การทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์
กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง

นายอภิชาติ ปัญญา

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
แขนงวิชาเทคโนโลยีสารสนเทศและการสื่อสาร สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช
พ.ศ. 2557

Data Mining for Website Development: A Case of Rajamangala University of Technology Lanna Lampang Website

Mr. Apichat Punya

A Thesis Submitted in Partial Fulfillment of the Requirements for
the Degree of Master of Science in Information and Communication Technology
School of Science and Technology
Sukhothai Thammathirat Open University
2014

หัวข้อวิทยานิพนธ์	การทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์ กรณีศึกษาเว็บไซต์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง
ชื่อและนามสกุล	นายอภิชาติ ปัญญา
แขนงวิชา	เทคโนโลยีสารสนเทศและการสื่อสาร
สาขาวิชา	วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช
อาจารย์ที่ปรึกษา	1. รองศาสตราจารย์สำรวย กมลายุตต์ 2. อาจารย์ ดร. ดวงดาว วิชาดากุล

วิทยานิพนธ์นี้ ได้รับความเห็นชอบให้เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรระดับปริญญาโท เมื่อวันที่ 30 กันยายน 2557

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ศิริภัทรา เหมือนมาลัย)

.....กรรมการ
(รองศาสตราจารย์สำรวย กมลายุตต์)

.....กรรมการ
(อาจารย์ ดร. ดวงดาว วิชาดากุล)

.....ประธานกรรมการบัณฑิตศึกษา
(ศาสตราจารย์ ดร. สิริวรรณ ศรีพหล)

ชื่อวิทยานิพนธ์ การทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์
กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง
ผู้วิจัย นายอภิชาติ ปัญญา **รหัสนักศึกษา** 2549600225 **ปริญญา** วิทยาศาสตรมหาบัณฑิต (เทคโนโลยีสารสนเทศและการสื่อสาร) **อาจารย์ที่ปรึกษา** (1) รองศาสตราจารย์สำรวย กมลายุตต์ (2) อาจารย์ ดร. ดวงดาว วิชาดากุล **ปีการศึกษา** 2557

บทคัดย่อ

งานวิจัยนี้นำเสนอการทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์ กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ซึ่งมีวัตถุประสงค์เพื่อ 1) สร้างคลังข้อมูลการเข้าใช้งานเว็บไซต์ 2) ทำเหมืองข้อมูลวิเคราะห์จำแนกกลุ่มการใช้งานเว็บไซต์โดยใช้อัลกอริทึมการจัดกลุ่ม 3) ทำเหมืองข้อมูลพยากรณ์จำนวนผู้ใช้งานเว็บไซต์โดยใช้อัลกอริทึมอนุกรมเวลา 4) ทำเหมืองข้อมูลวิเคราะห์หน้าเว็บที่มีความเกี่ยวข้องกันโดยใช้อัลกอริทึมกฎความสัมพันธ์5) ทำเหมืองข้อมูลวิเคราะห์จำแนกกลุ่มการใช้งานเว็บไซต์โดยใช้อัลกอริทึมการจัดกลุ่มโดยใช้ลำดับ

เครื่องมือที่ใช้ในงานวิจัย ได้แก่ โปรแกรม SQL Server 2008 และ SQL Service Analysis กระบวนการเริ่มจากการรวบรวมล็อกไฟล์การเข้าใช้งานเว็บไซต์ตั้งแต่วันที่ 1 มกราคม 2556 ถึง 31 ธันวาคม 2556 และข้อมูลเว็บไซต์หน่วยงานต่าง ๆ ของมหาวิทยาลัยฯ มาสร้างคลังข้อมูลด้วยกระบวนการอีทีแอล จากนั้นนำข้อมูลจากคลังข้อมูลมาสร้างคิวบ์ เพื่อนำเสนอรายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป และนำข้อมูลจากคลังข้อมูลมาทำเหมืองข้อมูลโดยใช้แบบจำลองคริสพ์-ดีเอ็ม ซึ่งเทคนิคการทำเหมืองข้อมูลประกอบด้วย 4 อัลกอริทึม ได้แก่ 1) การจัดกลุ่ม 2) อนุกรมเวลา 3) กฎความสัมพันธ์ 4) การจัดกลุ่มโดยใช้ลำดับ

ผลการวิจัยที่ได้คือคลังข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์มหาวิทยาลัยฯ ที่ใช้โครงสร้างแบบสโนว์เฟลค และผลจากการทำเหมืองข้อมูลพบว่าการจัดกลุ่มของการเข้าใช้เว็บไซต์แบ่งออกเป็น 6 กลุ่ม อนุกรมเวลาทำให้สามารถพยากรณ์ปริมาณผู้ใช้บริการเว็บไซต์ได้ ส่วนกฎความสัมพันธ์ที่เกิดขึ้นทั้งหมด 16 กฎ และการจัดกลุ่มโดยใช้ลำดับ พบว่าจัดกลุ่มออกเป็น 66 กลุ่มซึ่งกลุ่มที่น่าสนใจมีทั้งหมด 6 กลุ่ม การทำเหมืองข้อมูลสามารถทำให้ได้องค์ความรู้ใหม่ ซึ่งสามารถนำมาใช้ในการวางแผนพัฒนาปรับปรุงและดูแลเว็บไซต์ได้ และยังสามารถทำให้ผู้บริหารได้รับรายงานหลายมิติที่สามารถนำมาประกอบการตัดสินใจ เพื่อประโยชน์ในการบริหารเว็บได้อย่างมีประสิทธิภาพมากขึ้น

คำสำคัญ การทำเหมืองข้อมูล ล็อกไฟล์ การจัดกลุ่ม อนุกรมเวลา กฎความสัมพันธ์ การจัดกลุ่มโดยใช้ลำดับ

Thesis title: Data Mining for Website Development: A Case of Rajamangala University of Technology Lanna Lampang Website
Researcher: Mr. Apichat Punya; **ID**: 2549600225; **Degree:** Master of Science (Information and Communication Technology);
Thesis advisors: (1) Sumruay Komlayut, Associate Professor;
(2) Dr. Duangdao Wichadakul; **Academic year:** 2014

Abstract

This research was titled the data mining for website development case of Rajamangala University of Technology Lanna Lampang Website which was aimed to create the data warehouse for using the website, to analyze the data mining by classifying the website user group with clustering algorithm, to predict amount of the user by Time Series algorithm, to analyze the web page with the related pages by association rule algorithm, to, and to classify the website use groups by Sequence clustering algorithm.

The research instrument were SQL Server 2008 program and SQL Service Analysis, the methodology has been started by collecting the log file of the website using on January 1 to December 31, 2013 and the other website of each department in the university was created the data warehouse by ETL process, the data from data warehouse was created cube for presenting the report from evaluating the analyzing data or OLAP. The data mining was also created from the data of data warehouse by using the model of CRISP-DM, the data mining technic was consisted of 4 Algorithms; 1) Clustering 2)Time Series 3) Association Rules and 4) Sequence Clustering

The result was the data warehouse form the log file the using the university website in Snowflake Schema pattern and it was found that the Clustering of website using were divided into 6 cluster, Time Series which was able to predict the quantities of the website user, Association Rules was detected 16 rules, the Sequence Clustering with series was Cluster of 66 Cluster which were 6 interesting cluster. The data mining was able to detect the new knowledge and was able to plan, develop and look offer the websites, it was reported to the administrator in many dimensions which was able to engage in the decision for more website administration efficiency.

Keywords: Data mining, Log file, Clustering, Time Series,Association Rules, Sequence Clustering

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้ด้วยความอนุเคราะห์จากผู้ทรงคุณวุฒิหลายท่าน โดยเฉพาะอย่างยิ่ง รองศาสตราจารย์สำรวย กมลายุตต์ อาจารย์ที่ปรึกษาหลัก และอาจารย์ ดร.ดวงดาว วิชาดากุล อาจารย์ที่ปรึกษาร่วมวิทยานิพนธ์ฉบับนี้ ที่ได้เสียสละเวลาในการให้คำปรึกษาให้ความรู้ ตลอดถึงแนวทางการปฏิบัติ และข้อเสนอแนะต่างๆ จนวิทยาวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงด้วยดี ผู้วิจัยขอขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบคุณมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง รวมถึงผู้ที่มีส่วนเกี่ยวข้อง ที่ได้อนุเคราะห์ข้อมูล เพื่อใช้ในงานวิจัยครั้งนี้

นอกจากนี้ผู้วิจัยขอขอบพระคุณคณาจารย์ และผู้ทรงคุณวุฒิ ประจำสาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช ที่ค่อยสั่งสอน อบรม มอบความรู้ และให้คำปรึกษาด้วยดีตลอดมาร่วมถึงเพื่อนนักศึกษา ที่ค่อยให้กำลังใจจนกระทั่งงานวิจัยฉบับนี้เสร็จสมบูรณ์ด้วยดีและสิ่งที่สำคัญอย่างยิ่งผู้วิจัยขอขอบพระคุณ บิดา มารดา และครอบครัว ที่ค่อยให้โอกาสในการศึกษา และค่อยให้คำปรึกษา ด้วยความรักและห่วงใยตลอดมา

อภิชาติ ปัญญา

กันยายน 2557

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

สารบัญภาพ (ต่อ)

หน้า

สารบัญภาพ (ต่อ)

หน้า

บทที่ 1
บทนำ

1. ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันการให้บริการข้อมูลสารสนเทศเป็นสิ่งสำคัญสำหรับทุกองค์กรที่จำเป็นต้องประชาสัมพันธ์ข้อมูลข่าวสารการดำเนินกิจกรรมต่างๆ ที่เกิดขึ้นหรือเกี่ยวข้องภายในองค์กร จึงได้มีการนำเทคโนโลยีสารสนเทศเข้ามามีบทบาทในการเผยแพร่ข้อมูลดังกล่าวผ่านทาง “WWW” (World Wide Web) หรือเว็บ ซึ่งเป็นรูปแบบหนึ่งของระบบการเชื่อมโยงเครือข่ายข่าวสาร ใช้ในการค้นหาข้อมูลข่าวสารบนอินเทอร์เน็ต จากแหล่งข้อมูลหนึ่ง ไปยังอีกแหล่งข้อมูลหนึ่ง เว็บจะแสดงผลอยู่ในรูปแบบของเอกสารที่เรียกว่า ไฮเปอร์เท็กซ์ (Hyper Text) ซึ่งเป็นฐานข้อมูลชนิดหนึ่งที่ทำหน้าที่รวบรวมข่าวสารข้อมูลที่อยู่กระจัดกระจายในที่ต่างๆ ทั่วโลกให้สามารถนำมาใช้งานได้เสมือนอยู่ในที่เดียวกัน โดยใช้เว็บเบราเซอร์ซึ่งเป็นโปรแกรมที่ช่วยอ่านข้อมูลเหล่านั้น จึงทำให้ผู้ใช้ได้รับข้อมูลข่าวสารทั่วโลกได้อย่างรวดเร็ว การออกแบบโครงสร้างเว็บไซต์เป็นสิ่งสำคัญที่ทำให้ผู้ใช้สามารถเข้าถึงข้อมูลได้อย่างรวดเร็วและตรงตามความต้องการ แต่ละกลุ่มผู้ใช้มีความต้องการข้อมูลที่แตกต่างกันไป การออกแบบโครงสร้างเว็บไซต์ที่ไม่สอดคล้องกับความต้องการของผู้ใช้ จะส่งผลทำให้การค้นหาข้อมูลจากเว็บไซต์ล่าช้า ทำให้ผู้ใช้เกิดความเบื่อหน่าย อีกทั้งยังไม่น่าสนใจสำหรับกลุ่มผู้ใช้ด้วย

มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง เป็นมหาวิทยาลัยที่เปิดให้บริการด้านการศึกษาในระดับ ประกาศนียบัตรวิชาชีพชั้นสูง (ปวส). ถึงระดับปริญญาตรี และการถ่ายทอดองค์ความรู้สู่ชุมชน จึงต้องให้บริการข้อมูลการประชาสัมพันธ์ข่าวสารอยู่ตลอดเวลา อาทิ ข้อมูลข่าวสาร กิจกรรม การรับสมัครนักศึกษา การประกาศรับสมัครงาน การอบรมเผยแพร่ความรู้ ทางมหาวิทยาลัยจึงได้จัดทำเว็บไซต์เพื่อให้บริการข้อมูลข่าวสารดังกล่าวผ่านทาง “www” ใช้ชื่อ www.lpc.rmutl.ac.th เพื่อให้ข้อมูลการดำเนินกิจกรรมเข้าถึงผู้ใช้อย่างรวดเร็ว โดยได้มีการออกแบบโครงสร้างเว็บไซต์เพื่อรองรับการใช้งานเบื้องต้นจากผู้ใช้ทุกกลุ่ม แต่เนื่องจากเว็บไซต์ดังกล่าวให้บริการข้อมูลแบบสาธารณะ หมายถึง ใครก็สามารถเข้าใช้บริการได้ จึงทำให้ไม่ทราบถึงกลุ่มเป้าหมายที่เข้ามาใช้บริการ ดังนั้นการปรับปรุงแก้ไขโครงสร้างเว็บไซต์จึงไม่สามารถกำหนดรูปแบบที่ทำให้มั่นใจได้ว่าจะสามารถตอบสนองกับความต้องการของผู้ใช้กลุ่มใดกลุ่มหนึ่งได้ หรืออาจเกิดข้อผิดพลาดในการกำหนดนโยบายในการบริการข้อมูลข่าวสารที่ไม่ตรงตามกลุ่มเป้าหมาย ดังนั้นหากสามารถจำแนกกลุ่มผู้ใช้งานตามลักษณะการใช้งานจะทำให้การกำหนดนโยบายและการวางแผนการพัฒนาปรับปรุงเว็บไซต์เป็นไปอย่างมีประสิทธิภาพและสอดคล้องตรงตามความต้องการของผู้ใช้แต่ละกลุ่ม

ดังนั้นผู้ทำการวิจัยจึงมีความสนใจที่จะศึกษาและจำแนกกลุ่มผู้ใช้งานเว็บไซต์ โดยใช้เทคนิคการทำเหมืองข้อมูล เพื่อนำผลที่ได้มาปรับปรุงเว็บไซต์ ทั้งด้านการออกแบบโครงสร้าง และด้านการนำเสนอเนื้อหาที่ตรงตามความต้องการของกลุ่มผู้ใช้งานแต่ละกลุ่ม

2. วัตถุประสงค์การวิจัย

2.1 เพื่อสร้างคลังข้อมูลการเข้าใช้งานเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง

2.2 วิเคราะห์จำแนกกลุ่มการใช้งานเว็บไซต์โดยใช้อัลกอริทึมการจัดกลุ่ม

2.3 พยากรณ์จำนวนผู้ใช้งานเว็บไซต์ โดยใช้อัลกอริทึมอนุกรมเวลา

2.4 วิเคราะห์การใช้หน้าเว็บที่มีความเกี่ยวข้องกันโดยใช้อัลกอริทึมกฎความสัมพันธ์

2.5 วิเคราะห์จำแนกกลุ่มการใช้งานเว็บไซต์โดยใช้อัลกอริทึมการจัดกลุ่มโดยใช้ลำดับ

3. ขอบเขตของการวิจัย

ข้อมูลที่ใช้ในงานวิจัยนี้เป็นการรวบรวมข้อมูลการเข้าใช้งานเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง จากล็อกไฟล์ (Log file) การเข้าใช้งานเว็บไซต์ที่จัดเก็บในเครื่องให้บริการเว็บ (Server-Side) โดยเก็บรวมนวมข้อมูลตั้งแต่วันที่ 1 มกราคม 2556 ถึง 31 ธันวาคม 2556 โดยนำมาสร้างคลังข้อมูล และนำไปใช้ประโยชน์ในการวิเคราะห์เชิงตารางมิติและทำเหมืองข้อมูล

ในงานวิจัยนี้จะวิเคราะห์ข้อมูลของการเข้าใช้งานเว็บไซต์ของมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง โดยใช้แบบจำลองคริสพ์-ดีเอ็ม (CRISP-DM) ซึ่งเป็นแบบจำลองการทำเหมืองข้อมูล เพื่อจำแนกประเภทของกลุ่มผู้ใช้ การค้นหาความสัมพันธ์ และการพยากรณ์จำนวนผู้ใช้บริการในอนาคต

4. กรอบแนวคิดของการวิจัย

กรอบแนวความคิดของการทำวิจัยประกอบด้วย ตัวแปรต้นคือข้อมูลเกี่ยวกับการเข้าใช้งานเว็บไซต์ของทางมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ในช่วงระยะเวลา 1 ปี ตั้งแต่วันที่ 1 มกราคม 2556 ถึง 31 ธันวาคม 2556 และนำเครื่องมือซอฟต์แวร์ SQL Server 2008 มาจัดการกับข้อมูล ทำให้ได้ตัวแปรตามหรือผลลัพธ์ที่ได้จากการวิจัย ได้แก่ คลังข้อมูลล็อกไฟล์ของการเข้าใช้งานเว็บไซต์ รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป และผลจากการทำเหมืองข้อมูลด้วยเทคนิคต่างๆ ที่สอดคล้องกับกลุ่มผู้ใช้งานเว็บไซต์ และได้ความรู้ใหม่ๆ ที่เกี่ยวข้องกับการเยี่ยมชมเว็บไซต์ของมหาวิทยาลัย

5. นิยามศัพท์

5.1 เหมืองข้อมูล (Data Mining) หมายถึง กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น และเป็นเทคนิคเพื่อค้นหารูปแบบของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูล (โดยส่วนใหญ่

จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

5.2 เหมืองข้อมูลเว็บ (Web Mining) หมายถึง การใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาและสกัดข้อมูลและสารสนเทศจากเอกสารเว็บและบริการบนเว็บโดยอัตโนมัติ เพื่อนำความรู้ที่ได้มาแก้ปัญหาที่ต้องการทั้งทางตรงและทางอ้อม

5.3 เว็บไซต์ คือหน้าเว็บเพจหลายหน้า ซึ่งเชื่อมโยงกันผ่านทางไฮเปอร์ลิงก์ ส่วนใหญ่จัดทำขึ้นเพื่อนำเสนอข้อมูลผ่านคอมพิวเตอร์ โดยถูกจัดเก็บไว้ในเวิลด์ไวด์เว็บ หน้าแรกของเว็บไซต์ที่เก็บไว้ที่ชื่อหลักจะเรียกว่า โฮมเพจ เว็บไซต์โดยทั่วไปจะให้บริการต่อผู้ใช้ฟรี แต่ในขณะเดียวกันบางเว็บไซต์จำเป็นต้องมีการสมัครสมาชิกและเสียค่าบริการเพื่อที่จะดูข้อมูล ในเว็บไซต์นั้น ซึ่งได้แก่ข้อมูลทางวิชาการ ข้อมูลตลาดหลักทรัพย์ หรือข้อมูลสื่อต่างๆ ผู้ทำเว็บไซต์มีหลากหลายระดับ ตั้งแต่สร้างเว็บไซต์ส่วนตัว จนถึงระดับเว็บไซต์สำหรับธุรกิจหรือองค์กรต่างๆ การเรียกดูเว็บไซต์โดยทั่วไปนิยมเรียกดูผ่านซอฟต์แวร์ในลักษณะของ เว็บเบราว์เซอร์

5.4 ล็อกไฟล์ คือ ข้อมูลจราจรคอมพิวเตอร์ เป็นข้อมูลเกี่ยวกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ แสดงถึงแหล่งกำเนิด ต้นทาง ปลายทาง เส้นทาง เวลา วันที่ ปริมาณ ระยะเวลา ชนิดของบริการ หรืออื่นๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์

5.5 ความรู้ คือความรู้ที่แฝงอยู่ในข้อมูลเป็นสิ่งที่สามารถสกัดจากสารสนเทศที่มีรูปแบบน่าสนใจเป็นจริงสาหรับข้อมูลใหม่หรือข้อมูลที่ไม่เคยเห็นมาก่อนเป็นรูปแบบใหม่ที่มนุษย์ไม่เคยเห็นมาก่อนซึ่งผลลัพธ์สุดท้ายจากการวิเคราะห์สารสนเทศจะได้เป็นความรู้ที่เป็นประโยชน์ต่อผู้ใช้ได้

6. ประโยชน์ที่ได้จากการวิจัย

6.1 ได้คลังข้อมูลการเข้าใช้งานเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง

6.2 สามารถจำแนกกลุ่มของการใช้จากการค้นหาความสัมพันธ์ เพื่อนำผลลัพธ์ไปปรับปรุงแก้ไขโครงสร้างเว็บไซต์ให้สอดคล้องกับกลุ่มผู้ใช้งานกลุ่มต่างๆ

