การพัฒนาเครื่องมือสำหรับการทำ DATA PREPARATION

DEVELOPING TOOL FOR DATA PREPARATION

โดย

มนันญา ภาคศักดิ์ศรี

MANANYA PAKSAKSRI

อาจารย์ที่ปรึกษา

ผศ.ดร. พรฤดี เนติโสภากุล

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2549

DEVELOPING TOOL FOR DATA PREPARATION

MANANYA PAKSAKSRI

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

2/2006

หัวข้อ	การพัฒนาเครื่องมือสำหรับการทำ Data Preparation
นักศึกษา	นางสาวมนันญา ภาคศักดิ์ศรี
รหัสนักศึกษา	47066131
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	ผศ.ดร. พรฤดี เนติโสภากุล

บทคัดย่อ

การเตรียมข้อมูลก่อนการประมวลผล เพื่อนำข้อมูลนั้นๆ ไปใช้ในกระบวนการทำคาต้าไมนิ่งนับว่ามีความสำคัญเป็นอย่างมาก เนื่องจากข้อมูลที่เก็บอยู่ในฐานข้อมูลในปัจจุบันมีมากมาย และแต่ละแหล่งข้อมูลก็มีการเก็บข้อมูลที่ใช้รูปแบบที่มีความแตกต่างกัน บางครั้งข้อมูลที่ได้มานั้นก็ไม่มีความถูกต้อง และไม่เหมาะสมในการที่จะมาใช้ในการทำคาต้าไมนิ่ง

ในโครงการพัฒนาระบบนี้ จะทำการศึกษาทฤษฎีต่างๆที่เกี่ยวข้องกับการเตรียมข้อมูล (Data Preparation) และพัฒนาเครื่องมือสำหรับการทำ Data Preparation ซึ่งเครื่องมือนี้จะทำให้ผู้ใช้สามารถเตรียมข้อมูลให้มีความเหมาะสมเพื่อนำข้อมูลเหล่านั้นไปใช้ในกระบวนการคาต้าไมนิ่งต่อไป

Title	Developing tool for Data Preparation
Student	Miss Mananya Paksaksri
Student ID.	47066131
Degree	Master of Science in Information Technology
Programme	Information Science
Academic Year	2006
Advisor	Asst.Prof.Dr.Ponrudee Natisopakul

ABSTRACT

Data Preparation is the most important step before passing the data in data mining process. There are large amount of data that is stored in database. The data, which is stored in different sources sometime have different formats. So that is very necessary to prepare all of the selected data before use in data mining process, in order to get a good result.

In this project, we study a related theory of data preparation. Then, we apply a theory to develop the tool for data preparation. This developed tool can help miners to prepare their data before going through data mining process.

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณ ผศ.ดร.พรฤดี เนติโสภากุล อาจารย์ที่ปรึกษาวิชาโครงการพัฒนาระบบงาน ที่ได้กรุณาให้ความรู้ คำปรึกษาและคำแนะนำต่างๆ ที่เป็นประโยชน์ต่อการพัฒนาระบบและสละเวลาในการตรวจสอบแก้ไขข้อบกพร่อง ทำให้งานต่างๆ ผ่านไปได้ด้วยดี

ขอขอบพระคุณบิดาของข้าพเจ้า ที่เป็นพลังให้กับข้าพเจ้าในทุกๆ สิ่ง ถึงแม้ว่าท่านจะไม่มีโอกาสได้เห็นความสำเร็จของข้าพเจ้า ณ วันนี้ แต่คำสอนทุกคำสอน กำลังใจทุกกำลังใจที่ท่านได้ให้ไว้นั้น เป็นแรงผลักดันและเป็นพลังให้ข้าพเจ้ามีกำลังใจในการทำโครงการพัฒนาระบบนี้จนสำเร็จด้วยความสามารถที่ข้าพเจ้ามี นอกจากนี้ข้าพเจ้ายังขอขอบพระคุณมารดา และบุคคลภายในครอบครัว ที่ได้ให้การส่งเสริม สนับสนุน และเป็นกำลังใจในการศึกษาเล่าเรียนตลอดมา ขอขอบคุณสองแม่ครัว, พี่เถี่ยว, ห้องKmake และเพื่อนๆ IS17.1 ทุกคนที่เป็นเพื่อนที่ดีและช่วยเหลือกันตลอดมา

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าโครงงานพัฒนาระบบงานนี้ จะเป็นประโยชน์แก่ผู้ที่สนใจในงานทางด้านการเตรียมข้อมูล ไม่มากก็น้อย อีกทั้งในโครงการพัฒนาระบบงานชิ้นนี้ หากมีข้อผิดพลาดประการใด ข้าพเจ้าขออภัยไว้ ณ ที่นี้

มนันญา ภาคศักดิ์ศรี

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญรูป

สารบัญรูป(ต่อ)

รูปที่ หน้า

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของโครงการ

ในปัจจุบันข้อมูลเป็นจำนวนมากมักถูกเก็บอยู่ในฐานข้อมูล และมีรูปแบบในการเก็บข้อมูลที่มีความสัมพันธ์กัน หรือที่เรียกว่า Relation database แต่การเก็บข้อมูลในฐานข้อมูลคนละแหล่งกันนั้น มักมีรูปแบบในการเก็บข้อมูลที่มีความแตกต่างกัน จึงทำให้การสืบค้นข้อมูลที่มาจากหลายๆ แหล่งข้อมูลเกิดปัญหาขึ้น โดยเฉพาะการนำข้อมูลดังกล่าวมาเพื่อทำกระบวนการดาต้าไมนิ่ง จะทำให้เกิดปัญหามาก ซึ่งจะส่งผลให้ผลลัพธ์จากการทำดาต้าไมนิ่ง นั้น ไม่มีประสิทธิภาพเท่าที่ควร หรือเกิดความคลาดเคลื่อนของข้อมูลที่จะนำไปใช้ในการตัดสินใจ ดังนั้นการเตรียมข้อมูลก่อนที่จะมีการนำไปใช้ในกระบวนการดาต้าไมนิ่ง จึงมีความสำคัญเป็นอย่างมาก เนื่องจากข้อมูลที่ถูกเลือกนำมาใช้นั้น อาจเป็นข้อมูลที่ยังไม่มีความสมบูรณ์ เช่น บาง Record อาจจะมีข้อความที่หายไป หรือมีค่าข้อมูลที่เป็นไปไม่ได้ในความเป็นจริง ดังนั้นการนำข้อมูลมาผ่านกระบวนการเตรียมข้อมูล (Data Preparation) ก่อน จะทำให้ได้ข้อมูลมีความสมบูรณ์มากยิ่งขึ้น และสามารถนำข้อมูลดังกล่าวไปใช้ในกระบวนการต่างๆต่อไปได้อย่างมีประสิทธิภาพ ดังนั้นโครงการพัฒนาระบบงานในหัวข้อเรื่อง การพัฒนาเครื่องมือสำหรับการทำ Data Preparation จึงมีความสำคัญในแง่ที่ว่า ทำให้ผู้พัฒนาระบบได้นำเอาความรู้ที่ได้จากการศึกษาข้อมูลและทฤษฎีที่มีความเกี่ยวข้องมาประยุกต์ใช้ รวมทั้งศึกษาถึงปัญหาต่างๆที่จะเกิดขึ้นเมื่อมีการนำข้อมูลจากหลายๆ แหล่งมาใช้ และนำเอาทฤษฎีที่ได้ทำการศึกษามาประยุกต์ใช้เพื่อแก้ปัญหาที่เกิดขึ้น และทำการพัฒนาแอบพลิเคชั่นเพื่อใช้ในการเตรียมข้อมูลก่อนมีการนำข้อมูลเหล่านั้นไปใช้งาน เพื่อให้ได้ข้อมูลที่มีความถูกต้องและมีประสิทธิภาพเมื่อนำข้อมูลเหล่านั้นไปใช้งานในกระบวนการดาต้าไมนิ่งต่อไป

1.2 วัตถุประสงค์ของโครงการ

โครงการพัฒนาระบบงานในหัวข้อเรื่อง การพัฒนาเครื่องมือสำหรับการทำ Data Preparation มีวัตถุประสงค์หลักดังนี้

1. ศึกษาวิธีการเตรียมข้อมูลให้มีความเหมาะสมก่อนที่จะนำข้อมูลดังกล่าวไปใช้งานในกระบวนการดาต้าไมนิ่ง หรือกระบวนการอื่นๆต่อไป
2. เพื่อให้สามารถนำความรู้ที่ได้ศึกษา มาประยุกต์ใช้ในการแก้ปัญหาต่างๆที่เกิดขึ้นกับข้อมูลเพื่อให้ข้อมูลในฐานข้อมูลนั้นๆมีความเหมาะสมในการทำดาต้าไมนิ่ง ซึ่งเป็นกระบวนการในขั้นตอนต่อไป

3. ทำให้เกิดความเข้าใจในกระบวนการต่างๆ ในเรื่องของการเตรียมข้อมูล ซึ่งเป็นขั้นตอนที่มีความสำคัญมากก่อนที่มีการนำข้อมูลไปใช้งาน

4. โครงการพัฒนาระบบที่ได้จัดทำขึ้นมานั้นสามารถที่จะนำไปใช้ในการเตรียมข้อมูล เพื่อให้ข้อมูลดังกล่าวมีความถูกต้องและเหมาะสมกับกระบวนการที่จะนำข้อมูลเหล่านี้ไปใช้

1.3 ขอบเขตของการพัฒนาระบบ

โครงการพัฒนาระบบงานในหัวข้อเรื่อง การพัฒนาเครื่องมือสำหรับการทำ Data Preparation นี้จะเป็นการนำเอาขั้นตอนในการเตรียมข้อมูล หรือ Data Preparation ซึ่งเป็นขั้นตอนหนึ่งที่มีความสำคัญมากในกระบวนการค้นพบความรู้ในฐานข้อมูล (Knowledge Discovery in Database) มาประยุกต์ใช้ในการพัฒนาระบบ เนื่องจากกระบวนการการเตรียมข้อมูลนั้นมีความสำคัญและมีความจำเป็นเป็นอย่างมากในการที่จะใช้ในการจัดเตรียมข้อมูลก่อนที่จะนำข้อมูลดังกล่าวนั้น เข้าสู่ขั้นตอนของการทำคาต้าไมนิ่ง ต่อไป ดังนั้นในโครงการพัฒนาระบบนี้จึงมีขอบเขตการพัฒนาระบบในเรื่องของการเตรียมข้อมูลเป็นสำคัญ ซึ่งจะใช้ทฤษฎีในเรื่องของ Data Preparation มาใช้จัดการปัญหาต่างๆที่เกิดขึ้นกับข้อมูล โดยกระบวนการที่มีความเกี่ยวข้องกับการเตรียมข้อมูลนั้นประกอบด้วยกระบวนการหลักๆ 3 กระบวนการคือ Data Selection, Data Preprocessing และ Data Transformation

1.4 การดำเนินโครงการ

การดำเนินโครงการพัฒนาระบบงานในหัวข้อเรื่อง การพัฒนาเครื่องมือสำหรับการทำ Data Preparation นั้น มีขั้นตอนในการดำเนินโครงการ ดังต่อไปนี้

1. กำหนดวัตถุประสงค์และขอบเขตของโครงการที่จะพัฒนา

2. ศึกษาทฤษฎีที่มีความเกี่ยวข้อง ซึ่งสำหรับโครงการพัฒนาระบบงานในหัวข้อนี้ จะต้องศึกษาขั้นตอนและหลักการในการทำ Data Praparation เพื่อใช้เป็นทฤษฎีและเป็นแนวทางในการพัฒนาระบบ

3. ทำการออกแบบระบบเพื่อให้ครอบคลุมวัตถุประสงค์ของการพัฒนาระบบ

4. กำหนดแหล่งข้อมูลเพื่อใช้ในการศึกษาในโครงการนี้

5. ทำการพัฒนาระบบซึ่งใช้ในการทำ Data Preparation โดยใช้ทฤษฎีที่ศึกษามาเป็นแนวทางในการพัฒนา

6. จัดทำขั้นตอนและคู่มือในการใช้งานระบบที่พัฒนาขึ้นมา

7. ทำการประเมินผลและวิเคราะห์สิ่งที่ได้รับจากการพัฒนาโครงการนี้ขึ้นมา

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการพัฒนาระบบ

1. ทำให้สามารถนำความรู้ที่ได้จากการศึกษามาประยุกต์ใช้ในการแก้ปัญหาหรือปรับเข้าใช้กับการพัฒนาโครงการทางด้านเทคโนโลยีได้

2. ทำให้เข้าใจหลักการเกี่ยวกับการเตรียมข้อมูลก่อนนำข้อมูลไปใช้ในกระบวนการทำดาต้าไมนิ่ง อย่างละเอียด

3. จากทฤษฎีในเรื่องของการเตรียมข้อมูล (Data Preparation) ทำให้ทราบว่าการเตรียมข้อมูลก่อนการทำดาต้าไมนิ่ง นั้นมีความสำคัญเป็นอย่างมาก ดังนั้น โครงการพัฒนาระบบงานนี้จะอำนวยความสะดวกให้กับผู้ใช้ในการที่จะทำการจัดเตรียมข้อมูลที่มาจากหลายๆแหล่งข้อมูลให้มีความเหมาะสมและมีความถูกต้อง ก่อนที่จะนำข้อมูลไปใช้ในการทำดาต้าไมนิ่ง ซึ่งข้อมูลที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วนั้นจะมีผลทำให้ผลลัพธ์ของการทำดาต้าไมนิ่ง มีความถูกต้องและส่งผลต่อความแม่นยำในการใช้ข้อมูลนั้นๆ ในการตัดสินใจในเรื่องต่างๆต่อไป

4. ทำให้เกิดทักษะและมีความรู้มากยิ่งขึ้นในการวิเคราะห์, ออกแบบ และพัฒนาระบบโดยใช้ทฤษฎีและหลักการต่างๆที่ได้ศึกษามา

บทที่ 2
กระบวนการเตรียมข้อมูลและทฤษฎีที่เกี่ยวข้อง

2.1 ภาพรวมของกระบวนการทำดาต้าไมนิ่งและความสำคัญของการเตรียมข้อมูล

2.1.1 ดาต้าไมนิ่งคืออะไร

ดาต้าไมนิ่งเป็นขั้นตอนในการขุดค้นและวิเคราะห์กลุ่มของข้อมูลต่างๆ เพื่อหารูปแบบข้อมูลที่ยังไม่ถูกค้นพบและทำการค้นหาความหมายของข้อมูลด้วย ซึ่งจะเกี่ยวข้องกับการอธิบายแนวทางของข้อมูลในอดีตและทำนายความเป็นไปในอนาคต เพื่อให้เราทำการตัดสินใจในการดำเนินการทางธุรกิจ ตัวอย่างเช่น ในการค้าปลีกสามารถที่จะทำการจัดตำแหน่งพื้นที่ในการวางขายสินค้าได้ เช่น ถ้าต้องการที่จะเพิ่มยอดขายอุปกรณ์กอล์ฟ ก็จะมีการตรวจสอบรายการขายย้อนหลังในช่วงหลายๆปีที่ผ่านมา และได้ข้อสังเกตมาว่าลูกค้ามักจะซื้อรองเท้าสุภาพบุรุษพร้อมกับการซื้ออุปกรณ์กอล์ฟ ทำให้ห้างสรรพสินค้าตัดสินใจวางตำแหน่งกอล์ฟคลับถัดจากแผนกรองเท้าสุภาพบุรุษเพื่อเป็นการเพิ่มโอกาสในการสร้างยอดขาย เป็นต้น

2.1.2 กระบวนการทำดาต้าไมนิ่ง (Data Mining Process)

โดยทั่วไป เมื่อพูดถึงการทำดาต้าไมนิ่ง คนเรามักจะมุ่งไปที่การทำไมนิ่งและแนวทางในการค้นพบโดยตรง แต่ในความเป็นจริงแล้วนั้น การทำไมนิ่งเป็นเพียงขั้นตอนหนึ่งในกระบวนการทำดาต้าไมนิ่งทั้งหมดเท่านั้น แสดงได้ดังรูปที่ 2.1

2.1.2.1 การระบุวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

ขั้นตอนนี้เป็นขั้นตอนในการกำหนดปัญหาทางธุรกิจให้ชัดเจน ซึ่งเป็นขั้นตอนที่มีความสำคัญกับโครงงานใดๆ ก็ตามที่เกี่ยวข้องกับกระบวนการดาต้าไมนิ่ง ซึ่งขั้นตอนนี้เป็นขั้นตอนเริ่มต้นในการเริ่มทำโปรเจ็ค อย่างน้อยที่สุดสิ่งที่จะได้มาในขั้นตอนนี้ก็คือ การทำความเข้าใจกับปัญหาทางธุรกิจ และระดับความสัมพันธ์ของฝ่ายบริหาร และสามารถระบุวัตถุประสงค์ทางธุรกิจได้ ในขั้นตอนนี้อาจจะก่อให้เกิดปัญหาได้หากไม่มีการจัดการอย่างเหมาะสม บ่อยครั้งที่ไม่มีการกำหนดวัตถุประสงค์ทางธุรกิจ ซึ่งจะทำให้ไมเนอร์ไม่สามารถทราบถึงวิธีการแก้ปัญหาโดยใช้กระบวนการทางดาต้าไมนิ่งได้อย่างเหมาะสมและถูกต้องซึ่งสิ่งที่จะต้องนำมาทำการพิจารณามีดังต่อไปนี้

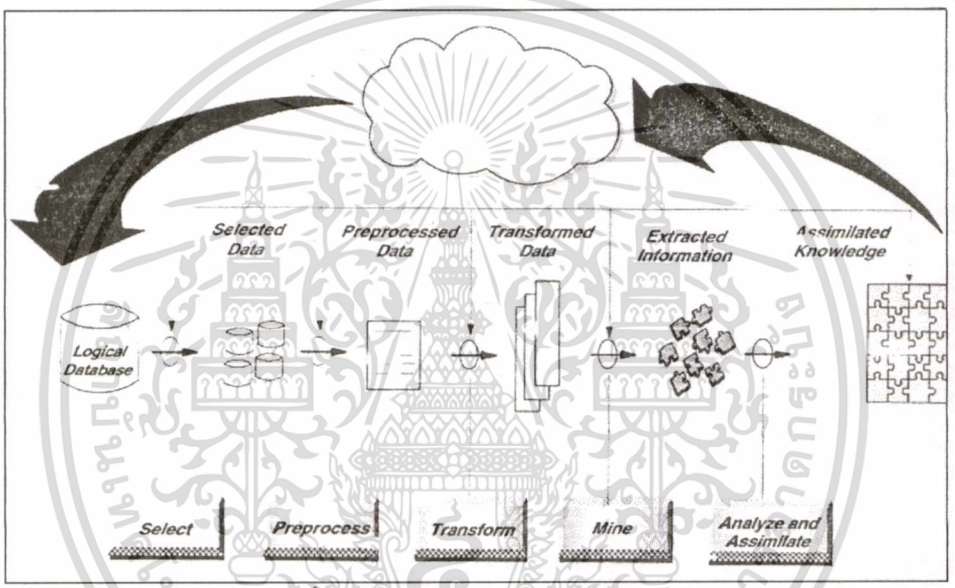
1. การระบุปัญหาหรือตั้งสมมุติฐาน เพื่อนำมาใช้ในการออกแบบผลลัพธ์ให้เป็นที่ยอมรับได้

2. หาเครื่องมือช่วยในการทำงานในช่วงของดาต้าไมนิ่ง ซึ่งในการที่จะเลือกเครื่องมือให้เหมาะสมนั้นต้องมองไปถึงระดับความต้องการของผู้ที่ต้องการความรู้นั้นๆด้วยว่าเป็นผู้ที่มีความเชี่ยวชาญอยู่แล้วหรือไม่มีมาก่อน

3. ประเมินค่าใช้จ่ายที่จะต้องใช้ตลอดการทำงาน

4. พิจารณาประเด็นทางกฎหมาย ข้อบังคับต่างๆที่จำเป็นต้องปฏิบัติตาม และพยายามระบุข้อบังคับทางกฎหมายที่อาจเกิดขึ้นหลังจากมีการนำเอาความรู้ใหม่ไปใช้งานแล้ว

5. แผนการบำรุงรักษาระบบการทำงานที่ต้องมีการเตรียมพร้อมอย่างเหมาะสม เมื่อความรู้ใหม่ถูกนำออกใช้



รูปที่ 2.1 กระบวนการทำดาต้าไมนิ่ง

อย่างไรก็ตาม การทำความเข้าใจและการนิยามความต้องการทางธุรกิจนั้นเป็นสิ่งที่ทำได้ยาก ซึ่งกระบวนการนี้ต้องอาศัยความร่วมมือจากนักวิเคราะห์ทางธุรกิจซึ่งมีความรู้ในด้านธุรกิจและนักวิเคราะห์ข้อมูลซึ่งมีความสามารถที่จะเปลี่ยนวัตถุประสงค์ทางธุรกิจไปประยุกต์เป็นการทำดาต้าไมนิ่งได้

2.1.2.2 การเตรียมข้อมูล (Data Preparation)

การเตรียมข้อมูลนั้นเป็นขั้นตอนที่ต้องใช้เวลาในการทำงานนานมากที่สุด เมื่อเทียบกับกระบวนการอื่นๆ ในกระบวนการดาต้าไมนิ่ง โดยทั่วไปนั้นการเตรียมข้อมูลต้องใช้เวลามากกว่า 60% ของการทำงานในโปรเจ็คหนึ่งๆ ซึ่งการเตรียมข้อมูลนั้น สามารถแบ่งการทำงานเป็นขั้นตอนย่อยๆได้ 3 ขั้นตอน คือ

1. Data Selection เป็นขั้นตอนในการระบุข้อมูล

2. Data Preprocessing เป็นขั้นตอนในการสุ่มข้อมูลและทดสอบคุณภาพของข้อมูล

3. Data Transformation เป็นขั้นตอนในการแปลงข้อมูลให้มีความเหมาะสมกับโมเดลในการวิเคราะห์

2.1.2.3 การทำดาต้าไมนิ่ง (Data Mining)

ขั้นตอนนี้เป็นขั้นตอนการทำไมนิ่งข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล ซึ่งเป็นกระบวนการสำคัญในการเลือกเทคนิคดาต้าไมนิ่งที่เหมาะสม เพื่อหารูปแบบที่ซ่อนอยู่ออกมา และเป็นการประมวลผลข้อมูลตามอัลกอริทึมที่ได้กำหนดไว้

2.1.2.4 การวิเคราะห์ผลลัพธ์ (Analysis of Results)

ในขั้นตอนนี้ จะเป็นขั้นตอนในการแปลงและวิเคราะห์ผลลัพธ์ที่ได้จากการทำไมนิ่งในขั้นตอนที่ 3 ซึ่งบางครั้งอาจมีการนำเอาเทคนิคของการทำ Visualization เข้ามาช่วยในการวิเคราะห์ด้วย

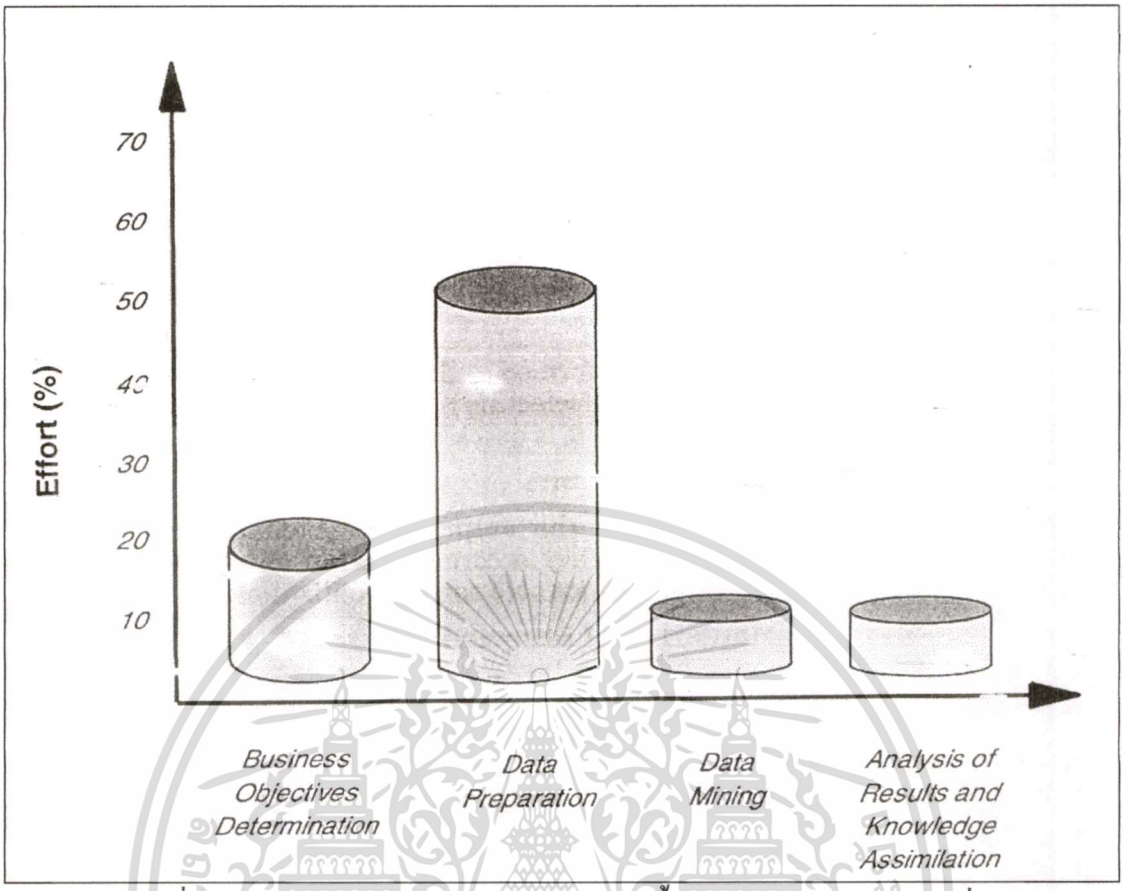
2.1.2.5 การนำความรู้ไปใช้จริง (Assimilation of Knowledge)

ในขั้นตอนนี้จะนำเอาความรู้ทางธุรกิจที่ได้จากขั้นตอนที่ 4 เข้าไปประยุกต์ใช้ในองค์กรทางธุรกิจและระบบสารสนเทศ

โดยจากขั้นตอนต่างๆข้างต้น จะเห็นว่าดาต้าไมนิ่งเป็นเพียงแค่ขั้นตอนหนึ่งในกระบวนการทั้งหมดที่มีอยู่เท่านั้น ส่วนขั้นตอนที่มีความสำคัญในการสืบค้นความรู้จากฐานข้อมูลนั้นเป็นขั้นตอนในการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่ง และเป็นขั้นตอนที่ใช้เวลาในการทำงานมากที่สุดอีกด้วย เนื่องมาจากอาจจะต้องมีการรวบรวมข้อมูลมาจากหลายๆแหล่งด้วยกัน เพื่อที่จะดูความสัมพันธ์ของข้อมูล ซึ่งข้อมูลที่ได้จากขั้นตอนการเตรียมจะต้องมีความชัดเจน และมีความถูกต้องด้วย

2.1.3 ความสำคัญของการเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลนี้เป็นขั้นตอนที่สำคัญมาก เนื่องจากขั้นตอนนี้จะเป็นการจัดเตรียมข้อมูลเพื่อส่งต่อไปยังกระบวนการการไมนิ่ง ถ้าเรามีการเตรียมข้อมูลที่ไม่ดีหรือเกิดข้อผิดพลาดจากการเตรียมข้อมูล จะส่งผลให้การทำไมนิ่งนั้นผิดไปจากวัตถุประสงค์ที่ตั้งไว้ ดังนั้นขั้นตอนนี้จึงสำคัญและจำเป็นต้องใช้เวลาในการทำมากที่สุดถึง 60 เปอร์เซ็นต์ของการทำดาต้าไมนิ่ง ดังแสดงในรูปที่ 2.2



รูปที่ 2.2 แสดงระยะเวลาการทำงานในแต่ละขั้นตอนของการทำดาต้าไมนิ่ง

จากรูปที่ 2.2 ข้างต้น จะเห็นได้ว่ากระบวนการในการทำดาต้าไมนิ่งนั้นถูกแบ่งออกเป็น 4 ช่วง ในช่วงของการกำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination) นั้นจะใช้เวลาและความพยายามในการทำงาน 20 % ของระยะเวลาในการทำดาต้าไมนิ่งทั้งหมด ส่วนช่วงของการทำดาต้ามานิ่ง (Data mining) และช่วงของการวิเคราะห์และนำความรู้ไปใช้จริง (Analysis of Results and Knowledge Assimilation) นั้นใช้เวลาและความพยายามในการทำงานช่วงละ 10 % ของระยะเวลาในการทำดาต้าไมนิ่งทั้งหมด ส่วนช่วงเวลาในการทำการเตรียมข้อมูล (Data Preparation) นั้น ใช้เวลาและความพยายามในการทำงานถึง 60 % ซึ่งมากกว่าครึ่งหนึ่งของระยะเวลาในการทำดาต้าไมนิ่งทั้งหมด ดังนั้นจึงเห็นได้ว่าการเตรียมข้อมูลก่อนการนำไปใช้ในการทำไมนิ่งนั้น มีความสำคัญและต้องใช้ระยะเวลานานมากที่สุดในการทำงาน ผู้วิเคราะห์ข้อมูลจึงต้องให้ความสำคัญกับขั้นตอนการเตรียมข้อมูลมากที่สุด ซึ่งถ้าข้อมูลที่นำไปใช้ในการทำไมนิ่งมีคุณภาพและได้ข้อมูลตรงตามวัตถุประสงค์ที่ต้องการ ผลที่ได้จากขั้นตอนในการทำไมนิ่งก็จะมีประสิทธิภาพ และได้ผลลัพธ์ตรงตามวัตถุประสงค์ที่ต้องการด้วย

2.2 กระบวนการต่างๆในขั้นตอนของการเตรียมข้อมูล

จากที่ได้กล่าวไปในข้างต้นว่า กระบวนการเตรียมข้อมูลนั้น สามารถแบ่งการทำงานเป็นขั้นตอนย่อยๆได้ 3 ขั้นตอน คือ Data selection, Data Preprocessing และ Data Transformation ต่อไปนี้จะกล่าวถึงรายละเอียดของแต่ละขั้นตอนโดยละเอียด

2.2.1 การเลือกข้อมูล (Data Selection)

เป้าหมายของการทำ Data Selection คือ ทำการระบุแหล่งข้อมูลที่มีอยู่ และทำการหาข้อมูลซึ่งมีความจำเป็นในการวิเคราะห์เบื้องต้น เพื่อเป็นการเตรียมการสำหรับการทำคาต้าไมนิ่งต่อไป

ในขั้นตอนนี้จะเป็นการเลือกชุดของข้อมูลที่สามารถใช้การได้ ซึ่งถือว่าเป็นส่วนสำคัญของการทำ Data Mining project ใดๆ ก็ตามให้สำเร็จ ซึ่งโดยทั่วไปชุดข้อมูลเป้าหมายนั้นจะถูกดึงมาจาก 3 แหล่งที่มาด้วยกันคือจาก Data Warehouse, Transaction Database และ Flatfile ซึ่งข้อมูลใน Data Mining ส่วนใหญ่จะต้องให้ข้อมูลนำเข้ามาอยู่ในรูปแบบของ Flatfile หรือ Speadsheet มากกว่า

