การพัฒนาเว็บเซอร์วิสสำหรับเหมืองข้อมูล

PROPOSAL OF WEB SERVICES FOR DATA MINING

โดย

นพปฎล ถาวรรัฐ

NOPPAPADON THAWORNRATH

อาจารย์ที่ปรึกษา

รศ.ดร. วรพจน์ กรีสุระเดช

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน
หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 1 ปีการศึกษา 2549

PROPOSAL OF WEB SERVICES FOR DATA MINING

NOPPAPADON THAWORNRATH

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

1 / 2006

หัวข้อ	การพัฒนาเว็บเซอร์วิสสำหรับเหมืองข้อมูล
นักศึกษา	นายนพปฎล ถาวรรัฐ
รหัสนักศึกษา	47066414
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2549
อาจารย์ที่ปรึกษา	รศ.ดร.วรพจน์ กรีสุระเดช

บทคัดย่อ

โครงงานฉบับนี้เสนอวิธีการพัฒนาเว็บเซอร์วิสซึ่งให้บริการการทำเหมืองข้อมูล โดยใช้ฟังก์ชันการทำงานของเหมืองข้อมูลเป็นการให้บริการบนเว็บเซอร์วิส ทำให้แอพพลิเคชันของเหมืองข้อมูลไม่เป็นการยึดติดกับแพลตฟอร์มใดๆ ซึ่งจะช่วยให้การใช้งานและการพัฒนาแอพพลิเคชันเหมืองข้อมูลเป็นไปได้กว้างขวางมากยิ่งขึ้น โดยจะเลือกใช้การทำงานเหมืองข้อมูลแบบแบ่งกลุ่ม (Clustering) และใช้อัลกอริทึม K-Means และ ISODATA การพัฒนาโปรแกรมจะใช้จาวาเป็นภาษาในการพัฒนาร่วมกับ Apache Axis ซึ่งเป็นเครื่องมือในการพัฒนาเว็บเซอร์วิส ซึ่งระบบการใช้งานเว็บเซอร์วิสจะใช้ภายในเครือข่ายองค์กรเท่านั้น

Title	Web services for data mining
Student	Mr. Noppapadon Thawornrath
Student ID.	47066414
Degree	Master of Science
Programme	Information Science
Academic Year	2006
Advisor	Assoc. Prof. Dr. Worapoj Kresuradej

ABSTRACT

This proposal propose web services for data mining with data mining operation is a service for mining. The web services for data mining can allow most data mining applications which use web services are independent from any platform and accommodation for application development or used in organization is easily. In this application use clustering as a task for mining and use K-Means and ISODATA algorithm. Java is programming language for development this application and use Apache Axis for create and publish web services. Web services system is used in organization network.

กิตติกรรมประกาศ

โครงงานนี้สำเร็จได้อย่างดี ด้วยคำแนะนำและคำปรึกษาจาก รศ.ดร.วรพจน์ กรีสุระเดช ซึ่งเป็นอาจารย์ที่ปรึกษาโครงงาน ขอขอบพระคุณเป็นอย่างสูงที่ให้คำแนะนำที่เป็นประโยชน์ในการทำโครงงานนี้

ขอกราบขอบพระคุณคณาจารย์ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุก ๆ ท่านที่ได้ประสิทธิ์ประสาทวิชาให้กับข้าพเจ้า

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ ในภาควิชาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทุกคนที่ให้คำแนะนำต่างๆ และคอยให้กำลังใจเสมอมา

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณ บิดา มารดา และครอบครัวของข้าพเจ้าที่เป็นกำลังใจและให้การสนับสนุนในทุกเรื่องๆ ทำให้ข้าพเจ้าสามารถทำโครงงานฉบับนี้สำเร็จลุล่วงด้วยดี

คุณค่าและประโยชน์อันพึงมาจากโครงงานฉบับนี้ ข้าพเจ้าขอมอบแด่ผู้มีพระคุณทุกท่าน

นพปฎล ถาวรรัฐ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเหมืองข้อมูลเป็นกระบวนการที่สำคัญที่ถูกใช้ในองค์กรต่างๆ ซึ่งเป็นการค้นหาข้อมูลที่มีอยู่อย่างมหาศาลในฐานข้อมูลเพื่อหาความสัมพันธ์ของรูปแบบข้อมูลที่ไม่เคยรู้มาก่อนหรือใช้ในการทำนายแนวโน้มของข้อมูล ถือได้ว่าเป็นการค้นพบองค์ความรู้ใหม่จากข้อมูลในฐานข้อมูลซึ่งมีอยู่เป็นจำนวนมาก โดยเหมืองข้อมูลได้เป็นที่นิยมใช้ในงานต่างๆ เช่น วิเคราะห์ข้อมูลการขาย ค้นหาความสัมพันธ์ของลูกค้า ทำนายแนวโน้มการขายสินค้า หรือการนำเอาไปใช้กับธุรกิจอัจฉริยะ (business intelligence) และอีกหลายการทำงานที่เกี่ยวข้องในการวิเคราะห์และทำนายข้อมูล เป็นต้น และในการที่ข้อมูลได้ถูกเก็บอยู่ในแหล่งต่างๆ ทำให้การจัดรวบรวมข้อมูลหรือการทำการวิเคราะห์เป็นไปด้วยความยากลำบากเพราะแอพพลิเคชันที่ใช้จัดการอาจแตกต่างกัน ซึ่งในแต่ละผลิตภัณฑ์นั้นๆ การทำงานและอัลกอริทึมของเหมืองข้อมูลจะถูกเขียนหรือติดตั้งอยู่ในตัวของตัวมันเอง ทำให้ในการใช้งานเหมืองข้อมูลจากเครื่องลูกข่ายอื่นๆ จำเป็นต้องใช้แอพพลิเคชันตัวเดียวกันหรือใช้โซลูชันของผลิตภัณฑ์นั้นๆ หรือการพัฒนาแอพพลิเคชันขึ้นมารองรับเป็นไปด้วยความยากลำบาก ถึงแม้จะมี API ในการติดต่อระหว่างแอพพลิเคชัน แต่ก็ยังไม่เป็นมาตรฐาน จึงทำให้การใช้งานแอพลิเคชันทางด้านเหมืองข้อมูลยังมีข้อจำกัดทางด้านการนำมาใช้งานและการพัฒนาแอพพลิเคชันซึ่งยังไม่เป็นที่แพร่หลายนัก

เว็บเซอร์วิสเป็นเทคโนโลยีที่ให้บริการโปรแกรมการทำงานอย่างใดอย่างหนึ่ง ซึ่งถูกเรียกใช้โดยแอพพลิเคชันผ่านทางหน้าเว็บ โดยมีภาษากลางในการสื่อสารคือ XML เมื่อนำมาประยุกต์ใช้ร่วมกับเหมืองข้อมูลโดยสร้างบริการทางด้านการทำงานหรือฟังก์ชันของเหมืองข้อมูลทำให้นักพัฒนาสามารถพัฒนาแอพพลิเคชันเพื่อใช้ประโยชน์จากเหมืองข้อมูลได้โดยเรียกใช้งานผ่านทางเว็บเซอร์วิส ซึ่งจะไม่มีการยึดติดกับแพลตฟอร์มใดๆ หรือภาษาใดๆ โดยจะใช้ XML เป็นภาษากลางในการสื่อสารกันของข้อมูล จึงทำให้การใช้งานของเหมืองข้อมูลมีความหลากหลายมากยิ่งขึ้นและนักพัฒนาจะไม่ต้องเสียเวลาในการพัฒนาแอพพลิเคชันเพิ่มเติมอีกทั้งมีความสะดวกในการปรับปรุงหรือเปลี่ยนแปลงการทำงานของเหมืองข้อมูลได้โดยง่าย โดยไม่ส่งผลกระทบต่อระบบเดิม

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

โครงงานฉบับนี้มุ่งหวังเพื่อศึกษาและพัฒนาเว็บเซอร์วิสสำหรับเหมืองข้อมูล เพื่อให้การพัฒนาแอพพลิเคชันของเหมืองข้อมูลเป็นไปได้อย่างสะดวกรวดเร็วซึ่งนักพัฒนาไม่จำเป็นที่จะต้องพัฒนาฟังก์ชันการทำงานในการคำนวณจากอัลกอริทึมของเหมืองข้อมูล โดยจะใช้การเรียกใช้ฟังก์ชันการทำงานของอัลกอริทึมของเหมืองข้อมูลผ่านเว็บเซอร์วิส และยังช่วยให้การปรับเปลี่ยนเพิ่มเติมอัลกอริทึมและการทำงานของเหมืองข้อมูลเป็นไปได้โดยสะดวกซึ่งจะไม่เกิดผลกระทบต่อแอพพลิเคชันเดิม โดยการใช้เว็บเซอร์วิสจะกระทำภายในเครือข่ายในองค์กร

1.3 สมมติฐานของการศึกษา

ข้อดีของเทคโนโลยีเว็บเซอร์วิสคือสามารถใช้ได้จากสภาวะแวดล้อมใดๆ ก็ได้ ไม่ยึดติดกับแพลทฟอร์ม โดยเป็นการแลกเปลี่ยนข้อมูลกันระหว่างแอพลลิเคชัน ซึ่งผู้ใช้ไม่จำเป็นต้องรู้ถึงข้อมูลภายใน อีกทั้งยังมีค่าใช้จ่ายในการพัฒนารวมถึงการปรับเปลี่ยนการทำงานที่น้อยกว่าการสร้างแอพพลิเคชันมาใช้เฉพาะงาน และยังสามารถเข้าถึงจากที่ใดก็ได้เพราะเป็นการทำงานบนระบบเครื่อข่าย

จากข้อดีของเว็บเซอร์วิสที่ได้กล่าวมาข้างต้น จึงได้เกิดแนวความคิดในการนำเทคโนโลยีของเว็บเซอร์วิสมาใช้ทำงานร่วมกับเหมืองข้อมูล ซึ่งเหมืองข้อมูลในปัจจุบันเริ่มมีบทบาทมากยิ่งขึ้นในการค้นหาองค์ความรู้ขององค์กร ดังนั้นจึงสร้างเว็บเซอร์วิสที่ให้บริการการคำนวณจากอัลกอริทึมของเหมืองข้อมูล โดยการเรียกใช้งานจากแอพพลิเคชันเหมืองข้อมูลที่พัฒนาให้สามารถติดต่อกับเว็บเซอร์วิสได้

1.4 ทฤษฎีหรือแนวคิดที่ใช้ในการวิจัย

การทำงานของเหมืองข้อมูลด้วยเทคนิคการแบ่งกลุ่ม (Clustering) จะถูกนำมาใช้เป็นบริการบนเว็บเซอร์วิส โดยการพัฒนาเว็บเซอร์วิสจะใช้เครื่องมือช่วยในการสร้าง เพื่อความสะดวกรวดเร็ว โดยสร้างอัลกอริทึมการคำนวณเพื่อให้แอพพลิเคชันเรียกใช้ฟังก์ชันผ่านเว็บเซอร์วิส ซึ่งอัลกอริทึมที่ใช้สำหรับสร้างแบบจำลองการแบ่งกลุ่มก็คือ K-Means และ ISODATA ตามลำดับ โดย K-Means เป็นวิธีการแบ่งกลุ่มที่มีขั้นตอนไม่ซับซ้อนมากนักเหมาะแก่การเรียนรู้เพื่อให้เข้าใจการทำงานของวิธีการแบ่งกลุ่ม ในขณะที่ ISODATA เป็นอัลกอริทึมที่ซับซ้อนขึ้นมากกว่าเดิม โดยสามารถกำหนดจำนวนกลุ่มข้อมูลที่เหมาะสม โดยการรวมกลุ่มเข้าด้วยกันเมื่อกลุ่มใกล้เคียงใกล้กันเกินไป หรือการแตกกลุ่มออกมาเมื่อถึงจุดที่พิจารณาว่าสมควรแตกกลุ่ม

1.5 ขอบเขตการพัฒนา

ในโครงงานนี้เป็นการพัฒนาเว็บเซอร์วิสสำหรับเหมืองข้อมูล ใช้ภาษาจาวาในการพัฒนา โดยขั้นตอนการพัฒนาแบ่งเป็น 2 ส่วน โดยส่วนแรกจะพัฒนาส่วนไคลเอ็นท์ขึ้นมาเพื่อแสดงการเรียกใช้งานข้อมูล การเตรียมข้อมูลเพื่อเข้าสู่กระบวนการทำเหมืองข้อมูลได้แก่ การแก้ไขข้อมูลที่ผิดพลาด การนอมอลไลซ์ข้อมูล และ การแสดงผลการดำเนินการ ซึ่งข้อมูลที่นำมาใช้สามารถใช้ได้จากไฟล์และจากฐานข้อมูล โดยฐานข้อมูลที่ใช้คือ PostgreSQL และในส่วนที่สองจะพัฒนาเว็บเซอร์วิสโดยใช้ Apache Axis ซึ่งเป็นเครื่องมือสำหรับใช้ในการสร้างเว็บเซอร์วิส โดยฟังก์ชันการทำงานของเหมืองข้อมูลจะใช้การทำงานแบบแบ่งกลุ่ม (Clustering) เป็นส่วนให้บริการบนเว็บเซอร์วิส

1.6 ขั้นตอนของการศึกษา

โครงงานฉบับนี้ได้แบ่งเนื้อหาออกเป็น 6 บทด้วยกันคือ

บทที่ 1 กล่าวถึงความเป็นมาของงานวิจัย ความมุ่งหมายและวัตถุประสงค์ สมมติฐาน ทฤษฎีที่ใช้ ขอบเขตของการพัฒนา และขั้นตอนการศึกษา

บทที่ 2 กล่าวถึงทฤษฎีพื้นฐานที่ใช้ในการพัฒนา ได้แก่พื้นฐานของเหมืองข้อมูลและเว็บเซอร์วิส

บทที่ 3 กล่าวถึงทฤษฎีเทคนิคการแบ่งกลุ่ม (Clustering) อัลกอริทึมที่ใช้ได้แก่ K-Means และ ISODATA และวิธีการสร้างเว็บเซอร์วิสโดยใช้ Apache Axis

บทที่ 4 กล่าวถึงการวิเคราะห์และออกแบบระบบ

บทที่ 5 กล่าวถึงการนำเสนอการทำงานของโปรแกรมและการใช้บริการผ่านเว็บเซอร์วิส

บทที่ 6 เป็นบทสรุปผลการพัฒนาและข้อเสนอแนะ

บทที่ 2
ทฤษฎีพื้นฐานที่ใช้ในการพัฒนาเว็บเซอร์วิสสำหรับเหมืองข้อมูล

ในหัวข้อนี้จะกล่าวถึงทฤษฎีพื้นฐานต่างๆ ที่ใช้ในโครงงานนี้ซึ่งได้แก่ ทฤษฎีที่เกี่ยวข้องในการพัฒนาเว็บเซอร์วิส เทคโนโลยีของเว็บเซอร์วิส ทฤษฎีของเหมืองข้อมูล ซึ่งเนื้อหาทั้งหมดนี้มีความสำคัญต่อการพัฒนาเว็บเซอร์วิส และความเข้าใจในเรื่องของเหมืองข้อมูล

2.1 เหมืองข้อมูล

เหมืองข้อมูล (Data Mining) เป็นวิธีการสืบค้นองค์ความรู้และสิ่งที่น่าสนใจในฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases : KDD) หรือกล่าวได้ว่า เป็นเทคนิคเพื่อจัดการกับข้อมูลที่มีขนาดใหญ่ เพื่อนำมาวิเคราะห์หาความสัมพันธ์ หรือลักษณะเด่น และใช้ในการทำนายสิ่งที่จะเกิดขึ้น เช่น การค้นหารายชื่อกลุ่มลูกค้าที่สนใจสินค้าประเภทเดียวกัน จากฐานข้อมูลลูกค้าและการสั่งซื้อของเป็นต้น ทั้งนี้ เพื่อนำไปใช้ในการตัดสินใจ เพื่อให้เกิดประโยชน์ต่อการเพิ่มประสิทธิภาพการใช้งานข้อมูลกลุ่มนั้น ๆ โดยเหมืองข้อมูลจะช่วยเพิ่มมูลค่าให้แก่องค์กรธุรกิจที่นำไปใช้ ยกตัวอย่างได้เช่น ในปัจจุบันเทคโนโลยีการเก็บข้อมูลสามารถเก็บข้อมูลได้เพิ่มมากขึ้น ซึ่งถ้าองค์กรมีการเก็บข้อมูลที่เพิ่มมากขึ้นเรื่อยๆ แล้ว ก็ยังสามารถค้นหารูปแบบข้อมูลที่ซ่อนอยู่ ทำให้เกิดองค์ความรู้ใหม่จากข้อมูลที่มีอยู่เป็นจำนวนมาก ซึ่งถ้าไม่มีการนำเหมืองข้อมูลมาใช้อาจจะไม่สามารถใช้ประโยชน์จากข้อมูลที่มีอยู่ได้เลย นอกจากนี้เหมืองข้อมูลยังมีความพร้อมในการเติบโตของเทคโนโลยีเพื่อตอบสนองต่อการแข่งขันกันทางธุรกิจที่สูงขึ้น สามารถวิเคราะห์ถึงปัจจัยความต้องการของลูกค้าบนเครือข่ายอินเตอร์เน็ทและระบบการสื่อสารข้อมูลต่างๆ ได้อย่างดี นอกจากนี้เทคโนโลยีของเหมืองข้อมูลยังสามารถพัฒนาให้มีประสิทธิภาพเพิ่มมากยิ่งขึ้น เพื่อตอบสนองแก่ผู้ใช้และนักพัฒนาแอพพลิเคชันต่างๆ เช่นการเพิ่มความสามารถด้านการคำนวณ ความรวดเร็วในการทำงาน หรือการมี APIs ซึ่งก็คือส่วนติดต่อระหว่างแอพพลิเคชันเพื่อเป็นมาตรฐานให้นักพัฒนาพัฒนาระบบงานได้ง่ายขึ้น

เทคนิคของเหมืองข้อมูลสามารถนำไปใช้ตอบปัญหาหรือแก้ไขปัญหาทางธุรกิจได้หลากหลายประเภท ยกตัวอย่างเช่น [4]

การวิเคราะห์ความไม่มั่นคงของลูกค้า (Churn analysis) : เป็นการวิเคราะห์พฤติกรรมของลูกค้าที่มีการเปลี่ยนแปลงการใช้บริการ หรือเปลี่ยนไปใช้บริการกับผู้อื่น ซึ่งมีผลต่อการแข่งขันกันทางด้านการตลาด การวิเคราะห์นี้ช่วยให้นักการตลาดมีความเข้าใจถึงเหตุผลในการเปลี่ยนแปลงการบริโภคของลูกค้าได้ดียิ่งขึ้น และนำไปปรับปรุงให้เกิดความสัมพันธ์ที่ดีต่อลูกค้าและทำให้ลูกค้ามีความจงรักภักดีต่อบริษัทตลอดไป

การขายสินค้าที่เกี่ยวข้องกัน (Cross selling) : ช่วยส่งเสริมการขายสินค้าแก่ลูกค้า ซึ่งลูกค้าที่ซื้อสินค้าชนิดหนึ่งอาจซื้อสินค้าอีกชนิดหนึ่งที่น่าสนใจหรือเป็นสินค้าที่มีความเกี่ยวข้องกันด้วยก็ได้ โดยเหมืองข้อมูลจะช่วยในการวิเคราะห์หาความสัมพันธ์กันของสินค้าที่ลูกค้าสนใจ

การตรวจหากลโกง (Fraud detection) : ปัญหาที่เกิดจากการขอสิทธิประกันภัยของลูกค้า ต่อบริษัทประกันภัยต่างๆ ซึ่งเหมืองข้อมูลจะช่วยในการตรวจสอบหาว่าการขอสิทธิประกันใดเป็นเท็จหรือไม่ โดยดูจากข้อมูลที่ลูกค้าร้องเรียนเข้ามา

การจัดการความเสี่ยง (Risk management) : เหมืองข้อมูลสามารถช่วยในการคำนวณและแสดงผลข้อมูลที่เกี่ยวกับความเสี่ยงที่จะเกิดขึ้นของลูกค้าต่อบริษัท ซึ่งอาจจะก่อให้เกิดปัญหาทตามมาของบริษัทได้ ทำให้บริษัทสามารถตัดสินใจได้ทันท่วงที

การแบ่งกลุ่มลูกค้า (Customer segmentation) : การแบ่งกลุ่มลูกค้าช่วยให้บริษัทสามารถแบ่งกลุ่มพฤติกรรมของลูกค้า เพื่อให้การขายสินค้าและบริการต่างๆ ตรงกับกลุ่มของลูกค้ามากยิ่งขึ้น

การพยากรณ์การขาย (Sell forcast) : เหมืองข้อมูลช่วยให้สามารถทำนายการขายสินค้าได้ตรงตามวัตถุประสงค์ในอนาคตได้