6.3 สามารถหาความสัมพันธ์ของการเชื่อมโยงลิงค์หน้าเว็บต่างๆ ภายในเว็บไซต์ของมหาวิทยาลัย

6.4 สามารถจำแนกผู้ใช้แต่ละกลุ่มตามความสัมพันธ์ของการเชื่อมโยงลิงค์หน้าเว็บต่างๆ

6.5 สามารถพยากรณ์จำนวนผู้เข้าใช้งานเว็บไซต์ในอนาคต

บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

การวิจัยครั้งนี้ผู้วิจัยได้ศึกษาค้นคว้าข้อมูลที่เกี่ยวข้องทั้งในส่วนของทฤษฎีและงานวิจัยในและต่างประเทศจากแหล่งข้อมูลต่างๆ ได้แก่ ห้องสมุด ระบบอินเทอร์เน็ต โดยมีรายละเอียด ดังนี้

1. แนวคิดเกี่ยวกับเว็บไซต์ และล็อกไฟล์ (Concept of website and log file)
2. คลังข้อมูล (Data warehouse)
3. การทำเหมืองข้อมูล
4. ซอฟต์แวร์พัฒนาเหมืองข้อมูล
5. งานวิจัยที่เกี่ยวข้อง

เทคนิคการทำเหมืองข้อมูลเว็บเพื่อจำแนกกลุ่มผู้ใช้งานเว็บไซต์ กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปางจำเป็นอย่างยิ่งที่ต้องศึกษาหาข้อมูลในส่วนของทฤษฎีและผลงานทางวิชาการต่างๆ ที่เกี่ยวข้อง ซึ่งมีรายละเอียดดังนี้

1. แนวคิดทั่วไปเกี่ยวกับเว็บไซต์ และล็อกไฟล์

1.1 แนวคิดทั่วไปเกี่ยวกับเว็บไซต์

1.1.1 เว็บไซต์ หมายถึง หน้าเว็บเพจหลายหน้า ซึ่งเชื่อมโยงกันผ่านทางไฮเปอร์ลิงก์ ส่วนใหญ่จัดทำขึ้นเพื่อนำเสนอข้อมูลผ่านคอมพิวเตอร์โดยถูกจัดเก็บไว้ในเวิลด์ไวด์เว็บ หน้าแรกของเว็บไซต์ที่เก็บไว้ที่ชื่อหลักจะเรียกว่า โฮมเพจ เว็บไซต์โดยทั่วไปจะให้บริการต่อผู้ใช้ฟรีแต่ในขณะเดียวกันบางเว็บไซต์จำเป็นต้องมีการสมัครสมาชิกและเสียค่าบริการเพื่อที่จะดูข้อมูล ในเว็บไซต์นั้น ซึ่งได้แก่ข้อมูลทางวิชาการ ข้อมูลตลาดหลักทรัพย์ หรือข้อมูลสื่อต่างๆ ผู้ทำเว็บไซต์มีหลากหลายระดับ ตั้งแต่สร้างเว็บไซต์ส่วนตัว จนถึงระดับเว็บไซต์สำหรับธุรกิจหรือองค์กรต่างๆ การเรียกดูเว็บไซต์โดยทั่วไปนิยมเรียกดูผ่านซอฟต์แวร์ในลักษณะของเว็บเบราว์เซอร์ (ที่มา เว็บไซต์ http://th.wikipedia.org/wiki/เว็บไซต์ ,2013)

1.1.2 โครงสร้างของเว็บไซต์ อนงค์ หลอดแก้วได้กล่าวว่า โครงสร้างเว็บไซต์ที่ดีจะช่วยให้ผู้ชมไม่สับสน และค้นหาข้อมูลที่ต้องการได้อย่างรวดเร็ว ไม่ควรเป็นลำดับที่ลึกหลายชั้นเกินไป เพราะผู้ใช้จะเบื่อเสียก่อน กว่าจะค้นหาเจอหน้าที่ต้องการและมีส่วนประกอบของหน้าเว็บเพจแบ่งออกเป็น 3 ส่วนหลักๆ คือ

1) ส่วนหัวของเว็บเพ็จ (Page Header) เป็นส่วนที่อยู่ตอนบนสุดของหน้าและเป็นส่วนที่สำคัญที่สุดของหน้า เพราะเป็นส่วนที่ดึงดูดผู้ชมให้ติดตามเนื้อหาภายในเว็บไซต์ มักใส่ภาพกราฟฟิกเพื่อสร้างความประทับใจ ส่วนใหญ่ประกอบด้วย

(1) โลโก้ (Logo) เป็นสิ่งที่เว็บไซต์ควรมี เป็นตัวแทนของเว็บไซต์ได้เป็นอย่างดี และยังทำให้เว็บน่าเชื่อถือ

(2) ชื่อเว็บไซต์

*(3) เมนูหลัก*หรือลิงค์เป็นจุดเชื่อมโยงไปสู่เนื้อหาของเว็บไซต์

2) ส่วนของเนื้อหา (Page Body) เป็นส่วนที่อยู่ตอนกลางของหน้าใช้แสดงข้อมูลเนื้อหาของเว็บไซต์ ซึ่งประกอบด้วยข้อความ, ตารางข้อมูล ภาพกราฟิก วีดีโอ และอื่นๆ และอาจมีเมนูหลัก หรือเมนูเฉพาะกลุ่มวางอยู่ในส่วนนี้ด้วย สำหรับส่วนเนื้อหาควรแสดงใจความสำคัญที่เป็นหัวเรื่องไว้บนสุด ข้อมูลมีความกระชับ ใช้รูปแบบตัวอักษรที่อ่านง่าย และจัด Layout ให้เหมาะสมและเป็นระเบียบ

3) ส่วนท้ายของหน้า (Page Footer) เป็นส่วนที่อยู่ด้านล่างสุดของหน้าจะมีหรือไม่มีก็ได้ มักวางระบบนำทางที่เป็นลิงค์ข้อความง่ายๆ และอาจแสดงข้อมูลเพิ่มเติมเกี่ยวกับเนื้อหาภายในเว็บไซต์ เช่น เจ้าของเว็บไซต์, ข้อความแสดงลิขสิทธิ์, วิธีการติดต่อกับผู้ดูแลเว็บไซต์, คำแนะนำการใช้เว็บไซต์ เป็นต้น

1.1.3 ประโยชน์ของเว็บไซต์ เว็บไซต์เป็นการให้บริการบนเครือข่ายอินเตอร์เน็ต ซึ่งมีประโยชน์ในหลายๆ ด้าน อาทิ 1) ช่วยส่งเสริมศักยภาพการแข่งขันในด้านธุรกิจ 2) ช่วยเผยแพร่ข้อมูลข่าวสารและบริการต่างๆ ให้เป็นที่รู้จักอย่างแพร่หลาย 3) ช่วยอำนวยความสะดวกให้กับลูกค้าในการให้บริการรูปแบบแบบออนไลน์ เป็นต้น

1.2 แนวคิดทั่วไปเกี่ยวกับล็อกไฟล์

1.2.1 ความหมายของล็อกไฟล์ พระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ กล่าวว่า “ข้อมูลจราจรทางคอมพิวเตอร์” หมายความว่า ข้อมูลเกี่ยวกับการติดต่อสื่อสารของระบบคอมพิวเตอร์ ซึ่งแสดงถึงแหล่งกำเนิด ต้นทาง ปลายทาง เส้นทาง เวลา วันที่ ปริมาณ ระยะเวลา ชนิดของบริการ หรืออื่นๆ ที่เกี่ยวข้องกับการติดต่อสื่อสารของระบบคอมพิวเตอร์นั้น

1.2.2 ที่ตั้งของล็อกไฟล์ L.K. Joshila Grace1 และคณะ (2011) ได้กล่าวไว้ในบทความงานวิจัยเรื่อง “ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING” ซึ่งมีการตีพิมพ์ในวารสาร “International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011” กล่าวว่า ล็อกไฟล์เป็นไฟล์ที่ถูกเขียนข้อมูลตามที่ผู้ใช้เข้าใช้งานหน้าเว็บไซต์หรือแหล่งข้อมูลต่างๆ ภายในเว็บไซต์ ซึ่งมีการเก็บข้อมูลแต่ละที่ตั้ง ดังนี้

1) Web server ล็อกไฟล์ที่อยู่ในเว็บเซิร์ฟเวอร์บันทึกกิจกรรมของผู้ใช้ ที่เข้าชมเว็บไซต์ผ่านเบราว์เซอร์

2) Web Pro Servers คือเซิร์ฟเวอร์กลางที่มีอยู่ระหว่างผู้ใช้และเว็บเซิร์ฟเวอร์ก่อนมีถ้าเว็บเซิร์ฟเวอร์ที่ได้รับการร้องขอของผู้ใช้ผ่านทางเซิร์ฟเวอร์พร็อกซี่แล้วรายการไปยังแฟ้มบันทึกจะเป็นข้อมูลของเซิร์ฟเวอร์พร็อกซี่

3) Client browsers ล็อกไฟล์จะถูกเก็บไว้ในโปรแกรมเบราว์เซอร์ของผู้ใช้ ซึ่งแล้วแต่ความสามารถในการจัดเก็บข้อมูลของแต่ละโปรแกรม อาทิ Google chrome , Internet Explorer และจะบันทึกข้อมูลการเข้าใช้เว็บไซต์ไว้ในเครื่องคอมพิวเตอร์ของผู้ใช้นั้นเอง

1.2.3 เนื้อหาของล็อกไฟล์ ล็อกไฟล์ในแต่ละ Server มีความแตกต่างการขึ้นอยู่กับข้อมูลพื้นที่ที่สามารถกำหนด หรือผู้ดูแลระบบได้กำหนดข้อมูลเบื้องต้นที่ต้องการ ซึ่งมีรายละเอียดดังนี้

1) Username จะระบุผู้เข้าชมเว็บไซต์ ส่วนมากมักจะเก็บในรูปแบบของ IP address ที่กำหนดโดยผู้ให้บริการอินเทอร์เน็ต (ISP) หรือบางที่จะเก็บในรูปแบบของชื่อผู้ใช้งานระบบ จะทำให้ทราบถึงรายละเอียดของผู้ใช้ ขึ้นอยู่กับระบบของเว็บไซต์ที่ต้องการจัดเก็บ username ตามประเภทที่ต้องการ

2) Visiting Path เส้นทางที่ถ่ายโดยผู้ใช้ในขณะที่การเยี่ยมชมเว็บไซต์ นี่อาจจะเป็นโดยใช้ URL โดยตรงหรือโดยการคลิกที่ลิงค์

3) Path Traversed จะระบุเส้นทางที่ถ่ายโดยผู้ใช้ที่มีอยู่ในเว็บไซต์ที่ใช้เชื่อมโยงต่างๆ

4) Time stamp เวลาที่ใช้โดยผู้ใช้ในแต่ละหน้าเว็บในขณะที่ท่องผ่านทางเว็บสถาน นี้ถูกระบุว่าเป็นเซสชั่น

5) Page last visited หน้าเว็บที่มีการเข้าชมโดยผู้ใช้ก่อนที่เขาหรือเธอออกจากเว็บ

6) Success rate อัตราความสำเร็จของเว็บไซต์จะถูกกำหนดโดยจำนวนของการดาวน์โหลดและการทำกิจกรรมของผู้ใช้

7) User Agent เบราว์เซอร์จากที่ผู้ใช้จะส่งคำขอไปยังเว็บเซิร์ฟเวอร์ จะบอกข้อมูลชนิดเบราว์เซอร์ของและรุ่นของเบราว์เซอร์ซอฟแวร์ถูกนำมาใช้

8) URL แหล่งทรัพยากรหรือแหล่งข้อมูลที่เข้าถึงโดยผู้ใช้ เช่น หน้า HTML

9) Request type วิธีที่ใช้สำหรับการถ่ายโอนข้อมูลที่ถูกตั้งข้อสังเกตว่าวิธีการเช่น GET,POST

1.2.4 ประเภทล็อกไฟล์ที่เก็บใน Web server จะบันทึกเป็นไฟล์ข้อความธรรมดา (ASCII) และเป็นอิสระจากเซิร์ฟเวอร์ มีความแตกต่างบางอย่างระหว่างซอฟต์แวร์เซิร์ฟเวอร์ มี 4 ประเภทคือ

1) Access log file คือข้อมูลที่มีการร้องขอเข้ามาทั้งหมดและข้อมูลเกี่ยวกับลูกค้าของเซิร์ฟเวอร์ เข้าบันทึกเข้าสู่ระบบการร้องขอทั้งหมดที่มีการประมวลผลโดยเซิร์ฟเวอร์

2) Agent Log คือข้อมูลเกี่ยวกับเบราว์เซอร์ของผู้ใช้ อาทิ เบราว์เซอร์ , รุ่น

3) Error Log คือรายการของข้อผิดพลาดภายใน เมื่อใดก็ตามที่มีข้อผิดพลาดหน้ามีการร้องขอจากผู้ใช้ไปยังเว็บเซิร์ฟเวอร์ รายการจะทำในบันทึกข้อผิดพลาด แล้วทำการบันทึกการเข้าถึงและข้อผิดพลาดไปที่เซิร์ฟเวอร์

4) Referrer Log คือไฟล์ที่ข้อมูลเกี่ยวกับการเชื่อมโยงและเปลี่ยนเส้นทางของผู้เข้าชมเว็บไซต์

1.2.5 รูปแบบของล็อกไฟล์ ล็อกไฟล์เป็นแฟ้มข้อความธรรมดาซึ่งบันทึกข้อมูลเกี่ยวกับผู้ใช้แต่ละคน การแสดงผลของข้อมูลล็อกไฟล์ในรูปแบบที่แตกต่างกัน มี 3 รูปแบบดังนี้

1) Extended log file format รูปแบบ W3C เป็นค่าเริ่มต้นรูปแบบแฟ้มบันทึกบน IIS เซิร์ฟเวอร์ ข้อมูลจะถูกคั่นด้วยช่องว่าง, เวลาบันทึกเป็น GMT และสามารถปรับแต่งรูปแบบได้ โดยดูแลระบบสามารถเพิ่มหรือเอาเขตข้อมูลขึ้นอยู่กับข้อมูลที่ต้องการบันทึก ในรูปแบบของ W3C ประกอบด้วย

(1) Software ชื่อของซอฟแวร์

(2) Version รุ่นของ IIS ที่กำลังทำงานอยู่

(3) Date บันทึกวันที่เวลาของรายการที่บันทึกไว้เป็นครั้งแรก

(4) Fields นี้ไม่ได้เป็นรูปแบบมาตรฐานเพราะผู้ดูแลระบบสามารถปรับแต่งในรูปแบบนี้ฟิลด์นี้ อาทิ วันเวลา ที่อยู่ IP, วิธีการส่งค่า URI, Browsers, รุ่นโปรโตคอล

```
#Software: Microsoft Internet Information Services 6.0
#Version: 1.0
#Date: 2002-05-02 17:42:15
#Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status cs(User-Agent)
2002-05-02 17:42:15 172.22.255.255 - 172.30.255.255 80 GET /images/picture.jpg - 200 Mozilla/4.0+(compatible;MSIE+5.5;+Windows+2000+Server)
```

ภาพที่ 2.1 แสดงตัวอย่างล็อกไฟล์รูปแบบ Extended log file format

ที่มา : https://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/ffdd7079-47be-4277-921f-7a3a6e610dcb.mspx?mfr=true

2) NCSA common log file format มีการบันทึกข้อมูลพื้นฐานเกี่ยวกับผู้ขอ เช่น ชื่อผู้ใช้และชื่อโฮสต์ระยะไกล, วันเวลา, ชนิดขอรหัสสถานะ HTTP และตัวเลขของไบต์ส่งโดยเซิร์ฟเวอร์ NCSA การแก้ไขรูปแบบไม่สามารถกำหนดเองได้ และสามารถใช้ได้สำหรับเว็บไซต์ การกำหนดรูปแบบของปีเป็น DD / MMM / YYYY เขตข้อมูลจะถูกคั่นด้วยช่องว่าง, เวลาขณะนี้ตามเวลาท้องถิ่น

```
10.3.5.1 -  Nwtraders.com  [18/Aug/2006:13:17:37 - 0800] "MAIL FROM -? FROM: someone HTTP/1.0"  250 0
10.3.5.1 - Nwtraders.com[18/Aug/2006:13:17:37 - 0800] "RCPT TO -? TO: someone@example.com HTTP/1.0" 250 0
10.3.5.1 - Nwtraders.com[18/Aug/2006:13:17:37 - 0800] "DATA -?000B72730441764MAIL01 HTTP/1.0" 250 97
10.3.5.1 - Nwtraders.com[18/Aug/2006:13:17:37 - 0800] "QUIT -?someone HTTP/1.0" 0 0
```

ภาพที่ 2.2 แสดงตัวอย่างล็อกไฟล์รูปแบบ NCSA common log file format

ที่มา: http://winintro.ru/mail.en/html/3581adb1-c526-4169-b2d8-1d46c1611c34.htm

3) IIS log file format รูปแบบ IIS ไม่ได้รับการปรับแต่งเป็นรูปแบบ ASCII คงที่ เขตข้อมูลจะถูกคั่นด้วยเครื่องหมายจุลภาคและง่ายต่อการอ่าน เวลาที่บันทึกไว้ในเวลาท้องถิ่น การบันทึกข้อมูลจะมีการบันทึกมากกว่ารูปแบบ NCSA ฟิลด์ใน IIS จะประกอบไปด้วย ที่อยู่ IP เครื่อง

ไคลเอนต์,ชื่อผู้ใช้,วันและเวลา,การให้บริการ,ชื่อเซิร์ฟเวอร์,ที่อยู่ IP ของเซิร์ฟเวอร์,เวลา, การส่งข้อมูลไบต์ของเครื่องไคลเอนต์, การส่งข้อมูลไบต์ของเครื่องเซิร์ฟเวอร์, รหัสสถานะบริการ (ค่าของ 200 บ่งชี้ว่าการร้องขอก็สำเร็จที่ประสบความสำเร็จ), รหัสสถานะของ Windows (ค่า 0 บ่งชี้ว่าการร้องขอก็สำเร็จที่ประสบความสำเร็จ), ชนิดของการร้องขอ, เป้าหมายของการดำเนินงาน, พารามิเตอร์

```
192.168.114.201, -, 03/20/01, 7:55:20, W3SVC2, SALES1, 172.21.13.45, 4502, 163,
3223, 200, 0, GET, /DeptLogo.gif, -,

172.16.255.255, anonymous, 03/20/01, 23:58:11, MSFTPSVC, SALES1,
172.16.255.255, 60, 275, 0, 0, 0, PASS, /Intro.htm, -,
```

ภาพที่2.3 แสดงตัวอย่างล็อกไฟล์รูปแบบ IIS log file format
ที่มา: http://msdn.microsoft.com/en-us/library/ms525807(v=vs.90).aspx

1.2.6 สถานะโค้ดการส่งข้อมูลโดยเซิร์ฟเวอร์

เว็บไซต์ http://en.wikipedia.org/wiki/List_of_HTTP_status_codes กล่าวว่า รหัสสถานภาพในการตอบรับเอชทีทีพีจากเครื่องให้บริการ ซึ่งมีทั้งรหัสที่กำหนดโดยมาตรฐานอินเทอร์เน็ตของคณะทำงานเฉพาะกิจด้านวิศวกรรมอินเทอร์เน็ต (IETF) และกำหนดโดยเอกสารขอความเห็น (RFC) เอกสารลักษณะเฉพาะอื่นๆ และรหัสที่มีการใช้งานโดยทั่วไปเพิ่มเข้ามา ตัวเลขแรกของรหัสสถานภาพ (หลักร้อย) เป็นตัวระบุประเภทของการตอบรับหนึ่งในห้าประเภท ซึ่งเครื่องลูกข่ายเอชทีทีพีสามารถรับรู้ประเภททั้งห้านี้ได้เป็นอย่างน้อย ซึ่งรหัสสถานภาพของเอชทีทีพีประกอบไปด้วย

1) *1xx* เป็น code ที่ใช้บอกถึงข้อมูลทั่วไป อาทิ เหตุการณ์ต่างๆ ที่เกิดขึ้นในการสื่อสารระหว่าง Client และ Server
2) *2xx* เป็น code ที่บ่งบอกว่าการร้องขอสำเร็จ
3) *3xx* เป็น code ที่บ่งบอกการเปลี่ยนทาง
4) *4xx* เป็น code ที่บ่งบอกถึงความผิดพลาดที่เกิดขึ้นจากส่วนของ Client
5) *5xx*เป็น code ที่บ่งบอกถึงความผิดพลาดที่เกิดขึ้นจากส่วนของ Server