Database Management System (DBMS) เป็นที่ที่ใช้เก็บข้อมูลและกระทำการกับ transaction data ซึ่งข้อมูลที่อยู่บน DBMS มักจะอยู่ในรูปโครงสร้างแบบ Relation Model ดังนั้นจึงมีลักษณะการเก็บข้อมูลเป็นแบบตาราง (table) จุดประสงค์ของ Relation Model ก็คือลดความซ้ำซ้อนของข้อมูล ถ้ามี table ใดที่มีความซ้ำซ้อนก็จะทำการแตกออกเป็นสองตารางย่อยๆ ในทางกลับกัน วัตถุประสงค์ของการทำคาต้าไมนิ่ง ก็เพื่อเปิดเผยความซ้ำซ้อนของข้อมูล และพยายามที่จะลดความซ้ำซ้อนโดยการรวมความสัมพันธ์เหล่านั้นเข้าด้วยกัน เพื่อที่จะปรับโครงสร้างของข้อมูลที่จะรับเข้ามาอยู่ในรูปแบบที่คาต้าไมนิ่งจะสามารถยอมรับได้

ข้อมูลที่เก็บอยู่ใน Data Warehouse จะมีการเก็บประวัติการใช้ และถูกออกแบบมาเฉพาะเพื่อสนับสนุนการตัดสินใจ ซึ่งโครงสร้างของ Data Warehouse สามารถช่วยลดความซ้ำซ้อนของข้อมูลได้เช่นกัน

ข้อมูลใน Transaction Database จะไม่มีการเก็บข้อมูลซ้ำซ้อน ทำให้สามารถแก้ไขและทำการดึงความรู้ออกมาได้เร็วกว่า ดังนั้นข้อมูลใน Transaction Database จึงต้องมีการปรับเปลี่ยนใหม่ให้เหมาะสมก่อนที่มีการนำไปใช้

จะเห็นว่า Target Data มาจากหลายๆแหล่ง ซึ่งแต่ละแหล่งมีรูปแบบการเก็บข้อมูลที่แตกต่างกัน ทำให้การส่งผ่านข้อมูลทำได้ช้าและใช้เวลานานตัวอย่างเช่น ถ้ามีฐานข้อมูลหนึ่งเก็บโครงสร้างของเพศในลักษณะ Male=1 และ Female =2 แต่อีกฐานข้อมูลหนึ่งมีการเก็บข้อมูลเป็น Male = M และ Female = F แล้วนั้น ก็จะทำให้การส่งผ่านทำได้ช้า เนื่องจากรูปแบบของการเก็บข้อมูลไม่เหมือนกัน

ในการเลือกข้อมูลนั้น จะเปลี่ยนแปลงไปตามวัตถุประสงค์ทางธุรกิจ หรืออาจเปลี่ยนแปลงตามชนิดของแอพพลิเคชั่นที่ใช้ นอกจากนี้Metadata จำเป็นที่จะต้องเข้าใจค่าที่ถูกเลือกด้วยว่า

หมายความว่าอย่างไร Metadata ไม่เพียงแต่เกี่ยวข้องกับนิยามทางธุรกิจเท่านั้น แต่ยังจะต้องเก็บค่ารายละเอียดของชนิดของข้อมูล รวมทั้งค่าที่เป็นไปได้, แหล่งข้อมูลเดิม, รูปแบบของข้อมูล และลักษณะอื่นๆ ด้วย ซึ่งค่าของข้อมูลสามารถแบ่งออกได้เป็น 2 ชนิดหลักๆ คือ

- **Categorical** : ค่าที่เป็นไปได้นั้นมีขอบเขตจำกัด ซึ่งสามารถแบ่งออกได้เป็น 2 ประเภทคือ Nominal และ Ordinal

  - Nominal เป็นค่าที่บอกถึงชนิดของ Object ซึ่งข้อมูลประเภทนี้ลำดับไม่มีความสำคัญ ตัวอย่างของข้อมูลเช่น สถานภาพการสมรส (โสด, แต่งงาน, หย่าร้าง, ไม่ระบุ) เพศ (ชาย, หญิง) และระดับการศึกษา (มหาวิทยาลัย, วิทยาลัย, มัธยมศึกษา) เป็นต้น

  - Ordinal เป็นชนิดของข้อมูลแบบ Categorical ที่ลำดับมีความสำคัญ ตัวอย่างเช่น ระดับความน่าเชื่อถือของลูกค้า (ดี, ปานกลาง, แย่)

- **Quantitative** : เป็นค่าที่มีความแตกต่างที่สามารถวัดได้ระหว่างค่าที่เป็นไปได้ ซึ่งแบ่งออกได้เป็น 2 ประเภท คือ Continuous, Discrete

  - Continuous เป็นค่าจำนวนจริง ตัวอย่างของค่าแบบ Continuous เช่น รายรับ , ค่าเฉลี่ยของการซื้อ, ภาษี

  - Discrete เป็นค่าที่เป็นตัวเลข ตัวอย่างของค่าแบบ Discrete เช่น จำนวนพนักงาน และ เวลาของปี (เดือน, ฤดูกาล) เป็นต้น

ค่าที่ถูกเลือกมาใช้สำหรับการทำดาต้าไมนิ่งนั้น เรียกว่า Active Variable ซึ่งจะเป็นค่าที่ถูกดึงมาใช้เพื่อทำการทำนาย หรือดำเนินงานกับระบบดาต้าไมนิ่งอื่นๆ นอกจากนี้ อัลกอริทึมดาต้าไมนิ่งบางอัลกอริทึมยังยอมให้มีการใช้ค่าข้อมูลเพิ่มเติม ซึ่งส่วนที่เพิ่มเติมนั้นไม่ได้ถูกนำมาใช้ในการทำดาต้าไมนิ่งโดยตรง แต่จะใช้ในการทำ Visualization และการอธิบายผลลัพธ์ที่ได้จากการทำไมนิ่ง

สิ่งสำคัญที่ต้องคำนึงถึงในการทำการเลือกข้อมูลคือช่วงอายุของข้อมูลที่เราเลือกมาใช้งาน นั่นคือจะต้องทำการกำหนดขอบเขตของการเปลี่ยนแปลงของสภาพแวดล้อมทั้งภายในและภายนอก ตัวอย่างเช่น เปอร์เซ็นต์ของลูกค้าจะมีการเปลี่ยนแปลงทุกๆปี ดังนั้น ในการวิเคราะห์จะต้องมีการตรวจสอบข้อมูลดังกล่าวเป็นระยะๆ

ในขั้นตอนของการเลือกข้อมูลนี้ นักวิเคราะห์จะต้องเริ่มมองไปยังอัลกอริทึมดาต้าไมนิ่งที่ตรงกับแอพพลิเคชั่นทางธุรกิจแล้ว ซึ่งการกระทำข้างต้นเป็นสิ่งที่มีความสำคัญที่ควรคำนึงถึงด้วย ในองค์กรใหญ่ๆ ที่ต้องเกี่ยวข้องกับข้อมูลเป็นจำนวนมาก มักจะมีทีมงานเพื่อจัดการกับข้อมูลโดยเฉพาะเพื่อทำหน้าที่ในการค้นหาข้อมูลและจัดการกับข้อมูลเพียงอย่างเดียวเท่านั้น เนื่องจากการทำงานกับข้อมูลที่มีจำนวนมากๆ เป็นหมื่นๆ แสนๆ เรคคอร์ดนั้น มีความยากในการจัดการ และต้องอาศัยเวลาและประสบการณ์ของทีมงานเป็นอย่างมาก

นอกจากนี้ยังมีหลักเกณฑ์ที่ต้องพิจารณาเพิ่มเติมเกี่ยวกับข้อมูลที่จะนำมาใช้ ดังนี้

- **ระดับของข้อมูลที่พิจารณา** สิ่งที่นำมาช่วยตัดสินใจว่าข้อมูลที่นำมาใช้ควรเป็นข้อมูลระดับรายการ (Item) หรือว่าเป็นข้อมูลที่ได้ทำการสรุปแล้ว คือวัตถุประสงค์ในการทำดาต้าไมนิ่ง เช่น

  - การทำไมนิ่งเกี่ยวกับการโทรศัพท์ ถ้าจุดประสงค์ของเราต้องการเน้นไปที่พฤติกรรมการใช้โทรศัพท์ของลูกค้า ข้อมูลที่จัดเก็บโดยปกติแล้วจะมีการจัดเก็บเป็นลักษณะรายละเอียดของแต่ละชุมสาย การเคลื่อนย้ายของอิเล็กตรอนไปยังสวิตชิ่ง ข้อมูลเหล่านี้จะไม่มีประโยชน์เลย เพราะจุดประสงค์ของเราสนใจสิ่งที่อยู่ภายใต้การควบคุมของลูกค้าและมีผลต่อการตลาด ดังนั้นข้อมูลที่เราสนใจจะเป็นเบอร์โทรศัพท์, เวลาเริ่มต้นที่ใช้ในการโทร และเวลาที่ใช้ในการโทรแต่ละครั้ง
  - ข้อมูลที่ยังไม่สรุป ทำให้จัดการได้ยาก รวมทั้งเกิดจำนวนการ Combination สูง เมื่อใช้เทคนิคของ Association Discovery เพราะข้อมูลของร้านค้าปลีกย่อมมีรายการสินค้ามาก ดังนั้นการนำเอาหน่วยวัดในการจัดเก็บสินค้าในคลัง (Stock Keeping Unit) เข้ามาช่วยจะสามารถลดจำนวนการ Combination ลงได้

- **ลักษณะของข้อมูลที่จัดเก็บ** การจัดเก็บข้อมูลด้วยคอมพิวเตอร์ที่แต่ละระบบปฏิบัติการเลือกใช้จะมีความแตกต่างกัน ทำให้ข้อมูลที่นำมาวิเคราะห์มีผลกระทบ

- **ความแตกต่างของข้อมูลแต่ละแหล่ง** เมื่อข้อมูลที่นำมาวิเคราะห์มาจากหลายแหล่ง ซึ่งแต่ละแหล่งมีรูปแบบการจัดเก็บข้อมูลที่ต่างกัน เช่นการวิเคราะห์การใช้ข้อมูลทางโทรศัพท์ เพื่อหาเบอร์โทรศัพท์ที่ใช้ฝากข้อความเข้า Voice Mailbox ด้วยต้นทางและปลายทาง แต่อีกเมืองหนึ่งอาจเก็บเบอร์โทรศัพท์ที่ไม่รู้ด้วยเบอร์ปลายทาง อีกเมืองหนึ่งอาจเก็บเบอร์โทรศัพท์ที่โทรเข้า Voice Mailbox จริงๆ ดังนั้น จึงจำเป็นต้องทำข้อมูลเหล่านี้ให้ออกมาในรูปแบบมาตรฐานเดียวกัน เพื่อที่จะได้เข้าใจถึงความแตกต่างในการเก็บข้อมูลของแต่ละแหล่งได้

- **ข้อมูลที่เป็นข้อความ** ข้อมูลที่จัดเก็บแบบ Text อาจก่อให้เกิดความสับสน เช่น '_no' กับ 'no_' ซอฟท์แวร์ที่ใช้ในการทำดาต้าไมนิ่งย่อมมองข้อมูลเหล่านี้ไม่เหมือนกัน ในทางแก้ไขคือสร้างตารางเก็บค่าที่ถูกต้องและแทนที่ข้อมูลที่นำมาวิเคราะห์ด้วย index ตัวอย่างที่เห็นได้ชัดคือ Relational Database มีการแทนที่ข้อมูลที่เป็น Product_name ด้วย Product_code ซึ่งมีความเป็น unique มากกว่า

ดังนั้นการทำงานในขั้นตอนนี้ คือต้องทำการจัดระเบียบข้อมูลให้อยู่ในรูปแบบที่เป็นมาตรฐานเดียวกัน เพื่อหาคุณลักษณะที่เด่นชัดของชุดข้อมูลนั้นๆ ที่ได้ทำการเลือกมาให้เรียบร้อยเสียก่อน

2.2.2 การปรับปรุงคุณภาพของข้อมูลให้ดีขึ้น (Data Preprocessing)

จุดมุ่งหมายของการทำ Data Preprocessing นั้น เพื่อให้แน่ใจว่าข้อมูลที่ถูกเลือกนั้นมีคุณภาพที่ดี นั้นคือข้อมูลจะต้องถูกปรับปรุงคุณภาพและทำให้ข้อมูลเข้าใจได้ง่ายขึ้น ซึ่งขั้นตอนนี้มีความจำเป็นและมีความสำคัญที่จะต้องกระทำก่อนที่จะทำการไมนิ่งข้อมูล ข้อมูลที่มีคุณภาพจะทำให้ผลลัพธ์ของการทำไมนิ่งถูกต้องและเกิดประสิทธิภาพมากขึ้น รวมทั้งให้การทำไมนิ่งง่ายขึ้นด้วย

ขั้นตอนในการทำ Data Preprocessing นี้เป็นขั้นตอนที่มีปัญหาเกิดขึ้นมากที่สุดในกระบวนการเตรียมข้อมูล โดยทั่วไปเพราะว่าข้อมูลที่ถูกเลือกมาใช้ในการทำดาต้าไมนิ่งส่วนใหญ่นั้นไม่ได้อยู่ในรูปแบบที่ตรงกับความต้องการในการทำดาต้าไมนิ่ง เช่นข้อมูลในรูปแบบ Point of sale หรือข้อมูลที่ถูกจับด้วยขั้นตอนอิเล็กทรอนิก ซึ่งข้อมูลดังกล่าวนี้จะอยู่ในรูปแบบของเอกสารที่ไม่สมบูรณ์ บางครั้งข้อมูลมีความขัดแย้งกัน รวมทั้งเป็นข้อมูลที่มาจากหลายแหล่งข้อมูล ดังนั้นจึงต้องเสียเวลาในการจัดการกับข้อมูลที่ไม่สมบูรณ์นี้ให้มีคุณภาพพร้อมที่จะทำการไมนิ่ง นอกจากนี้การพยายามที่จะรวมเอาข้อมูลจากภายนอกและข้อมูลภายในที่มีอยู่มารวมกันก็เป็นงานที่มีความยากมาก เนื่องจากรูปแบบการเก็บข้อมูลมักไม่เหมือนกัน ถ้าไม่มีการปรับปรุงคุณภาพของข้อมูลที่ดี ผลที่ได้ก็คือข้อมูลที่มีคุณภาพแย่ และบูรณภาพของข้อมูลที่ไม่สมบูรณ์

การทำ Data Preprocessing จะเริ่มจากการพิจารณาโครงสร้างของข้อมูลและวัดคุณภาพของข้อมูล ถึงแม้จะมีหลายแนวทางในขั้นตอนของการทำการปรับปรุงข้อมูล แต่เมื่อพิจารณาดูแล้วนั้นแนวทางส่วนใหญ่มักจะเกี่ยวข้องกับการรวมเอาวิธีการทางสถิติและเทคนิคการทำ Visualization เข้าด้วยกัน

กระบวนการในการปรับปรุงคุณภาพของข้อมูลสามารถแบ่งเป็นกิจกรรมย่อยๆ ได้ 3 กิจกรรม นั่นคือ Data Cleaning, Data Integration และ Data Reduction

2.2.2.1 การปรับคุณภาพของข้อมูล (Data Cleaning)

ในส่วนของการทำ Data Cleaning นี้จะช่วยรับประกันว่าข้อมูลที่ใช้จะมีคุณภาพและมีความถูกต้อง (Integrity) ปัญหาที่พบเมื่อมีการย้ายข้อมูลก็คือ โดยปรกติการอ้างอิง Object หนึ่งๆ จะอ้างอิงด้วยชื่อเพียงชื่อเดียว แต่เมื่อชื่อที่ใช้ในการอ้างอิงข้อมูลมีการเปลี่ยนแปลงหรือมีความคลาดเคลื่อนไปเพียงเล็กน้อยก็จะมีผลทำให้มีการอ้างอิงค่ามากกว่า 1 ค่าภายใน record นั้นๆ โดยความสัมพันธ์ดังกล่าวนี้คอมพิวเตอร์ไม่สามารถแยกแยะความแตกต่างและตรวจสอบได้ ตัวอย่างดังเช่นในตารางที่ 1

ตารางที่ 2.1 แสดงข้อมูลที่มีรูปแบบที่ต่างกันแต่อ้างถึงสิ่งเดียวกัน

John Q. Smith	Mr. Jsmith	Mr. John Smith	Mr.John Smith	Mr.Smith
John Smith	Mr.J.Q. Smith	John Smith	J. Smith	J.Q. Smith
Johnny Smith	Smith,John	Smith,J	Smith,J.Q.	Smith J Q

ข้อมูลที่ถูกเลือกมาจากหลายแหล่งข้อมูล เช่น ในสคมภ์ที่เก็บรายชื่อของเครื่องดื่มเป๊ปซี่ อาจสามารถเก็บค่าได้หลายค่า ดังนี้ "Pepsi", "Pepsi Cola", หรือ "Cola" เป็นต้น โดยค่าเหล่านี้อ้างอิงไปยังเครื่องดื่มชนิดเดียวกัน แต่ว่าคอมพิวเตอร์ไม่สามารถแยกแยะความแตกต่างได้ ต้องทำการเปลี่ยนให้เป็นค่าเดียวให้เป็นมาตรฐานที่เหมือนกัน โดยเฉพาะการเลือกข้อมูลมาจากหลายแหล่ง แต่ละแหล่งข้อมูลอาจจะใช้ค่าหรือรูปแบบการจัดเก็บข้อมูลที่มีความแตกต่างกันแต่หมายถึงค่าเดียวกัน ในการนำมาใช้ต้องปรับด้วย

การ "Clean" อีกแบบก็คือสถานะของข้อมูล (State Data) เช่นข้อมูลที่เกี่ยวกับที่อยู่ที่ใช้ในการส่ง Mail list ต้องตรวจสอบว่าที่อยู่นั้นเป็นที่อยู่ที่ยังถูกใช้งานอยู่หรือเปล่า หรือได้ทำการย้ายที่อยู่แล้ว และสิ่งที่ต้องระวังอีกอย่างคือความผิดพลาดในการสะกดอักษรและชนิดของข้อมูล เพราะถ้าเกิดความผิดพลาดขึ้นคอมพิวเตอร์ก็จะตีความไปเป็นความหมายอื่นที่ต่างออกไป รวมถึงการแปลงหน่วยของข้อมูลที่อ้างอิงลักษณะสำคัญเดียวกันให้มีหน่วยเดียวกันด้วย

วิธีการทำความสะอาดข้อมูลจะใช้ในการแก้ไขปัญหาข้อมูลมีความคลาดเคลื่อน (Noisy Data), ปัญหาข้อมูลที่ขาดหายไป (Missing Data) และปัญหาความไม่สอดคล้องกันของข้อมูล

1.1 ข้อมูลรบกวน (Noisy Data)

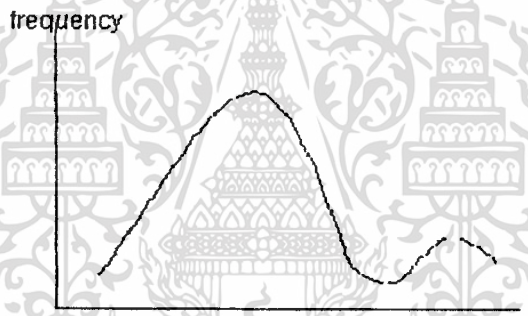
ข้อมูลรบกวน คือ ค่าของข้อมูลของตัวแปรหนึ่งหรือมากกว่าที่มีค่าแตกต่างจากกลุ่มที่ควรจะเป็นสำหรับค่าๆนั้น ซึ่งความคลาดเคลื่อนที่เกิดขึ้นนี้ อาจเกิดจากหลายสาเหตุ เช่น เกิดจากการเก็บข้อมูลผิดพลาด ตัวอย่างเช่น ต้องการเก็บข้อมูลอายุพนักงาน 40 ปี แต่ใส่ข้อมูลผิดเป็น 400 ปี กรณีนี้จะส่งผลให้ข้อมูลคลาดเคลื่อนไป โดยที่ค่าที่คลาดเคลื่อนนี้เรียกว่า outlier ดังนั้น จึงต้องทำการแก้ไขข้อมูลที่ผิดพลาดเหล่านี้ให้ถูกต้อง ซึ่ง outlier ที่เกิดขึ้นนั้น สามารถมองได้ทั้งแง่บวก และแง่ลบ แง่บวกของการค้นพบ outlier นั้น คือสามารถแสดงให้เห็นถึงโอกาสบางอย่างหรือแนวโน้มใหม่ๆ ที่อาจจะเกิดขึ้นในอนาคต แต่ในแง่ลบก็คือแสดงให้เห็นว่าค่าข้อมูลดังกล่าวอาจจะเป็นข้อมูลที่ไม่ถูกต้อง ดังนั้นข้อคำนึงในการจัดการกับ Noisy Data สามารถแบ่งได้เป็น

- หา Record ที่มีข้อมูลซ้ำได้อย่างไร
- หาตำแหน่งที่มีค่าใน Attribute ผิดเจอได้อย่างไร
- อะไรคือสิ่งที่ควรนำมาใช้เพื่อทำให้ข้อมูลราบเรียบ

การหาตำแหน่งที่มีค่าใน attribute ผิดอยู่ให้เจอนั้น จะทำได้ยากมาก หากว่าเซตของชุดข้อมูลที่เลือกมามีขนาดใหญ่มาก แต่ค่าที่ผิดมีปรากฏอยู่เพียงเล็กน้อย ซึ่ง Attribute ที่มักจะผิดอยู่เสมอ คือ attribute ของอายุหรือน้ำหนักที่มีค่าเป็น 0 ทั้งๆที่ค่าไม่มีทางเป็น 0 ได้ เป็นต้น

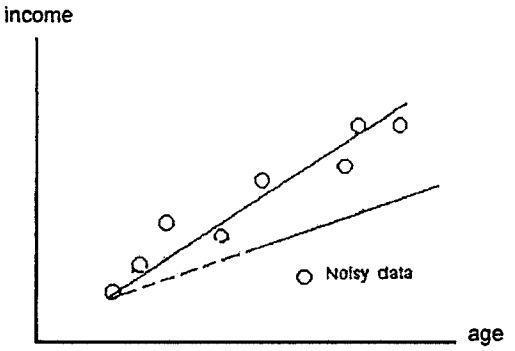
Outlier มีหลายแบบ ซึ่งเกิดได้หลายทาง แบบหนึ่งของ outlier เป็นผลมากจากความผิดพลาดของมนุษย์ เช่นการบันทึกข้อมูลที่ผิดพลาด เช่นใส่ข้อมูลอายุคนจาก 65 เป็น 650 ปี หรือการบันทึกอัตราเงินเดือนที่เป็นจำนวนลบ ดังนั้นในการบันทึกค่าควรมีการตรวจสอบขอบเขตของข้อมูลที่เป็นไปได้ด้วย ในกรณีที่เป็นข้อมูลจาก warehouse การเช็คตรงจุดนี้จะมีการทำตั้งแต่การนำข้อมูลเข้า Warehouse ตั้งแต่แรกแล้ว

Outlier อีกแบบคือ Outlier ที่เกิดขึ้นเมื่อมีการเปลี่ยนแปลงระบบการทำงาน (Operation) แต่ไม่ได้สะท้อนการทำงานนั้นไปยังระบบการทำ Mining ตัวอย่างเช่น มีรหัสของผลิตภัณฑ์ใหม่เข้าระบบการทำงาน ซึ่งจะมองเป็น Outlier สิ่งที่ควรทำคือการปรับปรุง Metadata



รูปที่ 2.3 แสดง Histrogram จำนวนประชากรที่สัมพันธ์กับรายได้

ในบางกรณีการเกิด Outlier แสดงให้เห็นถึงคุณภาพของข้อมูลที่เราใช้ว่าเหมาะสมเพียงไร เช่น เมื่อใช้ Histrogram แสดงในรูปที่ 2.3 เพื่อแสดงจำนวนประชากรในกลุ่มเป้าหมาย โดยแสดงให้เห็นว่าประชากรส่วนใหญ่เป็นคนมีรายได้น้อยและมีเพียงส่วนน้อยที่มีรายได้สูง จากผลที่ได้ถ้าเป็น Outliers ในแง่บวกอาจจะแสดงว่า Outliers ที่เกิดขึ้นนั้นอาจจะเป็นกลุ่มของบุคคลในระดับผู้บริหารที่มีรายได้สูงและ Outliers ในแง่ลบเป็นผลมาจากการเก็บข้อมูลที่ไม่มีประสิทธิภาพ ตัวอย่างเช่นกลุ่มเป้าหมายที่ต้องการหลักต้องการคนที่ออกจากงานแล้ว แต่การเก็บข้อมูลได้ข้อมูลของคนที่ยังทำงานเข้ามาโดยไม่ได้ตั้งใจ หรือแสดงให้เห็นว่าการเก็บข้อมูลนั้น ยังกระจายการเก็บข้อมูลไม่เพียงพอ เก็บข้อมูลเพียงบางกลุ่มเท่านั้น



รูปที่ 2.4 แสดง Noisy Data

จากรูปที่ 2.4 จะเห็นได้ว่ามี Outliers เกิดขึ้น สิ่งที่ต้องทำคือหาสาเหตุของการเกิด Outliers นั้น ในกรณีที่สามารถตรวจสอบหาสาเหตุได้ จะทำการแก้ไขข้อมูลนั้นให้ถูกต้อง แต่ถ้าไม่สามารถหาสาเหตุได้ หรือไม่พบว่าเกิดข้อผิดพลาดเกิดขึ้น ข้อมูลส่วนนั้นก็ควรจะตัดทิ้งไป เพราะถ้านำมาใช้งานอาจทำให้เกิดข้อผิดพลาดในการทำนายได้ จากรูปที่ 2.4 เส้นประแสดงให้เห็นถึงทิศทาง (Trend) ของข้อมูลที่ควรจะเป็น ในกรณีที่นำข้อมูลที่เป็นOutliers มาคิดด้วย จะเห็นได้ว่าทิศทางของข้อมูลที่ควรเป็นนั้น ผิดพลาดไปจากข้อมูลจริง ทำให้ผลการทำนายมีความถูกต้องน้อย ส่วนเส้นตรงธรรมดาแสดงให้เห็นถึงทิศทางของข้อมูลที่ควรเป็นเมื่อตัดข้อมูลที่เป็น Outliers ออก

จะเห็นได้ว่าทิศทางใกล้เคียงกับข้อมูลจริงมากกว่า หรือถ้าไม่ตัดข้อมูลที่เป็น Outliers อาจใช้วิธีการปรับค่าของข้อมูลตรงที่เป็น Outliers ให้เหมาะสม เช่นโดยใช้ค่าที่เป็นค่าเฉลี่ยของจุดก่อนและหลังจุดที่เป็น Outliers หรือจากรูปที่ 2.4 จะใช้จุดบนเส้นที่ได้จากการประมาณทิศทางที่ควรเป็น ที่ตรงกับจุดที่เกิด Outliers เป็นค่าแทน

1.2 การขาดหายของข้อมูล (Missing Data)

Missing Data คือ การที่ข้อมูลบาง attribute ในบางเรคคอร์ดได้ขาดหายไป ซึ่งวิธีแก้ปัญหานี้ก็มีหลายวิธี เช่น ถ้าข้อมูลเกิดขาดหายไปประมาณ 20-30 เปอร์เซ็นต์และ attribute นั้นไม่ค่อยมีความจำเป็นอาจจะใช้วิธีการตัด attribute นั้นออกไปเลย แต่ถ้าข้อมูลนั้นมีความจำเป็นอาจจะต้องใช้วิธีการเติมค่าที่ขาดหายไป

การขาดหายของข้อมูล จะรวมถึงข้อมูลที่ไม่ได้แสดงในข้อมูลที่เลือกมา เช่น ข้อมูลที่เลือกมาหายไปในบางช่วงเวลา และรวมถึงข้อมูลที่เป็นข้อมูลที่ไม่มีอยู่จริง (Invalid data) ที่เกิดจากการตัดทิ้งเมื่อทำการตรวจสอบข้อมูลรบกวน หรือตัดทิ้งเนื่องจากค่าไม่มีความถูกต้อง หรือสูญหายเนื่องจากความผิดพลาดของมนุษย์ ในขั้นตอนการป้อนข้อมูล หรือเกิดจากการนำข้อมูลจากหลายๆแหล่งมารวมกัน ทำให้ข้อมูลไม่ตรงกัน

Missing Data เป็นปัญหาที่จัดการได้ด้วยวิธีบางวิธี ซึ่งส่วนใหญ่แล้วค่า Missing Attribute เป็นการบ่งบอกถึง information ที่หายไป และในการที่จะส่งค่าให้ data mining ทำงานต่อได้นั้น ต้องมีการจัดการกับข้อมูลที่หายไปนี้เสียก่อน สามารถแบ่งวิธีการในการจัดการกับ missing data ออกได้เป็นข้อๆดังนี้

- ทิ้ง record ที่มีค่าที่หายนั้นไป โดยวิธีการนี้เหมาะสมมากที่สุด หากจำนวนเทียบเป็นเปอร์เซ็นต์ของค่าที่หายไปทั้งหมดมีเพียงเล็กน้อยและค่าที่หายไปนั้นมาจาก information ที่หายไปตั้งแต่แรกอย่างแน่นอนจริงๆ

- สำหรับค่าของข้อมูลที่เป็นจำนวนจริง (Real-Values) ให้แทนค่าที่หายไปด้วยค่าเฉลี่ยของ Class (Class mean) ซึ่งส่วนใหญ่แล้ววิธีนี้จะเหมาะสมมากกับ attribute ที่มีค่าเป็นตัวเลข

- แทนที่ค่า attribute ที่หายไปด้วยค่าที่สามารถพบได้ในค่าคงที่ใกล้เคียงกับค่าอื่นๆแต่ต้องไม่เกินค่าสูงสุดของ attribute นั้น วิธีนี้เหมาะสมสำหรับกรณีที่ไม่มีเงื่อนไขหรือแบบที่เป็นตัวเลข

- แทนที่ค่าที่ขาดหายไปด้วยค่า Median ของ attribute นั้นๆ

ในกรณีที่เป็นตัวแปรประเภทตัวเลข(Quantitative Variable) มักจะแทนค่าด้วยค่าที่ใกล้เคียง เช่น ค่าเฉลี่ย หรือค่าที่เป็นฐานนิยม และข้อมูลที่แสดงประเภท (Categorical Variable) อาจจะแทนข้อมูลที่ขาดหายไปด้วยค่าที่เป็นฐานนิยมหรือค่าที่สร้างขึ้นมาใหม่สำหรับคุณลักษณะนั้น เช่นค่า “unknown” เป็นตัวอย่าง หรือถ้ามากกว่านั้น การหาค่ามาแทนทำได้โดยการสร้างแบบจำลอง (Model) ขึ้นมาทำนายค่าที่ควรจะเป็น

การเลือกใช้วิธีใดนั้น ควรนึกไว้เสมอว่าแต่ละวิธีก็จะมี Cost เกิดขึ้น และต้องดูถึงความเหมาะสมและความถูกต้องของผลที่จะได้รับ และคุ้มค่ากับผลที่จะได้รับด้วย

2.2.2.2 การรวมข้อมูล (Data Integration)

ในการเลือกข้อมูลนั้น เราสามารถเลือกข้อมูลเลือกข้อมูลมาจากแหล่งข้อมูลเดียวหรือเลือกข้อมูลมาจากหลายๆแหล่งข้อมูล แต่เพื่อให้ได้ข้อมูลที่มีความหลากหลายและเพื่อนำข้อมูลดังกล่าวไปใช้ประโยชน์ในทางธุรกิจ การเลือกข้อมูลเพื่อนำไปใช้จากหลายๆแหล่งจึงมีประสิทธิภาพมากกว่า เมื่อมีการเลือกข้อมูลมากจากหลายๆ แหล่งข้อมูลก็อาจจะทำให้เกิดปัญหาขึ้นตามมา เช่น ข้อมูลมีความซ้ำซ้อนกัน หรืออาจเกิดปัญหาข้อมูลไม่มีความสัมพันธ์กัน ดังนั้นก่อนที่จะนำข้อมูลไปใช้ในกระบวนการปรับคุณภาพของข้อมูล (Data Preprocessing) จึงต้องทำการรวมข้อมูลจากหลายๆแหล่งข้อมูลให้เป็นข้อมูลก้อนเดียวกันเสียก่อน ซึ่งเราเรียกกระบวนการนี้ว่า Data Integration

