ระบบการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง

DATA PREPARATION AND EXPLORATION SYSTEM FOR DATA MINING

โดย

ชนินทร์ พงษ์ลิมานนท์

CHANIN PONGLIMANON

อาจารย์ที่ปรึกษา

รศ.ดร.วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2551

DATA PREPARATION AND EXPLORATION SYSTEM FOR DATA MINING

CHANIN PONGLIMANON

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

ใบรับรองโครงการพัฒนาระบบงาน (SYSTEM DEVELOPMENT PROJECT)

เรื่อง

ระบบการเตรียมข้อมูลและการสำรวจ เพื่อการทำดาต้าไมนิ่ง

DATA PREPARATION AND EXPLORATION SYSTEM FOR DATA MINING

นายชนินทร์ พงษ์ลิมานนท์

รหัสประจำตัว 49066421

ขอรับรองว่ารายงานฉบับนี้ข้าพเจ้าไม่ได้คัดลอกมาจากที่ใด
รายงานฉบับนี้ได้รับการตรวจสอบและอนุมัติให้เป็นส่วนหนึ่งของ
การศึกษาวิชาโครงการพัฒนาระบบงาน หลักสูตรวิทยาศาสตร์มหาบัณฑิต(เทคโนโลยีสารสนเทศ)
ภาคเรียนที่ 2 ปีการศึกษา 2551

.....อาจารย์ที่ปรึกษา
(รศ.ดร. วรพจน์ กรีสุระเดช)

.....กรรมการสอบ
(ผศ.ดร. พรฤดี เนติโสภากุล)

.....กรรมการสอบ
(ผศ.ดร. ธนารัตน์ ชลิดาพงศ์)

หัวข้อ	ระบบการเตรียมข้อมูลและสำรวจ สำหรับการทำดาต้าไมนิ่ง
นักศึกษา	นายชนินทร์ พงษ์ลิมานนท์
รหัสนักศึกษา	49066421
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2551
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

การดำเนินงานทางธุรกิจในยุคที่มีการแข่งขันกันสูง จำเป็นจะต้องมีการใช้เทคนิค กลยุทธ์ กลวิธีต่าง ๆ เพื่อช่วยให้การดำเนินธุรกิจอยู่เหนือคู่แข่งขัน จึงได้มีการนำเทคนิคดาต้าไมนิ่งมาใช้เพื่อวิเคราะห์ข้อมูลในฐานข้อมูลให้ได้สารสนเทศที่ซ่อนอยู่

กระบวนการที่มีบทบาทสำคัญต่อการทำดาต้าไมนิ่ง คือ การเตรียมข้อมูล เพื่อให้นำข้อมูลเข้าสู่อัลกอริทึมของดาต้าไมนิ่งได้ หากมีการเตรียมข้อมูลที่ไม่ดีแล้ว อาจทำให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งไม่มีคุณภาพด้วยเช่นกัน โครงการพัฒนาระบบงานการพัฒนาระบบการเตรียมข้อมูลและสำรวจสำหรับการทำดาต้าไมนิ่งนี้ ได้นำเทคนิคในการเตรียมข้อมูลมาใช้เพื่อปรับปรุงคุณภาพของข้อมูลให้เหมาะสมที่จะนำเข้าสู่กระบวนการดาต้าไมนิ่งต่อไปได้

Title	Data Preparation and Exploration System for Data Mining
Student	Mr. Chanin Ponglimanon
Student ID.	49066421
Degree	Master of Science
Programme	Information Science
Academic Year	2008
Advisor	Assoc. Prof. Dr.Worapoj Kreesuradej

ABSTRACT

Doing business in an era of high competitiveness, it is necessary to using techniques, strategies and artifices helps to improve business gaining advantage over competitors. This is the initiative to manipulate data mining techniques for database analysis as provide concealed information.

Data preparation is held as the key to successful data mining. In order to access data through data mining algorithms, if data preparation is defective; data mining results maybe non-qualify. Development of data preparation and exploration for data mining system is applied data preparation techniques to provide data streams of suitable quality for data mining process.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ รศ.ดร.วรพจน์ กรีสุระเดช อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงาน ที่ได้กรุณาให้ความรู้ ให้คำปรึกษาและคำแนะนำทางเทคนิคต่างๆ ที่เป็นประโยชน์ต่อการพัฒนาระบบ และสละเวลาในการ ตรวจสอบแก้ไขข้อบกพร่องของโครงการนี้

ขอกราบขอบพระคุณมารดาที่ให้โอกาสทางการศึกษากับข้าพเจ้า และเป็นกำลังใจหลักในการทำงานครั้งนี้ และขอบคุณทุก ๆ กำลังใจจากคนในครอบครัวที่ทำให้การพัฒนาระบบงานชิ้นนี้บรรลุผลสำเร็จได้เป็นอย่างดี

ขอบคุณเพื่อนๆ IS21.1 ที่เป็นกำลังใจและเป็นที่ปรึกษาในการพัฒนาระบบงานนี้

ขอบคุณคณาจารย์คณะเทคโนโลยีสารสนเทศที่ได้ประสิทธิประสาทวิชาความรู้ให้

ท้ายที่สุดนี้ คุณความดีและกุศลที่พึงบังเกิดมีจากโครงการพัฒนาระบบนี้ ข้าพเจ้าขออุทิศให้แก่นายชาญชัย เชื้อจันอัด ที่หยิบยื่นโอกาสทางการศึกษาให้กับข้าพเจ้า และสอนข้าพเจ้าว่าโอกาสทางการศึกษาไม่ได้มีกันทุกคน และอย่าละทิ้งโอกาสนั้น

ชนินทร์ พงษ์ลิมานนท์

สารบัญ

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญภาพ

รูปที่ หน้า

สารบัญภาพ(ต่อ)

รูปที่ หน้า

สารบัญภาพ(ต่อ)

รูปที่ หน้า

สารบัญภาพ(ต่อ)

รูปที่ หน้า

บทที่ 1

บทนำ

1.1 ความเป็นมา

การทำดาต้าไมนิ่ง(Data mining) ให้มีประสิทธิภาพนั้นเวลาส่วนหนึ่งต้องถูกนำมาจัดการกับกระบวนการในการเตรียมข้อมูล (Data Preparation) เพื่อทำการจัดการกับข้อมูลเหล่านั้นให้นำเข้าสู่กระบวนการของดาต้าไมนิ่งได้ หากข้อมูลที่จะนำมาวิเคราะห์ด้วยเทคนิคดาต้าไมนิ่งเป็นข้อมูลที่ไม่สมบูรณ์ จะส่งผลให้ผลลัพธ์ที่ได้จากการวิเคราะห์ในกระบวนการไมนิ่งไม่มีประสิทธิภาพดีพอต่อองค์กร

ข้อมูลที่นำมาใช้ในกระบวนการวิเคราะห์อาจเป็นข้อมูลที่รวบรวมมาจากหลายแหล่งข้อมูล ส่งผลให้ข้อมูลเหล่านั้นมีความซ้ำซ้อน ขาดความเป็นมาตรฐานเดียวกัน ค่าของข้อมูลในแต่ละแอตทริบิวมีความหลากหลาย มีขอบเขตที่กว้างเกินไป เป็นต้น ซึ่งเทคนิคในการเตรียมข้อมูล จะช่วยเพิ่มคุณภาพของข้อมูล ความถูกต้องแม่นยำ และประสิทธิภาพในกระบวนการไมนิ่ง

1.2 วัตถุประสงค์

โครงการพัฒนาระบบงานเรื่องการพัฒนาระบบการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง มีวัตถุประสงค์คือ ลดระยะเวลาในการเตรียมข้อมูลในการทำดาต้าไมนิ่ง และพัฒนาโปรแกรมเพื่อใช้เป็นเครื่องมือช่วยในการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง ซึ่งทำให้ผลลัพธ์ที่ได้จากการเตรียมข้อมูลมีคุณภาพกว่าการเตรียมข้อมูลจากการแทนค่าข้อมูลโดยไม่มีขอบเขตหรือใช้วิธีการแทนค่าข้อมูลแบบสุ่มเดา ช่วยเพิ่มมาตรฐานให้กับข้อมูลในฐานข้อมูลที่ไม่สมบูรณ์ ซึ่งจะส่งผลให้ผลลัพธ์ของการทำดาต้าไมนิ่งมีประสิทธิภาพดีเพิ่มขึ้น

1.3 ขอบเขตการดำเนินงาน

1. ข้อมูลที่จะนำมาใช้ต้องเป็นข้อมูลที่จัดเก็บใน Microsoft SQL Server 2005 เท่านั้น
2. โปรแกรมพัฒนาระบบนี้สามารถเลือกข้อมูลจากหลายตารางภายในฐานข้อมูลเดียวกันได้ โดยผู้ใช้ระบบจะต้องทราบความสัมพันธ์ของข้อมูลในแต่ละตาราง และสามารถใช้คำสั่ง SQL พื้นฐานในการเชื่อมความสัมพันธ์เหล่านั้น เพื่อเลือกข้อมูล(Data Selection) มาสร้างตารางใหม่ตามรูปแบบของโปรแกรม
3. ขั้นตอนการปรับแต่งข้อมูลที่ใช้ทำดาต้าไมนิ่ง(Data Cleansing) จะทำการกับเรคคอร์ดที่มีค่าว่าง (Null)

4. ขั้นตอนการแปลงข้อมูล (Data Transformation) สามารถแปลงค่าที่เป็น Numericให้เป็นCategorical และแปลงค่าที่เป็น Categorical ให้เป็น Numeric ได้
5. แสดงการสำรวจข้อมูลในรูปแบบของกราฟแท่ง(Bar Chart) และกราฟวงกลม(Pie Chart) ได้

1.4 ขั้นตอนและวิธีการดำเนินงาน

เพื่อให้การศึกษาเป็นไปตามวัตถุประสงค์ และขอบเขตที่กำหนด จึงได้กำหนดขั้นตอนในการดำเนินงานไว้ ดังนี้

1. ศึกษาทฤษฎีที่เกี่ยวข้องกับการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง (Data Preparation for Data Mining)
2. กำหนดวัตถุประสงค์ในการพัฒนาระบบ
3. ออกแบบระบบเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง
4. พัฒนาระบบเตรียมข้อมูลเพื่อการทำดาต้าไมนิ่ง
5. ทดสอบการใช้งานระบบ
6. สรุปผลการศึกษาและข้อเสนอแนะ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถลดระยะเวลาในการเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง
2. ได้เครื่องมือในการเตรียมข้อมูลที่มีคุณภาพ และข้อมูลมีความถูกต้องเหมาะสมต่อการนำไปใช้งาน
3. ได้เครื่องมือในการเตรียมข้อมูลที่สามารถเตรียมข้อมูลได้หลากหลาย ให้เหมาะกับแต่ละอัลกอริทึมของดาต้าไมนิ่งที่ต้องการข้อมูลในการวิเคราะห์แตกต่างกัน

1.6 เครื่องมือที่ใช้ในการพัฒนาระบบ

- Visual Studio.NET เป็นเครื่องมือที่ใช้พัฒนาภาษา ASP.NET บนระบบปฏิบัติการ Windows ซึ่งเป็นภาษาที่พัฒนาสำหรับการทำงานบนอินเตอเน็ต

- Microsoft SQL Server 2005 เป็นโปรแกรมการจัดการฐานข้อมูลในตระกูล Microsoft ถูกพัฒนาขึ้นภายใต้การใช้ภาษา SQL ที่เป็นสากล

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง (Data Mining)

ดาต้าไมนิ่งเป็นวิธีการที่ใช้ในการวิเคราะห์ข้อมูลจำนวนมากเพื่อหาแนวโน้มหรือความสัมพันธ์ของข้อมูลที่มีอยู่ข้อมูลที่ได้มาจากทำดาต้าไมนิ่งไม่ได้เกิดจากการคาดคะเนหรือจากการสมมติฐานแต่เป็นข้อมูลที่มีความสัมพันธ์ที่ซ่อนอยู่ภายใต้ข้อมูลที่เรามีอยู่ดังนั้นในการทำดาต้าไมนิ่งจึงไม่ได้เป็นการตั้งสมมติฐานแต่เป็นการดูผลลัพธ์ที่ได้จากการทำงานมากกว่าจะเห็นได้ว่าการทำดาต้าไมนิ่งนั้นเป็นวิธีการที่แตกต่างไปจากวิธีการวิเคราะห์ข้อมูลทางสถิติในแบบอื่นๆในการทำดาต้าไมนิ่งนั้นผลลัพธ์ที่เกิดขึ้นถือได้ว่าเป็นข้อมูลที่มีประโยชน์เป็นอย่างมากโดยสามารถที่จะนำข้อมูลเหล่านี้ไปใช้เป็นแนวทางในการตัดสินใจที่ก่อให้เกิดผลดีในการทำธุรกิจ

โดยทั่วไปแล้วในการทำ ดาต้าไมนิ่งนั้นมีด้วยกันอยู่สองบรรทัดฐานด้วยกัน คือ การค้นหาความรู้ (Knowledge Discovery: KD) และการสร้างแบบจำลองการคาดการณ์ (Predictive Modeling: PM) ในทางปฏิบัติแล้วจะทำการประยุกต์ใช้ AI หรือ เทคโนโลยีในการเรียนรู้ ของเครื่องจักร ในการวิเคราะห์ข้อมูลในฐานข้อมูลขนาดใหญ่ จุดประสงค์ของทั้งสองบรรทัดฐานนี้ก็คือ พยายามที่จะสร้างกระบวนการที่เป็นแบบอัตโนมัติให้มากที่ สุดเท่าที่จะเป็นไปได้ ซึ่งในทางปฏิบัติแล้ว การทำดาต้าไมนิ่งไม่ใช่ระบบอัตโนมัติอย่างสมบูรณ์ทั้งหมด แต่เป็นกระบวนการแบบกึ่งอัตโนมัติเท่านั้น

ปัจจัยที่ทำให้ดาต้าไมนิ่งเป็นที่ใช้งานกันอย่างกว้างขวางเป็นผลเนื่องมาจาก

1. ขนาดของข้อมูลมีขนาดใหญ่และขยายตัวอย่างรวดเร็ว การสืบค้นข้อมูลจะมีประโยชน์ก็ต่อเมื่อฐานข้อมูลที่ใช้มีขนาดใหญ่มาก

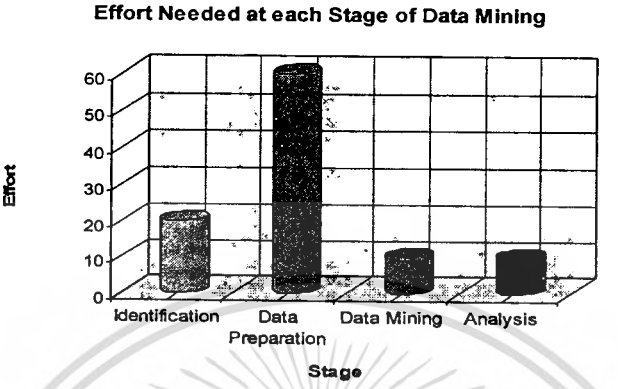
2. ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบสนับสนุนการตัดสินใจ เพื่อเป็นการง่ายต่อการนำข้อมูลมาใช้ในการวิเคราะห์ ส่วนมากข้อมูลจะถูกจัดเก็บอยู่ในรูปของ Data Warehouse ซึ่งเป็นการง่ายต่อการนำไปใช้ในการสืบค้นความรู้

3. เทคนิคดาต้าไมนิ่งประกอบไปด้วยอัลกอริทึมที่มีความซับซ้อนจึงจำเป็นต้องใช้งานกับระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงด้วย ซึ่งในปัจจุบันระบบคอมพิวเตอร์ที่มีประสิทธิภาพสูงในท้องตลาดก็มีราคาถูกลงมาก จึงเป็นสาเหตุให้มีการนิยมใช้ดาต้าไมนิ่งกันมากขึ้น

4. ในวงการธุรกิจมีการแข่งขันกันสูง จึงมีข้อมูลเกิดขึ้นเป็นจำนวนมากแต่ไม่มีการนำมาใช้ให้เกิดประโยชน์ จึงมีความจำเป็นที่จะต้องนำเทคนิคดาต้าไมนิ่งมาใช้เพื่อให้ได้ความรู้นั้น

2.2 ขั้นตอนการทำดาต้าไมนิ่ง

กระบวนการของการทำดาต้าไมนิ่งประกอบไปด้วยขั้นตอนหลักที่สามารถแบ่งได้ 5 ขั้นตอน ดังนี้



รูปที่ 2.1 แสดงขั้นตอนในการทำดาต้าไมนิ่ง

ขั้นตอนที่1: กำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจจะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะจะเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่งและสามารถกำหนดได้ว่าเมื่อไหร่จะใช้ดาต้าไมนิ่งในการแก้ปัญหา เนื่องจากในทุกปัญหาไม่สามารถแก้ไขได้ด้วยหลักการดาต้าไมนิ่งทั้งหมด ซึ่งในส่วนนี้จะประกอบไปด้วยการวิเคราะห์ทางธุรกิจและวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลอะไรและต้องการอะไรจากข้อมูลที่มีอยู่

ขั้นตอนที่ 2: การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ต้องใช้เวลานานที่สุดประมาณ 60% ของการทำดาต้าไมนิ่ง เพราะเป็นส่วนที่มีความสำคัญที่สุดในการทำดาต้าไมนิ่ง เนื่องจากข้อมูลที่นำมาใช้ในการทำดาต้าไมนิ่งเป็นข้อมูลที่มาจากฐานข้อมูลขนาดใหญ่ที่อาจมาจากฐานข้อมูลหลายๆแหล่งมารวมกัน ข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความถูกต้องเพื่อส่งผลให้ผลลัพธ์ในการทำไมนิ่งมีประสิทธิภาพ รายละเอียดของการเตรียมข้อมูลอยู่ในบทที่ 3 การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

ขั้นตอนที่ 3: การทำดาต้าไมนิ่ง (Data Mining)

เป็นขั้นตอนการประมวลผลข้อมูลตามอัลกอริทึมที่กำหนดไว้ ซึ่งเกี่ยวข้องกับการเลือกอัลกอริทึมในการทำดาต้าไมนิ่งซึ่งจะต้องพิจารณาลักษณะของปัญหาเป็นหลัก เพราะในแต่ละปัญหาต้องเลือกใช้อัลกอริทึมที่เหมาะสมจึงจะได้ผลการวิเคราะห์ที่ถูกต้อง ซึ่งอาจใช้หลายอัลกอริทึมเพื่อเปรียบเทียบผลลัพธ์ได้

ขั้นตอนที่ 4: การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง (Analysis of Result)

เป็นการวิเคราะห์และตีความหมายจากผลที่ได้ เช่น ศึกษาพฤติกรรมของลูกค้าไม่ให้ผลออกมาตรงๆ แต่จะได้ความสัมพันธ์จำนวนมาก ผู้ใช้ต้องนำมาวิเคราะห์และประเมินกฎเหล่านี้เอง ตัวอย่างเช่น การแบ่งส่วนข้อมูล ผลที่ได้จะรู้ข้อมูลกลุ่มไหนๆแต่ต้องวิเคราะห์เองว่าแต่ละกลุ่มหมายถึงอะไร ซึ่งวิธีการนี้ผลที่ได้จะแปลความหมายยากและใช้เวลานาน วิธีเลือกสิ่งที่น่าสนใจจากผลลัพธ์ของดาต้าไมนิ่งเป็นวิธีที่วัดว่าผลที่ได้นั้นน่าสนใจแค่ไหนจะเลือกโดยดูจากผลที่ได้ง่ายต่อความเข้าใจ และ เป็นสารสนเทศที่ใหม่ สมเหตุสมผล

ขั้นตอนที่ 5: การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of Knowledge)

การนำความรู้ที่ได้ไปใช้เป็นขั้นตอนสุดท้ายของกระบวนการทั้งหมด ซึ่งเป็นการรวบรวมความเข้าใจในแบบจำลองที่เป็นผลมาจากขั้นตอนการวิเคราะห์ผลลัพธ์ที่ได้ มารวมเข้ากับส่วนความรู้ทางธุรกิจเพื่อที่จะนำเสนอถึงวิธีการที่จะนำผลที่ได้นี้ไปใช้ให้เกิดประโยชน์ ในขั้นตอนนี้จะมีหลักอยู่ 2 ประการคือ

1. แสดงแนวคิดทางธุรกิจที่ค้นพบใหม่
2. กฎเกณฑ์ที่จะใช้ความรู้ใหม่ที่พบให้ได้ประโยชน์สูงสุด

2.3 เทคนิคการทำดาต้าไมนิ่ง

ดาต้าไมนิ่งมีเทคนิคและอัลกอริทึมที่สามารถนำมาใช้งานหลายประเภท ขึ้นอยู่กับเอพลิเคชั่น (Application) ที่ต้องการนำมาใช้งาน แบ่งออกเป็นรูปแบบต่าง ๆ ได้ดังตารางที่ 2.1

ตารางที่ 2.1 เทคนิคของดาต้าไมนิ่ง

Predictive Modeling	Classification
	Value Prediction
Database Segmentation	Demographic Clustering
	Neural Clustering
Link Analysis	Associations Discovery
	Sequential Pattern Discovery
Deviation Detection	Visualization
	Statistics

2.3.1 การสร้างแบบจำลองพยากรณ์ (Predictive Modeling)

เป็นการทำนายความเป็นไปได้ โดยใช้การสังเกตจากรูปแบบของข้อมูลที่มีอยู่ คือจะใช้วิธีนี้ในการวิเคราะห์ฐานข้อมูลที่มีอยู่เพื่อตัดสินใจเลือกลักษณะข้อมูลที่ต้องการ โดยมีลักษณะเป็นการเรียนรู้จากกลุ่มข้อมูลที่ได้กำหนดไว้ แล้วจึงนำไปวิเคราะห์กลุ่มข้อมูลที่ต้องการ ซึ่งวิธีนี้เรียกว่า Supervised Learning ดังนั้นข้อมูลที่มีอยู่ต้องสมบูรณ์ จึงจะทำให้ผลลัพธ์ออกมาถูกต้อง เพราะเราต้องนำข้อมูลในอดีตมาสร้างแบบจำลอง การทำงานจะแบ่งออกเป็น 2 ขั้นตอน คือ

Training Phase คือขั้นตอนการสร้างแบบจำลองขึ้นมาโดยใช้ข้อมูลในอดีต ซึ่งจะใช้ข้อมูลประมาณ 80% ของข้อมูลทั้งหมด

Testing Phase คือขั้นตอนที่ใช้ทำการทดสอบแบบจำลองที่สร้างว่ามีความเหมาะสมหรือไม่โดยจะนำข้อมูลส่วนที่เหลือ 20% มาใช้ทดสอบแบบจำลองที่สร้างขึ้น

การสร้างแบบจำลองพยากรณ์ สามารถแบ่งย่อยได้อีก เป็น 2 ประเภท คือ

- Classification เป็นการทำนายว่าสิ่งนั้นควรอยู่ในกลุ่มไหน ซึ่งเป็นการแบ่งกลุ่มของข้อมูลตามชนิดของกลุ่มข้อมูลที่ควรจะเป็น และสามารถแบ่งกลุ่มข้อมูลได้อย่างชัดเจน เช่น การจัดกลุ่มของลูกค้าเพื่อพิจารณาว่าควรจะให้วงเงินสินเชื่อเพิ่มขึ้นหรือไม่ เป็นต้น ซึ่งวิธีที่นิยมใช้คือ Tree Induction และ Neural Induction

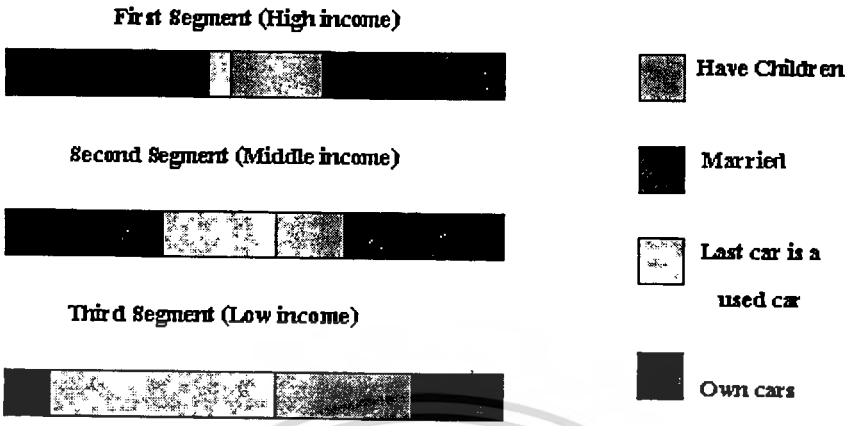
- Value Prediction เป็นการทำนายถึง ค่าความต่อเนื่องของข้อมูล เป็นการทำนายค่าที่เป็น Numeric เช่น การทำนายราคาหุ้น เป็นต้น โดยมีวิธีที่ใช้คือ Linear Regression และ Nonlinear Regression

2.3.2 การแบ่งส่วนฐานข้อมูล (Database Segmentation)

จะเป็นการแบ่งหรือจัดกลุ่มของข้อมูลที่มีลักษณะคล้ายกัน หรือมีคุณสมบัติใกล้เคียงกันในหลายๆ ด้าน ให้เป็นข้อมูลกลุ่มเดียวกัน ซึ่งแต่ละกลุ่มจะถูกเรียกว่าเซกเมนต์ (Segments) หรือคลัสเตอร์ (Clusters) การแบ่งกลุ่มข้อมูลนี้เราจะไม่สามารถกำหนดได้ว่าข้อมูลใดควรจะอยู่กลุ่มใด แต่จะเป็นการกำหนดกลุ่มของข้อมูลจากธรรมชาติของข้อมูลเอง ไม่ได้ใช้ความรู้สึกหรือประสบการณ์ในการตัดสินใจแบ่งกลุ่มข้อมูล และข้อมูลจะถูกจัดการโดยอัลกอลิทึมที่เหมาะสม จึงเรียกว่าเป็นรูปแบบของ Unsupervised Learning ซึ่งสามารถแบ่งย่อยตามวิธีที่ใช้ เช่น Demographic Clustering และ Neural Clustering ยกตัวอย่าง เช่น บริษัทจำหน่ายรถยนต์แห่งหนึ่งได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1. กลุ่มผู้มีรายได้สูง (>$80,000)
2. กลุ่มผู้มีรายได้ปานกลาง ($25,000 to $ 80,000)
3. กลุ่มผู้มีรายได้ต่ำ (less than $25,000)

และภายในแต่ละกลุ่มยังแยกออกเป็น Have Children, Married, Last car, is a used car, Own cars



รูปที่ 2.2 การแยกกลุ่มลูกค้าของบริษัทรถยนต์แห่งหนึ่ง

จากข้อมูลข้างต้นทำให้ทางบริษัททราบว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

2.3.3 การวิเคราะห์ความสัมพันธ์ (Link Analysis)

เป็นการหาความสัมพันธ์ของข้อมูล เช่น ลูกค้าเข้าร้านซื้อสินค้าอะไรบ้าง, วันที่เข้าร้าน (มักจะนำดาต้าไมนิ่งไปใช้กับพวกธุรกิจค้าปลีก) ซึ่งจะมีเทคนิค ได้แก่

- Associations Discovery เป็นหลักการค้นหาสิ่งที่มีความสัมพันธ์กัน

- Sequential Pattern Discovery เป็นการศึกษาว่าเหตุการณ์ใดเกิดแล้วเหตุการณ์ใดจะเกิดตามมา เช่น การกู้จะกู้เพื่อการศึกษาก่อน จากนั้นแล้วจะกู้เพื่อแต่งงานเป็นต้น หาลำดับว่ามีรูปแบบ (Pattern) เหล่านี้ เช่น กู้ซื้อบ้านแล้วต้องกู้ซื้อรถด้วย เป็นต้น

