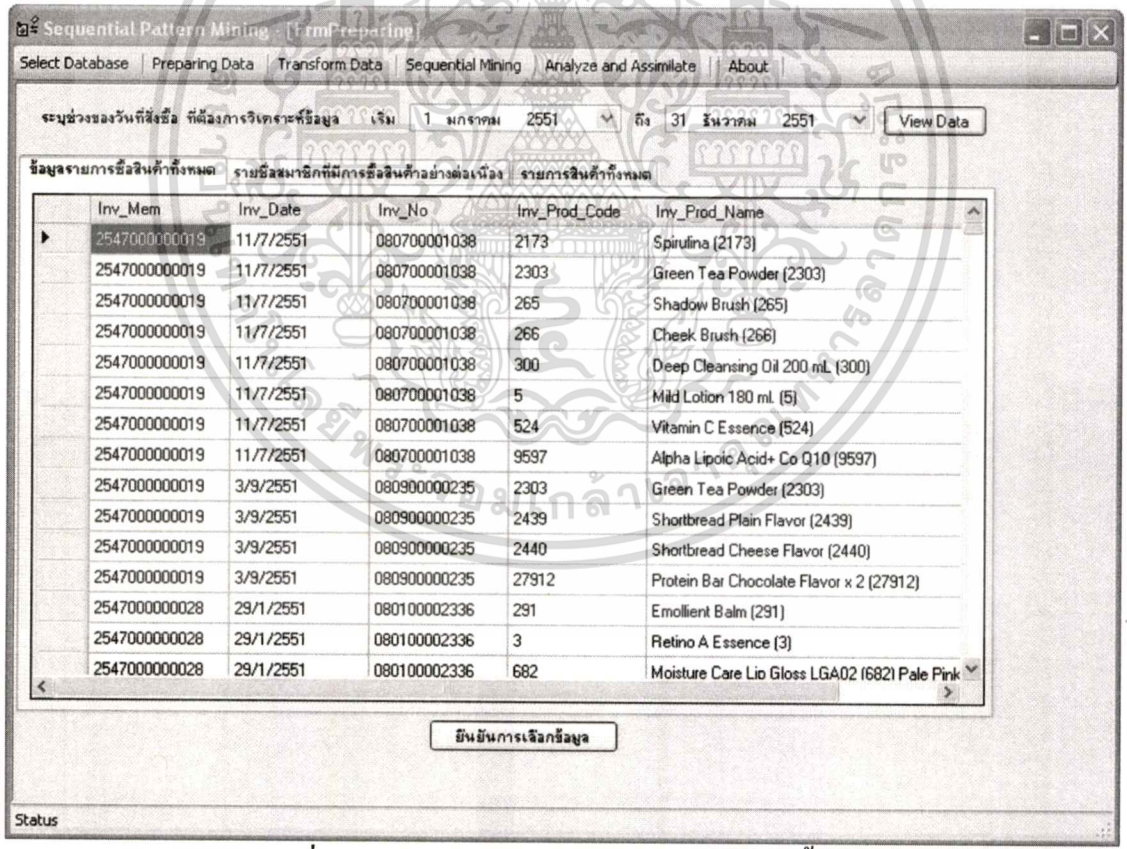
ภาพที่ 5.6 หน้าจอ Preparing Data กำหนดขอบเขตของข้อมูล

ภาพที่ 5.6 คือ หน้าจอ Preparing Data ซึ่งจะมีฟิลเตอร์ให้ผู้ใช้ระบุช่วงเวลาของใบสั่งซื้อเพื่อกำหนดขอบเขตของข้อมูลที่จะนำมาวิเคราะห์ ดังตัวอย่างเลือกข้อมูลวันที่ 1 มกราคม 2551 ถึงวันที่ 31 ธันวาคม 2551 เมื่อกดปุ่ม "View Data" ระบบจะไปดึงข้อมูลมาแสดง 3 ส่วน คือ

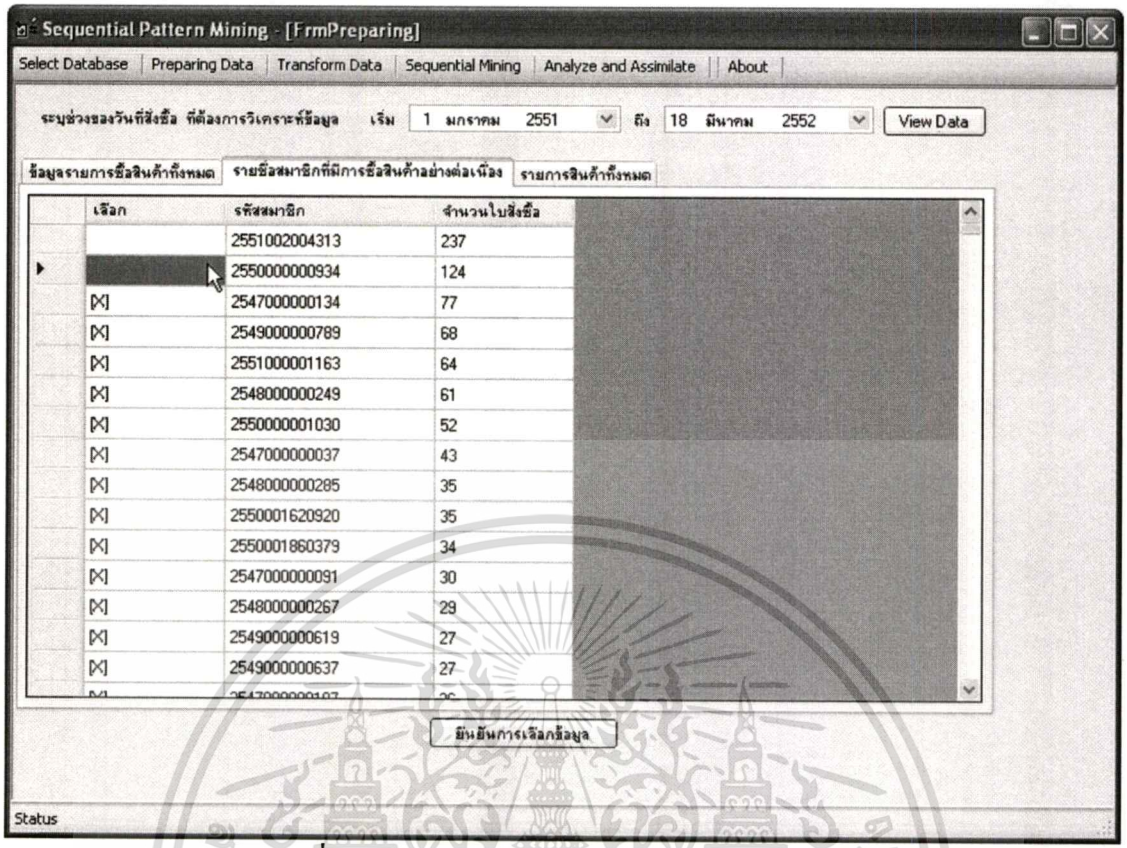
1) แท็บข้อมูลรายการซื้อสินค้าทั้งหมด จะแสดงข้อมูลทั้งหมดจากช่วงเวลาที่ได้ระบุไว้ ดังภาพที่ 5.7