ในกระบวนการทำ Data Integration นั้น จะมีการตรวจจับความซ้ำซ้อนของข้อมูล ซึ่งจะใช้วิธีการวิเคราะห์สหสัมพันธ์หรือเรียกว่า Correlational Analysis ซึ่งสูตรของ Correlational Analysis เป็นดังนี้

$$r_{a,b} = \frac{\sum (a - mean(a))(b - mean(b))}{(n-1)\sigma a \sigma b} \quad (2.1)$$

โดยที่ a = ค่าของข้อมูลชุดที่ 1

b = ค่าของข้อมูลชุดที่ 2

$r_{a,b}$ = ค่าความสัมพันธ์สหสัมพันธ์

mean(a) = ค่าเฉลี่ยของข้อมูลชุดที่ 1

mean(b) = ค่าเฉลี่ยของข้อมูลชุดที่ 2

n = จำนวนข้อมูลทั้งหมด

σa = ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูลที่ 1

σb = ค่าเบี่ยงเบนมาตรฐานของชุดข้อมูลที่ 2

ข้อมูลจะมีความสัมพันธ์กันก็ต่อเมื่อค่า $r_{a,b}$ มีค่าเท่ากับ 1 ถ้าข้อมูลทั้งสองชุดไม่มีความสัมพันธ์กัน $r_{a,b}$ จะมีค่าเข้าใกล้ 0 หรือมีค่าเท่ากับ 0 ซึ่งจากสูตรการหาความสัมพันธ์สหสัมพันธ์ข้างต้น สามารถแสดงตัวอย่างในการคำนวณหาความสัมพันธ์ได้ดังต่อไปนี้

a	b
1	2
2	4
3	6

$$\text{Mean (a)} = \frac{1+2+3}{3} = 2$$

$$\text{Mean (b)} = \frac{2+4+6}{3} = 4$$

$$\sigma a = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{(3-1)}} = \sqrt{\frac{1+1}{2}} = 1$$

$$\sigma b = \sqrt{\frac{(2-4)^2 + (4-4)^2 + (6-4)^2}{(3-1)}} = \sqrt{\frac{4+4}{2}} = 2$$

$$r_{a,b} = \frac{(1-2)(2-4) + (2-2)(4-4) + (3-2)(6-4)}{(3-1)(1)(2)} = \frac{2+0+2}{2(1)(2)} = 1$$

จากตัวอย่างการคำนวณข้างต้นสามารถสรุปได้ว่าข้อมูล a และ b มีความสัมพันธ์กัน

2.2.2.3 การลดจำนวนข้อมูล (Data Reduction)

ข้อมูลที่จะนำมาทำไมนิ่งนั้นส่วนใหญ่จะได้มาจากคลังข้อมูล ซึ่งข้อมูลที่อยู่ในคลังข้อมูลนั้นมีขนาดใหญ่และมีความซับซ้อนมาก ดังนั้นถ้านำข้อมูลจากคลังข้อมูลมาทำกาต้าไมนิ่งทั้งหมดจะทำให้เสียเวลาในการทำไมนิ่งเป็นอย่างมาก จึงจำเป็นที่จะต้องทำการลดจำนวนข้อมูล ซึ่งสามารถทำได้ 2 แบบ

3.1 ลดจำนวน attribute (ลดตามแนว column) วิธีการลดจำนวน attribute นั้นสามารถทำได้ 3 วิธี

- วิธีที่ 1 (Step- wise forward seletion) : เริ่มจาก 1 attribute แล้วค่อยๆเพิ่มข้อมูลเข้าไปทีละ attribute ไปเรื่อยจนกว่าค่า error จะเกินกว่าที่จะยอมรับได้
- วิธีที่ 2 (Step – wise backward seletion) : เริ่มจากทุก attribute แล้วค่อยๆตัดออกทีละ attribute จนกระทั่ง error จะเกินกว่าที่รับได้
- วิธีที่ 3 (decision – tree induction) : ใช้ decision tree ในการทำนายว่า attribute ใด ไม่จำเป็นต้องใช้ แล้วจึงตัด attribute นั้นออก

3.2 ลดปริมาณข้อมูล (ลดตามแนว row) วิธีการลดปริมาณข้อมูลจะทำโดยใช้วิธี Sampling ข้อมูลขึ้นมา

2.2.3 กระบวนการแปลงข้อมูลให้สอดคล้องกับโมเดล (Data Transformation)

ในขั้นตอนนี้จะมีการแปลงข้อมูลโดยอาจต้องมีการเพิ่มหรือกำจัดค่า attribute และ instance บางตัวที่อยู่ในชุดของข้อมูลเป้าหมาย ซึ่งงานในขั้นตอนนี้ได้แก่ การทำ Normalization, การแปลงไฟล์และการทำข้อมูลให้ง่ายต่อการทำความเข้าใจ และให้ได้ออกมาเป็นผลลัพธ์ที่ดีที่สุด

การแปลงข้อมูลมีวัตถุประสงค์ 2 อย่างคือ ทำให้ไมนิ่งมีประสิทธิภาพมากขึ้นและทำให้รูปแบบของข้อมูลสอดคล้องกับโมเดลที่จะนำไปใช้ เนื่องจากข้อมูลที่จะนำมาใช้กับดาต้าไมนิ่งในบางครั้งอยู่ในรูปแบบที่ไม่เหมาะสมกับอัลกอริทึมที่เลือกใช้ ดังนั้นจึงจำเป็นที่จะต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับอัลกอริทึมนั้นๆก่อน โดยวิธีการแปลงข้อมูลมีอยู่หลายวิธีซึ่งขึ้นอยู่กับปัญหาของข้อมูล

2.2.3.1 วิธี Normalization : เป็นวิธีที่แปลงข้อมูลให้อยู่ในช่วงๆหนึ่ง เช่น

การทำ Data normalization โดยทั่วไปจะเกี่ยวข้องกับการเปลี่ยนแปลงค่าของตัวเลข เป็นการทำให้มีความน่าสนใจมากขึ้น และเกี่ยวกับการแบ่งหมวดหมู่ โดยสามารถแบ่งการทำงานในการทำ Data Normalization ได้เป็น 4 วิธีคือ

- Decimal Scaling : มีการแบ่งค่าตัวเลขแต่ละค่าตามเลขยกกำลัง 10 เช่น ถ้าค่าตัวเลขใน attribute อยู่ในช่วงระหว่าง –1000 ถึง 1000 แล้วสามารถเปลี่ยนเป็นช่วง –1 ถึง 1 ได้โดยการหารแต่ละค่าด้วย 1000

- Min-Max Normalization : เป็นเทคนิคที่เหมาะสมจะทำเมื่อค่าที่น้อยที่สุดและค่าที่มากที่สุดของ attribute นั้นมีค่าเท่าไรแน่นอน มีสูตรว่า

$$V' = \frac{v\text{-}min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A \qquad (2.2)$$

โดยที่ V' = ค่าข้อมูลที่ได้หลังจากการแปลง

V = ข้อมูลที่จะนำมาทำการแปลง

min_A = ค่าต่ำสุดของข้อมูลใน attribute A

max_A = ค่าสูงสุดของข้อมูลใน attribute A

new_min_A = ค่าต่ำสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

new_max_A = ค่าสูงสุดของข้อมูลที่ต้องการทำการแปลงข้อมูลของ attribute A

ซึ่งการเปลี่ยนแปลงนี้มีประโยชน์กับ Neural Networks เมื่อมีการออกแบบค่าของช่วงเป็น [0,1] ในกรณีนี้สูตรอย่างง่าย คือ

$$newValue = \frac{originalValue - oldMin}{oldMax - oldMin} \qquad (2.3)$$

- Normalization Using Z-Scores : เป็นวิธีในการแปลงค่าให้เป็นมาตรฐานโดยการลบค่าด้วย attribute mean (μ) และทำการหารโดยให้ค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) เป็นตัวหาร มีสูตรว่า

$$newValue = \frac{originalValue - \mu}{\sigma} \qquad (2.4)$$

เทคนิคนี้เหมาะสมในการใช้งานเมื่อได้ทราบค่าสูงสุดและต่ำสุดที่แท้จริง

- Logarithmic normalization : เป็นการแทนที่เซตของค่าด้วยเลขยกกำลัง ซึ่งจะมีผลต่อช่วงของค่า แต่ไม่ทำให้ข้อมูลหายไป เช่น ค่าของ 2 ยกกำลังอะไรจึงจะได้ 64 คำตอบคือ 6 เพราะเมื่อเอา $2^6 = 2x2x2x2x2x2 = 64$ เป็นต้น

2.2.3.2 วิธี Discretization : เป็นวิธีที่แปลงข้อมูลที่ต่อเนื่องให้เป็นข้อมูลที่ไม่ต่อเนื่อง เช่น อุณหภูมิเป็นข้อมูลที่ต่อเนื่อง เราอาจจะจัดแบ่งเป็นช่วงๆ คือ ช่วง 0-20 องศาเซลเซียส เป็นช่วงอากาศเย็น ช่วง 21-30 องศาเซลเซียส เป็นช่วงอากาศอุ่น ถ้าอุณหภูมิเป็น 20.1 องศาเซลเซียสจะถูกปัดเป็น 21 และจัดให้อยู่ในกลุ่มของอากาศอุ่น ซึ่งความจริงแล้ว 20.1 องศาเซลเซียสไม่ต่างกับ 20 องศาเซลเซียส ควรจัดให้อยู่ในกลุ่มของอากาศเย็นมากกว่า ดังนั้นการแก้ไขปัญหานี้สามารถทำได้โดยการแบ่งช่วงให้ละเอียดมากขึ้นแต่ก็ไม่ควรที่จะละเอียดเกินไป

2.2.3.3 วิธี Generalization : เป็นวิธีที่แปลงข้อมูลโดยมองเป็นภาพรวม ตัวอย่างเช่น จัดกลุ่มถนนเป็นเขต จัดกลุ่มเขตเป็นจังหวัด จัดกลุ่มจังหวัดเป็นประเทศ เป็นต้น

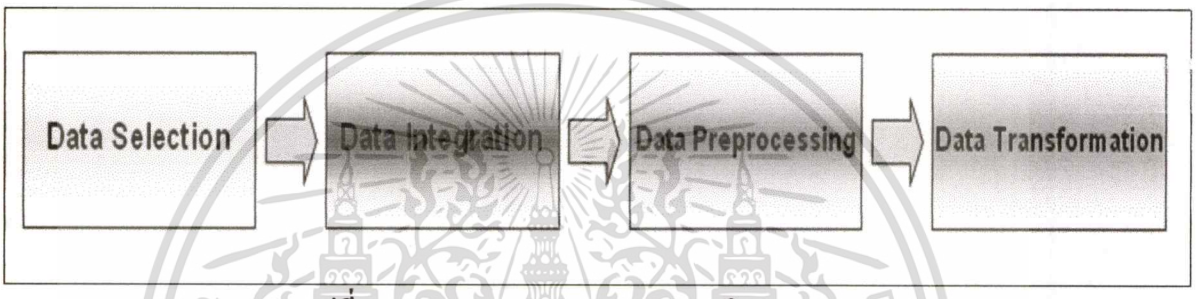
2.2.3.4 วิธี Attribute/Feature construction : เป็นวิธีแปลงข้อมูลโดยการสร้างข้อมูลใหม่จากข้อมูลเดิม เช่น พื้นที่หาได้จากกว้าง x ยาว

เมื่อข้อมูลได้ผ่านกระบวนการเตรียมข้อมูลเรียบร้อยแล้ว ข้อมูลที่ได้จะเป็นข้อมูลที่อยู่ในรูปแบบที่พร้อมที่จะนำไปประมวลผลในกระบวนการของการทำดาต้าไมนิ่งแล้ว ซึ่งกระบวนการทำดาต้าไมนิ่งนั้นเป็นขั้นตอนหนึ่งในกระบวนการทั้งหมดที่มีอยู่เท่านั้น ส่วนขั้นตอนที่มีความสำคัญในการสืบค้นความรู้จากฐานข้อมูลนั้น เป็นขั้นตอนในการเตรียมข้อมูลสำหรับทำดาต้าไมนิ่งและเป็นขั้นตอนที่ใช้เวลาในการทำงานมากที่สุดอีกด้วย ซึ่งได้อธิบายรายละเอียดทั้งหมดไปแล้วในข้างต้น

บทที่ 3
การวิเคราะห์ระบบ

3.1 ระบบงานโดยรวม

จากกระบวนการทำดาต้าไมนิ่งทั้งกระบวนการ โครงงานพัฒนาระบบนี้ได้เลือกนำเอาขั้นตอนของการทำ Data Preparation มาทำเป็นโครงงานพิเศษ ซึ่งระบบงานโดยรวมของการทำ Data Preparation สามารถแสดงได้ดังรูปที่ 3.1



รูปที่ 3.1 ระบบงาน Data Preparation โดยรวม

งานโดยรวมของระบบนี้จะประกอบไปด้วยระบบงานย่อยๆ ดังต่อไปนี้

1. ระบบงานของ Data Selection เป็นส่วนที่จัดการเกี่ยวกับข้อมูล เพื่อให้ผู้ใช้ทำการเลือกข้อมูลจากฐานข้อมูลต่างๆ ไม่ว่าจะเลือกข้อมูลมาจากหนึ่งฐานข้อมูล หรือจากหลายฐานข้อมูล แต่ในส่วนของการทำงานในระบบที่ทำการพัฒนานี้จะสามารถทำการเลือกข้อมูลได้จากสองแหล่งข้อมูลเท่านั้น

2. ระบบงานของ Data Integration เป็นส่วนที่ทำการรวมข้อมูลที่มาจากสองแหล่งข้อมูลให้เป็นกลุ่มข้อมูลเดียวกันเดียว

3. ระบบงานของ Data Preprocessing เป็นส่วนที่ทำหน้าที่ในการปรับปรุงคุณภาพของข้อมูลที่ได้จากการเลือกข้อมูล โดยการตรวจสอบหาข้อผิดพลาดของข้อมูล นั่นคือการตรวจสอบและกำจัดข้อมูลที่ขาดหายไป (Missing Value) โดยในขั้นตอนนี้จะใช้วิธีการแทนที่ข้อมูลที่ขาดหายไปดังนี้คือ

- การใช้ค่า Mean แทนที่ค่าที่ขาดหายไป
- การใช้ค่า Mode แทนที่ค่าที่ขาดหายไป
- การใช้ค่า Median แทนที่ค่าที่ขาดหายไป
- การใช้ค่าที่กำหนดโดยผู้ใช้แทนที่ค่าที่ขาดหายไป
- การใช้การลบเรคอร์ดทั้งเรคอร์ดที่มีค่าที่ขาดหายไป

ซึ่งวิธีการต่างๆที่กล่าวข้างต้นนี้ จะมีการกล่าวโดยละเอียดอีกครั้งหนึ่งในภายหลัง

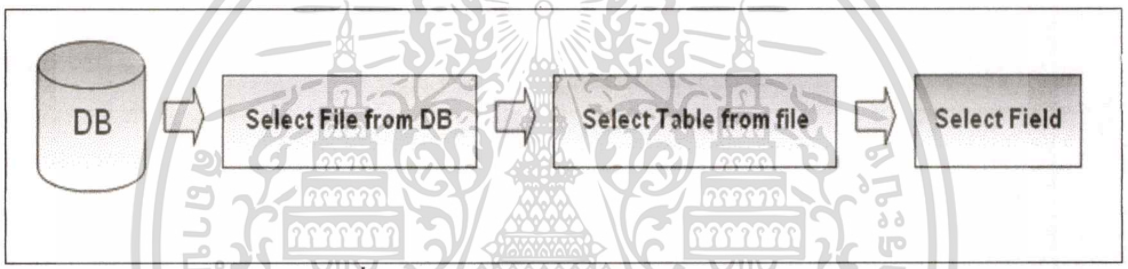
4.ระบบงานของ Data Transformation เป็นส่วนที่ทำหน้าที่ในการแปลงข้อมูลที่ได้จากการทำการปรบปรุงข้อมูลแล้วให้มีความเหมาะสมกับโมเดลดาต้าไมนิ่งที่จะใช้ในการทำไมนิ่ง

3.1.1 ระบบงานของการทำ Data Selection

ในขั้นตอนของการทำ Data Selection นั้น จะเป็นการเลือกข้อมูลจากฐานข้อมูล ซึ่งในการเลือกข้อมูลนั้น สามารถเลือกข้อมูลได้จากฐานข้อมูลเดียว และเลือกข้อมูลสองแหล่งฐานข้อมูล

3.1.1.1 ระบบการเลือกข้อมูลจากฐานข้อมูลเดียว

กระบวนการทำงานของระบบ เมื่อทำการเลือกข้อมูลจากฐานข้อมูลเดียว สามารถแสดงได้ดังรูปที่ 3.2



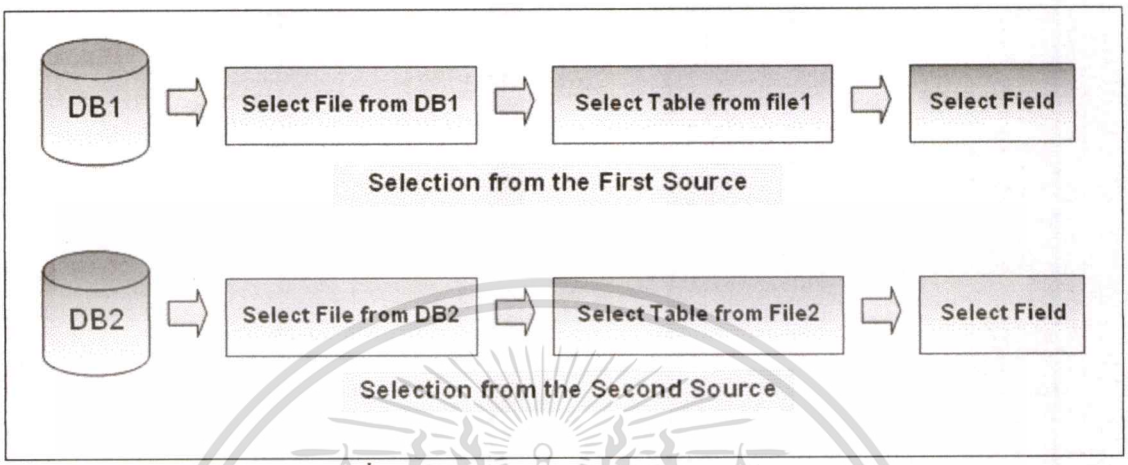
รูปที่ 3.2 การเลือกข้อมูลจากฐานข้อมูลเดียว

รายละเอียดในแต่ละขั้นตอนของรูปที่ 3.2 สามารถอธิบายได้ดังต่อไปนี้

- Select File from DB หมายถึง การเลือกไฟล์จากฐานข้อมูล ซึ่งในที่นี้หมายถึง Microsoft Access 2000
- Select Table from file หมายถึง การเลือกตารางจากฐานข้อมูลที่ได้ทำการเลือกขึ้นมา ในการเลือกตารางขึ้นมาทำงานนั้น สามารถเลือกตารางขึ้นมาทำงานได้เพิ่งหนึ่งตารางเท่านั้น
- Select Field หมายถึง การเลือกฟิลด์จากตารางที่ได้ทำการเลือก ในการเลือกฟิลด์ สามารถเลือกฟิลด์จากตารางขึ้นมาทำงานได้มากกว่าหนึ่งฟิลด์ ขึ้นอยู่กับความต้องการของผู้ใช้งาน

3.1.1.2 ระบบการเลือกข้อมูลจากหลายฐานข้อมูล

กระบวนการทำงานของระบบ เมื่อทำการเลือกข้อมูลจากหลายฐานข้อมูล สามารถแสดงได้ดังรูปที่ 3.3



รูปที่ 3.3 การเลือกข้อมูลจากสองฐานข้อมูล

รายละเอียดในแต่ละขั้นตอนของรูปที่ 3.3 สามารถอธิบายได้ดังต่อไปนี้

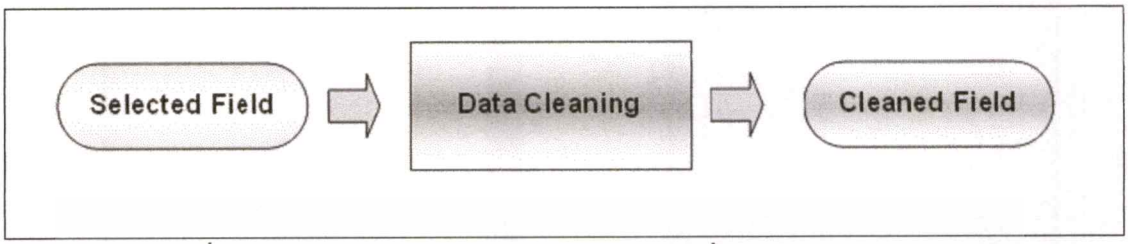
- Select File from DB1 และ Select File from DB2 หมายถึง การเลือกไฟล์จากฐานข้อมูลแต่จะเป็นการเลือกข้อมูลจากแหล่งข้อมูลสองแหล่งโดยข้อมูลทั้งสองแหล่งนี้ จะเป็นข้อมูลจาก Microsoft Access 2000
- Select Table from File1 และ Select Table from file2 หมายถึง การเลือกตารางของแต่ละไฟล์ฐานข้อมูลขึ้นมาทำงาน ซึ่งในการเลือกตารางของแต่ละไฟล์นั้นสามารถเลือกตารางขึ้นมาทำงานได้เพียงไฟล์ละหนึ่งตารางเท่านั้น
- Select Field หมายถึง การเลือกฟิลด์ที่ต้องการตรวจสอบขึ้นมา ซึ่งในการเลือกฟิลด์ขึ้นมาทำการตรวจสอบนั้น ผู้ใช้สามารถทำการเลือกฟิลด์ได้มากกว่าหนึ่งฟิลด์ของแต่ละตารางที่ได้ทำการเลือกขึ้นมา

3.1.2 ระบบงานในขั้นตอนของ Data Preprocessing

ในขั้นตอนของการทำ Data Preprocessing นั้น จะเป็นขั้นตอนในการปรับปรุงคุณภาพของข้อมูลให้ดีขึ้น ซึ่งงานหลักในขั้นตอนนี้คือการกำจัดค่าที่ขาดหายไป (Missing Value) และการทำการรวมข้อมูลในกรณีที่มีข้อมูลมาจากหลายแหล่ง

3.1.2.1 ระบบการปรับปรุงคุณภาพของข้อมูลที่มาจากฐานข้อมูลเดียว

กระบวนการทำการปรับปรุงคุณภาพของข้อมูลที่ได้มาจากฐานข้อมูลเดียวนั้น สามารถแสดงได้ ดังรูปที่ 3.4



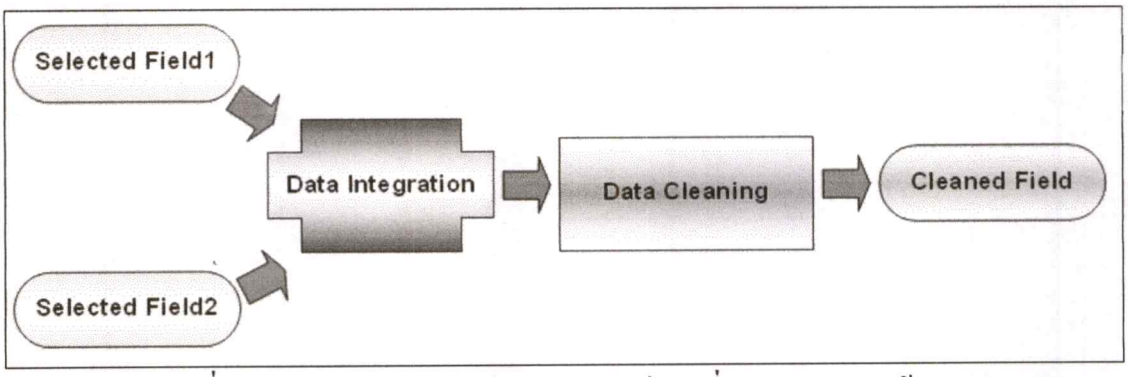
รูปที่ 3.4 การทำการปรับปรุงคุณภาพของข้อมูลที่มาจากฐานข้อมูลเดียว

รายละเอียดในแต่ละขั้นตอนของรูปที่ 3.4 สามารถอธิบายได้ดังต่อไปนี้

- Data Cleaning หมายถึง การทำความสะอาดข้อมูล ซึ่งจะเกี่ยวข้องกับการตรวจสอบและแก้ไขค่าที่หายไปที่ปรากฏอยู่ในฐานข้อมูล ซึ่งผลลัพธ์หลังจากการทำกระบวนการนี้จะได้ข้อมูลที่มีคุณภาพที่ดีขึ้น ซึ่งในการทำความสะอาดข้อมูลนั้น สามารถเลือกใช้งาน
  - การใช้ค่า Mean : จะใช้ค่าเฉลี่ยของฟิลด์ที่ต้องการทำความสะอาดข้อมูลในการแทนที่ทุกเรคอร์ดที่เป็นค่าว่าง
  - การใช้ค่า Mode : จะใช้ค่าฐานนิยมของฟิลด์ที่ต้องการทำความสะอาดข้อมูลในการแทนที่ทุกเรคอร์ดที่เป็นค่าว่าง
  - การใช้ค่า Median : จะใช้ค่าเปอร์เซ็นไทม์ที่ 50 ของฟิลด์ที่ได้ทำการเรียงค่าเรียบร้อยแล้วในการแทนที่ค่าว่าง
  - การใช้ค่าที่ผู้ใช้กำหนด : จะใช้ค่าที่ผู้ใช้กำหนดผ่านทางส่วนติดต่อกับผู้ใช้ในการแทนที่ค่าว่างภายในฟิลด์ที่ต้องการทำความสะอาด
  - การลบเรคอร์ดที่มีค่าว่างอยู่ : จะทำการลบเรคอร์ดที่ทั้งเรคอร์ดที่มีค่าว่างอยู่

3.1.2.2 ระบบการปรับปรุงคุณภาพของข้อมูลที่มาจากหลายฐานข้อมูล

กระบวนการทำการปรับปรุงคุณภาพของข้อมูลที่ได้มาจากหลายฐานข้อมูลนั้น จะต้องมีการทำการรวมข้อมูลดังกล่าวเข้าด้วยกันก่อน สามารถแสดงได้ ดังรูปที่ 3.5



รูปที่ 3.5 การทำการปรับปรุงคุณภาพของข้อมูลที่มาจากสองฐานข้อมูล

รายละเอียดในแต่ละขั้นตอนของรูปที่ 3.5 สามารถอธิบายได้ดังต่อไปนี้

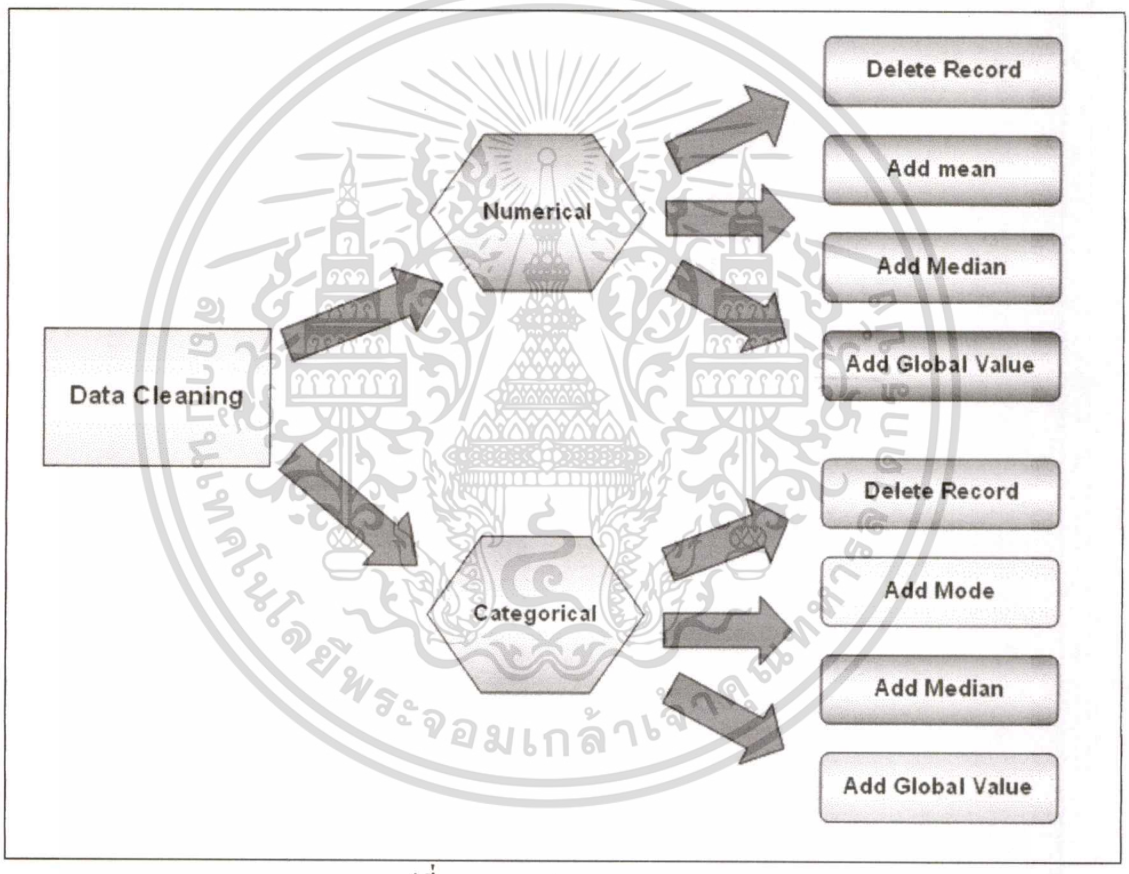
- Data Integration หมายถึง การรวมข้อมูลเข้าด้วยกัน ซึ่งในการรวมข้อมูลนี้จะเกิดขึ้นก็ต่อเมื่อมีการเลือกข้อมูลมาจากหลายฐานข้อมูลซึ่งในระบบที่จะทำการพัฒนานี้จะกำหนดให้มีการเลือกข้อมูลได้สองแหล่งข้อมูล และฟิลด์ที่เลือกขึ้นมาทำงานนั้นเป็นฟิลด์ที่ที่มีการเก็บข้อมูลเดียวกัน เช่น เก็บข้อมูลเพศเหมือนกัน เป็นต้น จึงจะสามารถทำการรวมข้อมูลเข้าด้วยกันได้ การรวมข้อมูลนี้จะทำให้ได้ข้อมูลที่ไม่ซ้ำซ้อนกัน และผลลัพธ์ในการทำดาต้าไมนิ่งก็จะมีประสิทธิภาพที่ดีขึ้น
- Data Cleaning หมายถึง การทำความสะอาดข้อมูล ซึ่งจะเกี่ยวข้องกับการตรวจสอบและแก้ไขค่าที่หายไปที่ปรากฏอยู่ในฐานข้อมูล ซึ่งผลลัพธ์หลังจากการทำกระวนการนี้จะได้ข้อมูลที่มีคุณภาพที่ดีขึ้น

3.1.2.3 ระบบงานของการทำ Data Cleaning

กระบวนการทำงานของขั้นตอนการทำ Data Cleaning สามารถแสดงได้ดังรูปที่ 3.6 ซึ่งจากรูปสามารถแสดงรายละเอียดในแต่ละขั้นตอนได้ดังต่อไปนี้

- การปรับปรุงคุณภาพของข้อมูลประเภท Numerical
  - Delete Record หมายถึง ทำการลบ Record ดังกล่าวทิ้งไปในกรณีที่พบค่าที่หายไป
  - Add Mean หมายถึง ใส่ค่าเฉลี่ยแทนที่ค่าที่หายไป
  - Add Median หมายถึง ทำการแทนที่ค่าที่หายไปด้วยค่าเปอร์เซ็นต์ไทต์ที่ 50 (ค่า Median)
  - Add Global Value หมายถึง แทนที่ค่าที่หายไปด้วยค่าที่ผู้ใช้กำหนด