- Similar Time Sequence Discovery เป็นการศึกษาพฤติกรรมของข้อมูลที่เกิดขึ้นทั้งหมดหรือเกิดขึ้นในช่วงเวลาเดียวกัน เพื่อหาความสัมพันธ์ระหว่างกลุ่มของข้อมูลเหล่านี้

2.3.4 การตรวจสอบค่าเบี่ยงเบน (Deviation Detection)

เป็นเทคนิคที่ใช้ทำการหาค่าที่มีความแตกต่างไปจากค่ามาตรฐานว่ามีค่ามากน้อยเพียงใด เป็นแบบจำลองที่ใช้เทคนิคทางสถิติ (Statistics) เพื่อใช้วัดความน่าเชื่อถือของข้อมูล และการแสดงให้เห็นภาพ (Visualization) ซึ่งเป็นการสรุปข้อมูลให้แสดงผลออกมาในรูปแบบกราฟิก เช่น แผนภูมิแท่ง หรือ แผนภูมิวงกลม เป็นต้น เพื่อให้สามารถเข้าใจได้ง่าย นอกจากนี้ยังสามารถนำไปใช้ร่วมกับเทคนิคอื่นๆ โดยใช้ในการแสดงผลที่ได้ในรูปแบบของกราฟิก

2.4 การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

เป้าหมายในการเตรียมข้อมูล เพื่อทำให้ข้อมูลมีความเหมาะสมกับอัลกอริทึมและเพื่อเพิ่มประสิทธิภาพในการทำดาต้าไมนิ่ง ดังนั้นจะเห็นได้ว่ากระบวนการเตรียมข้อมูลมีความสำคัญเป็นอย่างมากและเวลาส่วนใหญ่จึงถูกใช้ไปในกระบวนการนี้

2.4.1 การเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง

เป็นขั้นตอนในการทำข้อมูลดิบที่เราได้รับมา ซึ่งจะอยู่ในรูปแบบที่หลากหลายแตกต่างกันไปให้อยู่ในรูปที่พร้อมจะใช้งาน เพื่อให้ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่งมีความถูกต้องแม่นยำมากยิ่งขึ้น เป็นขั้นตอนที่ใช้เวลานาน เนื่องจากปริมาณข้อมูลมีเป็นจำนวนมากและข้อมูลที่ได้รับมาจากหลายแหล่ง รูปแบบของข้อมูลจึงแตกต่างกัน จึงต้องมีการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกันเพื่อให้พร้อมใช้งาน โดยขั้นตอนในการเตรียมข้อมูลนี้แบ่งออกเป็น 3 ขั้นตอนได้ดังนี้

1. การเลือกข้อมูล: Data Selection
2. การเตรียมข้อมูล: Data Preprocessing
3. การแปลงข้อมูล: Data Transformation

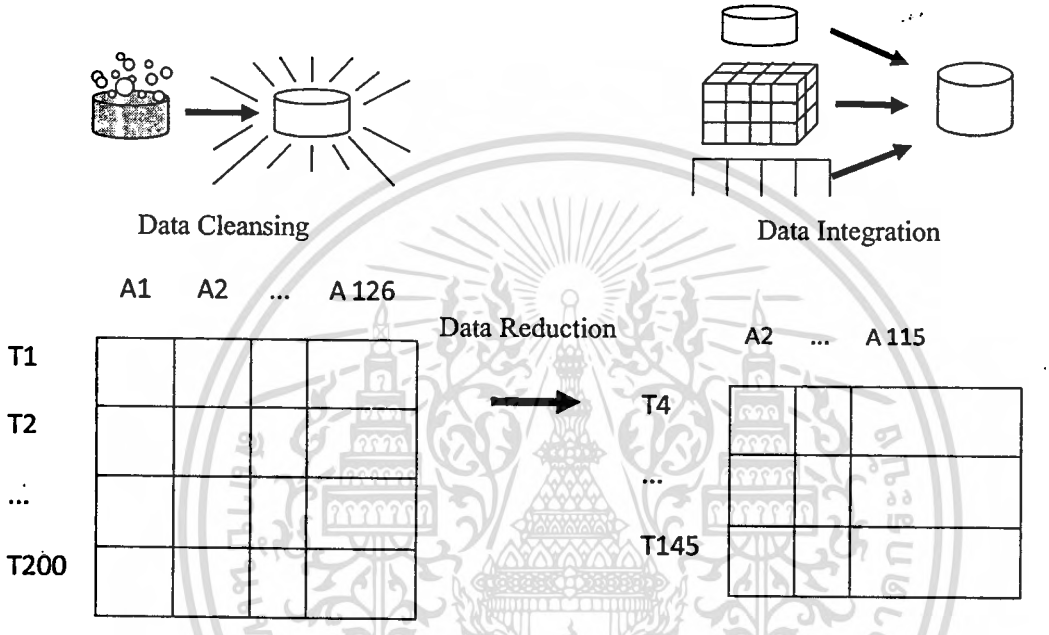
2.4.2 การเลือกข้อมูล: Data Selection

การเลือกข้อมูล เป็นการเลือกข้อมูลที่สำคัญออกมาจากฐานข้อมูลเพื่อทำการวิเคราะห์ในขั้นตอนต่อไป ข้อมูลที่นำมาวิเคราะห์นั้นต้องขึ้นอยู่กับวัตถุประสงค์ทางธุรกิจขององค์กรที่ได้กำหนดไว้ การเลือกข้อมูลจำเป็นจะต้องเข้าใจความหมายประเภทข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบและลักษณะอื่นๆ ของข้อมูล โดยแบ่งลักษณะของข้อมูลได้เป็น 2 ลักษณะ คือ

- ข้อมูลแบบแบ่งประเภท (Categorical)
    - Nominal: ตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น เพศ (ชาย, หญิง)
    - Ordinal: ตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น ลำดับของสินค้า (ดี, ไม่ดี)
- ข้อมูลแบบปริมาณ (Quantitative)
    - Continuous: ค่าที่เก็บเป็นเลขจำนวนจริง หรือเป็นค่าต่อเนื่อง เช่น จำนวนเงิน
    - Discrete: ค่าที่เก็บเป็นเลขจำนวนเต็ม เช่น จำนวนบุตร

2.4.3 การเตรียมข้อมูล: Data Preprocessing

วัตถุประสงค์ของการเตรียมข้อมูล คือ การแก้ไขปัญหาที่พบในข้อมูลเพื่อทำให้ข้อมูลมีคุณภาพก่อนที่จะนำข้อมูลไปประมวลผล ซึ่งในขั้นตอนนี้จะแบ่งออกเป็น 3 ขั้นตอนย่อย ดังนี้ Data Cleansing ที่ใช้ขจัดข้อมูลที่มีค่าผิดจากค่าที่ควรจะเป็น (noise) และขจัดข้อมูลที่ขัดแย้งกัน (inconsistencies), Data integration ใช้เพื่อรวมข้อมูลจากหลายๆ แหล่งให้เป็นข้อมูลก้อนเดียวกัน, Data Reduction ทำการลดขนาดของข้อมูลที่จะใช้ทำไมนิ่ง



รูปที่ 2.3 รูปแบบของการเตรียมข้อมูล

Data Cleansing เป็นขั้นตอนในการเลือกข้อมูลที่ต้องการและเอาข้อมูลที่ไม่ต้องการออกจากแหล่งข้อมูล ซึ่งข้อมูลส่วนใหญ่ในฐานข้อมูลนั้นมักจะไม่สมบูรณ์ (Incomplete) โดยการเติมข้อมูลใหม่แทนข้อมูลเดิมที่ขาดหาย (missing values), ขจัดข้อมูลที่ผิดจากค่าที่ควรจะเป็น (noisy data) หรือ ลบข้อมูลที่อยู่นอกเหนือขอบเขตของข้อมูล (remove outliers), ขจัดปัญหาข้อมูลที่ขัดแย้งกัน (resolve inconsistencies) ซึ่งวิธีการของ Data Cleansing ที่นำมาใช้ในการกำจัด Missing Value มีดังนี้

ข้อมูลที่ขาดหาย ไม่มีข้อมูลในบางแอทริบิว ซึ่งอาจเกิดจากการบันทึกข้อมูลผิดพลาดหรือกรอกข้อมูลไม่ครบถ้วน วิธีการในการเติมข้อมูลให้กับค่าที่ขาดหายไปนั้น มีดังนี้

1. Ignore the tuple: โดยไม่สนใจทิ้งข้อมูลแถว (record) นั้นไปเลย ซึ่งเป็นวิธีที่ไม่มีคุณภาพนัก นอกเสียจากว่าข้อมูลในแถวนั้นมีแอทริบิวว่างเป็นจำนวนมาก

2. Fill in the missing value manually: วิธีการนี้คือเติมค่าของแอทริบิวลงไปเอง ซึ่งเป็นวิธีที่ไม่เหมาะกับฐานข้อมูลขนาดใหญ่ๆ ที่มีข้อมูลขาดหายเป็นจำนวนมาก

3. Use a global constant to fill in the missing value: วิธีการนี้จะเติมค่าคงที่หรือตัวแทนของค่าที่กำหนดขึ้นลงในแอทริบิวเช่น "Unknown" หรือ "∞" แต่การเติมข้อมูลเหล่านี้อาจส่งผลให้ผลจากการทำดาต้าไมนิ่งเกิดความผิดพลาดและไม่มีประสิทธิภาพได้

4. Use the attribute mean to fill in the missing value: คือการใช้ค่าเฉลี่ยของแอทริบิวนั้นมาเติมลงในค่าที่ขาดหายไป เช่น การนำค่าเฉลี่ยรวมของรายได้มาเติม

5. Use the attribute mean for all samples belonging to the same class as the given tuple: ตัวอย่างเช่น หากมีการจัดกลุ่มของลูกค้าตาม credit_risk แล้วก็ให้นำค่าเฉลี่ยรวมของรายได้ของลูกค้าในแต่ละกลุ่มเติมลงในค่าที่ขาดหายไปในลูกค้ากลุ่มเดียวกัน

6. Use the most probable value to fill in the missing value: โดยวิธีการเติมค่าที่คาดว่าจะเป็นไปได้ลงในแอทริบิวโดยใช้ Decision tree ในการทำนายค่าที่ขาดหายไป

2.4.4 การแปลงข้อมูล: Data Transformation

การแปลงข้อมูลมีวัตถุประสงค์ 2 อย่างคือ ทำให้ไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำมาใช้ เนื่องจากข้อมูลที่จะนำมาใช้ทำดาต้าไมนิ่งในบางครั้งอยู่ในรูปแบบที่ไม่เหมาะสมกับอัลกอริทึมที่เลือกใช้ ดังนั้นจึงจำเป็นที่จะต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมนั้นๆก่อน โดยวิธีการแปลงข้อมูลมีอยู่หลายวิธีซึ่งขึ้นอยู่กับปัญหาของข้อมูล

1. Normalization: การทำให้ข้อมูลในแอทริบิวมีค่าไม่เกินขอบเขตที่กำหนด เช่น -1.0 ถึง 1.0, 0.0 ถึง 1.0

2. Attribute construction: สร้างแอทริบิวใหม่เพิ่มในแอทริบิวเซต เพื่อช่วยในกระบวนการดาต้าไมนิ่ง

Attribute ที่ทำการ ด้วยการกำหนดค่าให้อยู่ในขอบเขตที่กำหนด เช่น 0.0 ถึง 1.0 ซึ่งการทำ normalize นั้นเป็นประโยชน์ต่อการจัดกลุ่มเพื่อใช้ใน Algorithm Neural Network, Nearest Neighbor Classification และ Clustering การ normalize ค่าของอินพุท จะทำให้กระบวนการในการหาความรู้ (Learning Phase) ทำได้เร็วขึ้น

2.4.5 เทคนิคในการ Normalization

การ Normalize เป็นวิธีการแปลงข้อมูลให้อยู่ในช่วงหนึ่ง ๆ ช่วยทำให้ค่าในแอทริบิวมีขอบเขตที่ไม่กว้าง และหลากหลายเกินไป กระบวนการในการ Normalize มีหลากหลายวิธี ดังนี้

1. Min-max normalization:

เป็นการเปลี่ยนช่วงของข้อมูลให้แสดงเป็นแบบ linear สูตรในการคำนวณหาค่าของข้อมูล ที่มีค่า $\min_A$ และ $\max_A$ เป็นค่าที่น้อยที่สุดและค่าที่มากที่สุดในแอทริบิวนั้น โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A \qquad (2.1)$$

ตัวอย่างเช่น มีค่า Minimum และ Maximum ของแอทริบิว income เป็น 12000$ และ 98000$ ตามลำดับ และมีค่า income เป็น 73600$ และเราต้องการแปลงค่า income ให้อยู่ในช่วง [0.0, 1.0] โดยใช้ Min-Max Normalization สามารถคำนวณค่า income ใหม่ (v') ได้จากสูตร

$$v' = \frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716 \qquad (2.2)$$

2. Z-score normalization:

หรือ Zero-mean Normalize การทำ Z-score Normalization จะเหมาะกับเหตุการณ์ที่เราไม่สามารถรู้ค่า min – max ที่แท้จริงได้โดยค่าที่ได้สำหรับแอทริบิว A เป็นค่ากึ่งกลางและค่าเบี่ยงเบนมาตรฐานของ A กำหนดให้ A' เป็นค่าเฉลี่ย และ σ_A เป็นค่าเบี่ยงเบนมาตรฐาน โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v - \bar{A}}{\sigma_A} \qquad (2.3)$$

ตัวอย่างเช่น ให้ค่าเฉลี่ย (mean) ของแอทริบิว income = 54000$, Standard Deviation = 16000$, income = 73600$ จะสามารถคำนวณค่า income ใหม่ v' ได้จากสูตร

$$v' = \frac{73{,}600 - 54{,}000}{16{,}000} = 1.225 \qquad (2.4)$$

3. Normalization by decimal scaling:

เป็นการเติมจุดทศนิยมให้กับแอทริบิว A ซึ่งขึ้นอยู่กับค่าสูงสุดของ $|A|$ โดยการแปลงค่าเดิม v ให้เป็นค่าที่อยู่ในขอบเขต v'

$$v' = \frac{v}{10^j} \tag{2.5}$$

*Where j is the smallest integer such that $Max(|A'|) < 1$.

2.4.6 การหาฟิลด์ที่เหมาะสม (Gain Attribute)

ID3 Algorithm

เป็นอัลกอริธึมพื้นฐานที่ใช้ในการสร้างการตัดสินใจแบบโครงสร้างต้นไม้ ที่ใช้หลักการของการใช้ทฤษฎีข่าวสาร(Information Theory) และค่าที่วัดได้จะนำมาใช้ตัดสินใจว่าจะใช้ตัวแปรใดในการทำนาย หรือแบ่งประเภทของข้อมูล โดยมีชุดข้อมูลตัวอย่างที่ใช้ในการเรียนรู้ (Training Sample) และมีตัวแปรเป้าหมาย (Target Attribute) ซึ่งเป็นตัวแปรที่นำค่าไปใช้ในการทำนายผลในโครงสร้างต้นไม้ และแอททริบิวต์ (Attributes) ซึ่งตัวแปรอื่นๆ ที่ใช้ในการสร้างโหนดในต้นไม้ และไม่ใช่ตัวแปรเป้าหมาย (Target Attribute) สูตรที่ใช้คือ

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2.6}$$

บทที่ 3

การวิเคราะห์ระบบ

3.1 ระบบงานของการเตรียมข้อมูลและการสำรวจ สำหรับการทำดาต้าไมนิ่ง

1. การเลือกข้อมูล (Data Selection) เป็นขั้นตอนแรกในเตรียมข้อมูลสำหรับการทำดาต้าไมนิ่ง ซึ่งผู้ใช้สามารถเลือกตารางที่ต้องการ และเลือกฟิลด์ที่ต้องการได้

2. การแก้ไขข้อมูล (Data Cleansing) เป็นขั้นตอนที่สองสำหรับการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่ง ซึ่งจะให้ผู้ใช้ทำการแก้ไขค่าว่างโดยวิธีต่างๆ เพื่อให้ข้อมูลที่จะนำเข้าการทำดาต้าไมนิ่งนั้นมีประสิทธิภาพ

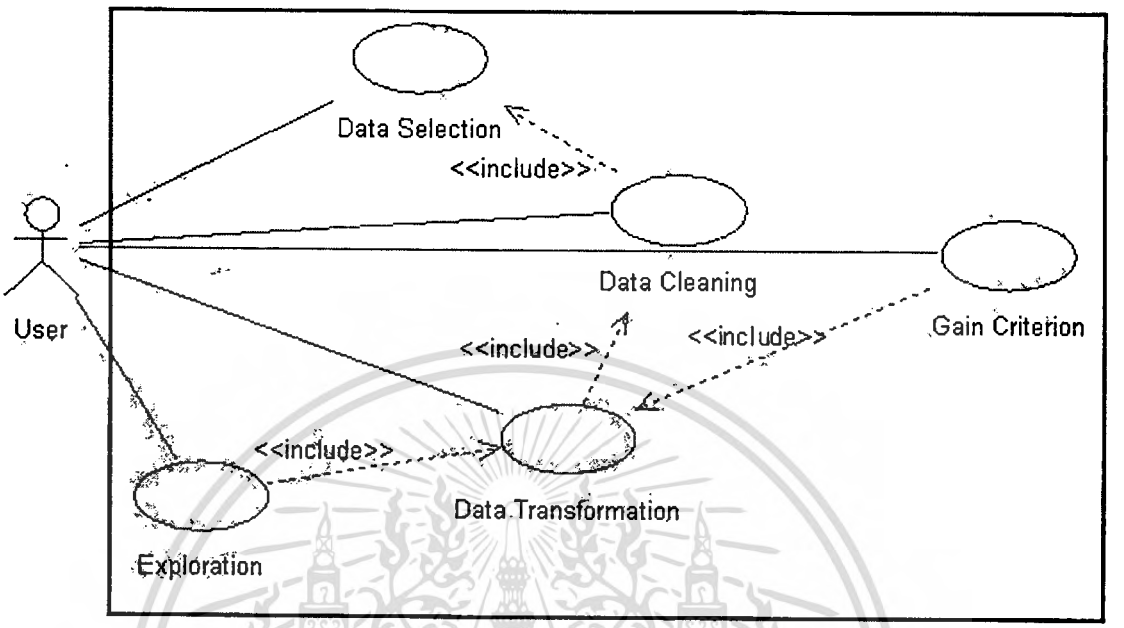
3. การปรับเปลี่ยนข้อมูล (Data Transformation) เป็นขั้นตอนสุดท้ายสำหรับการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่ง จะเป็นการปรับเปลี่ยนข้อมูลของ Numerical ให้อยู่ในช่วงๆหนึ่ง โดยผู้ใช้จะทำการกำหนดเองว่าต้องการปรับให้ข้อมูล Numerical อยู่ในช่วงใด เพื่อให้ผลลัพธ์ข้อมูลระหว่าง ฟิลด์ต่างๆ หลังจากประมวลผลดาต้าไมนิ่งแล้วค่าจะไม่ต่างกันเกินไป

4. การสำรวจข้อมูล (Data Exploration) เป็นการสำรวจข้อมูลขั้นสุดท้าย ก่อนที่จะนำเข้าการทำดาต้าไมนิ่งแสดงผลในรูปแบบของกราฟแท่ง กราฟวงกลม

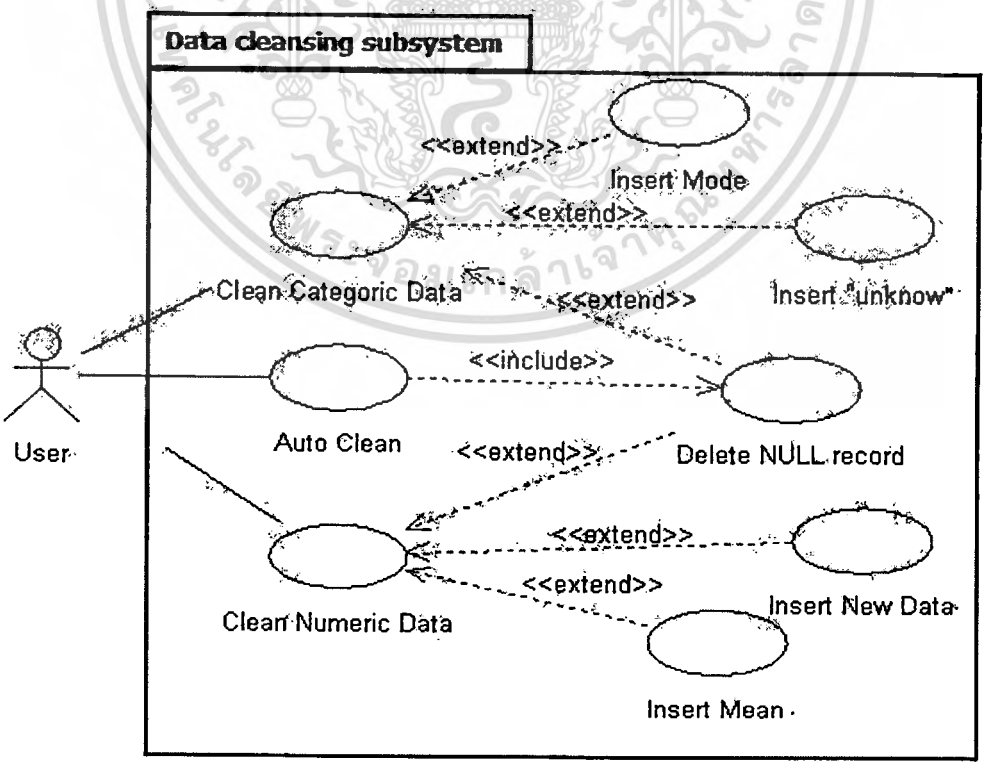
5. การหาฟิลด์ที่เหมาะสม (Gain Attribute) เป็นวิธีการสำหรับหาฟิลด์ที่เหมาะสมที่มีความเกี่ยวข้องกับข้อมูลที่ต้องการมากที่สุด

3.2 วิเคราะห์ออกแบบระบบโดยใช้แบบจำลองยูเอ็มแอล

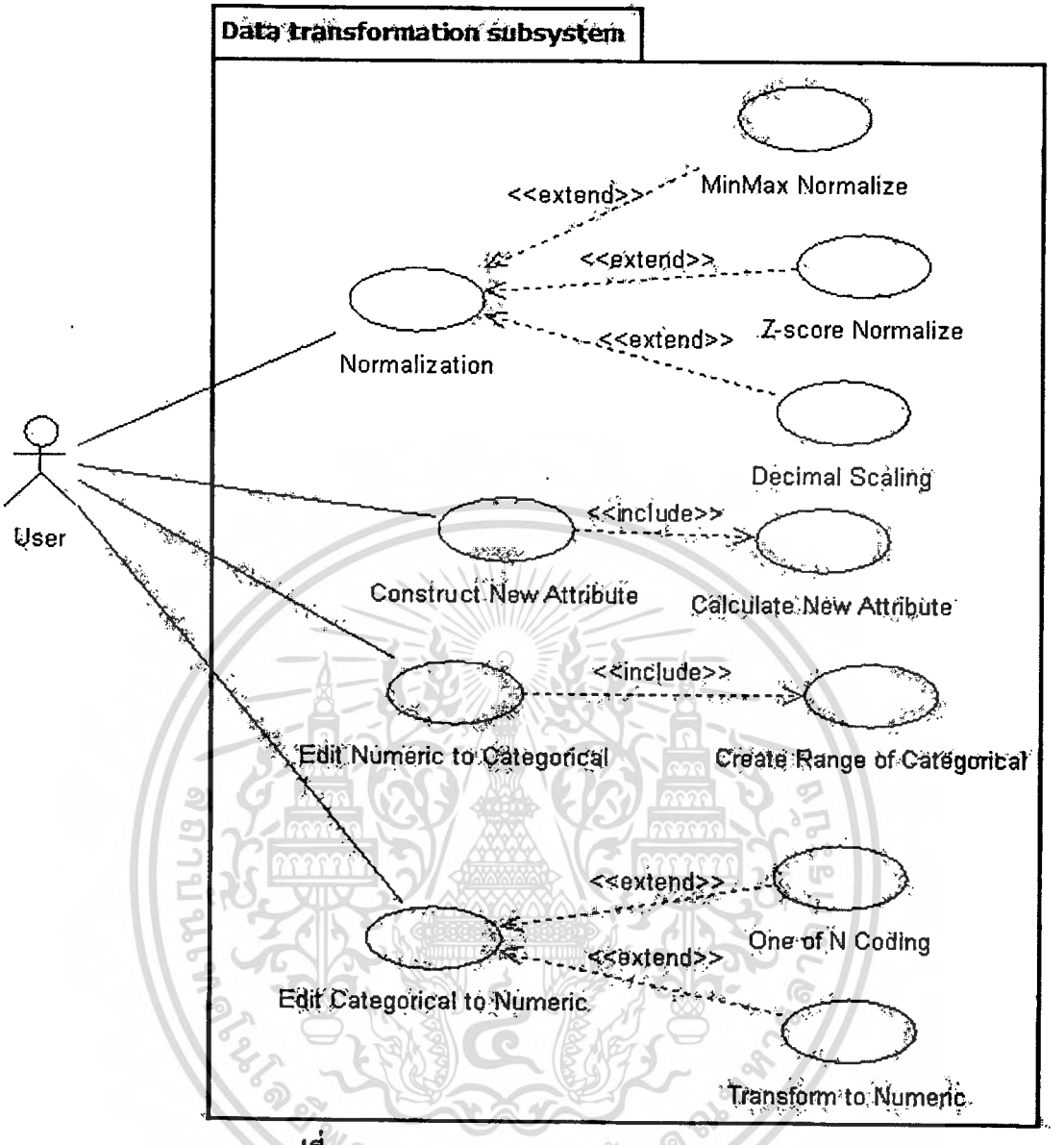
ยูสเคสไดอะแกรม (Use – Case Diagram)



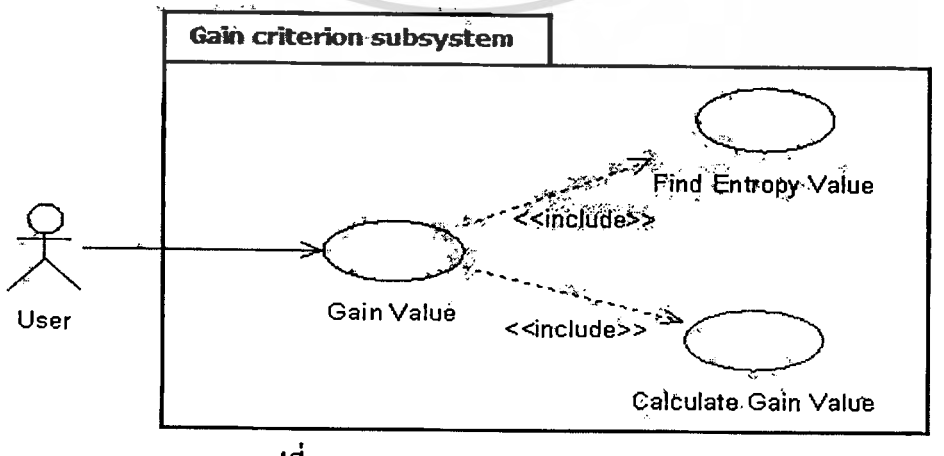
รูปที่ 3.2.1 Use case: Data Preparation System