1.2.7 การเตรียมข้อมูลจากล็อกไฟล์ Priyanka Patil and Ujwala Patil (2012) ได้กล่าวว่า การเตรียมข้อมูลเป็นขั้นตอนสำคัญในการกรองข้อมูลละจัดระเบียบข้อมูลที่เหมาะสมในการทำเหมืองข้อมูล การนำข้อมูลจากล็อกไฟล์ไปใช้งานในการทำเหมืองข้อมูลต้องทำการเตรียมข้อมูลที่ต้องการใช้ และนำข้อมูลเหล่านั้น มาทำการแปลงข้อมูลเพื่อลดจำนวนข้อมูลขยะและกำหนดรูปแบบที่ต้องการ ซึ่งมีข้อมูลตามขั้นตอนต่อไปนี้

1) *Field Extraction* การสกัสฟิลด์ในล็อกไฟล์นั้นมีรายการที่บันทึกซึ่งมีข้อมูลต่างๆ ที่จำเป็นต้องแยกออกมาเพื่อสำหรับการประมวลผล กระบวนการแยกข้อมูลจากล็อกไฟล์ซึ่งมีรูปแบบการใช้ตัวอักษรที่แตกต่างกันในการคั่นระหว่างข้อมูล โดยส่วนมากจะใช้อักษร “,” ซึ่งสามารถใช้อัลกอริทึมในการสกัดข้อมูลดังนี้

2) Data cleansing ขั้นตอนนี้เป็นขั้นตอนที่ขจัดเนื้อหาที่ไม่พึ่งประสงค์ อาทิ มัลติมีเดียไฟล์ภาพ, ไฟล์สไตล์เพจ, ลบรหัสสถานะ หรือข้อมูลอื่นๆ ที่ไม่ต้องการ ซึ่งสามารถใช้อัลกอริทึมสำหรับการจัดการเนื้อหาดังต้อไปนี้

3) User identification ขั้นตอนนี้เป็นขั้นตอนการระบุตัวตนของผู้ใช้ โดยใช้ที่อยู่ IP ของแต่ละผู้ใช้ และถ้าหากที่อยู่ IP ของผู้ใช้เหมือนกันซึ่งผู้เข้าใช้แตกต่างกัน ก็สามารถแยกแยะการระบุตัวตนของผู้ใช้ นั้นโดยการตรวจสอบที่รุ่นเบราว์เซอร์หรือระบบปฏิบัติการที่มีความแตกต่างกัน

4) Session identification ขั้นตอนนี้เป็นการระบุเซสชั่นของผู้ใช้ในการเข้าเยี่ยมชมหน้าเว็บไซต์ในช่วงเวลาที่แตกต่างกัน

2. คลังข้อมูล

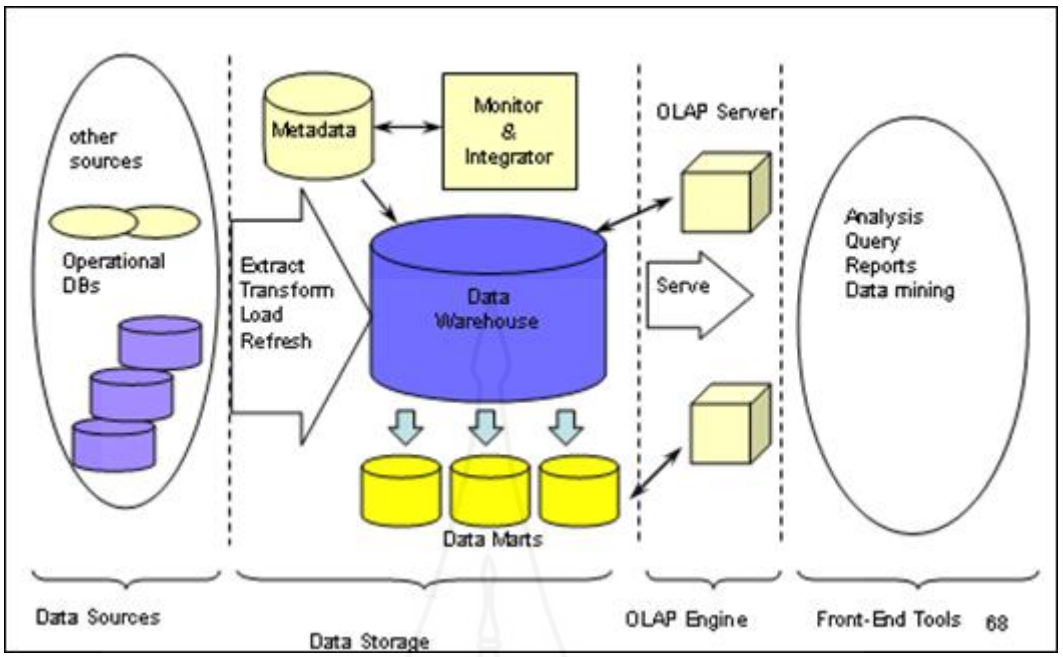
กิตติพงศ์ กลมกล่อม ได้เขียนหนังสือ “การออกแบบและพัฒนาคลังข้อมูล (Data warehouse)” และได้กล่าวว่า คลังข้อมูล คือ หลักการหรือวิธีการ ซึ่งมีที่มาจากการทำ Overall Business integration ที่สามารถช่วยให้การวิเคราะห์ข้อมูลเพื่อการตัดสินใจ (Decision Marking) เพื่อการบริหารงานในองค์กรเป็นไปได้อย่างมีประสิทธิภาพ และได้กล่าวว่า “การคลังข้อมูล (Data Warehousing)” เป็นศูนย์รวมของหลักการและวิธีการมากมายหลากหลาย อาทิเช่น การออกแบบและสร้างโครงสร้างของข้อมูลใน Data Warehouse วิธีการเพื่อได้มาซึ่งข้อมูล วิธีการสร้างผลลัพธ์จากข้อมูลที่มีรวมไปถึงวิธีการดูแลรักษาวิธีการปรับปรุงประสิทธิภาพ เป็นต้น

2.1 กระบวนการในการพัฒนาคลังข้อมูล

สุวรรณี อัศวกุลชัย (2552) ได้กล่าวว่า เนื่องจากข้อมูลพื้นฐานของฐานข้อมูลในคลังข้อมูลประกอบด้วยข้อมูลจำนวนมากจะต้องมีการออกแบบคลังข้อมูลเพื่อรวบรวมข้อมูลและวิเคราะห์หารูปแบบข้อมูลในคลังข้อมูล โดยกล่าวถึงกระบวนการในการพัฒนาคลังข้อมูล และการทำงานข้องข้อมูล ประกอบด้วยขั้นตอนต่างๆ ประกอบด้วย 1) การศึกษากระบวนการทางธุรกิจ 2) การศึกษาความต้องการของผู้ใช้ 3) การพัฒนาแบบจำลอง 4) กระบวนการออกแบบการดึงข้อมูล 5) การศึกษาวิเคราะห์ ออกแบบ และจัดทำระบบคลังข้อมูลกลาง 6) การแสดงรายงาน 7) การทดสอบคลังข้อมูล

2.2 องค์ประกอบหลักของคลังข้อมูล

สุวรรณี อัศวกุลชัย (2552) ได้กล่าวว่าองค์ประกอบของคลังข้อมูลประกอบด้วย 4 ส่วนหลักๆ ได้แก่ แหล่งข้อมูล ส่วนเก็บข้อมูล ส่วนของกลไกโอแลป และส่วนเครื่องมือผู้ใช้



ภาพที่ 2.4 แสดงองค์ประกอบของคลังข้อมูล
ที่มา: http://www.no-poor.com/dssandos/Chapter3-dss.htm

2.2.1 แหล่งข้อมูล เป็นส่วนการเตรียมความพร้อมก่อนนำเข้าสู่คลังข้อมูล โดยการรวบรวมข้อมูล จากฐานข้อมูลหลายแหล่ง รวมถึงข้อมูลที่ได้จากการประมวลผลรายงานทรายแซกชัน ซึ่งเป็นการเพิ่ม ลด ปรับปรุง และเรียกดูข้อมูลจากฐานข้อมูล ถ้าระบบมีผู้ใช้ระบบจำนวนมาก และใช้งานพร้อมกันจะเรียกว่า การประมวลธุรกรรมเชิงรายการแบบออนไลน์หรือโอแอลทีพี

2.2.2 ส่วนเก็บข้อมูล เป็นส่วนการทำความสะอาด โดยการดำเนินการกับข้อมูลที่เรียกว่าอีทีแอล ซึ่งเป็นกระบวนการที่ชาวย้ายข้อมูลจากหลายแหล่งที่มีรูปแบบที่แตกต่างกันให้อยู่ในรูปแบบใหม่ ในพื้นที่สำหรับเก็บข้อมูลต่างๆ โดยความคุมการทำงานและควบคุมข้อมูลในคลังข้อมูล ข้อมูลที่ผ่านกระบวนการในข้างต้นเรียบร้อยแล้วก็จะจัดเก็บในคลังข้อมูล

2.2.3 ส่วนของกลไกโอแลปการเก็บบันทึกข้อมูลและผลลัพธ์ต่างๆ ที่ได้จากการประมวลผลเชิงวิเคราะห์หรือโอแลปเมื่อประมวลผลจากคลังข้อมูลแล้ว จะดึงมาเก็บไว้ที่เซร์ฟเวอร์ของโอแลป

2.2.4 ส่วนเครื่องมือสำหรับผู้ใช้งาน มีเครื่องมือหลากหลายรูปแบบที่ช่วยในการแสดงผลในรูปแบบต่าง ๆ ดังนี้

1) แสดงผลการวิเคราะห์ต่างๆ เช่น การวิเคราะห์ความต้องการลูกค้า เพื่อย้ายการสั่งผลิตสินค้าในปีหน้า เป็นต้น

2) แสดงการสอบถามข้อมูล เช่นการสอบถามข้อมูลรายชื่อลูกค้าที่มียอดซื้อสูงสุดในเดือนนี้ เพื่อมองรางวัลลูกค้าดีเด่นเป็นต้น

3) แสดงรายงานต่างๆ เช่น เป็นลักษณะกราฟหรือตารางหลายมิติ เป็นต้น

4) เครื่องมือสำหรับการทำเหมืองข้อมูล

2.3 สถาปัตยกรรมคลังข้อมูล (Data Warehouse Architecture - DWA)

เลิศ เลิศศิริโสภณ (2541) ได้กล่าวว่า สถาปัตยกรรมคลังข้อมูลเป็นโครงสร้างมาตรฐานที่ใช้อธิบาย เพื่อให้เข้าใจแนวคิด และกระบวนการของคลังข้อมูลนั้นๆ ซึ่งโดยทั่วไปแล้วคลังข้อมูลแต่ละระบบอาจจะมีรูปแบบที่ไม่เหมือนกันได้ เพื่อให้เหมาะสมกับองค์กรนั้นๆ ทั้งนี้ส่วนประกอบต่างๆ ภายใน DWAที่สำคัญ ได้แก่

2.3.1 Operational database หรือ external database layer ทำหน้าที่จัดการกับข้อมูลในระบบงานปฏิบัติการหรือแหล่งข้อมูลภายนอกองค์กร

2.3.2 Information access layer เป็นส่วนที่ผู้ใช้ปลายทางติดต่อผ่านโดยตรง ประกอบด้วยฮาร์ดแวร์และซอฟต์แวร์ ที่ใช้ในการแสดงผลเพื่อการวิเคราะห์ โดยมีเครื่องมือช่วย เป็นตัวกลางที่ผู้ใช้ใช้ติดต่อกับคลังข้อมูล โดยในปัจจุบันเครื่องมือที่ได้รับความนิยมเพิ่มขึ้นอย่างรวดเร็วนั่นคือ Online Analytical Processing Tool หรือ OLAP tool ซึ่งเป็นเครื่องมือที่มีความสามารถในการวิเคราะห์ที่ซับซ้อน และแสดงข้อมูลในรูปแบบหลายมิติ

2.3.3 Data access layer เป็นส่วนต่อประสานระหว่าง Information access layer กับ operational layer

2.3.4 Data director (metadata) layer เพื่อให้เข้าถึงข้อมูลได้ง่ายขึ้น และเป็นการเพิ่มความเร็วในการเรียกและดึงข้อมูลของคลังข้อมูล

2.3.5 Process management layer ทำหน้าที่จัดการกระบวนการทำงานทั้งหมด

2.3.6 Application messaging layer เป็นมิดเดิลแวร์ทำหน้าที่ในการส่งข้อมูลภายในองค์กรผ่านทางเครือข่าย

2.3.7 Data warehouse (physical) layer เป็นแหล่งเก็บข้อมูลของทั้ง information data และ external dataในรูปแบบที่ง่ายแก่การเข้าถึงและยืดหยุ่นได้

2.3.8 Data staging layer เป็นกระบวนการแก้ไข และดึงข้อมูลจาก external database

2.4 วิธีการออกแบบคลังข้อมูล

ในปี 1996Ralph Kimball ได้เสนอวิธีการออกแบบฐานข้อมูลสำหรับจัดเก็บคลังข้อมูล เรียกว่าระเบียบวิธี 9 ขั้น หรือ Nine-Step Methodology (Connolly, 2002) โดยวิธีการนี้เริ่มจากการออกแบบจากส่วนย่อยที่แสดงถึงแต่ละระบบงานขององค์กร หรือเรียกอีกอย่างหนึ่งว่าดาต้ามาร์ท (data mart) โดยเมื่อออกแบบแต่ละส่วนสำเร็จแล้ว จึงนำมารวมกันเป็นคลังข้อมูลขององค์กรในขั้นสุดท้าย ซึ่งขั้นตอนทั้ง 9 ขั้นตอนมีรายละเอียดดังนี้

2.4.1 กำหนดดาต้ามาร์ท คือการเลือกว่าจะสร้างดาต้ามาร์ทของระบบงานใดบ้าง และระบบงานใดเป็นระบบงานแรกโดยองค์กรจะต้องสร้าง E-R model ที่รวมระบบงานทุกระบบขององค์กรไว้ แสดงการเชื่อมโยงของแต่ละระบบงานอย่างชัดเจน และสิ่งที่ต้องคำนึงถึงในการเลือกระบบงานที่จะเป็นดาต้ามาร์ทแรกนั้น มี 3 ปัจจัยที่เกี่ยวข้อง ได้แก่ จะต้องสามารถพัฒนาออกมาได้ทันตามเวลาที่ต้องการ โดยอยู่ในงบประมาณที่กำหนดไว้และต้องตอบปัญหาทางธุรกิจให้แก่องค์กรได้ ดังนั้นดาต้ามาร์ทแรกควรจะเป็นของระบบงานที่นำรายได้เข้ามาสู่องค์กรได้ เช่น ระบบงานขาย เป็นต้น

2.4.2 กำหนดfact table ของดาต้ามาร์ท คือการกำหนดเนื้อหาหลักที่ควรจะเป็นของดาต้ามาร์ท โดยการเลือกเอนทิตีหลักและกระบวนการที่เกี่ยวกับเอนทิตีนั้นๆออกมาจาก E-R model ขององค์กร นั่นหมายถึงว่าจะทำให้เราทราบถึง dimension table ที่ควรจะมีด้วย

2.4.3 กำหนดแอตทริบิวต์ที่จำเป็นในแต่ละ dimension table คือการกำหนดแอตทริบิวต์ที่บอกหรืออธิบายรายละเอียดของ dimension ได้ ทั้งนี้แอตทริบิวต์ที่เป็น primary key ควรเป็นค่าที่คำนวณได้ กรณีที่มีดาต้ามาร์ทมากกว่าหนึ่งดาต้ามาร์ทมี dimension เหมือนกัน นั่นหมายถึงว่าแอตทริบิวต์ใน dimension นั้นจะต้องเหมือนกันทุกประการ แต่นั่นก็ไม่อาจจะแก้ไขปัญหาการจัดเก็บข้อมูลซ้ำซ้อน อันจะนำมาสู่ความแตกต่างกันของข้อมูลชุดเดียวกัน ปัญหานี้จึงเป็นการดีที่จะมีการใช้ dimension table ร่วมกันในแต่ละ fact table ที่จำเป็นต้องมี dimension ดังกล่าว โดยเรียก dimension table ลักษณะแบบนี้ว่า conformed และเรียก fact tableว่า fact constellation เราสามารถกำหนดข้อดีของการใช้ dimension table ร่วมกันได้ดังนี้

1) แน่ใจได้ว่าในแต่ละรายงานจะออกมาสอดคล้องกัน
2) สามารถสร้างดาต้ามาร์ทในเวลาต่างๆ กันได้
3) สามารถเข้าถึงดาต้ามาร์ทโดยผู้พัฒนากลุ่มอื่นๆ
4) สามารถรวบรวมดาต้ามาร์ทหลายๆ อันเข้าด้วยกัน
5) สามารถออกแบบคลังข้อมูลร่วมกันได้

2.4.4 กำหนดแอตทริบิวต์ที่จำเป็นใน fact table โดยแอตทริบิวต์หลักใน fact table จะมาจาก primary keyในแต่ละ dimension table นอกจากนี้แล้ว ยังสามารถมีแอตทริบิวต์ที่จำเป็นอื่นๆ ประกอบอยู่ด้วย เช่นแอตทริบิวต์ที่ได้จากการคำนวณค่าเบื้องต้นที่จำเป็นสำหรับการคงอยู่ของแอตทริบิวต์อื่นใน fact tableเรียกอีกอย่างหนึ่งว่า measure การกำหนดแอตทริบิวต์นี้ไม่ควรจะเลือกแอตทริบิวต์ที่คำนวณค่าไม่ได้ เช่นเป็นตัวหนังสือหรือไม่ใช่ตัวเลข เป็นต้น และไม่ควรเลือกแอตทริบิวต์ที่ไม่เกี่ยวข้องกับเนื้อหาของ fact tableที่เราสนใจด้วย

2.4.5 จัดเก็บค่าการคำนวณเบื้องต้นใน fact table คือการจัดเก็บค่าที่ได้จากการคำนวณให้เป็นแอตทริบิวต์หนึ่งใน fact table ถึงแม้ว่าจะสามารถหาค่าได้จากแอตทริบิวต์อื่นๆ ก็ตาม ทั้งนี้เพื่อให้การสอบถามมีประสิทธิภาพมากขึ้น สามารถทำงานด้วยความเร็วที่เพิ่มขึ้นเนื่องจากไม่ต้องคำนวณค่าใหม่ทั้งหมด ถึงแม้ว่าจะเกิดความซ้ำซ้อนของข้อมูลในการจัดเก็บบ้างก็ตาม

2.4.6 เขียนคำอธิบายของ dimension table ทั้งนี้ก็เพื่อให้ผู้ใช้สามารถใช้งานดาต้ามาร์ทได้อย่างมีประสิทธิภาพเพราะเกิดความเข้าใจอย่างดีในส่วนต่างๆ

2.4.7 กำหนดระยะเวลาในการจัดเก็บข้อมูลในฐานข้อมูล โดยอาจจะเป็นการจัดเก็บเพียงช่วงระยะเวลา 1-2 ปี หรือนานกว่านั้น ขึ้นอยู่กับความต้องการขององค์กร เนื่องจากองค์กรแต่ละประเภทมีความต้องการในการจัดเก็บข้อมูลต่างช่วงเวลากัน ทั้งนี้ขึ้นอยู่กับความจำเป็นหรือข้อกำหนดในการดำเนินธุรกิจมีข้อสังเกตอยู่ 2 ประการที่น่าสนใจและสำคัญสำหรับการออกแบบแอตทริบิวต์ในเรื่องของการจัดเก็บข้อมูล ดังนี้

1) ข้อมูลที่ถูกจัดเก็บไว้นานเกินไปมักเกิดปัญหาการอ่าน หรือแปลข้อมูลนั้นๆ จากแฟ้มหรือเทปเก่า

2) เมื่อมีการนำรูปแบบเก่าของ dimension table มาใช้อาจเกิดปัญหา การเปลี่ยนแปลงของ dimensionอย่างช้าๆ ได้

2.4.8 การติดตามปัญหาการเปลี่ยนแปลงของ dimension คือ การเปลี่ยนแอตทริบิวต์ของ dimension table เก่ามาใช้แล้วส่งผลกระทบต่อข้อมูลปัจจุบันของ dimension table โดยสามารถแบ่งประเภทของปัญหาที่เกิดได้เป็น 3 ประเภท ดังนี้

1) เกิดการเขียนทับข้อมูลใหม่โดยข้อมูลเก่า
2) เกิดเรคอร์ดใหม่ๆ ขึ้นใน dimension
3) เกิดเรคอร์ดที่มีทั้งค่าเก่าและใหม่ปนกันไป