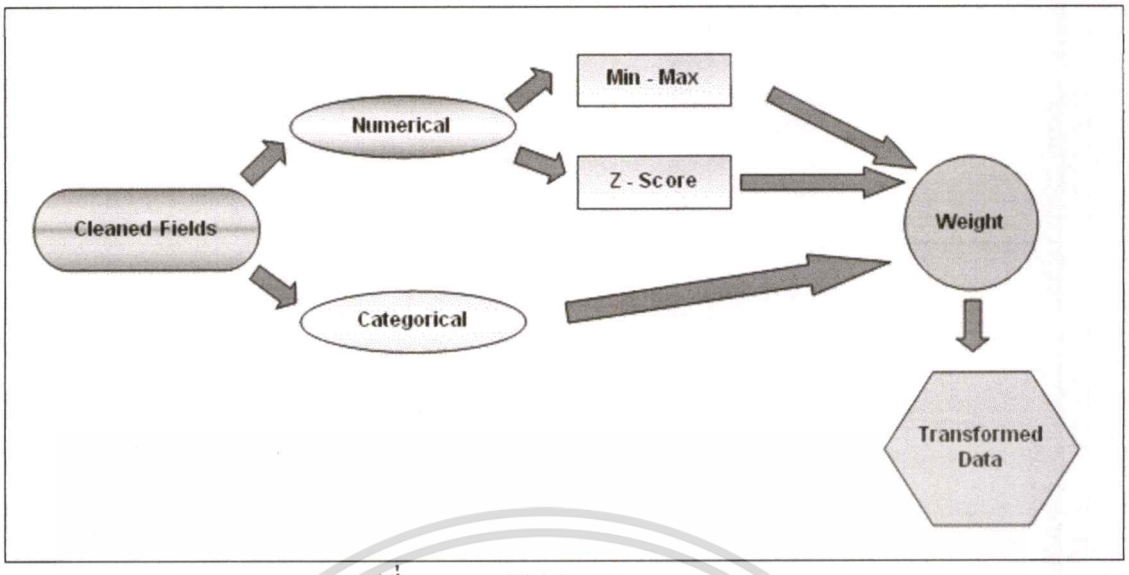
- การปรับปรุงคุณภาพของข้อมูลประเภท Categorical
  - Delete Record หมายถึง ทำการลบ Record ดังกล่าวทิ้งไปในกรณีที่พบค่าที่หายไป
  - Add Mode หมายถึง ใส่ค่าฐานนิยมแทนค่าที่หายไป
  - Add Median หมายถึง ทำการแทนที่ค่าที่หายไปด้วยค่าเปอร์เซ็นต์ไทล์ที่ 50 (ค่า Median)
  - Add Global Value หมายถึง แทนที่ค่าที่หายไปด้วยค่าที่ผู้ใช้กำหนด



รูปที่ 3.6 การทำ Data Cleaning

3.1.3 ระบบงานในขั้นตอนของ Data Transformation

ในขั้นตอนการทำ Data Transformation นี้ จะเป็นการแปลงข้อมูลที่ได้ในขั้นตอนของการทำการปรับปรุงข้อมูล ให้มีความเหมาะสมกับโมเดลการทำดาต้าไมนิ่งที่ทำการเลือก ซึ่งกระบวนการทำงานในขั้นตอนนี้ แสดงได้ดังรูปที่ 3.7



รูปที่ 3.7 การทำ Data Transformation

รายละเอียดในแต่ละขั้นตอนของรูปที่ 3.7 สามารถอธิบายได้ดังต่อไปนี้

- การแปลงข้อมูลประเภท Numerical
  - Min – Max Normalization หมายถึง การแปลงข้อมูลโดยการกำหนดค่าสูงสุด และต่ำสุดใหม่ให้กับข้อมูล ซึ่งในการคำนวณนั้น สามารถแสดงตัวอย่างในการคำนวณได้ ดังต่อไปนี้

A	b
1	2
2	4
3	6

สมมุติตารางที่ต้องการแปลงข้อมูลเป็นดังตารางข้างต้น ในการแปลงข้อมูลโดยใช้วิธีการ Min – Max Normalization สามารถคำนวณได้จากสูตร

$$V' = \frac{V - \text{Min A}}{\text{Max A} - \text{Min A}}(\text{new\_Max A} - \text{new\_Min A}) + \text{new\_Min A}$$

ถ้าสมมุติว่าค่าใหม่ที่ผู้ใช้กำหนด เป็นดังนี้

ค่า Min ที่กำหนดขึ้นมาใหม่ คือ 0

ค่า Max ที่กำหนดขึ้นมาใหม่ คือ 1

ดังนั้นเราสามารถทำการคำนวณหาค่าใหม่ของฟิลด์ A ได้ดังนี้

$$V'_1 = \frac{1-0\ (1-0)+0}{3-1}$$

$$= \frac{1\ (1)+0}{2} = 0.5$$

$$V'_2 = \frac{2-0\ (1-0)+0}{3-1}$$

$$= \frac{2\ (1)+0}{2} = 1$$

$$V'_3 = \frac{3-0\ (1-0)+0}{3-1}$$

$$= \frac{3\ (1)+0}{2} = 1.5$$

- Z – Score Normalization หมายถึง การแปลงข้อมูลโดยการใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานมาใช้ในการแปลงข้อมูล ซึ่งในการคำนวณนั้น สามารถแสดงตัวอย่างในการคำนวณได้ ดังต่อไปนี้

A	b
1	2
2	4
3	6

สมมุติตารางที่ต้องการแปลงข้อมูลเป็นดังตารางข้างต้น ในการแปลงข้อมูลโดยใช้วิธีการ Z – Score Normalization สามารถคำนวณได้จากสูตร

$$newValue = \frac{originalValue - \mu}{\sigma}$$

$$\mu = \frac{1+2+3}{3} = 2$$

$$\sigma = \sqrt{\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3}}$$

$$= \sqrt{\frac{1+0+1}{3}}$$

$$= \sqrt{\frac{2}{3}}$$

$$= 0.8165$$

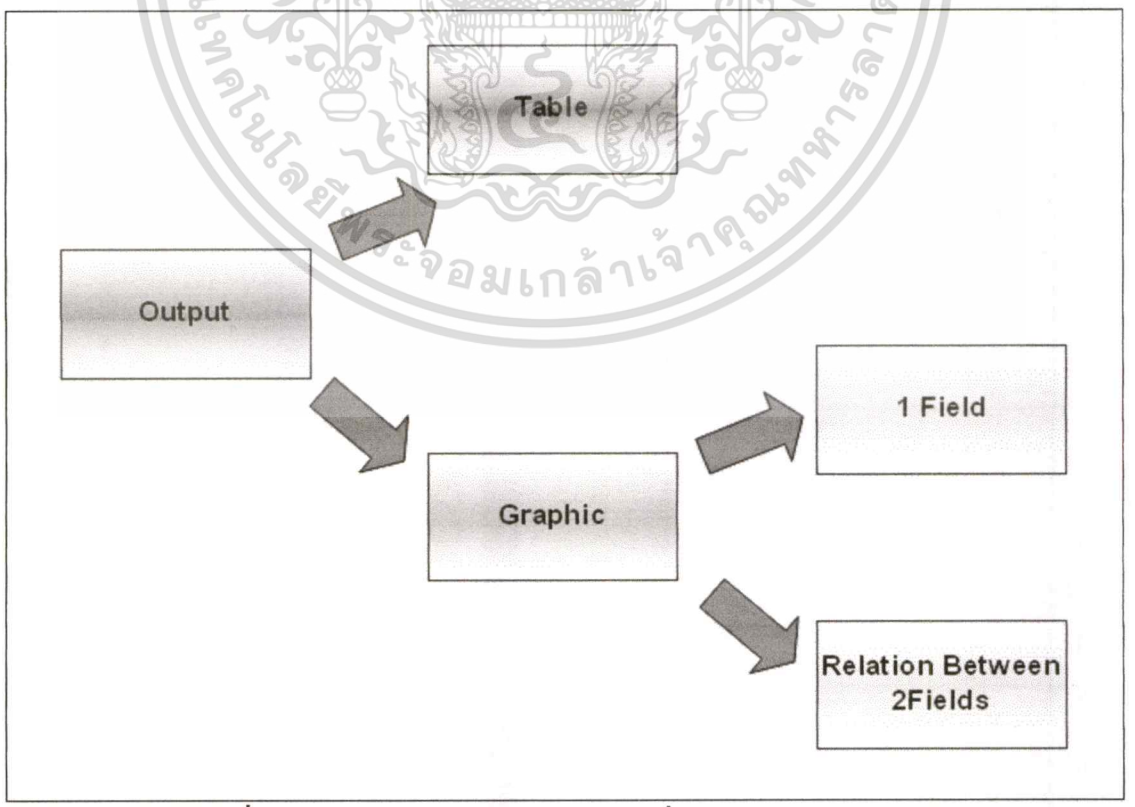
$$V'A(1) = \frac{1-2}{0.8165} = -1.224$$

$$V'A(2) = \frac{2-2}{0.8165} = 0$$

$$V'A(3) = \frac{3-2}{0.8165} = 2.449$$

- การ Weight หรือการให้ค่าน้ำหนักความสำคัญกับฟิลด์
  - ผู้ใช้สามารถให้ค่าน้ำหนักความสำคัญกับฟิลด์ได้ เพื่อกำหนดความสำคัญของเพื่อใช้ในการพิจารณาและการทำนายในขั้นตอนของการทำดาต้าไมนิ่ง

3.1.4 ภาพการแสดงผลลัพธ์ที่ได้จากการแปลงข้อมูล



รูปที่ 3.8 ภาพแสดงวิธีการแสดงผลลัพธ์ที่ได้จากการแปลงข้อมูล

รายละเอียดในแต่ละขั้นตอนของรูป 3.8 สามารถอธิบายได้ดังนี้

- output สามารถแสดงผลลัพธ์ได้ 2 รูปแบบ
  - Table หมายถึง การแสดงผลลัพธ์ในรูปแบบของตาราง ซึ่งจะสามารถเห็นข้อมูลทั้งหมด
  - Graphic หมายถึง การแสดงผลลัพธ์ในรูปแบบของกราฟ ซึ่งสามารถแสดงได้ 2 แบบ คือ แสดงกราฟที่ plot จากฟิลด์ข้อมูลเดียวที่ผู้ใช้เลือก หรือทำการแสดงกราฟจาก 2 ฟิลด์ข้อมูลเพื่อเปรียบเทียบความสัมพันธ์กัน

3.2 Functional Requirement ของระบบ

1. ระบบสามารถดึงข้อมูลมาจากแหล่งข้อมูลที่ได้มาจาก Microsoft Access 2000
2. ระบบทำการตรวจสอบข้อมูลจากไฟล์ Access เพื่อทำการตรวจสอบความถูกต้องของข้อมูลภายในฐานข้อมูล โดยแยกเป็นการดึงข้อมูลมาจากไฟล์ฐานข้อมูลเดียว และมาจาก 2 ฐานข้อมูล
3. ระบบสามารถทำการรวมข้อมูลที่มาจาก Access เพื่อให้เป็นกลุ่มข้อมูลเดียวเพื่อสะดวกในการตรวจสอบ
4. ในการแปลงข้อมูลเพื่อใช้ในการทำไมนิ่ง ระบบนี้จะใช้การแปลงข้อมูลโดยวิธีการ Normalization โดยใช้ Min – Max Value Normalization และ Z-Score Normalization
5. ระบบนี้รับข้อมูลประเภทตัวเลข และตัวอักษร ซึ่งเก็บไว้ในฐานข้อมูลแบบ Relation Database
6. ระบบจะทำความสะอาดข้อมูลโดยใช้การลบเรคอร์ด, การเติมค่า Mean ลงไป, การเติมค่า Median และใส่ค่าที่กำหนดลงไปในช่องของข้อมูลที่ขาดหายไป

3.3 Non-Functional Requirement

1. ระบบไม่สามารถทำการรวมข้อมูลที่มาจากหลายๆไฟล์ข้อมูลที่มีฟอร์แมทความแตกต่างกันได้ เช่น ไฟล์หนึ่งมาจาก Access อีกไฟล์หนึ่งมาจาก MySQL เป็นต้น
2. ระบบไม่สามารถรวมข้อมูลที่มีโครงสร้างข้อมูลแตกต่างกันได้

บทที่ 4
การออกแบบและพัฒนาระบบ

4.1 เครื่องมือที่ใช้ในการออกแบบและพัฒนาระบบ

เครื่องมือที่ใช้ในการออกแบบและพัฒนาเครื่องมือในการเตรียมข้อมูล (Data Preparation Tool) นี้ มีทั้งฮาร์ดแวร์ และซอฟแวร์ที่สำคัญ ดังต่อไปนี้

4.1.1 ซอฟแวร์ที่ใช้ในการออกแบบและพัฒนาระบบ

- ระบบปฏิบัติการ (Operating System) : Microsoft Windows XP Pack2
- ฐานข้อมูล (Database) : Microsoft Access 2000 ,Microsoft Access 2000
- Developing Software : Visual Basic 6.0
- โปรแกรมอื่นๆ : Microsoft Visio 2000, Microsoft Excel, Microsoft Powerpoint

4.1.2 ฮาร์ดแวร์ที่ใช้ในการออกแบบและพัฒนาระบบ

- ซีพียู : Intel core Duo processor T2300 1.66GHz
- แรม : 512MB DDR2
- Harddisk : 60GB 5400rpm

4.2 การออกแบบกระบวนการในการพัฒนาระบบ

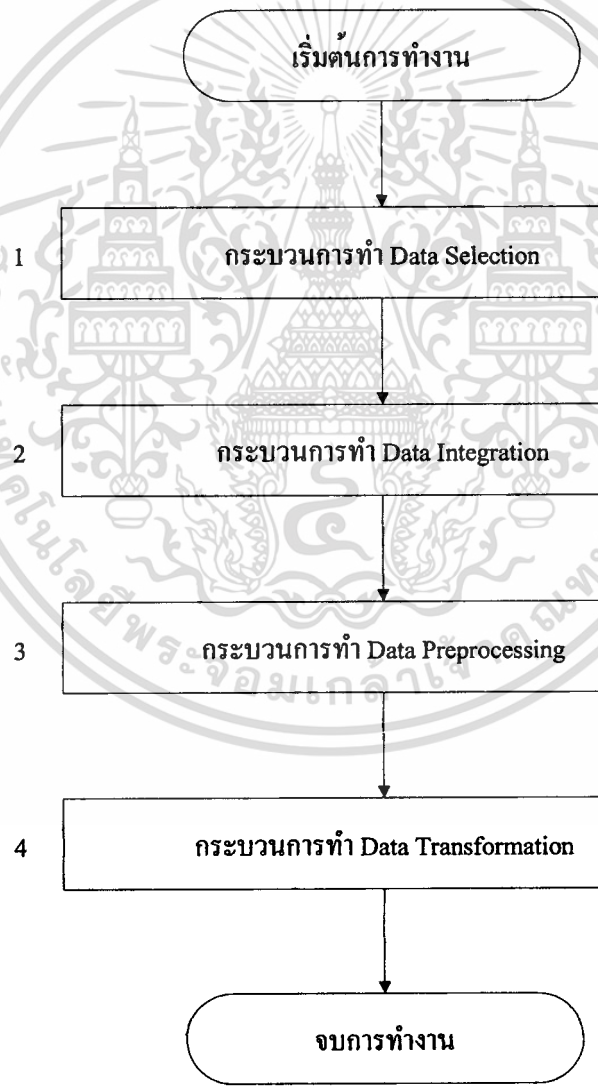
4.2.1 ภาพแสดงการทำงานหลักของระบบการเตรียมข้อมูล (Main Program of Data Preparation)

จากรูปที่ 4.1 แสดงการทำงานโดยรวมของระบบการเตรียมข้อมูล ซึ่งสามารถอธิบายขั้นตอนการทำงานของระบบโดยรวมได้ดังต่อไปนี้

1. กระบวนการทำ Data Selection เป็นขั้นตอนในการเลือกข้อมูลจากแหล่งข้อมูล ซึ่งในการเลือกข้อมูลในขั้นตอนนี้ สามารถเลือกข้อมูลได้จากแหล่งข้อมูลเดียว หรือจากสองแหล่งข้อมูล
2. กระบวนการทำ Data Integration เป็นขั้นตอนในการรวมข้อมูลที่ได้มาจากหลายแหล่งข้อมูลให้เป็นข้อมูลก้อนเดียวกัน ซึ่งในขั้นตอนของการรวมข้อมูลนี้จะเกิดปัญหาขึ้นได้หลายปัญหา ที่พบบ่อยคือความซ้ำซ้อนของข้อมูล และข้อมูลไม่มีความสัมพันธ์

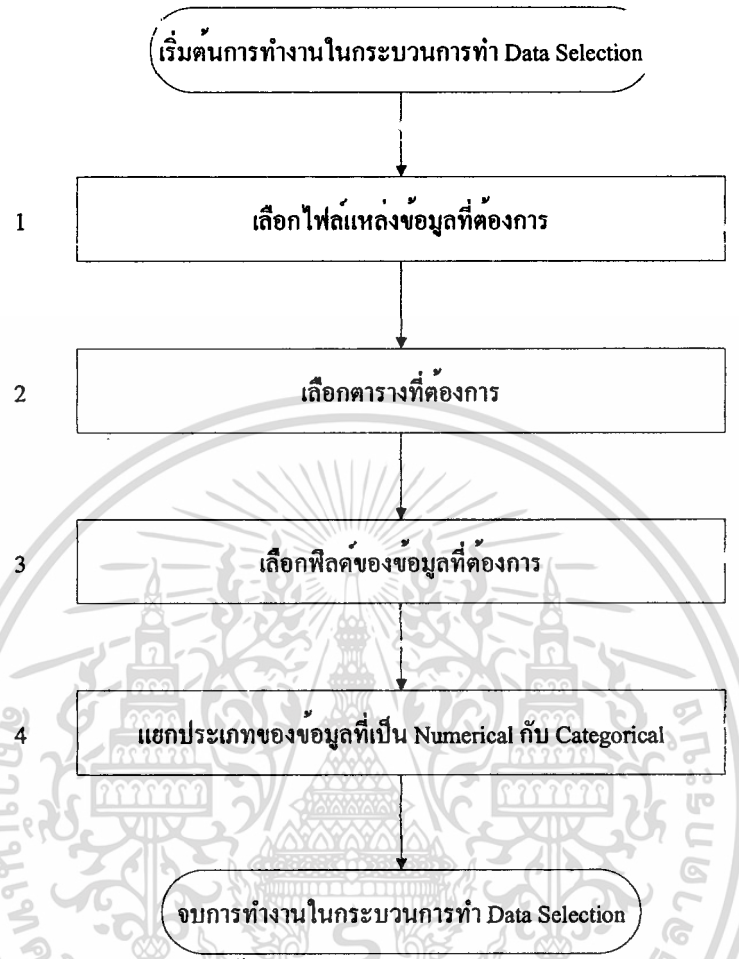
กัน ดังนั้นจึงมีการนำเอาการวิเคราะห์ความสัมพันธ์สหสัมพันธ์ (Correlation Analysis) มาใช้ในการแก้ปัญหาดังกล่าว

3. กระบวนการทำ Data Preprocessing เป็นขั้นตอนในการปรับปรุงข้อมูลให้มีคุณภาพดีขึ้นก่อนการนำข้อมูลดังกล่าวไปใช้ในขั้นตอนของการทำดาต้าไมนิ่งต่อไป ซึ่งในขั้นตอนจะเป็นขั้นตอนในการกำจัดค่าที่ขาดหายไป เพื่อให้สามารถนำข้อมูลไปใช้งานได้อย่างมีคุณภาพ
4. กระบวนการทำ Data Transformation เป็นขั้นตอนในการแปลงข้อมูลที่ได้จากการทำ Data Preparation ให้ได้ข้อมูลที่มีความเหมาะสมกับอัลกอริทึมที่จะเลือกใช้ในการทำดาต้าไมนิ่ง



รูปที่ 4.1 ภาพแสดงขั้นตอนการทำงานหลักของระบบการเตรียมข้อมูล

4.2.2 แผนภาพแสดงหลักการทำงานในกระบวนการเลือกข้อมูล (Data Selection)



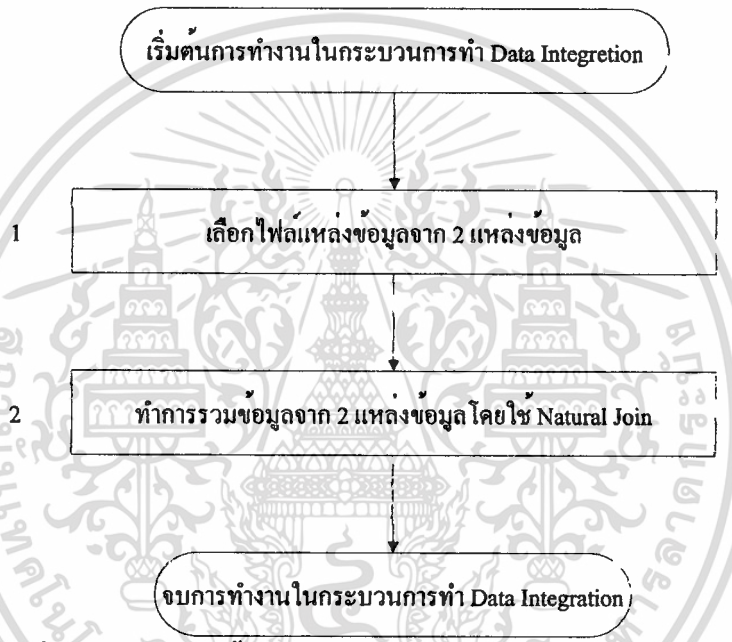
รูปที่ 4.2 ภาพแสดงขั้นตอนการทำงานในกระบวนการเลือกข้อมูล

จากรูปที่ 4.2 แสดงการทำงานของกระบวนการเลือกข้อมูลเข้าสู่ระบบ ซึ่งสามารถอธิบายแต่ละขั้นตอนย่อยๆ ในกระบวนการได้ดังต่อไปนี้

1. เลือกไฟล์แหล่งข้อมูลที่ต้องการ : เป็นการเลือกไฟล์ข้อมูลที่ต้องการเข้ามาใช้งานในระบบ ซึ่งไฟล์ที่ถูกเลือกมาใช้งานนั้น สามารถเลือกไฟล์ข้อมูลนั้น สามารถไฟล์หนึ่งไฟล์ หรือเลือกไฟล์ข้อมูลหลายไฟล์ เข้ามาใช้ในระบบก็ได้ (ในระบบจำลองนี้จำกัดให้สามารถนำเข้าไฟล์ฐานข้อมูลได้สูงสุดสองไฟล์ เพื่อทดสอบการทำงานในส่วนของการรวมข้อมูลเป็นก้อนเดียวกัน)
2. เลือกตารางที่ต้องการ : เป็นการเลือกตารางของข้อมูลที่ต้องการทำมางานในขั้นตอนของกระบวนการปรับปรุงคุณภาพของข้อมูล ซึ่งในการเลือกตารางนั้น จะสามารถเลือกตารางได้เพียงตารางเดียว

3. เลือกฟิลด์ของข้อมูลที่ต้องการ : เป็นการเลือกฟิลด์ที่อยู่ภายในตารางที่ได้ทำการเลือกมาเพื่อนำฟิลด์ที่เลือกไปการประมาณผลในขั้นตอนของการปรับปรุงคุณภาพข้อมูลต่อไป ซึ่งการเลือกฟิลด์นี้ สามารถเลือกได้ตามความต้องการในการใช้งาน
4. แยกประเภทของข้อมูลที่เป็น Numerical กับ Categorical : ทำการแยกประเภทของฟิลด์ เพื่อนำฟิลด์ที่แยกประเภทนี้ไปทำการประมวลผลในขั้นตอนของการปรับปรุงคุณภาพของข้อมูลตามประเภทของข้อมูลนั้นๆ

4.2.3 แผนภาพแสดงหลักการทำงานในกระบวนการรวมข้อมูล (Data Integration)



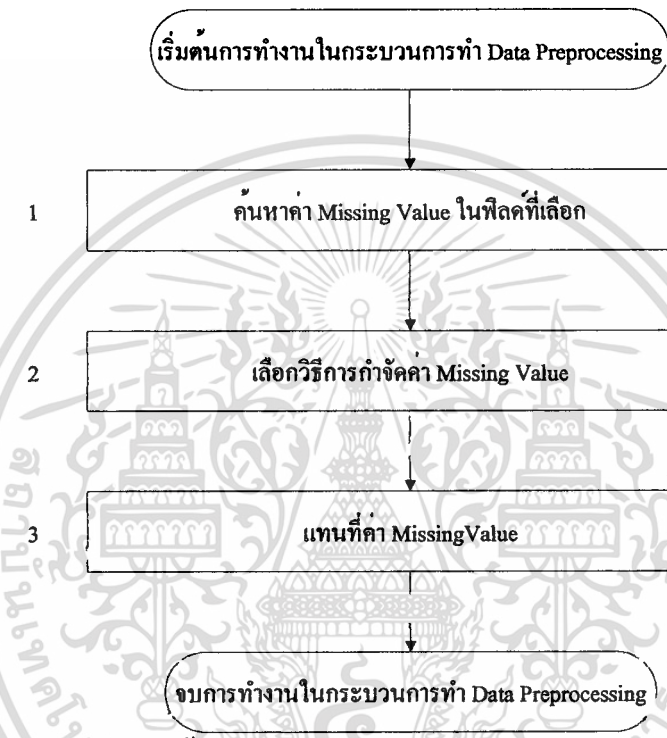
รูปที่ 4.3 ภาพแสดงขั้นตอนการทำงานในกระบวนการรวมข้อมูล

จากรูปที่ 4.3 เป็นกระบวนการทำการรวมข้อมูล (Data Integration) เมื่อมีการเลือกไฟล์ข้อมูลเข้ามาใช้งานในระบบมากกว่าหนึ่งไฟล์ข้อมูล ซึ่งไฟล์ข้อมูลทั้งสองจะต้องถูกรวมเป็นก้อนเดียวกันก่อน จึงจะสามารถนำมาทำงานกับระบบได้ ซึ่งสามารถอธิบายแต่ละขั้นตอนย่อยๆ ในกระบวนการได้ดังต่อไปนี้

1. เลือกไฟล์ข้อมูลจาก 2 แหล่งข้อมูล : ในขั้นตอนนี้มีความเกี่ยวข้องกับกระบวนการเลือกข้อมูล ซึ่งในการเลือกข้อมูลนั้น จะสามารถแบ่งการเลือกข้อมูลออกเป็นสองรูปแบบคือ การเลือกข้อมูลจากแหล่งข้อมูลเดียว และ การเลือกข้อมูลจากหลายแหล่งข้อมูล

2. ทำการรวมข้อมูลจาก 2 แหล่งข้อมูล : ทำการเปรียบเทียบข้อมูลที่ได้มาจากแหล่งข้อมูลทั้งสองแหล่งว่ามีความสัมพันธ์กันหรือไม่ เช่น เป็นแหล่งข้อมูลที่มีโครงสร้างเดียวกันหรือไม่ เป็นต้น แล้วทำการรวมแหล่งข้อมูลทั้งสองเข้าด้วยกันโดยใช้การ Join

4.2.4 แผนภาพแสดงหลักการทำงานในกระบวนการปรับปรุงข้อมูลให้ดีขึ้น (Data Preprocessing)

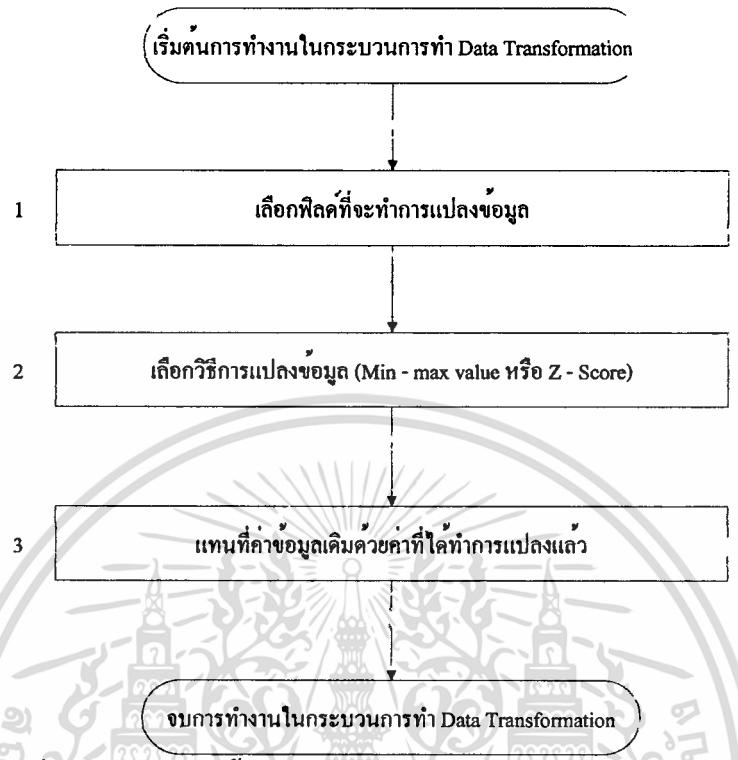


รูปที่ 4.4 ภาพแสดงขั้นตอนการทำงานในกระบวนการปรับปรุงข้อมูลให้ดีขึ้น

จากรูปที่ 4.4 เป็นกระบวนการในการปรับปรุงข้อมูลที่ได้ทำการเลือกให้มีคุณภาพของข้อมูลที่ดีขึ้น ซึ่งสามารถอธิบายแต่ละขั้นตอนย่อยๆ ในกระบวนการได้ดังต่อไปนี้

1. ค้นหาค่า Missing Value ในฟิลด์ที่เลือก : ทำการตรวจสอบฟิลด์แต่ละฟิลด์ที่ได้ทำการเลือกว่าข้อมูลในแต่ละฟิลด์มีค่าว่างหรือว่าค่าที่ขาดหายไปหรือไม่ ซึ่งในการตรวจสอบหาค่าว่างภายในฟิลด์นั้น ระบบจะทำการตรวจสอบทีละฟิลด์ที่ได้ถูกเลือก
2. เลือกวิธีการกำจัดค่า Missing Value : ในการแก้ปัญหาค่าว่างที่ปรากฏอยู่ภายในฟิลด์นั้น มีหลายวิธีการในการแทนที่ค่าว่างเหล่านั้น ขั้นตอนนี้เป็นขั้นตอนในการเลือกวิธีการในการหาค่าเพื่อใช้ในการแทนที่ค่าว่างที่เกิดขึ้น
3. แทนที่ค่า Missing Value : เป็นการนำค่าข้อมูลที่ได้ในขั้นตอนที่ 2 เข้าไปแทนที่ค่าว่างภายในฟิลด์ข้อมูลที่ถูกตรวจสอบ

4.2.5 แผนภาพแสดงหลักการทำงานในกระบวนการแปลงข้อมูล (Data Transformation)