2) แท็บรายชื่อสมาชิกที่มีการซื้อสินค้าอย่างต่อเนื่อง คือ ลิสต์ของสมาชิก ซึ่งจะเรียงลำดับจากผู้ที่มีจำนวนใบสั่งซื้อมากที่สุด ไปน้อยที่สุด ซึ่งในส่วนนี้ระบบจะเลือกเฉพาะผู้ที่มีใบสั่งซื้อมากกว่า 1 ใบมาใช้ในการวิเคราะห์ เพราะหมายถึงเป็นผู้ที่มีการซื้อสินค้าอย่างต่อเนื่องนั่นเอง แต่ผู้ใช้สามารถกำหนดได้ว่าจะ เลือก/ไม่เลือก สมาชิกคนใดได้ ดังภาพที่ 5.8 ซึ่งจะไม่เลือกลำดับที่ 1 และ ลำดับที่ 2 เพราะทั้งมีจำนวนใบสั่งซื้อมากเกินไป

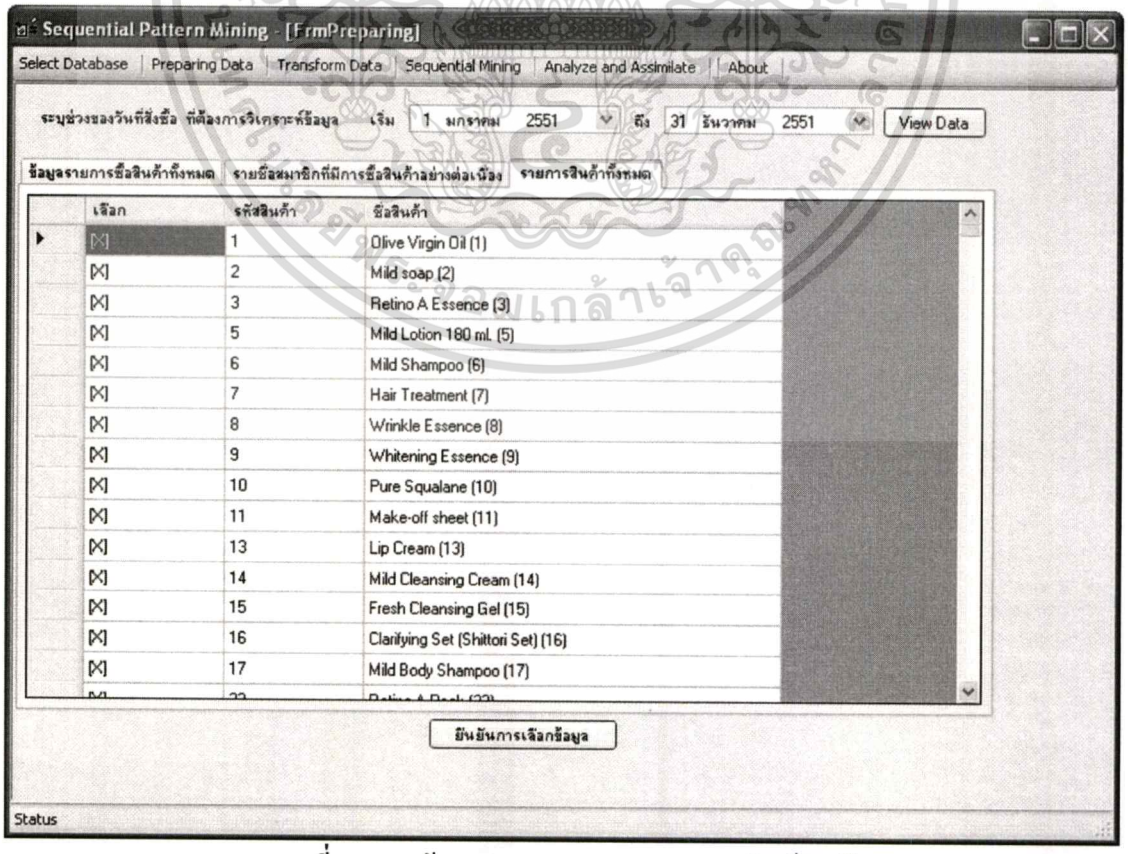
3) แท็บรายการสินค้าทั้งหมด คือ รหัสสินค้าและชื่อสินค้า ดังภาพที่ 5.9 ซึ่งผู้ใช้สามารถกำหนดได้ว่าจะ เลือก/ไม่เลือก สินค้ารายการใดมาวิเคราะห์ข้อมูลได้ด้วย



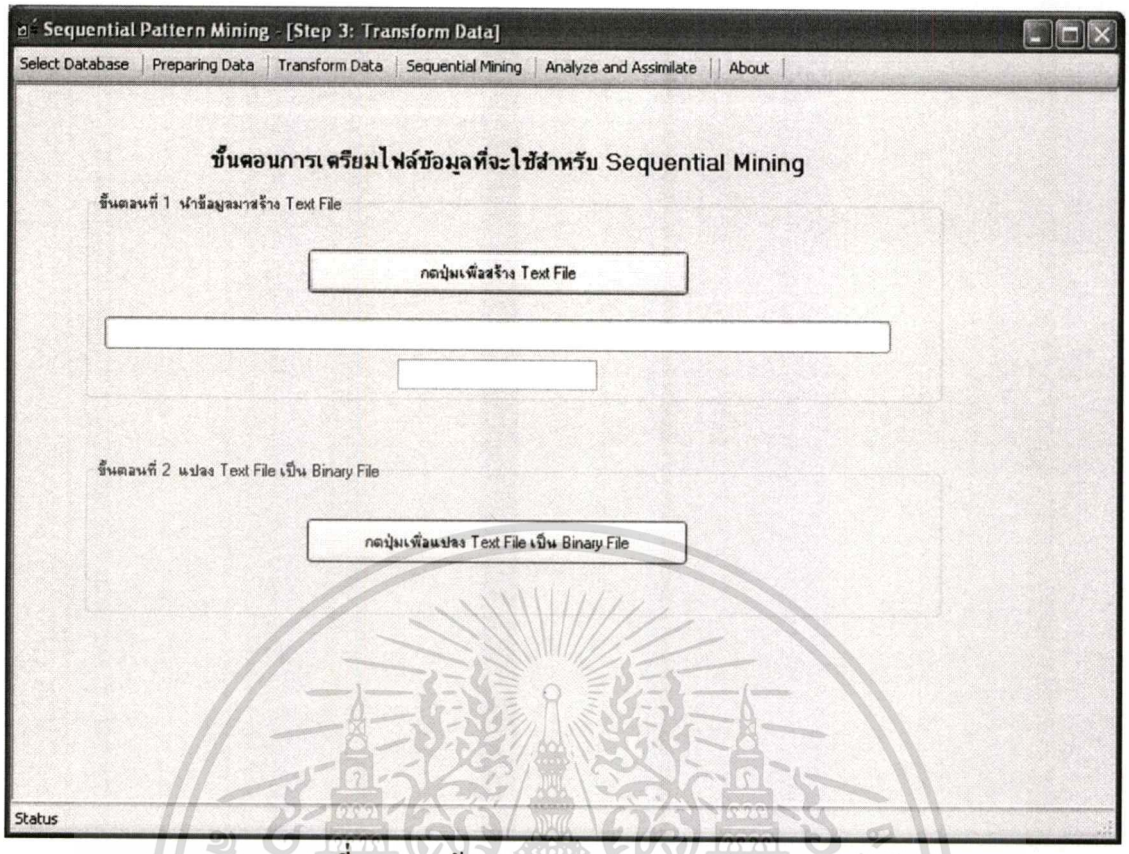
ภาพที่ 5.7 หน้าจอ Preparing Data แสดงข้อมูลทั้งหมด