ดังที่กล่าวมาข้างต้น เหมืองข้อมูลช่วยให้องค์กรสามารถกำหนดกลยุทธ์ทางการตลาดและเพิ่มความสามารถในการจัดการได้ดีขึ้น ซึ่งการนำเหมืองข้อมูลไปใช้ในการแก้ปัญหาหรือหาคำตอบในธุรกิจได้อย่างมีประสิทธิภาพแล้วก็ยิ่งจะทำให้องค์กรสามารถมีโอกาสในการเป็นผู้นำทางการแข่งขันได้มากขึ้นด้วย และในหัวข้อต่อไปนี้จะกล่าวถึงเทคนิคของเหมืองข้อมูลและวิธีการทำงานของเทคนิคนั้นๆ

2.1.1 การทำงานของเหมืองข้อมูล (Data Mining Tasks)

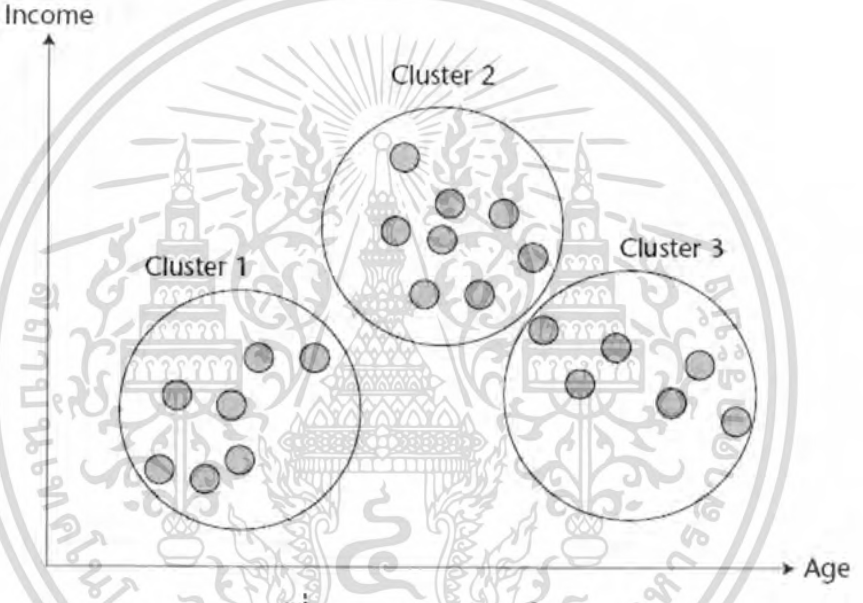
เหมืองข้อมูลสามารถแก้ปัญหาทางธุรกิจได้หลายวิธี ซึ่งเทคนิคการทำงานของเหมืองข้อมูลก็จะสนับสนุนการนำไปใช้แก้ปัญหาเหล่านั้น โดยการทำงานของเหมืองข้อมูลสามารถแบ่งออกได้ดังนี้

2.1.1.1 การจัดกลุ่ม (Classification)

การจัดกลุ่มเป็นวิธีการกำหนดกลุ่มของข้อมูล โดยใช้ข้อมูลที่ต้องการทำนายเป็นข้อมูลในการจัดกลุ่ม ซึ่งการเลือกแอททริบิวท์ที่ต้องการนำมาทำนายการจัดกลุ่ม เรียกว่า คลาส (class) และจะสร้างแบบจำลองเพื่อใช้ทำนายการจัดกลุ่ม การจัดกลุ่มเป็นการเรียนรู้แบบมีผู้สอน (supervised learning) ซึ่งจะต้องมีการกำหนดเป้าหมายซึ่งก็คือคลาสนั่นเอง เทคนิคอัลกอริทึมที่ใช้ได้แก่ ต้นไม้ตัดสินใจ (Dicision Trees) และ โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นต้น

2.1.1.2 การแบ่งกลุ่ม (Clustering)

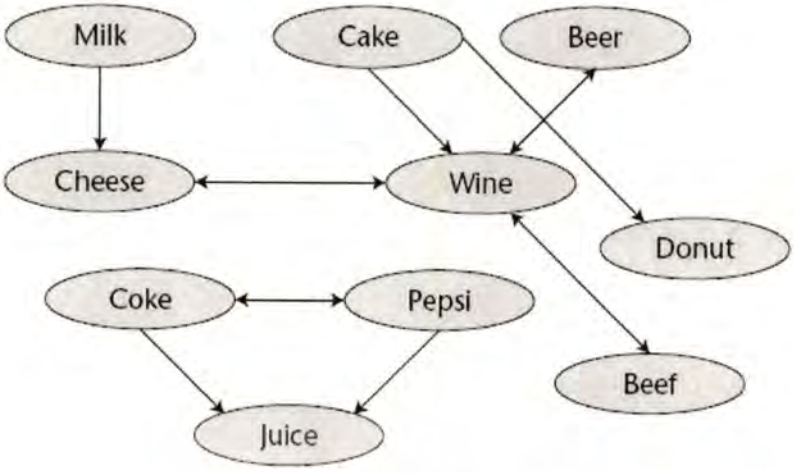
การแบ่งกลุ่ม หรือเรียกได้อีกชื่อหนึ่งว่า การแบ่งออกเป็นส่วน (Segmentation) จะใช้ในการแบ่งข้อมูลออกเป็นกลุ่ม ซึ่งเราไม่สามารถระบุข้อมูลได้ว่าข้อมูลมีกี่กลุ่มอะไรบ้าง การแบ่งกลุ่มนี้ข้อมูลที่อยู่กลุ่มเดียวกันจะมีลักษณะของข้อมูลที่เหมือนกันหรือมีความคล้ายกัน การแบ่งกลุ่มเป็นการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) ซึ่งจะไม่มีการกำหนดเป้าหมายให้กับการทำงาน มีการทำงานแบบทำซ้ำหลายๆครั้ง การเรียนรู้จะเสร็จสิ้นขั้นตอนก็ต่อเมื่อการแบ่งกลุ่มไม่สามารถแบ่งข้อมูลได้อีก ดังรูปที่ 2.1 แสดงการแบ่งข้อมูลโดยนำรายได้และอายุของลูกค้ามาแบ่งกลุ่ม



รูปที่ 2.1 การแบ่งกลุ่มข้อมูล

2.1.2.3 ความสัมพันธ์ (Association)

การทำงานนี้เป็นการหาความสัมพันธ์กันระหว่างข้อมูล บางครั้งเรียกว่าการวิเคราะห์ตลาด (Market Basket Analysis) เพื่อหาความสัมพันธ์ของสินค้าที่ลูกค้าจะสนใจซื้อพร้อมกัน ในทางธุรกิจจะใช้การวิเคราะห์นี้หาแนวโน้มเพื่อเพิ่มยอดการขาย วัตถุประสงค์ในการทำงานนี้คือการหาความถี่ของข้อมูล และการหากฎความสัมพันธ์ โดยในขั้นแรกจะหาความถี่ของข้อมูลที่สนใจ จากนั้นจึงนำไปหาความสัมพันธ์กับข้อมูลอื่นเพื่อหาความน่าจะเป็นที่จะนำมาสัมพันธ์กัน



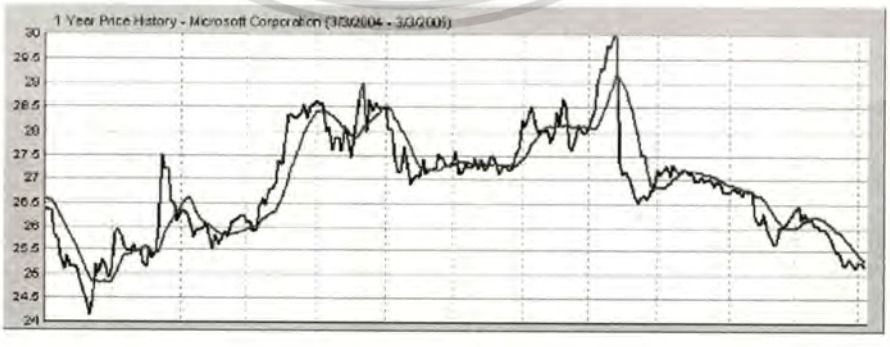
รูปที่ 2.2 ความสัมพันธ์ของข้อมูล

2.1.2.4 การวิเคราะห์ถดถอย (Regression) :

การทำงานนี้จะคล้ายกับการจัดกลุ่ม ซึ่งจะเน้นที่การทำนายค่าของข้อมูลที่จะเกิดขึ้นในอนาคต โดยใช้การคำนวณแบบเชิงเส้น เทคนิคที่ใช้ในการทำงานนี้ได้แก่ regression trees และโครงข่ายประสาทเทียม

2.1.2.5 การพยากรณ์ (Forecasting)

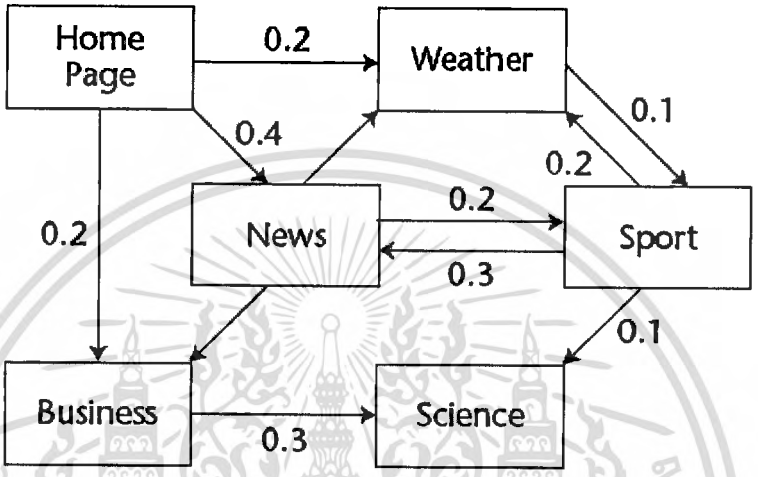
ใช้ในการทำนายหาผลลัพธ์ที่จะเกิดขึ้นในอนาคต โดยใช้กลุ่มข้อมูลอนุกรมเวลา (time series) เป็นข้อมูลนำเข้า ซึ่งก็คือแอททริบิวท์ที่เป็นข้อมูลแสดงลำดับเวลา โดยการวิเคราะห์อนุกรมเวลาจะดูการเคลื่อนไหวของข้อมูลในขอบเขตของเวลา ส่วนใหญ่จะแสดงผลลัพธ์เป็นกราฟเพื่อแสดงแนวโน้มของส่วนประกอบที่เกี่ยวข้องที่มีผลต่อการเปลี่ยนแปลงของเวลา โดยในรูปที่ 2.3 จะแสดงกราฟอนุกรมเวลาที่เกิดขึ้นในช่วงเวลานั้นๆ



รูปที่ 2.3 กราฟอนุกรมเวลา

2.1.2.6 การวิเคราะห์ลำดับ (Sequence Analysis)

การวิเคราะห์ลำดับเป็นการหารูปแบบข้อมูลที่ไม่ต่อเนื่องกัน ซึ่งจะแตกต่างจากอนุกรมเวลาที่เป็นข้อมูลแบบต่อเนื่อง โดยการทำงานนี้จะเน้นไปที่การวิเคราะห์การเปลี่ยนสถานะของข้อมูล วิเคราะห์หาความสัมพันธ์ของสถานะของข้อมูลที่เปลี่ยนแปลงไปตามลำดับ โดยรูปที่ 2.4 จะแสดงลำดับการเปิดหน้าเว็บในการเข้าไปดูเว็บไซต์ของลูกค้า



รูปที่ 2.4 การเปลี่ยนสถานะของข้อมูล

2.1.2.7 การวิเคราะห์ความเบี่ยงเบน (Deaviation Analysis)

การวิเคราะห์ความเบี่ยงเบนจะใช้สำหรับการหาตัวอย่างข้อมูลที่แตกต่างหรือเบี่งเบนออกจากข้อมูลอื่นๆ เช่น การตรวจจับข้อมูลบัตรเครดิตที่มีความปกติ การตรวจจับการบุกรุกในเครือข่าย เป็นต้น เป็นการหาข้อมูลที่ผิดปกติจากข้อมูลที่เกิดขึ้นประจำวันที่เกิดขึ้นเป็นจำนวนมาก

2.1.2 กระบวนการของเหมืองข้อมูล (Data Mining Process)

กระบวนการของเหมืองข้อมูล มีขั้นตอนวิธีตามกระบวนการของ CRISP-DM (CRoss Industry Standard Process for Data Mining) ซึ่งกลุ่มของผู้ร่วมพัฒนาแอพพลิเคชันเหมืองข้อมูลได้กำหนดขึ้นมาใช้ร่วมกัน ปัจจุบันอยู่ในเวอร์ชัน 1.0 โดยมีขั้นตอนต่างๆ ดังต่อไปนี้ [1][2]

2.1.2.1 การทำความเข้าใจกับปัญหา (Problem Understanding)

เป็นขั้นตอนแรกในการทำความเข้าใจกับวัตถุประสงค์ในสิ่งที่จะทำและมุมมองของงาน จากนั้นจะนำความรู้ที่ได้มาเข้าสู่การทำเหมืองข้อมูล และมีการวางแผนออกแบบการทำงานในแต่ละวัตถุประสงค์ โดยมีรายละเอียดดังต่อไปนี้

1. กำหนดวัตถุประสงค์ (Determining objective)

เป็นขั้นตอนแรกที่สำคัญในกระบวนการของเหมืองข้อมูล เป็นการทำความเข้าใจในปัญหา ระบุปัญหาและหาแนวทางการแก้ไขในแต่ละปัญหา ซึ่งปัญหาอาจมีด้วยกันหลายประการ เช่น การตอบรับคำร้องของลูกค้าในเรื่องการแข่งขันกันทางการตลาด การป้องกันการโกงบัตรเครดิต การตรวจจับการใช้งานของระบบคอมพิวเตอร์ เป็นต้น

2. กำหนดบรรทัดฐานของความสำเร็จ (Define success criteria)

เป็นขั้นตอนที่จะกำหนดขอบเขตในการทำเหมืองข้อมูลว่าแค่ไหนถึงจะเสร็จสิ้นการทำเหมืองข้อมูล มีความสำเร็จตามเป้าหมายในแต่ละปัญหาที่ได้กำหนดไว้

3. การประเมินสถานะ (Assess situation)

เป็นการประเมินความเป็นไปได้ในการเริ่มต้นทำเหมืองข้อมูล รู้ว่าความรู้ที่มีอยู่สามารถแก้ปัญหาได้ครอบคลุมหรือยัง หรือยังมีปัญหาอื่นที่ยังไม่สามารถแก้ไขได้ มีการกำหนดความหมายของคำศัพท์ต่างๆ เพื่อให้เกิดความเข้าใจตรงกันระหว่างปัญหากับการทำเหมืองข้อมูล

4. ตัดสินใจเป้าหมายของการทำเหมืองข้อมูล (Determine data mining goles)

เป็นการกำหนดเป้าหมายเพื่อแปลงการแก้ไขปัญหาเป็นวิธีการของเหมืองข้อมูล เช่น ต้องการทำยอดขายเพิ่มขึ้น จะต้องสร้างเหมืองข้อมูลที่บอกคุณลักษณะหรือพฤติกรรมของลูกค้าได้ โดยใช้เทคนิคของเหมืองข้อมูลที่เหมาะสม

5. สร้างแผนการทำงาน (Produce a project plan)

การวางแผนและกำหนดขั้นตอนการทำงานของเหมืองข้อมูล ในแต่ละขั้นตอนการทำงาน จะมีการวางแผนวิธีและเทคนิคที่จะเตรียมนำไปใช้

2.1.2.2 ความเข้าใจในข้อมูล (Data Understanding)

ขั้นตอนนี้เป็นการเริ่มต้นการรวบรวมข้อมูลและทำความเข้าใจกับข้อมูลที่จะนำไปใช้ในการทำงานของเหมืองข้อมูล กำหนดคุณภาพของข้อมูลที่จะใช้ และการเลือกใช้ข้อมูลที่เหมาะสมในการทำ โดยมีรายละเอียดดังต่อไปนี้

1. รวบรวมข้อมูล (Correct initial data)

เป็นขั้นตอนแรกของการนำข้อมูลมาใช้และเป็นขั้นตอนที่จำเป็นในการจัดเตรียมข้อมูลในกระบวนการต่อไป ซึ่งจะได้ข้อมูลและวิธีการใช้ข้อมูลทั้งหมดในการทำเหมืองข้อมูล

2. การอธิบายข้อมูล (Data description)

เมื่อได้ข้อมูลมาแล้วจำเป็นจะต้องมีการอธิบายของข้อมูล เช่น จำนวนข้อมูลและจำนวนแอททริบิวท์ที่ใช้ การระบุแอททริบิวท์และชนิดของข้อมูล

3. สำรวจข้อมูล (Explore data)

เป็นการหาโครงสร้างรูปแบบของข้อมูล เพื่อที่จะพิจารณานำไปใช้ในขั้นตอนของการสร้างแบบจำลอง ซึ่งในแต่ละเทคนิควิธีของแบบจำลองจะใช้ข้อมูลที่ไม่เหมือนกัน เช่น แอททริบิวท์ที่เป็นตัวเลข (numeric) และแอททริบิวท์ที่ไม่ใช่ตัวเลข (nominal) เป็นต้น

4. การตรวจสอบคุณภาพของข้อมูล (Verification data quality)

เป็นการตรวจสอบความถูกต้องของข้อมูล ดูว่าข้อมูลที่นำมาใช้เป็นชนิดและแอททริบิวท์ที่ถูกต้องหรือไม่ หรือตรวจสอบว่าข้อมูลที่จะใช้พร้อมใช้งานหรือยังซึ่งจะต้องไม่มีการสูญหายของข้อมูลหรือข้อมูลไม่ถูกต้องตามวัตถุประสงค์ ซึ่งการตรวจสอบคุณภาพของข้อมูลจะมีผลต่อประสิทธิภาพการทำงานของแบบจำลองของเหมืองข้อมูล

2.1.2.3 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนนี้เป็นการจัดเตรียมข้อมูลในขั้นสุดท้ายเพื่อให้พร้อมใช้งานแก่แบบจำลองต่างๆ โดยจะทำการแปลงข้อมูลหรือการทำความสะอาดข้อมูลเพื่อให้มีความเหมาะสมกับแบบจำลองนั้นๆ โดยมีรายละเอียดดังต่อไปนี้

1. เลือกข้อมูล (Select data)

เป็นการคำนึงถึงการเลือกข้อมูลที่มีความสมบูรณ์และถูกต้อง การกำหนดขนาดของข้อมูลและชนิดของข้อมูลเพื่อนำไปใช้กับแบบจำลอง

2. การทำความสะอาดข้อมูล (Data cleaning)

เป็นขั้นตอนการทำความสะอาดข้อมูลเพื่อให้ข้อมูลมีความสมบูรณ์ถูกต้องแก่การนำไปใช้ของแบบจำลอง เช่น การลดค่าข้อมูล (data normalization) เพื่อให้ข้อมูลมีค่าอยู่ในช่วง (0,1) การลดจำนวนข้อมูล (data reduction) ทำให้ข้อมูลมีขนาดที่เล็กลงเพื่อความรวดเร็วในการทำงาน การแก้ไขข้อมูลที่ผิดพลาดและข้อมูลรบกวน (treatment of missing value and noisy data) ทำให้ข้อมูลที่ไม่มีค่าหรือข้อมูลที่ผิดพลาดด้วยการแทนที่ด้วยจำนวนค่าหนึ่งเพื่อไม่ให้มีผลกระทบต่อการทำงาน และกำจัดข้อมูลที่ไม่เกี่ยวข้องต่อวัตถุประสงค์ออกไป

3. การจัดรูปแบบข้อมูล (Data formating)

หลังจากการเตรียมข้อมูลแล้ว ขั้นตอนนี้จะเป็นการจัดรูปแบบของข้อมูลเพื่อนำเข้าสู่แบบจำลอง ได้แก่ การจัดเรียงแอททริบิวท์ การกำจัดสัญลักษณ์พิเศษออกไปเช่น จุลภาค ช่องว่าง การตัดข้อความให้มีความพอดีกับขนาดของตัวอักษรที่กำหนด การแทนที่ข้อมูลที่ไม่ใช่ตัวเลขด้วยจำนวนค่าหนึ่ง เป็นต้น

2.1.2.4 การสร้างแบบจำลอง (Modeling)

ในขั้นตอนนี้เป็นการเลือกเทคนิคของแบบจำลองและการนำไปใช้ การกำหนดค่าพารามิเตอร์ต่างๆ ให้แก่แบบจำลอง ซึ่งแต่ละปัญหาอาจมีเทคนิคการใช้แบบจำลองที่หลากหลายจึงต้องมีการกลับไปทำขั้นตอนการเตรียมข้อมูลในบางครั้งเพื่อให้ได้ข้อมูลที่เหมาะกับแบบจำลอง ในขั้นตอนนี้มีรายละเอียดดังต่อไปนี้