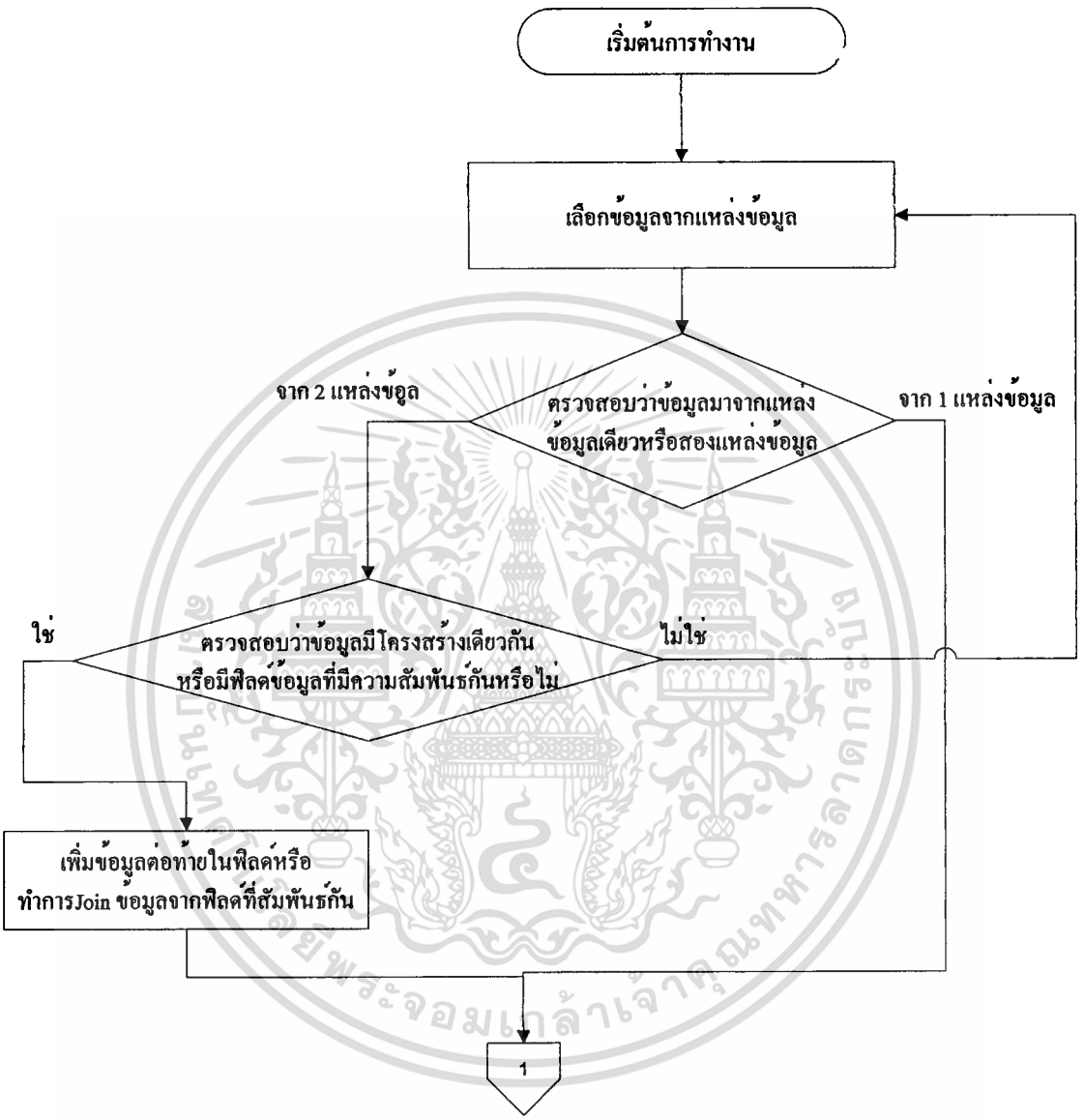
รูปที่ 4.5 ภาพแสดงขั้นตอนการทำงานในกระบวนการแปลงข้อมูล

จากรูปที่ 4.5 เป็นกระบวนการแปลงข้อมูลให้อยู่ในช่วงที่ต้องการเพื่อให้สามารถเข้าใจข้อมูลได้ง่ายขึ้น และให้มีความเหมาะสมกับอัลกอริทึมดาต้าไมนิ่งที่เลือกใช้ ซึ่งสามารถอธิบายแต่ละขั้นตอนย่อยๆ ในกระบวนการได้ดังต่อไปนี้

1. เลือกฟิลด์ที่จะทำการแปลงข้อมูล : ทำการเลือกฟิลด์ของข้อมูลที่ผ่านกระบวนการปรับปรุงข้อมูลแล้ว เพื่อทำการแปลงข้อมูลให้มีความเหมาะสมกับอัลกอริทึมดาต้าไมนิ่ง
2. เลือกวิธีการแปลงข้อมูล : เมื่อทำการเลือกฟิลด์ที่ต้องการที่จะทำการแปลงข้อมูลได้แล้ว ในขั้นตอนนี้จะเป็นขั้นตอนในการเลือกวิธีการในการแปลงข้อมูลภายในฟิลด์เหล่านั้น ซึ่งจะมีวิธีการในการแปลงข้อมูล 2 วิธีการ คือ Min – Max Value Normalization และ Z-Score Normalization
3. แทนที่ค่าข้อมูลเดิมด้วยค่าที่ได้ทำการแปลงแล้ว : เป็นขั้นตอนในการแทนที่ค่าที่ได้จากขั้นตอนที่ 2 มาแทนที่ค่าข้อมูลเดิม

4.3 Flow Chart ของระบบ

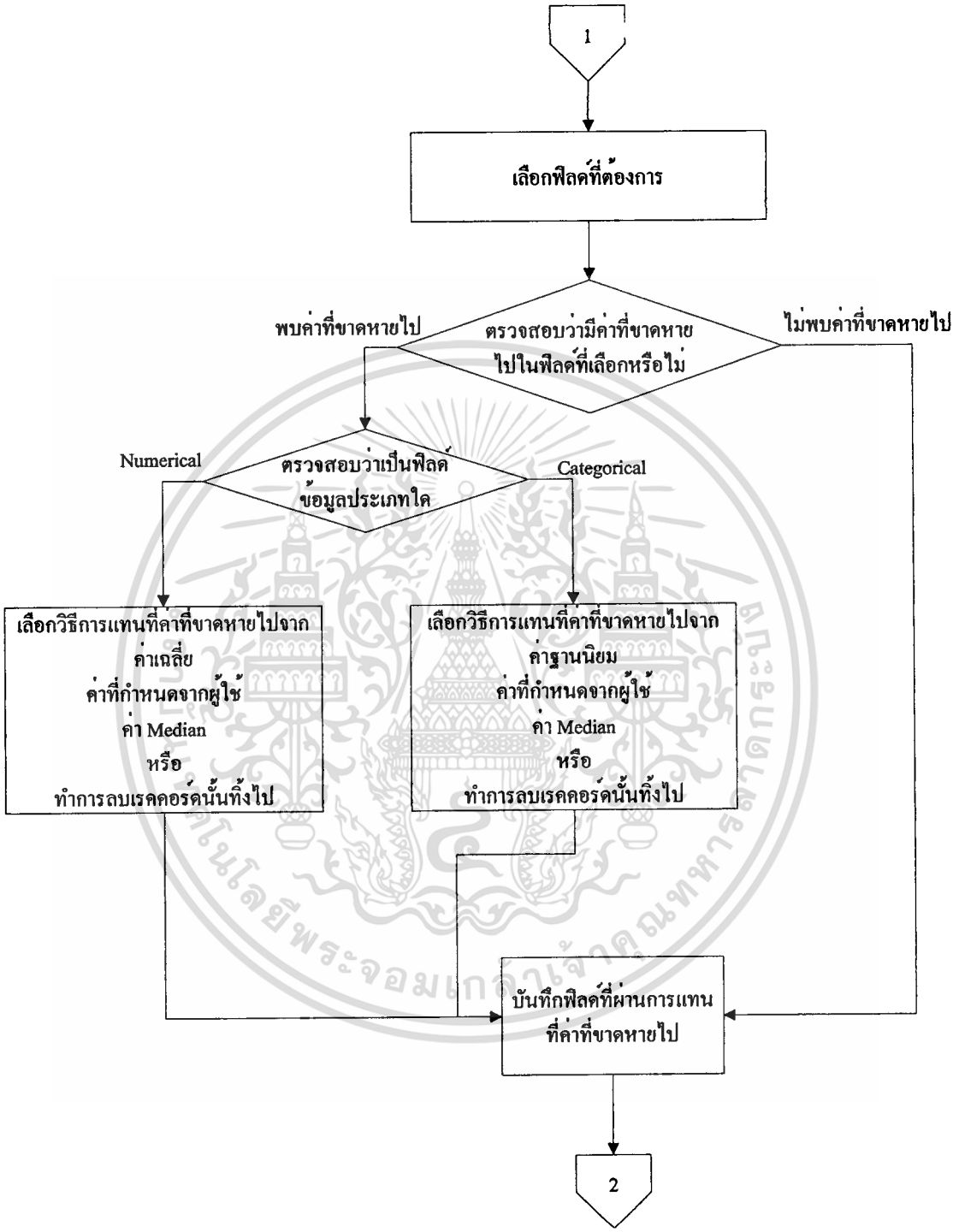
4.3.1 ผังการทำงานในขั้นตอนของการทำ Data Integration สามารถแสดงได้ดังรูปต่อไปนี้



รูปที่ 4.6 ผังการทำงานขั้นตอนของการทำ Data Integration

จากผังการทำงานข้างต้น เมื่อระบบเริ่มการทำงาน ข้อมูลจะถูกเลือกมาจากแหล่งข้อมูลต่างที่กัน หลังจากนั้นระบบจะทำการตรวจสอบว่าฟิลด์ที่ได้ทำการเลือกมาจากฐานข้อมูลแหล่งต่างๆนั้น เป็นฟิลด์ข้อมูลเดียวกันหรือไม่ ถ้าเป็นฟิลด์ข้อมูลเดียวกัน ระบบจะทำการเพิ่มข้อมูลต่อท้ายเข้าไปโดยไม่ต้องทำการสร้างฟิลด์ข้อมูลใหม่ขึ้นมา ทำการเพิ่มข้อมูลต่อท้ายข้อมูลเดิม แต่ถ้าฟิลด์ข้อมูลที่ได้ทำการเลือกต่างกันก็จะทำการสร้างฟิลด์ใหม่ขึ้นมาเพื่อทำการเก็บข้อมูล

4.3.2 ผังการทำงานในขั้นตอนของการทำ Data Cleaning สามารถแสดงได้ดังรูปที่ 4.7



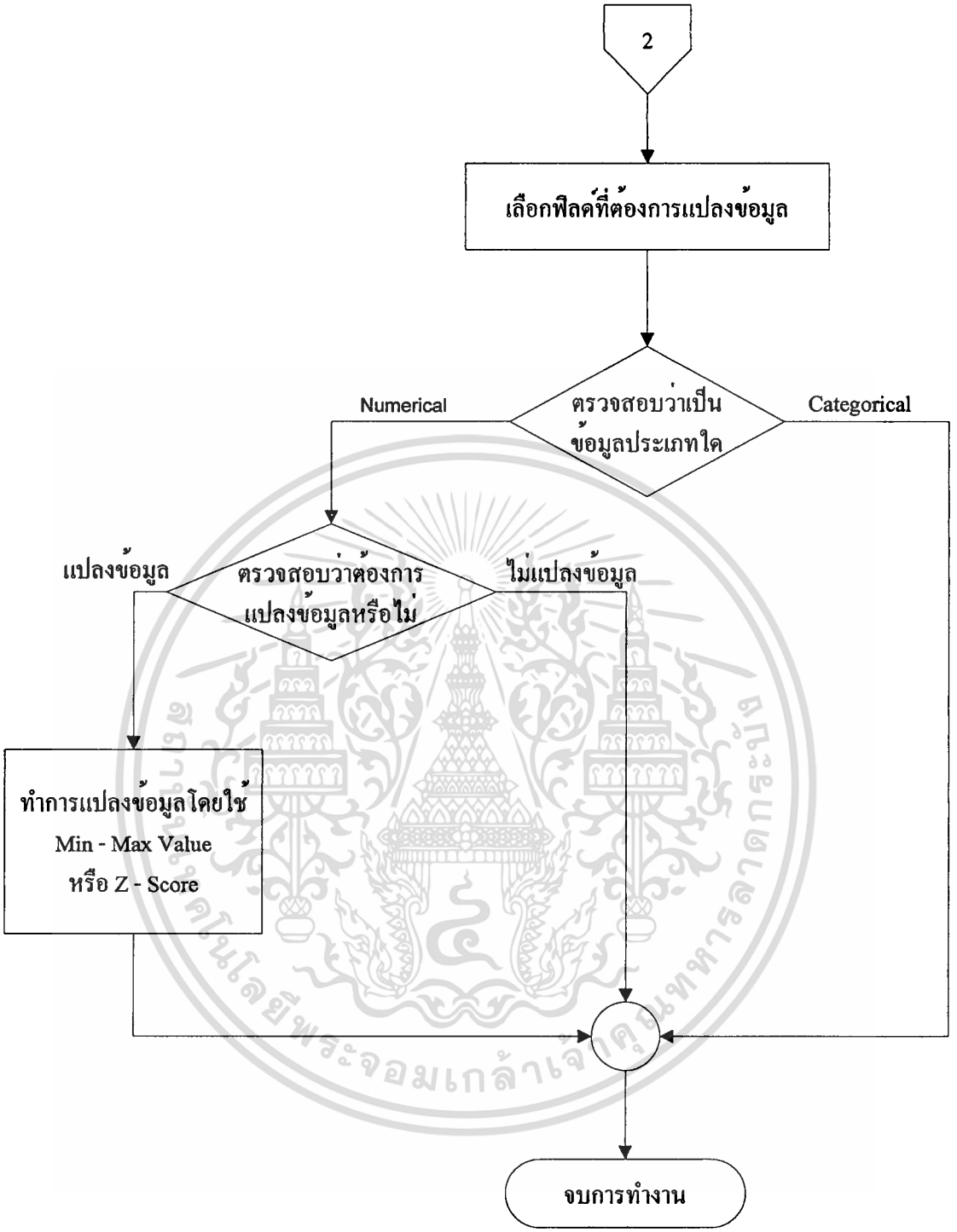
รูปที่ 4.7 ผังการทำงานในขั้นตอนของการทำ Data Cleaning

จากผังการทำงานข้างต้น เป็นขั้นตอนการทำงานของระบบในการทำ Data Cleaning ซึ่งเป็นกระบวนการลำดับต่อจากการทำ Data Integration ซึ่งในขั้นตอนของการทำ Data Cleaning นี้ ระบบจะทำการรับข้อมูลมาจากขั้นตอนของการทำ Data Integration เพื่อนำข้อมูลนั้นๆ มาทำการ

ตรวจสอบว่าข้อมูลมีความถูกต้องหรือไม่ และทำการแก้ไขข้อมูลให้มีความถูกต้องเพื่อส่งข้อมูลเหล่านั้นไปทำงานในขั้นตอนต่อไป โดยเมื่อระบบรับเอาข้อมูลจากขั้นตอนของการทำ Data Integration มาแล้ว จะทำการตรวจสอบว่าแต่ละฟิลด์นั้นๆ มีค่าที่ขาดหายไปหรือไม่ ถ้าฟิลด์ใดมีค่าที่ขาดหายไป ก็จะทำการแยกประเภทของฟิลด์นั้นๆ ว่าเป็น Numerical หรือว่าเป็น Categorical เพื่อทำการกำจัดค่าที่ขาดหายไปตามวิธีการของประเภทของฟิลด์นั้นๆ ซึ่งถ้าเป็นข้อมูลแบบ Numerical ระบบจะทำการแทนที่ค่าที่ขาดหายไปโดยค่าเฉลี่ย, ค่าที่กำหนดขึ้น, ค่า Median หรือว่าทำการลบเรคอร์ดนั้นๆ ทิ้งไป โดยแต่ละวิธีจะขึ้นอยู่กับการเลือกของผู้ใช้งานและความเหมาะสม ซึ่งต้องอาศัยประสบการณ์และการพิจารณาของผู้ใช้ ส่วนถ้าเป็นข้อมูลแบบ Categorical นั้นระบบจะทำการแทนที่ค่าที่ขาดหายไปด้วยค่าฐานนิยม, ค่าที่กำหนดขึ้น, ค่า Median หรือทำการลบเรคอร์ดนั้นๆ ทิ้งไป ซึ่งต้องอาศัยการพิจารณาและประสบการณ์ของผู้ใช้งานเช่นเดียวกัน

4.3.3 ผังการทำงานในขั้นตอนของการทำ Data Transformation สามารถแสดงได้ดังรูป

จากรูปที่ 4.8 เป็นขั้นตอนการทำงานของระบบในขั้นตอนของการทำ Data Transformation ซึ่งในขั้นตอนนี้จะไม่เกิดขึ้น หากผู้ใช้ไม่ต้องการที่จะทำการแปลงข้อมูล แต่หากมีความต้องการในการแปลงข้อมูลเพื่อเตรียมข้อมูลให้มีความพร้อมที่จะทำงานกับอัลกอริทึมของดาต้าไมนิ่ง กระบวนการนี้ก็จะถูกเรียกใช้งาน ซึ่งเมื่อได้รับข้อมูลมาแล้ว ระบบจะทำการตรวจสอบว่าเป็นฟิลด์ข้อมูลประเภทใด ถ้าเป็นแบบ Numerical ก็จะทำการแปลงข้อมูลโดยใช้วิธีการ Min-Max Value หรือ Z-Score ซึ่งระบบจะใช้วิธีการใดมาทำการแปลงข้อมูลนั้นก็จะขึ้นอยู่กับผู้ใช้งานว่าต้องการที่จะใช้วิธีการใด

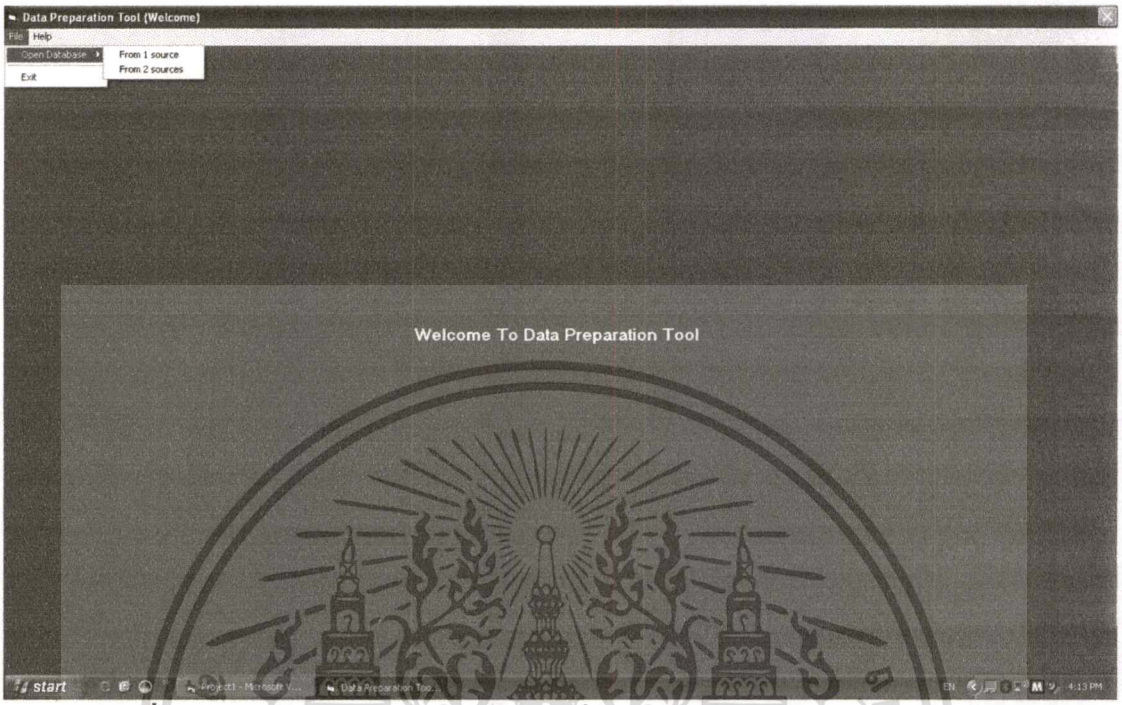


รูปที่ 4.8 ผังการทำงานในขั้นตอนของการทำ Data Transformation

4.4 การออกแบบหน้าจอติดต่อกับผู้ใช้ (User Interface)

การออกแบบส่วนติดต่อกับผู้ใช้ เป็นขั้นตอนที่มีความสำคัญ เนื่องจากส่วนติดต่อกับผู้ใช้นี้ จะทำหน้าที่ในการติดต่อกับผู้ใช้โดยตรงในการทำงานกับเครื่องมือที่พัฒนาขึ้น ดังนั้นเพื่อให้เครื่องมือสำหรับการทำ Data Preparation มีความสามารถในการรองรับการใช้งานจากผู้ใช้และครอบคลุมฟังก์ชันการทำงานทั้งหมดของระบบ จึงมีส่วนติดต่อกับผู้ใช้หรือหน้าจอที่สำคัญดังนี้

4.4.1 การออกแบบหน้าเมนูหลักของเครื่องมือสำหรับการทำ Data Preparation



รูปที่ 4.9 การออกแบบหน้าจอหลักของเครื่องมือสำหรับการทำ Data Preparation

การออกแบบหน้าเมนูหลักของระบบการทำ Data Preparation นี้ จะมีเมนู หลัก 2 เมนู คือ เมนู File และ เมนู Help ดังรูปที่ 4.9 โดยส่วนประกอบของเมนูหลักทั้ง 2 เมนูมีดังต่อไปนี้

เมนู File ประกอบด้วยเมนูย่อย

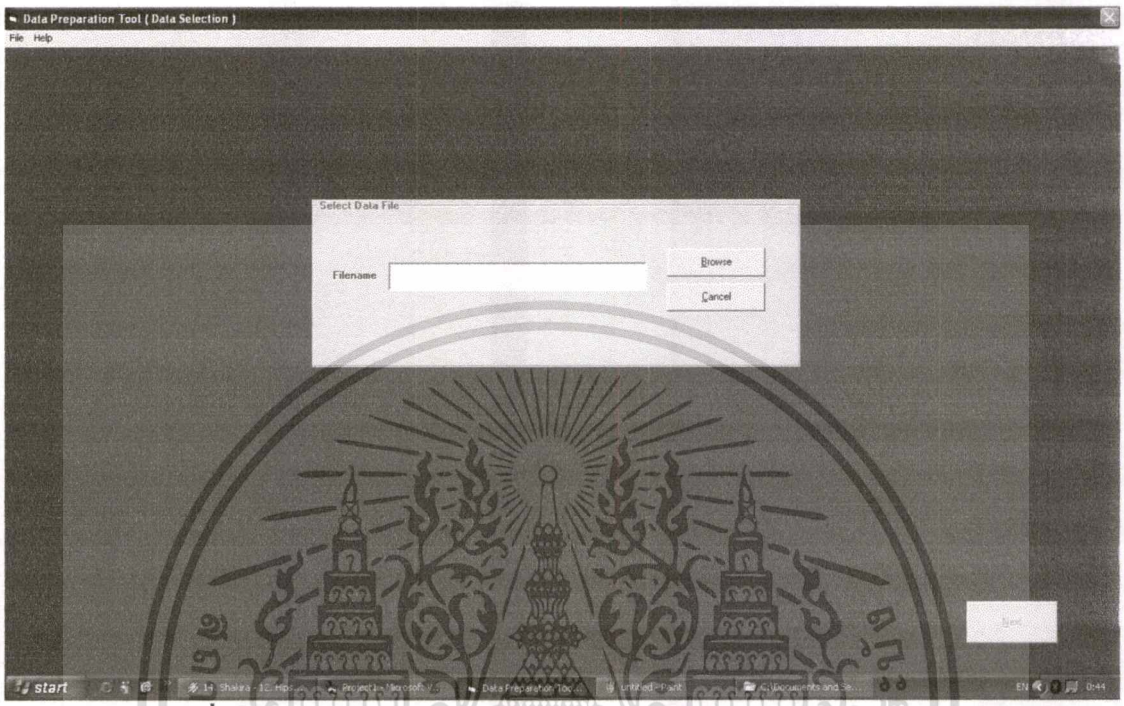
- Open Database เป็นเมนูที่ใช้สำหรับการเปิดไฟล์ฐานข้อมูลที่ผู้ใช้ต้องการ ซึ่งในเมนู Open Database นี้ จะมีเมนูย่อยอีก 2 เมนูคือ
  - From 1 Source เป็นเมนูย่อยที่ทำหน้าที่เปิดไฟล์ข้อมูลจากฐานข้อมูลเดียว
  - From 2 Sources เป็นเมนูย่อยที่ทำหน้าที่เปิดไฟล์ข้อมูลจากสองฐานข้อมูล
- Exit เป็นเมนูที่ใช้ในการออกจากระบบ

เมนู Help ประกอบด้วยเมนูย่อย

- About Data Preparation แสดงรายละเอียดย่อๆ เกี่ยวกับการทำ Data Preparation
- About This Tool แสดงรายละเอียดเกี่ยวกับระบบงานที่พัฒนา

4.4.2 การออกแบบหน้าจอเพื่อทำการเลือกไฟล์ข้อมูล

4.4.2.1 การออกแบบหน้าจอเพื่อทำการเลือกไฟล์ข้อมูลจากฐานข้อมูลเดียว

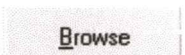


รูปที่ 4.10 การออกแบบหน้าจอการเลือกไฟล์ข้อมูลจากฐานข้อมูลเดียว

จากรูปที่ 4.10 แสดงการออกแบบหน้าจอเพื่อให้ผู้ใช้ทำการเลือกเปิดไฟล์ข้อมูลที่ผู้ใช้ต้องการ ซึ่งจะมีปุ่ม Browse เพื่อให้ผู้ใช้เปิดหน้าต่างเพื่อทำการเลือกไฟล์ข้อมูลขึ้นมา ส่วนปุ่ม Cancel เป็นปุ่มที่ทำหน้าที่ยกเลิกไฟล์ข้อมูลที่ผู้ใช้ได้เลือกขึ้นมา ใช้ในกรณีที่ผู้ใช้ต้องการยกเลิกไฟล์ข้อมูลที่เลือกก่อนหน้าและทำการเลือกไฟล์ข้อมูลใหม่ ส่วนปุ่ม Next จะใช้เมื่อทำการเลือกไฟล์ข้อมูลเรียบร้อยแล้ว และต้องการทำงานในขั้นตอนต่อไป

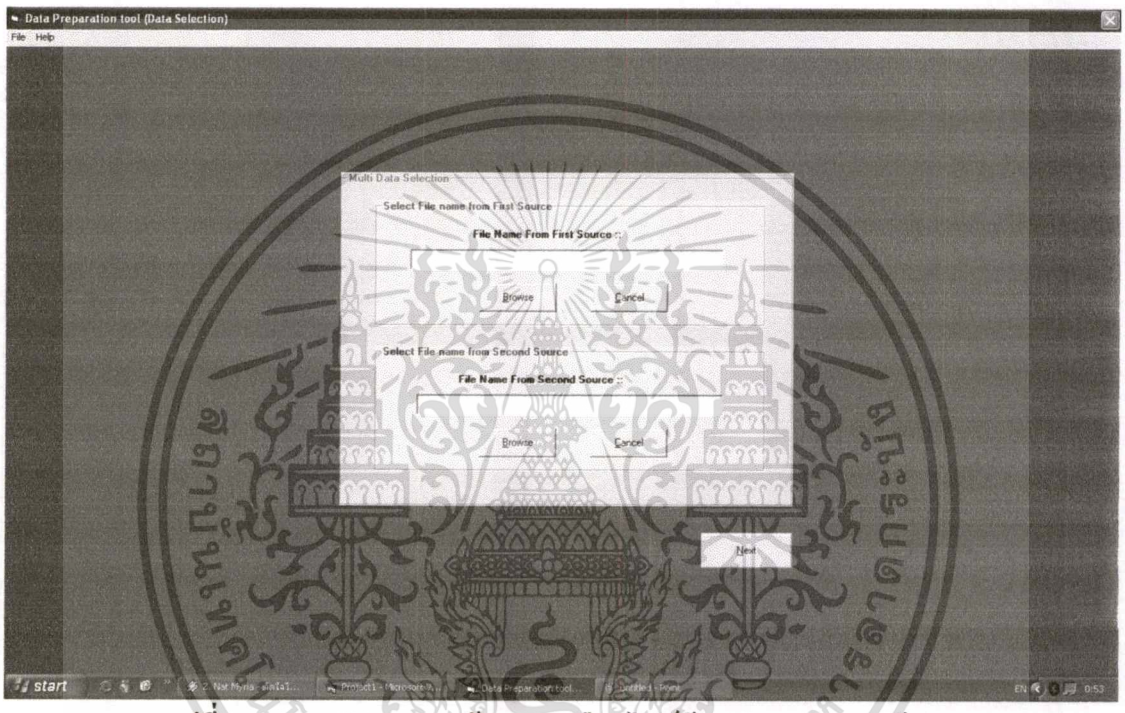
4.4.2.2 การออกแบบหน้าจอเพื่อทำการเลือกไฟล์ข้อมูลจาก 2 ฐานข้อมูล

จากรูปที่ 4.11 แสดงการออกแบบส่วนติดต่อกับผู้ใช้เพื่อทำการเลือกไฟล์ฐานข้อมูลจากสองแหล่งข้อมูล ซึ่งผู้ใช้สามารถเลือกไฟล์ฐานข้อมูลจาก Microsoft Access 2000 ซึ่งส่วนประกอบของหน้าจอนี้ สามารถอธิบายได้ดังต่อไปนี้

- TextBoxของแหล่งข้อมูลที่ 1 และ 2 เป็นส่วนที่ใช้ในการแสดงชื่อของไฟล์ที่ผู้ใช้ได้ทำการเลือก
- ปุ่ม  เป็นปุ่มที่ใช้ในการเลือกไฟล์จากฐานข้อมูล Microsoft Access 2000

- ปุ่ม Cancel เป็นปุ่มที่ใช้ในการยกเลิกไฟล์ที่ได้ทำการเลือก ซึ่งบ่รากฏอยู่บ.. TextBox

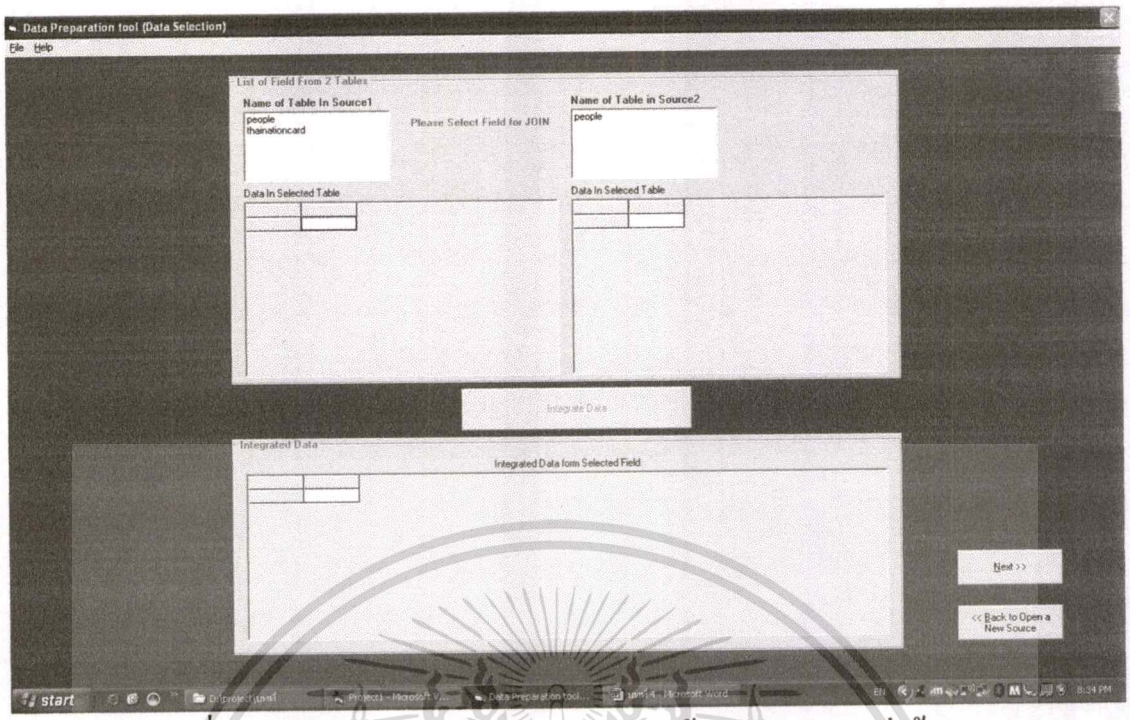
- ปุ่ม Next เป็นปุ่มที่ใช้ในการเลือกเพื่อจะทำงานในขั้นตอนต่อไป ซึ่งเป็นขั้นตอนในการรวมข้อมูลให้เป็นข้อมูลเดียวกัน (Data Integration) หลังจากได้ทำการเลือกไฟล์จากแหล่งข้อมูลทั้งสองแหล่งข้อมูลเรียบร้อยแล้ว