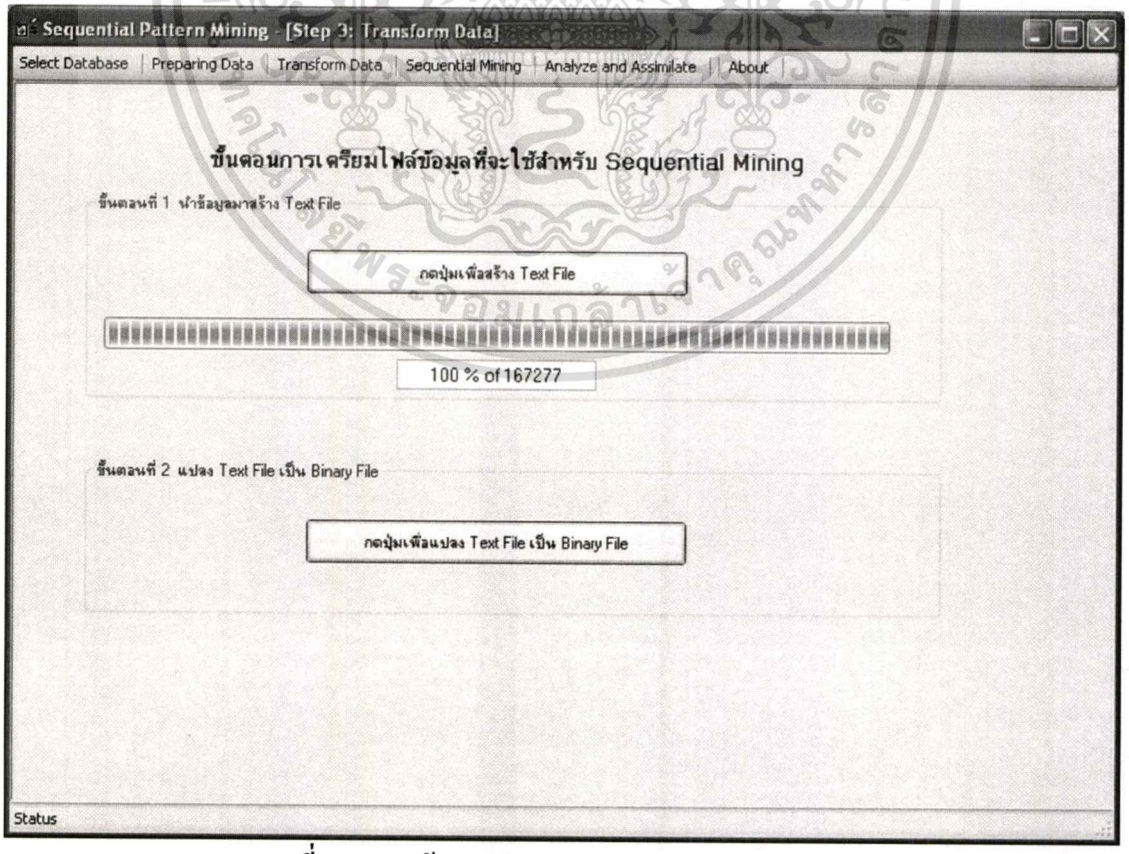
ภาพที่ 5.8 หน้าจอ Preparing Data แสดงลิสต์สมาชิก



ภาพที่ 5.9 หน้าจอ Preparing Data แสดงลิสต์สินค้า



ภาพที่ 5.10 หน้าจอ Transform Data ก่อนทำงาน

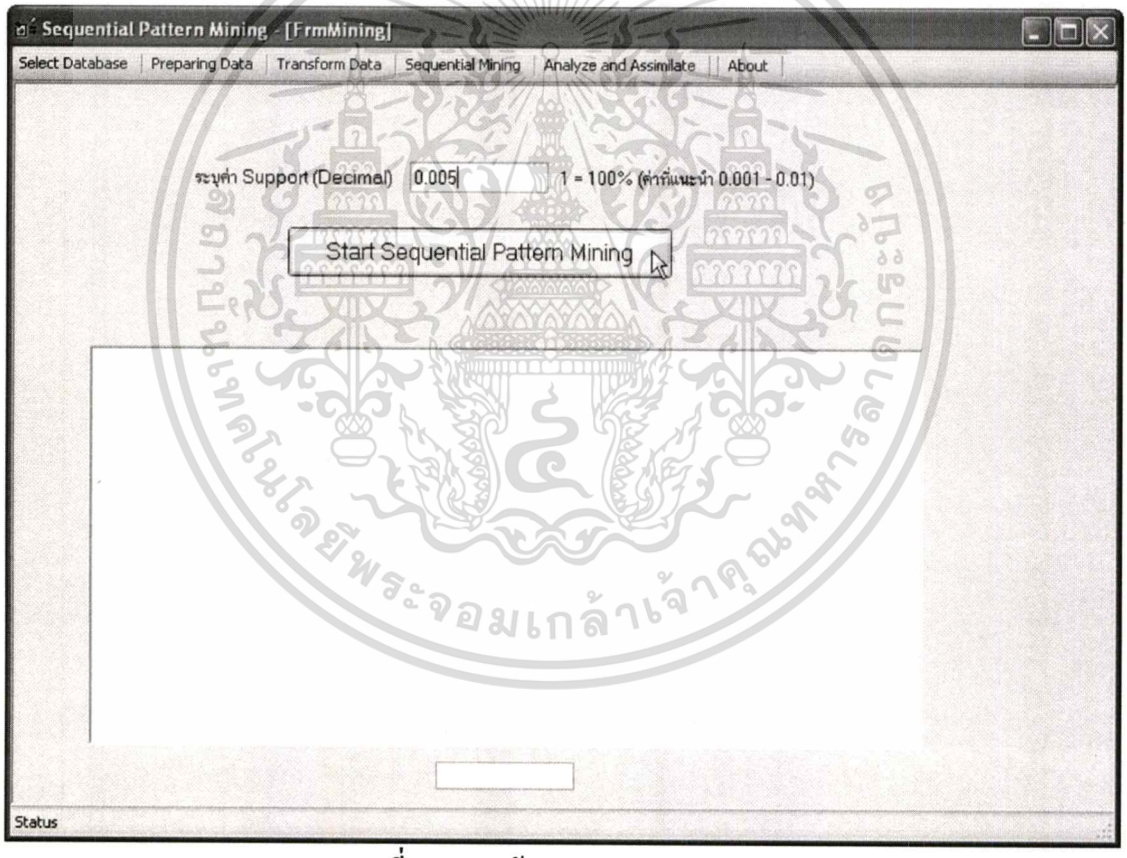


ภาพที่ 5.11 หน้าจอ Transform Data ระว่างการทำงาน

ภาพที่ 5.10 คือ หน้าจอ Transform Data เมื่อผู้ใช้กดปุ่ม เพื่อสร้าง Text File แล้วระบบจะวนลูปอ่านข้อมูลที่ได้เลือกเอาไว้จากข้างต้นมาสร้าง Text File เพื่อใส่ข้อความคั่นแต่ละใบสั่งซื้อและแต่ละรหัสสมาชิกตามรูปแบบที่ได้กำหนดไว้ ระหว่างการทำงานระบบจะแสดงสถานะการอ่านข้อมูลเป็นแถบ Progress Bar เปอร์เซนต์ของการทำงานและจำนวนเรคคอร์ดทั้งหมด พื่อให้ผู้ใช้ได้ทราบ