รูปที่ 3.2.2 Use case: Data Cleansing



รูปที่ 3.2.3 Use case: Data Transformation



รูปที่ 3.2.4 Use case: Gain Criterion

Use-Case Description

ตารางที่ 3.1 คำอธิบายยูสเคส Data Selection

Project :	Data Preparation System
Use Case Name :	Data Selection
Actors :	User
Use Case Reference :	
Abstract :	ผู้ใช้ทำการเลือกข้อมูลที่ต้องการจะเตรียมสำหรับการทำเหมืองข้อมูล
Basic Flow :	1. ผู้ใช้เลือกโหมดที่ต้องการ 2. ระบบแสดงตารางที่มีอยู่ทั้งหมดในระบบ ถ้าเลือกโหมด Basic 3. ผู้ใช้เลือกฟิลด์ที่ต้องการของแต่ละตาราง 4. ผู้ใช้สามารถเลือกเป็นโหมด advance ซึ่งสามารถเขียน SQL เองได้ 5. ผู้ใช้กดปุ่ม Execute
Alternate Flow :	5.1 ถ้าผู้ใช้พิม SQL ผิด หรือว่ามีข้อมูลอยู่ในฐานข้อมูลที่จะสร้างใหม่แล้ว ระบบจะแจ้งเตือนกลับมา "Error SQL Command"

ตารางที่ 3.2 คำอธิบายยูสเคส Data Cleansing

Project :	Data Preparation System
Use Case Name :	Data Cleansing
Actors :	User
Use Case Reference :	Data Selection
Abstract :	ผู้ใช้เลือกฟิลด์ที่ต้องการ Clean ข้อมูล แล้วทำการเลือกวิธีในการ Clean
Basic Flow :	1. ระบบจะแสดงฟิลด์ที่ได้ทำการเลือกไว้ใน Data Selection 2. ผู้ใช้เลือกฟิลด์ที่ต้องการClean 3. ระบบแสดงค่าที่คำนวณได้ เช่น Mean, Max, Min, Stdv, Missing (สำหรับข้อมูลที่เป็นตัวเลข) Mode, Missing(สำหรับข้อมูลที่เป็นตัวอักษร) 4. ผู้ใช้เลือกวิธีที่ต้องการจะ Clean 5. ผู้ใช้กดปุ่ม Clean หรือว่า Auto Clean
Alternate Flow :	

ตารางที่ 3.3 คำอธิบายยูสเคส Data Transformation

Project :	Data Preparation System
Use Case Name :	Data Transformation
Actors :	User
Use Case Reference :	Data Cleansing
Abstract :	หลังจากที่ผู้ใช้ผ่านการทำ Data Cleansing ผู้ใช้สามารถที่จะแปลงข้อมูลเป็นลักษณะข้อมูลตามที่ต้องการได้
Basic Flow :	1. ผู้ใช้เลือกวิธีการแปลงข้อมูล 2. ระบบแสดงฟิลด์ที่ผ่านการ Clean มาแล้วให้ผู้ใช้เห็น 3. ผู้ใช้ทำการแปลงข้อมูลไปตามที่ต้องการ 4. ผู้ใช้กดปุ่ม Transform เพื่อแปลงข้อมูลต่างๆ
Alternate Flow :	

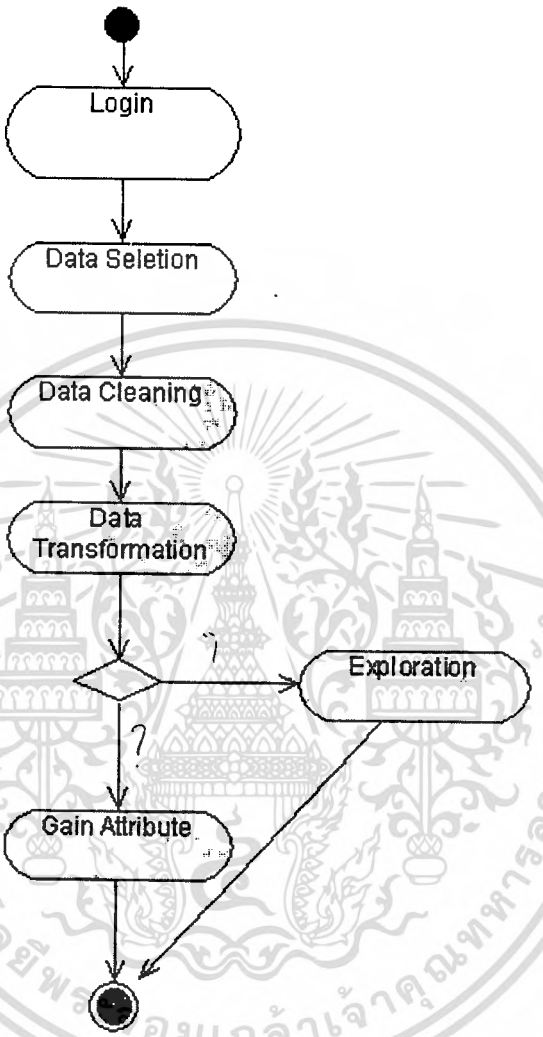
ตารางที่ 3.4 คำอธิบายยูสเคส Exploration

Project :	Data Preparation System
Use Case Name :	Exploration
Actors :	User
Use Case Reference :	Data Transformation
Abstract :	ผู้ใช้สามารถเรียกดูข้อมูลฟิลด์ต่างๆได้ หลังจากผ่านการ Transform ข้อมูลแล้ว
Basic Flow :	1. ผู้ใช้เลือกฟิลด์ที่ต้องการจะดู 2. ระบบแสดงค่าของฟิลด์ที่ได้ทำการเลือก
Alternate Flow :	

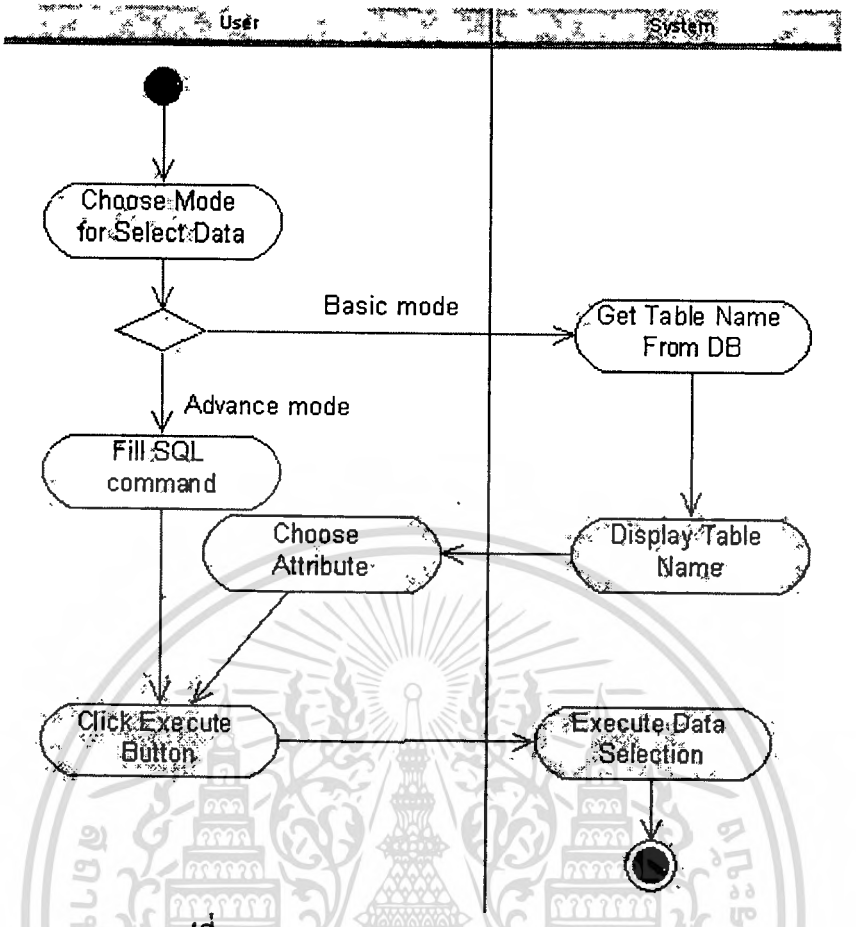
ตารางที่ 3.5 คำอธิบายยูสเคส Gain Criterion

Project :	Data Preparation System
Use Case Name :	Gain Criterion
Actors :	User
Use Case Reference :	Data Transformation
Abstract :	ผู้ใช้สามารถหาความเกี่ยวข้องระหว่างข้อมูลเป้าหมายกับฟิลด์อื่นๆได้ว่าฟิลด์ไหนมีความเกี่ยวข้องมากที่สุด
Basic Flow :	1. ผู้ใช้เลือกฟิลด์ที่ต้องการ 2. ระบบจะทำการคำนวณหาค่า Entropy 3.ผู้ใช้เลือกฟิลด์ที่ต้องการหาความสัมพันธ์กับฟิลด์ที่ต้องการ 4. ผู้ใช้กดปุ่ม Gain Value 5. ระบบจะแสดงค่าที่คำนวณได้มาแสดงโดยเรียงลำดับจากข้อมูลที่มีความสัมพันธ์มากที่สุดไปน้อยที่สุด 6. ผู้ใช้เลือกฟิลด์ที่ต้องการ นำมาสร้างตารางใหม่ แล้วกดปุ่ม Execute
Alternate Flow :	

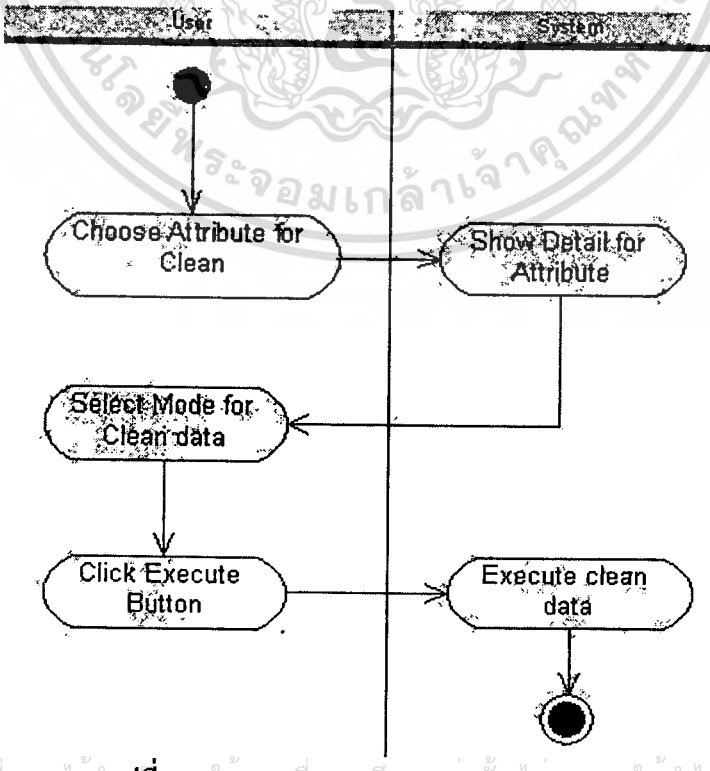
Activity Diagram



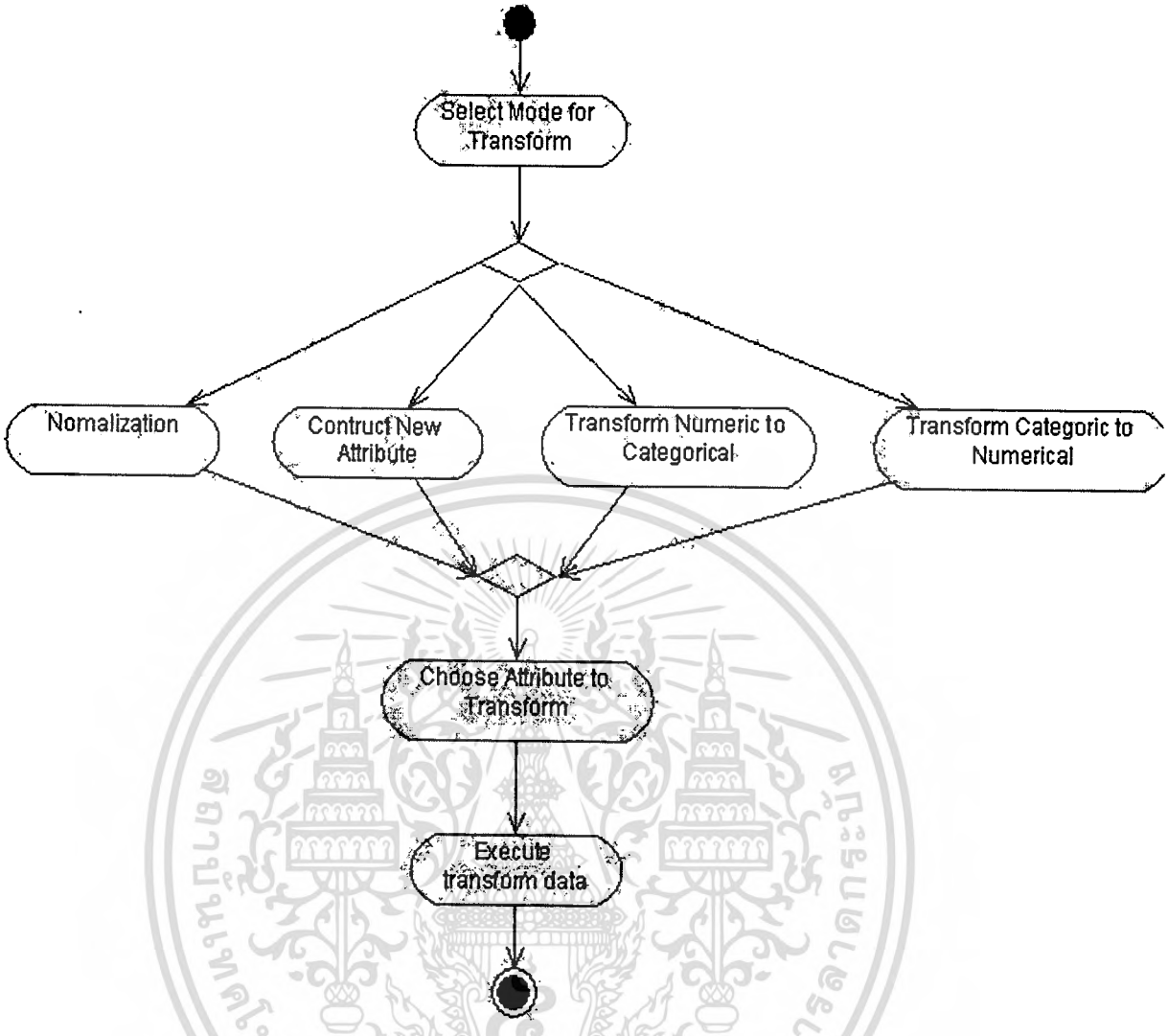
รูปที่ 3.2.5 Activity Diagram: Data Preparation System



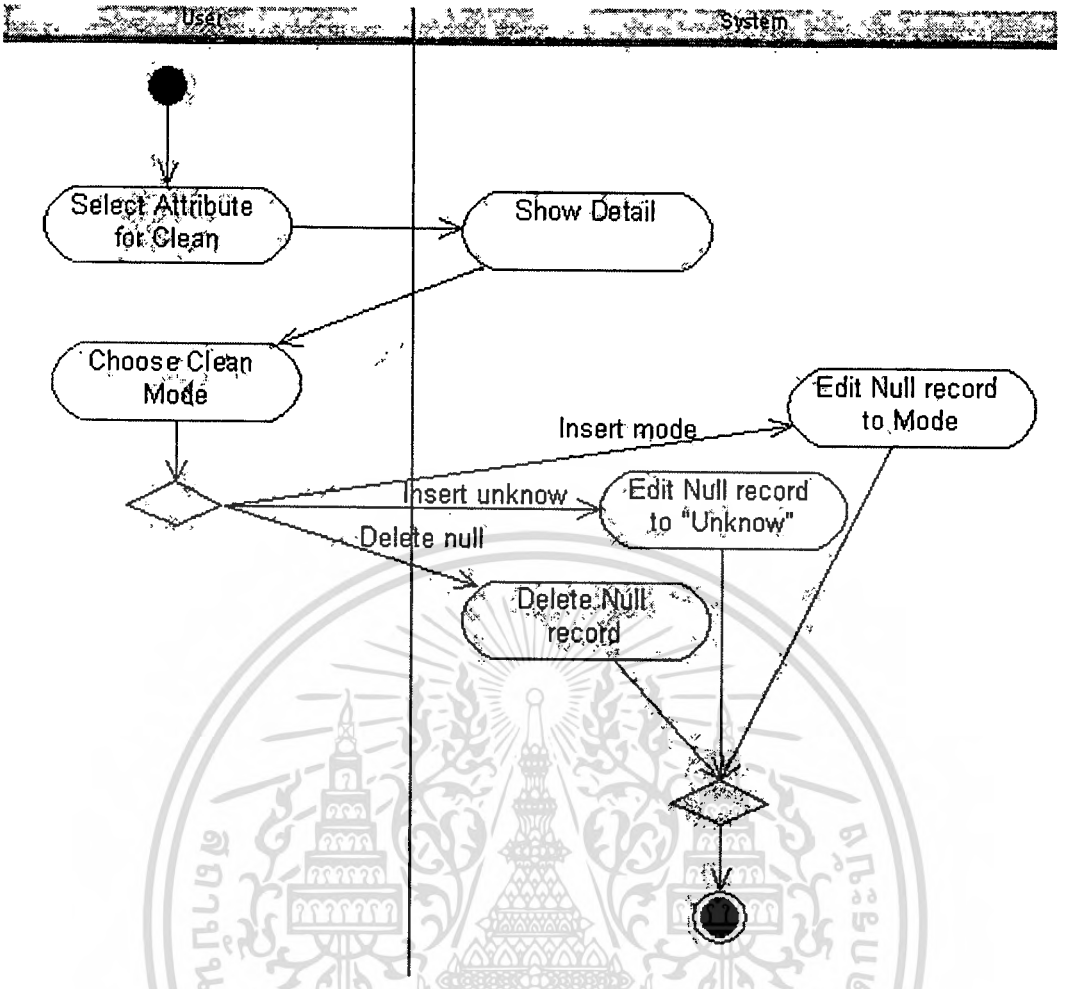
รูปที่ 3.2.6 Activity Diagram: Data Selection



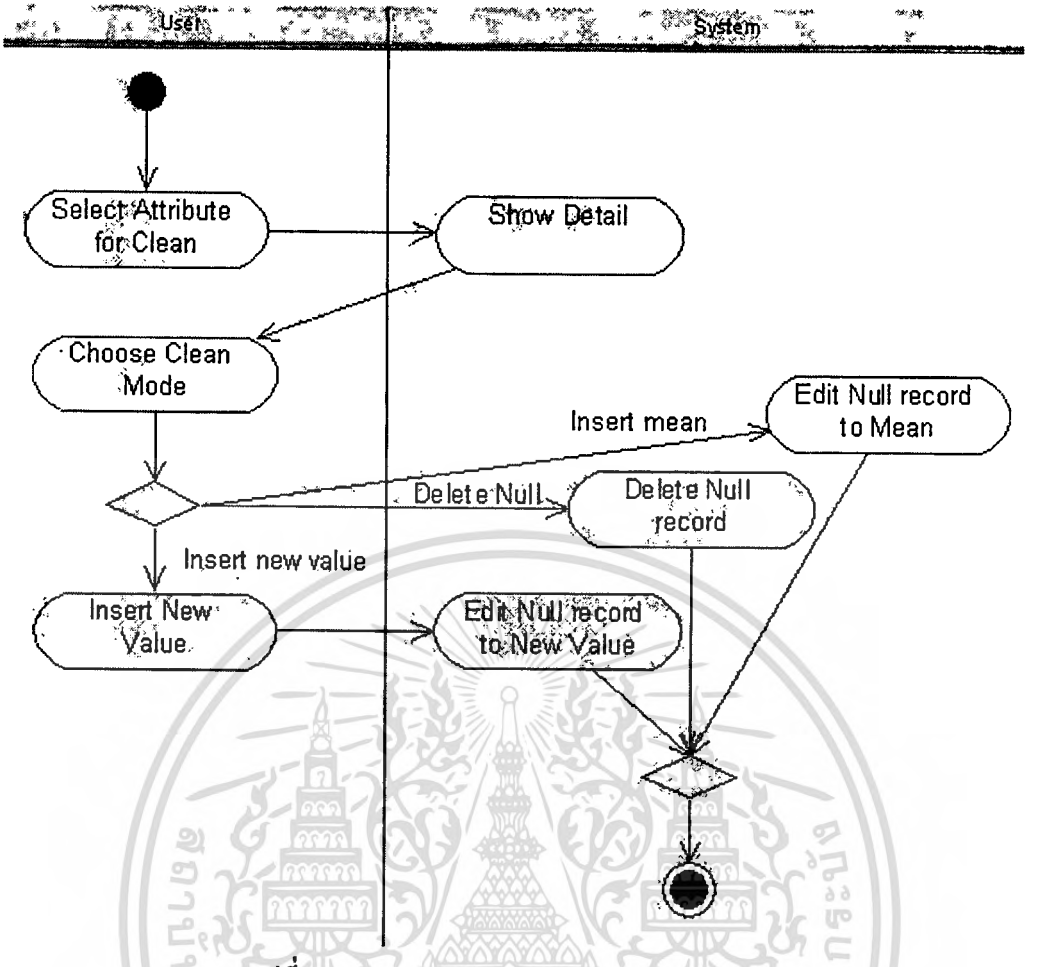
รูปที่ 3.2.7 Activity Diagram: Data Cleansing



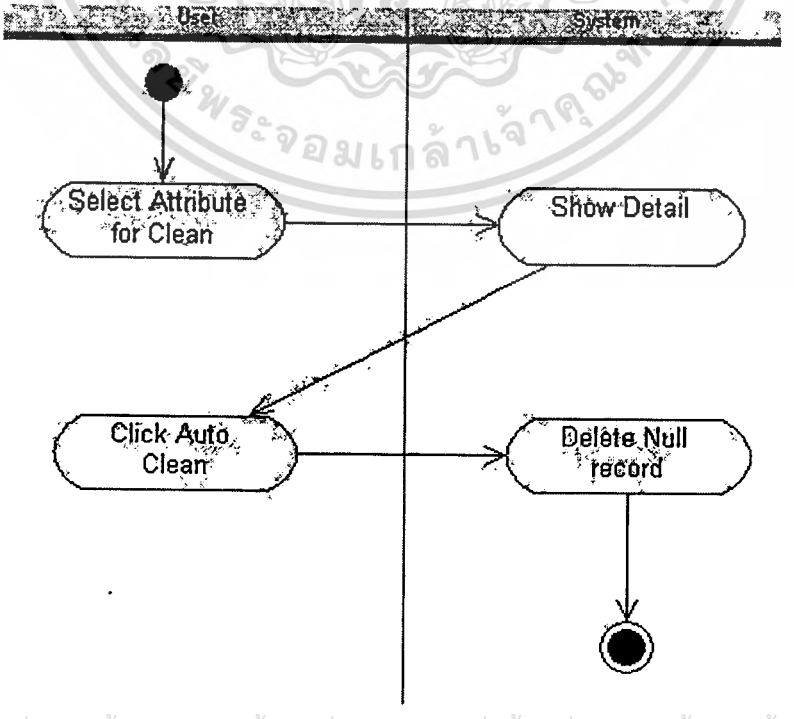
รูปที่ 3.2.8 Activity Diagram: Data Transformation



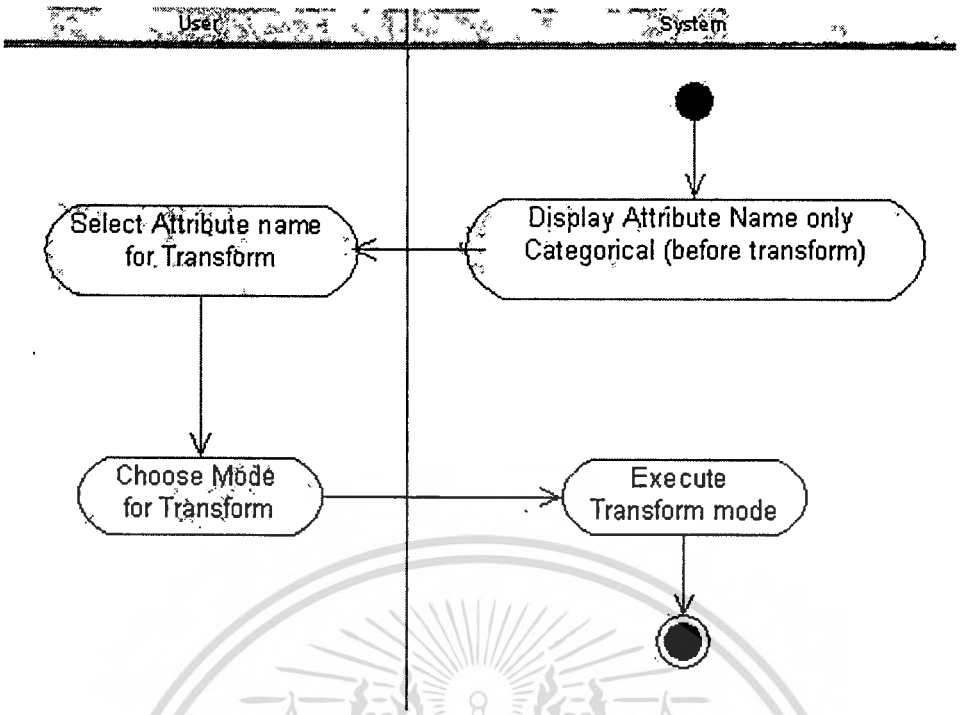
รูปที่ 3.2.9 Activity Diagram: Clean Categorical



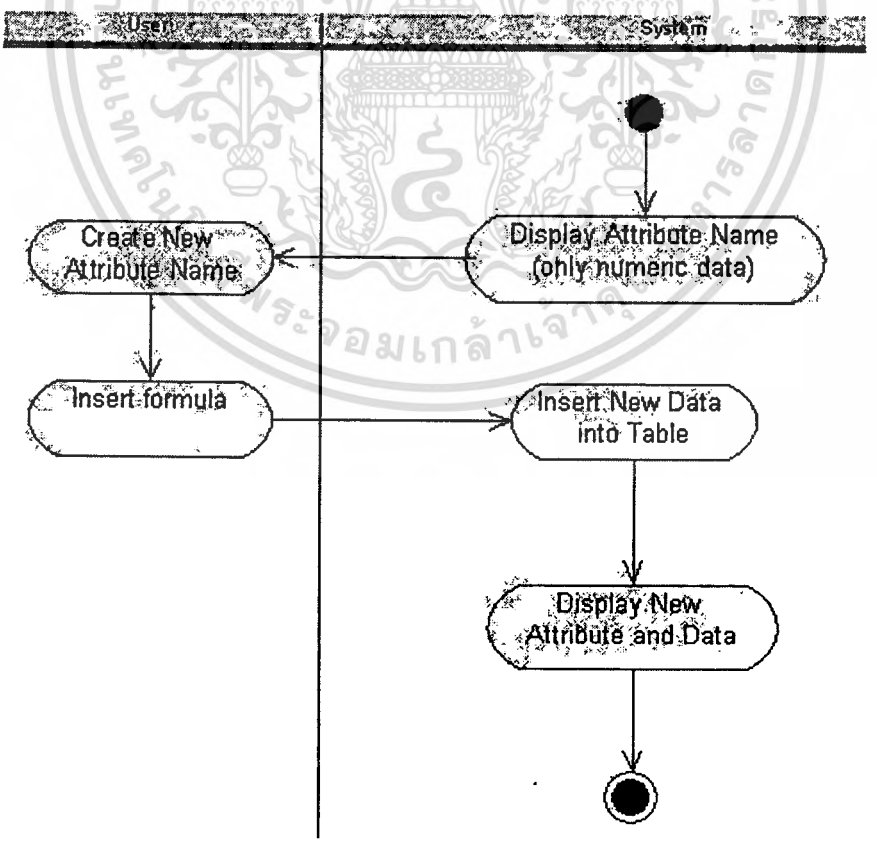
รูปที่ 3.2.10 Activity Diagram: Clean Numeric