2.4.9 กำหนดคิวรีเป็นการออกแบบด้านกายภาพ เพื่อให้ผู้ใช้เกิดความสะดวกในการใช้งานและสามารถทำงานได้อย่างมีประสิทธิภาพเมื่อดำเนินการทั้ง 9 ขั้นตอนสำหรับแต่ละดาต้ามาร์ทเสร็จแล้ว จึงจะนำทั้งหมดมารวมกันเป็นภาพของคลังข้อมูลขององค์กรต่อไป

3. การทำเหมืองข้อมูล

อดุลย์ ยิ้มงาม (2551) ให้ความหมายของ การทำเหมืองข้อมูล คือ กระบวนการที่การทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น ในปัจจุบันการทำเหมืองข้อมูลได้ถูกนำไปประยุกต์ใช้ในงานหลายประเภท ทั้งในด้านธุรกิจที่ช่วยในการตัดสินใจของผู้บริหาร ในด้านวิทยาศาสตร์และการแพทย์รวมทั้งในด้านเศรษฐกิจและสังคม

Srivastava, et al., (2000) ให้ความหมายของ การทำเหมืองข้อมูล การใช้งานเว็บที่กำหนดเป็นกระบวนการของการใช้เทคนิคการทำเหมืองข้อมูลเพื่อการค้นพบรูปแบบการใช้ข้อมูลจากเว็บล็อกซึ่งจะระบุพฤติกรรมของผู้ใช้เว็บ การทำเหมืองข้อมูลการใช้งานเว็บเป็นประเภทของกิจกรรมการทำเหมืองเว็บที่เกี่ยวข้องกับการค้นหาโดยอัตโนมัติของรูปแบบการเข้าถึงของผู้ใช้จากเว็บเซิร์ฟเวอร์หนึ่งหรือมากกว่าหนึ่ง

วิภา เจริญภัณฑารักษ์ (2555) ได้ให้ความหมาย การทำเหมืองข้อมูล หมายถึง กระบวนการค้นหาความรู้ ซึ่งเป็นรูปแบบและความสัมพันธ์ที่น่าสนใจ เพื่อสร้างตัวแบบ/แบบจำลอง มีเทคนิคหรือวิธีการต่างๆ เช่นวิธีการจัดกลุ่ม การค้นหาความสัมพันธ์ การพยากรณ์ เหตุผลของการทำเหมืองข้อมูลที่มีความซับซ้อนและมีข้อมูลจำนวนมาก และเพื่อสกัดความรู้ที่ซ่อนเร้นอยู่ในข้อมูลที่เกิดในกิจกรรมของหน่วยงาน

วิกิพีเดีย (th.wikipedia.org/wiki/การทำเหมืองข้อมูล, 2556) ได้กล่าวว่า การทำเหมืองข้อมูล (อังกฤษ: data mining) หรืออาจจะเรียกว่า การค้นหาความรู้ในฐานข้อมูล (อังกฤษ: Knowledge Discovery in Databases - KDD) เป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) ของจากข้อมูลจำนวนมหาศาลโดยอัตโนมัติโดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่องและการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูล (โดยส่วนใหญ่จะมีจำนวนมาก) เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

3.1 การทำเหมืองข้อมูลเว็บ

อดุลย์ ยิ้มงาม (2551) กล่าวว่า การทำเหมืองข้อมูลเว็บคือการใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาองค์ความรู้ และสกัดข้อมูลสารสนเทศจากเอกสารบนเว็บเพจ และการให้บริการเว็บไซต์โดยอัตโนมัติ เพื่อนำองค์ความรู้ที่สกัดมาใช้ประโยชน์ หรือแก้ปัญหาทั้งทางตรงและทางอ้อม การทำเหมืองข้อมูลเว็บได้แบ่งประเภทของการทำเหมืองข้อมูลเว็บโดยพิจารณาจากข้อมูลที่เพื่อนำมาวิเคราะห์ออกเป็น 3 ประเภทดังนี้คือ

3.1.1 Web Content Mining เป็นการค้นหาข้อมูลที่มีประโยชน์จากข้อมูลที่อยู่ภายในเว็บ อาทิ ข้อความ รูปภาพ โดยสามารถแบ่งออกเป็น 2 ประเภทตามมุมมอง ได้แก่ มุมมองทางด้านการสืบค้นสารนิเทศ (Information Retrieval) ซึ่งมุ่งแน่นเพื่อการปรับปรุงการหาข้อมูล หรือกรองข้อมูล ให้ผู้ใช้โดยพิจารณาจากข้อมูลที่ผู้ใช้อ้างอิง และมุมมองทางด้านฐานข้อมูล (Database) คือการพยายามจำลองข้อมูลบนเว็บและรวบรวมข้อมูลเหล่านั้น เพื่อให้การสอบถามทำงานดีขึ้นมากกว่าการใช้คำหลักเป็นตัวค้นหาเพียงอย่างเดียว

3.1.2 Web Structure Mining เป็นวิธีการที่พยายามค้นหารูปแบบโครงสร้างการเชื่อมโยงที่สำคัญและซ่อนอยู่ในเว็บ ซึ่งรูปแบบนี้จะขึ้นอยู่กับรูปแบบการเชื่อมโยงเอกสารภายในเว็บ โดยนำรูปแบบที่ได้มาใช้เพื่อจัดกลุ่มเว็บเพจและใช้สร้างข้อมูลสารสนเทศที่เป็นประโยชน์ เช่น นำมาใช้ในการปรับโครงสร้างของเว็บให้สามารถให้บริการผู้ใช้ได้อย่างรวดเร็ว

3.1.3 Web Usage Mining เป็นวิธีการที่พยายามค้นหาความหมายของข้อมูลที่สร้างจากช่วงการทำงานหนึ่งของผู้ใช้หรือสร้างจากพฤติกรรมของผู้ใช้เรียกอีกชื่อหนึ่งว่า Web Log Mining โดยในขณะที่ Web Content Mining และ Web Structure Mining ใช้ประโยชน์จากข้อมูลจริง หรือข้อมูลพื้นฐานบนเว็บแต่ Web Usage Mining ทำการค้นหาความรู้จากข้อมูลการติดต่อสื่อสารระหว่างกันของผู้ใช้ที่ติดต่อกับเว็บ โดย Web Usage Mining ทำการรวบรวมข้อมูลจากบันทึกในการดำเนินการต่างๆ เช่น บันทึกการใช้งานของ Proxy (Proxy Server Log) ข้อมูลการลงทะเบียน (Registration Data) หรือข้อมูลอื่นอันเป็นผลจากการทำงานร่วมกันมาใช้วิเคราะห์ ดังนั้น Web Usage Mining จึงเป็นวิธีการทำงานที่เน้นใช้เทคนิคที่สามารถทำนายพฤติกรรมของผู้ใช้ในขณะที่ผู้ใช้ทำงานกับเว็บ กระบวนการทำงานของ Web Usage Mining สามารถแบ่งออกเป็น 2 วิธีคือ

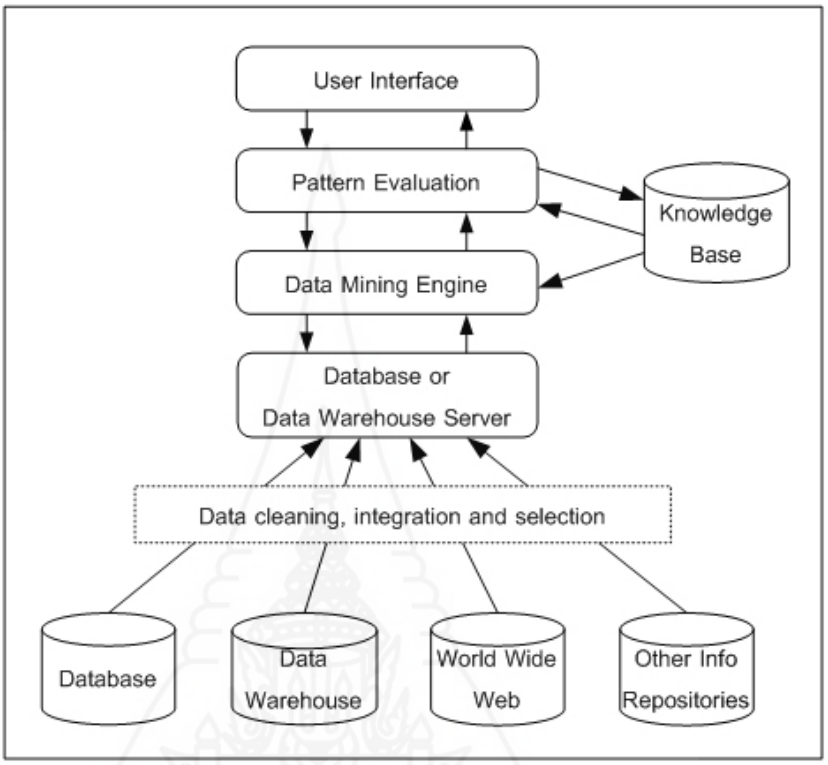
1) ทำการจับคู่ข้อมูล การใช้งานของเครื่องให้บริการเว็บให้อยู่ในรูปของตารางความสัมพันธ์ ก่อนที่นำข้อมูลนี้มาปรับใช้กับเทคนิคการทำเหมืองข้อมูลการใช้เว็บ

2) ใช้ประโยชน์จากข้อมูล ในบันทึกการใช้งานโดยตรงซึ่งจะใช้เทคนิคการเตรียมข้อมูล (Preprocessing) เพื่อเตรียมข้อมูลก่อนหาความสัมพันธ์ (Pattern Discovery) และวิเคราะห์รูปแบบ (Pattern Analysis)

วิภา เจริญภัณฑารักษ์ (2555) ได้เขียนหน่วยที่ 8 เรื่องหลักการพื้นฐานของการทำเหมืองข้อมูล ในประมวลสาระชุดวิชา “คลังข้อมูล เหมืองข้อมูล และธุรกิจอัจริยะ” มหาวิทยาลัยสุโขทัยธรรมาธิราช ได้กล่าวเกี่ยวกับ สถาปัตยกรรมระบบเหมืองข้อมูล,กระบวนการการทำเหมืองข้อมูล,ประเภทการทำเหมืองข้อมูล ไว้ดังต่อไปนี้

3.2 สถาปัตยกรรมระบบเหมืองข้อมูล

สถาปัตยกรรมของระบบการทำเหมืองข้อมูลประกอบด้วยส่วนประกอบที่สำคัญ ดังนี้



ภาพที่ 2.5 แสดงสถาปัตยกรรมการทำระบบเหมืองข้อมูล

ที่มา: http://siripornk.blogspot.com/2010/08/data - mining.html Retrieved January 20,2014

3.2.1 แหล่งข้อมูลที่ใช้ทำเหมืองข้อมูล (data sources) เป็นแหล่งข้อมูลที่อาจจะจัดเก็บไว้อยู่ในฐานข้อมูล คลังข้อมูล เวิลด์ไวด์เว็บ (WWW) และคลังจัดเก็บข้อมูลอื่นๆ เนื้อหาสาระภายในข้อมูลเป็นสิ่งที่ผู้ทำเหมืองข้อมูลมีความสนใจและต้องการค้นหาความรู้ที่ซ่อนเร้นอยู่ การดำเนินการเบื้องต้นเพื่อให้ข้อมูลเหล่านี้อยู่ในสภาพพร้อมใช้งาน นับตั้งแต่

1) การทำความสะอาดข้อมูล (data cleaning) เป็นการจัดการข้อมูลที่ไม่สมบูรณ์ รายละเอียดขาดหายไป เนื้อหาข้อมูลขัดแย้งกันเอง มาดำเนินการให้ข้อมูลใช้งานได้

2) การบูรณาการข้อมูล (data integration) การนำข้อมูลจากแหล่งต่างๆ ที่สัมพันธ์กันมาไว้ในที่เดียวกัน จำเป็นต้องปรับรูปแบบให้เป็นแบบเดียวกัน เพื่อให้ทำงานร่วมกันได้

3) การคัดเลือกข้อมูล (data selection/extraction) เป็นการนำข้อมูลที่ต้องการมาใช้งานมาจัดเก็บเพื่อเป็นแหล่งข้อมูล

4) เซิร์ฟเวอร์ฐานข้อมูลหรือคลังข้อมูล (Database/data warehouse server) เป็นระบบแม่ข่ายที่ใช้จัดเก็บข้อมูล เพื่อรองรับข้อมูลทั้งหมดในการทำเหมืองข้อมูล

5) ฐานความรู้ (knowledge base) หมายถึง ความรู้ความเข้าใจในสิ่งหรือเรื่องที่ต้องการวิเคราะห์หรือค้นหา ความรู้ความเข้าใจเหล่านี้เกิดจากการเรียนรู้และประสบการณ์ในเรื่องดังกล่าว มีความเข้าใจในธรรมชาติของข้อมูลที่มีอยู่และความรู้ในการเลือกวิธีการทำเหมืองข้อมูลได้อย่างเหมาะสมกับลักษณะงานหรือข้อมูลที่ต้องการดำเนินการ

6) กลไกการทำเหมืองข้อมูล (data mining engine) เป็นวิธีการทำเหมืองข้อมูล ได้แก่ การกำหนดคุณสมบัติข้อมูล การวิเคราะห์ความสัมพันธ์ของข้อมูล การจำแนกหมวดหมู่ การจัดกลุ่ม การค้นหาความผิดปกติของข้อมูล และเครื่องมือที่ใช้ในการทำเหมืองข้อมูลแต่ละประเภท เช่น วิธีการทางสถิติ

7) ส่วนประเมินรูปแบบ (pattern evaluation module) ในการทำเหมืองข้อมูล สิ่งที่สำคัญคือ การสร้างแบบจำลองหรือตัวแบบ (model) เพื่อแสดงถึงสมมติฐานได้อย่างชัดเจนให้สามารถค้นพบความรู้จากแบบจำลองที่ทำขึ้นมาได้ ทั้งนี้แบบจำลองที่ทำขึ้นควรต้องมีความเหมาะสมในการพยากรณ์ได้ และเครื่องมือหรือกลไกเหล่านี้ ยังใช้ตรวจสอบและประเมินแบบจำลองที่จัดทำขึ้นมาด้วย เพื่อให้ผลลัพธ์ที่ได้มีความน่าเชื่อถือ เหมาะสม และถูกต้องมากที่สุด

8) ส่วนที่ใช้ติดต่อกับผู้ใช้งาน (user interface) แบ่งเป็น

(1) ส่วนรองรับการป้อนคำสั่งเข้าของผู้ใช้งาน เช่น การค้นหา การใส่เงื่อนไขการทำเหมืองข้อมูล

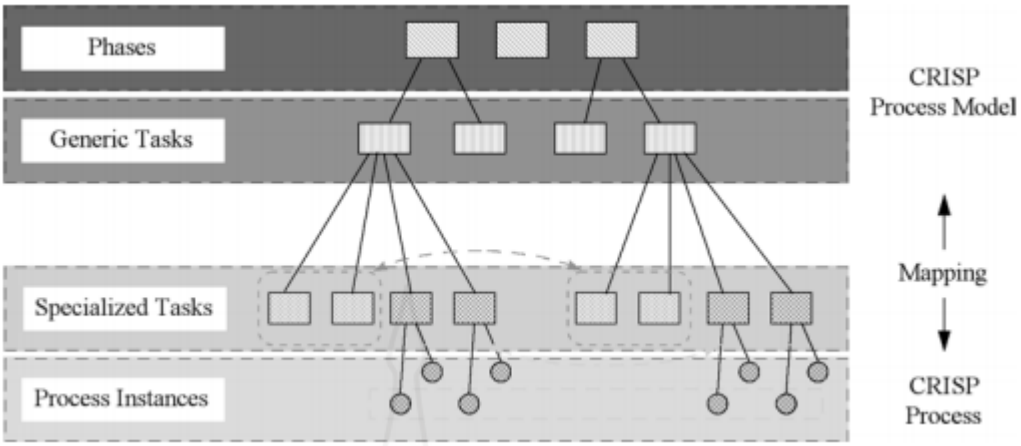
(2) ส่วนนำเสนอผลลัพธ์ที่ได้จากการทำเหมืองข้อมูลในรูปแบบต่างๆ เช่น ตาราง แผนภูมิ และรูปแบบความสัมพันธ์ของข้อมูลที่ค้นหา

3.3 กระบวนการทำเหมืองข้อมูล

การจัดทำเหมืองข้อมูลมีขั้นตอนหรือกระบวนการทำเหมืองข้อมูล ซึ่งประกอบด้วยรายละเอียดต่างๆ ในการดำเนินการแต่ละขั้นตอนอาจจะแตกต่างกันไป ดังนั้นเพื่อให้มีแนวทางการทำงานร่วมกันอย่างชัดเจน จึงได้มีกำหนดกระบวนการที่เป็นมาตรฐานในการดำเนินการ ซึ่งมาตรฐานที่ว่านี้ เรียกว่า คริสป์-ดีเอ็ม (CRISP-DM – Cross Industry Standard Process for Data Mining) ก่อเกิดขึ้นด้วยเหตุผลเพื่อให้มีขั้นตอนมาตรฐาน ที่มีกรอบงานแต่ละส่วนช่วยให้ผู้ทำเหมืองข้อมูลใช้เป็นแนวทาง เพื่อช่วยให้ทำงานมีประสิทธิภาพขึ้น

3.4 หลักการของคริสป์-ดีเอ็ม

ในการกำหนดวิธีของการทำเหมืองข้อมูลจะใช้หลักการพื้นฐาน โดยการแบ่งงานเป็นลำดับชั้น (Hierarchical breakdown) เป็นกลุ่มงานย่อย (set of tasks) 4 ระดับดังภาพที่ 2.6



ภาพที่ 2.6 แสดงการแบ่งระดับงานของคริสป์-ดีเอ็ม
ที่มา: http://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

3.4.1 ระดับที่ 1 งานหลัก (phase) เป็นส่วนบนสุด ประกอบด้วยงานย่อยต่างๆ ที่ต้องดำเนินการในการทำเหมืองข้อมูล ซึ่งเป็นงานทั่วไปทั้งหมด เพื่อให้เห็นภาพรวมของงานที่มีอยู่

3.4.2 ระดับที่ 2 งานทั่วไป (generic task) ในส่วนนี้จะเป็นการกำหนดว่างานแต่ละงานจะทำเหมืองข้อมูลด้วยเทคนิคอะไรบ้าง ซึ่งบางครั้งอาจมีปัญหาที่คาดไม่ถึงเกิดขึ้นในระหว่างการทำงานก็ตาม แต่ก็ให้เห็นภาพของงานที่จะต้องทำได้ในระดับหนึ่ง และจะเชื่อมโยงกับระดับถัดไปได้ ตัวอย่างเช่น ระบุว่ามีงานทำความสะอาดข้อมูล (clean data) จะต้องมีการกำหนดเทคนิคการทำเหมืองข้อมูลที่คาดว่าน่าจะเหมาะสมสำหรับงานนั้น

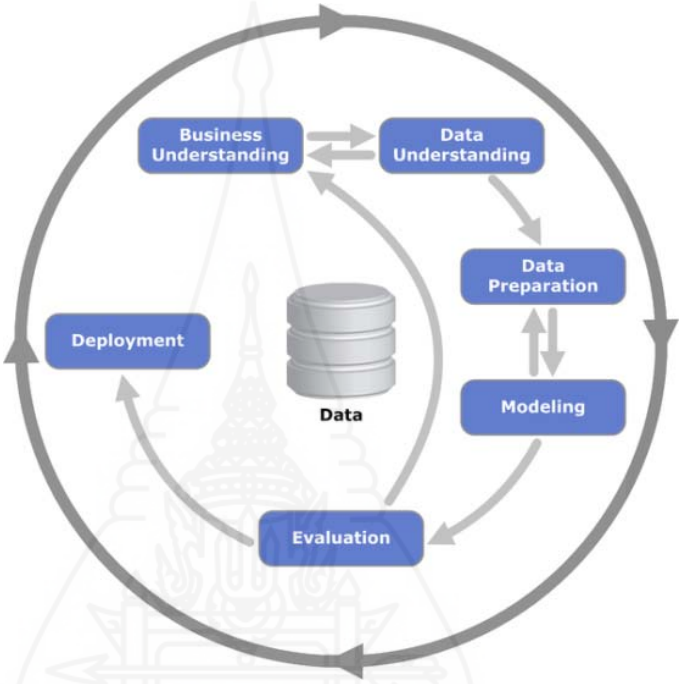
3.4.3 ระดับที่ 3 งานเฉพาะ (specific task) ในส่วนนี้เป็นการขยายรายละเอียดของงานย่อยในระดับ 2 ว่าแต่ละงานย่อยมีวิธีการทำอะไรบ้างในแต่ละสถานการณ์ เช่น การทำความสะอาดข้อมูลที่เป็นตัวเลข ดังนั้นหากมีข้อมูลประเภทอื่นที่ไม่ใช่ตัวเลขแล้วจะมีวิธีการจัดการอย่างไร หรือตัวอย่างจากระดับที่สอง ระบุเทคนิคการทำเหมืองข้อมูลให้ชัดเจนว่าจะใช้เทคนิคใด เช่น การจัดกลุ่ม (clustering) การจำแนกประเภท (classification)