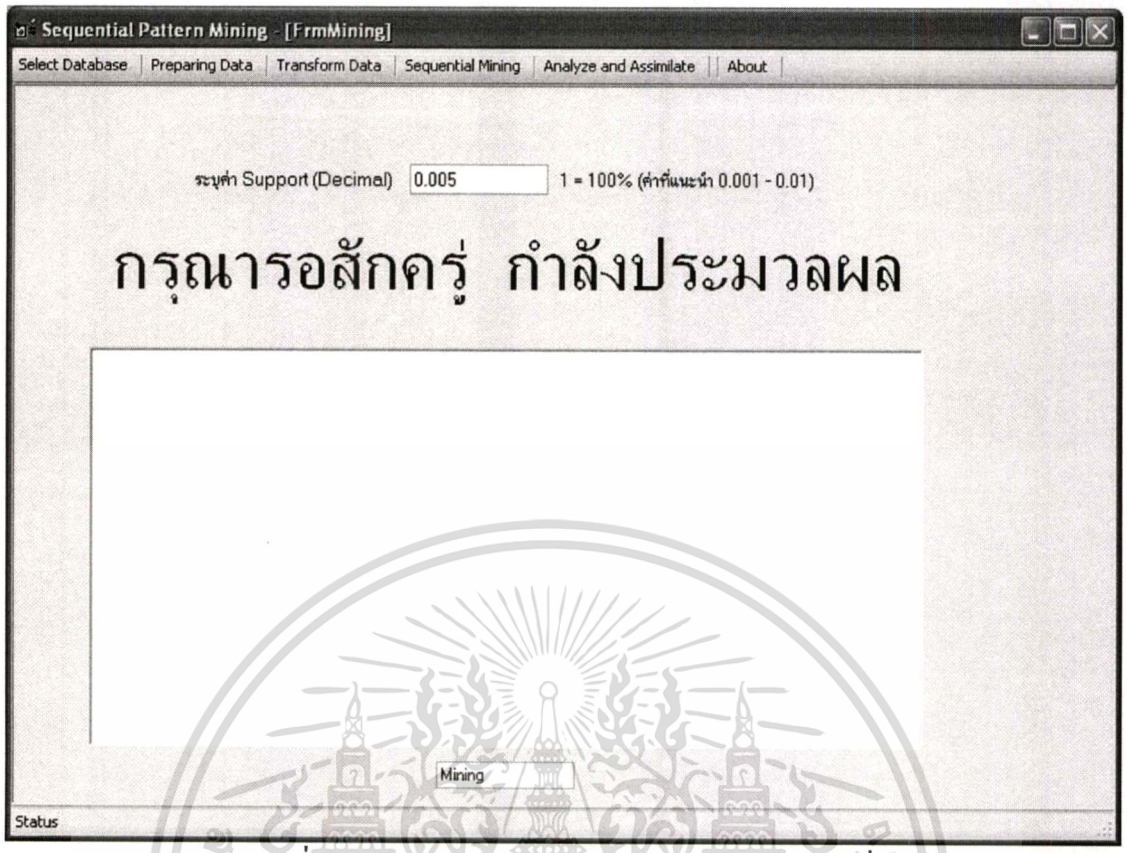
เมื่อระบบทำงานเสร็จ จะแสดงกล่องข้อความแจ้งให้ผู้ใช้ทราบ ขั้นตอนต่อไปให้ผู้ใช้กดปุ่มเพื่อแปลง Text File เป็น Binary File ตามภาพที่ 5.11

ภาพที่ 5.12 คือ หน้าจอ Sequential Mining ซึ่งจะมีกล่องข้อความให้ผู้ใช้ระบุค่า Support ที่จะใช้ในการไมน์นิ่งข้อมูล เมื่อใส่ค่าเรียบร้อยแล้วให้กดปุ่ม “Start Sequential Pattern Mining” ระบบส่งค่า Support และ Binary File ที่ได้เตรียมไว้ไปไมน์นิ่งข้อมูล