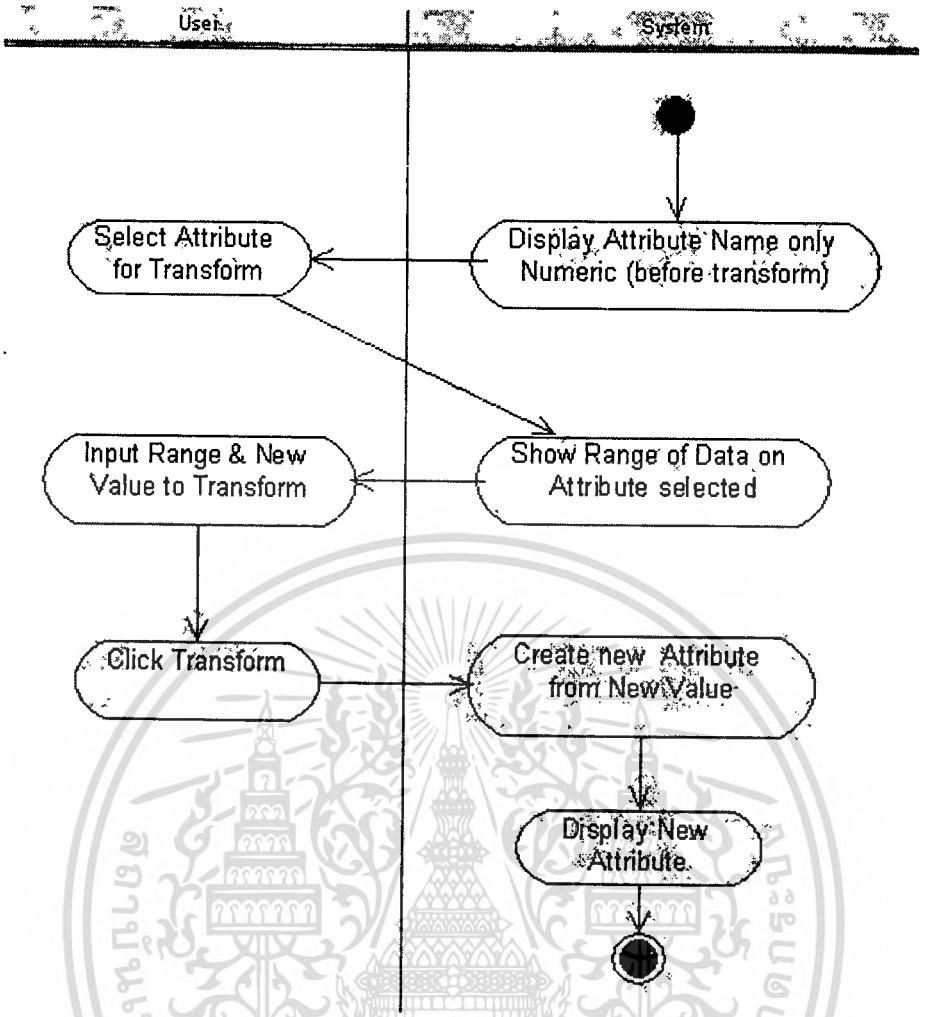
รูปที่ 3.2.11 Activity Diagram: Auto Clean



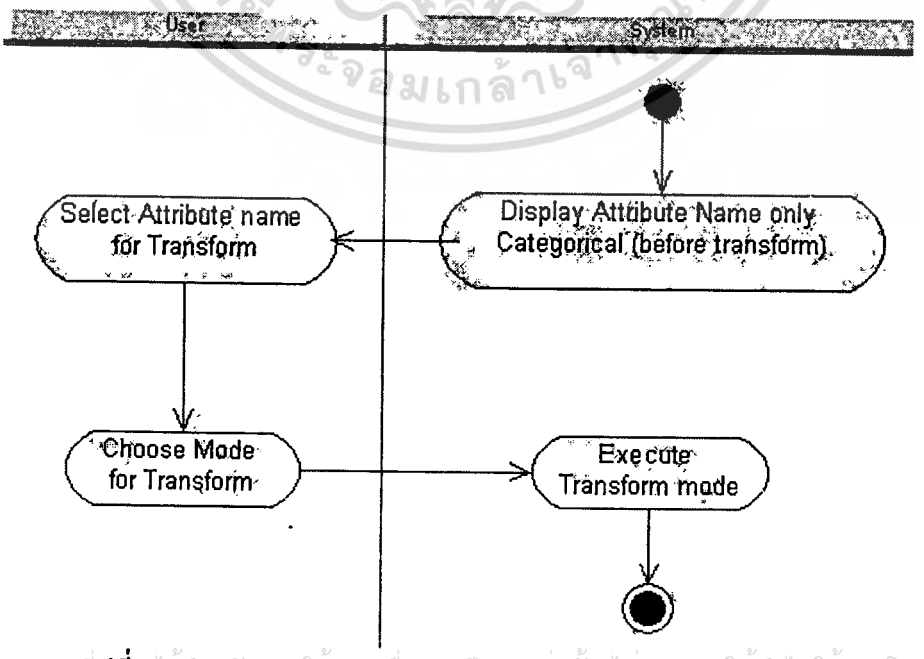
รูปที่ 3.2.12 Activity Diagram: Normalization



รูปที่ 3.2.13 Activity Diagram: Construct New Attribute



รูปที่ 3.2.14 Activity Diagram: Transform Numeric to Categorical



รูปที่ 3.2.15 Activity Diagram: Transform Categorical to Numeric

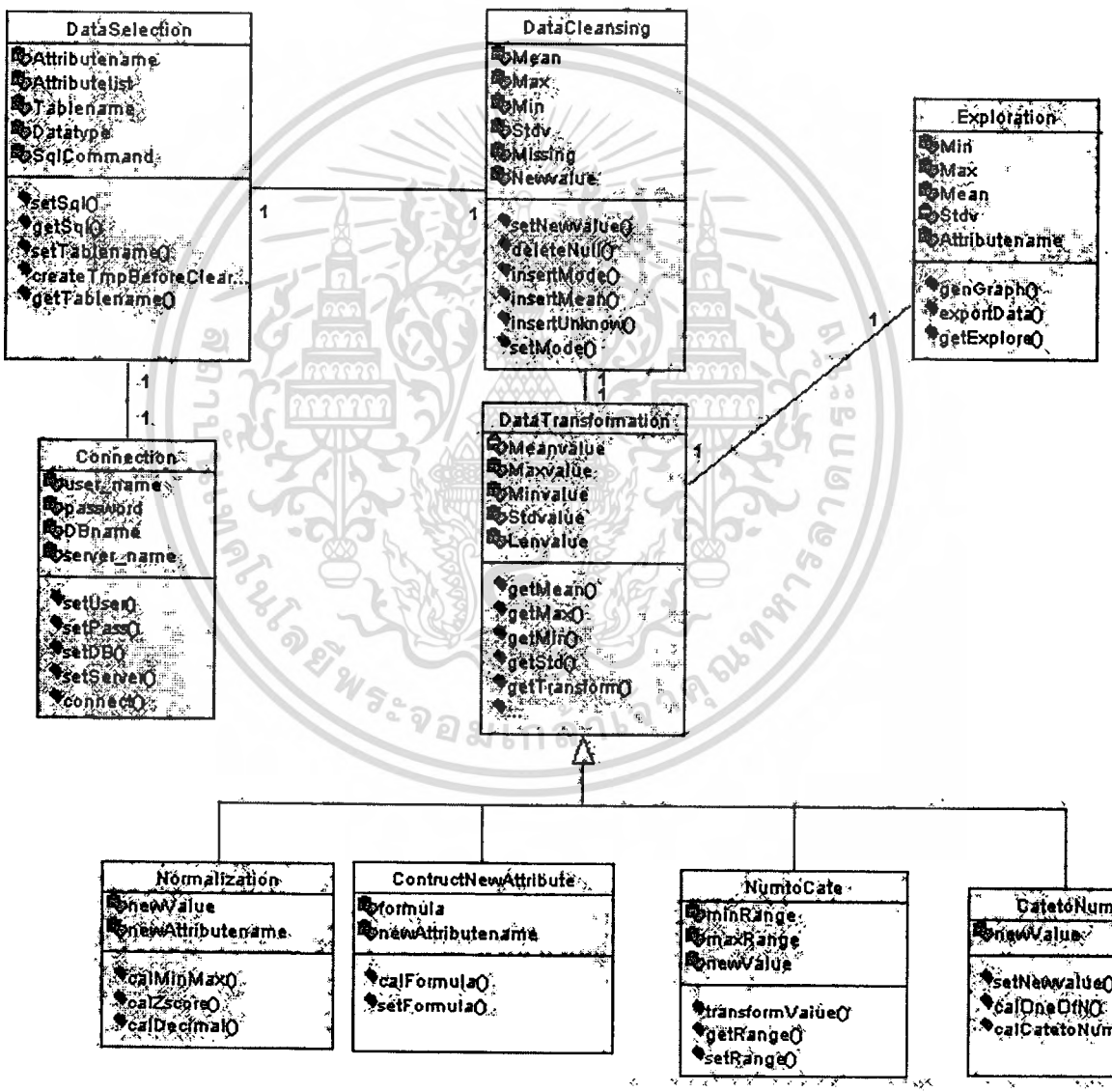
บทที่ 4

การออกแบบระบบ

4.1 การออกแบบระบบโดยใช้แบบจำลองยูเอ็มแอล

Class Diagram

รูปแสดงการทำงานของคลาสต่างๆ ที่มีในระบบ โดยมีรายละเอียดดังรูป



รูปที่ 4.1 Class Diagram

Class-Responsibility-Collaborator or CRC

ตารางที่ 4.1 CRC connection class

Class: Connection	
Description: use to connect to database server	
Responsibility:	**Collaborator:**
Connect to Database and get Table	DataSelection

ตารางที่ 4.2 CRC DataSelection class

Class: DataSelection	
Description: select attribute for data preparation	
Responsibility:	**Collaborator:**
Set sql for select attribute	
Choose attribute and create TmpBeforeClean DB	DataCleansing

ตารางที่ 4.3 CRC DataCleansing class

Class: DataCleansing	
Description: manage data was "null" to new value or delete that	
Responsibility:	**Collaborator:**
Choose attribute that have null record and cleansing	DataTransformation
Delete null record	
Insert mean, mode or unknown data represent null record	

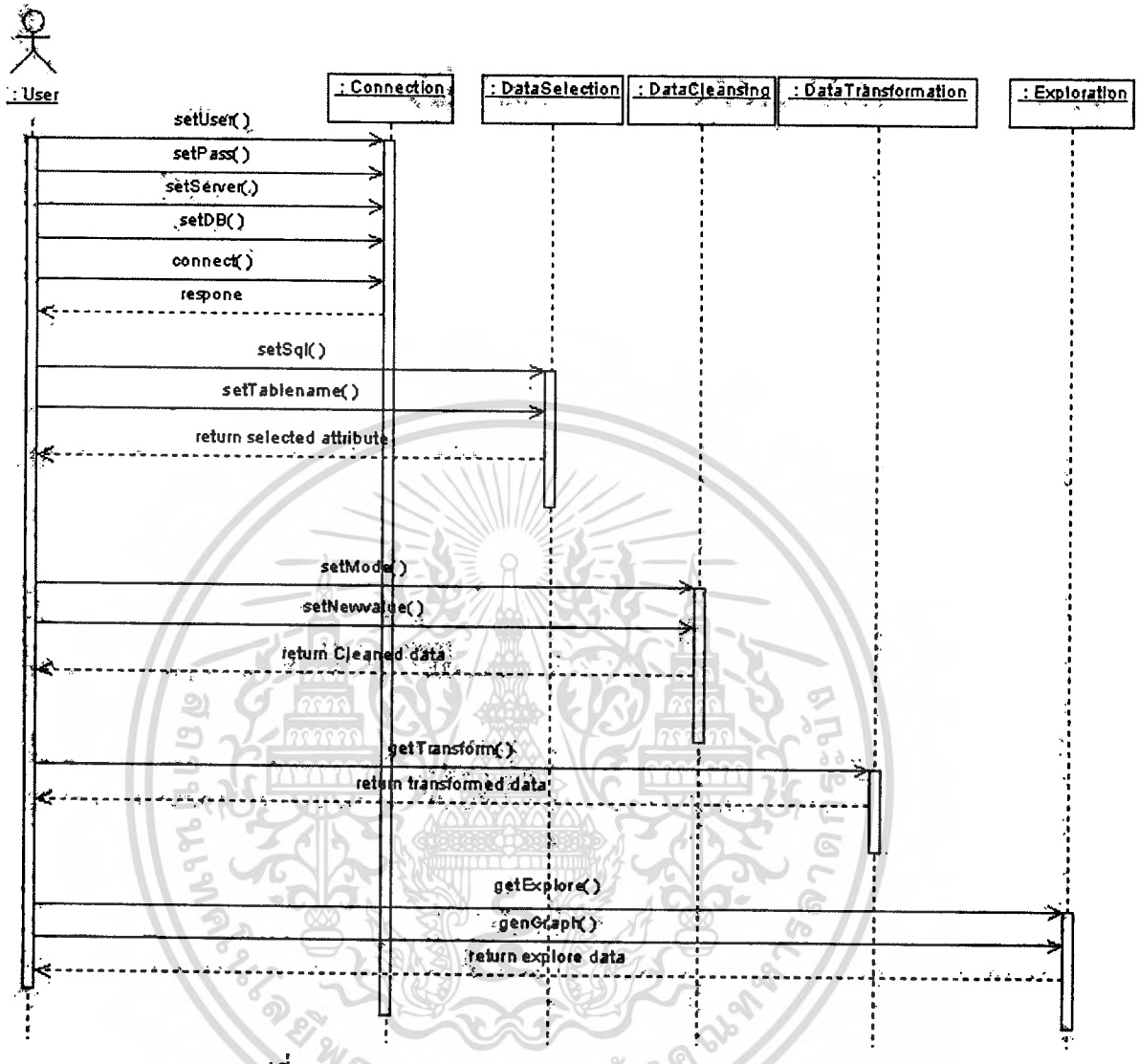
ตารางที่ 4.4 CRC DataTransformation class

Class: DataTransformation	
Description: transform data to new value	
Responsibility:	**Collaborator:**
Transform data to new value	Exploration

ตารางที่ 4.5 CRC Exploration class

Class: Exploration	
Description: explore date after transform	
Responsibility:	**Collaborator:**
Explore data	
Generate graph	
Export data	

Sequence Diagram

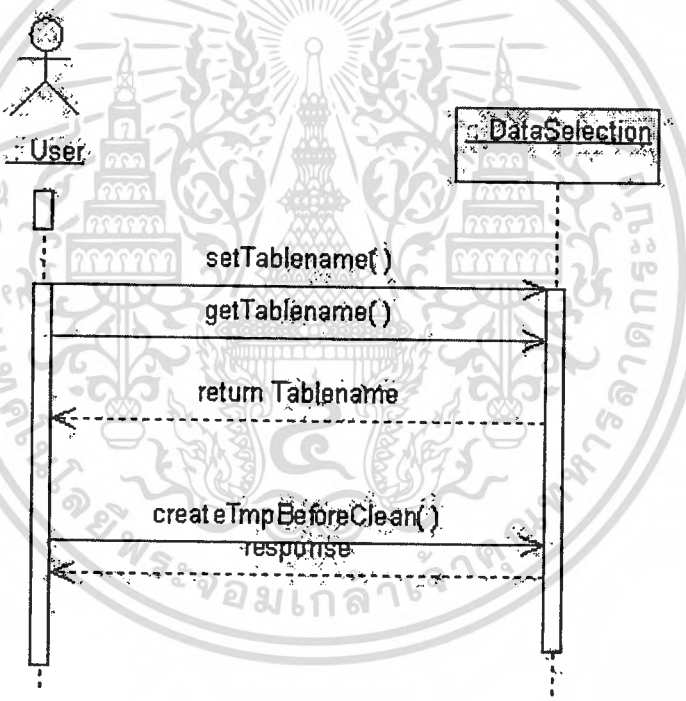


รูปที่ 4.2 Sequence diagram: Data Preparation

จากรูปที่ 4.2 สามารถอธิบายการทำงานได้ดังนี้

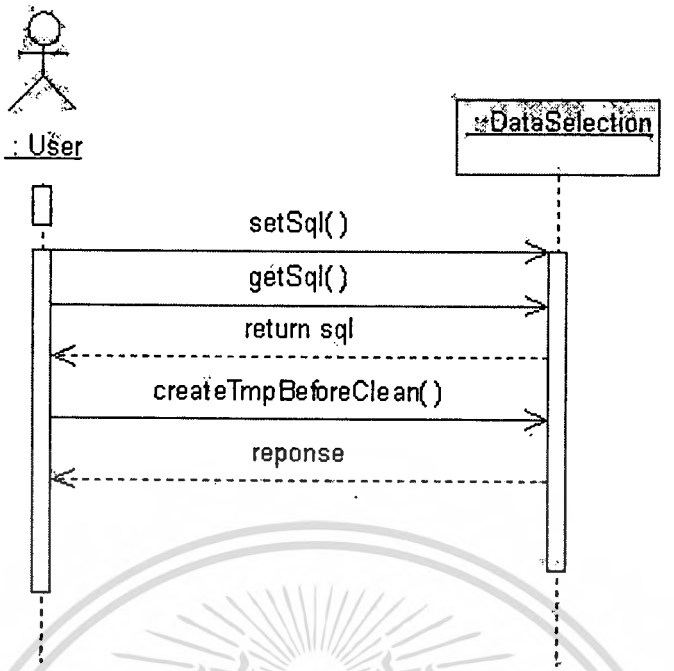
1. ผู้ใช้กำหนดชื่อเข้าใช้งานให้กับคลาสconnection
2. ผู้ใช้กำหนดพาสเวิร์ดให้กับคลาสconnection
3. ผู้ใช้กำหนดชื่อเซิฟเวอร์ให้กับคลาสconnection
4. ผู้ใช้กำหนดฐานข้อมูลที่ต้องการให้กับคลาสconnection
5. ผู้ใช้สั่งให้คลาสconnection ทำการติดต่อฐานข้อมูล
6. คลาสconnectionทำการส่งผลลัพธ์ของการติดต่อ
7. ผู้ใช้กำหนดsqlที่ใช้ในการทำงานให้กับคลาสDataSelection
8. ผู้ใช้เลือกตารางที่ต้องการแล้วกำหนดให้กับคลาสDataSelection

9. คลาสDataSelectionคืนค่าแอททริบิวต์ที่ได้ทำการเลือก
10. ผู้ใช้ทำการเลือกวิธีการที่ต้องคลีนข้อมูลให้กับคลาสDataCleansing
11. ผู้ใช้กำหนดค่าใหม่ที่ต้องการให้กับคลาสDataCleansing
12. คลาสDataCleansingคืนค่าข้อมูลที่ผ่านการคลีนแล้ว
13. ผู้ใช้ร้องขอค่าที่ผ่านการแปลงข้อมูลกับคลาสDataTransformation
14. คลาสDataTransformationส่งผลลัพธ์ที่ผ่านการแปลงข้อมูลแล้ว
15. ผู้ใช้ร้องขอให้แสดงข้อมูลกับคลาสExploration
16. ผู้ใช้ร้องขอให้ทำการสร้างกราฟกับคลาสExploration
17. คลาสExplorationส่งผลลัพธ์ข้อมูลที่ต้องการ



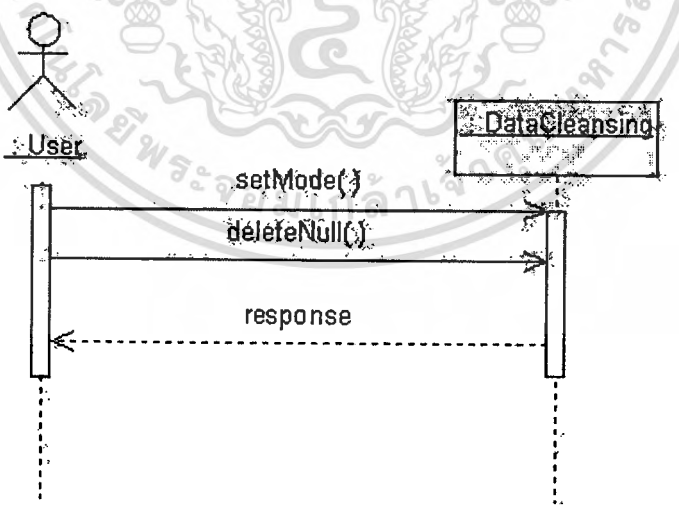
รูปที่ 4.3 Sequence Diagram: Data Selection basic mode

จากรูปที่ 4.3 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาสDataSelection ซึ่งเป็นกระบวนการเลือกข้อมูลแบบทั่วไป



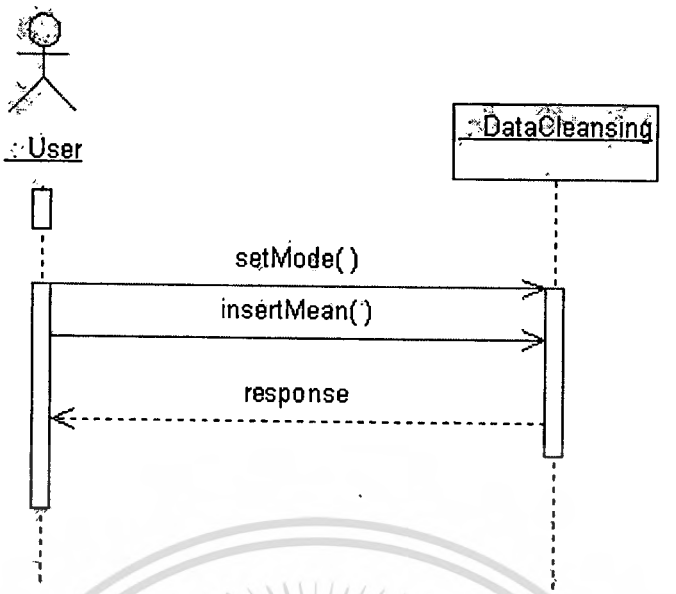
รูปที่ 4.4 Sequence Diagram: Data Selection advance mode

จากรูปที่ 4.4 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาสDataSelection ซึ่งเป็นกระบวนการเลือกข้อมูลแบบเชี่ยวชาญ



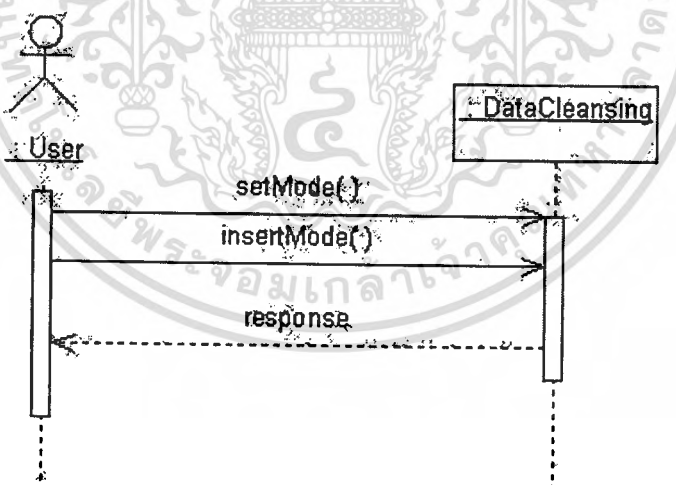
รูปที่ 4.5 Sequence Diagram: Data Cleansing delete null

จากรูปที่ 4.5 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataCleansing ซึ่งเป็นกระบวนการคลีนข้อมูลด้วยวิธีการลบข้อมูลที่มีค่าเป็น null ทิ้งไป



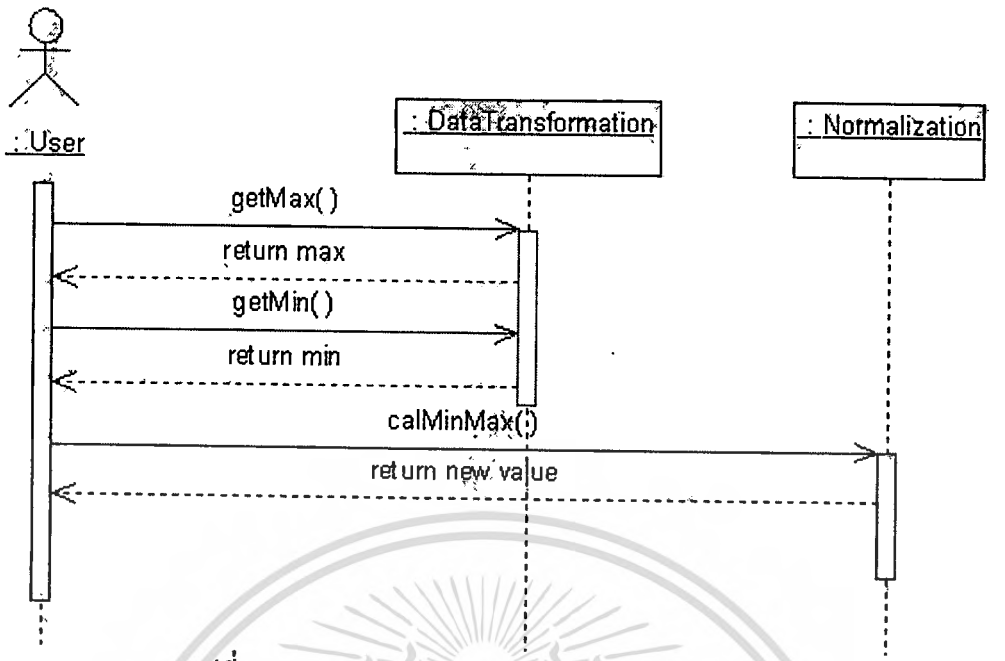
รูปที่ 4.6 Sequence Diagram: Data Cleansing insert mean

จากรูปที่ 4.6 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataCleansing ซึ่งเป็นกระบวนการคลีนข้อมูลด้วยวิธีการจัดการกับข้อมูลที่มีค่าเป็นnullด้วยค่าmean



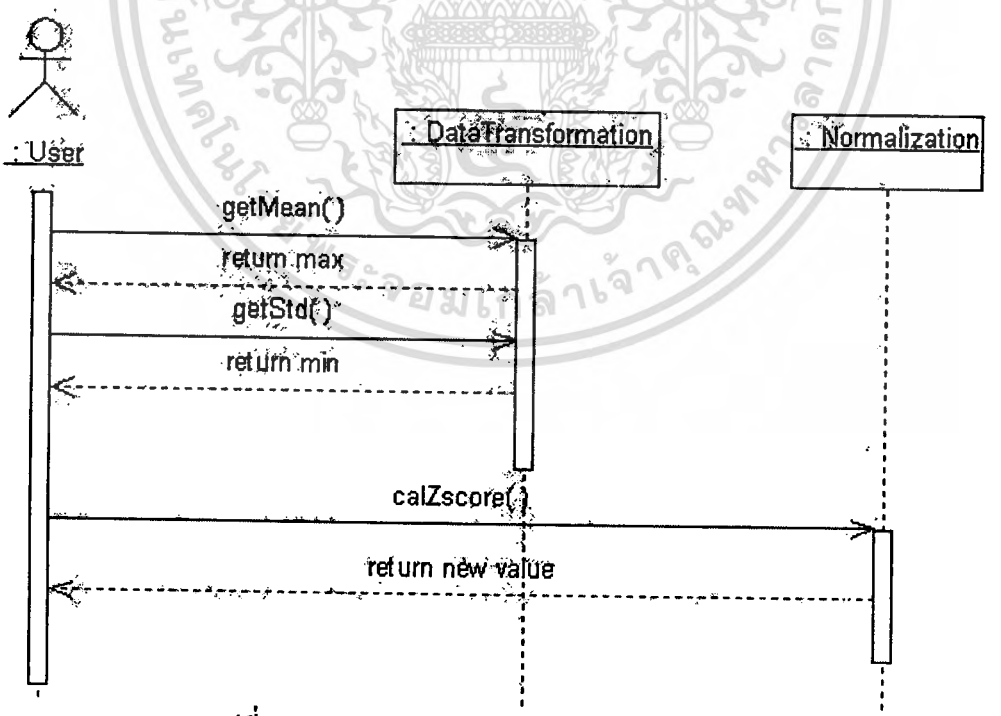
รูปที่ 4.7 Sequence Diagram: Data Cleansing insert mode