3.4.4 ระดับที่ 4 รายละเอียดวิธีการ (process instance) เป็นข้อมูลที่กำหนดรายละแต่ละงานย่อยในระดับที่เหนือกว่าให้มีรายละเอียดมากขึ้น ซึ่งจะเป็นรายละเอียดเกี่ยวกับการดำเนินการ (action) ของแต่ละงานย่อย เงื่อนไข (decision) ผลลัพธ์ที่เกิดขึ้น (result of the actions) ของงานย่อยแต่ละอัน

3.5 ขั้นตอนของ CRISP-DM

แบบจำลอง CRISP-DM เป็นแบบจำลองแสดงวงจรชีวิตของโครงการทำเหมืองข้อมูลหรือดาต้าไมนิ่งโดยแบบจำลอง CRISP-DM จะครอบคลุมระยะเวลาหรือเฟส การทำงานในโครงการตามลำดับ ของความสัมพันธ์ระหว่างงานทั้งหมดในโครงการ โดยวงจรชีวิตของโครงการทำเหมืองข้อมูล มีด้วยกัน 6 ระยะหรือเฟส โดยลำดับของงานในแต่ละเฟสไม่ตายตัว การเคลื่อนย้ายไปข้างหน้าหรือถอยหลังระหว่างเฟสสามารถกระทำได้ตลอดเวลา ผลที่ได้รับจากแต่ละเฟส ทำให้ทราบว่าต้องดำเนินการเฟสใดต่อไป ลูกศรที่แสดงในภาพชี้ให้เห็นความสำคัญและการขึ้นต่อกันระหว่างเฟส

ต่างๆ วงกลมที่อยู่รอบนอกเป็นสัญลักษณ์ของการทำเหมืองข้อมูลซึ่งต้องวนเป็นวงกลม โดยการทำเหมืองข้อมูลจะไม่สิ้นสุดเมื่อมีการนำผลไปใช้ แต่สิ่งที่ได้รับหรือเรียนรู้ในระหว่างกระบวนและจากการนำผลไปใช้จะสามารถก่อให้เกิดสิ่งใหม่ๆ ซึ่งมักจะมุ่งเน้นเพื่อตอบคำถามให้แก่การดำเนินงานของหน่วยงานได้ โดยผลที่ได้รับจากกระบวนการทำเหมืองข้อมูลจะเป็นประโยชน์มากเมื่อนำระสบการณ์ที่ได้รับจากการวนซ้ำของกระบวนนี้มาประกอบด้วยในกระบวนการทำเหมืองข้อมูลแบบคริสป์-ดีเอ็ม ได้กำหนดไว้ 6 ขั้นตอนดังภาพที่ 2.7



ภาพที่ 2.7 กระบวนการทำเหมืองข้อมูลแบบคริสป์-ดีเอ็ม
ที่มา: e.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf

3.5.1 ความเข้าใจธุรกิจ เป็นการวิเคราะห์ความต้องการธุรกิจ จะช่วยให้เข้าใจสถานะของธุรกิจในปัจจุบัน และประเด็นความต้องการที่ผู้บริหารนำมาใช้ตัดสินใจเพื่อช่วยให้การดำเนินธุรกิจประสบความสำเร็จ และตัดสินใจได้ดียิ่งขึ้น เช่น ความต้องการเกี่ยวกับพฤติกรรมลูกค้า เพื่อพยากรณ์ว่าลูกค้ารายใดน่าจะซื้อสินค้าในรายการสินค้าใหม่ ทำให้สามารถวางแผนธุรกิจได้ตรงกับความต้องการ

3.5.2 ความเข้าใจข้อมูล เป็นการวิเคราะห์ความต้องการข้อมูลที่เกี่ยวข้อง เนื่องจากแหล่งข้อมูล ที่นำมาจัดทำเหมืองข้อมูลนั้นมีมากมาย และจัดเก็บในคลังข้อมูลหรือฐานข้อมูลขนาดใหญ่ มีข้อมูลที่หลากหลายซึ่งมีทั้งข้อมูลที่จำเป็นต้องใช้ในการทำเหมืองข้อมูล ปะปนกับข้อมูลอื่นซึ่งไม่ต้องการในการทำเหมืองข้อมูลในบางครั้ง จึงต้องมีขั้นตอนการกำหนดรายการและประเภทของข้อมูลที่จะนำมาใช้ โดยมีการตรวจสอบในด้านของคุณภาพของข้อมูล จำนวน ปริมาณเนื้อหาและการเข้าถึงข้อมูล เพื่อกำหนดเป็นข้อมูลที่ตรงกับความต้องการทำเหมืองข้อมูล ซึ่งอาจจะมีการคัดเลือก

โดยการในคำสั่งสอบถามข้อมูลจากคลังข้อมูลที่มีข้อมูลจำนวนมาก โดยเลือกตัวอย่างข้อมูลมาทำเหมืองก่อน ซึ่งช่วยลดค่าใช้จ่าย และประหยัดเวลาในการดำเนินการได้

3.5.3 การเตรียมข้อมูล เป็นขั้นตอนสำคัญในการกรองข้อมูลละจัดระเบียบข้อมูลที่เหมาะสมในการทำเหมืองข้อมูล เช่น

3.5.4 การจัดทำแบบจำลอง ในขั้นนี้เป็นการใช้เทคนิคการทำเหมืองข้อมูล แต่ละเทคนิคที่มีรูปแบบและวิธีการในการทำแตกต่างกัน เพื่อสร้างแบบจำลองขึ้นมา ซึ่งแบบจำลองที่ได้จะนำช่วยสกัดสาระสำคัญหรือความรู้ที่ซ่อนเร้นอยู่ออกจากฐานข้อมูล เช่น การค้นหาคำตอบว่าลูกค้าจะซื้อสินค้าหรือไม่ อาจต้องทำการวิเคราะห์โดยใช้เทคนิคการจัดกลุ่มของลูกค้า หรือเทคนิคการจำแนกประเภทลูกค้าแต่ละคนว่าจะซื้อหรือไม่ซื้อสินค้า

ในการสร้างแบบจำลองด้วยเทคนิคต่างๆ ในการทำเหมืองข้อมูล จะแบ่งกลุ่มข้อมูลที่เกี่ยวข้องคือ ข้อมูลที่ใช้ในการสร้างแบบจำลอง (training data set) และกลุ่มข้อมูลที่ใช้ในการทดสอบแบบจำลอง เพื่อตรวจสอบความเที่ยงตรงหรือความน่าเชื่อถือของตัวแบบจำลองที่สร้างขึ้น การใช้เทคนิคต่างๆ ทำให้ได้แบบจำลองที่ช่วยตอบคำถามที่เกิดขึ้น ทั้งนี้ในขณะที่ทำเหมืองข้อมูลอาจมีความจำเป็นต้องเข้าถึงข้อมูลอื่นในคลังข้อมูล รวมทั้งการแปล หรือทดลองนำผลไปประยุกต์ใช้กับธุรกิจเพื่อทดสอบความน่าเชื่อถือ มีการนำเอาสารสนเทศที่ทำเหมืองข้อมูลที่ได้มาวิเคราะห์เพื่อตอบคำถามที่ผู้ตัดสินใจกำลังค้นหาคำตอบ ซึ่งการวิเคราะห์ในส่วนนี้จะครอบคลุมการกรองสารสนเทศที่เหมาะสมกับผู้ใช้งานและการแปลผล

3.5.5 การประเมินผล เกี่ยวข้อง 2 ประเด็น คือ

1) ประเมินความรู้ที่ค้นพบจากการทำเหมืองข้อมูล ว่าตรงกับที่ต้องการหรือไม่ หรือครอบคลุมประเด็นปัญหาที่มีหรือไม่

2) ประเมินวิธีการนำเสนอ ผลลัพธ์ในรูปแบบที่เหมาะสมและง่ายต่อการใช้งาน เช่น แผนภูมิต่างๆ ในเชิงรูปภาพถ้าผลที่ได้ไม่เป็นที่พอใจ ก็จะต้องทำขั้นตอนนี้ซ้ำอีก รวมทั้งขั้นตอนก่อนหน้าด้วย หากเป็นคำตอบที่พอใจหรือน่าเชื่อถือ ก็จะสามารถดำเนินการขั้นต่อไป

3.5.6 การนำแบบจำลองนำไปใช้งาน เป็นขั้นตอนสุดท้ายในการทำเหมืองข้อมูล หลังจากการตรวจสอบจนแน่ใจว่าแบบจำลองที่สร้างขึ้นมาน่าเชื่อถือ ผู้ใช้งานซึ่งเป็นผู้ทำงานเกี่ยวข้องกับการวิเคราะห์และตัดสินใจ สามารถนำไปใช้งานได้

3.6 ประเภทการทำเหมืองข้อมูล

การทำเหมืองข้อมูลเป็นการค้นหารูปแบบที่เกิดขึ้นกับข้อมูลที่สนใจ ซึ่งแบ่งตามลักษณะงานที่ดำเนินการ (Data mining operation/functionality) เป็นประเภทต่างๆ ได้ดังนี้

3.6.1 การค้นหาคุณลักษณะ/รายละเอียดของข้อมูล เป็นการค้นหาคุณลักษณะเฉพาะของข้อมูล (class/concept description) ที่เราให้ความสนใจหรือต้องการดำเนินการ เพื่อเข้าใจในรายละเอียดที่เกี่ยวข้องกับข้อมูลเหล่านั้น การทำเหมืองข้อมูลประเภทนี้ แบ่งได้ 2 ลักษณะ คือ

1) การกำหนดคุณลักษณะเฉพาะข้อมูล (data characterization) เป็นการค้นหาคุณลักษณะเฉพาะของข้อมูล เพื่ออธิบายรายละเอียดเจาะลึกของข้อมูลนั้น ทำให้เข้าใจภาพรวมของข้อมูลในเรื่องนั้น ตัวอย่างเช่น การค้นหาคุณลักษณะ/ประวัติของลูกค้า (customer profile) ที่ซื้อสินค้ากับบริษัท “STOU Electronics” มากกว่า $1,000 ต่อปี โดยใช้คำสั่งเอสคิวแอล

(SQL) เพื่อคัดกรองข้อมูลที่ต้องการในเบื้องต้น พบว่า ลูกค้ามีอายุระหว่าง 40-50 ปี มีงานทำ และมีประวัติ/เครดิต (credit rating) อยู่ในระดับดี เป็นต้น

2) การกำหนดคุณลักษณะแตกต่าง (data discrimination) เป็นการค้นหาและศึกษารายละเอียดของข้อมูลที่เราสนใจเปรียบเทียบกับข้อมูลเดิมที่มีคุณสมบัติที่แตกต่างหรือตรงกันข้าม ซึ่งการเปรียบเทียบข้อมูล มักแบ่งเป็น 2 กลุ่มคือ กลุ่มหลัก และกลุ่มที่แตกต่าง ทั้งนี้ผู้ทำเหมืองข้อมูลต้องเป็นผู้กำหนดเงื่อนไขเอง เช่น การศึกษาเปรียบเทียบรายละเอียดข้อมูลของกลุ่มลูกค้าที่ซื้อสินค้ากับบริษัท STOU Electronics มากกว่า 3 ครั้งต่อเดือน เทียบกับกลุ่มลูกค้าที่ซื้อสินค้าน้อยกว่า 3 ครั้งต่อปี เป็นต้น ผลที่ได้จากการค้นหารายละเอียดคือประวัติการซื้อสินค้าของลูกค้า เช่น กลุ่มที่ซื้อสินค้าประจำ จะเป็นผู้มีอายุระหว่าง 20-40 ปี จบการศึกษาระดับปริญญาตรี ขณะที่กลุ่มที่แตกต่าง จะมีประวัติของลูกค้าที่มีอายต่ำกว่า 20 หรือ มากกว่า 60 ปี และจบการศึกษาในระดับที่ต่ำกว่าปริญญาตรี เป็นต้น ทั้งนี้อาจจะใช้ข้อมูลที่จัดเก็บในคลังข้อมูลมาดำเนินการโดยใช้คำสั่งเอสคิวแอล (SQL) ในการค้นหาตามเงื่อนไขที่ต้องการได้เครื่องมือหรือวิธีการในการทำเหมืองข้อมูลแบบนี้ ใช้วิธีการทางสถิติ เช่น ค่าฉลี่ย (mean) ค่ามัธยฐาน (median) ค่าฐานนิยม (mode) การหาค่าเบี่ยงเบนมาตรฐาน (standard deviation) การเขียนกราฟต่างๆ การหาความสัมพันธ์ระหว่างตัวแปรที่สนใจ และการเปรียบเทียบข้อมูลด้วยตารางหลายมิติ (multidimensional data cube) เป็นต้น การนำเสนอผลลัพธ์ที่ได้อาจจะเป็นแผนภูมิแบบต่างๆ เช่น แผนภูมิแท่ง เส้น และวงกลม เป็นต้น

3.6.2 การค้นหาความสัมพันธ์ เป็นการทำเหมืองข้อมูลแบบการค้นหาความสัมพันธ์ (frequent pattern and association) ประกอบด้วย การค้นหารูปแบบที่เกิดควบคู่กัน (frequent pattern) และสร้างความสัมพันธ์ที่เกิดขึ้น (association rule) มีรายละเอียดดังนี้

*1) การค้นหารูปแบบของข้อมูล*ที่เกิดขึ้นควบคู่ในรายการเดียวกันเสมอ (frequent item set) การทำเหมืองข้อมูลประเภทนี้นิยมใช้ในการวางแผนการตลาด การวางแผนการสั่งซื้อสินค้าว่าควรสั่งซื้อสินค้าใดร่วมกันบ้าง การจัดชั้นวางสินค้าอย่างเหมาะสม การจัดรายการส่งเสริมการขายสินค้า จึงมักนิยมเรียกการทำเหมืองข้อมูลประเภทนี้ว่า การวิเคราะห์ตะกร้าสินค้า (market basket analysis) เป็นการวิเคราะห์หรือค้นหาความสัมพันธ์ข้อมูลที่เกิดควบคู่กันในแต่ละรายการ (transaction) เช่น เมื่อซื้อนม มักจะซื้อ ขนมปัง ไข่ ร่วมกันด้วยเสมอ เป็นต้น

*2) การค้นหาความสัมพันธ์ (association)*หลังจากการค้นหารูปแบบที่เกิดขึ้นควบคู่กัน (frequent item set) ได้แล้ว จะสร้างกฎความสัมพันธ์ของข้อมูล โดยการกำหนดค่า “support” ซึ่งเป็นจำนวนรายการข้อมูลที่นำมาดำเนินการ และค่า “confidence” เป็นค่าความน่าจะเป็นที่ใช้ในการวิเคราะห์ความสัมพันธ์ เช่น STOU Electronics ต้องการค้นหารายการข้อมูล (transaction) ในการซื้อสินค้าของลูกค้า ตามกฎความสัมพันธ์ต่อไปนี้

buys (X, “computer”) => buys (X, “software”) (support=1%, confidence=50%)

หมายความว่า 1% ของจำนวนทรานแซกชันที่นำมาดำเนินการค้นหา (support=1%) พบว่าโอกาสหรือความน่าจะเป็นของลูกค้าที่ซื้อ “computer” แล้วจะซื้อ “software” มี 50% (confidence=50%) ผลที่ได้ทำให้ผู้วิเคราะห์มีความเข้าใจความเชื่อมโยงสัมพันธ์ของข้อมูลได้ดียิ่งขึ้น

ตัวอย่างการค้นหาความสัมพันธ์ด้วยอัลกอริทึมอะไพรโอริ

หลักการของอัลกอริทึมอะไพรโอริ เป็นเทคนิคที่ใช้หารายกรสินค้าที่เกิดร่วมกันบ่ายโดยนำเอาจำนวนนับหรือความถี่ของทรานแซกชัน มาเปรียบเทียบกับค่าเกณฑ์ชั้นต่ำโดยจะเอาเฉพาะค่าจำนวนนับหรือความถี่ของทรานแซกชัน ที่เท่ากับหรือมากกว่าค่าเกณฑ์ชั้นต่ำมาดำเนินการเท่านั้น นั่นคือ ถ้ารายการสินค้าใดไม่ผ่านเกณฑ์ชั้นต่ำถือว่าจำนวนการเกิดรายการสินค้าในทรานแซกชันไม่มากพอ ดังนั้นถ้าหากมีรายการสินค้านั้นปรากฏในรายการทรานแซกชันใดๆ ให้ถือว่ารายการสินค้ามีผ่านเกณฑ์ขั้นต่ำ และไม่ต้องนำรายการสินค้าเหล่านั้นมาดำเนินการหากฎความสัมพันธ์อีก ทำให้ประหยัดเวลาในการประมวลผลได้อย่างมาก

ทั้งนี้เทคนิคอัลกอริทึมอะไพรโอริแสดงด้วยสัญลักษณ์ทางคณิตศาสตร์ได้ดังนี้คือ

“If an itemset Z is not frequent then for any item A, Z ∪A will not be a frequent”

นั่นคือ ถ้าทรานแซกชันที่มีรายการสินค้า Z เป็นค่าที่ไม่เกิดขึ้นบ่าย (เพราะไม่ผ่านเกณฑ์ขั้นต่ำแล้วหากมีรายการสินค้าอื่น เช่น รายการสินค้า A เกิดร่วมกับสินค้า Z แล้ว (นั่นคือ Z ∪A) ให้ถือว่ารายการที่เกิดขึ้นร่วมกันระหว่าง Z และ A มีจำนวนการเกิดไม่บ่อย (not frequent) ด้วยเช่นกันและไม่นำทรานแซกชันของการเกิดรายการสินค้าร่วมกันระหว่าง A และ Z มาหาความสัมพันธ์

ขั้นตอนการทำงานของเทคนิคอัลกอริทึมอะไพรโอริ มีขั้นตอนดังนี้

1) กำหนดรายการสินค้าในแต่ละทรานแซกชัน โดยเริ่มจากสินค้า 1 รายการ
2) นับจำนวนรายการสินค้าในแต่ละทรานแซกชัน ที่มี 1 รายการ
3) เปรียบเทียบจำนวนนับรายการสินค้า กับค่าเกณฑ์ขั้นต่ำ เป็นค่าที่ผู้วิเคราะห์กำหนด ซึ่ง

(1) ถ้าหากค่าจำนวนนับที่มีค่าเท่ากับหรือมากกว่าค่าเกณฑ์ขั้นต่ำถือว่ารายการสินค้านั้นเกิดขึ้นบ่ายเมื่อเกิดร่วมกับรายการสินค้าใดๆ ก็จะถือว่ารายการสินค้าที่เกิดร่วมมีค่าเกิดขึ้นบ่อยด้วยเช่นกัน

(2) ถ้าหากรายการสินค้าไม่ผ่านเกณฑ์ขั้นต่ำ ถือว่ารายการสินค้านั้นมีจำนวนนับไม่มากพอเมื่อเกิดร่วมกับรายการสินค้าใดๆ ก็จะถือว่ารายการสินค้าที่เกิดร่วมมีจำนวนนับไม่มากด้วย ให้ตัดรายการนั้นออกไป รวมทั้งรายการสินค้าอื่นที่เกิดขึ้นร่วมกันให้ตัดออกเช่นกันไม่นำมาดำเนินการต่อในขั้นถัดไป

หลังจากนั้นเพิ่มการเกิดร่วมกันของรายการสินค้า 2 รายการและทำซ้ำขั้นตอนขั้นต้น และเพิ่มการเกิดร่วมกันของรายการสินค้า 3 รายการ และ 4 รายการ ไปเรื่อยจนกว่าจะทำต่อไม่ได้ หรือไม่มีรายการสินค้าในการจับคู่ร่วมกันอีกต่อไป

การสร้างกฎความสัมพันธ์ เป็นการสร้างความสัมพันธ์ของการเกิดเหตุการณ์หนึ่งแล้วจะเกิดอีกเหตุการณ์ตามมา เช่น เมื่อซื้อสินค้ารายการหนึ่งแล้ว จะซื้อสินค้าอีกรายการหนึ่งด้วย เป็นต้น ในการพิจารณากฎความสัมพันธ์ว่ามีความเป็นไปได้มากน้อยเพียงใด จะพิจารณาจากค่า