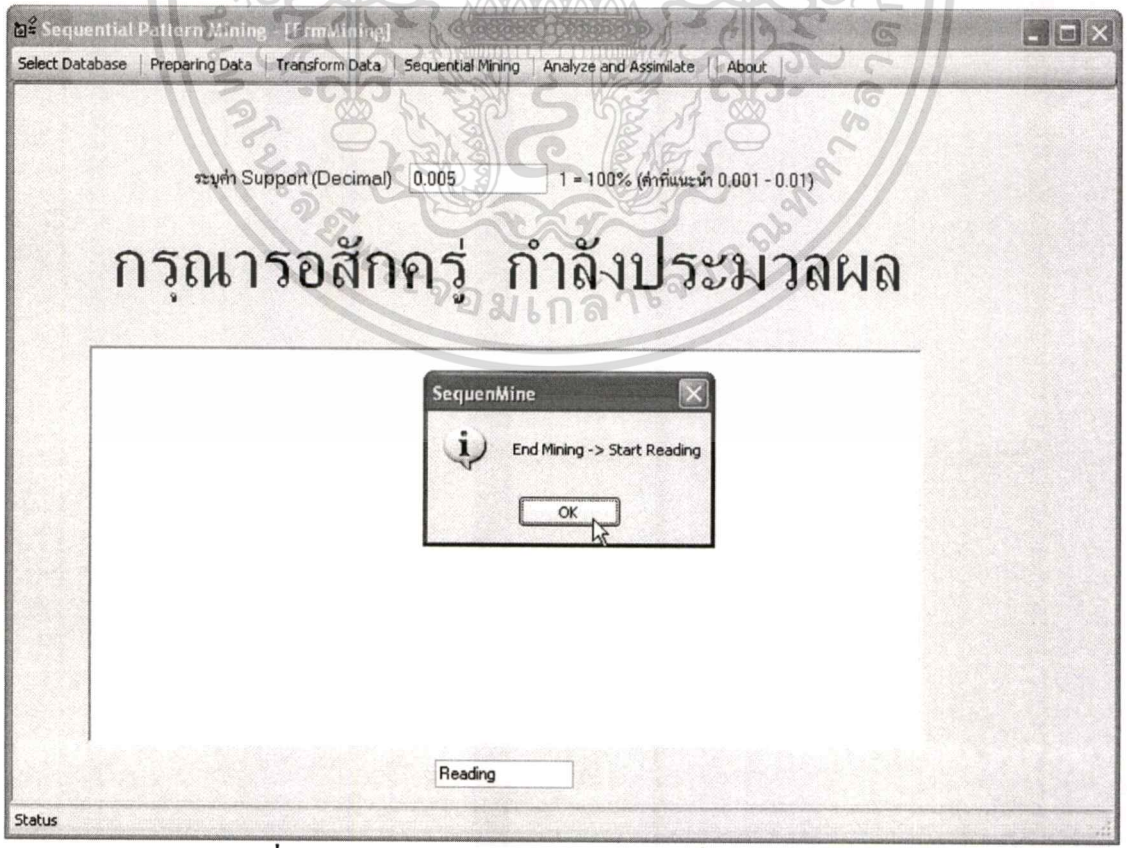


ภาพที่ 5.12 หน้าจอ Sequential Mining

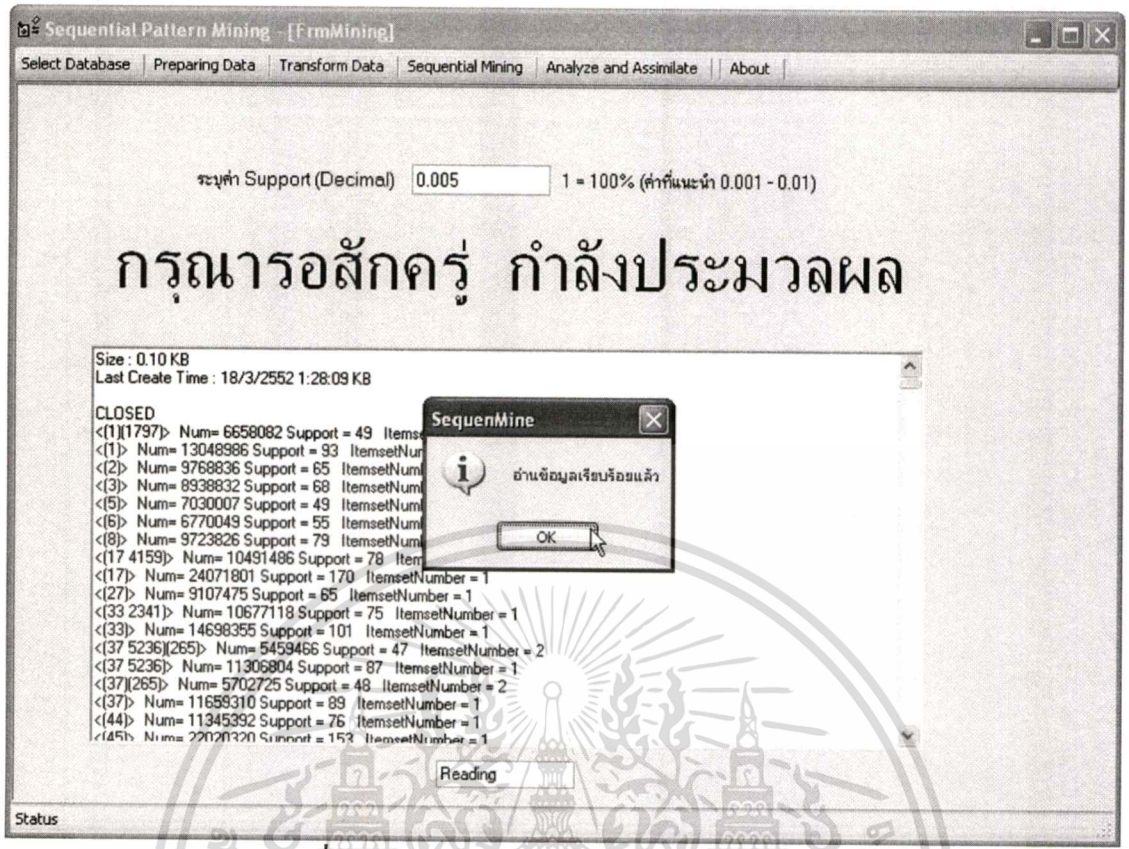
ภาพที่ 5.13 คือ หน้าจอ Sequential Mining ระหว่างการประมวลผล จะมีข้อความด้านล่างของหน้าจอแจ้งสถานะการทำงานปัจจุบันให้ผู้ใช้ทราบ และเมื่อกระบวนการไมน์นิ่งข้อมูลเสร็จแล้ว จะมีกล่องข้อความ “End Mining -> Start Reading” แจ้งตามภาพที่ 5.14 แล้วระบบจะเริ่มทำการอ่านข้อมูลผลลัพธ์ของการไมน์นิ่งจาก Text File เข้าไปเก็บในฐานข้อมูล และเมื่ออ่านข้อมูลเสร็จเรียบร้อยแล้ว จะแสดงข้อมูลและมีกล่องข้อความแจ้งผู้ใช้ตามภาพที่ 5.15



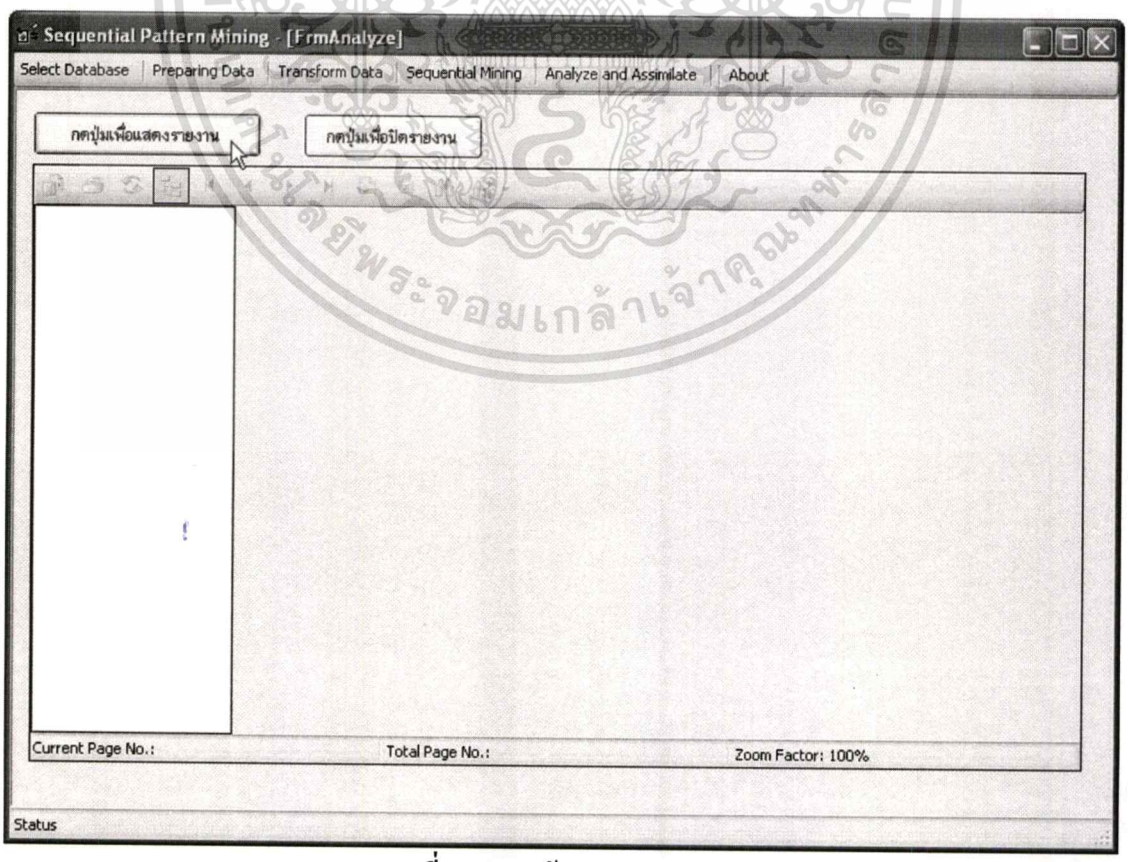
ภาพที่ 5.13 หน้าจอ Sequential Mining ระหว่างไมน์นิ่ง