จากรูปที่ 4.7 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataCleansing ซึ่งเป็นกระบวนการคลีนข้อมูล ด้วยวิธีการจัดการกับข้อมูลที่มีค่าเป็น null ด้วยค่า mean



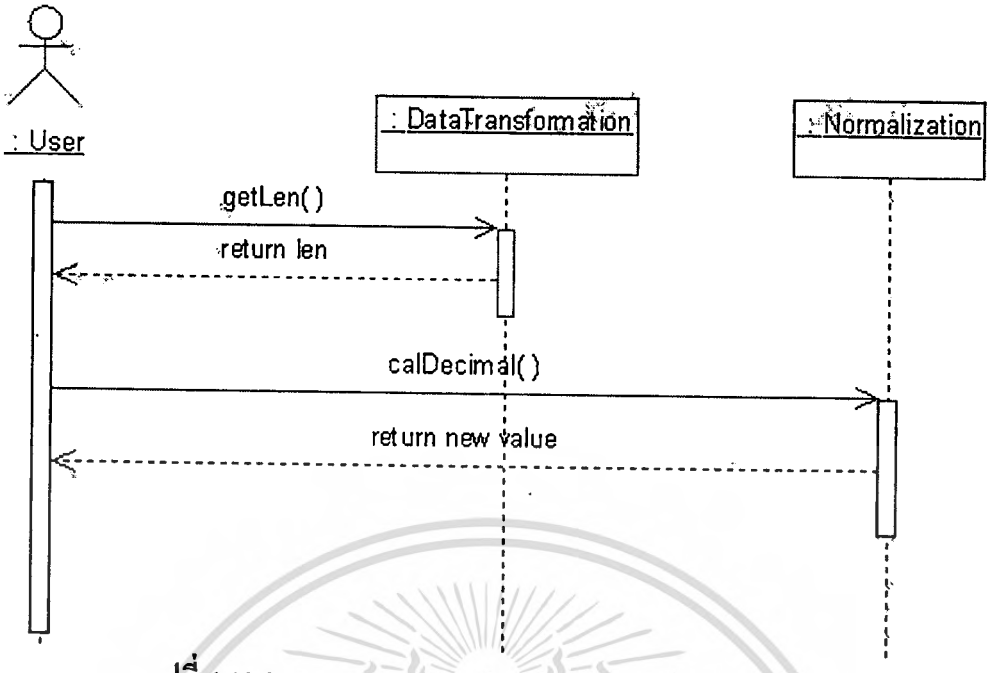
รูปที่ 4.8 Sequence Diagram: min-max normalization

จากรูปที่ 4.8 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส Normalization ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการ Min-max normalize



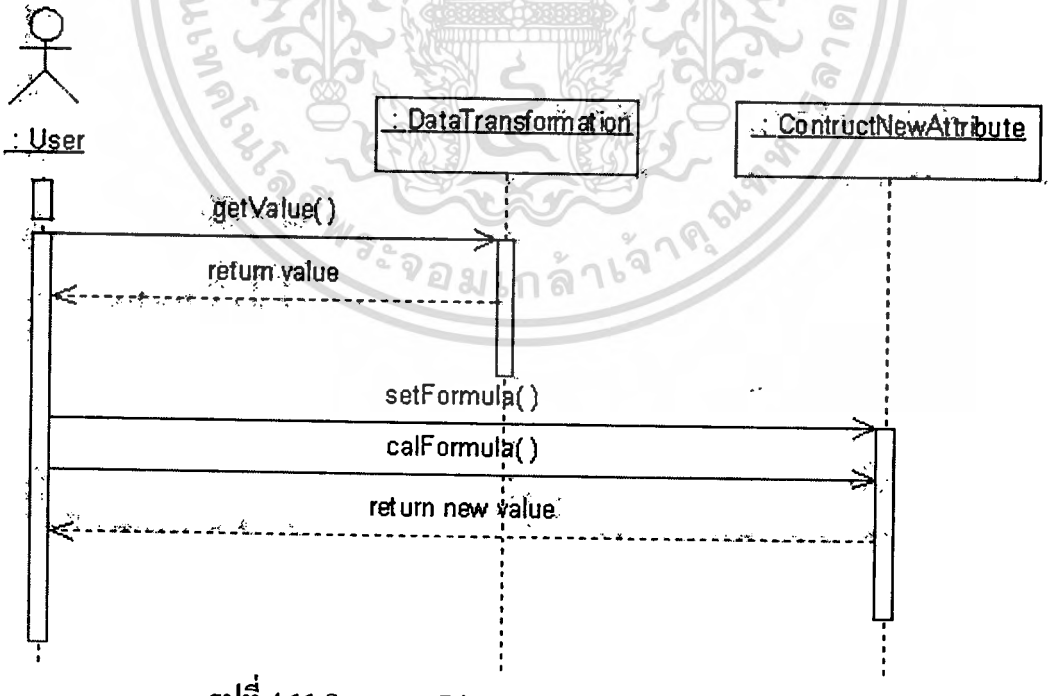
รูปที่ 4.9 Sequence Diagram: zscore normalization

จากรูปที่ 4.9 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส Normalization ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการ zscore normalize



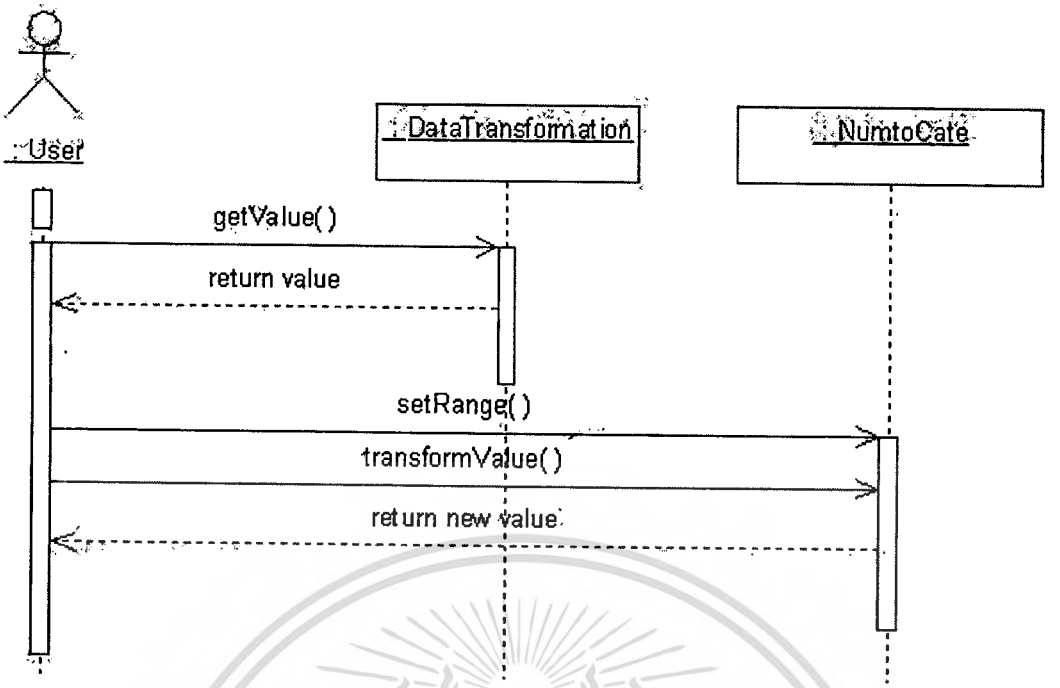
รูปที่ 4.10 Sequence Diagram: decimal scaling normalization

จากรูปที่ 4.10 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส Normalization ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการ decimal scaling normalize



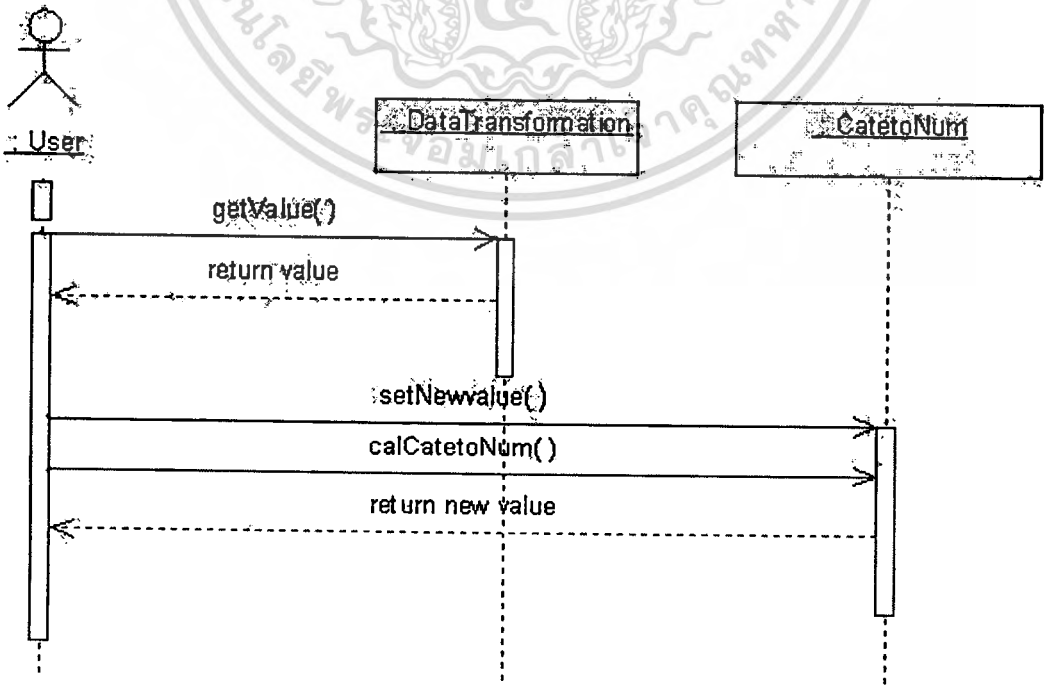
รูปที่ 4.11 Sequence Diagram: construct new attribute

จากรูปที่ 4.11 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส ConstructNewAttribute ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการสร้างแอททริบิวต์ขึ้นใหม่



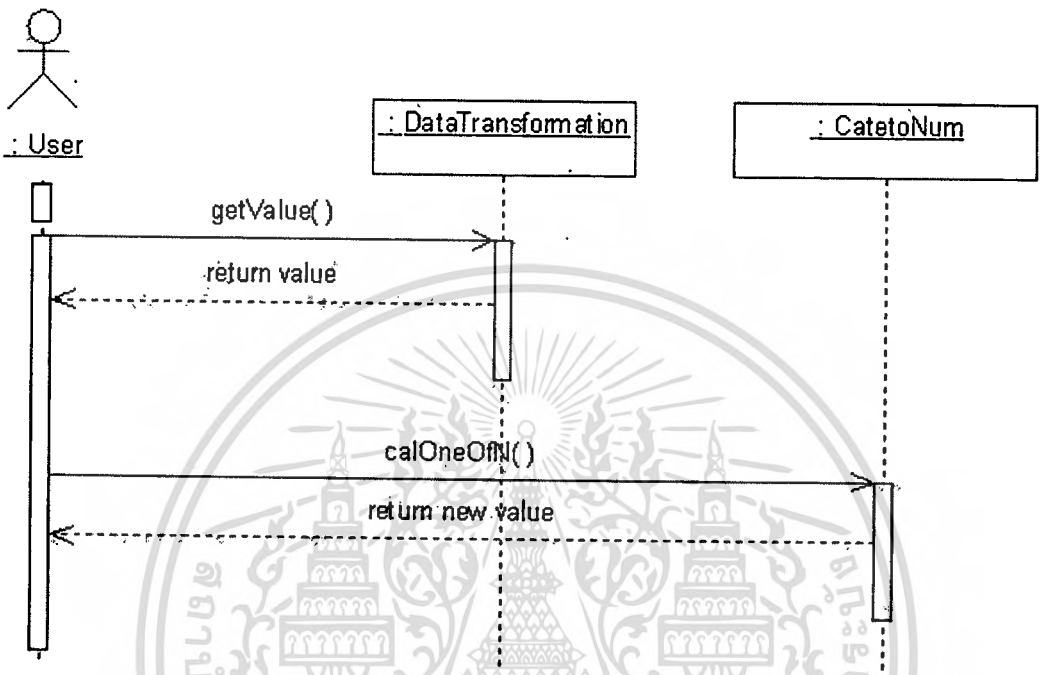
รูปที่ 4.12 Sequence Diagram: numeric to category

จากรูปที่ 4.12 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส NumtoCate ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการแปลงข้อมูลที่อยู่ในรูปแบบตัวเลขให้อยู่ในรูปแบบตัวอักษร



รูปที่ 4.13 Sequence Diagram: category to numeric (normal)

จากรูปที่ 4.13 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส CatetoNum ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการแปลงข้อมูลที่อยู่ในรูปแบบอักษรให้อยู่ในรูปแบบตัวเลข



รูปที่ 4.14 Sequence Diagram: category to numeric (one of n coding)

จากรูปที่ 4.14 เป็นกระบวนการทำงานระหว่างผู้ใช้กับคลาส DataTransformation และคลาส CatetoNum ซึ่งเป็นกระบวนการแปลงข้อมูล โดยใช้หลักการแปลงข้อมูลที่อยู่ในรูปแบบอักษรให้อยู่ในรูปแบบตัวเลขโดยใช้หลักการ One of N coding

บทที่ 5

การพัฒนาระบบ

5.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

5.1.1 Hardware

การพัฒนาระบบใช้งานเครื่องคอมพิวเตอร์ที่มีคุณสมบัติดังนี้

- CPU Core duo 1.6 GHz
- Hardisk 80 GB
- RAM 1.5 GB

5.1.2 Software

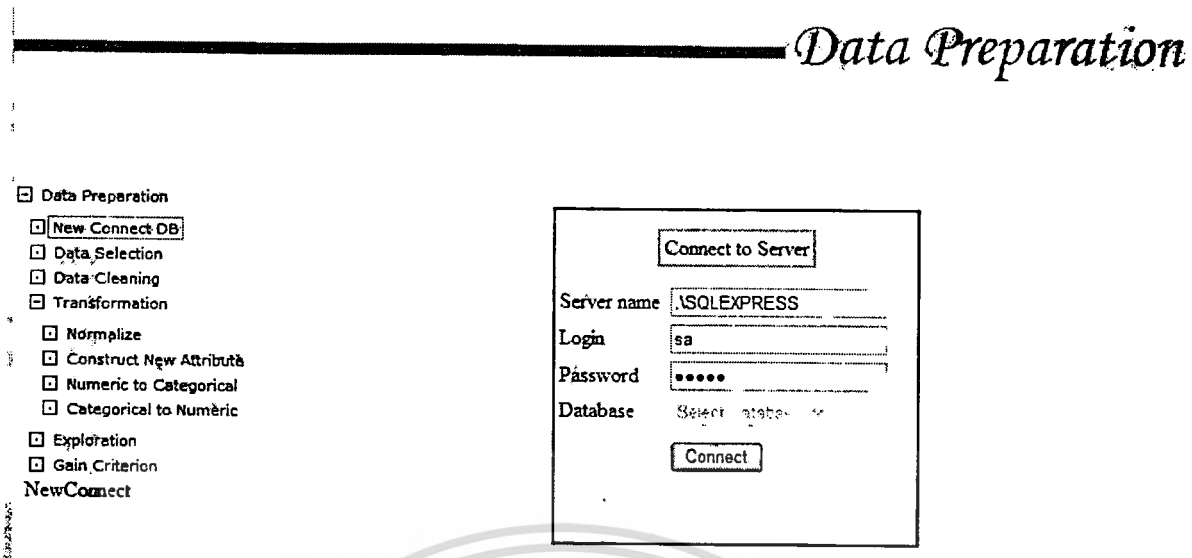
เครื่องมือ และโปรแกรมที่ใช้ในการพัฒนาระบบมีดังนี้

- OS Windows XP service pack 2
- Microsoft Visual Studio.NET 2005
- Microsoft Visual Web Developer 2008 Express Edition
- Microsoft SQL server 2005
- ComponentArt WebChart2006 for ASP.NET

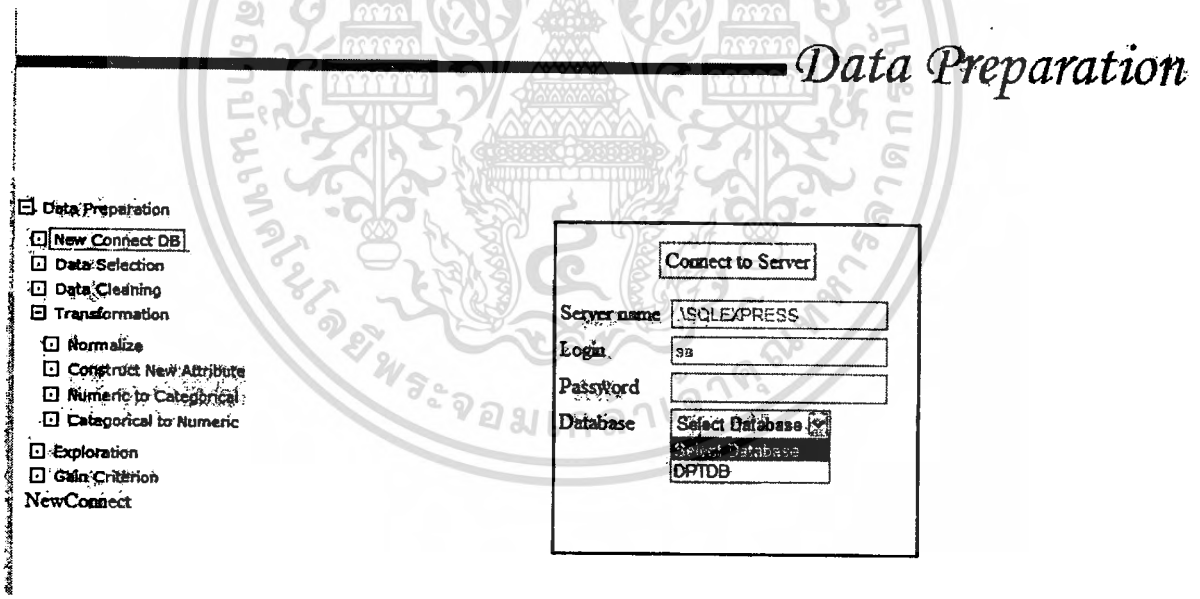
5.2 การติดต่อกับฐานข้อมูล

ขั้นตอนแรกก่อนที่จะเข้าสู่กระบวนการทำคาต้าไมนิ่งต้องทำการติดต่อกับฐานข้อมูล โดยฐานข้อมูลที่จะทำการติดต่อด้วยคือ Microsoft SQL Server 2005

1. ผู้ใช้ระบบต้องทำการกรอกข้อมูลคือ ชื่อเซิฟเวอร์ ชื่อผู้ใช้ และ พาสเวิร์ด
2. กดปุ่ม Connection เพื่อทดสอบการเชื่อมต่อกับเซิฟเวอร์
3. ทำการเลือกฐานข้อมูลที่ต้องการใช้งาน
4. กดปุ่ม New Connection เพื่อเปลี่ยนการเชื่อมต่อเซิฟเวอร์ หรือฐานข้อมูล
5. ระบบทำการเปลี่ยนหน้าเข้าสู่หน้าเลือกข้อมูล



รูปที่ 5.1 ขั้นตอนการติดต่อกับเซิฟเวอร์



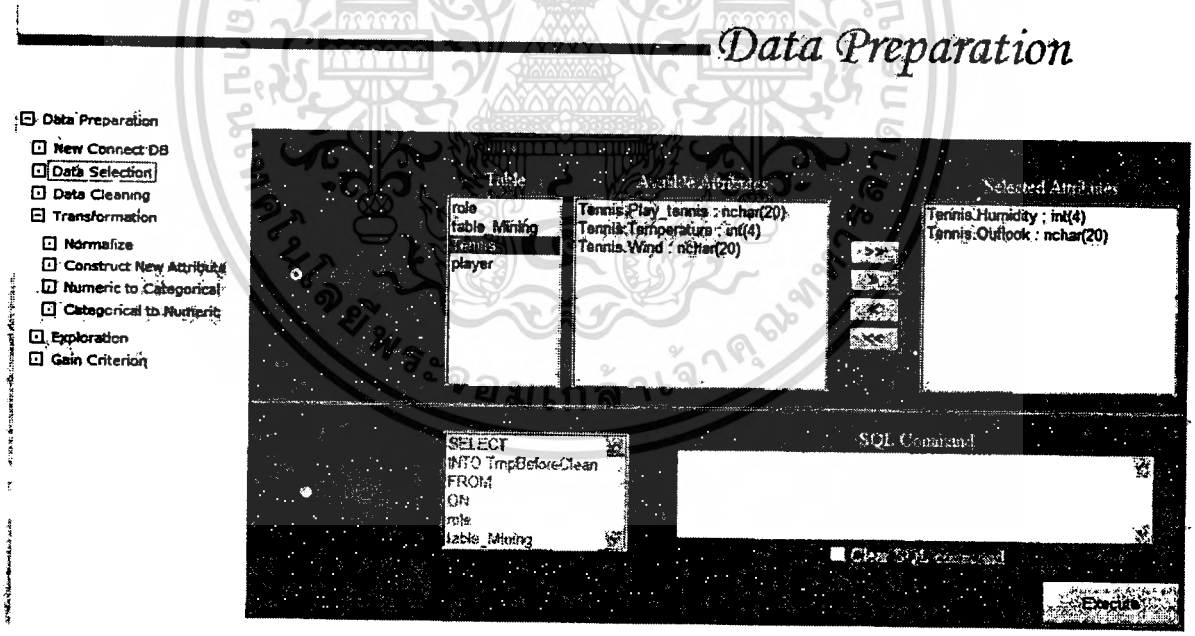
รูปที่ 5.2 ขั้นตอนการเลือกฐานข้อมูล

5.3 การเลือกข้อมูล (Data Selection)

การเลือกข้อมูลสามารถกระทำได้สองวิธีคือ

5.3.1 การเลือกข้อมูลจากหนึ่งตาราง

1. หลังจากติดต่อกับฐานข้อมูลที่เราระบุเรียบร้อยแล้ว หน้าจอระบบจะแสดงรายชื่อตารางที่อยู่ในฐานข้อมูลในช่อง Table คลิกที่ชื่อของตาราง ภายในช่อง Attributes จะแสดงรายชื่อแอตทริบิวของตารางนั้น
2. ช่อง Available Attributes ด้านซ้ายแสดงรายชื่อของแอตทริบิวจากตารางที่เลือก
3. ช่อง Selected Attributes แสดงรายชื่อแอตทริบิวที่ผู้ใช้ระบบต้องการนำไปใช้ในการทำคาต้าไมนิ่งต่อไป
4. คลิกที่ปุ่ม Execute สิ้นสุดขั้นตอนการเลือกข้อมูล เพื่อเข้าสู่การเตรียมข้อมูลในขั้นตอนต่อไป

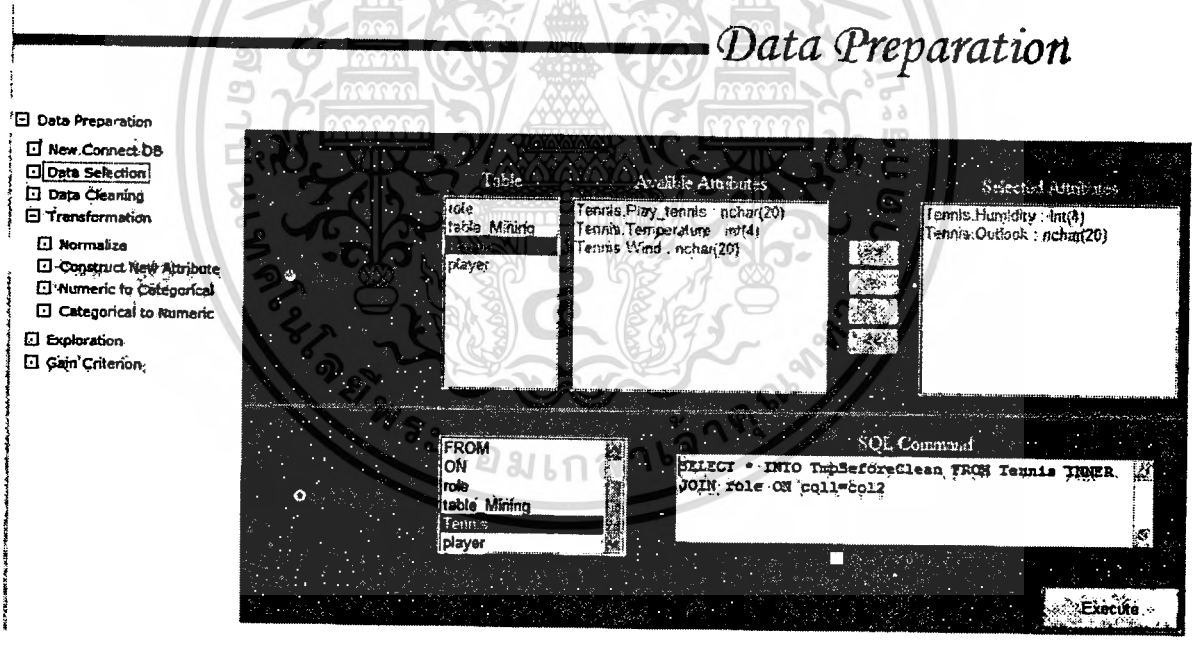


รูปที่ 5.3 ขั้นตอนการเลือกข้อมูลจากหนึ่งตาราง

5.3.2 การเลือกข้อมูลจากหลายตาราง

ในการเลือกข้อมูลสามารถเลือกข้อมูลสามารถเลือกข้อมูลที่มาจากหลายตารางภายในฐานข้อมูลเดียวกันได้ดังรูปที่ 5.4 โดยมีขั้นตอนการทำงานดังนี้คือ

1. เลือกช่อง Join Table สามารถพิมพ์คำสั่งของ SQL ลงใน SQL Command เพื่อเลือกข้อมูลที่ต้องการ โดยผู้ใช้ระบบจะต้องทราบความสัมพันธ์ของข้อมูลในแต่ละตาราง และสามารถใช้คำสั่ง SQL พื้นฐานในการเชื่อมความสัมพันธ์เหล่านั้น เพื่อเลือกข้อมูลที่จะนำมาทำดาต้าไมนิ่งได้
2. คลิกที่ปุ่ม Execute สิ้นสุดขั้นตอนการเลือกข้อมูล เพื่อเข้าสู่การเตรียมข้อมูลในขั้นตอนต่อไป
3. การเลือกข้อมูลในการทำดาต้าไมนิ่งผู้ใช้ระบบ จะเลือกข้อมูลจากหนึ่งตาราง หรือหลายตารางจากการ Join ได้วิธีใดวิธีหนึ่งเท่านั้น