รูปที่ 4.11 การออกแบบหน้าจอการเลือกไฟล์ข้อมูลจากสองฐานข้อมูล

4.4.3 การออกแบบหน้าจอเพื่อทำการรวมข้อมูล (Data Integration)

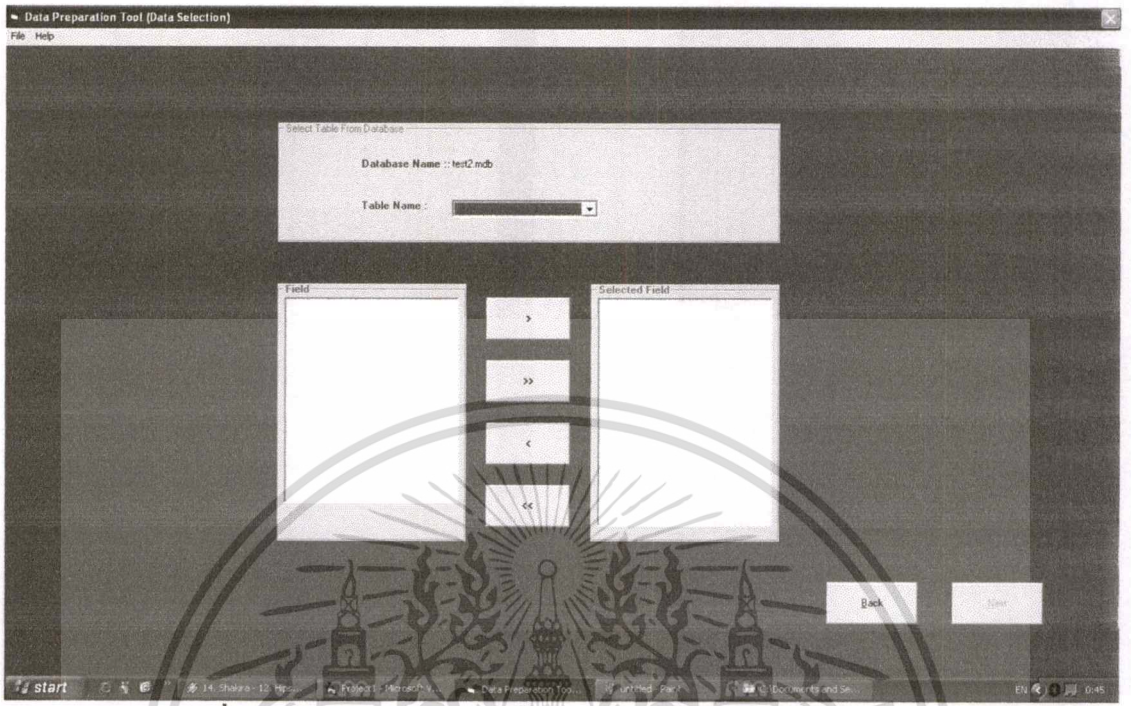
จากรูปที่ 4.12 แสดงการออกแบบหน้าจอการรวมข้อมูลจากสองแหล่งข้อมูล ให้เป็นข้อมูลกลุ่มก้อนเดียวกันซึ่งจากหน้าจอที่แสดงดังรูป สามารถอธิบายส่วนประกอบต่างๆ ของหน้าจอการรวมข้อมูลได้ ดังต่อไปนี้



รูปที่ 4.12 การออกแบบหน้าจอแสดงการรวมข้อมูลจากสองแหล่งข้อมูล

- ListBoxของแหล่งข้อมูลที่ 1 และ 2 เป็นส่วนที่ใช้ในการแสดงรายการของตารางที่มีในแหล่งข้อมูลที่ผู้ใช้ทำการเลือกจากแหล่งข้อมูลทั้งสองแหล่ง
- MSHFlexGrid ของแหล่งข้อมูลที่ 1 และ 2 เป็นส่วนที่ใช้ในการแสดงข้อมูลทั้งหมดของตารางที่ผู้ใช้ทำการเลือก โดยจะแสดงข้อมูลในรูปแบบของตาราง
- ปุ่ม Integrate Data เป็นปุ่มที่ใช้ในการรวมข้อมูลโดยจะใช้วิธีการ Join หรือ ทำการตรวจสอบข้อมูลว่ามีโครงสร้างเดียวกันหรือไม่เพื่อทำการรวมข้อมูลซึ่งเมื่อทำการรวมข้อมูลเรียบร้อยแล้ว ข้อมูลจะแสดงไว้ที่ MSHFlexGrid ทางด้านล่างของปุ่ม Integrate Data
- ปุ่ม Next เป็นปุ่มที่ใช้ในการเลือกเพื่อที่จะทำงานในขั้นตอนต่อไป นั้นคือขั้นตอนของการเลือกข้อมูล (Data Selection)
- ปุ่ม << Back to Open a New Source เป็นปุ่มที่ใช้ในการกลับไปทำการเลือกแหล่งข้อมูลใหม่ในกรณีที่ไฟล์ที่ได้ทำการเลือกมาใช้งานนั้นไม่สามารถที่จะทำการรวมกันได้ เนื่องจากเป็นไฟล์ข้อมูลที่ไม่มีความสัมพันธ์กัน หรืออาจเป็นข้อมูลคนละโครงสร้างกัน

4.4.4 การออกแบบหน้าจอเพื่อทำการเลือกตารางจากไฟล์ข้อมูล



รูปที่ 4.13 การออกแบบหน้าจอเพื่อทำการเลือกตารางจากไฟล์ข้อมูล

จากรูปที่ 4.13 แสดงการออกแบบหน้าจอของการเลือกตารางจากไฟล์ข้อมูลซึ่งส่วนประกอบของหน้าจอนี้ สามารถอธิบายได้ดังต่อไปนี้

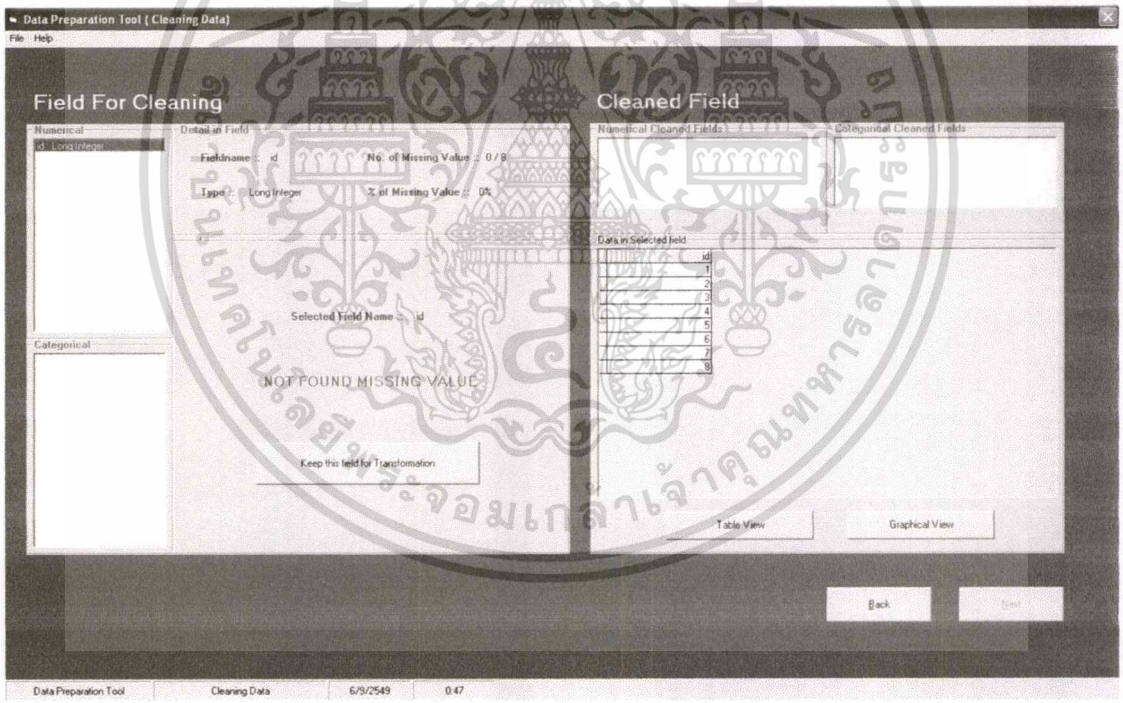
- ComboBox เป็นส่วนที่ใช้เก็บชื่อตารางต่างๆ จากไฟล์ข้อมูลที่ได้ทำการเลือก
- ListField เป็นส่วนที่ใช้แสดงชื่อฟิลด์ต่างๆ ที่อยู่ภายในตารางที่ได้ทำการเลือกจาก ComboBox
- ปุ่ม > เป็นปุ่มคำสั่งเพื่อใช้ทำการเลือกฟิลด์จาก ListField มาไว้ที่ SelField เพื่อนำฟิลด์ที่ได้ทำการเลือกนั้นไปประมวลผลในขั้นตอนต่อไป ซึ่งสามารถเลือกฟิลด์ได้ทีละฟิลด์
- ปุ่ม >> เป็นปุ่มคำสั่งเพื่อใช้ทำการเลือกฟิลด์ทั้งหมดจาก ListField มาไว้ที่ SelField เพื่อนำฟิลด์ที่เลือกนี้ไปทำการประมวลผลในขั้นตอนต่อไป
- ปุ่ม < เป็นปุ่มคำสั่งเพื่อทำการย้ายฟิลด์ที่ไม่ต้องการนำไปประมวลผลมาเก็บไว้ที่เดิม นั่นคือ ที่ ListField ซึ่งเป็นการย้ายฟิลด์ทีละฟิลด์
- ปุ่ม << เป็นปุ่มคำสั่งเพื่อทำการย้ายฟิลด์ทั้งหมดที่ไม่ต้องการนำไปประวลผลมาเก็บไว้ที่เดิม นั่นคือ ListField

- ปุ่ม Back เป็นปุ่มคำสั่งเพื่อย้อนกลับไปทำการเลือกไฟล์ข้อมูลใหม่
- ปุ่ม Next เป็นปุ่มคำสั่งเพื่อสั่งให้ระบบทำงานในขั้นตอนต่อไป นั่นคือการทำการปรับปรุงข้อมูลให้มีคุณภาพดีขึ้น

4.4.5 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูล (Data Cleaning)

4.4.5.1 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูล กรณีไม่พบค่าที่ขาดหายไป

จากรูปที่ 4.14 เป็นการออกแบบหน้าจอการทำงานเมื่อระบบทำการตรวจสอบค่าที่ขาดหายไปภายในฟิลด์ที่ถูกเลือก โดยฟิลด์ที่ผู้ใช้เลือกจะถูกแบ่งประเภทตามชนิดของฟิลด์ นั่นคือ จะถูกแบ่งเป็นข้อมูลประเภท Numerical และ Categorical เมื่อฟิลด์ที่ผู้ใช้ทำการเลือกนั้นไม่มีค่าที่ขาดหายไป หน้าจอการทำงานจะปรากฏดังรูปที่ 4.14 คือ แสดงสถานะของฟิลด์ว่าไม่พบค่าที่ขาดหายไป

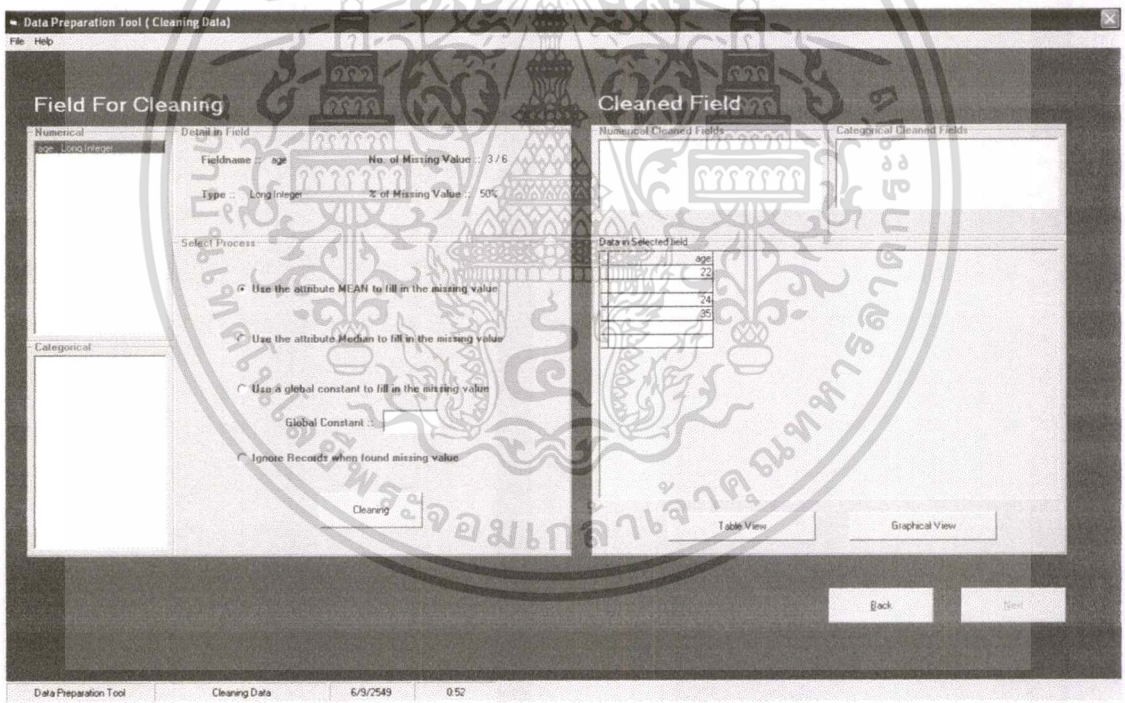


รูปที่ 4.14 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูล กรณีไม่พบค่าที่ขาดหายไป

4.4.5.2 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูล กรณีค่าฟิลด์ที่เลือกเป็นข้อมูลประเภทNumerical

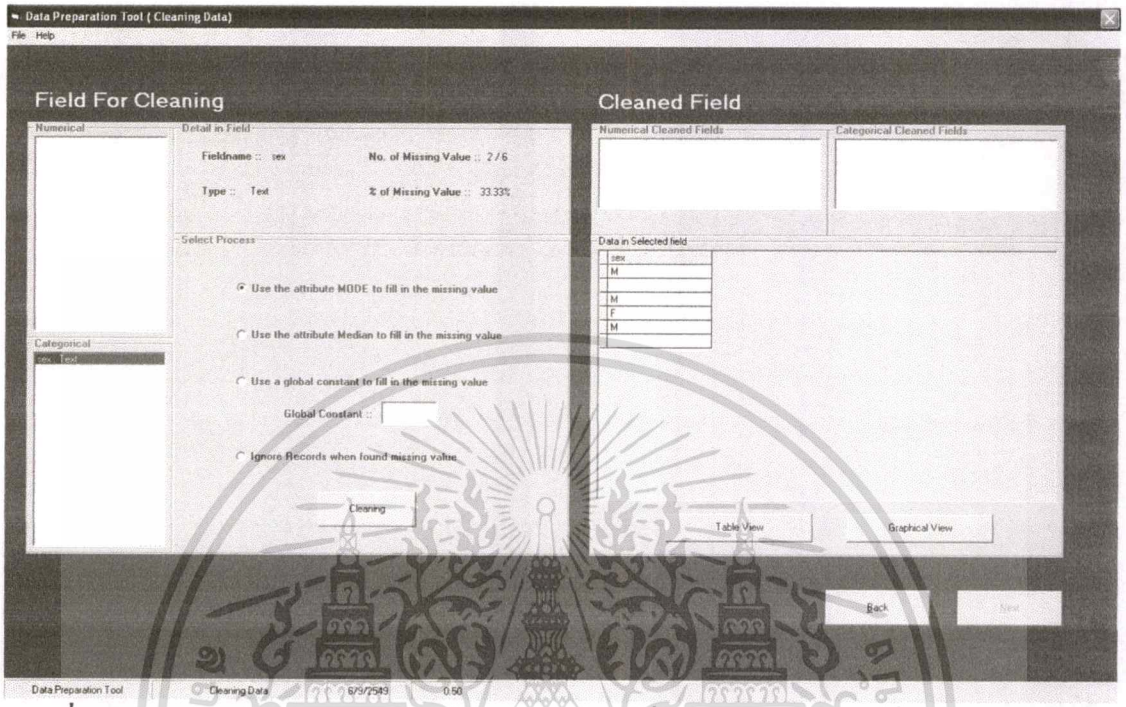
จากรูปที่ 4.15 แสดงการออกแบบหน้าจอการทำงานของระบบเมื่อพบค่าที่ขาดหายไป โดยฟิลด์ที่เลือกเป็นฟิลด์ข้อมูลประเภท Numerical ซึ่งส่วนประกอบของหน้าจอสามารถอธิบายได้ดังต่อไปนี้

- Ignore Records when found missing value จะทำการลบเรคอร์ดที่พบค่าที่ขาดหายไป
- Use a global constant to fill in the missing value ให้ผู้ใช้ทำการใส่ค่าที่ต้องการแทนที่ค่าที่ขาดหายไปลงไป
- Use the attribute MEAN to fill in missing value ใช้ค่าเฉลี่ยเพื่อแทนที่ค่าที่ขาดหายไป
- ปุ่ม Cleaning เป็นปุ่มที่ใช้ทำการทำความสะอาดข้อมูล โดยจะทำการทำความสะอาดข้อมูลตามวิธีที่ผู้ใช้ได้ทำการเลือก
- ปุ่ม Table View เป็นปุ่มที่ใช้ในการแสดงข้อมูลที่ผ่านการทำความสะอาดข้อมูลแล้วในรูปแบบของตาราง
- ปุ่ม Graphical View เป็นปุ่มที่ใช้ในการแสดงข้อมูลที่ผ่านการทำความสะอาดข้อมูลแล้วในรูปแบบของกราฟ



รูปที่ 4.15 การออกแบบหน้าจอปรับปรุงคุณภาพข้อมูลเมื่อค่าฟิลด์เป็นข้อมูลประเภทNumerical

4.4.5.3 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูล กรณีค่าฟิลด์ที่เลือกเป็นข้อมูลประเภท Categorical



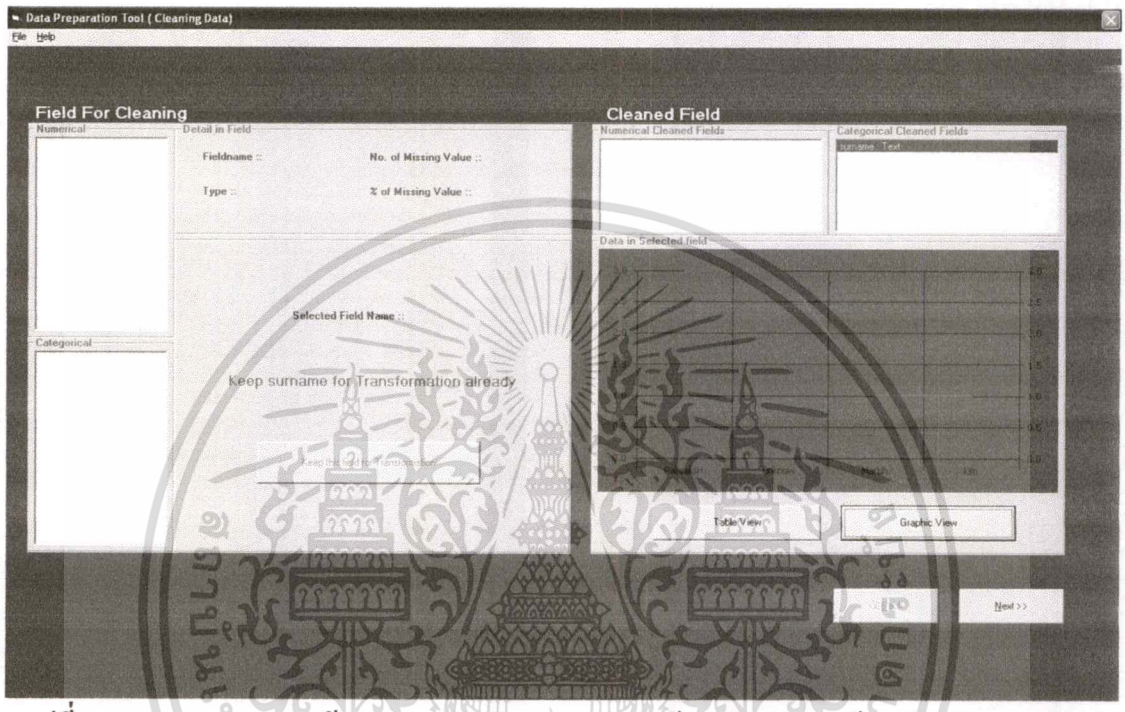
รูปที่ 4.16 การออกแบบหน้าจอปรับปรุงคุณภาพข้อมูลเมื่อค่าฟิลด์เป็นข้อมูลประเภทCategorical

จากรูปที่ 4.16 แสดงการออกแบบหน้าจอการทำงานของระบบเมื่อพบค่าที่ขาดหายไป โดยฟิลด์ที่เลือกเป็นฟิลด์ข้อมูลประเภท Categorical ซึ่งส่วนประกอบของหน้าจอสามารถอธิบายได้ดังต่อไปนี้

- Ignore Records when found missing value จะทำการลบเรคอร์ดที่พบค่าที่ขาดหายไป
- Use a global constant to fill in the missing value ให้ผู้ใช้ทำการใส่ค่าที่ต้องการแทนที่ค่าที่ขาดหายไปลงไป
- Use the attribute MODE to fill in missing value ใช้ค่าฐานนิยมเพื่อแทนที่ค่าที่ขาดหายไป
- ปุ่ม Cleaning เป็นปุ่มที่ใช้ทำการทำความสะอาดข้อมูล โดยจะทำการทำความสะอาดข้อมูลตามวิธีที่ผู้ใช้ได้ทำการเลือก
- ปุ่ม Table View เป็นปุ่มที่ใช้ในการแสดงข้อมูลที่ผ่านการทำความสะอาดข้อมูลแล้วในรูปแบบของตาราง

- ปุ่ม Graphical View เป็นปุ่มที่ใช้ในการแสดงข้อมูลที่ผ่านการทำความสะอาดข้อมูลแล้วในรูปแบบของกราฟ

ส่วนรูปที่ 4.17 เป็นการออกแบบส่วนติดต่อกับผู้ใช้เพื่อใช้ในการแสดงข้อมูลของฟิลด์ต่างๆ ในลักษณะของกราฟ



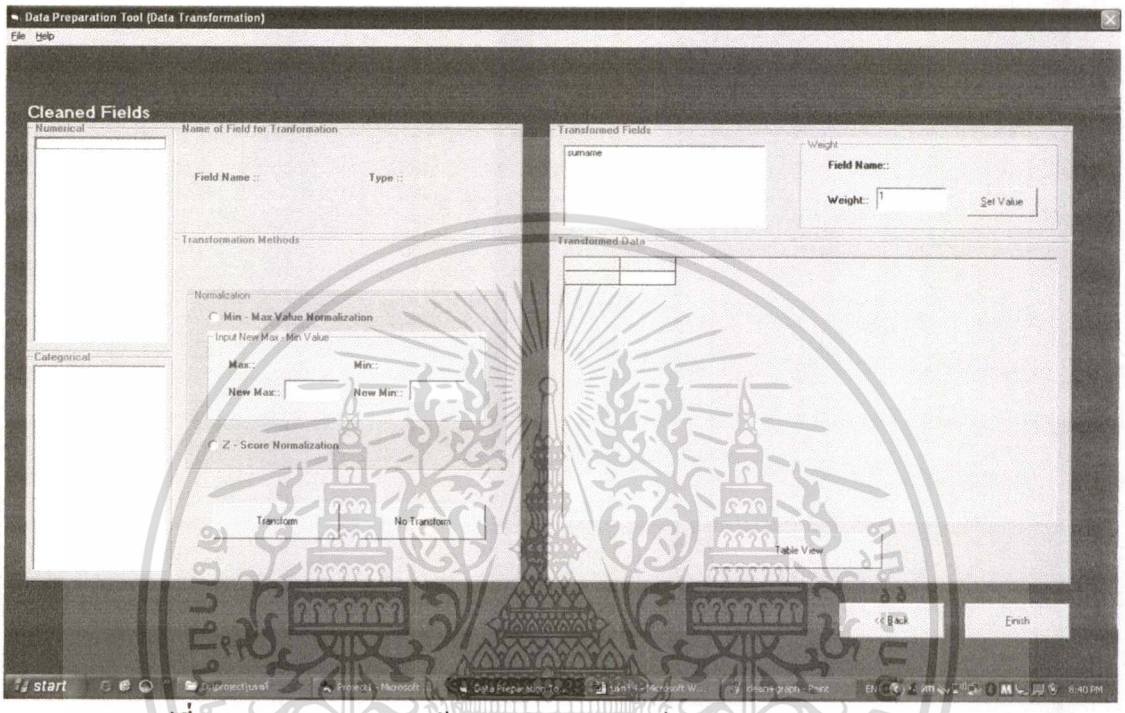
รูปที่ 4.17 การออกแบบหน้าจอการปรับปรุงคุณภาพของข้อมูลโดยแสดงข้อมูลในลักษณะกราฟ

4.4.6 การออกแบบหน้าจอของการแปลงข้อมูล (Data Transformation)

รูปที่ 4.18 แสดงการออกแบบหน้าจอของการทำการแปลงข้อมูลเพื่อให้เหมาะสมกับอัลกอริทึมดาต้าไมนิ่งที่ได้ทำการเลือกใช้ซึ่งส่วนประกอบต่างๆของหน้าจอในการแปลงข้อมูลสามารถอธิบายได้ดังต่อไปนี้

- ListBox แสดงฟิลด์ของ Numerical และ Categorical ที่ทำความสะอาดแล้ว
- Transformation Methods เป็นส่วนที่ให้ผู้ใช้ทำการเลือกวิธีในการแปลงข้อมูล โดยจะมีวิธีให้เลือกอยู่ 2 วิธีคือ Min – Max Normalization ซึ่งวิธีการนี้ผู้ใช้จะต้องทำการกำหนดค่ามากที่สุดและค่าน้อยที่สุดใหม่ ส่วนวิธีการที่สองคือ การแปลงโดยใช้วิธีการ Z – Score Normalization
- ปุ่ม Transform เป็นปุ่มที่ใช้ในการแปลงค่าข้อมูลโดยจะทำการแปลงข้อมูลตามวิธีการที่ผู้ใช้เลือก

- ปุ่ม No Transform ใช้เมื่อผู้ใช้ไม่ต้องการแปลงข้อมูลในฟิลด์นั้นๆ
- ปุ่ม Set Value เป็นปุ่มที่ใช้ในการตั้งค่าน้ำหนักให้กับฟิลด์ที่เลือกตามที่ผู้ใช้กำหนด



รูปที่ 4.18 การออกแบบหน้าจอของการแปลงข้อมูล (Data Transformation)

4.5 Source Codeของระบบบางส่วน

ในหัวข้อนี้จะเป็นการนำเสนอส่วนของโปรแกรมบางส่วน ซึ่งเป็นกระบวนการหลักในการทำงานของเครื่องมือสำหรับการทำ Data Preparation ซึ่งรายละเอียดของแต่ละกระบวนการจะได้อธิบายในลำดับต่อไป

4.5.1 Source Code ของระบบในส่วนของการทำ Data Selection

4.5.1.1 Source Code ในการเปิดไฟล์ฐานข้อมูลจาก Ms Access 20000

```
Dialog.Filter = " Microsoft Access (*.mdb | *.mdb)"
Dialog.Showopen
```

จากอัลกอริทึมข้างต้น เป็นการกรองไฟล์ฐานข้อมูลที่เป็นฐานข้อมูลชนิด Microsoft Access โดยใช้ Property ที่ชื่อว่า Filter

4.5.1.2 Source Code ในการดึงชื่อของตารางจากไฟล์ฐานข้อมูลที่เลือกมาเก็บไว้ใน Recordset

```
With Recordset
If .Recordset < > 0 the
    Do Until .EOF
        If (Left(.Fields("Table_name").value,4) < > "MSys" ) then
            ComboTable.Additem  .Fields("Table_name")
            .Movenext
        else
            .Movenext
        end if
    Loop
End if
End with
```

ในการดึงชื่อตารางจากไฟล์ฐานข้อมูลที่เราเลือกมาทำการแสดงเพื่อให้ผู้ใช้ได้เลือกทำงานกับตารางต่างๆ นั้น จะต้องทำการดึงชื่อตารางมาเก็บไว้ที่ Recordset ก่อน และก่อนที่จะทำการแสดงชื่อตารางให้กับผู้ใช้นั้น จะต้องทำการกรองเอาเพียงแค่ตารางที่อยู่ในไฟล์ฐานข้อมูลจริงๆเท่านั้น ซึ่งจะต้องทำการตรวจสอบว่าชื่อตารางที่เก็บไว้ใน Recordset นั้น จะต้องไม่ได้ขึ้นต้นด้วย "MSys" ถ้าตรงกับเงื่อนไขก็จะทำการเพิ่มชื่อตารางที่ตรงกับเงื่อนไขไว้ที่ Bound Control

4.5.1.3 Source Code ในการดึงฟิลด์จากตารางมาเก็บไว้ใน Recordset

```
Recordset.open  Tablename,connection,adOpenKeyset,adLockOptimistic,adCmdTable
```

ในการดึงชื่อฟิลด์ข้อมูลเพื่อนำมาแสดงใน Bound Control จะต้องมีการดึงข้อมูลมาเก็บไว้ที่ Recordset ก่อน โดยการดึงชื่อฟิลด์จากตารางที่ต้องการนั้น จะใช้ Property "Open"

4.5.2 Source Code ของระบบในส่วนของการทำ Data Cleaning

4.5.2.1 Source Code ในการ Cleaning โดยใช้ค่า Mean

```
query = "SELECT avg("ชื่อฟิลด์") as Average FROM ชื่อตาราง"
with Recordset
    .Activeconnection = Connection
    .CursorType = adOpenForwardOnly
    .CursorLocation = adUseClient
```

```
.Open query
average = Recordset.Fields("Average")
end with
query = "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' "& average &" ' WHERE ชื่อฟิลด์ IS NULL"
connection.Execute query
```

อัลกอริทึมในการทำความสะอาดข้อมูลโดยใช้วิธีการแทนที่เรคอร์ดที่ขาดหายไปโดยใช้ค่าเฉลี่ยของฟิลด์นั้นๆ จะต้องทำการหาค่าเฉลี่ยของฟิลด์นั้น โดยใช้คำสั่ง "SELECT avg ("ชื่อฟิลด์") as Average FROM ชื่อตาราง" เพื่อให้ได้ค่าเฉลี่ยของฟิลด์ที่ต้องการทำความสะอาด เมื่อได้ค่าเฉลี่ยมาเก็บไว้ในตัวแปรแล้ว ซึ่งในที่นี้คือตัวแปรชื่อว่า "Average" ก็จะทำการแทนที่ค่าเฉลี่ยในเรคอร์ดที่ขาดหายไปโดยใช้คำสั่ง "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' "& average &" ' WHERE ชื่อฟิลด์ IS NULL" ซึ่งคำสั่งนี้จะทำการแทนที่ค่าตัวแปร "Average" ลงในเรคอร์ดที่มีค่า NULL