1. การเลือกเทคนิคแบบจำลอง (Selection of modeling technique)

เป็นการเลือกชนิดของเทคนิคที่จะใช้ ซึ่งจะขึ้นอยู่กับปัญหาที่เราสนใจ เช่น ใช้เพื่อทำนายหรือเพื่อวิเคราะห์ข้อมูล ยกตัวอย่างเช่น การจัดกลุ่ม (classification) จะใช้ dicision trees หรือ โครงข่ายประสาทเทียม การแบ่งกลุ่ม (clustering) จะใช้วิธีการแบ่งกลุ่มแบบต่างๆ และ โครงข่ายประสาทเทียม เป็นต้น

2. ออกแบบการทดสอบ (Generate test design)

ก่อนที่จะสร้างแบบจำลองเราจะต้องมีการออกแบบการทดสอบก่อน โดยจะต้องมีการทดสอบดูค่าผิดพลาด ปรับค่าพารามิเตอร์ต่างๆ และประเมินคุณภาพของแบบจำลองก่อนที่จะนำไปใช้สร้างจริง

3. สร้างแบบจำลอง (Building model)

การสร้างแบบจำลองสามารถสร้างออกมาได้หลายแบบขึ้นอยู่กับค่าพารามิเตอร์ต่างๆ ที่ได้ปรับแต่งจนเป็นที่พอใจแล้ว และการสร้างแบบจำลองควรมีกระบวนการทำหลายๆ ครั้งเพื่อให้ได้ผลลัพธ์ที่พอใจที่สุด

4. การประเมินแบบจำลอง (Model assessment)

เป็นการทดสอบและประเมินแบบจำลองที่ถูกสร้างขึ้นมา ดูว่าสามารถแก้ปัญหาได้ครอบคลุมวัตถุประสงค์ที่ต้องการหรือไม่ ความถูกต้องและความรวดเร็วในการทำงานเป็นอย่างไร เพื่อตัดสินใจในการนำแบบจำลองไปใช้

2.1.2.5 การประเมินแบบจำลอง (Evaluation)

ขั้นตอนนี้จะประเมินความสามารถในการทำงานของแบบจำลองก่อนจะไปสู่ขั้นตอนการนำไปใช้ ตรวจสอบทุกขั้นตอนการทำงานของแบบจำลองเพื่อให้มั่นใจว่าจะไม่เกิดความผิดพลาด และสามารถแก้ปัญหาได้ถูกต้องตรงตามวัตถุประสงค์ที่กำหนดไว้ ในขั้นตอนนี้มีรายละเอียดดังนี้

1. การประเมินผล (Evaluate results)

เป็นการประเมินว่าแบบจำลองสามารถแก้ปัญหาได้มากน้อยอย่างไร พิจารณาการสร้างแบบจำลองมีสิ่งที่ผิดพลาดอย่างไร ควรนำแบบจำลองที่ได้ไปทดสอบกับปัญหาจริงที่เกิดขึ้น เพื่อดูผลลัพธ์ที่ได้และหาแนวทางในการพัฒนาการทำแบบจำลองต่อไปในอนาคต

2. ตรวจสอบกระบวนการ (Review process)

ขั้นตอนนี้จะเป็นการตรวจสอบการทำงานของแบบจำลอง เพื่อหาสิ่งผิดพลาดที่เกิดขึ้นซึ่งอาจมองข้ามไป

3. ตัดสินใจทำขั้นต่อไป (Determine next steps)

เป็นการตัดสินใจในการทำในขั้นตอนต่อไปว่าแบบจำลองที่ได้นั้นมีความพร้อมที่จะนำไปใช้หรือไม่ ถ้าพิจารณาแล้วว่ายังไม่สามารถนำไปใช้ได้ อาจจะต้องกลับไปทำยังขั้นตอนเริ่มต้นใหม่

2.1.2.6 การนำไปใช้ (Deployment)

เป็นขั้นตอนสุดท้ายในกระบวนการทำเหมืองข้อมูล สรุปผลที่ได้รับจากการใช้เหมืองข้อมูลว่าตรงกับความต้องการหรือไม่ และสร้างรายงานเพื่อเป็นประโยชน์ต่อนักวิเคราะห์ข้อมูลที่จะนำไปใช้ต่อไป ในขั้นตอนนี้มีรายละเอียดดังต่อไปนี้

1. วางแผนนำไปใช้ (Planing deployment)

เป็นการวางแผนกลยุทธ์การนำเหมืองข้อมูลไปใช้ให้ได้ตรงกับวัตถุประสงค์

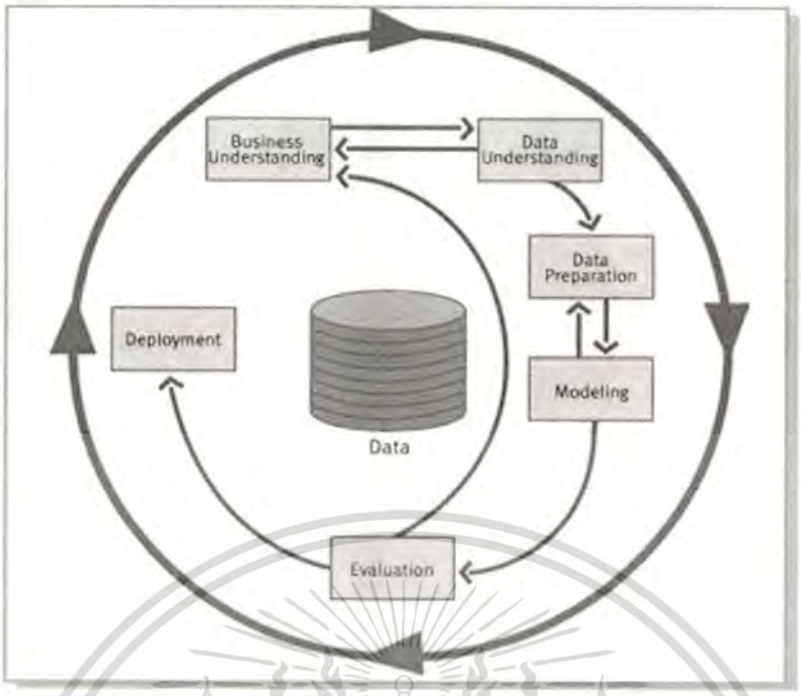
2. ตรวจและปรับปรุง (Monitoring and maintenance)

ในการนำเหมืองข้อมูลไปใช้งาน จะต้องมีการตรวจสอบการทำงานในแต่ละวันว่ามีข้อผิดพลาดอย่างไร ซึ่งถ้ามีข้อผิเพลาดเกิดขึ้นจะต้องนำกลับมาแก้ไขแบบจำลองที่ใช้ใหม่

3. รายงานผล (Final report)

ขั้นตอนสุดท้ายของกระบวนการ เป็นการสรุปงานทั้งหมดที่ได้ดำเนินการไปเพื่อเพิ่มประสบการณ์และความรู้ในการนำเหมืองข้อมูลไปใช้ในแต่ละปัญหา รู้ว่าส่วนใดในแต่ละปัญหามีความสำคัญอย่างไร

เนื่องจากกระบวนการของเหมืองข้อมูลซึ่งได้กล่าวไปข้างต้น บางครั้งการทำเพียงรอบเดียวอาจจะยังไม่ได้ให้ข้อมูลที่เหมาะสมและตรงตามความต้องการนัก ดังนั้น กระบวนการจึงต้องมีการย้อนกลับไปทำใหม่เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องและนำไปใช้ในการตัดสินใจให้ได้มากที่สุด ในรูปที่ 2.5 แสดงกระบวนการเหมืองข้อมูลตามวิธีการของ CRISP-DM v.1.0



รูปที่ 2.5 กระบวนการของเหมืองข้อมูล

2.2 เว็บเซอร์วิส (Web Services)

โดยทั่วไปเทคโนโลยีการสื่อสารบนเว็บสามารถแบ่งออกได้เป็น 2 ประเภทตามการพัฒนาเว็บคือ static web และ dynamic web ซึ่งการพัฒนาเว็บแบบ static web นั้นจะเป็นการเขียนเว็บสำหรับใช้การแสดงผลหน้าเว็บธรรมดา โดยใช้ HTML เป็นภาษาหลัก และอาจจะมีลูกเล่นหรือการแสดงผลที่น่าสนใจเพิ่มขึ้นโดยการเขียนสคริปต์เช่น Java Script VBScript หรือ Java Applet เป็นต้น ซึ่งการเขียนสคริปต์เหล่านี้เรียกว่า client-side script โดยจะประมวลผลทางฝั่งไคลเอ็นต์อย่างเดียว ส่วนการพัฒนาเว็บแบบ dynamic web นั้น เป็นการพัฒนามาจาก static web โดยจะให้การประมวลผลที่ฝั่งเซิร์ฟเวอร์โดยใช้ภาษาในการสร้างสคริปต์ เรียกว่า server-side script โดยจะใช้ความสามารถของเซิร์ฟเวอร์ในการจัดการข้อมูล หรือติดต่อกับฐานข้อมูล และจะส่งข้อมูลกลับมาให้ฝั่งไคลเอ็นต์ต่อไป โดยภาษาที่ใช้ได้แก่ CGI PERL ASP.NET PHP JSP เป็นต้น

ในการติดต่อสื่อสารผ่านเว็บในทางธุรกิจนั้นบางครั้งในแต่ละระบบอาจมีความแตกต่างกันขึ้นอยู่กับแอพพลิเคชันที่ใช้ ในความต้องการที่แท้จริงในทางธุรกิจนั้นคือทำอย่างไรเพื่อให้การติดต่อกันระหว่างข้อมูลและการเรียกใช้งานระหว่างองค์กรนั้นจะไม่ยึดติดกับแอพพลิเคชันหรือแพลตฟอร์มใดๆ เพื่อลดความซับซ้อนและค่าใช้จ่ายที่จะเกิดขึ้น จึงได้เกิดแนวคิดหา

มาตรฐานกลางขึ้นมาเพื่อทำหน้าที่ให้บริการการทำงานหนึ่งๆ บนเว็บ ซึ่งไม่ว่าที่ใดก็สามารถเรียกใช้งานได้ โดยมาตรฐานกลางที่เกิดขึ้นนี้เรียกว่าเว็บเซอร์วิส

2.2.1 สถาปัตยกรรมของเว็บเซอร์วิส (Web Services Architecture)

ในความหมายของเว็บเซอร์วิสนั้น เว็บเซอร์วิส คือ ระบบซอฟต์แวร์ที่ถูกออกแบบมาเพื่อสนับสนุนการทำงานร่วมกันระหว่างคอมพิวเตอร์บนเครื่อข่าย โดยมีการติดต่อประสานงานกันของคอมพิวเตอร์ในรูปแบบภาษาที่ใช้อธิบายและเรียกใช้การทำงานของเว็บเซอร์วิส (WSDL : Web Services Description Language) และระบบที่ใช้ติดต่อกับเว็บเซอร์วิสจะถูกอธิบายเป็นแบบข้อความ SOAP ด้วยภาษากลางที่ใช้คือ XML และข้อความจะถูกส่งไปบน HTTP โดยเว็บเซอร์วิสจะใช้สถาปัตยกรรมการให้บริการที่เรียกว่า Service Oriented Architecture (SOA) ซึ่ง SOA จะมีส่วนประกอบหลัก 3 ส่วนคือ ผู้ให้บริการ (Service Provider) ผู้ขอบริการ (Service Requester) และตัวแทนของผู้ให้บริการ (Service Broker) ซึ่งส่วนประกอบหลักทั้ง 3 ส่วนนี้จะติดต่อถึงกันโดยใช้ฟังก์ชันพื้นฐาน คือการประกาศ (publish) การค้นหา (find) และการเรียกใช้ (bind) โดยการทำงานร่วมกันของทั้ง 3 ฟังก์ชันนี้สามารถอธิบายได้คือ ผู้ให้บริการจะประกาศการบริการ ไปยังตัวแทนของผู้ให้บริการหรือเรียกได้อีกอย่างหนึ่งว่าไดเรกทอรีของการบริการ ในขณะที่ผู้ขอรับบริการจะค้นหาบริการที่ต้องการจากตัวแทนของผู้ให้บริการ และเมื่อพบบริการแล้วจะจึงเรียกใช้บริการจากผู้ให้บริการนั้น โดยสถาปัตยกรรมของ SOA จะแสดงได้ดังรูปที่ 2.6

รูปที่ 2.6 สถาปัตยกรรมของ SOA

แนวคิดของ SOA ได้ถูกนำมาประยุกต์ใช้กับเว็บเซอร์วิส โดยมีเทคโนโลยีที่เกี่ยวข้องในการพัฒนาเว็บเซอร์วิสได้แก่ [7]

2.2.1.1 XML (The Extensible Markup Language)

XML เป็นภาษามาตรฐานที่ใช้ในการแลกเปลี่ยนข้อมูลบนอินเตอร์เน็ต ซึ่งเป็นภาษา Markup มีลักษณะ tags เปิด และ tags ปิด เพื่อบอกว่าข้อมูลใน tags นั้นหมายถึงอะไร ผู้ใช้สามารถออกแบบ tags ได้อย่างอิสระ แล้วส่งข้อมูลเอกสาร XML ไปยังแอพพลิเคชันใดๆ ที่สามารถใช้ได้กับเอกสาร XML นี้ โดยผู้ที่มีหน้าที่รับผิดชอบ และกำหนดมาตรฐานของ XML คือ W3C (World Wide Web Comsortium)

2.2.1.2 SOAP (Simple Object Access Protocal)

เป็น XML-based โปรโตคอล ที่ใช้ในการแลกเปลี่ยนข้อมูลในสภาพแวดล้อมแบบกระจายศูนย์ ซึ่งจะอยู่บนโปรโตคอล HTTP โดยจะกำหนด messaging protocol ระหว่างผู้ให้บริการกับผู้ขอใช้บริการ ซึ่งจะกำหนดรูปแบบการแลกเปลี่ยนข้อมูลตามวิธีการของผู้ให้บริการนั้น

รูปที่ 2.7 โครงสร้างข้อความ SOAP

แสดงตัวอย่างข้อความ SOAP ได้ดังนี้

SOAP request :

```
<SOAP-ENV:Envelope
 xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
 SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
 <SOAP-ENV:Body>
  <m:GetEndorsingBoarder xmlns:m="http://namespaces.snowboard-info.com">
   <manufacturer>K2</manufacturer>
   <model>Fatbob</model>
  </m:GetEndorsingBoarder>
 </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

SOAP response :

```
<SOAP-ENV:Envelope
 xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
 SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/">
 <SOAP-ENV:Body>
  <m:GetEndorsingBoarderResponse xmlns:m="http://namespaces.snowboard-info.com">
   <endorsingBoarder>Chris Englesmann</endorsingBoarder>
  </m:GetEndorsingBoarderResponse>
 </SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

2.2.1.3 WSDL (Web Services Description Language)

เป็นภาษาที่ใช้อธิบายคุณลักษณะการใช้บริการและวิธีการติดต่อของเว็บเซอร์วิส โดยใช้ภาษา XML โดยมี tags โครงสร้างของ WSDL อธิบายได้ดังนี้

<portType> เป็นส่วนที่อธิบายโอเปอเรชันที่เว็บเซอร์วิสให้บริการ ในเว็บเซอร์วิสจะมีกี่โอเปอเรชันก็ได้

<operation> อธิบายรายละเอียดที่เกี่ยวกับโอเปอเรชัน

<message> อธิบายเกี่ยวกับข้อมูลที่ส่งเข้าและออกจากโอเปอเรชัน

<types> อธิบายข้อมูลที่เว็บเซอร์วิสใช้

<binding> อธิบายรายละเอียดของข้อมูลและโปรโตคอลที่ใช้ในแต่ละพอร์ตของเว็บเซอร์วิส

<service> อธิบายบริการของเว็บเซอร์วิส ชื่อการให้บริการ ในเว็บเซอร์วิสจะมีหลายบริการได้แต่ห้ามตั้งชื่อบริการซ้ำกัน

แสดงตัวอย่างข้อมูลของ WSDL ได้ดังนี้

```
<?xml version="1.0"?>
<definitions name="StockQuote"

targetNamespace="http://example.com/stockquote/service"
        xmlns:tns="http://example.com/stockquote/service"
        xmlns:soap="http://schemas.xmlsoap.org/wsdl/soap/"
        xmlns:defs="http://example.com/stockquote/definitions"
        xmlns="http://schemas.xmlsoap.org/wsdl/">

  <import namespace="http://example.com/stockquote/definitions"
         location="http://example.com/stockquote/stockquote.wsdl"/>

   <binding name="StockQuoteSoapBinding" type="defs:StockQuotePortType">
     <soap:binding style="document" transport="http://schemas.xmlsoap.org/soap/http"/>
     <operation name="GetLastTradePrice">
       <soap:operation soapAction="http://example.com/GetLastTradePrice"/>
       <input>
         <soap:body use="literal"/>
       </input>
       <output>
         <soap:body use="literal"/>
       </output>
     </operation>
   </binding>

   <service name="StockQuoteService">
     <documentation>My first service</documentation>
     <port name="StockQuotePort" binding="tns:StockQuoteBinding">
       <soap:address location="http://example.com/stockquote"/>
     </port>
   </service>
</definitions>
```

2.2.1.4 UDDI (Universal Description, Discovery and Integration)

เป็นมาตรฐานที่ใช้สำหรับค้นหาเว็บเซอร์วิส โดยจะเก็บรวบรวมเว็บเซอร์วิสต่างๆ ไว้ในแหล่งเดียวกันสำหรับให้ผู้ใช้บริการสามารถค้นหาได้ง่ายเปรียบได้กับสมุดหน้าเหลืองโทรศัพท์

รูปที่ 2.8 โครงสร้างของเว็บเซอร์วิส

บทที่ 3
การแบ่งกลุ่ม
และการเรียกใช้บริการผ่านเว็บเซอร์วิส

ในบทที่ 3 นี้จะกล่าวถึงเทคนิคเหมืองข้อมูลโดยใช้วิธีการแบ่งกลุ่มข้อมูล (clustering) โดยในหัวข้อต่อไปจะกล่าวถึงประเภทของการแบ่งกลุ่มและอัลกอริทึมการแบ่งกลุ่มที่ใช้ในโครงงานนี้ รวมถึงการใช้งานผ่านเว็บเซอร์วิส และขั้นตอนการสร้างและการประกาศการใช้บริการของเว็บเซอร์วิส

3.1 การแบ่งกลุ่ม (Clustering)

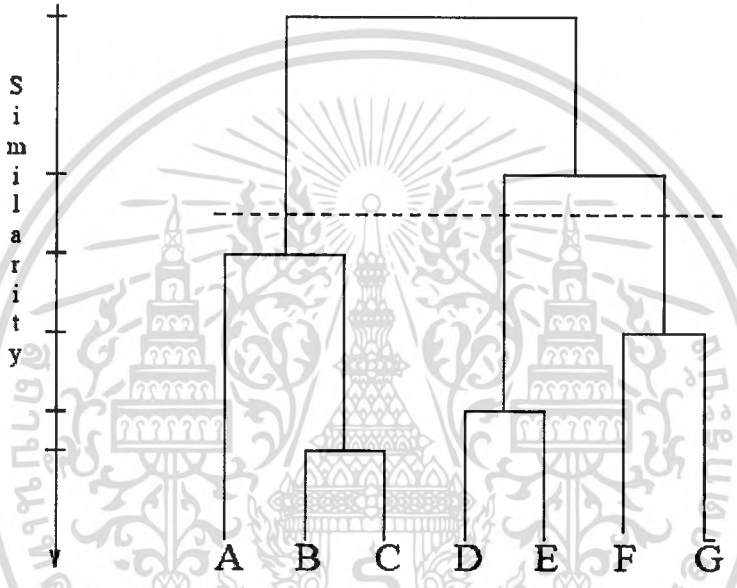
การแบ่งกลุ่ม คือ การแบ่งกลุ่มข้อมูลที่มีลักษณะที่เหมือนกันหรือคล้ายกันมาอยู่รวมในกลุ่มเดียวกัน (cluster) โดยจะใช้วิธีการวิเคราะห์ค้นหารูปแบบข้อมูลที่เราไม่สามารถรู้ในรายละเอียดและต้องการรู้ว่ามีข้อมูลอะไรบ้างในแต่ละกลุ่ม การแบ่งกลุ่มจัดเป็นการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) ซึ่งเป็นวิธีการเรียนรู้ของคอมพิวเตอร์ที่ไม่มีการกำหนดผลลัพธ์ไว้ล่วงหน้า โดยเทคนิคการแบ่งกลุ่มนี้ได้ถูกใช้อย่างกว้างขวาง ยกตัวอย่างเช่น นำไปใช้กับข้อมูลทางสถิติ การทำเหมืองข้อมูล การวิเคราะห์ข้อมูลลูกค้าและข้อมูลการตลาด การวิเคราะข้อมูลทางการแพทย์ และการจดจำรูปแบบ เป็นต้น ในโครงงานนี้กล่าวถึงการแบ่งกลุ่มข้อมูลในการทำเหมืองข้อมูลเพื่อใช้ในการแบ่งกลุ่มข้อมูลที่มีขนาดใหญ่ มีความหลากหลายของแอททริบิวต์และชนิดข้อมูลที่แตกต่างกัน