ความน่าจะเป็น 2 ค่า คือค่า support และ confidence โดยที่ค่า support เป็นค่าความน่าจะเป็นหรือโอกาสที่จะซื้อสินค้าหรือเกิดเหตุการณ์ร่วมกันและค่า confidence เป็นค่าความน่าจะเป็นหรือโอกาสที่เมื่อสินค้ารายการหนึ่งหรือเกิดเพตุการหนึ่งแล้ว จะซื้อสินค้ารายการอื่นหรือเกิดเหตุการณ์อื่นตามมา

ดังนั้นในการสร้างกฎความสัมพันธ์มักจะอยู่ในรูปแบบของกฎที่ประกอบไปด้วยเงื่อนไข IF-THEN และค่าความน่าจะเป็นของการการซื้อสินค้าทั้ง 2 รายการ (support) รวมทั้งความน่าจะเป็นหรือโอกาสที่ซื้อสินค้ารายการแรกแล้ว จะซื้อสินค้ารายการที่สองตามมา (confidence) ดังนี้

IF buy A THEN buy B (support = 70%, confidence 0.8) หรือ

IF buy A THEN buy B (support = 0.7, confidence 0.8)

หมายความว่า ค่า support= 70% หมายถึงความน่าจะเป็นหรือโอกาสที่จะซื้อสินค้า A และ B มี 70 เปอร์เซนต์นั้นคือในตะกร้าสินค้า 100 ตะกร้า จะพบว่ามี 70 ตะกร้าที่ซื้อสินค้า A และ B ค่า confidence= 80% หมายถึงความน่าจะเป็นหรือโอกาสที่เมื่อซื้อสินค้า A และจะซื้อสินค้า B มี 80 เปอร์เซนต์ นั้นคือในตะกร้าสินค้า 100 ตะกร้า ที่ซื้อสินค้า A จะพบว่าจะซื้อสินค้า B ตามมาด้วย 80 ตะกร้า

การหาค่า Support เป็นค่าความน่าจะเป็นในการซื้อสินค้าทั้งสองรายการนิยมนำเสนอเป็นร้อยละหรือเปอร์เซนต์ (%) สูตรที่ใช้ในการหาค่า support (A->B) โดย A แทนรายการสินค้าแรกและ B แทนรายการสินค้าที่ซื้อร่วม คือ

$$\text{support } (A \rightarrow B) = P(A \cap B)$$

$$= \frac{\text{จำนวนนับของรายการสินค้า A และ B}}{\text{จำนวนนับของทรานแซกชั้นทั้งหมด}}$$

การหาค่า confidence เป็นค่าความน่าจะเป็นเมื่อซื้อสินค้ารายการแรกแล้วโอกาสหรือความเชื่อมั่นจะซื้อสินค้ารายการถัดไปจะมีมากน้อยเพียงใด สูตรที่ใช้ในการหาค่า confidence (A->B) โดย A แทนรายการสินค้าแรก และ B แทนรายการสินค้ารายการที่จะซื้อถัดไปหลังจากที่ทราบค่า support แล้วคือ

$$\text{confidence } (A \rightarrow B) = P(A|B)$$

$$P(A \cap B) = \frac{\text{จำนวนนับของรายการสินค้า A และ B}}{\text{จำนวนนับของทรานแซกชั้นทั้งหมด}}$$

$$P(A) = \frac{\text{จำนวนนับของรายการสินค้า A}}{\text{จำนวนนับของทรานแซกชั้นทั้งหมด}}$$

$$\text{ดังนั้น confidence } (A \rightarrow B) = \frac{\text{จำนวนนับของรายการสินค้า A และ B}}{\text{จำนวนนับของรายการสินค้า A}}$$

เครื่องมือในการทำเหมืองข้อมูลประเภทนี้มักจะเป็นใช้เทคนิคหรืออัลกอริทึมมาช่วยรวบรวมและวิเคราะห์ความถี่ของการเกิดข้อมูล เช่น การหาจำนวนความถี่การเกิดรายการ (Frequent item set) ด้วยอัลกอริทึมอะไพรโอริ (Apriori) อัลกอริทึมฟรีเควนแพตเทิร์นโกรท (Frequent Pattern Growth – FP-growth) กฎความสัมพันธ์เป็นต้น

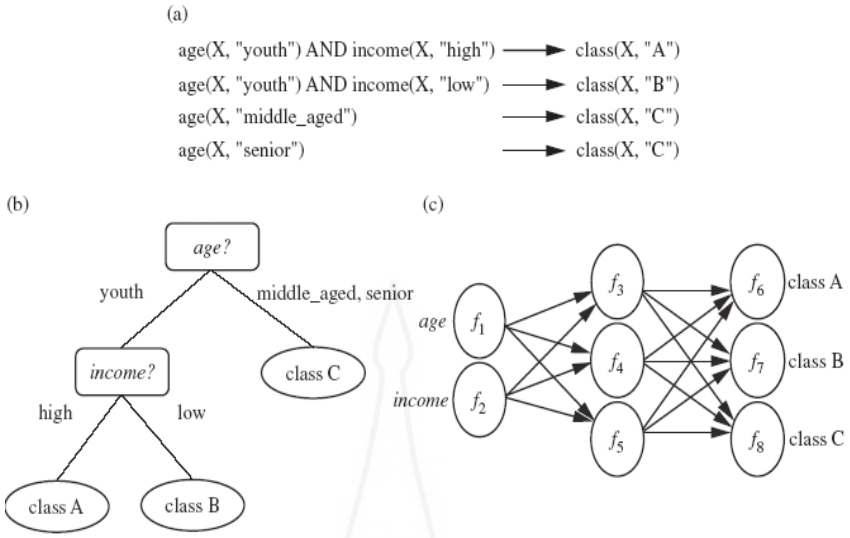
3.6.3 การจำแนกประเภทและการพยากรณ์ (Classification and Prediction)

1) การจำแนกประเภท (classification) เป็นการทำเหมืองข้อมูลเพื่อจำแนกว่าข้อมูลที่ต้องการวิเคราะห์ว่าอยู่ในประเภท (class) ใด โดยอาศัยแบบจำลองหรือตัวแบบที่สร้างขึ้นมา การจำแนกประเภทจะใช้ข้อมูลที่มีอยู่จำนวนหนึ่ง หรือชุดข้อมูลเพื่อสร้างแบบจำลอง (training data set) หลังจากที่ได้แบบจำลองแล้วก็จะต้องทำการทดสอบแบบจำลองดังกล่าวด้วยข้อมูลอีกกลุ่มเพื่อตรวจสอบความเที่ยงหรือความน่าเชื่อถือของแบบจำลองนั้น การทำเหมืองข้อมูลโดยวิธี จำแนกประเภทจะใช้สำหรับข้อมูลที่มีค่าไม่ต่อเนื่อง (discrete data) เช่น ข้อมูลที่ระบุเป็นผลการตัดสินใจว่า ใช่ หรือ ไม่ใช่ เสี่ยง หรือ ไม่เสี่ยง

ตัวอย่างเช่น ฝ่ายสินเชื่อของธนาคารต้องการวิเคราะห์ความเสี่ยงในการให้สินเชื่อลูกค้าของธนาคาร ซึ่งมีแบบจำลองที่สร้างขึ้นเพื่อกำหนดหรือใช้พยากรณ์ประเภท (class) โดยมีเงื่อนไข ดังนี้ หากเงินเดือนตั้งแต่ 20,000 ขึ้นไป และมีงานประจำ มีประเภทความเสี่ยงการให้สินเชื่อ คือ ไม่เสี่ยง แต่หากเงินเดือนต่ำกว่า 20,000 และงานที่ทำเป็นงานชั่วคราว มีประเภทความเสี่ยงการให้สินเชื่อ คือ เสี่ยง ดังนั้นหากต้องการพยากรณ์ ความเสี่ยงในการให้สินเชื่อแก่นาย A ซึ่งมีเงินเดือน 20,000 และมีอาชีพรับราชการ การพยากรณ์ประเภทความเสี่ยงของการให้สินเชื่อนาย A คือ ไม่เสี่ยง

เครื่องมือในการทำเหมืองข้อมูลแบบจำแนกประเภท ได้แก่ การใช้อัลกอริทึมสำหรับสร้างต้นไม้การตัดสินใจ (decision tree) ทฤษฎีเบย์ (Bayes Thorem) นาอีฟเบย์ (Naïve Bayes) โครงข่ายความเชื่อแบบเบย์ (Bayes Belief network) การวิเคราะห์แบบ rule base ในรูปแบบ IF_THEN rule โครงข่ายประสาทประดิษฐ์หรือนิวรัลเน็ตเวิร์ก (neural networks) เอสวีเอ็ม (SVM -Support Vector Machine) กฎความสัมพันธ์การหาคุณสมบัติที่มีความคล้ายคลึงหรือใกล้เคียงกันโดยวิธี k-nearest neighbor การหาเหตุและผลเพื่อตอบสมมติฐานโดยใช้วิธี case base reasoning และอัลกอริทึมที่ใช้หลักการเชิงพันธุกรรม (genetic alogorithm)

ภาพที่ 2.8 เป็นตัวอย่างเครื่องมือที่ใช้ในการจำแนกประเภท (classification) 8.7 (a) IF-THEN rule 8.7 (b) ต้นไม้การตัดสินใจ (decision tree) 8.7 (c) โครงข่ายประสาทประดิษฐ์ (neural networks)

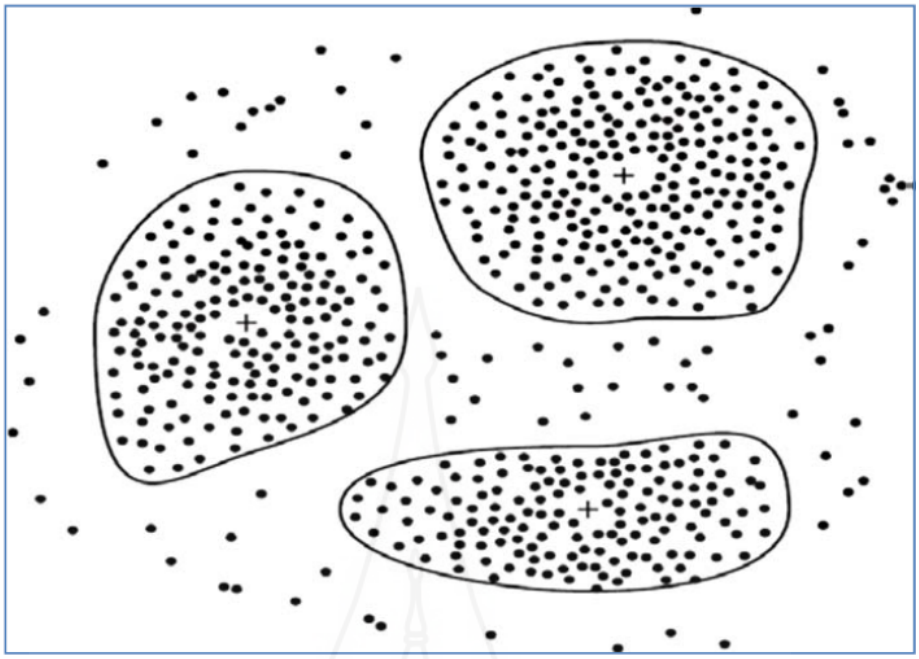


ภาพที่ 2.8 ตัวอย่างเทคนิคการทำเหมืองข้อมูลแบบจำแนกประเภท
ที่มา: วิภา เจริญภัณฑารักษ์ (2555: 26)

2) การพยากรณ์ (prediction) มีหลักการเช่นเดียวกับการจำแนกประเภท แต่จะใช้กับข้อมูลที่มีค่าต่อเนื่อง (continuous data) หรือค่าข้อมูลที่มีค่าเป็นลำดับการเกิดของข้อมูล (ordered data) นั้น จำแนกอยู่ในประเภทใด เช่น ข้อมูลอายุ เงินเดือน ความสูง เวลา รายได้จากการขายสินค้า เป็นต้น การทำเหมืองข้อมูลประเภทนี้เพื่อใช้พยากรณ์หรือประมาณการผลลัพธ์จากกลุ่มข้อมูลที่มีอยู่ในขณะนั้น เช่น การพยากรณ์ยอดขายสินค้าจากข้อมูลปัจจุบัน

เครื่องมือในการทำเหมืองข้อมูลแบบพยากรณ์ (Prediction) ได้แก่ การวิเคราะห์ความถดถอย (regression analysis) การวิเคราะห์และตรวจสอบด้วยการใช้ไคสแควร์ (Chi Square Automatic Interaction Detection – CHAID)

3.6.4 การจัดกลุ่มการทำเหมืองข้อมูลประเภทจัดกลุ่ม ไม่ได้กำหนดหรือแบ่งกลุ่มไว้ล่วงหน้า เป็นเทคนิคจัดกลุ่มสมาชิกที่มีความคล้ายคลึงกันสูงสุด (nearest neighbor) เพื่อจำแนกข้อมูลแต่ละหน่วยในชุดข้อมูล โดยวัดค่าความห่างจากจุดศูนย์กลาง (distance measurement) การวัดค่าเบี่ยงเบน (deviation measurement) เป็นการรวมหน่วยที่คล้ายกันมากที่สุดเข้าเป็นกลุ่มเดียวกัน ซึ่งบางครั้งเรียกเทคนิคนี้ว่า k-nearest neighbor ตัวอย่างเช่น การทำเหมืองข้อมูลเพื่อวิเคราะห์ตำแหน่งที่อยู่ของลูกค้าบริษัท STOU Electronics เพื่อจัดกลุ่มลูกค้าว่าอาศัยในในเขตพื้นที่ใดบ้าง ดังภาพที่ 2.9 ข้อมูลของลูกค้าที่อาศัยในพื้นที่ต่างๆ ซึ่งจัดได้ 3 กลุ่มได้ จะเห็นว่าการจัดกลุ่มเหล่านี้ไม่ได้มีการกำหนดว่าจะต้องเป็นกลุ่มใดล่วงหน้า ในการเกิดกลุ่มเหล่านี้เป็นเพราะข้อมูลในแต่ละกลุ่มมีคุณลักษณะเป็นไปในทางเดียวกันจึงรวมตัวกันเป็นกลุ่มๆ ได้ ซึ่งต่างจากการจำแนกประเภทที่มีการกำหนดกลุ่มหรือประเภทไว้ล่วงหน้า



ภาพที่ 2.9 ข้อมูลของลูกค้าที่อาศัยในพื้นที่ต่างๆ ซึ่งจัดได้ 3 กลุ่ม
ที่มา: วิภา เจริญภัณฑารักษ์ (2555: 27)

1) การจัดกลุ่มโดยใช้เทคนิค K-means เป็นการจัดกลุ่มที่แบ่งแต่ละกลุ่มออกจากกันอย่างสิ้นเชิงไม่มีการจัดเป็นชั้นๆ หลักการจัดกลุ่มด้วยมีดังนี้

(1) กำหนดจำนวนคลัสเตอร์ที่ต้องการจัดกลุ่ม

(2) ให้นำข้อมูลที่ต้องการจัดกลุ่มไว้ในแต่ละกลุ่ม กลุ่มละ 1 จำนวน

(3) หาค่าเฉลี่ยของข้อมูลในแต่ละกลุ่ม (การหาค่าเฉลี่ยอาจจะใช้สูตรในการคำนวณที่แตกต่างกันได้ ทั้งนี้ขึ้นกับผู้วิเคราะห์และลักษณะข้อมูล)

(4) นำข้อมูลที่เหลือแต่ละจำนวนจัดเข้ากลุ่มหรือคลัสเตอร์ที่มีอยู่ โดยค่าของข้อมูลใกล้เคียงกับค่าเฉลี่ยของกลุ่มที่จะจัดเข้ามากที่สุด

(5) ให้หาค่าเฉลี่ยใหม่ของแต่ละกลุ่มหรือคลัสเตอร์ โดยใช้ข้อมูลที่มีอยู่ในคลัสเตอร์จากนั้นจะเปรียบเทียบข้อมูลแต่ละจำนวนกับค่าเฉลี่ยใหม่ และปรับข้อมูลเข้ากลุ่มใหม่ไปเรื่อยจนกระทั่งค่าเฉลี่ยไม่เปลี่ยนแปลง

ตัวอย่างการจัดกลุ่มโดยใช้เทคนิค K-means เช่น มีข้อมูลในการจัดกลุ่ม คือ {2, 4, 10, 12, 3, 20, 30, 11, 25} และกำหนดจำนวนกลุ่มที่ต้องการจัดกลุ่มคือ 2 กลุ่มหรือคลัสเตอร์ โดยมีขั้นตอนดังนี้

(1) กำหนดจำนวนคลัสเตอร์ที่ต้องการจัดกลุ่ม มี 2 คลัสเตอร์ คือ C1, C2

(2) จัดข้อมูลเข้ากลุ่มทำให้คลัสเตอร์ C1 ={2} และคลัสเตอร์ C2={4}และหาค่าเฉลี่ยของแต่ละคลัสเตอร์ คลัสเตอร์ C1 มีค่าเฉลี่ย = 2 คลัสเตอร์ C2 มีค่าเฉลี่ย = 4

(3) เลือกข้อมูลที่เหลืออยู่ในรายการที่กำหนด คือ {10, 12, 3, 20, 30, 11, 25} มาจัดเข้ากลุ่ม C1 หรือ C2 โดยเปรียบเทียบว่าข้อมูลแต่ละจำนวนกับค่าเฉลี่ยของแต่ละคลัสเตอร์ โดยข้อมูลที่ใกล้เคียงกับค่าเฉลี่ยของคลัสเตอร์ใด ก็ให้เข้าคลัสเตอร์นั้น

(4) หาค่าเฉลี่ยและปรับกลุ่มใหม่ โดยหาค่าเฉลี่ยของแต่ละคลัวเตอร์ใหม่ตามข้อมูลชุดใหม่ที่มีในคลัสเตอร์นั้น และปรับข้อมูลในแต่ละคลัสเตอร์ให้ใกล้เคียงกับค่าเฉลี่ยใหม่ที่คำนวณได้ของแต่ละคลัสเตอร์ นั้น จนกว่าว่าค่าเฉลี่ยแต่ละคลัสเตอร์ไม่เปลี่ยนแปลง นั่นคือได้คลัสเตอร์ที่ต้องการจัดกลุ่ม

ตารางที่ 2.1 ขั้นตอนการจัดกลุ่มด้วยเทคนิค K-means จากข้อมูล {2, 4, 10, 12, 3, 20, 30, 11, 25}

ขั้นที่	คลัสเตอร์ C1	คลัสเตอร์ C2	ค่าเฉลี่ยของ C1	ค่าเฉลี่ยของ C2
1	{2}	{2}	2	4
2	{2,3}	{4,10,12,20,30,11,25}	(2 + 3) / 2 = 2.5	(4 + 10 +12 + 20 +30 +11 +25) / 7 =16
3	{2,3,4}	{10,12,20,30,11,25}	3	18
4	{2,3,4,10}	{12,20,30, 11,25}	4.75	19.6
5	{2,3,4,10,11}	{12,20,30,25}	6	21.75
6	{2,3,4,10,11,12}	{20,30,25}	7	25
7	{2,3,4,10,11,12}	{20,30,25}	7	25

สรุปการจัดกลุ่มด้วยเทคนิค K-means จากข้อมูลเริ่มต้น คือ {2, 4, 10, 12, 3, 20, 30, 11, 25} โดยกำหนดให้มี 2 กลุ่มหรือคลัสเตอร์ ได้ดังนี้ {2,3,4,10,11,12} และ {20,30,25}

2) ซีเควนซ์คลัสเตอริง (Sequence Clustering) เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูลที่มีลักษณะของการเรียงลำดับ (sequence) เช่น ลำดับของชุดวิชาที่นักศึกษาลงทะเบียนเรียน และนอกจากนี้ซีเควนซ์คลัสเตอริงยังใช้หลักการของแบบจำลองมาร์คอฟเชน (Markov Chain) มาเป็นพื้นฐานในอัลกอริทึมด้วย ทำให้สามารถวิเคราะห์การเปลี่ยนสถานะของการลงทะเบียนเรียนของนักศึกษากลุ่มนี้ได้ด้วยโดยอัลกอริทึมจะทำการจัดกลุ่มข้อมูลที่มีการเรียงลำดับคล้ายคลึงกันไว้ด้วยกัน

เครื่องมือในการทำเหมืองข้อมูลแบบการจัดกลุ่ม เช่น การใช้การวิเคราะห์ค่าเฉลี่ยแบบเค (k-mean analysis) ลำดับการจัดกลุ่ม (Microsoft Sequence Clustering Algorithm)