รูปที่ 5.4 ขั้นตอนการเลือกข้อมูลจากหลายตาราง

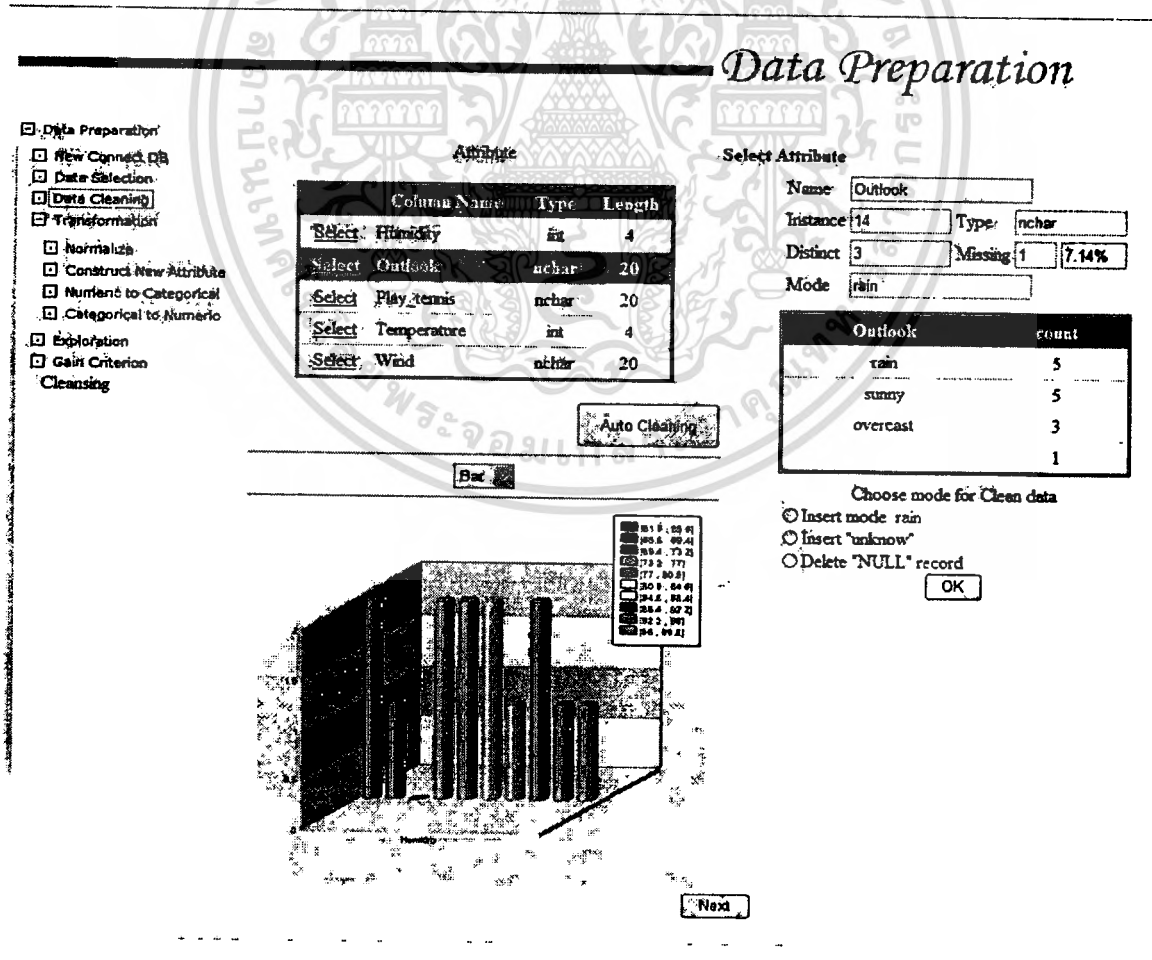
5.4 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูลคือ การแก้ไขปัญหาที่พบในข้อมูลเพื่อทำให้ข้อมูลมีคุณภาพก่อนที่จะนำข้อมูลไปประมวลผล ซึ่งในขั้นตอนนี้เป็นการทำ Data Cleansing

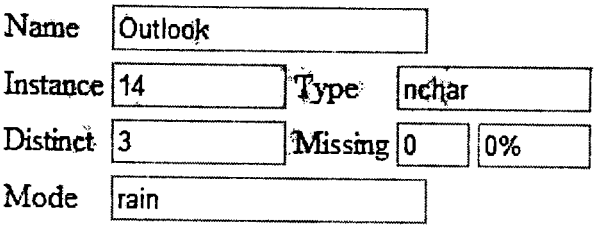
5.4.1 Data cleansing ข้อมูลที่เป็น Categorical

เป็นขั้นตอนในการจัดการข้อมูลที่ต้องการ ซึ่งข้อมูลส่วนใหญ่ในฐานข้อมูลนั้นมักจะไม่สมบูรณ์ ซึ่งมีการทำงานของระบบ ดังนี้

1. หลังจากขั้นตอนการเลือกข้อมูลเสร็จสิ้น ระบบจะเข้าสู่เมนู Data Cleansing ขึ้นมาโดยอัตโนมัติ
2. เลือกชื่อแอตทริบิวที่ใช้ทำดาต้าไมนิ่ง ระบบจะแสดงรายละเอียดของแอตทริบิวและแสดงกราฟของจำนวนข้อมูลในแอตทริบิวที่ผู้ใช้เลือก



รูปที่ 5.5 ขั้นตอนการทำ data cleansing



รูปที่ 5.6 รายละเอียดของแอตทริบิวที่เลือก

จากรูปที่ 5.6 แสดงรายละเอียดของแอตทริบิวที่ผู้ใช้ระบบคลิกเลือก โดยแสดงรายละเอียด ดังนี้

Name: ชื่อแอตทริบิว

Instance: จำนวนเรคคอร์ด

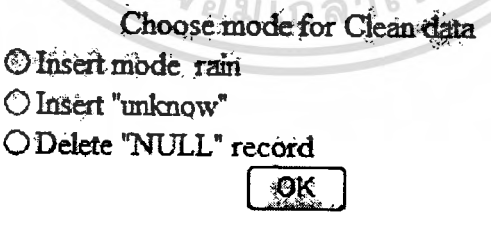
Type: ชนิดข้อมูล

Distinct: จำนวนข้อมูลที่ไม่ซ้ำกัน

Missing: จำนวนเรคคอร์ดที่ค่าหายไป และคิดเป็นเปอร์เซ็นต์

Mode: ค่าฐานนิยม

ผู้ใช้สามารถคลิกเลือกวิธีตามที่ต้องการได้ จากนั้นคลิก OK เพื่อทำการคลีนข้อมูลตามวิธีที่ได้เลือกไว้ ระบบจะทำการคลนีข้อมูลที่เลือก หลังจากนั้นทำการแก้ไขข้อมูลต่อไปจนครบทุกแอตทริบิว หากผู้ใช้ไม่ต้องการที่จะทำการแก้ไขข้อมูลครั้งละแอตทริบิว ให้คลิกที่ปุ่ม Auto Cleansing เพื่อให้ระบบลบเรคคอร์ดที่มี Null ทิ้งไปเป็นการเสร็จสิ้นขั้นตอนการทำ Data Cleansing



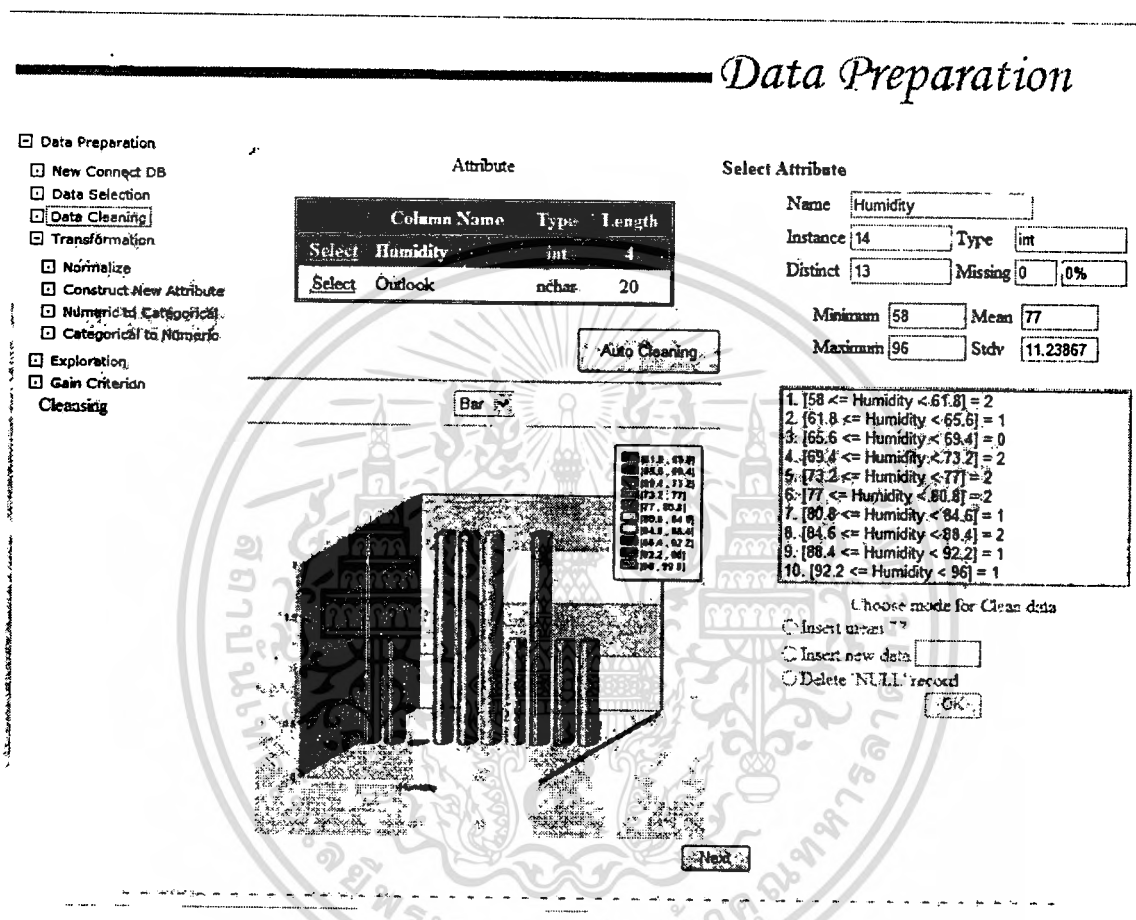
รูปที่ 5.7 ทางเลือกในการ clean ข้อมูลที่เป็น Category

จากรูปที่ 5.7 ระบบแสดงทางเลือกในการขจัดข้อมูลที่ขาดหายไป 3 ทางเลือกคือ

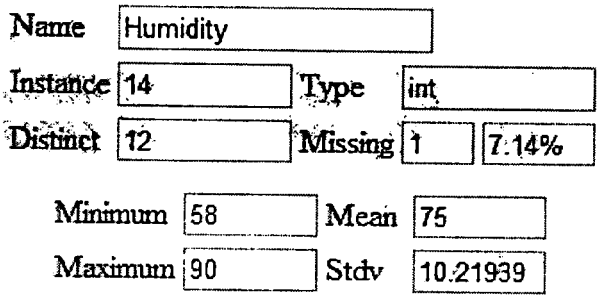
- เติมค่าฐานนิยม (mode) ลงในข้อมูลที่หายไป
- เติม Unknown ลงในข้อมูลที่หายไป
- ลบเรคคอร์ดที่มีค่า Null ทิ้งไป

5.4.2 Data cleansing ข้อมูลที่เป็น Numerical

เลือกชื่อแอตทริบิวที่ใช้ทำดาต้าไมนิ่ง จากตัวอย่างนี้ให้คลิกเลือกแอตทริบิวที่เป็น Numeric ระบบจะแสดงรายละเอียดของแอตทริบิว และแสดงกราฟข้อมูลในแอตทริบิว



รูปที่ 5.8 ขั้นตอนการขจัดข้อมูลที่เป็น Numeric ในเรคคอร์ดที่เป็น Null



รูปที่ 5.9 รายละเอียดของแอตทริบิวที่เป็น Numeric

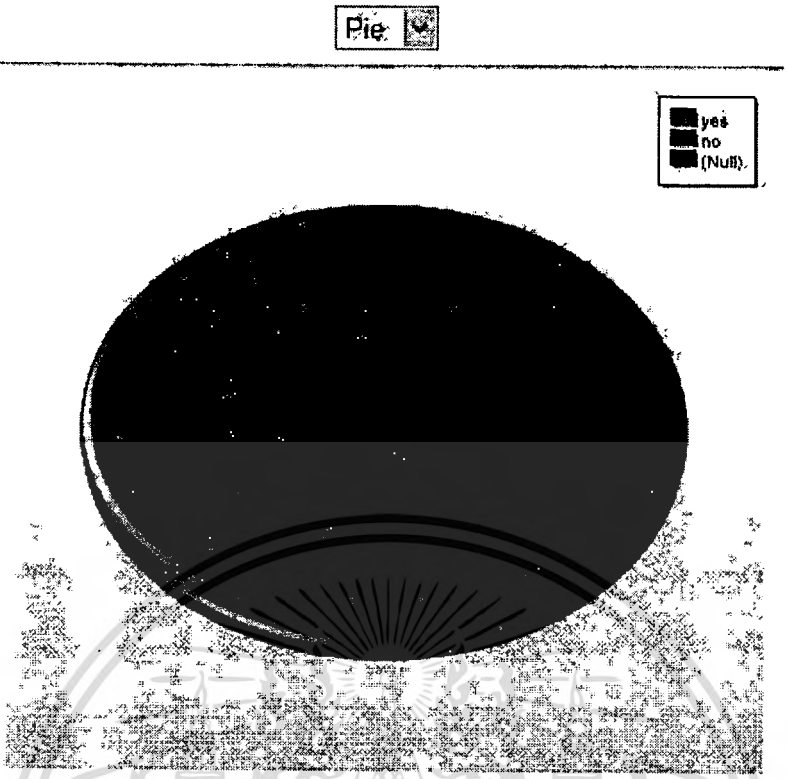
จากรูปที่ 5.9 แสดงรายละเอียดของแอตทริบิวที่ผู้ใช้ระบบคลิกเลือก โดยแสดงรายละเอียด ดังนี้

Name:	ชื่อแอตทริบิว
Instance:	จำนวนเรคคอร์ด
Type:	ชนิดข้อมูล
Distinct:	จำนวนข้อมูลที่ไม่ซ้ำกัน
Missing:	จำนวนเรคคอร์ดที่ค่าหายไป และคิดเป็นเปอร์เซ็นต์
Minimum:	ค่าที่น้อยที่สุด
Maximize:	ค่าที่มากที่สุด
Mean:	ค่าเฉลี่ย หรือค่ากลาง
Stdv:	ค่าเบี่ยงเบนมาตรฐาน

รูปที่ 5.10 แสดงช่วงของข้อมูลในแอตทริบิว age มีชนิดข้อมูลแบบ float ที่เป็น Numeric ซึ่งมีค่า Min = 58 และมีค่า Max =90 โดยกำหนดไว้ 10 ช่วง ช่วงละเท่า ๆ กัน เรียงจากน้อยไปมาก ตามลำดับ และแสดงจำนวนเรคคอร์ดที่นับได้ในแต่ละช่วง ซึ่งข้อมูลที่แสดงนี้จะสัมพันธ์กับกราฟ ดังรูปที่ 5.12

1. [58 <= Humidity < 61.2] = 2
2. [61.2 <= Humidity < 64.4] = 0
3. [64.4 <= Humidity < 67.6] = 1
4. [67.6 <= Humidity < 70.8] = 1
5. [70.8 <= Humidity < 74] = 1
6. [74 <= Humidity < 77.2] = 2
7. [77.2 <= Humidity < 80.4] = 2
8. [80.4 <= Humidity < 83.6] = 1
9. [83.6 <= Humidity < 86.8] = 1
10. [86.8 <= Humidity < 90] = 2

รูปที่ 5.10 ช่วงของข้อมูลในแอตทริบิว

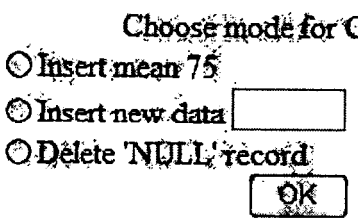


รูปที่ 5.12 กราฟข้อมูลในแอตทริบิว

รูปที่ 5.13 ระบบแสดงทางเลือกในการขจัดข้อมูลที่ขาดหายไป 3 ทางเลือกคือ

- เติมค่าเฉลี่ย (mean) ลงในข้อมูลที่หายไป
- ผู้ใช้ระบุตัวเลขที่ต้องการลงไป
- ลบเรคคอร์ดที่มีค่า Null ทิ้งไป

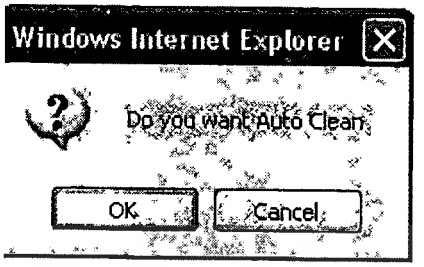
ผู้ใช้สามารถคลิกเลือกวิธีตามที่ต้องการได้ จากนั้นคลิก OK เพื่อทำการคลีนข้อมูลตามวิธีที่ได้เลือกไว้ หรือคลิกที่ปุ่ม Auto Cleansing ก็ได้เช่นเดียวกัน



รูปที่ 5.13 ทางเลือกในการ clean ข้อมูลที่เป็น Numeric

5.4.3 AUTO CLEANSING

เป็นวิธีการคลีนข้อมูลอัตโนมัติ ซึ่งวิธีการนี้จะใช้วิธีการลบเร็คคอร์ดที่มีค่า Null ออกทันที



รูปที่ 5.14 Auto Clean

5.5 การแปลงข้อมูล (Data Transformation)

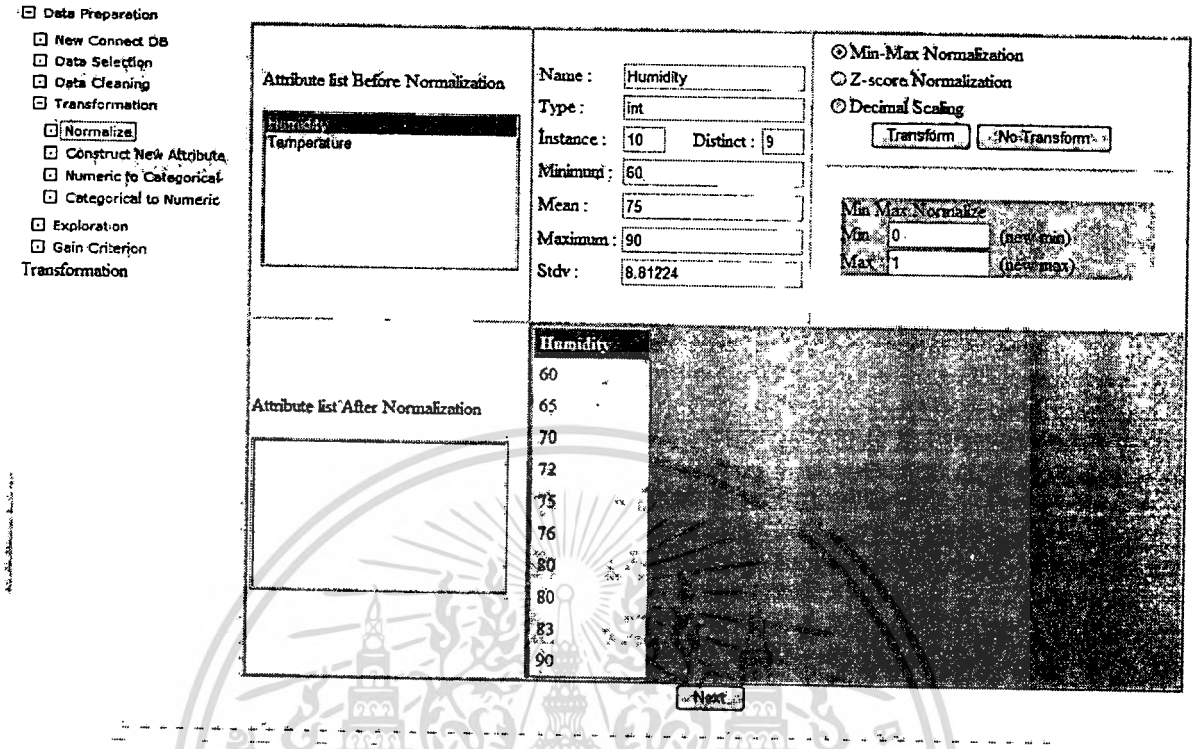
ขั้นตอนการแปลงข้อมูลคือ การทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำมาใช้

5.5.1 Normalization

การแปลงค่าทำให้ข้อมูลในแอทริบิวมีค่าไม่เกินขอบเขตที่กำหนด

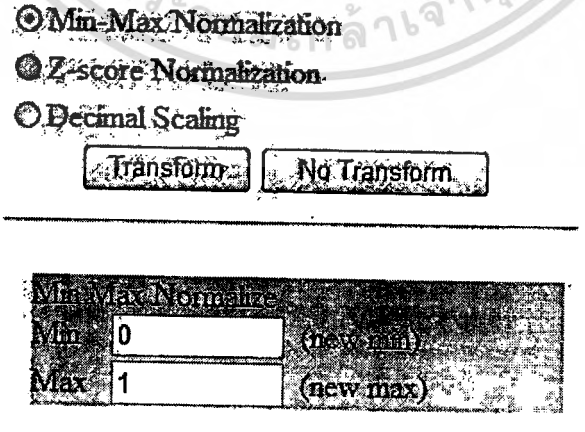
5.5.1.1 Min-Max Normalization

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Min- Max Normalization
3. กำหนดขอบเขตของค่าที่ต้องการแปลง Min คือช่วงต่ำสุดของค่าที่ต้องการ, Max คือช่วงสูงสุดของค่าที่ต้องการ
4. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
5. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Min-Max Normalization



รูปที่ 5.15 ขั้นตอนการแปลงข้อมูลแบบ Min- Max Normalization

รูปที่ 5.16 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Min- Max Normalization ผู้ใช้ต้องระบุขอบเขตของข้อมูลก่อน จากตัวอย่างกำหนด Min = 0 และ Max = 1 ค่าที่ได้จากการแปลงจะอยู่ในขอบเขตนี้



รูปที่ 5.16 การเลือกวิธี Min- Max Normalization

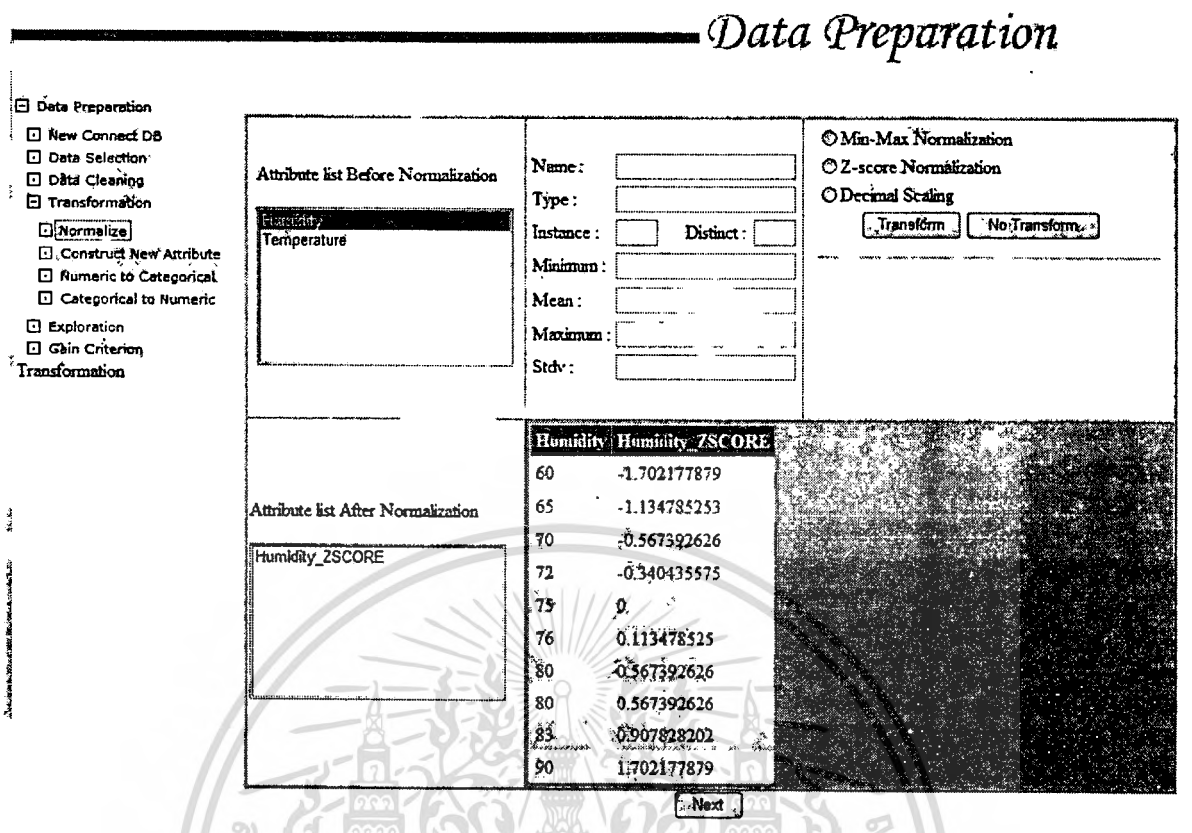
รูปที่ 5.17 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว Humidity ก่อนทำการแปลงค่าแบบ Min- Max Normalization โดยค่าที่แสดงเรียงลำดับจากน้อยไปมาก คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว Humidity_MinMax โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

Humidity	Humidity_MINMAX
60	0
65	0.1666666666666
70	0.3333333333333
72	0.4
75	0.5
76	0.5333333333333
80	0.6666666666666
80	0.6666666666666
83	0.7666666666666
90	1

รูปที่ 5.17 ข้อมูลที่ได้จากการแปลงโดยวิธี Min- Max Normalization

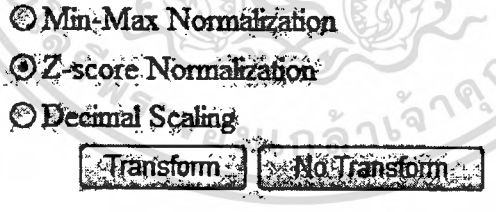
5.5.1.2 Z-score Normalization

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Z-score Normalization
3. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
4. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Z-score Normalization



รูปที่ 5.18 ขั้นตอนการแปลงข้อมูลแบบ Z-score Normalization

รูปที่ 5.19 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Z-score Normalization



รูปที่ 5.19 การเลือกวิธี Z-score Normalization

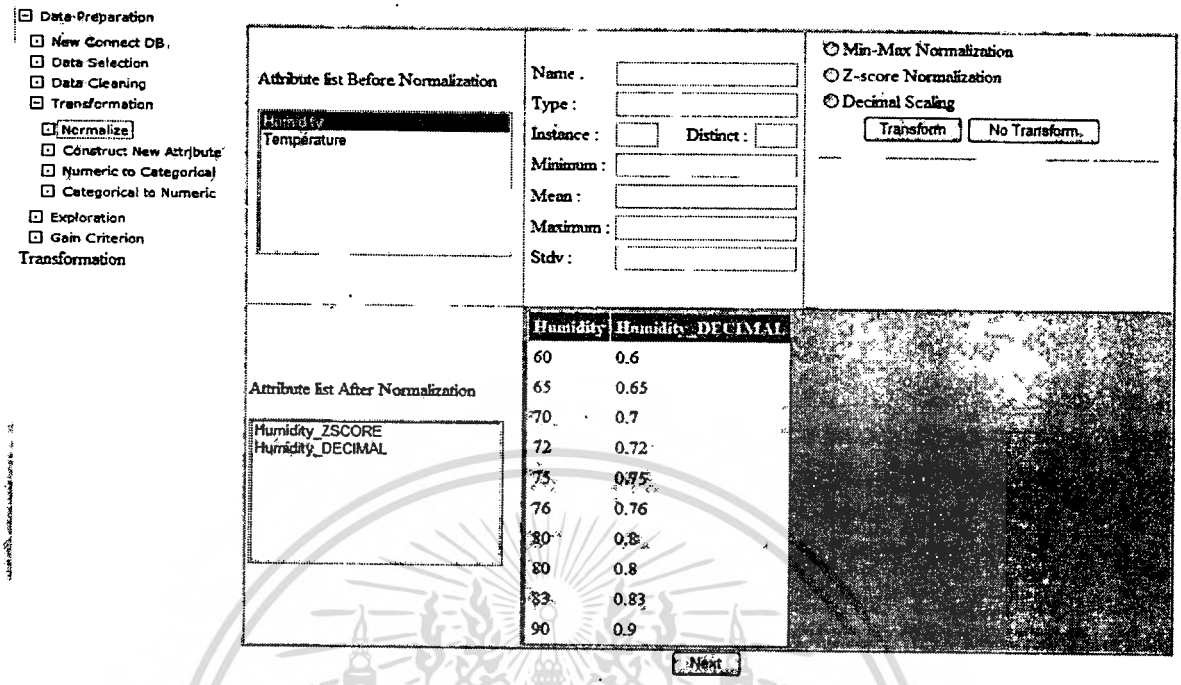
รูปที่ 5.20 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว Humidity ก่อนทำการแปลงค่าแบบ Z-score Normalization คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว Humidity_Zscore โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