การแบ่งกลุ่มสามารถจัดประเภทของการแบ่งออกเป็นประเภทใหญ่ๆ ได้ 3 ประเภท คือ การแบ่งกลุ่มตามลำดับชั้น (Hierarchical Clustering) การแบ่งกลุ่มแบบพาร์ทิชัน (Partitioning Clustering) และการแบ่งกลุ่มโดยใช้แบบจำลอง (Model-Based Clustering) โดยในแต่ละประเภทสามารถอธิบายได้ดังต่อไปนี้ [3]

3.1.1 การแบ่งกลุ่มแบบลำดับชั้น (Hierarchical Clustering)

การแบ่งกลุ่มประเภทนี้เป็นการแบ่งกลุ่มโดยใช้วิธีการเปรียบเทียบข้อมูลที่อยู่ใกล้กัน มาจัดเป็นกลุ่มเดียวกันในลักษณะเป็นแบบลำดับชั้นทำต่อไปเรื่อยๆ โดยวิธีการสร้างลำดับชั้นจะใช้ในลักษณะต้นไม้ หรือเรียกว่าเดนโดแกรม (dendogram) ซึ่งทุกๆ โหนดของคลัสเตอร์จะประกอบไปด้วยคลัสเตอร์ลูกและคลัสเตอร์ที่เกี่ยวข้องกันก็จะมีคลัสเตอร์โหนดพ่อเหมือนกัน เป็นไปในลักษณะเช่นนี้ของทุกลำดับชั้น การแบ่งกลุ่มตามลำดับชั้นแบ่งการทำงานออกได้ 2

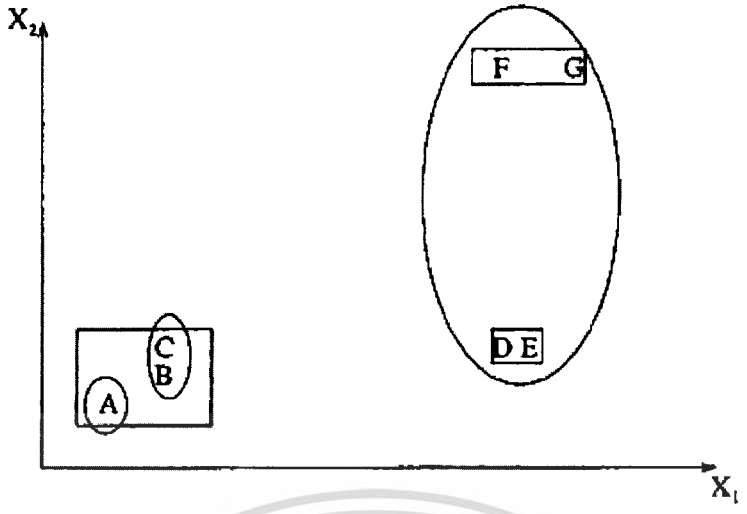
แบบด้วยกันคือ การทำงานแบบล่างขึ้นบนเรียกว่า agglomerative และการทำงานแบบบนลงล่างเรียกว่า divisive โดยการแบ่งกลุ่มตามลำดับชั้นแบบ agglomerasive จะเริ่มต้นด้วยคลัสเตอร์เพียง 1 คลัสเตอร์ และจะรวมคลัสเตอร์ที่เหมาะสมรวมเป็นคลัสเตอร์ใหญ่ต่อไปเรื่อยๆ ส่วนการแบ่งกลุ่มตามลำดับชั้นแบบ division นั้นจะเริ่มจากคลัสเตอร์ 1 คลัสเตอร์ของข้อมูลทั้งหมดรวมกัน และจะแตกกลุ่มออกไปเป็นแต่ละคลัสเตอร์ไปเรื่อยๆ ตามลำดับชั้น กระบวนการการทำงานจะดำเนินไปจนถึงจุดหนึ่งที่ได้กำหนดไว้ การแบ่งกลุ่มประเภทนี้ไม่เหมาะกับการแบ่งกลุ่มข้อมูลที่มีจำนวนมาก เพราะการทำงานจะใช้เวลามาก



รูปที่ 3.1 การแบ่งกลุ่มตามลำดับชั้น

3.1.2 การแบ่งกลุ่มแบบพาร์ทิชัน (Partitioning Clustering)

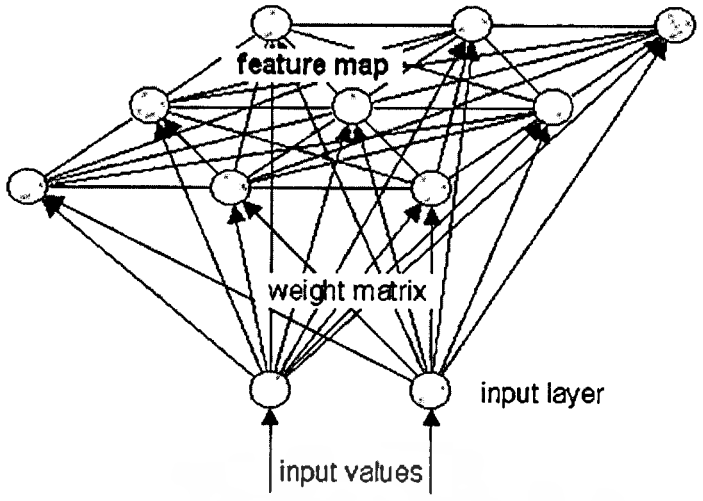
การแบ่งกลุ่มประเภทนี้จะใช้การกำหนดกลุ่มหรือคลัสเตอร์เป็นจำนวน k คลัสเตอร์ ของจำนวนข้อมูลทั้งหมดจำนวน n ตัว โดยที่ $k \leq n$ โดยในแต่ละพาร์ทิชันจะดูว่าข้อมูลที่มีระยะห่างที่ใกล้กันจะเป็นข้อมูลที่อยู่ในกลุ่มเดียวกันและเป็นข้อมูลที่มีลักษณะคล้ายกัน โดยในแต่ละคลัสเตอร์จะมีการวัดระยะห่างระหว่างจุดศูนย์กลางของคลัสเตอร์กับตำแหน่งของข้อมูล วิธีนี้จะเหมาะสำหรับข้อมูลที่มีขนาดใหญ่ และจะมีความรวดเร็วในการทำงาน โดยอัลกอริทึมที่สำคัญของการแบ่งกลุ่มประเภทนี้ได้แก่ k-means และ k-medoids เป็นต้น



รูปที่ 3.2 การแบ่งกลุ่มแบบพาร์ทิชัน

3.1.3 การแบ่งกลุ่มโดยใช้แบบจำลอง (Model-Based Clustering)

การแบ่งกลุ่มประเภทนี้จะใช้แบบจำลองทางคณิตศาสตร์มาใช้ในการคำนวณเพื่อหากลุ่มของข้อมูลที่เหมาะสม ตัวอย่างแบบจำลองที่นำมาใช้แบ่งกลุ่มเช่น โครงข่ายประสาทเทียม (Artificial Neural Networks) โดยมีวิธีการทำงานที่โดดเด่นอยู่ 2 วิธีคือ การเรียนรู้แบบการแข่งขัน (competitive learning) และ วิธี การเรียนรู้โดยใช้ลักษณะสำคัญ (self-organizing feature map) โดยดูผลลัพท์ที่ได้ว่าในแต่ละโหนดของเอาท์พุทของโครงข่ายโหนดไหนที่ใกล้กันจะเป็นข้อมูลที่อยู่กลุ่มเดียวกัน โครงข่ายประสาทเทียมที่นิยมใช้ เช่น Kohonen neural networks เป็นต้น ซึ่งความสำคัญของการใช้โครงข่ายประสาทเทียมในการแบ่งกลุ่ม คือ มีกระบวนการคำนวณเวคเตอร์ที่รับเข้ามาและแสดงผลลัพธ์ในรูปแบบหรือขนาดข้อมูลที่เหมาะสม เช่นการปรับปรุงค่าถ่วงน้ำหนักเพื่อให้การเรียนรู้ดีขึ้น และเป็นสถาปัตยกรรมที่มีการคำนวณแบบขนานและแบบกระจายได้ ซึ่งการนำไปใช้กับข้อมูลที่มีขนาดใหญ่มากๆ จะใช้เวลาในการเรียนรู้ค่อนข้างนานและมีความซับซ้อนทางการคำนวณสูง



รูปที่ 3.3 Kohonen neural networks

3.2 อัลกอริทึมของการแบ่งกลุ่ม (Clustering Algorithm)

ในโครงงานนี้จะใช้อัลกอริทึมการแบ่งกลุ่ม 2 วิธี คือ K-Means และ ISODATA โดยมีรายละเอียดดังต่อไปนี้

3.2.1 K-Means

เป็นอัลกอริทึมที่พิจารณาการแบ่งข้อมูลออกเป็นแต่ละคลัสเตอร์ (k) โดยหาระยะทางของข้อมูลที่มีระยะห่างจากจุดศูนย์กลางของคลัสเตอร์นั้น (centroid) มายังตำแหน่งข้อมูลนั้นเป็นระยะทางที่น้อยที่สุด โดยใช้ฟังก์ชันการคำนวณ เช่น Euclidean distance method ซึ่งจะมองว่าเป็นข้อมูลที่อยู่ในกลุ่มเดียวกันและมีลักษณะที่คล้ายกัน ในการทำงานของการแบ่งกลุ่มประเภทนี้จะคำนวณหาจุดศูนย์กลางของคลัสเตอร์และจะทำซ้ำไปเรื่อยๆ จนกระทั่งไม่สามารถเปลี่ยนแปลงจุดศูนย์กลางได้อีกจึงถือว่าการแบ่งกลุ่มเสร็จสิ้น โดยอัลกอริทึมของ K-Means แสดงได้ดังต่อไปนี้ [6]

ขั้นตอนที่ 1. กำหนดจำนวนกลุ่มที่ต้องการแบ่งข้อมูล โดยให้ k เป็นเซ็ตของคลัสเตอร์ และ M แทนจุดศูนย์กลางของคลัสเตอร์ จะได้ $M_1^{(0)}, M_2^{(0)}, ..., M_k^{(0)}$ และกำหนดจำนวนรอบ $l = 0$

ขั้นตอนที่ 2. กำหนดให้แต่ละจำนวน $\{X_i,\ i = 1,\ \ldots,\ N\}$ อยู่ในแต่ละคลัสเตอร์และคำนวณระยะทางระหว่างข้อมูลกับศูนย์กลางของคลัสเตอร์ :

$$X \sim \omega_j \ \ if \ \ D_L(X, M_j^{(l)}) = min\ \{\ D_L(X, M_i^{(l)}),\ \ i = 1,\ \ldots,\ k\ \} \qquad (3.1)$$

โดยให้ ω_j แทนคลัสเตอร์ที่ i ซึ่งมีคลัสเตอร์ $M_j^{(l)}$ ที่รอบ l ;

ขั้นตอนที่ 3. ปรับปรุงจุดศูนย์กลางโดยให้ $M_j^{(l+1)}$

$$M_j^{(l+1)} = \frac{1}{N_j} \sum_{X \sim \omega_j} X, \qquad (j = 1,\ \ldots,\ k) \qquad (3.2)$$

เมื่อ $N_j^{(l)}$ เป็นจำนวนข้อมูลของ $\omega_j^{(l)}$ ในรอบที่ l , และ

$$\sum_{j=1}^{k} N_j^{(l)} = N \qquad (3.3)$$

คำนวณผลรวมระยะทางทั้งหมดจากทุกจุด ω_j เพื่อหาระยะทางที่น้อยที่สุด

$$\sum_{X \sim \omega_j^{(l)}} D_L(X, M_j^{l+1}) \rightarrow min. \quad (j = 1,\ \ldots, k) \qquad (3.4)$$

ขั้นตอนที่ 4. สิ้นสุดการทำงานถ้าไม่มีการเปลี่ยนแปลงของจุดศูนย์กลาง

$$M_j^{(l+1)} = M_j^{(l)} \quad (j=1,\ \ldots,\ k) \qquad (3.5)$$

หรือสิ้นสุดจำนวนรอบการทำงานตามที่กำหนด
ถ้าไม่ใช่ ให้ $l \leftarrow l + 1$, กลับไปขั้นตอนที่ 2.;

3.2.2 ISODATA

เป็นอัลกอริทึมที่มีการทำงานคล้ายกับ K-Means แต่ต่างกันที่ ISODATA สามารถรวมกลุ่มคลัสเตอร์ได้หรือแตกคลัสเตอร์ออกมาได้อีก ซึ่งจะแบ่งกลุ่มได้ถูกต้องและมีความเหมาะสมมากกว่า K-Means ซึ่งเป็นวิธีการแบ่งกลุ่มอย่างง่าย หลักการการทำงานของอัลกอริทึมนี้คือจะพิจารณาว่าคลัสเตอร์ที่กำหนดมานั้นจะสมควรดำเนินการแตกคลัสเตอร์หรือรวมกันหรือไม่ โดยดูจากค่าเบี่ยงเบนมาตรฐาน (Standard Deviation) และค่าเธรชโชลด์ต่างๆ ที่เกี่ยวข้อง เป็นตัวกำหนดการดำเนินการ โดยอัลกอริทึมของ ISODATA แสดงได้ดังนี้ [5]

K = จำนวนคลัสเตอร์ที่ต้องการ;

I = จำนวนรอบสูงสุด;

P = จำนวนคู่ของคลัสเตอร์สูงสุดที่จะสามารถรวมเข้ากันได้;

θ_N = ค่าเธรชโชลด์เป็นจำนวนน้อยที่สุดของข้อมูลที่คลัสเตอร์สามารถมีได้ (ใช้ในการพิจารณาทิ้งคลัสเตอร์);

θ_S = ค่าเธรชโชลด์ของค่าเบี่ยงเบนมาตรฐาน (สำหรับการพิจารณาการแยกคลัสเตอร์);

θ_C = ค่าเธรชโชลด์ของระยะห่างของคลัสเตอร์แต่ละคู่ (สำหรับพิจารณาการรวมคลัสเตอร์);

อัลกอริทึม :

ขั้นตอนที่ 1. เริ่มจากการกำหนด k คลัสเตอร์ (ไม่เท่ากับจำนวน K) กำหนดให้จุดศูนย์กลางของแต่ละคลัสเตอร์เป็น : $M_1, M_2, ..., M_k$ จากชุดข้อมูล $\{X_i,\ i = 1, 2, ..., N\}$

ขั้นตอนที่ 2. ให้แต่ละข้อมูล N อยู่ในคลัสเตอร์ โดยการพิจารณาระยะห่างจากจุดศูนย์กลาง :

$$X \sim \omega_j \ \ if \ \ D_L(X, M_j) = max\ \{\ D_L(X, M_i),\ \ i = 1, 2, \ldots, N\} \qquad (3.6)$$

ขั้นตอนที่ 3. ทิ้งคลัสเตอร์ที่มีจำนวนข้อมูลน้อยกว่า θ_N, ตัวอย่างเช่น สำหรับทุกๆ $j, N_j < \theta_N$ แล้วทิ้ง ω_j และให้ $k \leftarrow k - 1$

ขั้นตอนที่ 4. ปรับปรุงจุดศูนย์กลางของคลัสเตอร์ใหม่ :

$$M_j = \frac{1}{N_j} \sum_{X \sim \omega_j} X \quad (j = 1, \ldots, k) \qquad (3.7)$$

ขั้นตอนที่ 5. คำนวณระยะทางเฉลี่ย D_j ของข้อมูลในแต่ละคลัสเตอร์ ω_j จากจุดศูนย์กลาง :

$$D_j = \frac{1}{N_j} \sum_{X \sim \omega_j} D_L(X, M_j) \quad (j = 1, \ldots, k) \qquad (3.8)$$

ขั้นตอนที่ 6. คำนวณหาระยะทางเฉลี่ยทั้งหมดระหว่างข้อมูลกับจุดศูนย์แต่ละคลัสเตอร์ :

$$D = \frac{1}{N}\sum_{j=1}^{k} N_j D_j \tag{3.9}$$

ขั้นตอนที่ 7. ถ้า $k \le K/2$ (คลัสเตอร์น้อยไป), ไปขั้นตอนที่ 8; หรือถ้า $k > 2K$ (คลัสเตอร์เยอะไป), ให้ไปขั้นตอนที่ 11; ถ้าไม่ใช่ ไปขั้นตอน 14.

(ขั้นตอนที่ 8 – 10 เป็นการแยกคลัสเตอร์, ขั้นตอนที่ 11 – 13 สำหรับการรวมคลัสเตอร์)

ขั้นตอนที่ 8. เริ่มต้นการแยกคลัสเตอร์ หาส่วนเบี่ยงเบนมาตรฐาน $\sum_j = \left[\sigma_1^{(j)},...,\sigma_n^{(j)}\right]^T$ สำหรับแต่ละคลัสเตอร์ :

$$\sigma_i^{(j)} = \sqrt{\frac{1}{N_j}\sum (x_i - m_i^{(j)})^2}, \quad (i = 1, \ldots, n,\ j = 1, \ldots, k) \tag{3.10}$$

โดยที่ $m_i^{(j)}$ คือคอมโพเน็นท์ที่ i ของ M_j และ σ_i คือส่วนเบี่ยงเบนมาตรฐานของข้อมูลใน ω_j , N_j เป็นจำนวนข้อมูลของ ω_j

ขั้นตอนที่ 9. หาค่าคอมโพเนนท์มากสุดของแต่ละ $\sum_j$ และแทนค่าด้วย $\sigma_{\max}^{(j)}$; ทำทั้งหมดสำหรับ $j = 1, \ldots, k$

ขั้นตอนที่ 10.

สำหรับทุกๆ $\sigma_{\max}^{(j)}$, $(j = 1, \ldots, k)$, ทุกๆ ข้อจะเป็นจริงก็ต่อเมื่อ

$\sigma_{\max}^{(j)} > \theta_S$,

$D_j > D$,

$N_j > 2\theta_N$

ดังนั้น แยก M_j เป็น 2 คลัสเตอร์ที่มีจุดศูนย์กลางเป็น M_j^+ และ M_j^- โดยการบวก $\pm\delta$ ในแต่ละคอมโพเนนท์ของ ที่ M_j สอดคล้องกับ $\sigma_{\max}^{(j)}$ โดย δ สามารถเป็น $\alpha\sigma_{\max}^{(j)}$ สำหรับ $\alpha > 0$
จากนั้นลบ M_j และให้ $k \leftarrow k+1$

จากนั้นไปยังขั้นตอนที่ 2 ถ้าไม่ใช่ ไปขั้นตอนที่ 14.