3.6.5 การค้นหาหน่วยข้อมูลที่ผิดปกติ เป็นการทำเหมืองข้อมูลเพื่อค้นหาหน่วยข้อมูลที่แตกต่างหรือผิดแปลก (outlier analysis) จากคุณสมบัติข้อมูลทั่วไปที่จัดเก็บในฐานข้อมูลอย่างมาก สาเหตุของการเกิดหน่วยเช่นนั้นอาจจะมาจากความผิดพลาดในการบันทึกข้อมูล หรือความผิดพลาดอื่น (noise) หรือเป็นลักษณะผิดปกติจริง (exception) การค้นพบหน่วยผิดปกติและจัดการ

กับข้อมูลดังกล่าวอย่างเหมาะสมในการทำเหมืองข้อมูลจะทำให้ข้อความรู้ที่ได้มีคุณค่ามากขึ้น เช่น ติดตามการฉ้อโกงในการใช้บัตรเครดิต ที่มีรายการและยอดการใช้จ่ายที่พุ่งสูงขึ้นอย่างรวดเร็ว เมื่อเทียบกับพฤติกรรมการใช้เวลาปกติทั่วไป

เครื่องมือที่ใช้ในการทำเหมืองข้อมูลประเภทนี้ มักจะเป็นการใช้สถิติเพื่อวิเคราะห์ความน่าจะเป็น การวัดค่าความห่างจากจุดศูนย์กลาง (distance measurement) และการวัดค่าเบี่ยงเบน (deviation measurement)

3.6.6 การวิเคราะห์การเปลี่ยนแปลงที่อิงตามกาลเวลา เป็นการทำเหมืองข้อมูลเพื่อค้นหารูปแบบข้อมูลที่มีการเปลี่ยนแปลงตามเวลา (evolution analysis) มักจะดำเนินการผสมผสานวิธีต่างๆ ที่กล่าวมาข้างต้นเพื่อค้นหารูปแบบเพื่อใช้ประโยชน์จากข้อมูลที่มีอยู่ เช่น การวิเคราะห์การเปลี่ยนแปลงของราคาหุ้นในแต่ละประเภทเพื่อการลงทุนที่เหมาะสม เป็นต้น

ตัวอย่างอัลกอริทึม ไทม์ซีรีส (Time Series) เป็นอัลกอริทึมที่ใช้เทคนิคออโตรีเกรสชัน (autoregression) และดีซีชันทรี (decision tree) มาผสมผสานกัน บางครั้งอัลกอริทึมนี้จึงถูกเรียกว่า ‘ออโตรีเกรสชันทรี (AutoRegtressionTree -ART)’ ไทม์ซีรีสเป็นเทคนิคในการพยากรณ์ค่าตัวเลขในอนาคต

เครื่องมือในการทำเหมืองข้อมูลแบบการวิเคราะห์การเปลี่ยนแปลงที่อิงตามกาลเวลาเช่น การวิเคราะห์อนุกรมเวลา (Time series analysis) ไมโครซอฟท์ไทม์ซีรีส (Microsoft Time Series Algorithm)

4. ซอฟต์แวร์ทำเหมืองข้อมูล

ในงานวิจัยนี้ใช้ ซอฟแวร์Microsoft SQL Server 2008 ซึ่งเป็นระบบจัดการฐานข้อมูลของบริษัทไมโครซอฟต์สำหรับพัฒนาคลังข้อมูลและในส่วนของการทำเหมืองข้อมูลใช้ซอฟต์แวร์ SQL Analysis Service ซึ่งมีฟังก์ชันสำหรับการทำเหมืองข้อมูล เช่น การหาความสัมพันธ์ (Association mining) การจำแนกประเภท (classification) ด้วยเทคนิคต้นไม้การตัดสินใจ (decision tree) การจัดกลุ่ม (clustering) การหารูปแบบตามลำดับเหตุการณ์ (sequence pattern) การหารูปแบบตามลำดับเหตุการณ์ (sequence pattern) การวิเคราะห์อนุกรมเวลา (Time series Analysis) เป็นต้น รวมทั้งยังสามารถทำงานร่วมกับซอฟต์แวร์ของบริษัทอื่นได้ ไมโครซอฟต์เอสคิวแอล เซิร์ฟเวอร์ 2008 (Microsoft SQL Server 2008) ประกอบด้วยเทคโนโลยีต่อไปนี้

4.1 SQL Server Database Engine เป็นเทคโนโลยีที่ใช้ในการเก็บการประมวลผลและรักษาความปลอดภัยให้ข้อมูล อีกทั้งยังทำให้สามารถควบคุมการเข้าถึง และการประมวลผลที่รวดเร็วเพื่อสนองต่อความต้องการขององค์กร

4.2 SQL Server Analysis Services เป็นเทคโนโลยีที่ใช้ในการทำ Online Analytical Processing (OLAP) และ การทำเหมืองข้อมูลเพื่อนำไปใช้ด้านธุรกิจ Analysis Service สนับสนุนการทำ OLAP โดยให้ผู้ใช้สามารถออกแบบ สร้าง และจัดการข้อมูลที่ได้จากการรวบรวมข้อมูลจากแหล่งอื่นๆ เข้าด้วยกัน สำหรับการทำเหมืองข้อมูลแล้ว Analysis Service ให้ผู้ใช้สามารถออกแบบ

สร้าง และดูตัวแบบเหมืองข้อมูลที่สร้างได้ โดยสามารถนำข้อมูลจากแหล่งอื่นๆ มาทำเหมืองข้อมูลด้วยขั้นตอนวิธี ที่เป็นมาตรฐานและใช้กันอย่างกว้างขวาง

4.3 SQL Server Integration Services เป็นเทคโนโลยีในการแปลงและรวบรวมข้อมูลระดับองค์กร ซึ่งผู้ ใช้สามารถนำเครื่องมือไปทำการสกัด แปลง และรวบรวมข้อมูลจากแหล่งข้อมูลหลายแหล่งที่อยู่แยกจากกันได้แล้วนำข้อมูลนี้ไปไว้ปลายทางที่เดียวกัน หรือแยกกันหลายๆ ที่ก็ได้

4.4 SQL Server Replication เป็นชุดของเทคโนโลยีที่มีไว้สำหรับการสำเนา และกระจายข้อมูลหรือวัตถุในฐานข้อมูลจากฐานข้อมูลหนึ่งไปยังฐานข้อมูลอีกแห่งหนึ่ง และทำให้ข้อมูลในฐานข้อมูลหลายๆ แห่งมีความต้องกัน (Consistency) นอกจากนี้ยั้งช่วยให้สามารถกระจายข้อมูลไปยังที่ต่างๆ และสามารถใช้งานข้อมูลเหล่านั้น ผ่านระบบเครือข่าย เชื่อมโยงระยะใกล้, เครือข่ายเชื่อมโยงระยะไกล, เครือข่ายไร้สาย หรือเครือข่ายอินเตอร์เน็ตได้

4.5 SQL Server Reporting Services เป็นเทคโนโลยีที่ทาให้สามารถสร้างรายงานจากข้อมูลที่อยู่ในแหล่งข้อมูลเชิงสัมพันธ์หรือแหล่งข้อมูลหลายมิติได้ โดยสามารถสร้างและจัดการรายงานที่อยู่ในรูปแบบ Tabular หรือ Matrix หรือ Graphic หรือ Free-Form ได้ ซึ่งรายงานที่สร้างขึ้นมานี้สามารถดูและจัดการผ่านเว็บเพ็จได้

4.6 SQL Server Notification Services เป็นเทคโนโลยีสาหรับสร้างข้อความแจ้งเตือนและนาส่งข้อความนั้นไปให้ผู้ใช้จำนวนมากในเวลาอันรวดเร็ว

4.7 SQL Server Service Broker เป็นเทคโนโลยีที่ช่วยให้งานที่กระทำบนฐานข้อมูลนั้นมีความไว้วางใจได้ สามารถปรับขนาดงานได้ และทำให้งานมีความปลอดภัย ซึ่ง Service Broker นั้นเป็นเทคโนโลยีที่อยู่ในส่วนของ Database Engine อีกที Service Broker จะใช้การสื่อสารด้วยข้อความเพื่องานต่างๆ ที่อยู่แยกกันให้ สามารถทำงานร่วมกันได้ นอกจากนี้ Service Broker ยังมีเครื่องมือต่างๆ ที่จำเป็นสำหรับสร้างงานบนฐานข้อมูลแบบกระจาย และเครื่องมือเหล่านี้ช่วยลดระยะเวลาในการพัฒนาฐานข้อมูล ลงได้มาก นอกจากนี้ Service Broker ยังช่วยให้ผู้ใช้สามารถปรับขนาดของงาน ให้เหมาะสมกับปริมาณงานที่ได้รับ

4.8 SQL Server Tools and UtilitiesSQL Server มีเครื่องมือต่างๆ ที่ช่วยให้ผู้ใช้ออกแบบ พัฒนา นาไปประยุกต์ใช้ และบริหารฐานข้อมูลเชิงสัมพันธ์ , Analysis Services Cube, Data Transformation Packages, Replication Technologies, Reporting Servers และ Notification Servers ได้

5. งานวิจัยที่เกี่ยวข้อง

การนำเหมืองข้อมูลมาใช้เพื่อสกัดหาองค์ความรู้ใหม่เพื่อใช้ประกอบการตัดสินใจในการดำเนินกิจกรรมขององค์กรนั้นๆ ได้เป็นที่นิยมพึ่งขึ้นอย่างต่อเนื่อง ทำให้ผู้สนใจเริ่มทำการศึกษาค้นคว้าด้านนี้มากขึ้น อาทิ

พิจิตราจอมศรี (2549) ได้ศึกษาการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปกร โดยมีวัตถุประสงค์เพื่อ 1) ศึกษาการทำงานของเหมืองข้อมูล 2) พัฒนาโมเดลเพื่อใช้ในการทำนายแนวโน้มการใช้งานเว็บในอนาคต โดยใช้เทคนิคการทำเหมืองข้อมูล 3) ทดสอบ

ความถูกต้องโมเดลที่พัฒนาขึ้น 4) เพื่อนำโมเดลที่พัฒนาขึ้นมาประยุกต์ใช้ โดยข้อมูลที่ใช้ในการทำวิจัย มีการจัดเก็บข้อมูล 2 ประเภท คือ จัดเก็บข้อมูลจริงจากการเรียกใช้งานเว็บไซต์ในระบบพร็อกซีเซิร์ฟเวอร์ของมหาวิทยาลัยศิลปกร วิทยาเขตพระราชวังสนามจันทร์ ระหว่างเดือนกันยายน – พฤษจิกายน พ.ศ. 2548 และข้อมูลจากการจัดทำฐานข้อมูลเว็บเพื่อจัดหมวดหมู่เว็บ วิธีการวิจัยโดยนำข้อมูลทั้ง 2 ส่วนมาหาความสัมพันธ์เพื่อสร้างต้นแบบ โดยพิจารณาตัวแบบจากค่าความเชื่อมั่นและค่าสนับสนุน และนำข้อมูลที่ได้ทำการทดสอบความถูกต้องของตัวแบบ ซึ่งจากงานวิจัยพบว่า โมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่ถูกเรียกใช้ได้โดยมีความถูกต้อง 66.67% และผู้วิจัยกล่าวว่า จากงานวิจัยนี้สามารถทำนายเนื้อหาเว็บที่จะถูกเรียนใช้ในอนาคตได้ และสามารถเพิ่มประสิทธิภาพการทำงานของระบบพร็อกซี่เซิร์ฟเวอร์ได้ ทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและลดปริมาณข้อมูลในระบบเครือข่ายได้ แต่อย่างไรก็ตามเทคนิคนี้ยังไม่สามารถครอบคลุมการทำงานในช่วงเหตุการณ์ไม่เป็นปกติ เช่น อุบัติภัย และเทศกาลต่างๆ ได้

พันธ์รัตน์ อักษรศรีกุล และศิฬาณี นุชิตประสิทธิ์ชัย (2552) ได้ศึกษาระบบคลังข้อมูลจาก Log File ของการใช้อินเทอร์เน็ต โดยมีวัตถุประสงค์เพื่อออกแบบคลังข้อมูลสำหรับจัดเก็บข้อมูล Log file ของการใช้อินเทอร์เน็ต ซึ่งได้ใช้ข้อมูลจากกลุ่มผู้ใช้ข้าราชการกองทัพเรือ ที่เป็นสมาชิกระบบอินเทอร์เน็ต 26,409 นาย (ข้อมูล ณ วันที่ 1 กรกฎาคม 2551) ซึ่งงานวิจัยนี้เสนอการออกแบบระบบสารสนเทศ และการพัฒนาคลังข้อมูลจากล็อกไฟล์ของการใช้อินเทอร์เน็ตที่อยู่ในลักษณะ Text-Based file ให้อยู่ในลักษณะฐานข้อมูลเชิงสัมพันธ์ และใช้เทคนิค Online-Analytic Processing (OLAP) ในการวิเคราะห์ข้อมูลในมุมมององค์ประกอบที่เกี่ยวข้อง โดยนำเสนอผ่านเว็บบราวเซอร์ และทำการประเมินตามวิธีของไลเคอร์ท พบว่าผลการประเมินด้านความสามารถในการทำงานตรงตามที่ต้องการมีค่าเฉลี่ย 4.07 ด้านหน้าที่ของระบบมีค่าเฉลี่ย 4.25 ด้านการใช้งานระบบมีค่าเฉลี่ย 4.11 ด้านประสิทธิภาพของระบบมีค่าเฉลี่ย 4.13 และด้านความปลอดภัยของระบบมีค่าเฉลี่ย 4.17 และสรุปได้ว่าพบว่าผู้ใช้งานระบบมีความพึงพอใจต่อระบบอยู่ในระดับดีระบบสามารถนำไปใช้งานได้จริง

ดาวพระศุกร์ ฤทธิบัณฑิตย์ (2554) ได้ศึกษาระบบสนับสนุนการวิเคราะห์ข้อมูลจาราจรคอมพิวเตอร์ โดยมีวัตถุประสงค์เพื่อพัฒนาระบบสนับสนุนการวิเคราะห์ข้อมูลจราจรคอมพิวเตอร์ โดยการเก็บรวบรวมข้อมูลล็อกไฟล์ในลักษณะของเท็กซ์ไฟล์จากเครื่องพร็อกซี่เซิร์ฟเวอร์ ได้ศึกษาจากกลุ่มตัวอย่างนักศึกษาและบุคลากรภายในหน่ายงานวิทยาลัยอาชีวศึกษาสระบุรี ประมาณ 3,500 คน โดยวิธีการค้นหารูปแบบความสัมพันธ์ของข้อมูล และได้พัฒนาระบบเว็บแอพพลิแคชั่นเพื่อใช้งานและใช้แบบสอบถามเป็นเครื่องมือในการประเมินและทดสอบระบบ ซึ่งแบบออกเป็น 2 กลุ่ม ได้แก่ กลุ่มผู้เชี่ยวชาญ 5 คน และกลุ่มผู้ใช้งานทั่วไป 30 คน พบว่าผลการประเมิน และการทดสอบคุณภาพของระบบโดยผู้เชียวชาญ มีคุณภาพเหมาะสมในระดับดี มีค่าเฉลี่ย 3.71 และส่วนเบี่ยงเบนมาตรฐาน 0.52 และผลการประเมินความพึงพอใจของผู้ใช้งาน พบว่าอยู่ในระดับดี มีค่าเฉลี่ย 3.59 และส่วนเบี่ยงเบนมาตรฐาน 0.60 และสามารถนำระบบไปประยุกต์ใช้ได้อย่างมีความเหมาะสมในระดับดี และสำเร็จตามสมมติฐาน

Houqun Yang, Jingsheng Lei, Fa Fu(2007) ได้ศึกษาวิจัยเรื่อง An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining โดยเสนอวิธีการจัดกลุ่มเพื่อวิเคราะห์โครงสร้างของเว็บไซต์ด้วยเทคนิคการทำเหมืองข้อมูลเว็บ ในการทำวิจัยครั้งนี้ได้ใช้ข้อมูล

จากการบันทึกจากล็อกไฟล์เว็บไซต์ของมหาวิทยาลัยประจำเดือนกุมภาพันธ์ 2005 ซึ่งมีข้อมูลมากกว่า 540,000 แถว และได้ทำการตรวจสอบความถูกต้องของข้อมูล (Clean) จนได้ข้อมูลที่ถูกต้องซึ่งเหลือเพียง 70,000 แถว และมีเพียง 2,864 ที่เกี่ยวข้องกับหน้าเว็บที่มีความแตกต่างกัน จากนั้นได้นำข้อมูลมาทำการวิเคราะห์หากฎความสัมพันธ์ของเส้นทางบนเว็บไซต์ผลของการค้นหาความสัมพันธ์มีทั้งหมด 10 กฎ และพบว่ากฎข้อแรกคือผู้ใช้เข้าหน้า default.asp จะเข้าหน้า jiaoxue.asp ต่อ โดยมีค่าสนับสนุน 22.15 % และค่าความเชื่อมั่น 65% และจากผลการวิจัยสามารถนำความรู้ที่ได้นำไปปรับปรุงโครงสร้างของเว็บไซต์เพื่อให้เหมาะสมกับการให้บริการได้

Azizul Azhar bin Ramli (2005) ได้ศึกษาเรื่อง WEB USAGE MINING USING APRIORI ALGORITHM: UUM LEARNING CARE PORTAL CASE โดยมีวัตถุประสงค์ เพื่อศึกษา 1) การเตรียมข้อมูลจากล็อกเซิร์ฟเวอร์ E-Learning ของเว็บไซต์ UUM Educare เพื่อกำหนดและค้นหารูปแบบการเข้าถึงของผู้ใช้ 2) การทำเหมืองข้อมูลโดยใช้กฎความสัมพันธ์ด้วยอัลกอริทึม Apriori เพื่อการผลิตรูปแบบการใช้โดยกำหนดสนใจของผู้ใช้ 3) การวิเคราะห์รูปแบบการใช้ และพฤติกรรมของผู้งานเว็บไซต์ ซึ่งข้อมูลที่นำมาใช้ในการวิเคราะห์เป็นข้อมูลล็อกไฟล์ระหว่างวันที่ 19 กุมภาพันธ์ 2004 ถึงวันที่ 13 มีนาคม 2004 ของเซิร์ฟเวอร์ UUM Educare (www.e-web.uum.edu.my) มีจำนวนข้อมูลทั้งหมด 10,578 ทรานเซชันและได้ทำข้อมูลมาวิเคราะห์หากฎความสัมพันธ์โดยใช้อัลกอริทึม Apriori และได้กำหนดค่าสนับสนุนเท่ากับ 15% และค่าความเชื่อมั่นเท่ากับ 70% พบว่า ผู้ที่เข้าใช้หน้า /announcement และ หน้า /main จะเข้าใช้หน้า /dms ต่อ ซึ่งมีค่าสนับสนุนมากที่สุด 22.0% และผู้ที่เข้าใช้หน้า /announcement และหน้า /dms จะเข้าใช้หน้า /main ต่อ มีค่าความเชื่อมั่นมากที่สุด 99.1% และจากผลการวิจัยนี้ผู้ดูแลเว็บสามารถนำไปประยุทธ์ใช้สำหรับการออกแบบและปรับปรุงเว็บไซต์ที่เหมาะสม

Karuna P.Joshi และคณะได้ทำวิจัยเรื่อง On Using aWarehouse to AnalyzeWeb Logs โดยศึกษาการสร้างคลังข้อมูลเพื่อใช้ในการวิเคราะห์ข้อมูลจากเว็บล็อกของเว็บไซต์มหาวิทยาลัยแมรี่แลนด์ (umbc.edu) และได้นำเสนอการรายงานผลการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลปและค้นหาแนวโน้มของการใช้งานเว็บไซต์ด้วยการทำเหมืองข้อมูลเว็บ เพื่อนำมาใช้ในการพัฒนาปรับปรุงเว็บไซต์ให้เหมาะสมกับความต้องการของผู้ใช้ โดยใช้เทคนิคการหากฎความสัมพันธ์ด้วยอัลกอริทึม Apriori และการจัดกลุ่มของผู้ใช้ด้วยใช้อัลกอริทึม C-medoids ที่พัฒนาโดย Krishnapuram และคณะ ซึ่งผลลัพธ์จากการทำเหมืองข้อมูลด้วยอัลกอริทึมทั้ง2 อัลกอริทึมมีดังนี้

1) การหากฎความสัมพันธ์ (Association rules) โดยการกำหนดค่าสนับสนุน (Support) 80% และค่าความเชื่อมั่น (Confidence) พบว่าหน้าเว็บไซต์ที่มีความสัมพันธ์ 3 ลำดับแรกคือ