4.5.2.2 Source Code ในการ cleaning โดยใช้ค่า Mode

```
For i = 1 to Recordset.Recordcount
        Data_array(i) = Reordset.Fields(0).value
        .movenext
Next i

For j = 1 to Recordset.Recordcount
If cate_data(1) = "" then
        Cate_data(1) = data_array(j)
        Cate_count(1) = 1
        Arraycount = 1
Else
        If cate_data(1) < > "" then
                For k = 1 to (arraycount + 1)
                        If data_array(j) = cate_data(k) then
                                Cate_count(k) = cate_count(k)+1
                                Keep = arraycount
                                K = arraycount + 1
                        Else
                        If k = (arraycount +1 ) then
```

```
                                    If cate_data(k) = “” then
                                        Cate_data(k) = data_array(j)
                                        Cate_count(k) = cate_count(k)+1
                                        Arraycount = arraycount+1
                                        Keep = arraycount
                                    End if
                                    End if
                                    End if
                                Next k
                                Arraycount = keep
                                End if
                                End if
                            Next j
For i = 1 to Recordset.Recrdcount
If cate_count(1) < cate_count(i) then
            Cate_count = cate_count(i)
            Cate_data(1) = cate_data(i)
            modeValue = cate_data(1)
else
            modeValue = cate_data(1)
end if
next i
```

query = “UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ‘ “ & modeValue & ” ’ WHERE ชื่อฟิลด์ IS NULL”
connection.execute query

อัลกอริทึมในการทำความสะอาดข้อมูลโดยใช้วิธีการแทนที่ค่าที่หายไปด้วยค่าฐานนิยม (Mode) ของฟิลด์นั้นๆ จะต้องทำการคำนวณหาค่าความถี่ของข้อมูลที่เก็บอยู่ในฟิลด์นั้นก่อนว่าข้อมูลแต่ละข้อมูล มีความถี่ของข้อมูลเท่าไรเพื่อหาค่าของข้อมูลที่มีความถี่มากที่สุดในการแทนที่ค่าข้อมูลที่ขาดหายไป

จากอัลกอริทึมเบื้องต้น จะนำค่าข้อมูลที่เก็บอยู่ใน Recordset มาเก็บไว้เป็นอาเรย์ โดยทำการวนลูปเท่ากับจำนวนเรคอร์ดที่เก็บอยู่ใน Recordset โดยเรียกใช้ Property “Recordcount” และทำ

การอ่านค่าข้อมูลที่ pointer ปัจจุบันชี้อยู่มาเก็บไว้ที่อาเรย์โดยใช้คอลเล็คชัน “Fields” ของออบเจ็ค Recordset และในการเลื่อน pointer ไปยังตำแหน่งถัดไปเพื่อทำการอ่านค่าในเรคอร์ดนั้น จะใช้เมธอด MoveNext ของออบเจ็ค Recordset

เมื่อทำการเก็บค่าข้อมูลทุกเรคอร์ดไว้ในอาเรย์เรียบร้อยแล้ว หลังจากนั้นจะทำการวนลูปเพื่อนับค่าความถี่ของข้อมูลแต่ละตัว โดยที่ค่าความถี่ของข้อมูลใดๆนั้นจะเก็บไว้ในอาเรย์ชื่อว่า “Cate_count” เมื่อได้ค่าความถี่ของข้อมูลแต่ละตัวแล้ว ก็จะหาค่าข้อมูลที่มีความถี่มากที่สุดเพื่อหาค่าฐานนิยมของฟิลด์นั้นๆ ซึ่งค่าที่มีความถี่มากที่สุดจะถูกเก็บไว้ในตัวแปร “modeValue” และจะทำการแทนที่ค่าฐานนิยมลงไปในเรคอร์ดที่มีค่าข้อมูลที่ขาดหายไป โดยใช้คำสั่ง “UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ‘ “ & modeValue & ” ’ WHERE ชื่อฟิลด์ IS NULL”

4.5.2.3 Source Code ในการ Cleaning โดยใช้ค่า Median

query = “SELECT ชื่อฟิลด์ FROM ชื่อตาราง WHERE ชื่อฟิลด์ IS NOT NULL order by ชื่อฟิลด์” ASC”

```
connection.Execute query

with Recordset
        .open query
end with
set MSHBackup.Datasource = Recordset
count = Recordset.Recordcount
middle = count / 2

if middle = Round(middle,0) then
        Position1 = MSHBackup.TextMatrix(middle,1)
        Position2 = MSHBackup.TextMatrix(middle + 1 , 1)
End if
Median = (Val(Position1)+ Val(Position2)) / 2
```

Query = “UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ‘ “ & Median & ” ’ WHERE ชื่อฟิลด์ IS NULL”

```
Connection.Execute query

If middle < > Round (middle,0) then
```

Mid = Round(middle,0)

Position1 = MSHBackup.TextMatrix(mid,1)

End if

Query = "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' " & Position1 & " ' WHERE ชื่อฟิลด์ IS NULL"

Connection.execute query

อัลกอริทึมในการทำความสะอาดข้อมูลโดยใช้วิธีการแทนที่ค่าที่ขาดหายไปโดยใช้ค่า Median จะต้องคำนวณหาค่าที่อยู่ที่ตำแหน่งเปอร์เซ็นไทม์ที่ 50 ของฟิลด์ที่ต้องการทำความสะอาด การหาค่า Median ของฟิลด์ ก็คือการหาเรคอร์ดตำแหน่งที่ (N+1)/2 ซึ่งจากอัลกอริทึมนี้ทำการหาตำแหน่งของเรคอร์ดที่อยู่ตรงกลางได้จากการนำเอาจำนวนเรคอร์ดทั้งหมดหารสองเพื่อทำการแบ่งครึ่งของข้อมูลทั้งหมด ผลหารจะถูกปัดขึ้นในทุกกรณี ซึ่งจากอัลกอริทึมนี้จะใช้คำสั่ง "Round" เพื่อทำการปัดค่าทศนิยมของผลลัพธ์จากการหาร ผลลัพธ์จากการหารที่ได้ จะเป็นค่าตำแหน่ง Pointer ของค่าที่จะนำมาแทนที่ค่าที่ขาดหายไปในฟิลด์ ส่วนถ้าจำนวนเรคอร์ดทั้งหมด หลังจากได้ค่า Median มาแล้วก็จะทำการแทนที่ค่าที่ขาดหายไปโดยใช้คำสั่ง "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' " & Position1 & " ' WHERE ชื่อฟิลด์ IS NULL"

4.5.2.4 Source Code ในการ Cleaning โดยการแทนค่าจากค่าที่ผู้ใช้กำหนด

query = "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' " ค่าที่ผู้ใช้กำหนด " ' WHERE ชื่อฟิลด์ IS NULL"

connection.Execute query

อัลกอริทึมในการทำความสะอาดข้อมูลโดยใช้การแทนค่าจากค่าที่ผู้ใช้กำหนด จะทำการรับ "ค่าที่ผู้ใช้กำหนด" จาก Bound Control ผ่านทางหน้าจอและจะนำค่านั้นแทนที่ค่าที่ขาดหายไปโดยใช้คำสั่ง "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ' " ค่าที่ผู้ใช้กำหนด " ' WHERE ชื่อฟิลด์ IS NULL"

4.5.2.5 Source Code ในการ Cleaning โดยการลบเรคอร์ดทิ้ง

query = "DELECT * FROM ชื่อตาราง WHERE ชื่อฟิลด์ IS NULL"

connection.Execute query

อัลกอริทึมในการทำความสะอาดข้อมูลโดยการลบเรคอร์ดทิ้ง จะทำการลบเรคอร์ที่มีค่าที่ขาดหายไปทิ้งทั้งเรคอร์ด โดยจะใช้คำสั่ง "DELECT * FROM ชื่อตาราง WHERE ชื่อฟิลด์ IS NULL"

4.5.3 Source Codeของระบบในส่วนของการทำ Data Transformation

4.5.3.1 Source Code ในการแปลงข้อมูลโดยใช้ Min – Max Value Normalizaion

```
query = "SELECT max("ชื่อฟิลด์") as maxValue FROM ชื่อตาราง"
with Recordset
        .open query
        maxValue = Recordset.Fields("maxValue")
end with

query = "SELECT min("ชื่อฟิลด์") as minValue FROM ชื่อตาราง"
with Recordset
        .open query
        minValue = Recordset.Fields("minValue")
end with

query = "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = (" & ชื่อฟิลด์ & " – " & minValue & " / (" & maxValue & " – " & minValue & ") * ( " & newMax & " – " & newMin & ") + " & newMin & "
connection.execute query
```

อัลกอริทึมในการแปลงข้อมูลโดยใช้วิธี Min – Max Value Normalization จะต้องทำการหาค่ามากที่สุด และค่าน้อยที่สุดในฟิลด์ที่ต้องการจะทำการแปลงข้อมูล เพื่อใช้เป็นค่าคงที่ในการแปลงข้อมูลในแต่ละเรคอร์ด ซึ่งค่ามากที่สุดสามารถหาได้จากคำสั่ง "SELECT max("ชื่อฟิลด์") as maxValue FROM ชื่อตาราง" และเก็บค่ามากที่สุดในตัวแปร จากอัลกอริทึมนี้ได้เก็บค่ามากที่สุดของฟิลด์ที่ต้องการแปลงข้อมูลไว้ในตัวแปรชื่อ "maxValue" ส่วนค่าน้อยที่สุดสามารถของฟิลด์หาได้จากคำสั่ง "SELECT min("ชื่อฟิลด์") as minValue FROM ชื่อตาราง" และเก็บค่าที่ได้ไว้ในตัวแปรชื่อ "minValue"

จากนั้นอัลกอริทึมข้างต้นจะทำการรับค่ามากที่สุด และค่าน้อยที่สุดค่าใหม่ตามที่ผู้ใช้ต้องการจากทาง Bound Control และทำการแปลงข้อมูลในฟิลด์ที่ได้ทำการเลือกโดยใช้คำสั่ง "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = (" & ชื่อฟิลด์ & " – " & minValue & " / (" & maxValue & " – " & minValue & ") * (" & newMax & " – " & newMin & ") + " & newMin & "

4.5.3.2 Source Code ในการแปลงข้อมูลโดยใช้ Z – Score Normalization

```
query = "SELECT avg (" & ชื่อฟิลด์ & ") as Average FROM ชื่อตาราง"
with Recordset
```

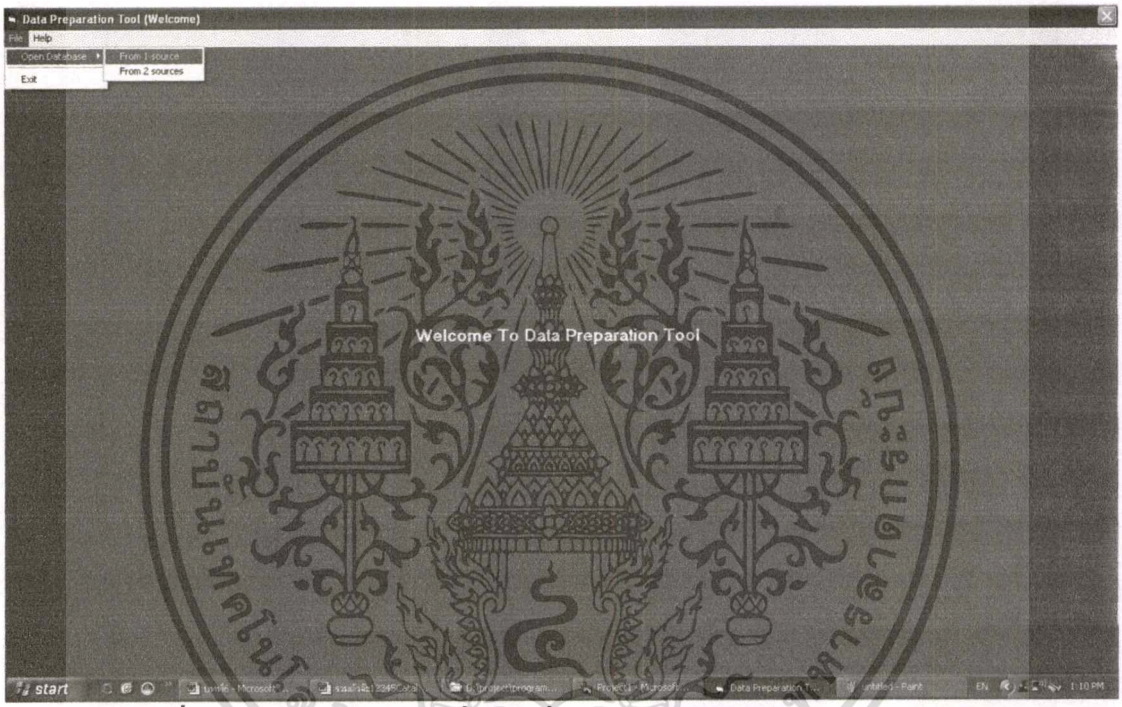
```
        .Open query
average = Recordset.Fields("Average")
end with

query = "SELECT ชื่อฟิลด์ FROM ชื่อตาราง"
with Recordset
        .Open query
end with
N = Recordset.Recordcount


For i = 1 to N
        X = Recordset.Fields(0).value
        Minus = x – average
        Minus2 = minus * minus
        Sum = sum + minus2
Next i

SD = sqr (sum / N)

Query = "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = ( " & ชื่อฟิลด์ & " – " & average & ") /
( "& SD &")
        Connection.execute query
```

อัลกอริทึมในการแปลงข้อมูลโดยใช้วิธี Z – Score Normalization จะต้องทำการคำนวณหาค่าเฉลี่ยของฟิลด์ที่ต้องการแปลงข้อมูล ซึ่งจากอัลกอริทึมข้างต้นใช้คำสั่ง "SELECT avg (" & ชื่อฟิลด์ & ") as Average FROM ชื่อตาราง" และเก็บค่าที่ได้จากการคำนวณไว้ในตัวแปร ซึ่งในอัลกอริทึมนี้ค่าเฉลี่ยดังกล่าวถูกเก็บไว้ในตัวแปรชื่อ "Average" หลังจากนั้นทำการนับจำนวนเรคอร์ดทั้งหมดของฟิลด์ที่ต้องการแปลงข้อมูล และทำการคำนวณหาค่าเบี่ยงเบนมาตรฐานของฟิลด์ข้อมูลนั้น ในอัลกอริทึมนี้จะเก็บค่าเบี่ยงเบนมาตรฐานที่คำนวณได้ไว้ที่ตัวแปรชื่อ "SD" เมื่อได้ค่าที่ต้องการครอบถ้วนแล้ว จะเรียกใช้คำสั่ง "UPDATE ชื่อตาราง SET ชื่อฟิลด์ = (" & ชื่อฟิลด์ & " – " & average & ") / ("& SD &") เพื่อทำการแปลงค่าข้อมูลในฟิลด์ที่เลือก

บทที่ 5
การทดลองใช้งานโปรแกรม

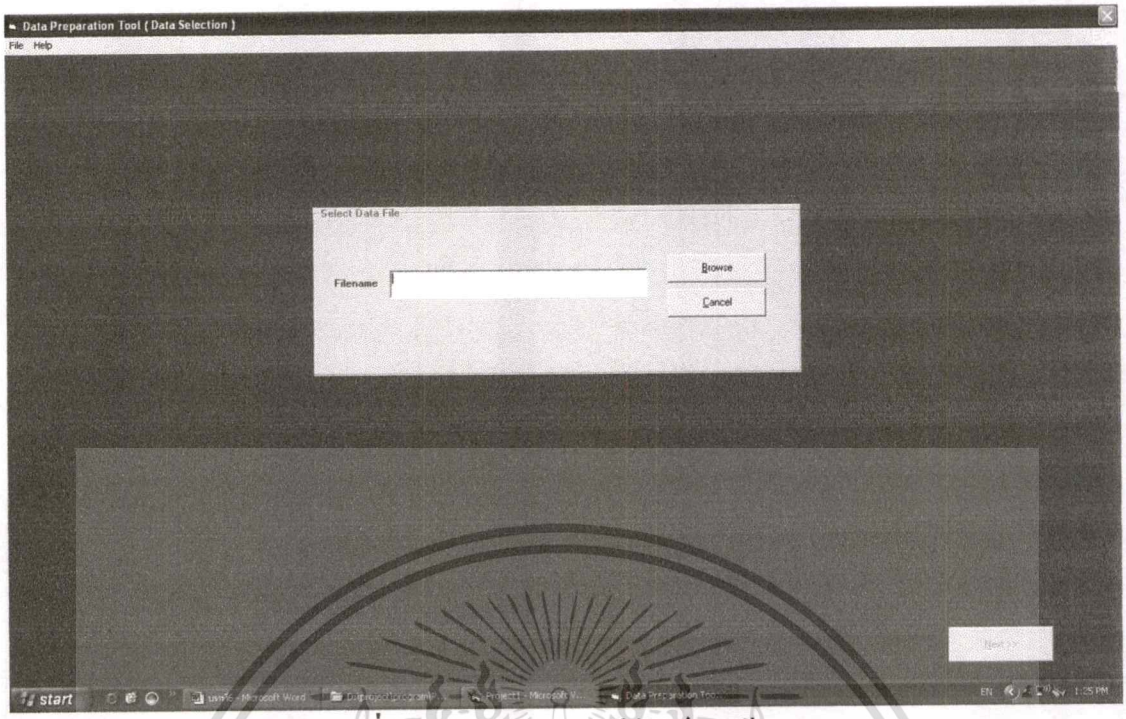
5.1 คู่มือและตัวอย่างการใช้งาน

เมื่อผู้ใช้ทำการเปิดเครื่องมือในการทำ Data Preparation ขึ้นมาใช้งาน จะปรากฏหน้าจอหลักของเครื่องมือดังแสดงในรูปที่ 5.1



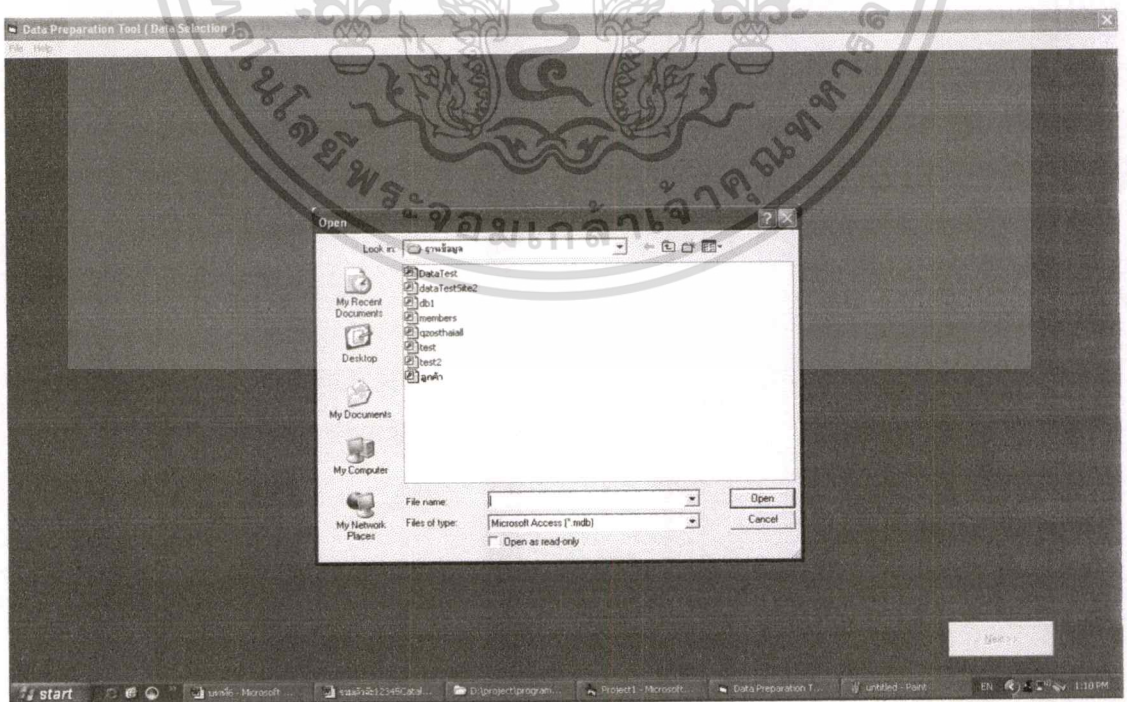
รูปที่ 5.1 หน้าจอหลักของเครื่องมือที่ทำการเลือกข้อมูลจากแหล่งข้อมูลเดียว

จากรูปที่ 5.1 ผู้ใช้จะสังเกตเห็นว่าเครื่องมือที่ปรากฏจะสามารถนำเข้าข้อมูลได้ 2 แบบ คือ นำเข้าข้อมูลจากแหล่งข้อมูลเดียว และนำเข้าข้อมูลจากสองแหล่งข้อมูล เมื่อผู้ใช้ทำการเลือกที่ From 1 source จากเมนูหลัก จะปรากฏหน้าจอของการนำเข้าข้อมูลดังรูปที่ 5.2 ซึ่งเป็นการนำเข้าข้อมูลจากฐานข้อมูลเดียว



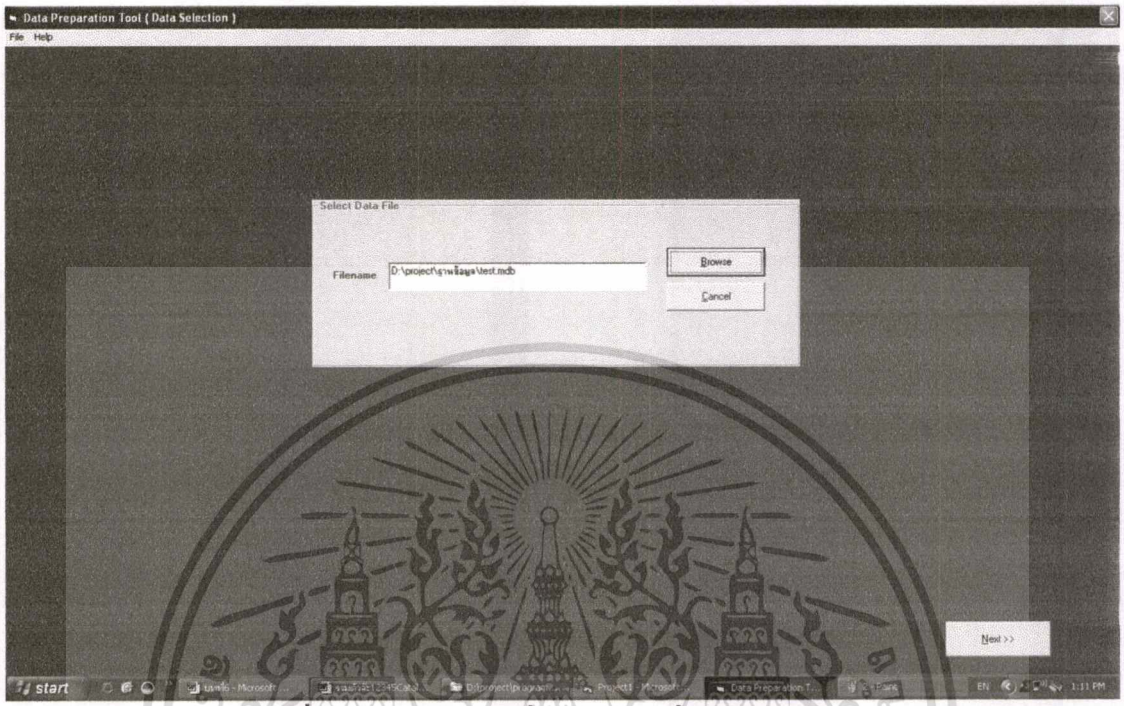
รูปที่ 5.2 หน้าจอการนำเข้าไฟล์ฐานข้อมูล

จากรูปที่ 5.2 ทำการเลือกไฟล์ฐานข้อมูลจากปุ่ม Browse ซึ่งเมื่อผู้ใช้ทำการกดปุ่ม Browse จะปรากฏหน้าต่างวินโดว์ขึ้นมาเพื่อให้ผู้ใช้ทำการเลือกไฟล์ฐานข้อมูลที่เป็น Microsoft Access ดังรูปที่ 5.3



รูปที่ 5.3 หน้าจอวินโดว์เพื่อทำการเลือกไฟล์

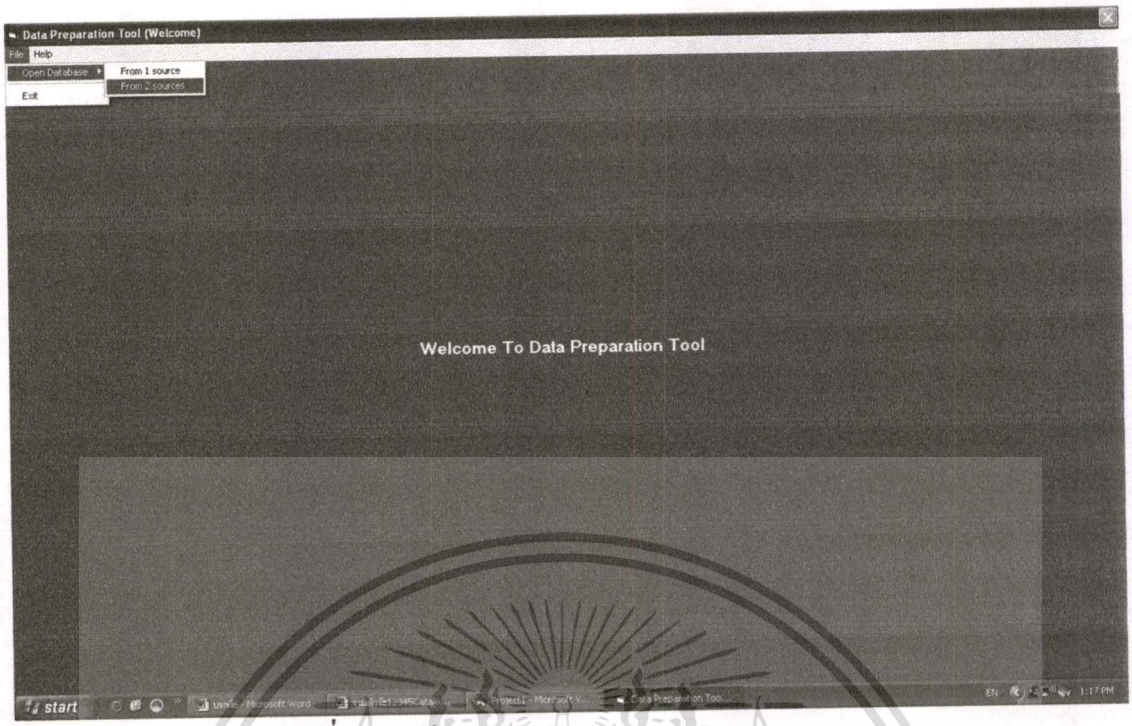
จากรูปที่ 5.3 เมื่อผู้ใช้ทำการเลือกไฟล์ฐานข้อมูลเรียบร้อยแล้ว ชื่อไฟล์ฐานข้อมูลและชื่อ path จะปรากฏที่ TextBox ดังแสดงให้เห็นดังรูปที่ 5.4



รูปที่ 5.4 หน้าจอแสดงชื่อ Path และชื่อไฟล์ฐานข้อมูล

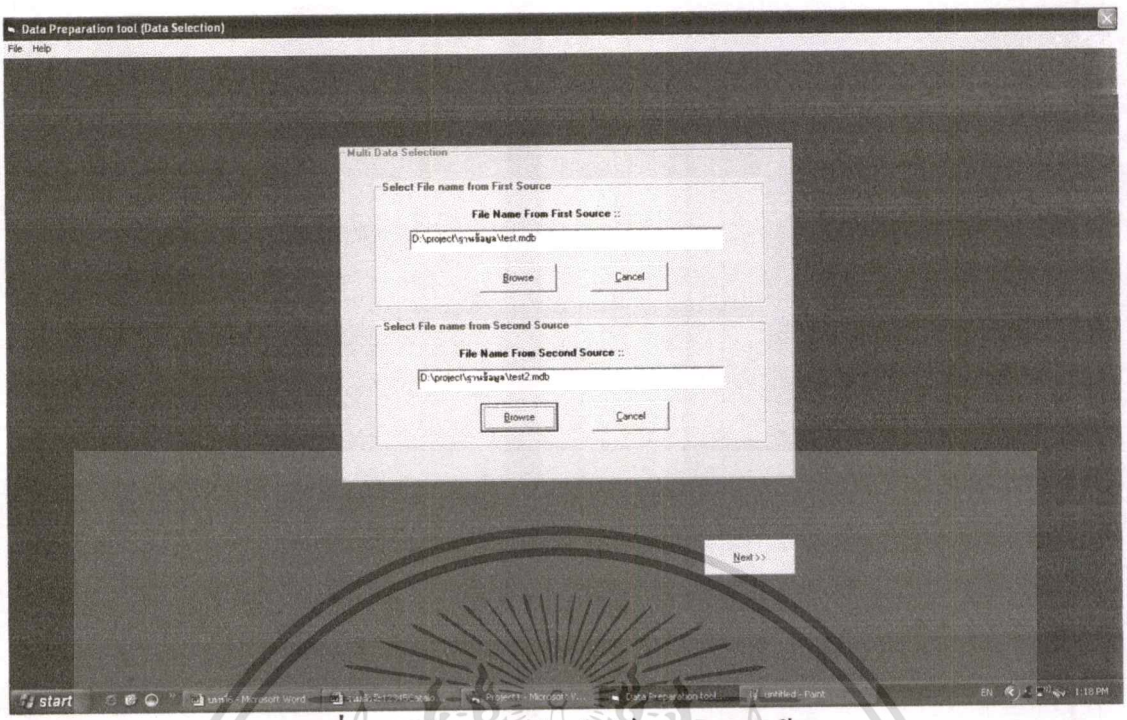
ถ้าผู้ใช้ต้องการเปลี่ยนไฟล์ฐานข้อมูลที่ได้ทำการเลือกไว้ใหม่ หรือไม่ต้องการไฟล์ฐานข้อมูลที่เลือกไว้ก่อนหน้านี้ ให้ผู้ใช้ทำการเลือกที่ปุ่ม Cancel เพื่อทำการลบชื่อไฟล์ที่เลือกไว้ และทำการเลือกไฟล์ฐานข้อมูลใหม่โดยการกดที่ปุ่ม Browse อีกครั้งหนึ่ง ถ้าผู้ใช้ไม่ต้องการที่จะเปลี่ยนฐานข้อมูล และต้องการที่จะดำเนินการกับข้อมูลในขั้นตอนต่อไป ให้ผู้ใช้ทำการคลิกที่ปุ่ม next เพื่อเข้าสู่กระบวนการในการเลือกตารางและฟิลด์จากฐานข้อมูลต่อไป

ในกรณีที่ต้องการเลือกไฟล์ฐานข้อมูลจากสองแหล่งข้อมูล ผู้ใช้สามารถทำการเลือกวิธีการนำเข้าไฟล์ฐานข้อมูลได้พร้อมกันสองแหล่งฐานข้อมูล ด้วยการเลือกคลิกที่ From 2 Source ที่เมนูหน้าจอหลัก ดังรูปที่ 5.5