ภาพที่ 5.14 หน้าจอ Sequential Mining เมื่อไมน์นิ่งเสร็จแล้ว

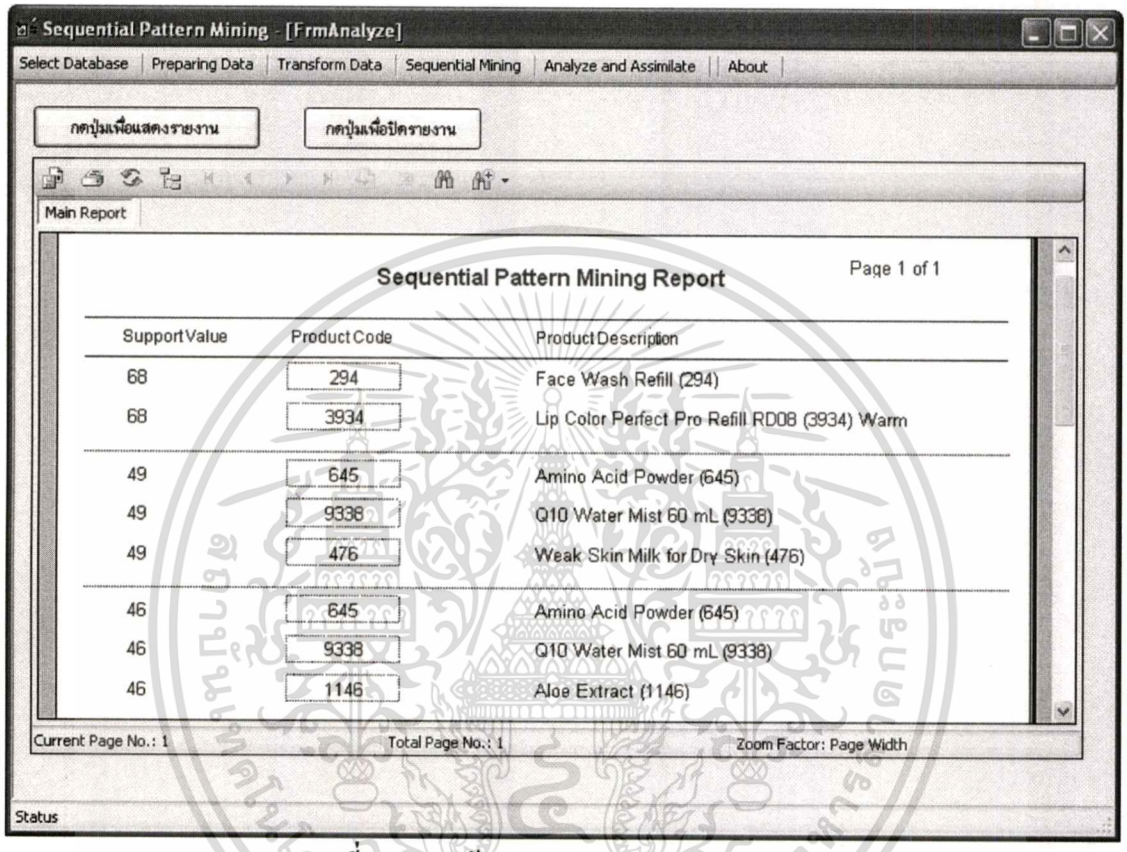


ภาพที่ 5.15 หน้าจอ Sequential Mining อ่านผลลัพธ์



ภาพที่ 5.16 หน้าจอ View Report

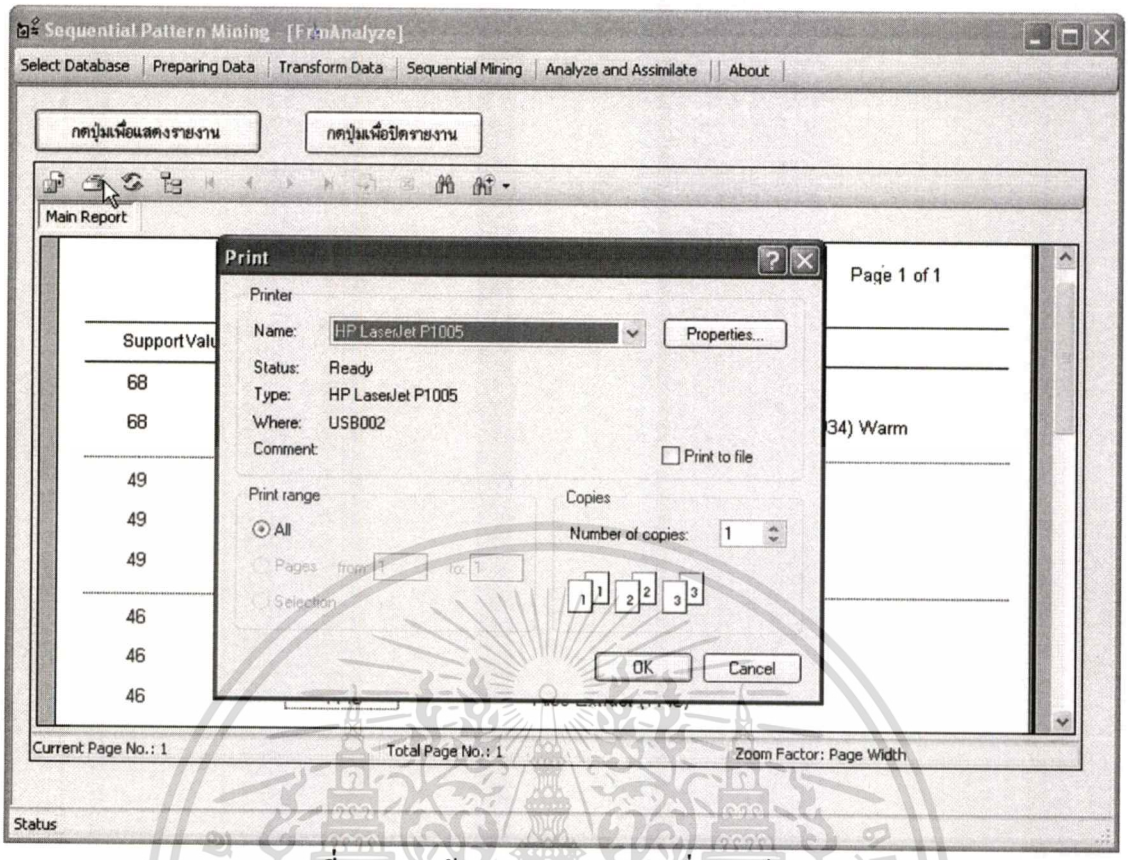
ภาพที่ 5.16 คือ หน้าจอ View Report เพื่อแสดงรายงานผลลัพธ์ที่ได้จากการไมน์นิ่งข้อมูล ซึ่งจะนำเฉพาะข้อมูลที่น่าสนใจมาแสดงรายงาน เมื่อผู้ใช้กดปุ่มเพื่อแสดงรายงาน ระบบจะแสดงรายงานขึ้นมา โดยจะมีค่า Support Value, Product Code และ Product Description แสดงในรายงาน ดังภาพที่ 5.17



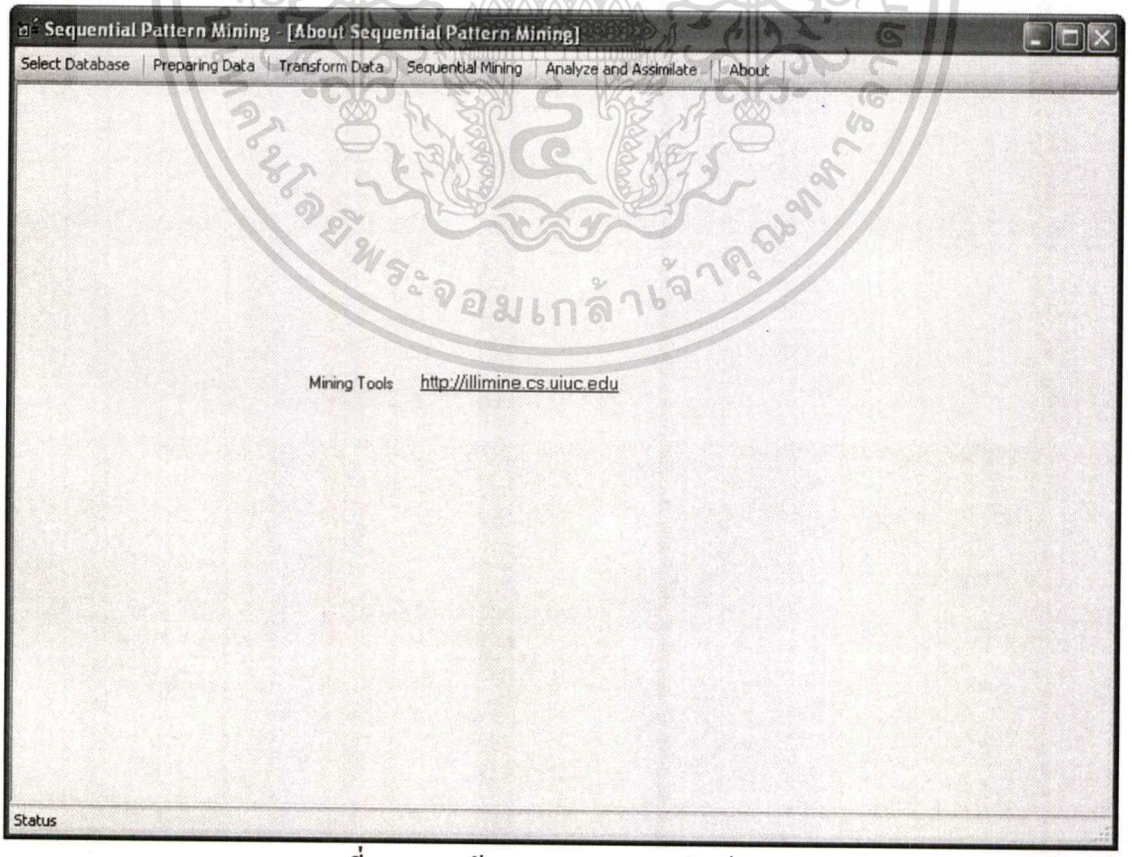
ภาพที่ 5.17 หน้าจอ View Report แสดงรายงาน