Humidity	Humidity_ZSCORE
60	-1.702177879
65	-1.134785253
70	-0.567392626
72	-0.340435575
75	0
76	0.113478525
80	0.567392626
80	0.567392626
83	0.907828202
90	1.702177879

รูปที่ 5.20 ข้อมูลที่ได้จากการแปลงโดยวิธี Z-score Normalization

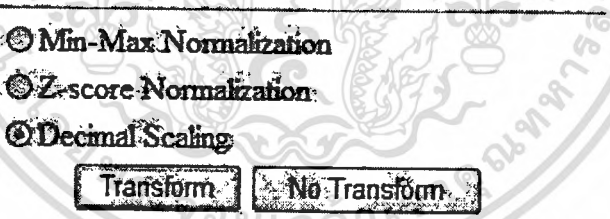
5.5.1.3 Decimal Scaling

1. คลิกเลือกชื่อแอตทริบิวที่ต้องการแปลงข้อมูลในช่อง Attribute list before Normalization
2. คลิกเลือก Decimal Scaling
3. คลิกปุ่ม Transform เพื่อแปลงค่าตามที่กำหนด
4. ระบบแสดงค่าแอตทริบิวเดิม และแอตทริบิวที่ได้หลังการแปลงข้อมูลแบบ Decimal Scaling



รูปที่ 5.21 ขั้นตอนการแปลงข้อมูลแบบ Decimal Scaling

รูปที่ 5.22 แสดงการเลือกวิธีการแปลงข้อมูลแบบ Decimal Scaling



รูปที่ 5.22 ขั้นตอนการแปลงข้อมูลแบบ Decimal Scaling

รูปที่ 5.23 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว Humidity ก่อนทำการแปลงค่าแบบ Decimal Scaling คอลัมน์ที่สองแสดงข้อมูลในแอตทริบิว Humidity_DECIMAL โดยชื่อของแอตทริบิวระบบจะทำการสร้างให้โดยอัตโนมัติซึ่งอ้างอิงจากชื่อแอตทริบิวเดิมแล้วต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล ซึ่งวิธีนี้เป็นการเติมทศนิยมให้กับข้อมูล

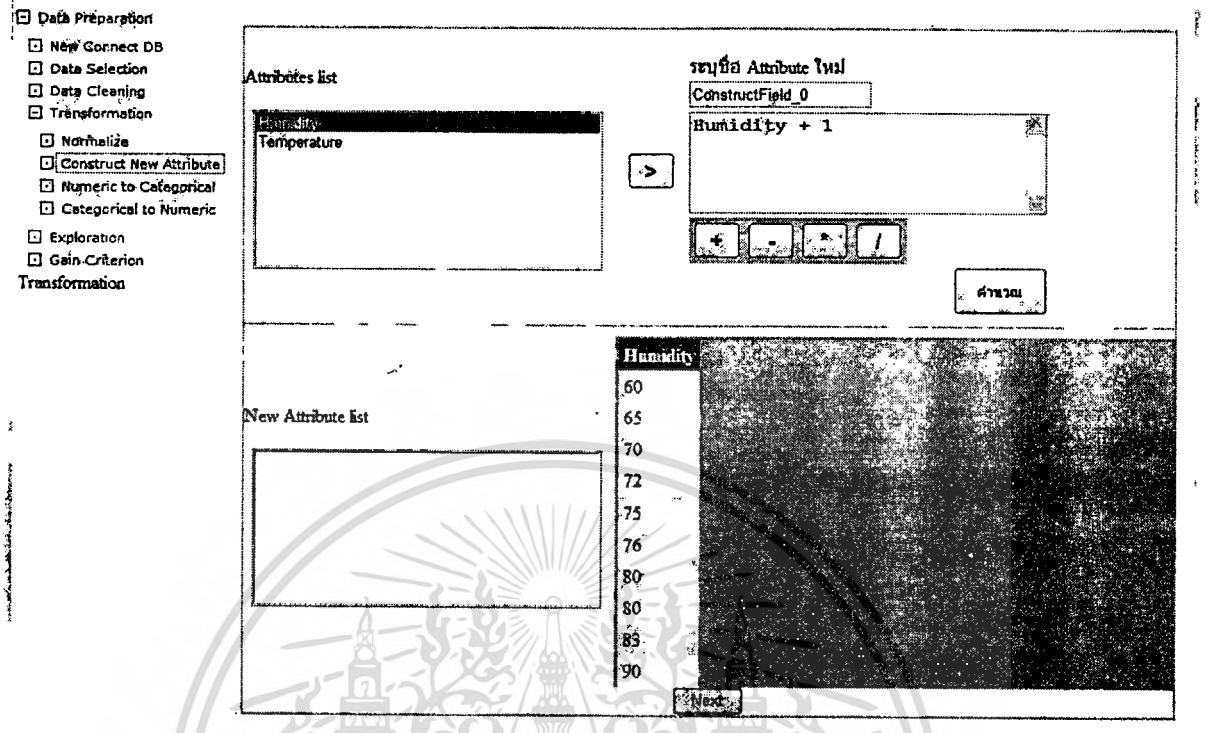
Humidity	Humidity_DECIMAL
60	0.6
65	0.65
70	0.7
72	0.72
75	0.75
76	0.76
80	0.8
80	0.8
83	0.83
90	0.9

รูปที่ 5.23 ข้อมูลหลังจากการแปลงแบบ Decimal Scaling

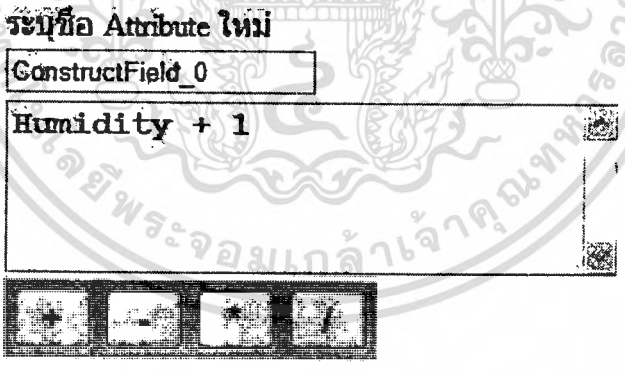
5.5.2 Construct New Attribute

การสร้างแอตทริบิวใหม่ที่ได้จากการคำนวณ มีขั้นตอนการทำงานดังนี้

1. คลิกที่เมนู Construct New Attribute
2. หากผู้ใช้ระบบต้องการระบุชื่อ แอตทริบิวใหม่ ให้พิมพ์ลงในช่อง ระบุชื่อ Attribute ใหม่ หรือใช้ชื่อตามที่ระบบตั้งให้
3. พิมพ์สูตรที่ต้องการคำนวณ
4. กดปุ่ม คำนวณ
5. ระบบแสดงแอตทริบิวใหม่พร้อมข้อมูลที่ได้จากการคำนวณ
6. ในช่อง Attribute list ด้านล่างแสดง ชื่อแอตทริบิวใหม่และสูตรที่คำนวณ



รูปที่ 5.24 ขั้นตอนการแปลงข้อมูลแบบ Construct New Attribute



รูปที่ 5.25 ขั้นตอนการแปลงข้อมูลแบบ Construct New Attribute

จากรูปที่ 5.25 คือการสร้างแอตทริบิวที่ได้จากการคำนวณ ผู้ใช้สามารถตั้งชื่อของแอตทริบิวได้เองในช่องระบุชื่อ Attribute ใหม่ หรือระบบตั้งให้ซึ่งขึ้นต้นด้วยคำว่า ConstructField แล้วตามด้วยหมายเลข ช่องด้านล่างให้ผู้ใช้ระบุสูตรที่ใช้ในการคำนวณแอตทริบิวที่สร้างขึ้นมาใหม่โดยผู้ใช้จะต้องระบุสูตรที่ถูกต้องลงไปเอง ขั้นสุดท้ายเมื่อระบุสูตรแล้ว คลิกที่ปุ่มคำนวณเป็นการเสร็จสิ้นการสร้างแอตทริบิวที่ได้จากการคำนวณ

รูปที่ 5.26 ในคอลัมน์แรกแสดงข้อมูลของแอตทริบิว credit_limit คอลัมน์ที่สองแสดงข้อมูลของแอตทริบิว ConstructField_0 ที่ได้จากการคำนวณตามสูตร ConstructField_0 = Humidity +1

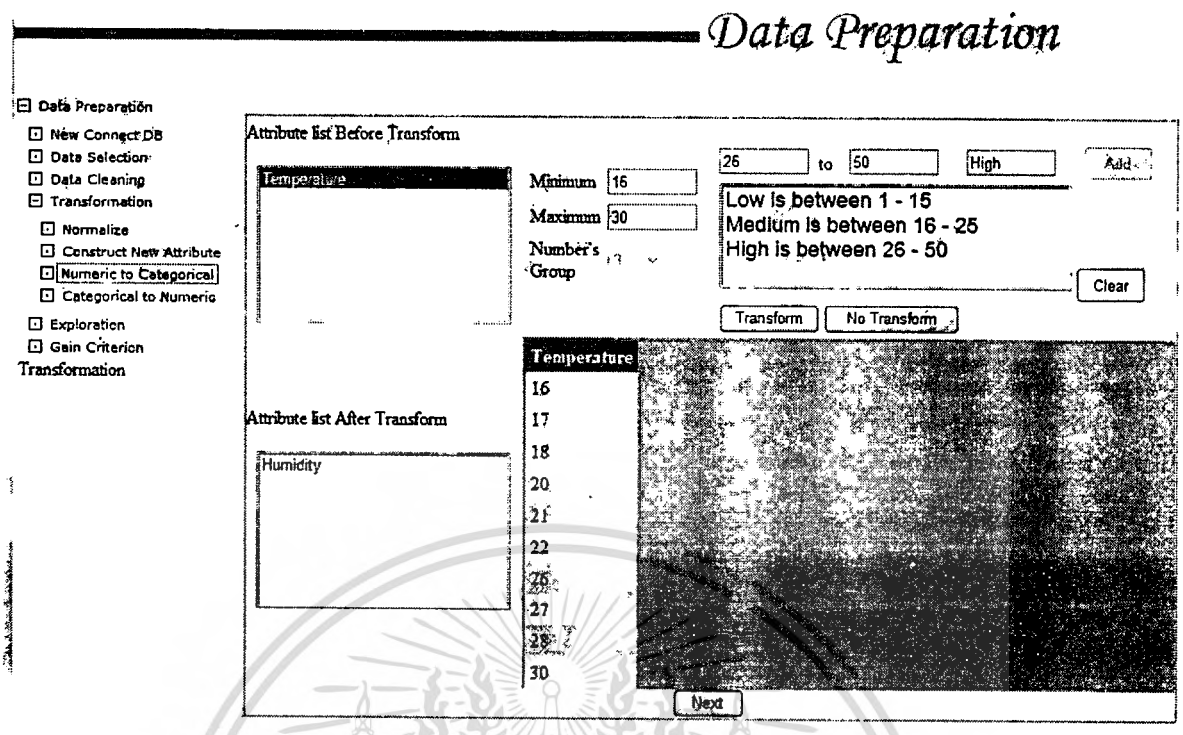
Humidity	ConstructField_0_CONSTRUCT
60	61
65	66
70	71
72	73
75	76
76	77
80	81
80	81
83	84
90	91

รูปที่ 5.26 ข้อมูลหลังการแปลงแบบ Construct New Attribute

5.5.3 Numeric to Categorical

การแปลงข้อมูลที่เป็น Numeric ให้เป็นข้อมูลที่เป็น Category มีขั้นตอนการแปลงข้อมูลดังนี้

1. คลิกที่เมนู Numeric to Categorical
2. คลิกเลือกจำนวนกลุ่มที่ต้องการ
3. ระบุช่วงที่ต้องการแปลงข้อมูล
4. ระบุคำที่ต้องการแปลง สำหรับช่วงที่กำหนด คลิกปุ่ม Add เพิ่มลงในระบบ
5. ทำซ้ำข้อ 3-5 จนครบจำนวนกลุ่มที่ต้องการแบ่ง
6. คลิกปุ่ม Transform เพื่อแปลงข้อมูล
7. ระบบแสดงข้อมูลของแอตทริบิวที่ต้องการแปลง และข้อมูลที่ได้หลังการแปลงแบบ Numeric to Categorical

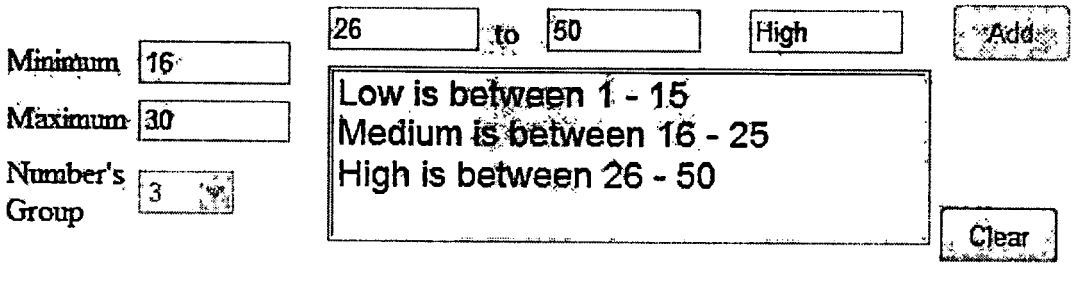


รูปที่ 5.27 ขั้นตอนการแปลงข้อมูลจากตัวเลขเป็นตัวอักษร

จากรูปที่ 5.27 ระบบแสดงค่า Min และค่า Max ของแอตทริบิวที่ต้องการแปลง ให้ผู้ใช้ระบุจำนวนกลุ่มที่ต้องการ จากนั้นให้กำหนดช่วงของตัวเลขที่ต้องการ แล้วกำหนดข้อความที่จะแทนลงไป คลิกที่ปุ่ม Add แล้วกำหนดแบบเดิมจนครบตามจำนวนกลุ่มที่ระบุ ขั้นตอนสุดท้ายให้คลิกที่ปุ่ม Transform เพื่อทำการแปลงค่าตามที่กำหนด

ตัวอย่าง การแปลงข้อมูลอุณหภูมิให้เป็นข้อความ 3 ข้อความตามช่วงอุณหภูมิ เช่น

- ช่วง 1-15 แปลงค่าเป็น Low
- ช่วง 16-25 แปลงค่าเป็น Medium
- ช่วง 26-50 แปลงค่าเป็น High



รูปที่ 5.28 ตัวอย่างการกำหนดข้อความให้กับข้อมูลที่ต้องการแปลง

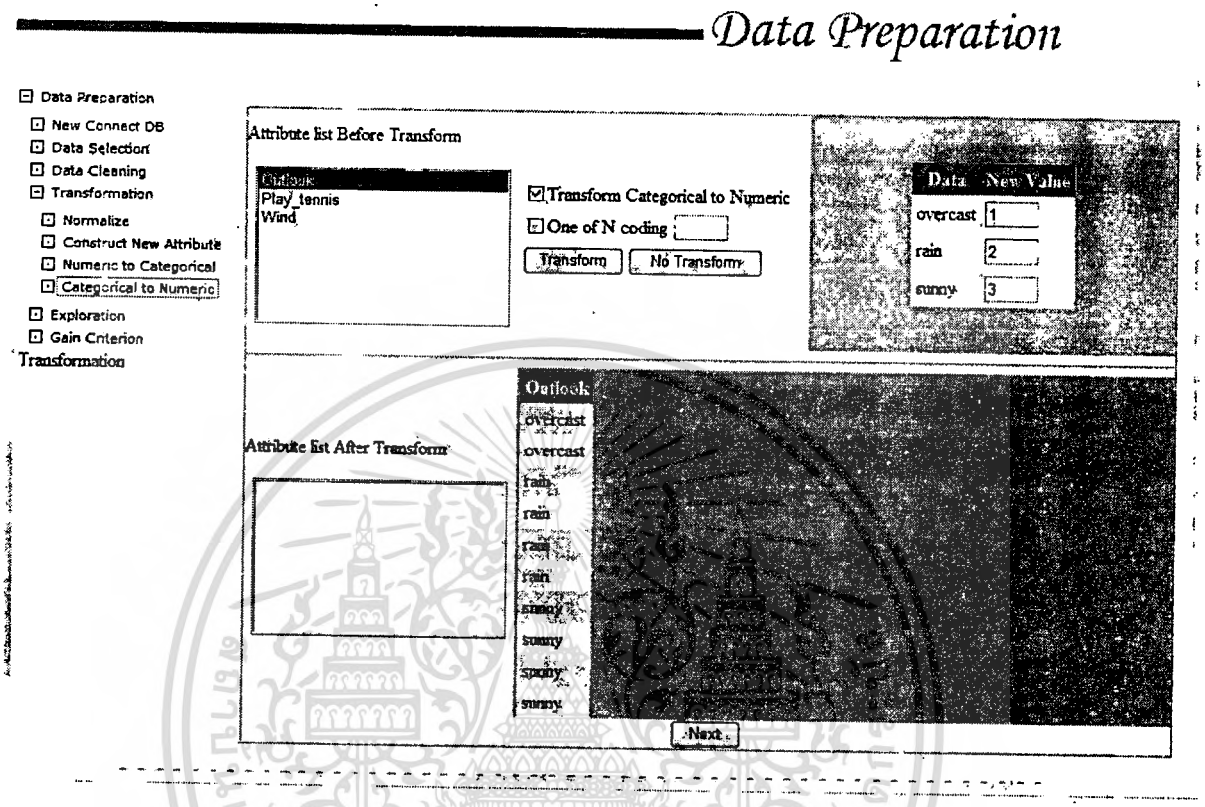
รูปที่ 5.29 คอลัมน์แรกแสดงข้อมูลในแอตทริบิว age ก่อนที่จะแปลงข้อมูล คอลัมน์ที่สองแสดงข้อมูลที่ได้หลังจากการแปลงจากข้อมูลแบบ Numeric เป็นข้อมูลแบบ Category โดยชื่อแอตทริบิว จะขึ้นต้นด้วยชื่อแอตทริบิวเดิมต่อด้วยวิธีที่ใช้ในการแปลงข้อมูล

Temperature	Temperature_CATEGORY
16	Medium
17	Medium
18	Medium
20	Medium
21	Medium
22	Medium
26	High
27	High
28	High
30	High

รูปที่ 5.29 ข้อมูลหลังการแปลงแบบ Numeric to Category

5.5.4 Categorical to Numeric

เป็นขั้นตอนการแปลงข้อมูลที่อยู่ในลักษณะตัวอักษรไปเป็นตัวเลข



รูปที่ 5.30 ขั้นตอนการแปลงข้อมูล Category to Numeric

5.5.4.1 One of N Coding

ตัวอย่าง การแปลงค่าในแอตทริบิว weak ที่มีค่า strong, weak ให้เป็นตัวเลข ได้ข้อมูลดังตารางที่ 5.1

ตารางที่ 5.1 ตัวอย่างของการแปลงค่าแบบ One of N Coding

Wind	Wind_0	Wind_1
strong	1	0
weak	0	1

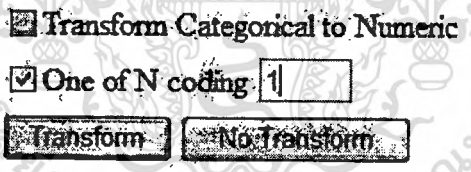
ขั้นตอนการใช้งานโปรแกรม มีดังนี้

1. คลิกที่เมนู Categorical to Numeric
2. คลิกเลือกที่แอตทริบิวที่ต้องการแปลงค่า

3. คลิกเลือก One of N Coding และใส่ค่าที่ต้องการ เช่น 1
4. คลิกที่ปุ่ม Transform
5. ระบบแสดงค่าของแอตทริบิวที่ได้หลังการแปลงแบบ One of N Coding

Wind	Wind_0_NUMERIC	Wind_1_NUMERIC
strong	1	0
strong	1	0
strong	1	0
strong	1	0
weak	0	1
weak	0	1
weak	0	1
weak	0	1
weak	0	1
weak	0	1

รูปที่ 5.31 การแปลงข้อมูลจากตัวอักษรเป็นตัวเลขวิธี One of N Coding



รูปที่ 5.32 การแปลงข้อมูลจากตัวอักษรเป็นตัวเลขวิธี One of N Coding

รูปที่ 5.32 แสดงการเลือกวิธีการแปลงข้อมูลแบบ One of N coding โดยคลิกเลือกที่วิธีที่สอง

5.5.4.2 การแปลงข้อมูลให้เป็นตัวเลข

การแปลงข้อมูลที่ไม่ใช่ตัวเลขให้เป็นตัวเลข ตัวอย่างการแปลงค่าในแอตทริบิว Wind ที่มีค่า strong,weak ให้เป็นตัวเลข ได้ข้อมูลดังตารางที่ 5.2

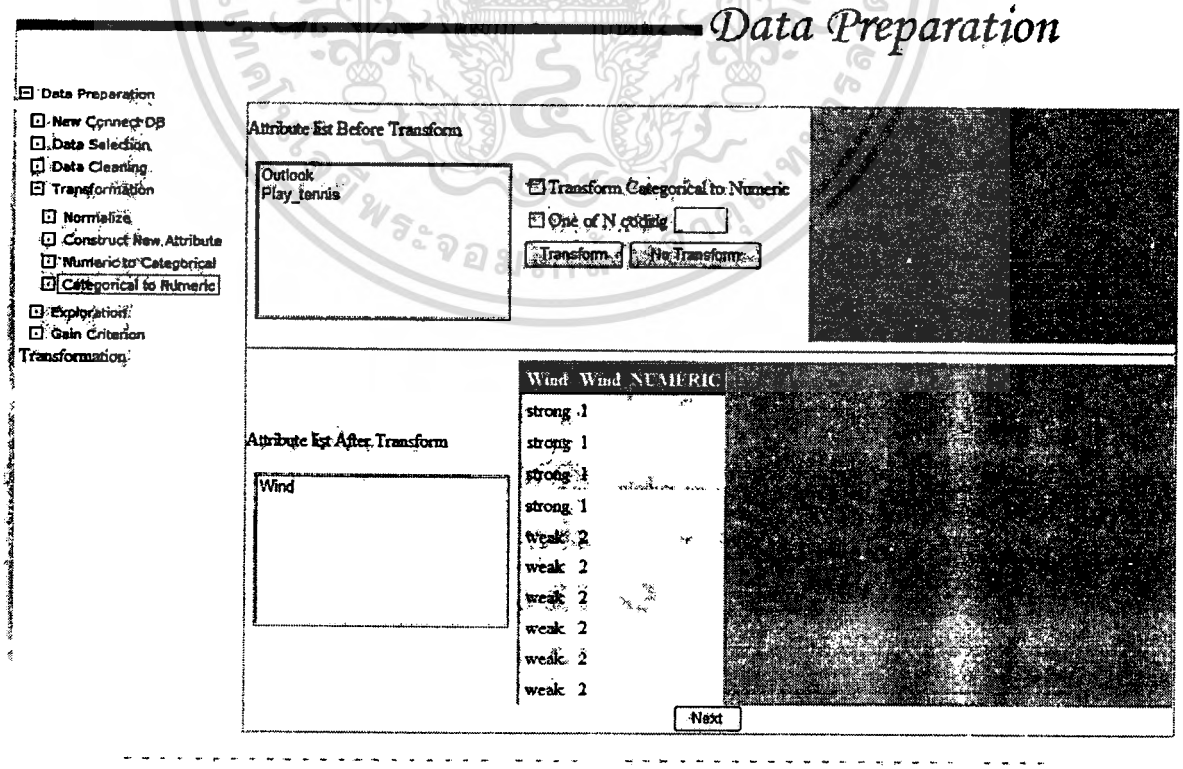
ตารางที่ 5.2 ตัวอย่างของการแปลงค่าให้เป็นตัวเลข

Wind	Wind_NUMERIC
strong	1
weak	2

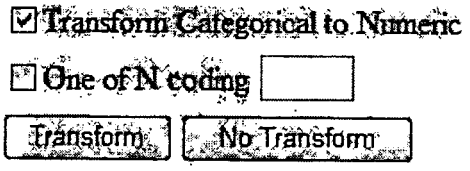
ขั้นตอนการใช้งานโปรแกรม มีดังนี้

คลิกที่แท็บ Categorical to Numeric

1. คลิกเลือกที่แอตทริบิวที่ต้องการแปลงค่า
2. คลิกเลือก แปลงตัวอักรเป็นตัวเลข
3. ระบบจะแสดงค่า DATA และ NEW VALUE โดยผู้ใช้ระบบสามารถแก้ไขค่า NEW VALUE ให้เป็นค่าตัวเลขตามที่ต้องการได้ คลิกที่ปุ่ม Transform
4. ระบบแสดงค่าของแอตทริบิวก่อนการแปลงข้อมูล และหลังการแปลงให้ทราบ

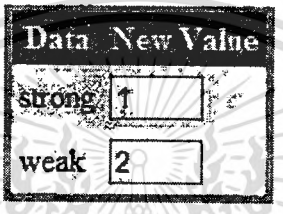


รูปที่ 5.33 ขั้นตอนการแปลงข้อมูลจากตัวอักษรเป็นตัวเลข



รูปที่ 5.34 ขั้นตอนการแปลงข้อมูลจากตัวอักษรเป็นตัวเลข

รูปที่ 5.35 ในคอลัมน์ Data แสดงข้อมูลแบบไม่ซ้ำของแอตทริบิว ก่อนทำการแปลงข้อมูล ในคอลัมน์ New Value ระบบทำการระบุตัวเลขให้กับข้อมูลโดยอัตโนมัติ ซึ่งผู้ใช้สามารถกำหนดตัวเลขอื่นให้กับข้อมูลได้