ขั้นตอนที่ 11. ขั้นแรกของการรวมคลัสเตอร์ คำนวณระยะห่างระหว่างจุดศูนย์กลางคลัสเตอร์ D_{ij} สำหรับทุกคู่คลัสเตอร์ :

$$D_{ij} = D_L(M_i, M_j), \quad (for\ all\ i \neq j) \tag{3.11}$$

และเรียงลำดับ $k(k-1)/2$ จากน้อยไปมาก

ขั้นตอนที่ 12. หาว่าไม่มีคู่จำนวนมากกว่า P คู่ที่ D_{ij} แต่ละคู่น้อยกว่า θ_C จากนั้นให้เรียงลำดับจากน้อยไปมาก:

$$D_{i_1 j_1} \leq D_{i_2 j_2} \leq \ldots \leq D_{i_p j_p}$$

ขั้นตอนที่ 13 รวมคู่ของคลัสเตอร์ : สำหรับ $l = 1, \ldots, P$, ทำขั้นตอนดังนี้ :
ถ้า M_{il} และ M_{jl} ไม่ถูกใช้ในการทำงานรอบนี้ ,
ดังนั้นให้รวมคลัสเตอร์ทั้งคู่เข้าด้วยกันจะได้จุดศูนย์กลางคลัสเตอร์ใหม่ :

$$M = \frac{1}{N_{i_l} + N_{j_l}} [N_{i_l} M_{i_l} + N_{j_l} M_{j_l}] \tag{3.12}$$

ลบ M_{il} และ M_{jl} , และให้ $k \leftarrow k+1$
กลับไปทำขั้นตอนที่ 2

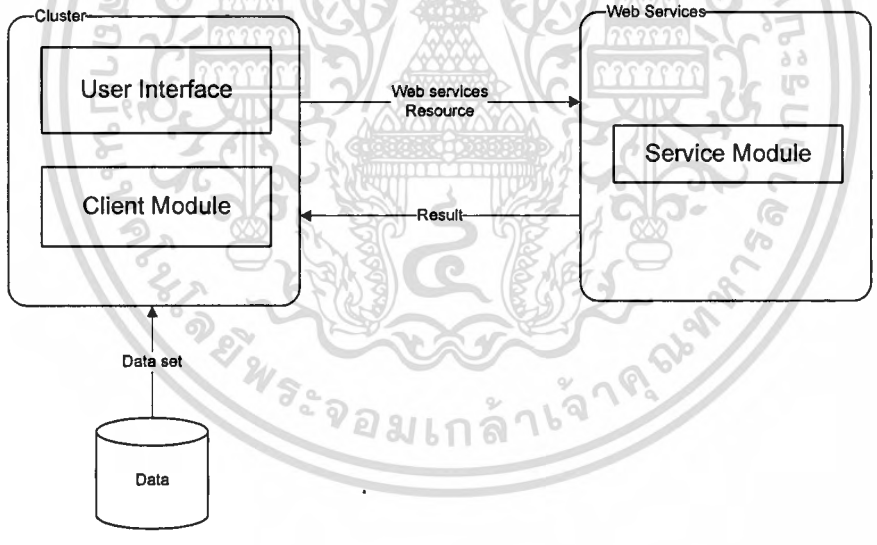
ขั้นตอนที่ 14. สิ้นสุดการดำเนินการถ้ามีจำนวนรอบเกิน l ถ้ายัง ทำขั้นตอนที่ 2 ต่อไป

3.3 การใช้บริการผ่านเว็บเซอร์วิส

ในหัวข้อนี้จะกล่าวถึงโครงสร้างการทำงานของการใช้งานเหมืองข้อมูลระหว่างไคลเอ็นท์และเว็บเซอร์วิส โดยจะแสดงขั้นตอนการทำงานในการเรียกใช้บริการ และการสร้างเว็บเซอร์วิสในการให้บริการเทคนิคของเหมืองข้อมูล

3.3.1 โครงสร้างการทำงาน

ในโครงงานนี้ได้แบ่งการพัฒนาออกเป็น 2 ส่วน คือ ส่วนแรกเป็นการพัฒนาไคลเอ็นท์ ซึ่งเป็นส่วนของแอพพลิเคชันที่ใช้งานเหมืองข้อมูล สามารถโหลดข้อมูลและจัดเตรียมข้อมูลก่อนเข้าสู่กระบวนการสร้างแบบจำลอง โดยไคลเอ็นท์สามารถสร้างแบบจำลองโดยการเรียกใช้บริการการสร้างแบบจำลองผ่านเว็บเซอร์วิส และในส่วนที่สองเป็นการพัฒนาเว็บเซอร์วิสซึ่งจะให้บริการการสร้างแบบจำลองและคำนวณผล ก่อนจะส่งข้อมูลผลลัพธ์กลับมาให้ไคลเอนท์ โดย รูปที่ 3.4 จะแสดงโครงสร้างของการทำงานร่วมกันระหว่างไคลเอ็นท์และเว็บเซอร์วิส



รูปที่ 3.4 โครงสร้างการทำงานของไคลเอ็นท์กับเว็บเซอร์วิส

3.3.2 โอเปอร์เรชันของเว็บเซอร์วิส

ในหัวข้อนี้ จะแสดงข้อมูลโอเปอร์เรชันของเว็บเซอร์วิสที่ให้บริการ โดยแสดงรายละเอียดของโอเปอร์เรชันและค่าพารามิเตอร์ที่รับเข้ามา โดยแสดงรายละเอียดได้ดังตารางที่ 3.1

ตารางที่ 3.1 รายละเอียดโอเปอร์เรชันของเว็บเซอร์วิส

Operation	Parameter	Description
KMeans	ModelOptions DataSet	ค่าสำหรับสร้างโมเดล ชุดข้อมูล
ISODATA	ModelOptions DataSet	ค่าสำหรับสร้างโมเดล ชุดข้อมูล

จากตารางที่ 3.1 ค่าพารามิเตอร์ของโอเปอร์เรชันของเหมืองข้อมูล คือ KMeans และ ISODATA จะมี 2 ประเภท คือ 1. ModelOption จะเป็นพารามิเตอร์ที่ใช้ในการสร้างแบบจำลองการแบ่งกลุ่ม เช่น จำนวนคลัสเตอร์ หรือ ค่าเธรชโชรด์ต่างๆ ที่ใช้ในอัลกอริทึม ISODATA และ 3. DataSet คือชุดของข้อมูลที่ส่งมาจากไคลเอนท์

3.3.3 ขั้นตอนการทำงาน

ในหัวข้อนี้จะอธิบายการทำงานระหว่างไคลเอ็นท์กับเว็บเซอร์วิส เมื่อไคลเอ็นท์ต้องการเรียกใช้บริการมายังเว็บเซอร์วิส โดยมีรายละเอียดขั้นตอนดังต่อไปนี้

1. ไคลเอ็นท์เลือกชุดข้อมูลในการทำเหมืองข้อมูล โดยอาจเลือกจากไฟล์หรือดิวรีจากฐานข้อมูล เมื่อได้มาแล้วจะเข้าสู่ขั้นตอนการเตรียมข้อมูล โดยการแก้ไขค่าที่ผิดพลาดและการทำการนอมอลไลซ์ข้อมูล
2. ไคลเอ็นท์เลือกการทำงานของเหมืองข้อมูล โดยเลือกที่การแบ่งกลุ่ม จากนั้นเลือกอัลกอริทึมที่จะใช้ และค่าพารามิเตอร์ต่างๆ ที่จำเป็นในการสร้างแบบจำลอง
3. ไคลเอ็นท์กำหนดแอดเดรสของเว็บเซอร์วิสเพื่อเรียกใช้บริการการทำเทคนิคการแบ่งข้อมูลของเหมืองข้อมูล
4. เมื่อตกลงการขอใช้บริการแล้วค่าพารามิเตอร์ต่างๆ จะถูกส่งไปยังเว็บเซอร์วิส เพื่อใช้ในการคำนวณ รวมถึงข้อมูลซึ่งแปลงเป็น XML
5. เว็บเซอร์วิสจะรับข้อมูล XML มาแปลงเป็นรูปแบบข้อมูลสำหรับการประมวลผล และเมื่อเว็บเซอร์วิสคำนวณผลเสร็จสิ้นแล้ว จะส่งพารามิเตอร์ที่แสดงผลลัพธ์กลับไปให้แก่ไคลเอ็นท์

3.3.4 การสร้างเว็บเซอร์วิสสำหรับเหมืองข้อมูล

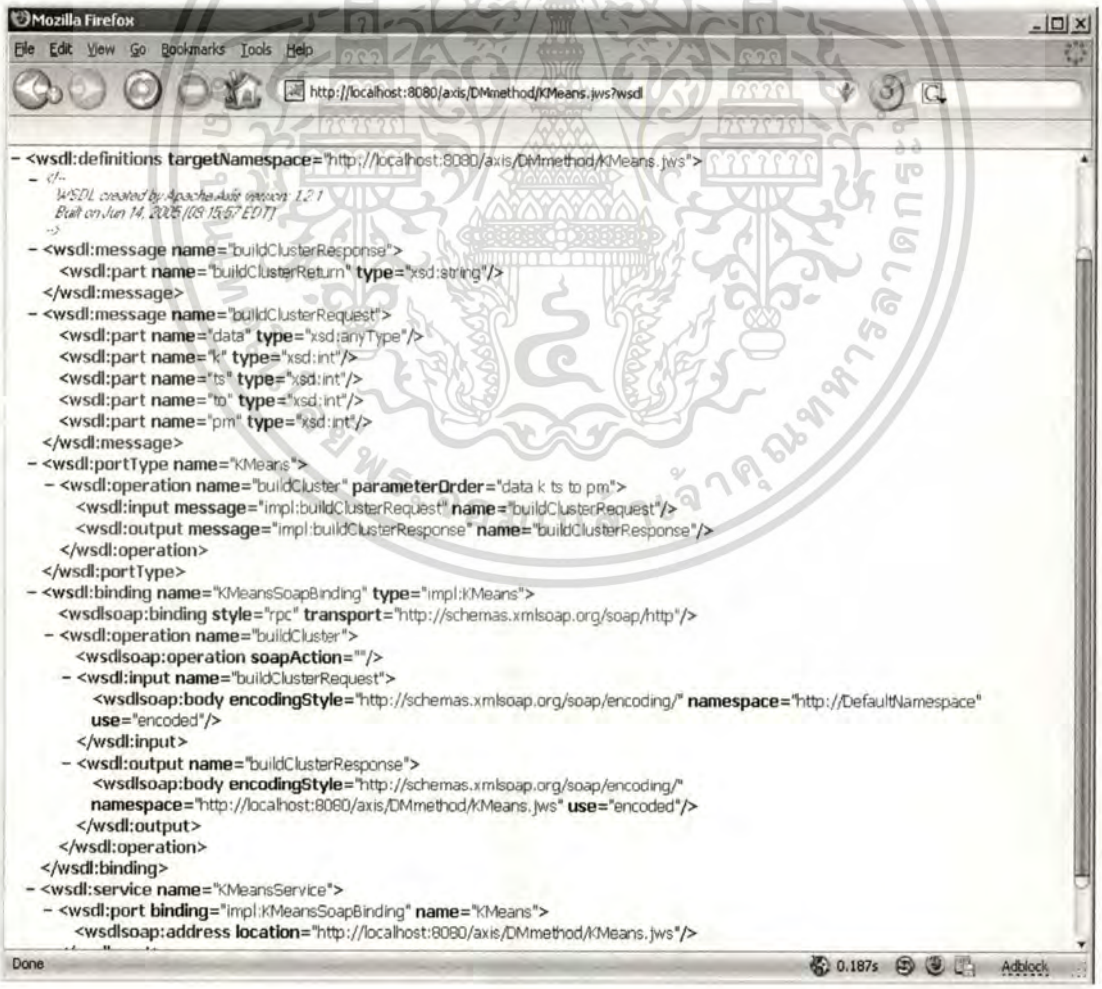
ในหัวข้อนี้จะแสดงการสร้างเว็บเซอร์วิสเพื่อให้บริการของเหมืองข้อมูล โดยเครื่องมือที่ใช้นั้นในโครงงานนี้ได้เลือก Apache Axis เป็นเครื่องมือสำหรับสร้างเว็บเซอร์วิส และจะใช้ Apache Tomcat เป็นเว็บเซิร์ฟเวอร์โดยมีขั้นตอนต่างๆ ดังนี้

3.3.4.1 การสร้างเว็บเซอร์วิส (Create web services)

ในการสร้างเว็บเซอร์วิสให้สามารถนำไปใช้ได้นั้น ในขั้นตอนแรกให้นำไฟล์ที่เป็นคลาสของอัลกอริทึมของการแบ่งกลุ่ม มาแปลงชื่อไฟล์ให้เป็นไฟล์ของ Java Web Service เช่น ไฟล์ชื่อ KMean.java เปลี่ยนเป็น KMeans.jws และเก็บไว้ในไดเรกทอรีสำหรับบริการ โดยจะเก็บไว้ที่

"%TOMCAT_HOME%/webapps/axis/DMmethod"

จากนั้นเปิด Apache Tomcat และเข้าไปที่ http://localhost:8080/axis/DMmethod/KMeans.jws Axis จะคอมไพล์ไฟล์ KMeans.jws ให้โดยอัตโนมัติ จะได้เอกสาร WSDL ของไฟล์ KMeans.jws โดยแสดงได้ดังรูปที่ 3.5



รูปที่ 3.5 เอกสาร WSDL ของเว็บเซอร์วิส

3.3.4.2 การเตรียมเว็บเซอร์วิส (Deploy web services)

การที่จะทำให้เว็บเซอร์วิสพร้อมให้บริการนั้น จะต้องมีการเตรียมเว็บเซอร์วิสก่อน เรียกว่า deploy web services โดยใช้ AdminClient ของ Axis (org.apache.axis.client.AdminClientr class) และใช้คำสั่ง java org.apache.axis.client.AdminClient deploy.wsdd จากนั้นจะได้การเปิดบริการเว็บเซอร์วิสดังรูปที่ 3.6

รูปที่ 3.6 การเปิดการบริการเว็บเซอร์วิส

บทที่ 4
การวิเคราะห์และออกแบบระบบ

ในบทนี้จะกล่าวถึงการออกแบบระบบและแสดงภาพรวมของการทำงานของระบบ ซึ่งจะแสดงความต้องการของระบบ และสภาพแวดล้อมของการทำงานของเว็บเซอร์วิสสำหรับเหมืองข้อมูล การออกแบบโปรแกรมส่วนไคลเอนท์โดยใช้ UML ซึ่งจะแสดงรายละเอียดของการวิเคราะห์และออกแบบระบบได้ดังต่อไปนี้

4.1 องค์ประกอบของระบบ

4.1.1 โปรแกรมส่วนไคลเอ็นท์

ในส่วนนี้จะเป็นโปรแกรมที่อยู่บนฝั่งไคลเอ็นท์ ซึ่งจะมีฟังก์ชันการทำงานตามกระบวนการของเหมืองข้อมูล ได้แก่ การเลือกข้อมูล การแก้ไขเปลี่ยนแปลงข้อมูล และการใช้บริการการแบ่งกลุ่มข้อมูลผ่านทางเว็บเซอร์วิส โดยฟังก์ชันต่างๆ สามารถอธิบายได้ดังต่อไปนี้

1. ฟังก์ชันการเลือกข้อมูล

การเลือกข้อมูลเป็นขั้นตอนแรกของการทำเหมืองข้อมูล ซึ่งในโปรแกรมฝั่งไคล์เอ็นท์นั้นจะสามารถเลือกข้อมูลได้ทางไฟล์และทางฐานข้อมูล โดยไฟล์ที่จะนำมาใช้นั้นจะเป็นไฟล์สกุล csv ซึ่งเป็นไฟล์ที่คั่นข้อมูลในแต่ละแอททริบิวท์ด้วยจุลภาค และการเลือกข้อมูลจากฐานข้อมูล ในโครงงานนี้ได้ใช้ฐานข้อมูล PostgreSQL 8.1 เป็น RDBMS ซึ่งสามารถใช้คำสั่งสืบค้นข้อมูลเรียกจากฐานข้อมูลมาแปลงเป็นข้อมูลที่จะใช้สำหรับการทำเหมืองข้อมูลต่อไป

2. ฟังก์ชันการแก้ไขข้อมูลที่ผิดพลาด

การแก้ไขข้อมูลที่ผิดพลาด (missing value) เป็นการแทนที่ค่าที่ว่างซึ่งเกิดจากการตรวจพบว่ามีค่าที่ผิดพลาด ขาดหาย หรือเป็นค่าที่ไม่เหมาะสมในการนำมาทำเหมืองข้อมูล ซึ่งจะลบข้อมูลดังกล่าวให้มีค่าว่าง แล้วใช้ฟังก์ชั้นนี้ในการแก้ไขให้เหมาะสม โดยอาจใช้ค่าความถี่มากสุดหรือค่าเฉลี่ยตามชนิดของข้อมูลซึ่งจะทำให้ค่าผลรวมของข้อมูลใกล้เคียงกับข้อมูลจริงทั้งหมด

3. ฟังก์ชันการแปลงข้อมูล

การแปลงข้อมูลจะใช้วิธีการลดขนาดของข้อมูล (Normalization) ซึ่งจะใช้กับข้อมูลที่เป็นตัวเลข (numeric) เท่านั้น โดยการแปลงจะแปลงให้อยู่ในช่วงค่า 0-1 ซึ่งการลดขนาดของข้อมูลทำให้การคำนวณของเหมืองข้อมูลไม่ต้องทำงานหนักมากเกินไป

4. ฟังก์ชันการใช้บริการผ่านเว็บเซอร์วิส

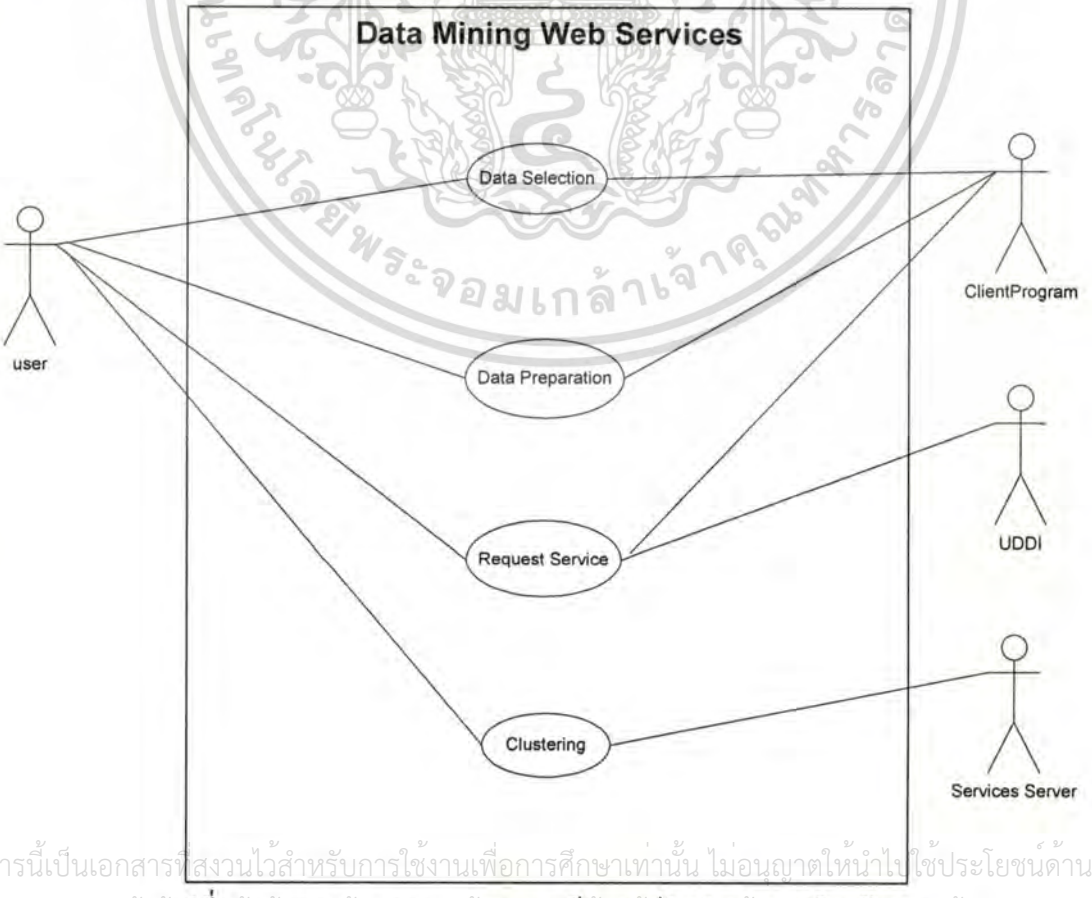
การเรียกใช้บริการผ่านเว็บเซอร์วิสนี้ในขั้นตอนแรกจะใช้การติดต่อไปยัง UDDI ซึ่งจะบอกข้อมูลที่อยู่และการใช้บริการของเว็บเซอร์วิสที่ต้องการจาก WSDL และใช้ AXIS ในการแปลงข้อมูล WSDL เป็นไฟล์สำหรับการเรียกใช้บริการเว็บเซอร์วิส เมื่อได้ข้อมูลการเรียกใช้ของเว็บเซอร์วิสนั้นแล้วโปรแกรมจะติดต่อไปยังเซอร์วิสที่ให้บริการนั้น โดยเลือกการใช้การแบ่งกลุ่มข้อมูลและส่งค่าพารามิเตอร์ต่างๆ ไปให้เซอร์วิสในการประมวลผล จากนั้นจึงรอรับข้อมูลผลลัพธ์ต่อไป

4.1.2 ส่วนเว็บเซอร์วิส

ในส่วนของเว็บเซอร์วิสนี้จะเป็นการบริการฟังก์ชันการแบ่งกลุ่มข้อมูล ซึ่งใช้อัลกอริทึม K-Means และ ISODATA ซึ่งจะรับข้อมูลที่ส่งมาจากไคลเอ็นท์ในรูปแบบของ XML และจะแปลงข้อมูล XML กลับไปเป็นรูปแบบข้อมูลที่ใช้ในการประมวลต่อไป จากนั้นจึงส่งผลลัพธ์ในรูปแบบ text กลับไปยังไคลเอ็นท์ต่อไป