ถ้าผู้ใช้ต้องการพิมพ์รายงานนี้ ให้กดปุ่ม Print ในลำดับที่ 2 ดังภาพที่ 5.18 ที่มีลูกศรชี้อยู่ ระบบจะขึ้นกล่อง Print ขึ้นมาให้ผู้ใช้ได้เลือกเครื่องพิมพ์ กำหนดเงื่อนไขในการพิมพ์รายงาน และสั่งพิมพ์รายงาน หรือถ้าผู้ใช้ต้องการจะปิดรายงาน ให้กดปุ่มเพื่อปิดรายงาน ระบบจะทำการปิดรายงานนี้

ภาพที่ 5.19 คือ หน้าจอ About ซึ่งได้นำลิงค์ของ Illimine เพื่อให้ผู้ใช้ได้ลิงค์ไปหาข้อมูลเพิ่มเติมได้



ภาพที่ 5.18 หน้าจอ View Report เพื่อพิมพ์รายงาน



ภาพที่ 5.19 หน้าจอ About แสดงลิงค์ Illimine

บทที่ 6
บทสรุป

โครงการพัฒนาระบบดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining จัดทำขึ้นเพื่อใช้เป็นเครื่องมือวิเคราะห์ข้อมูลที่มีรูปแบบต่อเนื่องด้วยเทคนิควิธีดาต้าไมน์นิ่ง และนำผลลัพธ์ที่ได้มาแสดงในรูปแบบที่ง่ายต่อการทำความเข้าใจ โดยผ่านแอพพลิเคชั่นนี้ ซึ่งออกแบบขึ้นมาให้ง่ายต่อการใช้งาน และเปิดกว้างให้ผู้ใช้ได้เลือกฐานข้อมูลที่จะนำเข้ามาใช้อีกด้วย

6.1 สรุปผลการพัฒนาระบบ

สำหรับการพัฒนาระบบงานใหม่นี้ มีวัตถุประสงค์ที่จะสร้างโปรแกรมขึ้นมา เพื่อใช้ทำดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining โดยเริ่มตั้งแต่การคัดเลือกข้อมูลจาก MS SQL Server และด้วยความสามารถของ DTS ของ MS SQL Server นั้นยังเปิดกว้างสำหรับข้อมูลจากฐานข้อมูลอื่นๆ ได้อีกด้วย จากนั้นโปรแกรมจะทำการตรวจสอบข้อมูลเพื่อทำความสะอาด และเตรียมข้อมูลให้อยู่ในรูปแบบที่จะใช้ส่งให้กับ Illimine ใช้การทำดาต้าไมน์นิ่งต่อไป เพื่อขุดค้นหารูปแบบของความสัมพันธ์แบบต่อเนื่องของข้อมูล และนำผลที่ได้มารายงานผลให้อยู่ในรูปแบบที่ง่ายต่อการทำความเข้าใจ

6.2 ประโยชน์ของการพัฒนาระบบ

ประโยชน์ที่ได้จากการพัฒนาระบบนี้ แบ่งออกเป็น 2 ส่วน คือ ด้านผู้พัฒนาระบบงาน และด้านการวิเคราะห์ข้อมูล

ด้านผู้พัฒนาระบบงานได้ประโยชน์ คือ ได้ศึกษาและทำความเข้าใจวิธีการทำดาต้าไมน์นิ่งไม่ได้จำกัดแค่เพียง Sequential Pattern Mining เท่านั้น แต่ยังรวมถึงเทคนิควิธีการทำดาต้าไมน์นิ่งแบบอื่นๆ อีกด้วย ในปัจจุบันนี้มีเครื่องมือที่ถูกพัฒนาขึ้นเพื่อทำดาต้าไมน์นิ่งมากมาย เช่น SAS, WEGA รวมทั้ง MS SQL Server 2005 ซึ่งผู้พัฒนาระบบจะนำความรู้และพื้นฐานจากการพัฒนาระบบงานไปต่อยอดและประยุกต์ใช้ต่อไปได้

ด้านการวิเคราะห์ข้อมูล คือ ข้อมูลที่ได้จากการทำดาต้าไมน์นิ่งนั้น เป็นข้อมูลที่ทำให้รู้จักพฤติกรรมของผู้บริโภคสินค้า ซึ่งเป็นไปตามความคาดหมายของการทำดาต้าไมน์นิ่งคือทำให้พบข้อมูลที่ซ่อนอยู่ในฐานข้อมูล ซึ่งไม่เคยรู้มาก่อนและคาดไม่ถึง

6.3 ปัญหาและอุปสรรคระหว่างการพัฒนาโปรแกรม

ปัญหาและอุปสรรคระหว่างการพัฒนาโปรแกรม คือ ข้อจำกัดทางด้านเวลา เนื่องจากผู้พัฒนาระบบต้องทำงานประจำที่รับผิดชอบฟังก์ชันหลักของบริษัท ซึ่งมีความยากและค่อนข้างซับซ้อน ทำให้เกิดความเหนื่อยล้าทางความคิดและมีเวลามาศึกษาพัฒนาระบบงานนี้น้อยลง ทำให้ไม่สามารถพัฒนาระบบนี้สำเร็จตามเป้าหมายและระยะเวลาที่กำหนดได้

6.4 ข้อเสนอแนะในการพัฒนาต่อ

การวิเคราะห์ข้อมูลการซื้อสินค้าจากฐานข้อมูลขนาดใหญ่นั้น เมื่อได้ผลลัพธ์จากการทำดาต้าไมน์นิ่งด้วยเทคนิค Sequential Pattern Mining แล้ว จะพบว่ารูปแบบของข้อมูลที่มีความต่อเนื่องนั้น มีความหลากหลายของข้อมูลค่อนข้างสูง ซึ่งอาจจะเกิดจากกลุ่มของลูกค้าหรือโปรโมชั่น ที่มีผลต่อการตัดสินใจในการเลือกซื้อสินค้า ข้อเสนอแนะในการพัฒนาระบบนี้ต่อไปแยกเป็น 2 ส่วน คือ