รูปที่ 5.5 หน้าจอเมนูหลักแสดงเมนู From 2 source

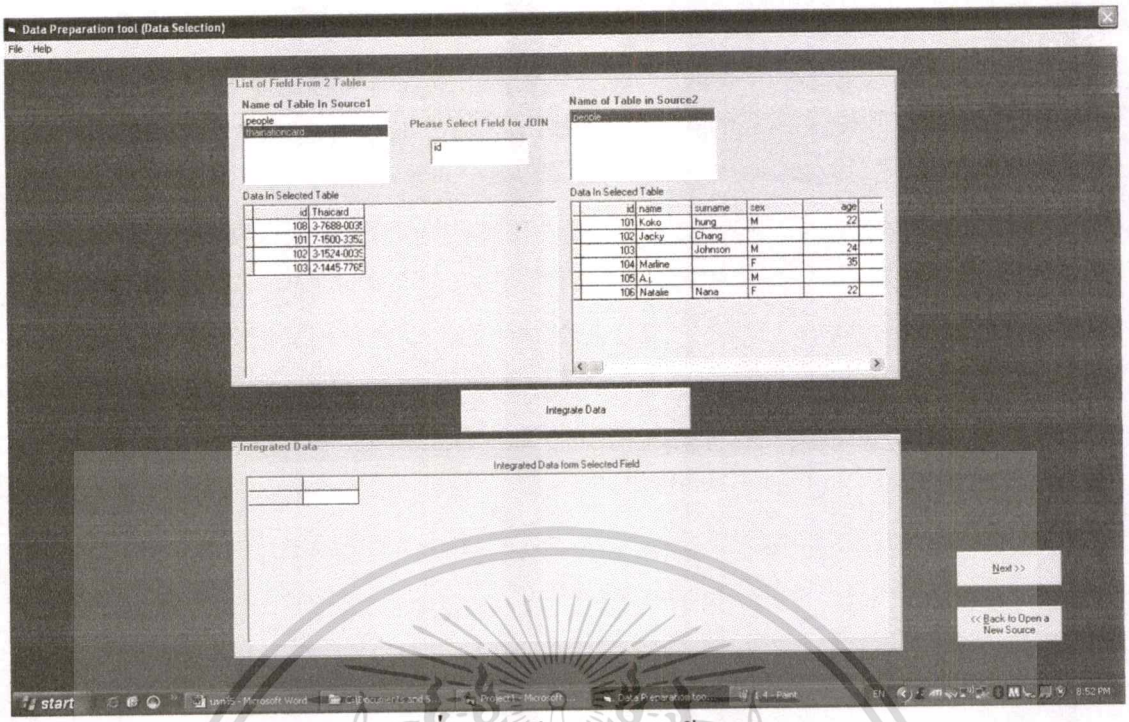
เมื่อผู้ใช้ทำการคลิกเลือกเมนู From 2 source แล้วนั้น จะปรากฏหน้าจอเพื่อทำการเลือกไฟล์ฐานข้อมูลจากสองแหล่งดังรูปที่ 5.6 โดยผู้ใช้ทำการเลือกไฟล์ฐานข้อมูลด้วยการคลิกเลือกที่ปุ่ม Browse เพื่อทำการเลือกไฟล์ ถ้าผู้ใช้ทำการเลือกไฟล์ที่ชื่อซ้ำกัน โปรแกรมจะทำการส่งข้อความเตือนเพื่อให้ผู้ใช้ทำการเลือกไฟล์ข้อมูลใหม่ ถ้าไฟล์ข้อมูลไม่มีการซ้ำกันและผู้ใช้ต้องการดำเนินการในขั้นตอนต่อไป ให้ผู้ใช้ทำการคลิกปุ่ม Next เพื่อเข้าสู่ขั้นตอนของการทำการรวมข้อมูล ดังรูปที่ 5.7



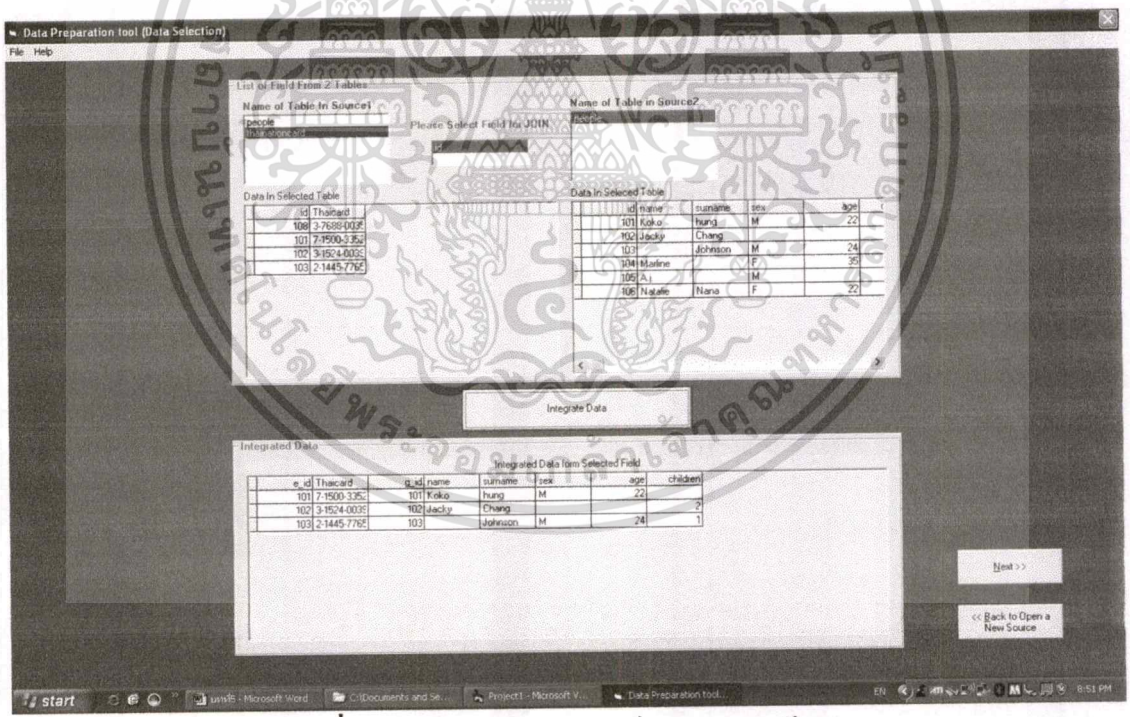
รูปที่ 5.6 หน้าจอการเลือกไฟล์จาก 2 แหล่งข้อมูล

จากรูปที่ 5.7 หน้าจอจะแสดงรายการของตารางที่อยู่ในไฟล์ฐานข้อมูลที่ผู้ใช้ทำการเลือก ผู้ใช้ทำการคลิกเลือกชื่อตารางจากแหล่งที่หนึ่ง เพื่อทำการเปรียบเทียบกับตารางจากแหล่งที่สอง ถ้าตารางมีโครงสร้างเดียวกันหรือมีความสัมพันธ์กัน ปุ่ม Integrate Data จะสามารถทำงานได้ แต่ถ้าตารางของทั้งสองแหล่งไม่มีความสัมพันธ์กัน จะปรากฏข้อความขึ้นมาเตือนและผู้ใช้ก็ต้องทำการเลือกตารางใหม่เพื่อทำการเปรียบเทียบต่อไป

ในกรณีที่ปุ่ม Integrate Data สามารถคลิกเพื่อทำงานได้ แสดงว่าข้อมูลของตารางจากแหล่งข้อมูลที่หนึ่งและแหล่งข้อมูลที่สองสามารถทำการรวมกันได้ ผู้ใช้สามารถทำการกดปุ่ม Integrate Data เพื่อทำการรวมข้อมูล ซึ่งผลลัพธ์ที่ได้สามารถแสดงได้ดังรูปที่ 5.8

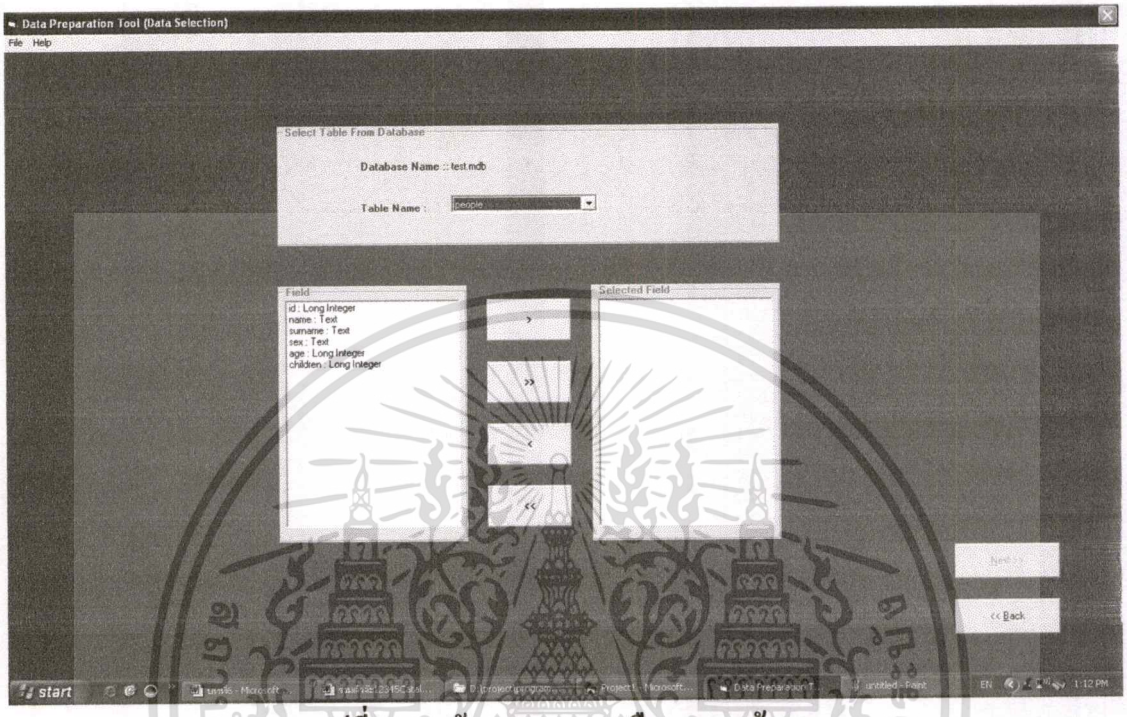


รูปที่ 5.7 หน้าจอการรวมข้อมูล

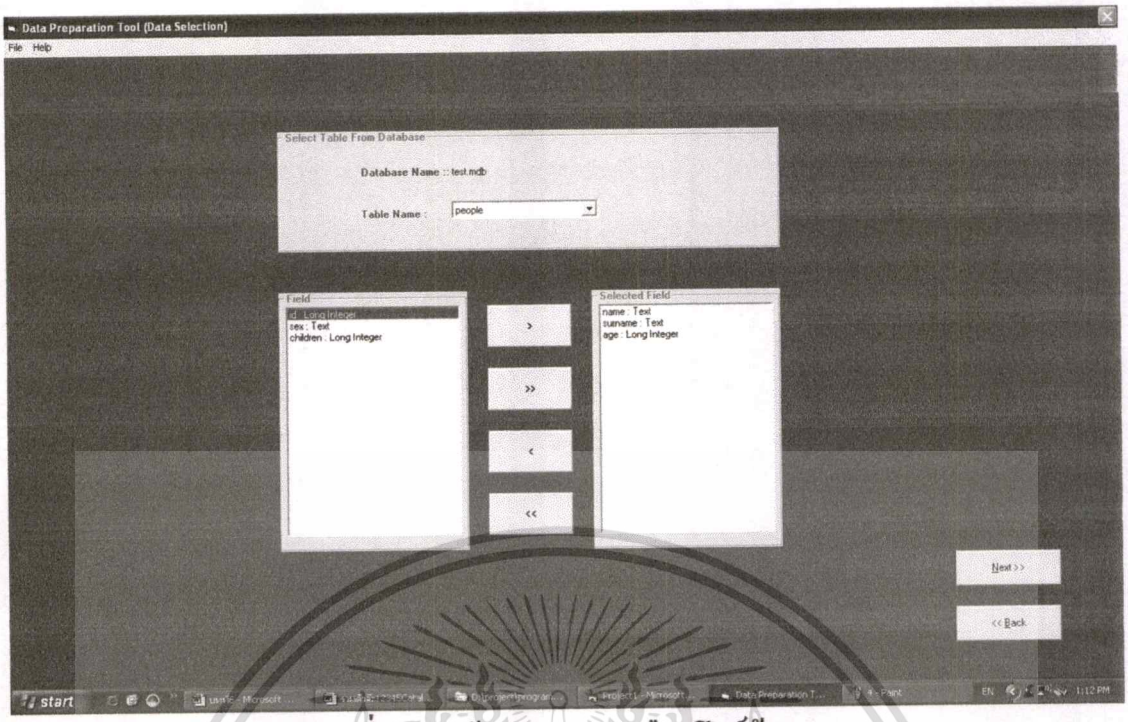


รูปที่ 5.8 หน้าจอแสดงผลลัพธ์ของการรวมข้อมูล

จากรูปที่ 5.4 และรูปที่ 5.8 เมื่อผู้ใช้ทำการกดปุ่ม Next จะเข้าสู่ขั้นตอนในการทำการเลือกฟิลด์ข้อมูลเพื่อใช้ในการทำงานในกระบวนการทำความสะอาดข้อมูลต่อไป ซึ่งหน้าจอที่ปรากฏหลังจากผู้ใช้ทำการกดปุ่ม Next จะแสดงได้ดังรูปที่ 5.9

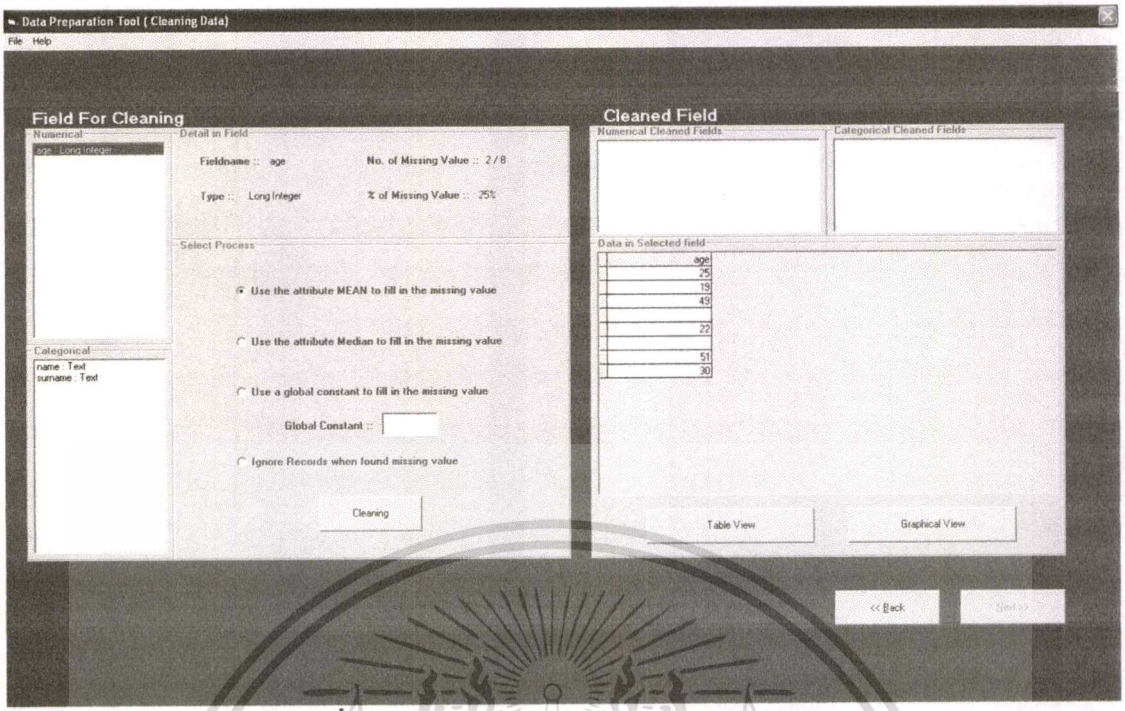


รูปที่ 5.9 หน้าจอแสดงการเลือกตารางข้อมูล



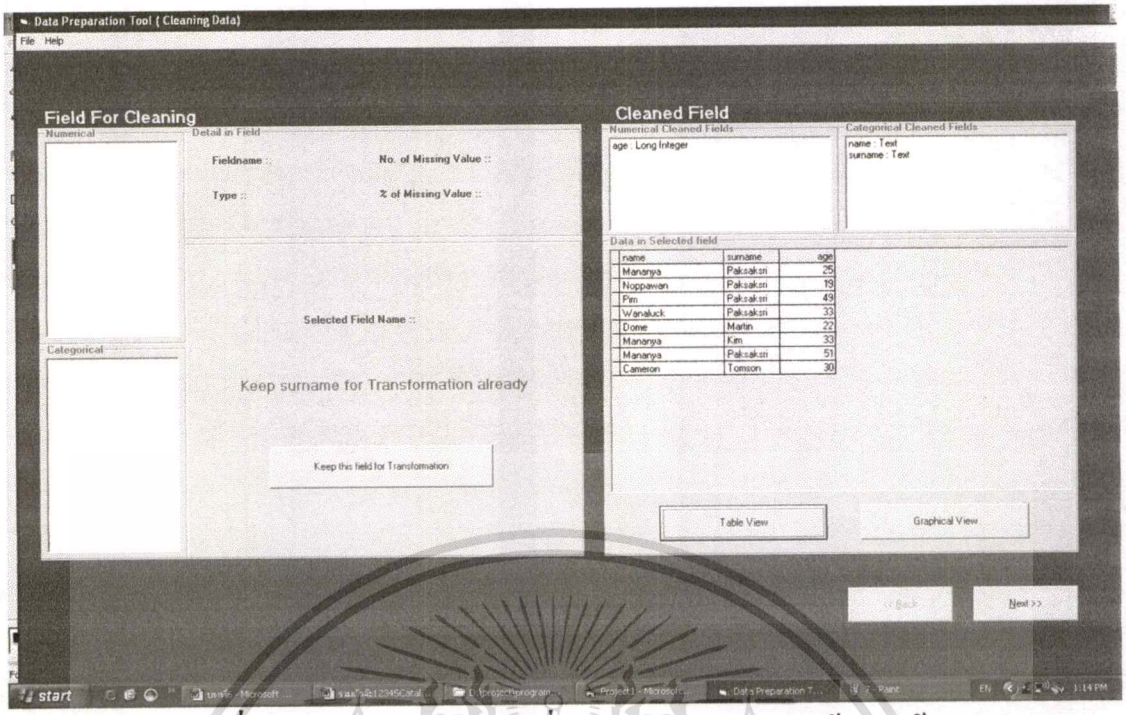
รูปที่ 5.10 หน้าจอแสดงการเลือกฟิลด์ข้อมูล

จากรูปที่ 5.10 ถ้าผู้ใช้ต้องการเลือกฟิลด์ใดเพื่อนำไปใช้งานก็ให้คลิกเลือกที่รายการฟิลด์ด้านซ้ายมือ และทำการคลิกปุ่ม [>] เพื่อทำการเลือกข้อมูลทีละรายการ แต่ถ้าผู้ใช้ต้องการที่จะเลือกฟิลด์ที่อยู่ทางด้านซ้ายมือทั้งหมดก็สามารถคลิกปุ่ม [>>] ได้ ในทางตรงกันข้าม ถ้าผู้ใช้ต้องการจะยกเลิกการเลือกฟิลด์ทีละรายการ ก็สามารถทำได้โดยการคลิกปุ่ม [<] หรือทำการคลิกุปุ่ม [<<] เพื่อทำการยกเลิกฟิลด์ข้อมูลที่เลือกทั้งหมด เมื่อทำการเลือกฟิลด์ข้อมูลที่ต้องการแล้ว ให้ผู้ใช้ทำการคลิกปุ่ม next เพื่อที่จะทำงานในขั้นตอนต่อไปนั้นคือ ขั้นตอนของการทำความสะอาดข้อมูล

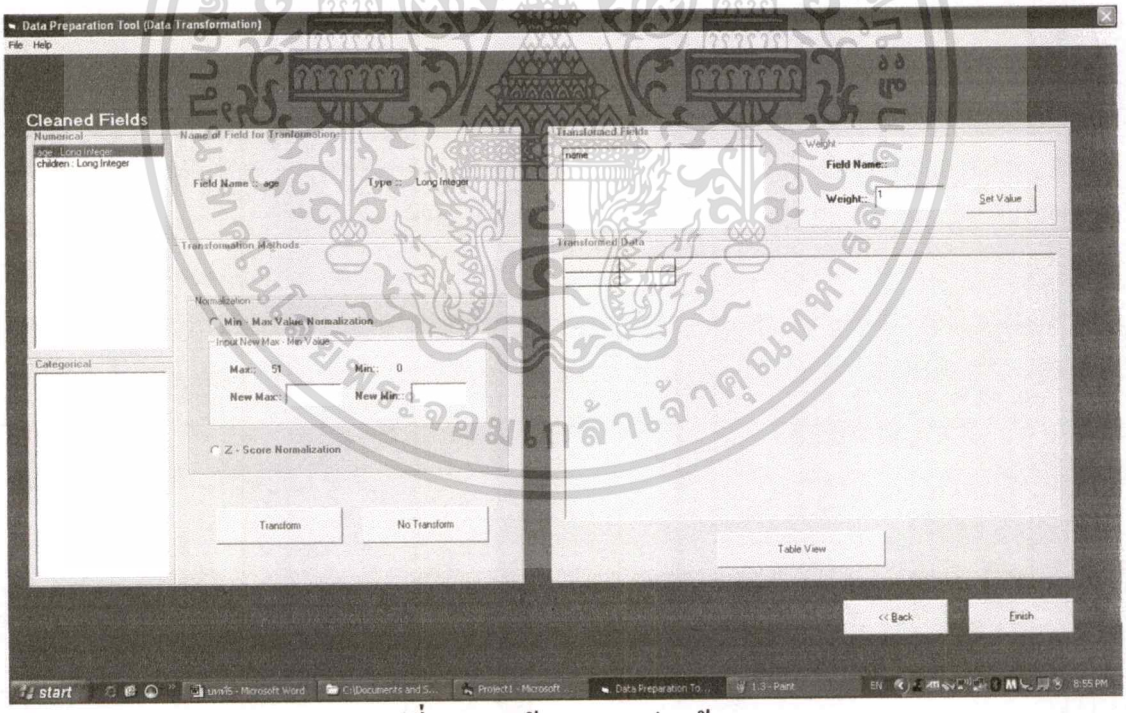


รูปที่ 5.11 หน้าจอของการทำความสะอาดข้อมูล

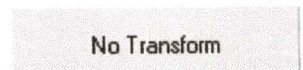
จากรูปที่ 5.11 เป็นหน้าจอของการทำความสะอาดข้อมูล โดยที่ถ้าผู้ใช้ต้องการทำความสะอาดข้อมูลฟิลด์ใด ก็ทำการเลือกฟิลด์นั้น ซึ่งฟิลด์ข้อมูลที่ได้ทำการเลือกมาจากรูปที่ 5.10 จะถูกแบ่งประเภทออกเป็นNumerical และ Categorical ซึ่งในการทำความสะอาดข้อมูลนั้น ขึ้นอยู่กับผู้ใช้ว่าต้องการเลือกวิธีการใดในการทำความสะอาดข้อมูล เมื่อผู้ใช้ทำการกำหนดวิธีการในการทำความสะอาดข้อมูลเรียบร้อยแล้วนั้น ให้ผู้ใช้ทำการคลิกที่ปุ่ม  เพื่อทำความสะอาดข้อมูลฟิลด์นั้นๆ ทำเช่นนี้ไปเรื่อยๆ จนกว่าจะทำความสะอาดข้อมูลครบทุกฟิลด์ เมื่อผู้ใช้ทำความสะอาดข้อมูลครบทุกฟิลด์แล้ว ปุ่ม next จะสามารถทำงานได้ดังรูปที่ 5.12 เพื่อเข้าสู่ขั้นตอนต่อไป นั้นคือการแปลงข้อมูลดังรูปที่ 5.13



รูปที่ 5.12 หน้าจอแสดงข้อมูลที่ผ่านการทำความสะอาดข้อมูลแล้ว

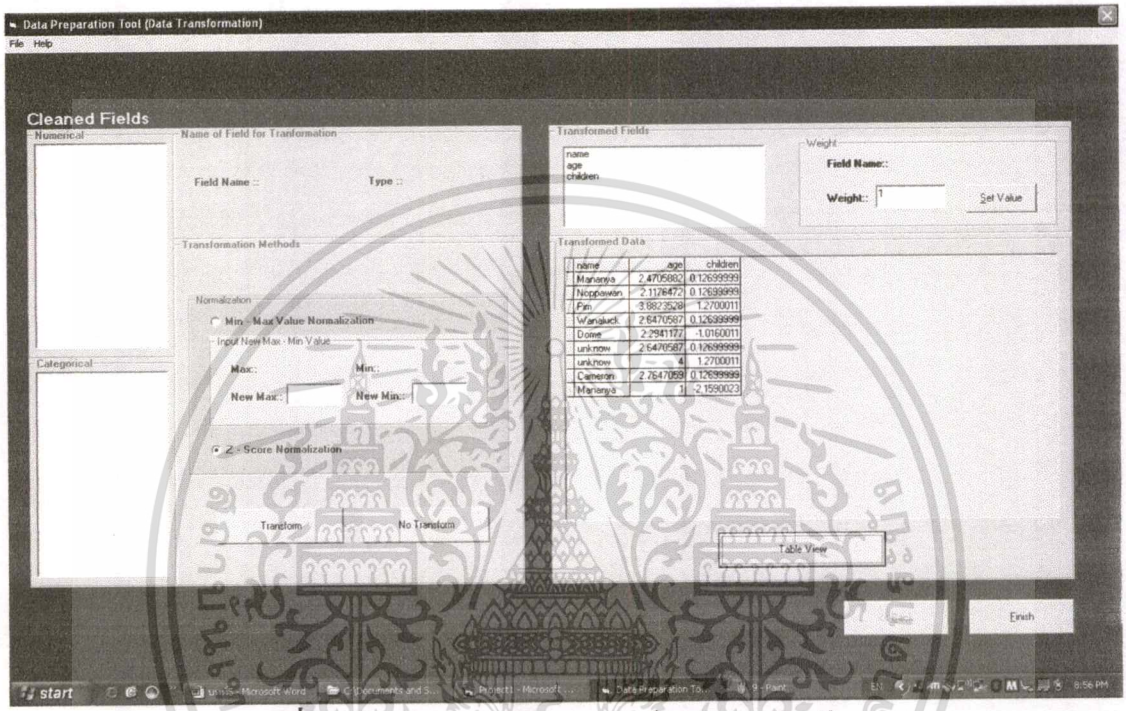


รูปที่ 5.13 หน้าจอการแปลงข้อมูล

จากรูปที่ 5.13 เป็นการแปลงข้อมูลโดยที่ผู้ใช้สามารถเลือกได้ว่าจะทำการแปลงข้อมูลหรือไม่ ถ้าผู้ใช้ไม่ต้องการแปลงข้อมูลฟิลด์ที่ทำการเลือก ก็ให้คลิกปุ่ม  ข้อมูล

ฟิลด์นั้นๆ ก็จะถูกส่งมาเก็บไว้ที่ Transformed Field เพื่อรอใส่ค่าน้ำหนักให้กับฟิลด์ แต่ผู้ใช้ต้องการแปลข้อมูลฟิลด์ใด ก็ให้ทำการคลิกเลือกฟิลด์นั้นๆ และทำการเลือกวิธีการในการแปลงข้อมูล เมื่อเลือกวิธีการในการแปลงข้อมูลเรียบร้อยแล้วให้ผู้ใช้คลิกปุ่ม 

เพื่อทำการแปลงข้อมูลด้วยวิธีการที่ได้เลือกไว้ซึ่งผลลัพธ์จากการแปลงข้อมูลสามารถแสดงได้ดังรูปที่ 5.14



รูปที่ 5.14 หน้าจอการแสดงผลลัพธ์ของการแปลงข้อมูล

บทที่ 6
บทสรุป

6.1 สรุปโครงการ

โครงการพัฒนาเครื่องมือในการทำ Data Praparation นี้ประสบผลสำเร็จตามความมุ่งหมายที่ได้ตั้งใจไว้ทั้งในส่วนของทฤษฎีและปฏิบัติ คือ

- เป็นโครงการพัฒนาระบบที่ทำการออกแบบและสร้างเครื่องมือในการทำการเตรียมข้อมูลให้มีความถูกต้องสมบูรณ์ก่อนที่จะนำข้อมูลดังกล่าวไปทำงานในขั้นตอนของการทำไมนิ่งต่อไป
- ทำการศึกษาทฤษฎีเกี่ยวกับกระบวนการเตรียมข้อมูล รวมทั้งรูปแบบวิธีการเขียนโปรแกรม Visual Basic 6.0 และ Microsoft Access 2000 ใช้ในการพัฒนาเครื่องมือในการทำ Data Preparation
- โครงการพัฒนาระบบงานนี้เริ่มด้วยการวิเคราะห์ระบบว่ามีกระบวนการในการทำงานเป็นอย่างไร มีขั้นตอนในการประมวลผลในแต่ละกระบวนการอย่างไร การออกแบบระบบที่รวมถึงการออกแบบฐานข้อมูลและส่วนติดต่อกับผู้ใช้ ในการพัฒนาระบบจะนำเอากระบวนการที่ได้ทำการวิเคราะห์เอาไว้มาทำการพัฒนาโปรแกรมเพื่อใช้งาน
- โครงการพัฒนาเครื่องมือในการทำ Data Preparation นี้สามารถทำการเตรียมข้อมูลเพื่อนำไปใช้ในการทำคาด้าไมนิ่งได้ ส่วนเครื่องมือที่ได้พัฒนาขึ้นมานั้น สามารถทำงานได้อย่างถูกต้องตรงตามความต้องการและมีประสิทธิภาพ

6.2 ประโยชน์ที่ได้รับจากโครงการพัฒนาระบบงาน

จากการที่ได้ศึกษาค้นคว้าและวิเคราะห์ออกแบบรวมถึงการพัฒนาโครงการพัฒนาระบบงานนั้นได้รับประโยชน์ดังนี้

- ได้รับความรู้ในเรื่องของกระบวนการในการเตรียมข้อมูล
- ได้รับความรู้อย่างละเอียดในเรื่องของขั้นตอนในแต่ละกระบวนการในการเตรียมข้อมูล
- ได้รับประสบการณ์ในการนำความรู้ทางสาขาเทคโนโลยีสารสนเทศมาประยุกต์ใช้ในโครงการพัฒนาระบบ
- ได้รับประสบการณ์ในการออกแบบระบบงานเพื่อให้สอดคล้องกับทฤษฎีที่ได้ศึกษามาเพื่อจะให้ได้ระบบที่ต้องตามความต้องการและมีประสิทธิภาพ

- ได้รับความรู้ในการพัฒนาเครื่องมือด้วยการใช้ Visual Basic 6.0 และ Microsoft Access 2000

6.3 ปัญหา ข้อจำกัด และข้อเสนอแนะ

จากการพัฒนาโครงการพัฒนาระบบงานนั้นมีแนวทางที่ต้องพัฒนาหรือศึกษาต่อดังนี้

- ทำการปรับปรุงเครื่องมือให้สามารถรับข้อมูลจากหลายๆแหล่งข้อมูลที่มีรูปแบบการเก็บข้อมูลที่แตกต่างกันได้
- ทำการปรับปรุงขั้นตอนในการรวมข้อมูลให้สามารถทำการรวมข้อมูลจากหลายๆ แหล่งข้อมูลที่มีรูปแบบการเก็บข้อมูลที่มีความแตกต่างกันได้

บรรณานุกรม

กิตติ ภักดีวัฒนะกุล และ จำลอง ครูอุตสาหะ. 2546. **Visual Basic6 ฉบับฐานข้อมูล**. พิมพ์ครั้งที่ 5. กรุงเทพฯ:เคทีพี คอมพ์ คอนซัลท์.

ฉันทวุฒิ พืชผล และคณะ. 2547. **คู่มือเรียน Visual Basic.** พิมพ์ครั้งที่ 11. กรุงเทพฯ:โปรวิชั่น.

วรรณวิภา ติคณะสิริ. 2548. **คู่มือเรียน SQL ด้วยตัวเอง.** พิมพ์ครั้งที่ 4. กรุงเทพฯ:โปรวิชั่น.

Larose,D.T. 2005. **Discovering Knowledge in Data.** New York:John Wiley and Sons.

Pyle,D. 1999. Data Preparation for Data Mining. USA:Academic Press.

Wang,J. 2003. **Data Mining:Opportunities and Challenge**. Hershey PA:Idea Group Pub.

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวมนันญา ภาคศักดิ์ศรี
วันเดือนปีเกิด	6 มกราคม 2524
สถานที่เกิด	ราชบุรี
ปริญญาตรี	มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ คณะวิทยาศาสตร์และเทคโนโลยี สาขาวิทยาการคอมพิวเตอร์
ปีที่สำเร็จการศึกษา	2545
มัธยมศึกษา	โรงเรียนชลราษฎรอำรุง จ.ชลบุรี
ประถมศึกษา	โรงเรียนมารดานฤมล จ.ฉะเชิงเทรา