4.2 การออกแบบระบบ

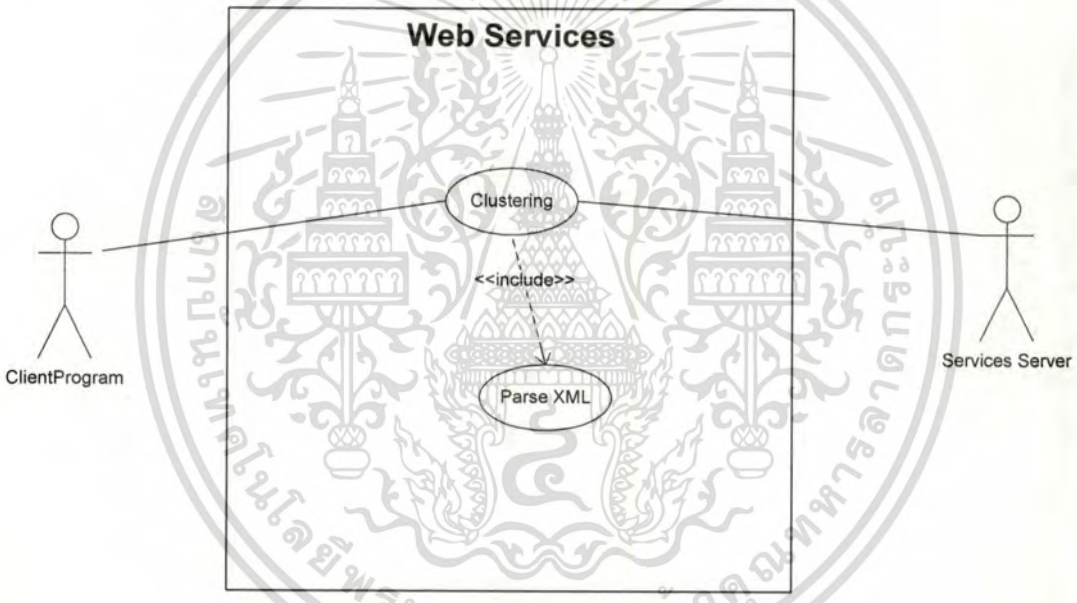
4.2.1 Use Case Diagram



รูปที่ 4.1 Use Case Diagram ระบบเว็บเซอร์วิสสำหรับเหมืองข้อมูล

จากรูปที่ 4.1 ในการออกแบบระบบจะใช้ Use Case Diagram ในการอธิบายภาพรวมทั้งหมดของระบบ โดยอธิบายได้ว่า ในระบบจะมี user เป็น primary actor ของระบบ และจะมี passive actor คือ ClientProgram UDDI และ ServiceServer ตามลำดับ โดยจะมี use case หลัก คือ Data Selection (การเลือกข้อมูล) Data Preparation (การเตรียมข้อมูล) Request Service (การขอใช้บริการเว็บเซอร์วิส) และ Clustering (การแบ่งกลุ่มข้อมูล)

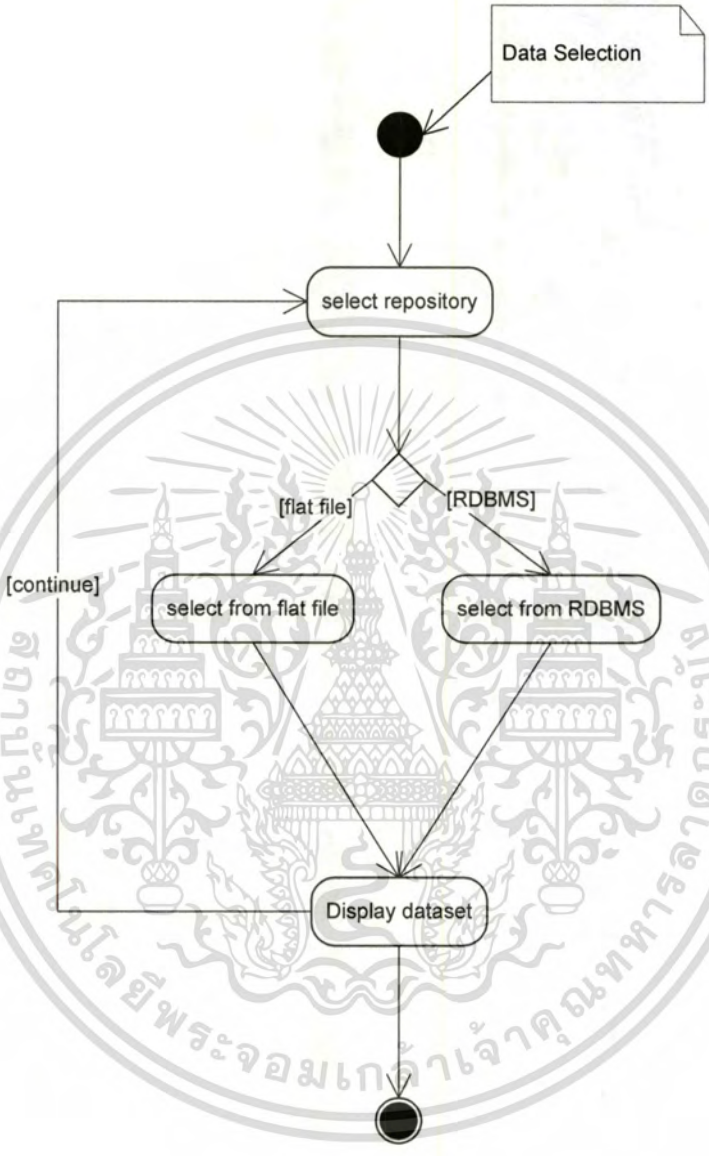
ในส่วนของระบบการให้บริการของเว็บเซอร์วิสนั้น จะเป็นการบริการการแบ่งกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Means และ ISODATA โดย primary actor คือ ClientProgram และ passive actor คือ ServiceServer โดยจะมี use case หลัก คือ Clustering (การแบ่งกลุ่มข้อมูล) ซึ่งจะเรียกใช้ Parse XML (การอ่านข้อมูล XML) โดยระบบการให้บริการของเว็บเซอร์วิสนั้นจะแสดงได้ดังรูปที่ 4.2



รูปที่ 4.2 Use Case Diagram ระบบการให้บริการของเว็บเซอร์วิส

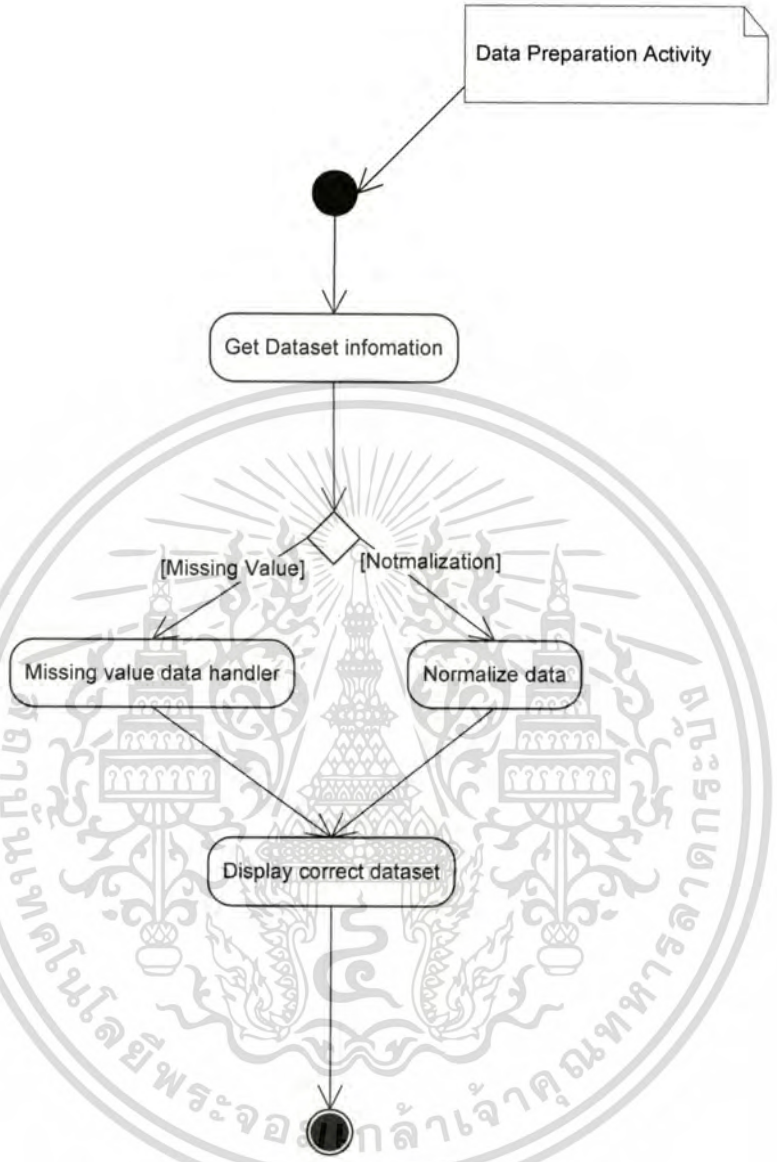
4.2.2 Activity Diagram

1. Activity ของการเลือกข้อมูล



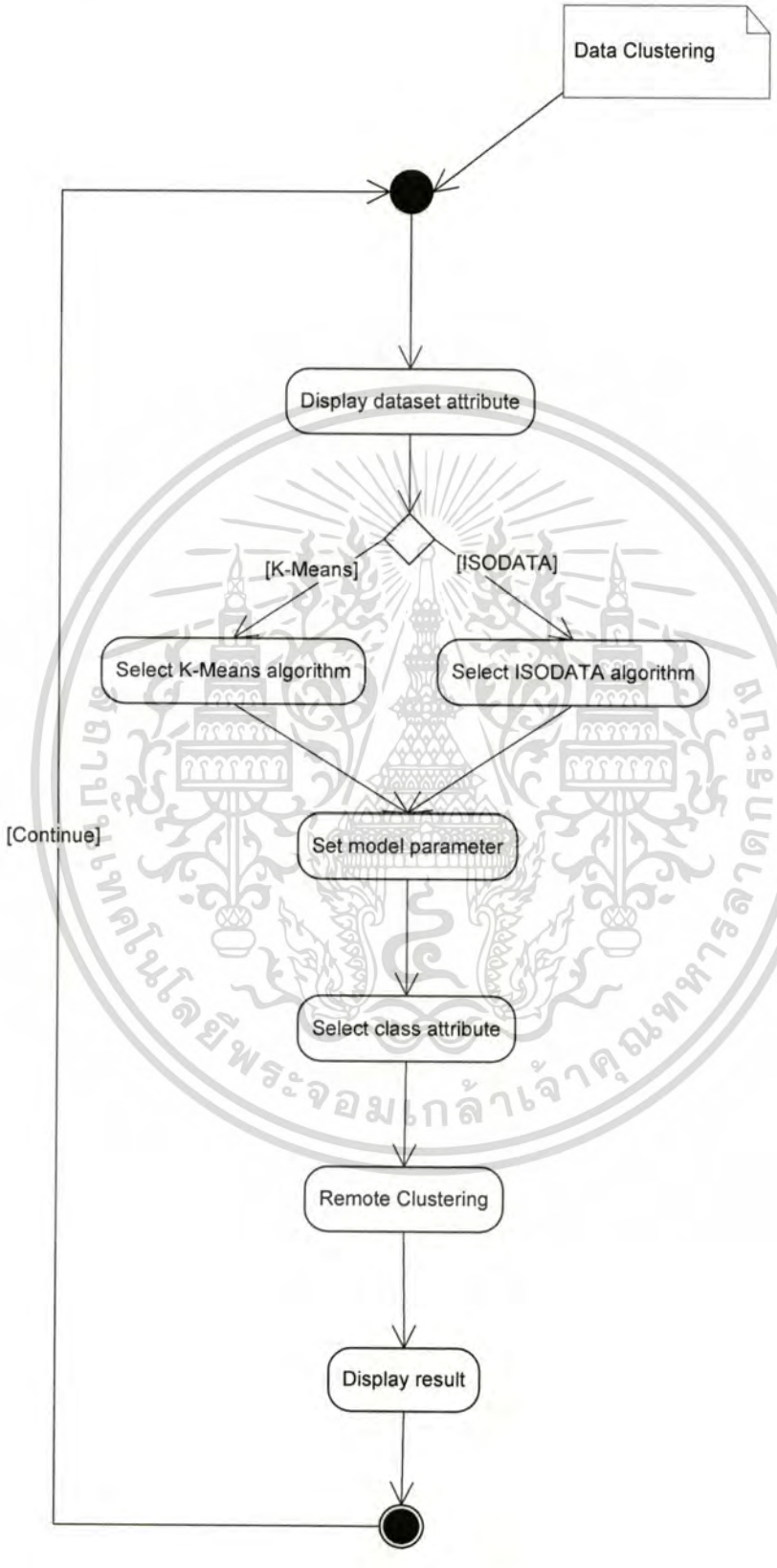
รูปที่ 4.3 Activity Diagram ของการเลือกข้อมูล

2. Activity Diagram ของการเตรียมข้อมูล



รูปที่ 4.4 Activity Diagram ของการเตรียมข้อมูล

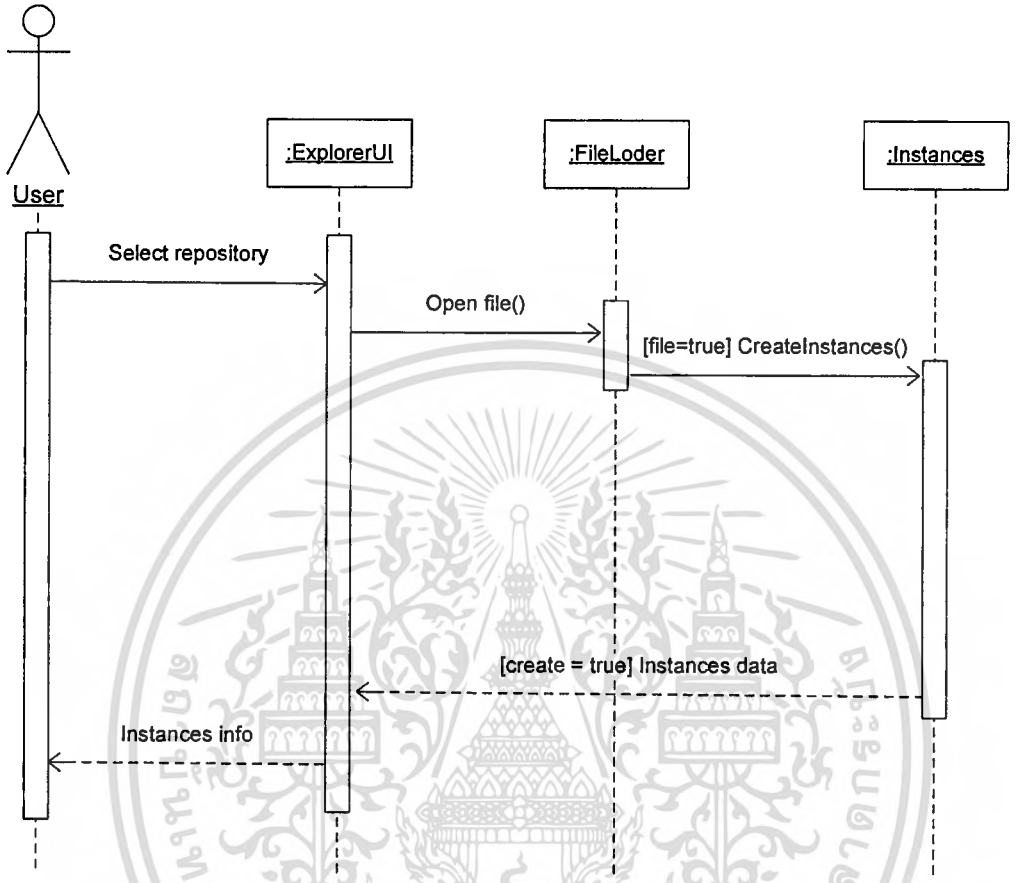
3. Activity Diagram ของการแบ่งกลุ่ม



รูปที่ 4.5 Activity Diagram ของการแบ่งกลุ่ม

4.2.3 Sequence Diagram

1. Sequence Diagram ของการเลือกข้อมูลจากไฟล์

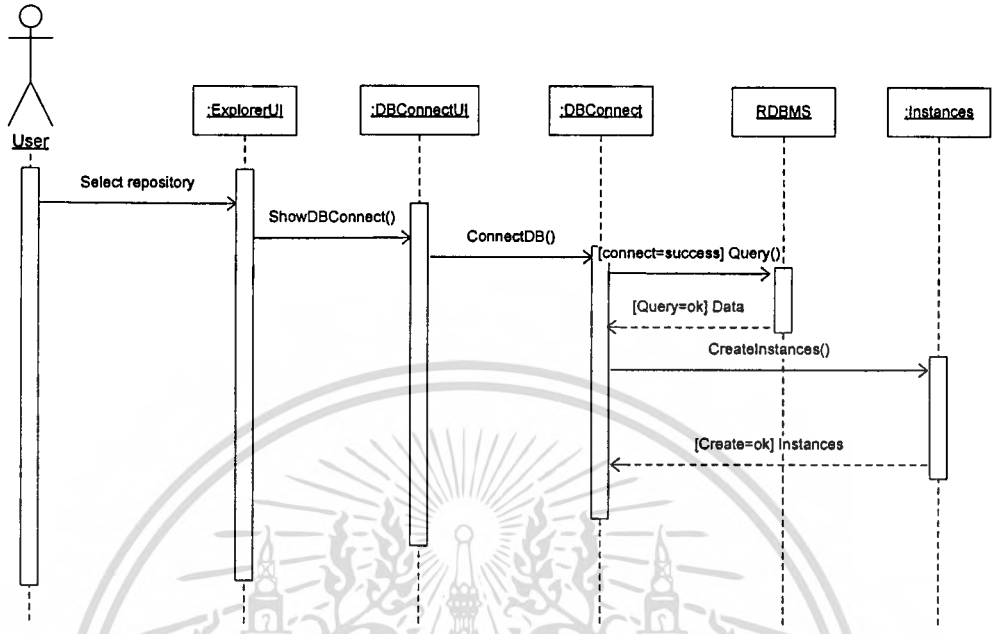


รูปที่ 4.6 Sequence Diagram การเลือกข้อมูลจากไฟล์

จากรูปที่ 4.6 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- ผู้ใช้เลือกการนำเข้าข้อมูลโดยเลือกการนำเข้าข้อมูลจากไฟล์
- หน้าจอหลักจะเรียกใช้ฟังก์ชันการเปิดไฟล์จากคลาส FileLoder
- เมื่อการปิดไฟล์สามารถเปิดได้ถูกต้องจะถูกนำมาสร้างเป็นข้อมูลสำหรับใช้ในโปรแกรมต่อไป
- เมื่อสร้างข้อมูลเรียบร้อยแล้วจะถูกส่งกลับมาแสดงข้อมูลและรายละเอียดของข้อมูลต่อไป

2. Sequence Diagram ของการเลือกข้อมูลจากฐานข้อมูล

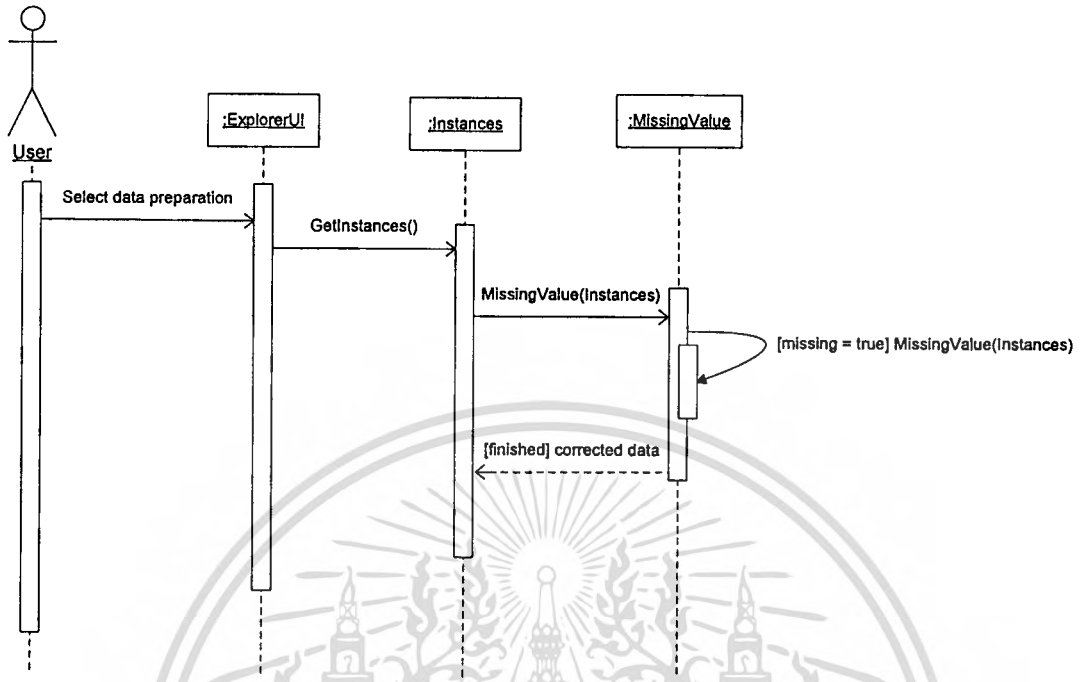


รูปที่ 4.7 Sequence Diagram การเลือกข้อมูลจากฐานข้อมูล

จากรูปที่ 4.7 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- ผู้ใช้เลือกการนำเข้าข้อมูลโดยเลือกการนำเข้าข้อมูลจากฐานข้อมูล
- หน้าจอหลักจะเปิดหน้าจอการสืบค้นข้อมูลจากฐานข้อมูล
- หน้าจอการติดต่อฐานข้อมูลจะติดต่อไปยังฐานข้อมูล
- เมื่อติดต่อฐานข้อมูลได้แล้วจะใช้คำสั่งสืบค้น และข้อมูลจะถูกส่งกลับไปยังหน้าจอการติดต่อฐานข้อมูล
- จากนั้นจะสร้างข้อมูลให้เป็นรูปแบบสำหรับใช้ในโปรแกรมต่อไป