1. เครื่องมือ CloSpan ของ Illinois ที่ได้นำมาใช้ในการขุดค้นนี้ เป็นโปรแกรมที่พัฒนาด้วย C++ ซึ่ง Open Source เปิดโอกาสให้ผู้สนใจได้ศึกษาอัลกอริทึม และโครงสร้างของคำสั่ง ที่จะนำมาพัฒนาต่อยอดให้โปรแกรมมีความสามารถสูงยิ่งขึ้นได้

2. เพิ่มการกรองข้อมูล เพื่อตัดข้อมูลที่ส่งผลให้เกิดการแปรปรวน หรือการเหวี่ยงข้อมูลออกไป เช่น โปรโมชั่นของการขายสินค้า

3. เพื่อลดความหลากหลายของรูปแบบของข้อมูล สำหรับสินค้าชนิดเดียวกัน แต่มีหลายๆ สี เช่น ลิปสติก, แป้ง หรือ อายแชโดว์ นั้น ควรจะกำหนดให้รหัสสินค้ามีเพียงรหัสเดียว เพื่อให้ผลลัพธ์ที่ได้มานั้น มีรูปแบบความต่อเนื่องเพิ่มมากยิ่งขึ้น

4. นำข้อมูลที่ได้จากการทำดาต้าไมน์นิ่งมาวิเคราะห์ในเชิงลึกต่อไป เช่น ถ้าหากได้รูปแบบของข้อมูลมาแล้ว ก็นำรูปแบบข้อมูลนั้นมาวิเคราะห์ต่อว่า รูปแบบที่เกิดขึ้นนั้นมาจากลูกค้ากลุ่มใดเพื่อจัดกลุ่มของลูกค้า เช่น อายุ เพศ รายได้ เป็นต้น ซึ่งบางทีอาจจะพบข้อมูลที่ซ่อนอยู่ว่าลูกค้าแต่ละกลุ่มที่แตกต่างกันนั้น มีผลต่อการตัดสินใจซื้อสินค้าในรูปแบบที่ไม่ซ้ำกันเลย

ภาคผนวก

เครื่องมือ CloSpan : Illimine

ในการพัฒนาระบบงานนี้ ได้นำเครื่องมือในการทำ Sequential Pattern Mining ของภาควิชา Computer Science ของ University of Illinois มาใช้ ซึ่งมีชื่อโปรแกรมว่า Illimine และชื่อ Package CloSpan ซึ่งเป็นโปรแกรมที่พัฒนาขึ้นด้วยภาษา C++ และ Open Source ซึ่งได้เขียนคำอธิบายวิธีการเรียกใช้โปรแกรม CloSpan ดังนี้

CloSpan: Mining Closed Sequential Patterns

Author: Xifeng Yan, University of Illinois at Urbana-Champaign

Contact: xifeng@gmail.com

[License Agreement]

By using the software enclosed in this package (CloSpan), you agree to become bound by the terms of this license.

(1).This software is for your internal use only. Please DO NOT redistribute it without the permission from the authors (Xifeng Yan and Jiawei Han).

(2).This software is for academic use only. No other usage is allowed without a written permission from the authors. It cannot be used for any commercial interest.

(3).The authors appreciate it if you can send us your feedback and test results including any bug report.

(4).The algorithm used in this software can be found in

"X. Yan, J. Han, R. Afshar, CloSpan: Mining Closed Sequential Patterns in Large Databases, Proc. 2003 SIAM Int. Conf. Data Mining (SDM'03), 166 - 177, 2003."

(5).The authors do not hold any responsibility for the correctness of this software, though we crosschecked all experimental results.

How-To:

CloSpan filename min_sup num_of_labels

Parameters: (1) filename, your binary data (2) min_sup, the minimum frequency of patterns (3) num_of_labels, the number of distinct item labels.

Example:

CloSpan D10N1B.data 0.1 1000

It mines all frequent sequences from "D10N1B.data", each of which should appear in at least 10% of the sequences in the dataset. 1000 means there are 1000 different symbols in this dataset.

Input Format:

1. The input is a set of sequences; each sequence has the following format

<(item_11, item_12, ..., item_1n)(item_21, item_22, ... item_2m)...>

---------- ----------

transaction 1 transaction 2 .....

Example:

<(ab)(c)(d)>

<(e)(acfh)>

...

The input is stored in a binary file, we use a 4-byte integer "-1" to separate transactions in each sequence and another 4-byte integer "-2" to separate sequences in a dataset. Each of items is encoded using a 4-byte integer. For example, <(ab)(c)(d)><(e)(acfh)> is stored as ab-1c-1d-1-2e-1acfh-1-2, where each symbol is a 4-byte integer (binary) and all of them are concatenated together.

Output:

Program status as it is executing and the final results (such as timing) are printed to stdout (console).

The discovered patterns are stored in a file named "ClosedPatterns", which is in a format of plain text.

The first column in the output file shows the discovered patterns.

The second column in the output file is the number of times that a pattern appears in the dataset.

บรรณานุกรม

ธวัชชัย งามสันติวงศ์. 2549. **การวิเคราะห์และออกแบบระบบงานเชิงวัตถุ UML2**. พิมพ์ครั้งที่ 1. กรุงเทพฯ: เซ็นจูรี.

ศุภชัย สมพานิช. 2550. **พัฒนาระบบฐานข้อมูลด้วย VB2005 & CV#2005**. พิมพ์ครั้งที่ 1. นนทบุรี: DEV BOOK.

Hong Cheng, Xifeng Yan and Jiawei Han. **"IncSpan: Incremental Mining of Sequential Pattern in Large Database"**. U.S. Mational Science Foundation NFS. University of Illinois at Urbana-Champaign.

ILLIMINE. Available URL : http://illimine.cs.uiuc.edu

Jay Ayres, Johannes Gehrke, Tomi Yiu, and Jason Flannick. **"Sequential PAttern Mining using A Bitmap Representation"**, Dept. of Computer Science,Cornell University.

Jian Pei and Jiawie Han. **"PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Project Pattern Growth"**. Intelligent Database System Research Lab. School of Computing Science. Simon Fraser University, Canada.

Jiawei Han and Micheline Kamber. 2000. **Data Mining: Concept and Techniques**. Morgan Kaufmann.

Pedro Gabriel Dias Ferreira. **"A SURVEY ON SEQUENCE PATTERN MINING ALGORITHMS"**. University of Minho,Departments of Informatics, Portugal.

Peter Cabena [et al.]. 1997. **Discovering data mining: from concept to implementation.** Prentice-Hall PTR.

Rakesh Agrawal Ramakrishnan Srikant. **"Mining Sequential Patterns"**. IBM Almaden Research Center, CA.

Xifeng Yan and Jiawei Han. **"CloSpan: Mining Closed Sequential Patterns in Large Datasets"**. U.S. Mational Science Foundation NFS. University of Illinois at Urbana-Champaign.

ประวัติผู้เขียน

ชื่อ – นามสกุล	นางสาวนันทวัน นาคศิริ
วัน เดือน ปีเกิด	12 กุมภาพันธ์ 2522
สถานที่เกิด	จังหวัดสุพรรณบุรี
ประวัติการศึกษา	วิทยาศาสตร์บัณฑิต สาขาวิทยาการคอมพิวเตอร์
	มหาวิทยาลัยแม่โจ้ จังหวัดเชียงใหม่
ประสบการณ์การทำงาน	
พ.ศ. 2547 – ปัจจุบัน	บริษัท ดีเอชซี (ประเทศไทย) จำกัด
	ตำแหน่ง Senior Programmer
พ.ศ. 2545 – 2547	บริษัท ยูสตาร์ จำกัด
	ตำแหน่ง Programmer