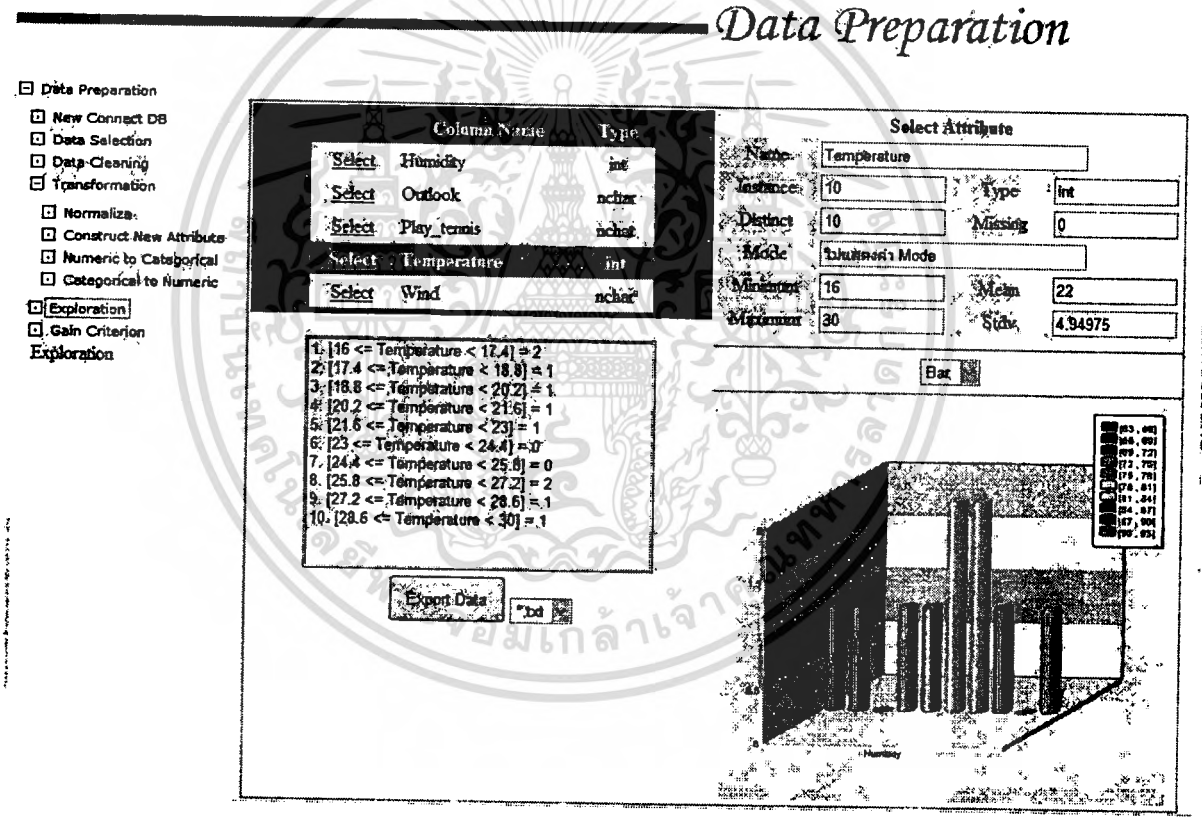
รูปที่ 5.35 ข้อมูลจากแอตทริบิวที่ต้องการแปลงและค่าใหม่ที่เป็นตัวเลข

รูปที่ 5.36 ข้อมูลจากการแปลงข้อมูล Category เป็น Numeric

5.6 การสำรวจข้อมูล (Data Exploration)

5.6.1 การสำรวจข้อมูลที่เป็น Numeric

1. หลังจากทำขั้นตอนการแปลงข้อมูลเรียบร้อยแล้ว คลิกที่แท็บ Exploration
2. ระบบแสดงรายชื่อแอตทริบิวทั้งหมดที่ได้จากการแปลงข้อมูล
3. คลิกเลือกชื่อแอตทริบิว
4. ระบบจะแสดง จำนวนเรคคอร์ด ประเภทของแอตทริบิว ค่าสูงสุด ค่าต่ำสุด ค่าเฉลี่ย และค่าเบี่ยงเบนมาตรฐาน
5. เลือกรูปแบบการแสดงกราฟได้ 3 แบบคือ แบบกราฟแท่ง กราฟเส้น และแบบวงกลม



รูปที่ 5.37 การสำรวจข้อมูลที่เป็น Numeric

รูปที่ 5.38 ระบบแสดงรายชื่อฟิลด์ทั้งหมดที่ได้จากขั้นตอน Data Selection และแอตทริบิวที่ได้หลังจากขั้นตอน Data Transformation

	Column Name	Type
Select	Humidity	int
Select	Outlook	nchar
Select	Play_tennis	nchar
Select	Temperature	int
Select	Wind	nchar

รูปที่ 5.38 รายชื่อแอตทริบิวทั้งหมดหลังขั้นตอน Data Transformation

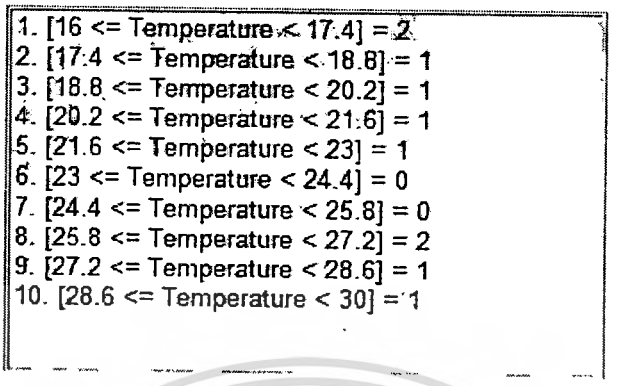
รูปที่ 5.39 เมื่อคลิกที่ชื่อแอตทริบิวที่ต้องการดูข้อมูลแล้ว ในส่วนนี้จะเป็นส่วนแสดงรายละเอียดของแอตทริบิวให้ผู้ใช้ทราบ

Select Attribute

Name	Temperature		
Instance	40	Type	int
Distinct	10	Missing	0
Mode	ไม่แสดงค่า Mode		
Minimum	15	Mean	22
Maximum	30	Stdv	4.94975

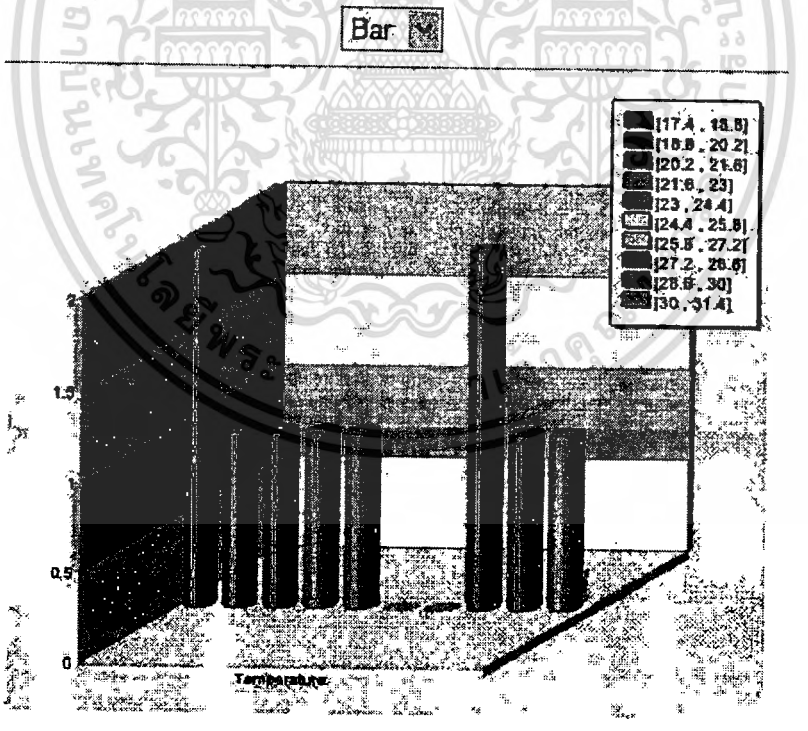
รูปที่ 5.39 รายชื่อแอตทริบิวทั้งหมดหลังขั้นตอน Data Transformation

รูปที่ 5.40 แสดงจำนวนเรคคอร์ดที่นับได้ในแต่ละช่วงข้อมูล โดยระบบกำหนดให้ข้อมูลที่เป็น Numeric มี 10 ช่วงดังรูป ซึ่งข้อมูลที่แสดงในส่วนนี้จะสัมพันธ์กับกราฟในรูปที่ 5.41



รูปที่ 5.40 รายละเอียดของข้อมูลที่เป็น Numeric

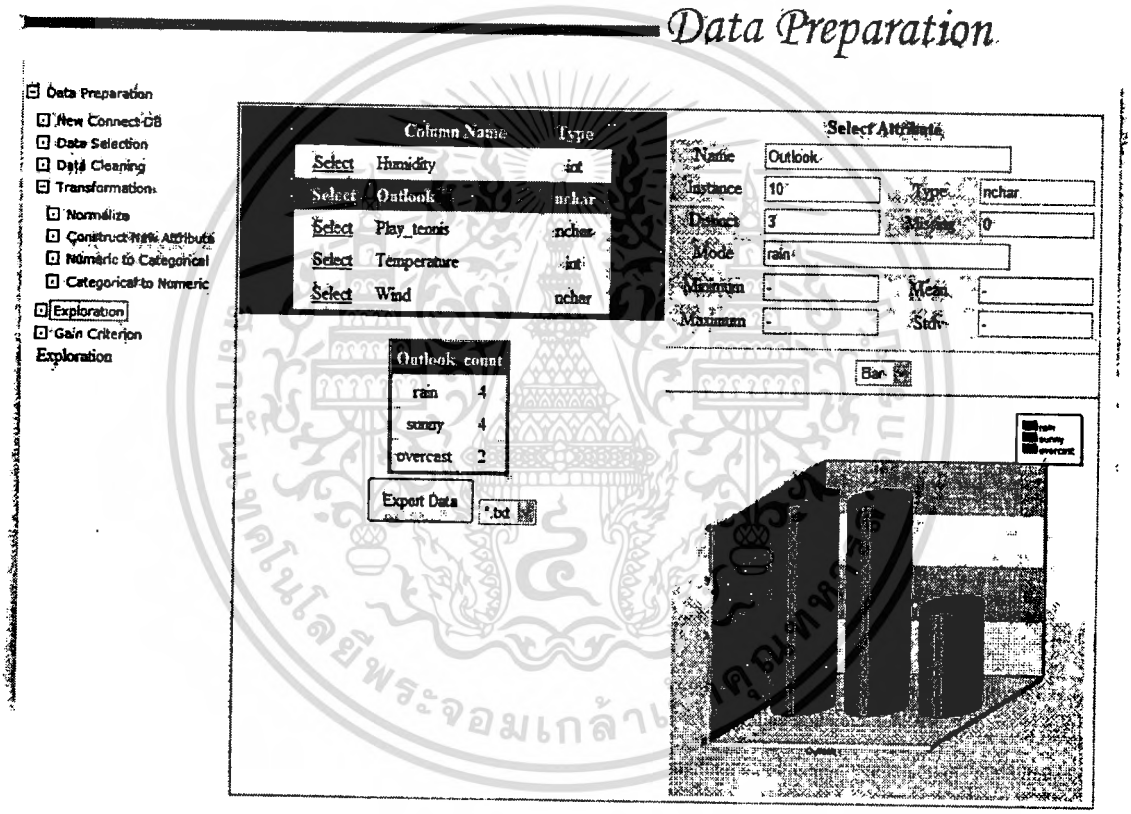
รูปที่ 5.41 แสดงกราฟแท่งของแอตทริบิว Temperature ซึ่งสามารถเปลี่ยนการแสดงผลได้ 3 รูปแบบ



รูปที่ 5.41 กราฟแท่งแสดงข้อมูลของแอตทริบิว Temperature

5.6.2 การสำรวจข้อมูลที่เป็น Category

1. คลิกเลือกชื่อแอตทริบิว
2. กราฟที่แสดงได้จากการหาค่าที่ต่างกัน แล้วนับจำนวนข้อมูลแต่ละตัว โดยการคลิกเลือกที่แอตทริบิว Outlook โดยค่าที่แสดง เป็นตัวอักษรแบ่งได้ 3 ค่า คือ
   - rain จำนวน 4 ค่า
   - Sunny จำนวน 4 ค่า
   - Overcast จำนวน 2 ค่า



รูปที่ 5.42 การสำรวจข้อมูลที่เป็น Categoryแสดงในรูปกราฟแท่ง

รูปที่ 5.43 เมื่อคลิกที่ชื่อแอตทริบิวที่ต้องการดูข้อมูลแล้ว ในส่วนนี้จะเป็นส่วนแสดงรายละเอียดของแอตทริบิวให้ผู้ใช้ทราบ

Select Attribute

Name	Outlook		
Instance	10	Type	nchar
Distinct	3	Missing	0
Mode	rain		
Minimum	-	Mean	-
Maximum	-	Stdv	-

รูปที่ 5.43 รายละเอียดของแอตทริบิว Outlook

รูปที่ 5.44 แสดงข้อมูลในแอตทริบิว Outlook ที่เลือกมาแบบไม่ซ้ำและจำนวนเรคคอร์ดที่นับได้

Outlook	count
rain	4
sunny	4
overcast	2

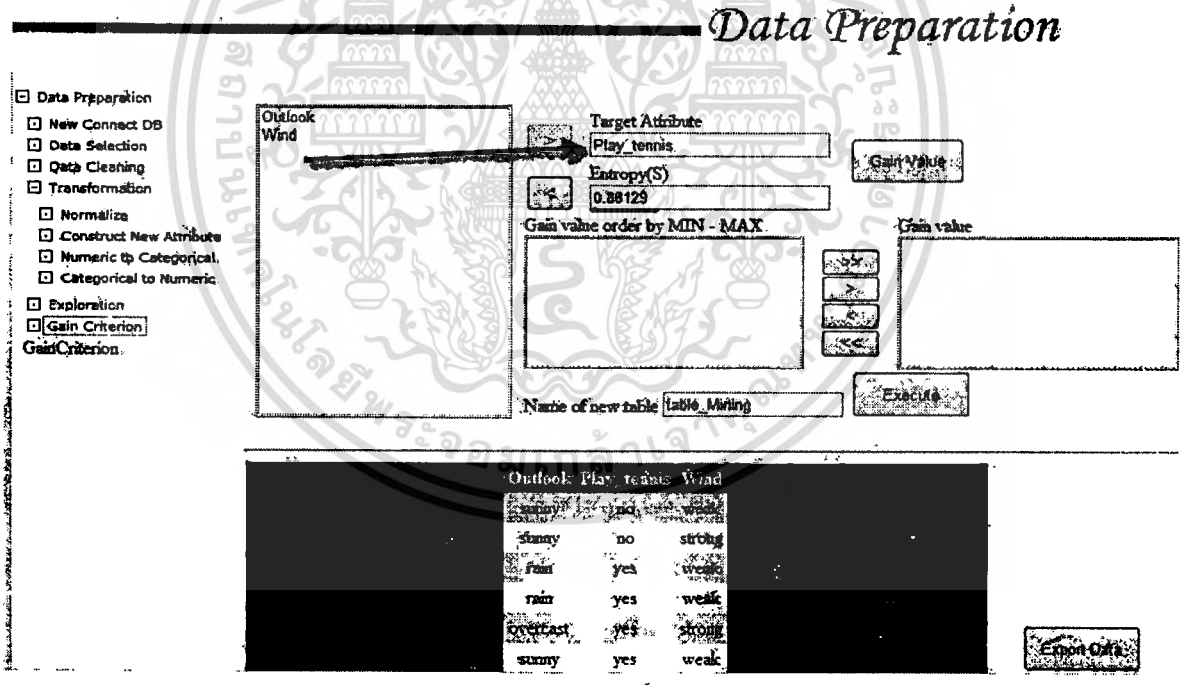
รูปที่ 5.44 ข้อมูลในแอตทริบิว Outlook

5.7 การหาค่า Information Gain สำหรับข้อมูลที่เป็น Categorical

คือวิธีการหาความเกี่ยวข้องของข้อมูล ว่าข้อมูลชุดไหนมีความสำคัญกับ Target attribute มากที่สุด ซึ่งการหา Information Gain นี้จะใช้ ID3 Algorithm มาใช้ในการหาที่ได้กล่าวไปแล้วในบทที่ 3

5.7.1 การหาค่า Entropy(s)

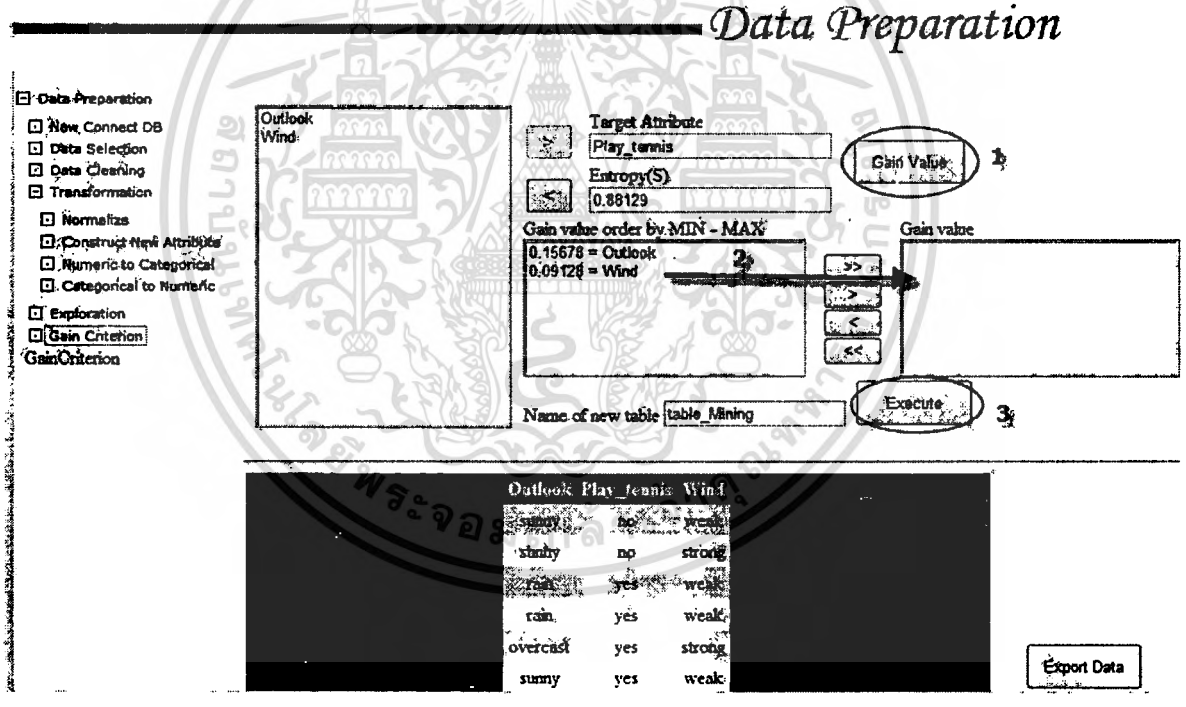
1. การหาค่า Information Gain สามารถหาได้จากแอตทริบิวที่เป็น Categorical เท่านั้น
2. ผู้ใช้เลือก Target Attribute จากรายชื่อแอตทริบิวด้านซ้าย ระบบจะคำนวณค่า Entropy แล้วแสดงให้ทราบ
3. จากตัวอย่าง เลือก Play_tennis เป็น Target Attribute คำนวณค่า Entropy(s) ได้ 0.88129



รูปที่ 5.45 การหาค่า Entropy(S)

5.7.6 การหาค่า Information Gain

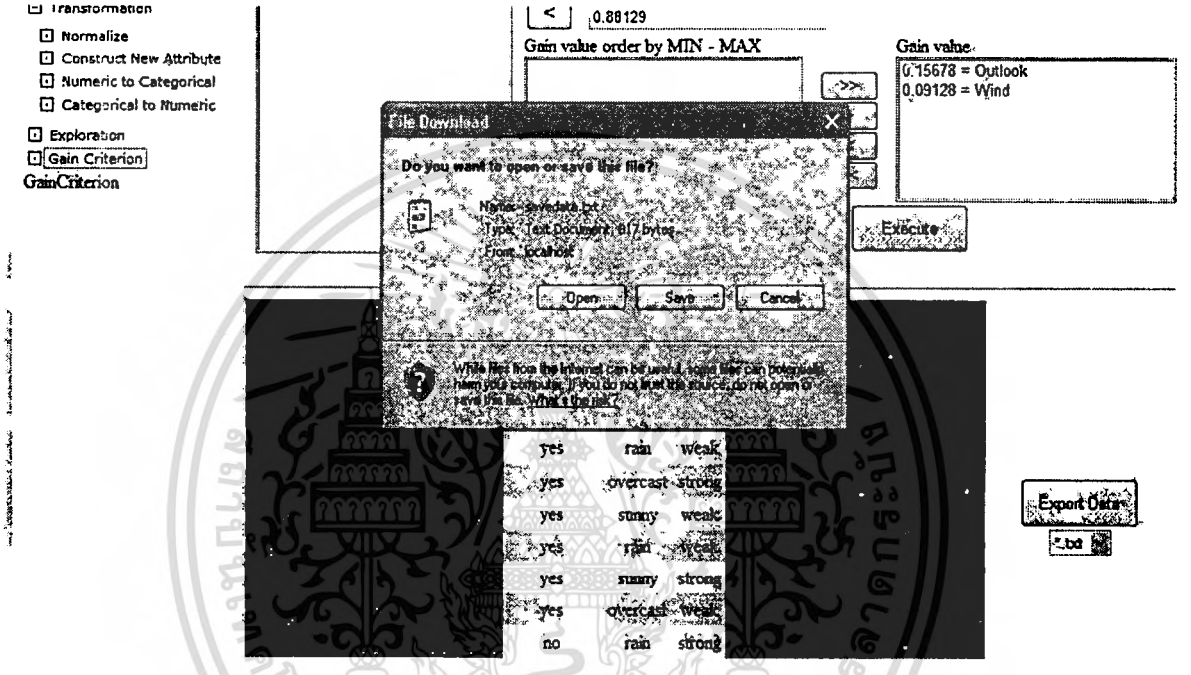
1. คลิกที่ปุ่ม Gain Value ระบบจะคำนวณค่า Gain ของแต่ละแอตทริบิวแล้วแสดงโดยเรียงลำดับค่า Gain จากมากมาน้อย
2. ผู้ใช้สามารถนำค่า Gain ที่ได้มาพิจารณาประกอบในการเลือกข้อมูลเพื่อนำไปใช้ในการทำดาต้าไมนิ่งต่อไป โดยพิจารณาจากค่า Gain ที่มีค่ามาก มาน้อย
3. ผู้ใช้เลือก แอตทริบิวด้านซ้าย โดยพิจารณาจากค่า Gain ที่มีค่ามากก่อน
4. จากตัวอย่าง เลือก Outlook และ Wind ที่มีค่า Gain จากมากมาน้อย ตามลำดับ
5. คลิกที่ปุ่ม Execute เพื่อสร้างตารางใหม่จากแอตทริบิวที่เลือกโดยผู้ใช้สามารถระบุชื่อตารางได้ตามต้องการ
6. ระบบแสดงข้อความว่า ได้สร้าง table ตามชื่อตารางที่กำหนด แล้ว



รูปที่ 5.46 การเลือกแอตทริบิวที่ได้จากการหาค่า Gain

5.7.7 การExport data

1. คลิกที่ปุ่ม Export Data เพื่อทำการ export ข้อมูลที่แสดงอยู่ในตาราง
2. ระบบแสดงข้อความให้ทำการเลือกตำแหน่งที่ต้องการที่จะบันทึกข้อมูลที่ export
3. กด open สำหรับเปิดดูข้อมูล หรือ กด save สำหรับเลือกตำแหน่งที่ต้องการบันทึกข้อมูล



รูปที่ 5.47 ข้อความแสดงตำแหน่งที่ต้องการบันทึกexport file

บทที่ 6

สรุปผลการศึกษาและข้อเสนอแนะ

6.1 สรุปผลการศึกษา

จากการศึกษาทฤษฎีของดาต้าไมนิ่ง ทำให้ได้เรียนรู้ว่าดาต้าไมนิ่งเป็นกระบวนการที่ค้นหาข้อมูลที่เป็นประโยชน์จากภายในฐานข้อมูลที่มีอยู่ ทำให้ได้รับสารสนเทศที่เป็นประโยชน์และสามารถนำสารสนเทศนั้นไปช่วยสนับสนุนการตัดสินใจและการประยุกต์นำไปใช้งานกับธุรกิจต่างๆได้ ซึ่งกระบวนการดังกล่าวจะเริ่มตั้งแต่การกำหนดวัตถุประสงค์ของการทำดาต้าไมนิ่ง จากนั้นก็มีขั้นตอนการเตรียมข้อมูลมาวิเคราะห์ ซึ่งจะประกอบไปด้วยการคัดเลือกข้อมูล การทำความสะอาดข้อมูล และการแปลงข้อมูลให้เหมาะสม หลังจากนั้นก็จะทำดาต้าไมนิ่งเมื่อข้อมูลผ่านการทำไมนิ่ง ก็จะได้ผลลัพธ์ที่เกิดประโยชน์ในทางธุรกิจ

ก่อนที่จะเข้าสู่ขั้นตอนของการทำดาต้าไมนิ่งได้นั้น การเตรียมข้อมูลถือได้ว่าเป็นขั้นตอนที่ใช้ระยะเวลาในการดำเนินการมากกว่าขั้นตอนอื่นๆ ของการทำดาต้าไมนิ่ง เนื่องจากปริมาณข้อมูลมีเป็นจำนวนมากและข้อมูลที่ได้รับมาจากหลายแหล่ง รูปแบบของข้อมูลแตกต่างกัน จึงต้องมีการเตรียมข้อมูลให้อยู่ในรูปแบบเดียวกัน เพื่อให้พร้อมใช้งาน แต่ละอัลกอริทึมของดาต้าไมนิ่งก็ต้องการการนำเข้าข้อมูลแตกต่างกัน อัลกอริทึมบางประเภทใช้เพื่อวิเคราะห์ข้อมูลที่เป็น Numeric เท่านั้น จึงต้องมีการแปลงข้อมูลเหล่านั้นให้เหมาะสมกับอัลกอริทึมแต่ละแบบ

6.2 ข้อเสนอแนะ

1. ระบบนี้เลือกข้อมูลได้ครั้งละหนึ่งฐานข้อมูลเท่านั้น จึงควรเพิ่มเติมในส่วนของการเลือกข้อมูลให้สามารถเชื่อมต่อได้มากกว่าหนึ่งฐานข้อมูล
2. ในการเลือกข้อมูลจากหลายตาราง ผู้ใช้ระบบจะต้องเข้าใจความสัมพันธ์ของแต่ละตาราง แต่ละแอตทริบิว เพื่อใช้คำสั่ง SQL ในการเลือกข้อมูลเข้าสู่ระบบนี้ได้อย่างถูกต้อง
3. ระบบที่พัฒนาขึ้นสามารถเชื่อมต่อกับฐานข้อมูลในระบบ Microsoft SQL Server 2005 เท่านั้น
4. ระบบนี้เมื่อผ่านกระบวนการอะไรแล้วจะไม่สามารถย้อนกลับไปแก้ไขได้ ต้องทำการติดต่อฐานข้อมูลใหม่เท่านั้น

บรรณานุกรม

Arthitaya Chuachan-ad. 2006. **Development of data preparation and exploration for data mining.** System development project. Faculty of information technology King Mongkut's institute of technology Ladkrabang

Dorain Pyle. 1999. **Data Preparation for Data Mining.** Morgan Kaufmann.

Ian H. Witten and Eibe Frank. 2005. **Data Mining Practical Machine Learning Tools and Techniques 2nd Edition.** Morgan Kaufmann.

Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada. **Data Mining: Concepts and Techniques Chapter 3.** [Online]. Available: www.cs.sfu.ca/~han/bk/3prep.ppt

Jiawei Han and Micheline Kamber. 2001. **Data Mining: Concepts and Techniques.** USA : Academic Press.

Kira Tarapanoff. 2001. **Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil.** [Online]. Available: http://informationr.net/ir/7-1/paper117.html

ประวัติผู้เขียน

ชื่อผู้เขียน	นายชนินทร์ พงษ์ลิมานนท์
วันเกิด	1 มกราคม 2527
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษาระดับปริญญาตรี	บริหารธุรกิจบัณฑิต
สถานที่สำเร็จการศึกษา	คณะวิทยาการจัดการ มหาวิทยาลัยสงขลานครินทร์
ปีที่สำเร็จการศึกษา	2549