3. Sequence Diagram ของการแก้ไขข้อมูล

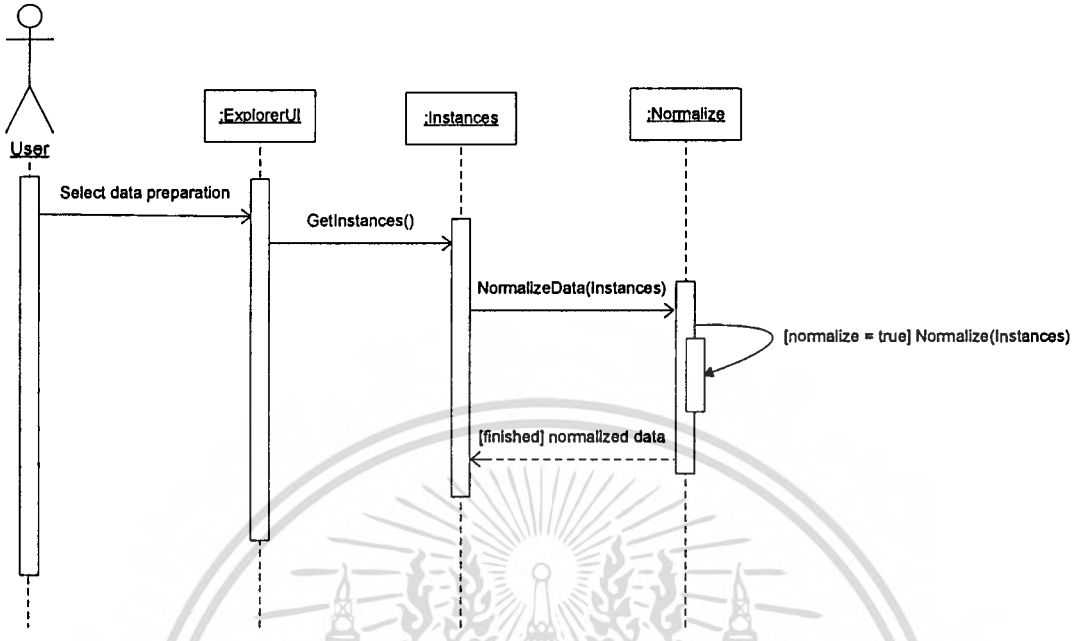


รูปที่ 4.8 Sequence Diagram การแก้ไขข้อมูล

จากรูปที่ 4.8 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- เมื่อผู้ใช้ได้แสดงข้อมูลทางหน้าจอแล้ว จะเลือกการแก้ไขข้อมูลค่าที่ผิดพลาด
- ข้อมูลจะถูกใช้ฟังก์ชัน MissingValue ในการแก้ไขข้อมูลโดยการแทนที่ค่าที่ถูกต้องแทนค่าที่ได้ว่างไว้
- ดำเนินการแก้ไขค่าให้ถูกต้องจนครบทั้งหมด
- จากนั้นจึงส่งข้อมูลที่ได้แก้ไขแล้วกลับมาแสดงยังหน้าจอหลักต่อไป

4. Sequence Diagram ของการแปลงข้อมูล

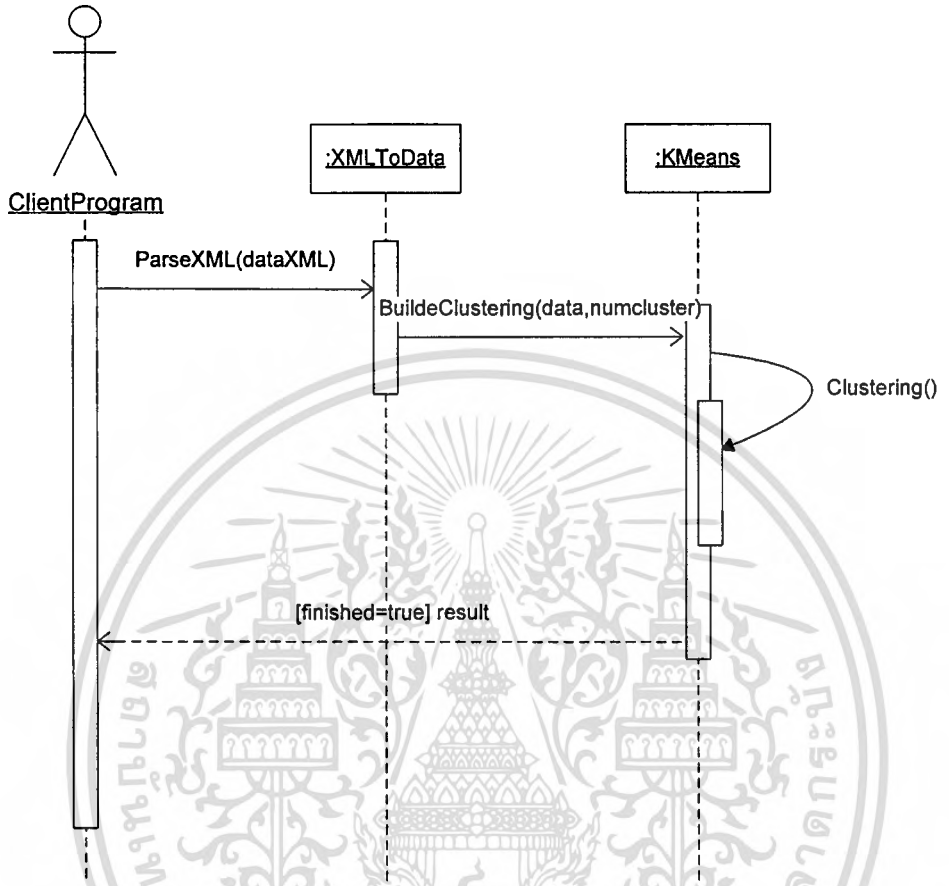


รูปที่ 4.9 Sequence Diagram การแปลงข้อมูล

จากรูปที่ 4.9 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- เมื่อผู้ใช้ได้แสดงข้อมูลทางหน้าจอแล้ว จะเลือกการแปลงข้อมูลโดยการลดขนาดข้อมูล
- ข้อมูลจะถูกใช้ฟังก์ชัน NormalizedData ในการแปลงข้อมูลให้อยู่ในชาวงที่กำหนด
- ดำเนินการแปลงข้อมูลให้ถูกต้องจนครบทั้งหมด
- จากนั้นจึงส่งข้อมูลที่ได้แปลงแล้วกลับมาแสดงยังหน้าจอหลักต่อไป

5. Sequence Diagram ของการแบ่งกลุ่มโดยใช้อัลกอริทึม K-Means

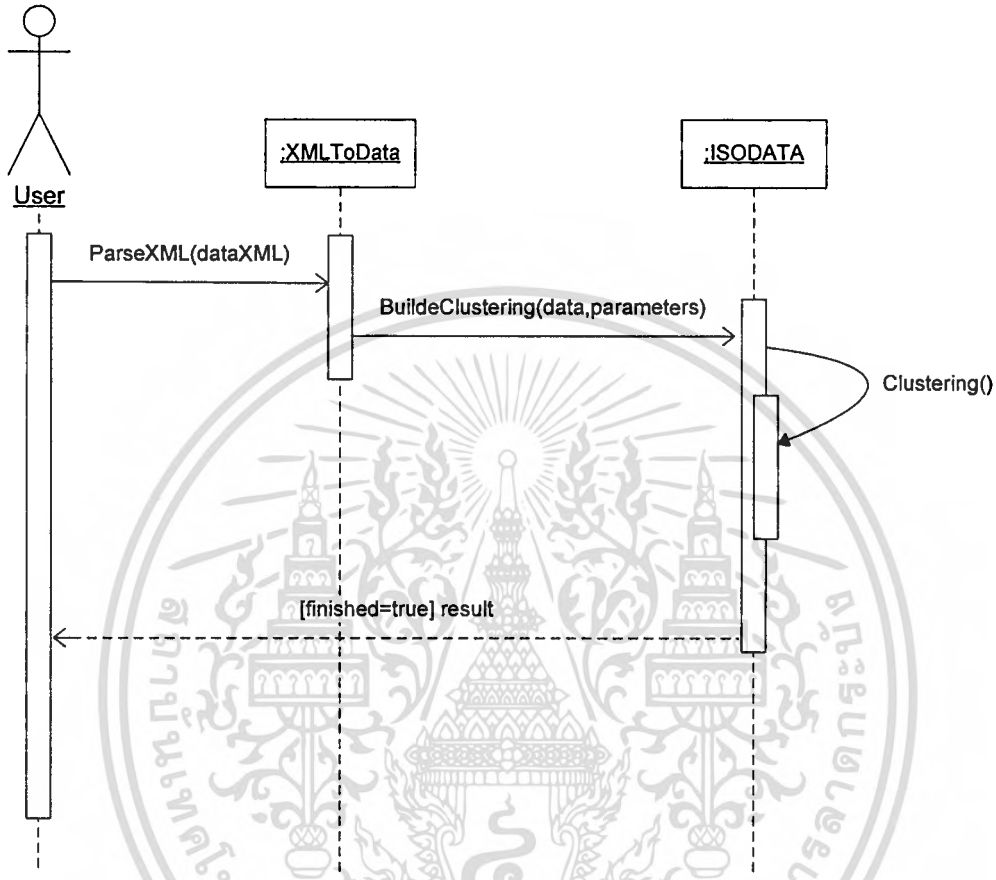


รูปที่ 4.10 Sequence Diagram การแบ่งข้อมูลโดยใช้อัลกอริทึม K-Means

จากรูปที่ 4.10 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- ClientProgram ส่งข้อมูลที่อยู่ในรูปแบบ XML มายังเว็บเซอร์วิส
- โปรแกรมอ่านข้อมูล XML และแปลงข้อมูล XML ให้เป็นรูปแบบข้อมูลที่ใช้สำหรับการแบ่งกลุ่ม
- โปรแกรมจะดำเนินการแบ่งกลุ่มจนได้จุดศูนย์กลางของแต่ละกลุ่มที่ชัดเจน
- จากนั้นจึงส่งผลลัพธ์ที่ได้จากการแบ่งกลุ่มแล้วกลับไปยังหน้าจอการแบ่งกลุ่มของ ClientProgram

6. Sequence Diagram ของการแบ่งกลุ่มโดยใช้อัลกอริทึม ISODATA

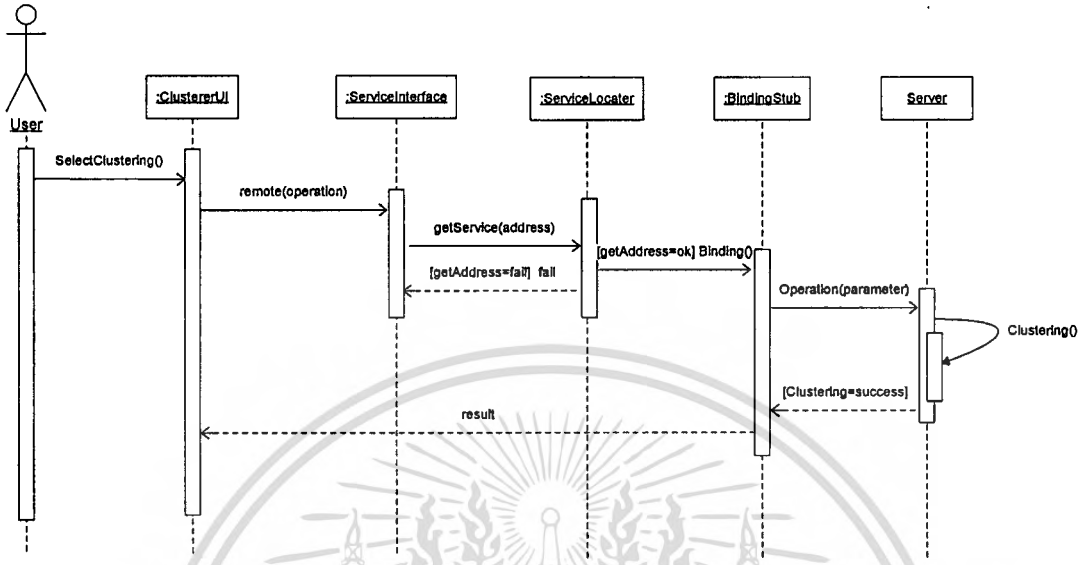


รูปที่ 4.11 Sequence Diagram การแบ่งข้อมูลโดยใช้อัลกอริทึม ISODATA

จากรูปที่ 4.11 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- ClientProgram ส่งข้อมูลที่อยู่ในรูปแบบ XML มายังเว็บเซอร์วิส
- โปรแกรมอ่านข้อมูล XML และแปลงข้อมูล XML ให้เป็นรูปแบบข้อมูลที่ใช้สำหรับการแบ่งกลุ่ม
- โปรแกรมจะดำเนินการแบ่งกลุ่มจนได้จุดศูนย์กลางของแต่ละกลุ่มที่ชัดเจน
- จากนั้นจึงส่งผลลัพธ์ที่ได้จากการแบ่งกลุ่มแล้วกลับไปยังหน้าจอการแบ่งกลุ่มของ ClientProgram

7. Sequence Diagram ของการเรียกใช้การแบ่งกลุ่มผ่านเว็บเซอร์วิส

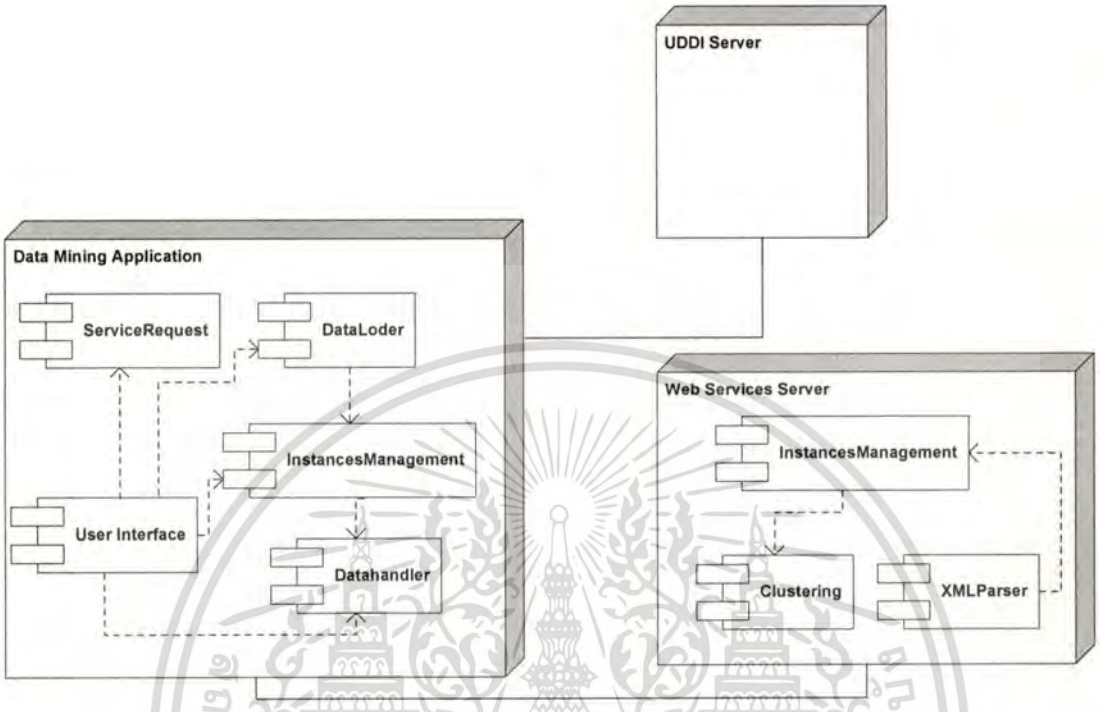


รูปที่ 4.12 Sequence Diagram การเรียกใช้การแบ่งกลุ่มผ่านเว็บเซอร์วิส

จากรูปที่ 4.12 อธิบาย Sequence Diagram ได้ดังต่อไปนี้

- ผู้ใช้เลือกอัลกอริทึมในการแบ่งกลุ่มข้อมูล
- หน้าจอการแบ่งกลุ่มจะเรียกฟังก์ชัน Remote() ของ ServiceInterface เพื่อเรียกใช้บริการอัลกอริทึมที่กำหนด
- โปรแกรมจะดำเนินการหาที่อยู่ของการบริการ ซึ่งถ้าพบการบริการจะทำการ binding การบริการนั้น แต่ถ้ามพบหรือผิดพลาดจะส่งข้อความการติดต่อล้มเหลวกลับมา
- การ binding จะเรียกใช้ operation ที่เลือกไว้พร้อมทั้งส่งค่าพารามิเตอร์ที่เกี่ยวข้องเพื่อให้เซิร์ฟเวอร์ทำการแบ่งกลุ่ม จนกระทั่งเสร็จสิ้นการดำเนินการ
- จากนั้นจึงส่งผลลัพธ์ที่ได้จากการแบ่งกลุ่มแล้วกลับไปยังหน้าจอการแบ่งกลุ่ม

4.2.4 Deployment Diagram



รูปที่ 4.13 Deployment Diagram ระบบเว็บเซอร์วิสของเหมืองข้อมูล

จากรูปที่ 4.13 แสดง Deployment Diagram ของระบบ โดยในแต่ละโหนดของฮาร์ดแวร์จะประกอบด้วยคอมโพเน็นท์ต่างๆ ซึ่งจะอธิบายได้ดังนี้ คือ

1. Data Mining Application

เป็นโหนดของไคลเอ็นท์โปรแกรมซึ่งจะประกอบด้วยคอมโพเน็นท์หลักดังนี้

- User Interace ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องของส่วนติดต่อกับผู้ใช้
- Data Loder ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องในการนำเข้าของข้อมูล
- InstancesManagment ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องของในการสร้างรูปแบบข้อมูลสำหรับใช้ในโปรแกรม
- Datahandler ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องของในการเตรียมข้อมูล
- SericeRequest ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องในการติดต่อไปยังเว็บเซอร์วิส

2. UDDI Server

เป็นโหนดของ UDDI ในการเก็บการบริการของเว็บเซอร์วิส

3. Web Services Server

เป็นโหนดของเซิร์ฟเวอร์ที่ให้บริการการแบ่งกลุ่มของข้อมูล โดยประกอบด้วยคอมโพเน็นท์ที่เกี่ยวข้อง คือ

- XMLParser ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องในการแปลงข้อมูล XML ไปเป็นรูปแบบข้อมูลที่ใช้สำหรับการประมวลผล
- InstancesManagment ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องของในการสร้างรูปแบบข้อมูลสำหรับใช้ในโปรแกรม
- Clustering ประกอบด้วยแพกเกจและคลาสที่เกี่ยวข้องของการแบ่งกลุ่มได้แก่ อัลกอริทึม K-Means และ ISODATA

บทที่ 5
การทำงานของโปรแกรม

ในบทนี้จะกล่าวถึงรายละเอียดของข้อมูลที่นำมาใช้ และขั้นตอนการทำงานของโปรแกรม โดยจะนำเสนอการใช้งานโปรแกรมจากส่วนติดต่อของผู้ใช้ของไคลเอ็นท์ และการใช้งานโดยเรียกบริการผ่านเว็บเซอร์วิส โดยมีรายละเอียดดังต่อไปนี้

5.1 ขั้นตอนการเลือกข้อมูล

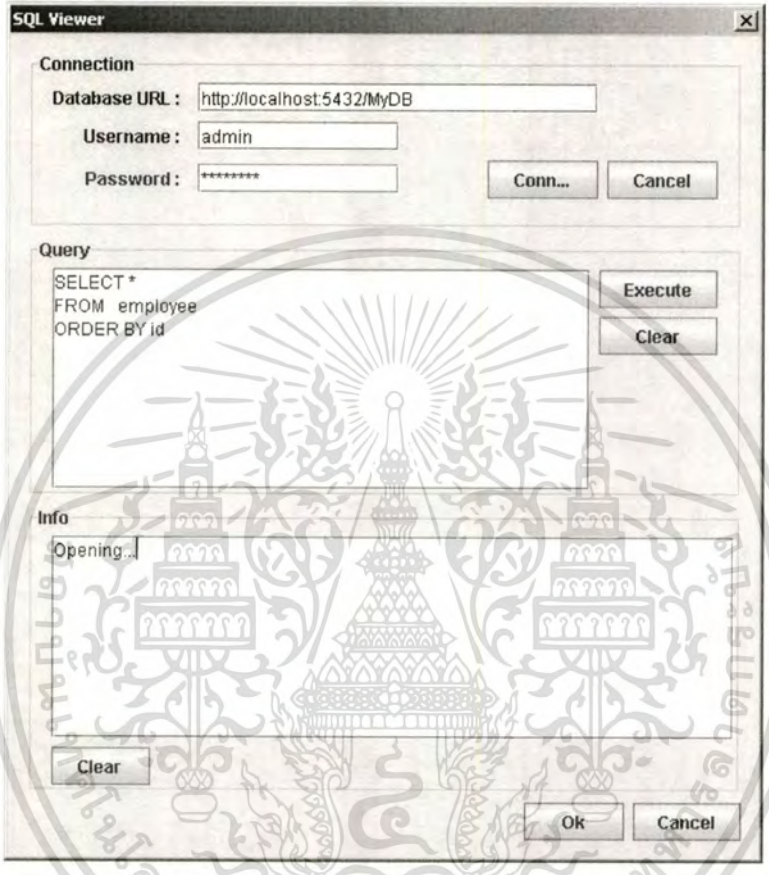
ขั้นตอนแรกจะเป็นการเลือกข้อมูลที่จะนำมาใช้กับโปรแกรม ซึ่งจะต้องมีการทำความสะอาดข้อมูลก่อน โดยทำการลบข้อมูลที่ผิดพลาด หรือข้อมูลรบกวนออกไป

5.2 ขั้นตอนการใช้งานส่วนไคลเอนท์

เมื่อเปิดโปรแกรมเหมืองข้อมูลขึ้นมาจะแสดงส่วนติดต่อกับผู้ใช้ โดยหน้อจอนี้จะมีเมนูการเลือกการจัดการไฟล์ และ การวิเคราะห์ข้อมูล โดยจะแสดงได้ดังรูปที่ 5.1

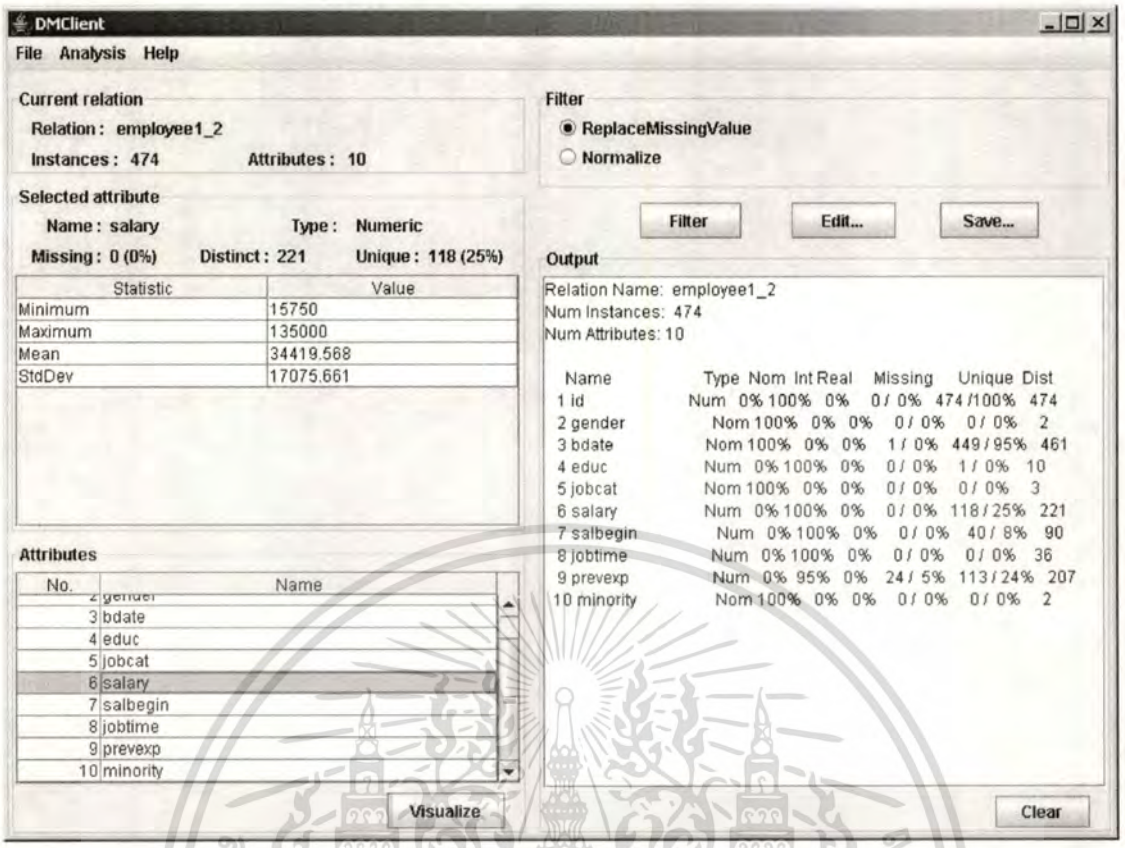
รูปที่ 5.1 หน้าจอแสดงรายละเอียดของข้อมูลและการจัดการข้อมูล

จากนั้นเลือกเปิดไฟล์ที่ต้องการนำมาทำไมนิ่ง โดยไปที่เมนู File->Open File จากนั้นกดเลือกไฟล์ หรือถ้าต้องการเรียกจากฐานข้อมูลสามารถทำได้โดยไปที่เมนู File->Open DB โดยหน้าจอการคิวรีข้อมูลจากฐานข้อมูล แสดงได้ดังรูปที่ 5.2



รูปที่ 5.2 หน้าจอการคิวรีข้อมูล

เมื่อได้เปิดข้อมูลจากไฟล์หรือคิวรีข้อมูลจากฐานข้อมูลแล้ว โปรแกรมจะแสดงรายละเอียดของข้อมูล โดยจะแสดงรายละเอียดของข้อมูลต่างๆ ค่าทางสถิติในแต่ละแอททริบิวท์ โดยแสดงได้ดังรูปที่ 5.3

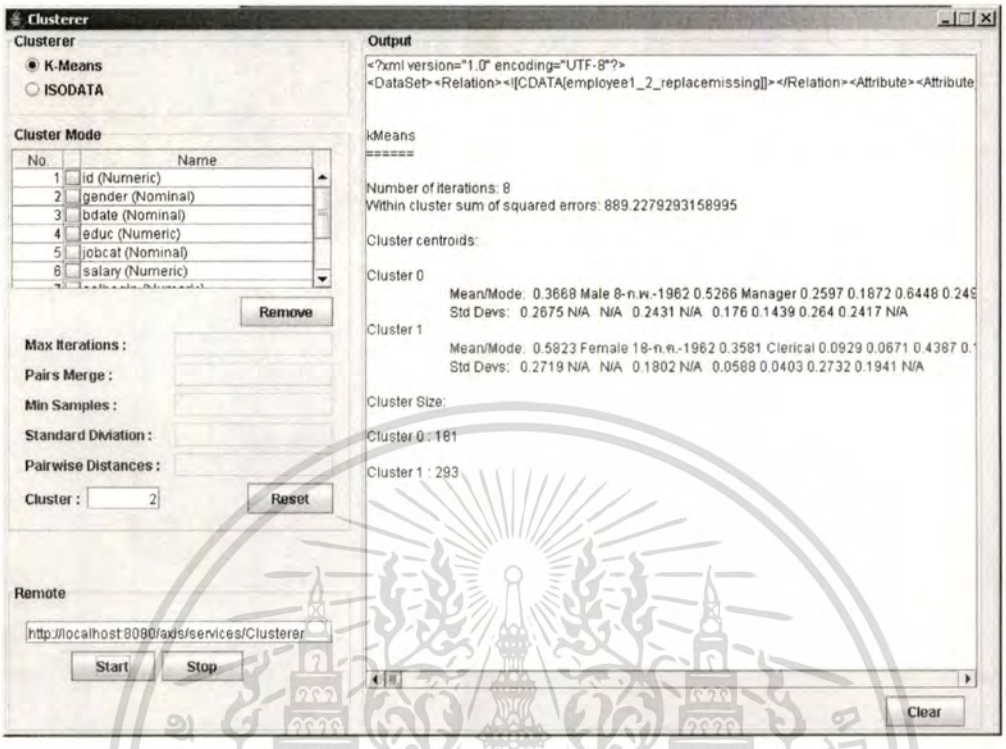


รูปที่ 5.3 หน้าจอรายละเอียดของข้อมูล

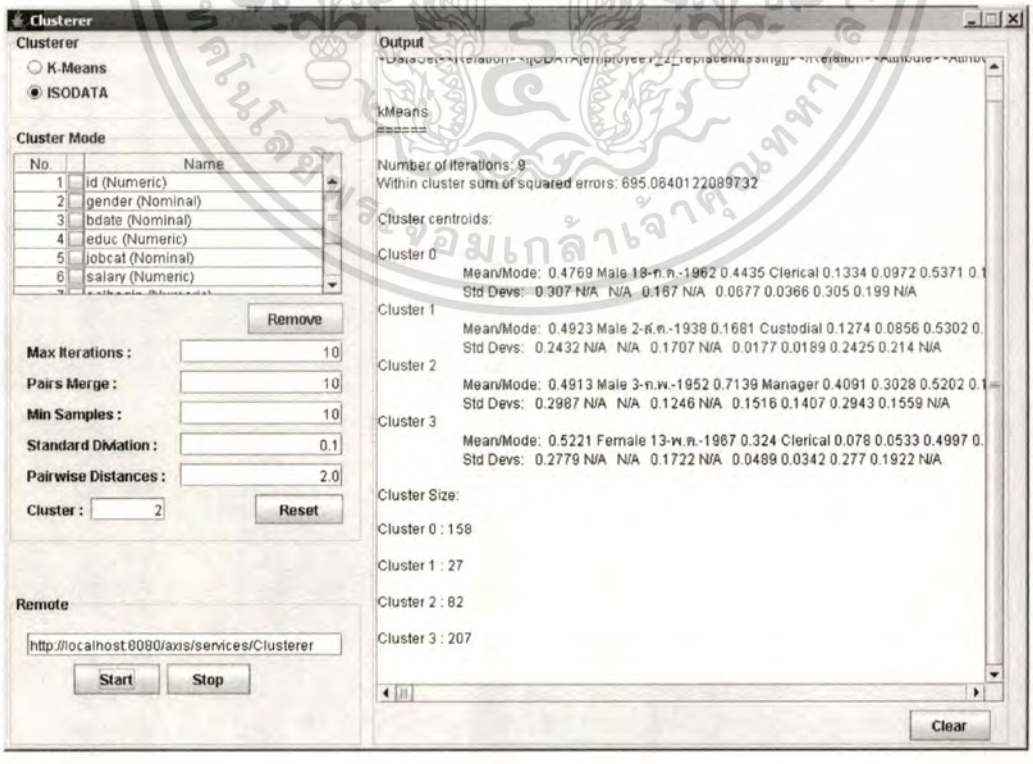
ในขั้นตอนต่อไปจะเป็นขั้นตอนการเตรียมข้อมูล ซึ่งจะต้องมีการแก้ไขข้อมูลที่ผิดพลาด โดยเลือกที่ ReplaceMissingValue โปรแกรมจะแก้ไขค่าที่ว่างโดยการใส่ข้อมูลให้มีค่า ซึ่งถ้าข้อมูลเป็นตัวเลขโปรแกรมจะคำนวณค่าเฉลี่ย (Mean) ของแอททริบิวท์นั้นและจะนำมาใส่ในค่าว่าง แต่ถ้าข้อมูลเป็นจำนวนที่ไม่ใช่ตัวเลข โปรแกรมจำนำค่าความถี่สูงสุด (Mode) ของแอททริบิวท์นั้นมาใส่แทน และเลือกที่ Normalize โปรแกรมจะนอมอลไลซ์ข้อมูล ให้ข้อมูลที่เป็นตัวเลขมีค่าอยู่ในช่วง 0-1 จากนั้นกดที่ปุ่ม filter เพื่อแปลงข้อมูล

5.3 ขั้นตอนการใช้งานผ่านเว็บเซอร์วิส

ในขั้นตอนนี้เป็นการแสดงการใช้บริการการทำเหมืองข้อมูลผ่านเว็บเซอร์วิส โดยในหน้าจอการแบ่งกลุ่ม ให้ใส่ URL ของเว็บเซอร์วิสที่ให้บริการ หลังจากนั้นกดปุ่ม Start โปรแกรมจะส่งข้อมูลไปให้เว็บเซอร์วิสคำนวณผล และเมื่อเสร็จสิ้นการคำนวณแล้ว โปรแกรมจะรับค่าผลลัพธ์กลับมาแสดง โดยการแบ่งกลุ่มสามารถเลือกใช้อัลกอริทึม K-Means และ ISODAT ดังรูปที่ 5.4 และ 5.5 ตามลำดับ



รูปที่ 5.4 การแบ่งกลุ่มโดยใช้อัลกอริทึม ISODATA



รูปที่ 5.5 การแบ่งกลุ่มโดยใช้อัลกอริทึม K-Means ผ่านเว็บเซอร์วิส

บทที่ 6
สรุปผล และข้อเสนอแนะ

ในการแข่งขันทางธุรกิจที่สูงขึ้นในปัจจุบัน การนำข้อมูลมาผ่านกระบวนการเปลี่ยนเป็นสารสนเทศนั้นมีความสำคัญมากขึ้น ซึ่งการใช้ประโยชน์จากข้อมูลที่มีอยู่นั้นถ้าองค์กรใดมีความสามารถในการนำข้อมูลที่มีอยู่เหล่านั้นมาใช้ให้เกิดประโยชน์มากที่สุด จะทำให้องค์กรนั้นจะมีกลยุทธ์ที่เหนือกว่าคู่แข่งอย่างมากมาย ในการที่จะใช้ข้อมูลที่มีอยู่เหล่านั้นให้เกิดประโยชน์สูงสุด ได้มีการนำเหมืองข้อมูลมาใช้ในองค์กร ซึ่งเหมืองข้อมูลได้ช่วยให้ผู้บริหารสามารถกำหนดกลยุทธ์ทางธุรกิจได้โดยใช้ข้อมูลที่ผ่านการประมวลผลแล้ว และแอพพลิเคชันทางเหมืองข้อมูลนั้นต่างก็มีเทคโนโลยีเป็นของตัวเอง ซึ่งในการใช้เทคโนโลยีใดๆ ต้องขึ้นอยู่กับผู้ผลิตแอพพลิเคชันนั้นๆ จึงทำให้ไม่มีอิสระทางการใช้แอพพลิเคชันทางเหมืองข้อมูล ดังนั้นการนำเทคโนโลยีเว็บเซอร์วิสมาใช้ร่วมกับเหมืองข้อมูลจึงเป็นอีกหนทางหนึ่งที่จะทำให้การใช้งานและการพัฒนาแอพพลิเคชันเหมืองข้อมูลไม่ยึดติดกับแพลตฟอร์มใดๆ

เหมืองข้อมูลเป็นเทคนิคในการค้นหาองค์ความรู้จากแหล่งข้อมูลที่มีข้อมูลอยู่เป็นจำนวนมาก ช่วยให้นักวิเคราะห์ข้อมูลสามารถนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์อย่างสูงสุด และทำให้องค์กรนั้นสามารถดำเนินกิจการหรือพัฒนาไปได้ตรงตามวัตถุประสงค์ที่กำหนดไว้ โดยเหมืองข้อมูลจะช่วยในการวิเคราะห์ข้อมูล ทำนายข้อมูลล่วงหน้า หรือหาความสัมพันธ์กันของข้อมูล และการที่จะนำเหมืองข้อมูลมาใช้งานนั้นจะต้องมีกระบวนการขั้นตอนในการนำเหมืองข้อมูลมาแก้ปัญหาทางธุรกิจให้เหมาะสมมากที่สุด และเมื่อนำมาใช้ร่วมกับเทคโนโลยีเว็บเซอร์วิสซึ่งเป็นการให้บริการการทำงานระหว่างแอพพลิเคชันจะช่วยให้การพัฒนาแอพพลิเคชันของเหมืองข้อมูลเป็นไปได้อย่างสะดวกและรวดเร็ว และทำให้การใช้เหมืองข้อมูลเกิดการขยายตัวมากขึ้น เนื่องจากองค์กรสามารถพัฒนามาใช้เองได้

วิธีการทำงานหรือเทคนิคของเหมืองข้อมูลที่สำคัญวิธีหนึ่งคือการแบ่งกลุ่ม ซึ่งเป็นวิธีที่จะดูลักษณะของข้อมูลว่าข้อมูลจะสามารถแบ่งออกได้กี่กลุ่ม ซึ่งข้อมูลที่อยู่กลุ่มเดียวกันจะมีความคล้ายกันช่วยให้การวิเคราะห์ข้อมูลในขั้นต่อไปทำได้สะดวกขึ้น โดยอัลกอริทึมพื้นฐานของการแบ่งกลุ่มคือ K-Means ซึ่งเป็นอลักอริทึมที่ง่าย ทำงานได้รวดเร็ว และอัลกอริทึม ISODATA เป็นอัลกอริทึมที่พัฒนาขึ้นมาช่วยให้การแบ่งกลุ่ม เนื่องจากสามารถกำหนดการแบ่งกลุ่มได้ด้วยตัวเองเพื่อให้กลุ่มข้อมูลมีความเหมาะสมมากที่สุด

ในโครงงานนี้ ได้ศึกษาถึงแนวทางการนำเทคโนโลยีเว็บเซอร์วิสมาใช้ร่วมกับเหมืองข้อมูล เพื่อหาความเป็นไปได้ในการทำให้การพัฒนาแอพพลิเคชันและการใช้งานเหมืองข้อมูลมีความสะดวกมากยิ่งขึ้น โดยการให้บริการของเว็บเซอร์วิสสำหรับเหมืองข้อมูลนั้น จะให้การ

บริการการคำนวณแบบจำลองสำหรับนำไปใช้ในการวิเคราะห์ข้อมูล โดยได้นำเสนอการสร้างแบบจำลองแบบแบ่งกลุ่ม เป็นตัวอย่างการให้บริการผ่านเว็บเซอร์วิสภายในองค์กร

วิธีการนำเสนอในโครงงานนี้ เป็นเพียงแนวทางเริ่มต้นในการศึกษาการใช้เทคโนโลยีร่วมกัน ซึ่งสามารถพัฒนาให้เว็บเซอร์วิสมีการบริการทางเหมืองข้อมูลที่หลายหลายเพิ่มมากขึ้น หรือเพิ่มวิธีการในการสร้างแบบจำลอง และปรับปรุงอัลกอริทึมให้ดีขึ้นได้ โดยไม่ส่งผลกระทบต่อแอพพลิเคชันที่เรียกใช้ อีกทั้งยังสามารถปรับปรุงในเรื่องประสิทธิภาพในการส่งข้อมูลให้มีปริมาณจำนวนที่มากขึ้นในขณะที่ใช้เวลาคำนวณและเวลาการส่งข้อมูลมีค่าที่น้อยลงยิ่งขึ้นตามลำดับ

บรรณานุกรม

CRoss Industry Standard Process for Data Mining. [Online]. Available : http://www.crisp-dm.org/. 2000

DM Methodology. [Online]. Available : http://dms.irb.hr/tutorial/tut_main.php

H. Xiangji, 2006. **"clustering analysis and algorithm"**. Encyclopedia of Data Warehousing and Mining, pp. 159-162.

T. ZhaoHui and M. Jamie, 2005. **Data Mining with SQL Server 2005**. Indianapolis USA: Wiley Publishing, Inc.,

W. Ruye, 2004. The Isodata Algorithm. [Online]. Available : http://fourier.eng.hmc.edu/e161/lectures/classification/node13.html

W. Ruye, 2004. The K-Means Algorithm. [Online]. Available : http://fourier.eng.hmc.edu/e161/lectures/classification/node12.html

Web Services. [Online]. Available : http://www.w3.org/2002/ws/. 2002

Web Services Axis. [Online]. Available : http://ws.apache.org/axis/. 2004

ประวัติผู้เขียน

ชื่อ – นามสกุล	นายนพปฎล ถาวรรัฐ
วัน เดือน ปีเกิด	27 มกราคม 2523 ที่กรุงเทพมหานคร
ที่อยู่	129/1 ถ. ตากสินมหาราช ต. ท่าประดู่ อ.เมือง จ.ระยอง 21000
	โทร 0-3887-0384, 08-1377-2080
ประวัติการศึกษา	2545 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์
	มหาวิทยาลัยบูรพา
ประสบการณ์ทำงาน	2545 – 2548 ตำแหน่งนักพัฒนาโปรแกรม บ. เมดดิโคซอฟท์ จำกัด