(1) หน้า /Search/ มีความสัมพันธ์กันกับ /Directory/,/StudentLink/, /FacAcademics/Depart/,/AboutUMBC/Schedule/summer1999/

(2) หน้า /AboutUMBC/Schedule/spring1999/ มีความสัมพันธ์กันกับหน้า/StudentLink/, /LibComp, /UnderGrad,/FacAcademics/Depart/, /AboutUMBC/Schedule/,/AboutUMBC/Schedule/summer1999/

(3) หน้า /Directory/ มีความสัมพันธ์กันกับหน้า /Search, /StudentLink/, /Admissions/, /FacAcademics/,/FacAcademics/departs.html, /FacAcademics/depart/

2) *การจัดกลุ่ม (Clustering)* โดยนำข้อมูลจากกลุ่มตัวอย่างที่เข้าใช้หน้าเว็บ Agents (http://www.csee.umbc.edu/agents/) พบว่าการจัดกลุ่มการเข้าใช้งานเว็บไซต์มีจำนวน 5 กลุ่ม ประกอบด้วย กลุ่มที่ 1 คือกลุ่มผู้ใช้ที่เข้าหน้าเพจ archive กลุ่มที่ 2 และกลุ่มที่ 3 เป็นกลุ่มที่มีขนาดเล็กที่สุดกลุ่มที่ 4 เป็นกลุ่มที่สนใจข้อมูลของข่าวสาร อาทิ News, FAQs, Technology กลุ่มที่ 5 เป็นกลุ่มของผู้ใช้ทั่วไปที่เข้าชมหน้าหลัก และการเชื่อมโยงลิงค์อยู่ในหน้าเว็บนั้นดังภาพที่2.9

Cluster no.	Cluster cardinality (no. of sessions)	No. of URLs in the cluster	URLs	Degree with which URL belongs to the cluster
1 Archives	19	56	{/agentslist/archive/digests} or {/agentslist/archive/digests/}	0.21
			{/agents/}	0.158
			{ /agentslist/archive/}	0.105
			Other URLs	<0.05
2	2	8	{/agents/commercial}	1.0
			{/agentslist/archive/1996..}	0.50
			{/agents/papers/migratory.html}	0.50
3	2	6	{/agents/}	1.0
			{/agents/groups}	0.5
			{/agents/commercial}	1.0
			{/agents/interface}	1.0
4 News, FAQs, Technology	61	151	{/agents/} or {/agents}	0.72
			{/agentslist/} or {/agentslist}	0.361
			{/agents/news/}	0.213
			{/agents/technology/}	0.18
			{/agents/faq/}	0.164
			{/agents/agentnews/}	0.115
			Other URLs	<0.1
5 General browser	83	257	{/agents/}	0.783
			{/agents/introduction/}	0.265
			{/agents/papers/}	0.169
			{/agents/web/}	0.132
			{/agents/papers/collections.shtml}	0.132
			{/agents/theory/}	0.12
			{/agents/mobile/}	0.12
			{/agents/faq/}	0.11
			{/agents/introduction/jennings98.pdf}	0.096
			{/agents/ standards/}	0.096
			{/agents/news/}	0.096
			Other URLs	<0.09

ภาพที่ 2.10 แสดงผลการจัดกลุ่ม (Clustering) การจัดกลุ่มการเข้าใช้งานเว็บไซต์

ซึ่งผลลัพธ์จากการทำวิจัยสามารถนำไปใช้ประโยชน์ในการจัดทำรายงาน การพัฒนาปรับปรุงและออกแบบโครงสร้างเว็บไซต์เพื่อให้ตรงตามความต้องการของผู้ใช้ในแต่ละกลุ่มได้

Chaofeng Li และคณะ ได้ทำวิจัยเรื่อง Similarity Measurement of Web Sessions by Sequence Alignment โดยมีวัตถุประสงค์เพื่อศึกษาการจัดกลุ่มผู้เข้าใช้เว็บไซต์ที่คล้ายคลึงกันด้วยวิธีการลำดับการจัดเรียง (Sequence) โดยวิเคราะห์ข้อบกพร่องของวิธีการแบบเดิมที่วัดความคล้ายคลึงกันระหว่างหน้าเว็บตามหน้าเพจ (URL) และวัดความคล้ายคลึงกันระหว่างหน้าเว็บตามเวลาที่เข้าใช้ (Viewing Time) และได้นำเสนอวิธีการใหม่ในการวัดความคล้ายคลึงกันโดย ใช้ข้อมูลการเข้าใช้งานเว็บไซต์จากเว็บเซิร์ฟเวอร์ของห้องสมุดของ South-Central University for Nationalities ระหว่างวันที่ 21 พฤษภาคม 2006 ถึง 28 พฤษภาคม 2006 ซึ่งมีขนาด 133 MB มีจำนวนข้อมูลดิบ 769,621 เรคอร์ด และดำเนินการขั้นตอนการตรวจสอบความถูกต้องสมบูรณ์และกำจัดข้อผิดพลาดของข้อมูลทำให้เหลือข้อมูลเพียง 115,645 เรคอร์ดแบ่งเป็นจำนวนของผู้ใช้ 53,094 เรคอร์ด จำนวนของ web sessions56,351 เรคอร์ด ผลจากการแบ่งกลุ่มสามารถแบ่งกลุ่มได้ 10 กลุ่ม และยังกล่าว

ว่าการจัดกลุ่มเว็บไซต์เป็นเทคนิคที่มีประโยชน์ที่สามารถจัดกลุ่มเว็บไซต์กลุ่มเดียวกันที่มีลักษณะคล้ายกัน และกลุ่มเว็บไซต์ที่มีลักษณะแตกต่างกันได้ และหลังจากการตรวจสอบประสิทธิภาพของอัลกอริทึมยังพบปัญหาเดิมอยู่ อาทิ การวิเคราะห์เวลาและหน่วยความจำที่ต้องใช้ในการประมวลผล (Time and Space Complexity)

บทที่ 3
วิธีดำเนินงานวิจัย

งานวิจัยนี้มุ่งเน้นการประยุกต์ใช้การทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์ กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง จากข้อมูลล็อกไฟล์ของการใช้งานเว็บไซต์ การทำเหมืองข้อมูลดำเนินการตามขั้นตอนของแบบจำลองคริสพ์ ซึ่งรายละเอียดของวิธีดำเนินการวิจัย ประกอบด้วย 3 ส่วน ได้แก่

1. ประชากรและกลุ่มตัวอย่าง
2. เครื่องมือที่ใช้ในการดำเนินงานวิจัย
3. ขั้นตอนการดำเนินงาน

1. ประชากรและกลุ่มตัวอย่าง

1.1 ประชากร

ข้อมูลผู้เข้าใช้บริการเว็บไซต์ของมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ได้แก่ บุคลากร นักศึกษา และบุคคลทั่วไปที่จัดเก็บในเครื่องเว็บเซิร์ฟเวอร์ (Web Server) ระหว่างวันที่ 1 มกราคม 2556 ถึงวันที่ 31 ธันวาคม 2556 มีจำนวน 55,685,313 เรคอร์ด

1.2 กลุ่มตัวอย่าง

ใช้ประชากรเป็นกลุ่มตัวอย่าง

2. เครื่องมือที่ใช้ในการดำเนินงานวิจัย

เครื่องมือที่ใช้ในงานวิจัยนี้ประกอบด้วย 2 ส่วน คือ

2.1 เครื่องมือการสร้างคลังข้อมูล (Data warehouse Development Tools) ใช้โปรแกรม Microsoft SQL Sever 2008 สร้างคลังข้อมูลโดยผ่านกระบวนการอีทีแอล (ETL Process) เพื่อจัดเก็บข้อมูลการใช้เว็บของมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง โดยข้อมูลดังกล่าวจะถูกจัดเก็บไว้ในล็อกไฟล์ที่เครื่องเว็บเซิร์ฟเวอร์

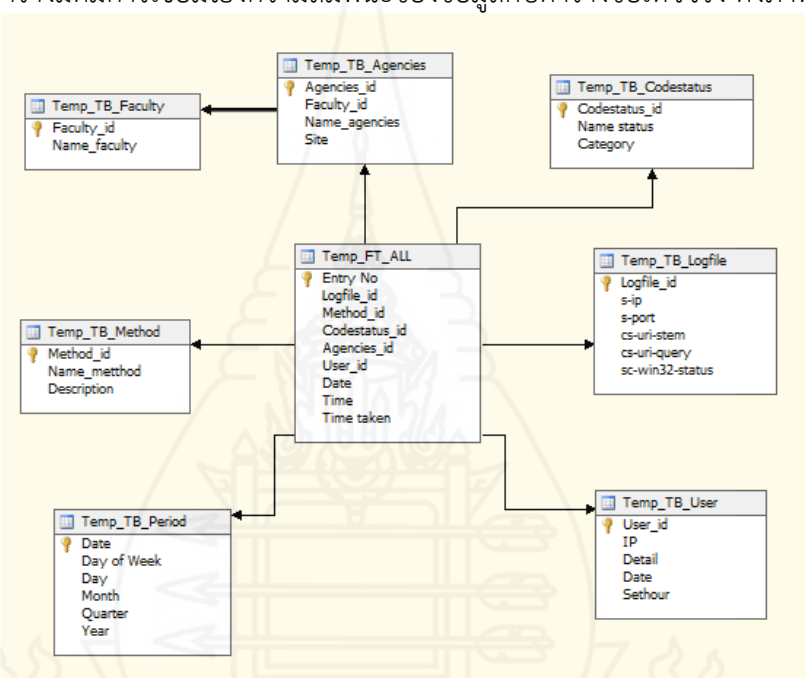
2.2 เครื่องมือทำเหมืองข้อมูล (Data Mining Tools) ใช้โปรแกรม Microsoft SQL Server 2008 Analysis Services (SSAS) เพื่อวิเคราะห์หาแพทเทิร์น (Pattern) ข้อมูลและความสัมพันธ์ของข้อมูล (Data association) รวมทั้งวิเคราะห์พฤติกรรมของผู้เข้าใช้บริการ

3. ขั้นตอนการดำเนินงาน

งานวิจัยนี้ประกอบด้วยขั้นตอนสำคัญ 3 ขั้นตอนคือ 1) การสร้างคลังข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์ 2) การสร้างคิวบ์ (cube) และ 3) การทำเหมืองข้อมูล

3.1 การสร้างคลังข้อมูลล็อกไฟล์การใช้งานเว็บไซต์

ขั้นตอนแรกของการสร้างคลังข้อมูลคือ ออกแบบคลังข้อมูลในการทำวิจัยครั้งนี้ ผู้วิจัยได้ออกแบบโครงสร้างของคลังข้อมูลโดยใช้แผนภาพอีอาร์ โดยใช้โครงสร้างคลังข้อมูลในรูปแบบสโนว์เฟลค สคีมา (Snowflake Schema) ที่มีตารางข้อเท็จจริงหนึ่งตาราง และมีตารางมิติอยู่รอบๆ ซึ่งตารางมิติมีการเชื่อมโยงความสัมพันธ์ของข้อมูลกับตารางข้อเท็จจริง ดังภาพที่ 3.1

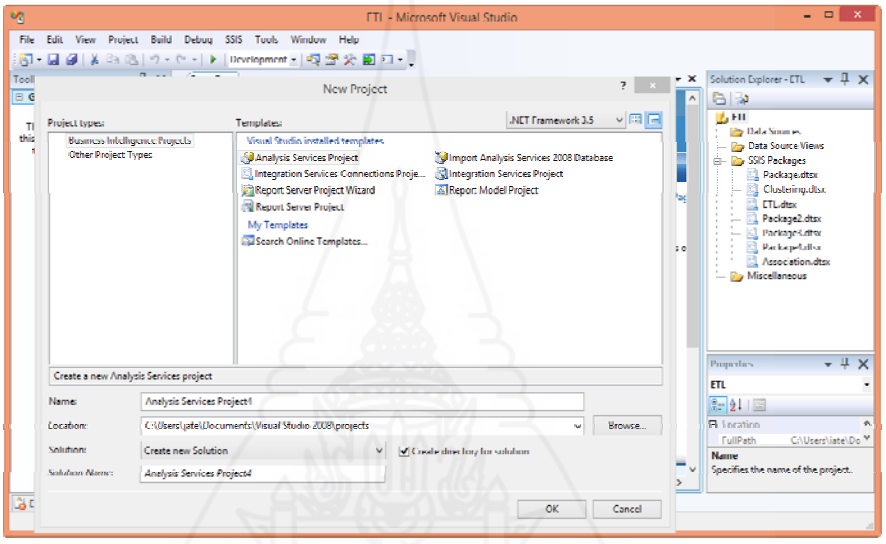


ภาพที่ 3.1 แผนภาพอีอาร์ โดยใช้โครงสร้างคลังข้อมูลในรูปแบบสโนว์เฟลค

เมื่อออกแบบโครงสร้างคลังข้อมูลเสร็จแล้ว ขั้นตอนต่อมาคือ กระบวนการนำเข้าข้อมูลสู่คลังข้อมูล คือกระบวนการอีทีแอล (Extract Transform and Load) ซึ่งประกอบด้วย 3 ขั้นตอน คือ 1) การคัดแยกข้อมูล (Extract - E) คือการดึงข้อมูลจากล็อกไฟล์การใช้งานเว็บจากเครื่องเว็บเซิร์ฟเวอร์ ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมาก เนื่องจากการดึงข้อมูลที่ถูกต้องจะทำให้การดำเนินงานในขั้นตอนต่อไปถูกต้องไปด้วย 2) การแปลงข้อมูล (Transform - T) คือการนำข้อมูลที่ได้จากขั้นตอนแรก “การคัดแยกข้อมูล” มาจัดอยู่ในรูปแบบที่สอดคล้องกัน การทำให้ข้อมูลที่มีความหมายเดียวกันแต่อยู่ในรูปแบบที่แตกต่างกันให้อยู่ในรูปแบบเดียวกัน รวมถึงการทำความสะอาดข้อมูล การตรวจสอบความถูกต้อง และการแก้ไขข้อมูลให้ถูกต้อง โดยกำจัดข้อมูลที่ผิดพลาดออกไป การแปลงข้อมูลยังรวมถึงการปรับปรุงรูปแบบของข้อมูลที่สามารถนำไปวิเคราะห์ได้ 3) การโหลด (Load – L) เป็นการนำข้อมูลล็อกไฟล์ที่ผ่านการแปลงข้อมูลที่ต้องกับความต้องการแล้ว เข้าสู่คลังข้อมูล (data warehouse)

ซึ่งผู้วิจัยได้นำข้อมูลล็อกไฟล์ของเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ซึ่งได้เก็บรวบรวมข้อมูลต่างๆ ของการเข้าถึงเว็บไซต์ของทางมหาวิทยาลัย นำมาสร้างเป็นคลังข้อมูลใหม่ โดยผ่านกระบวนการอีทีแอล เพื่อสกัดข้อมูลที่ต้องการและสร้างรูปแบบเพื่อที่จะนำไปใช้ในการทำเหมืองข้อมูลในขั้นตอนต่อไป

ผู้วิจัยได้ใช้เครื่องมือ Integration Service ของ Microsoft SQL Server 2008 เพื่อนำข้อมูลเข้าสู่คลังข้อมูลโดยใช้กระบวนการอีทีแอล คัดแยกข้อมูล การแปลงข้อมูลและการโหลดข้อมูล ดังภาพที่ 3.2 และมีรายละเอียดดังต่อไปนี้

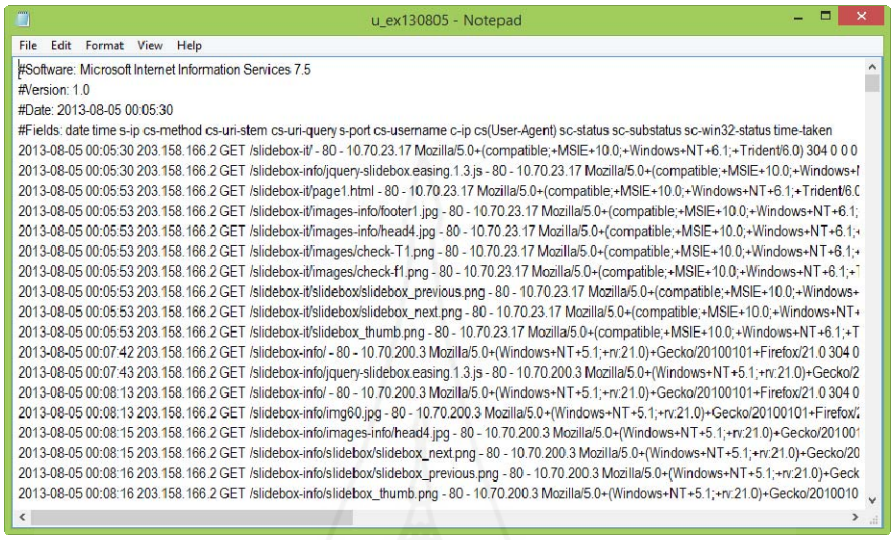


ภาพที่ 3.2 การใช้เครื่องมือ Integration Service ในกระบวนการอีทีแอล

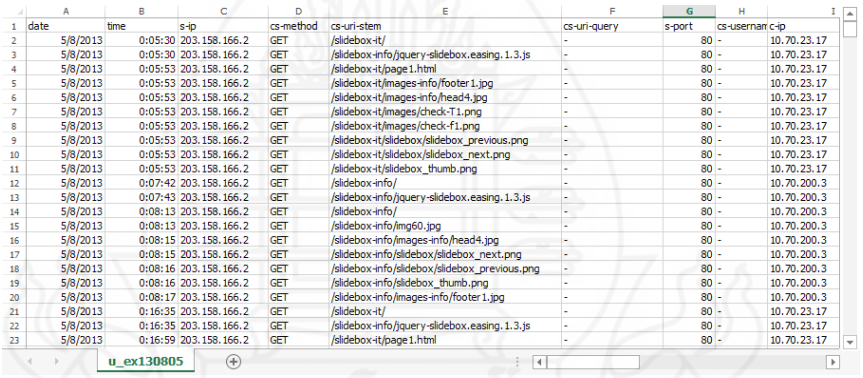
3.1.1 ขั้นตอนการคัดแยกข้อมูล (Extract - E) คือการดึงข้อมูลจากแหล่งข้อมูล ซึ่งข้อมูลในการทำวิจัยครั้งนี้เป็นข้อมูลล็อกไฟล์ (Log file)ของการเข้าใช้เว็บไซต์ของทางมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนาลำปาง ในระหว่างวันที่ 1 มกราคม 2556 –31 ธันวาคม2556 จำนวน 55,685,313 เรคอร์ดและข้อมูลอื่นๆ ที่เกี่ยวข้องโดยมีขั้นตอนในการดำเนินการต่อไปนี้

*1) จัดเตรียมข้อมูล*ที่ต้องการนำมาวิเคราะห์ตรงตามขอบเขตของงานวิจัย ซึ่งในงานวิจัยครั้งนี้ได้ใช้ข้อมูล ซึ่งประกอบไปด้วย 1) ล็อกไฟล์ (Logfile) ของการเข้าใช้งานเว็บไซต์ของทางมหาวิทยาลัย ในรูปแบบ W3C Extended log file format ที่อยู่ในเว็บเซิร์ฟเวอร์ ประกอบด้วย date, time, s-ip, cs-method, cs-uri-stem, cs-uri-query, s-port, cs-username, c-ip, cs(User-Agent), sc-status, sc-substatus, sc-win32-status, time-taken ข้อมูลเหล่านี้นำมาสร้างรูปแบบที่สามารถนำไปใช้ได้ โดยสร้างอยู่ในรูปแบบของไฟล์เอกสารเอ็กเซล (Excel) ดังภาพที่ 3.2 และภาพที่ 3.3 2) ข้อมูลหน่วยงานหลัก ซึ่งประกอบด้วย รหัสหน่วยงาน ชื่อหน่วยงานหลัก 3) ข้อมูลหน่วยงานย่อย ซึ่งประกอบด้วย รหัสหน่วยงานย่อย รหัสหน่วยงานหลัก ชื่อที่อยู่ของเว็บ 4) ข้อมูลรหัสสถานะของการรับส่งข้อมูล HTTP ซึ่งประกอบด้วย รหัส ชื่อสถานะ หมวดสถานะ 5) ข้อมูลวิธีการรับส่ง ซึ่ง

ประกอบด้วย รหัส ชื่อการรับส่ง รายละเอียด 6) ข้อมูลระยะเวลา (Period) ซึ่งประกอบด้วย ข้อมูลลำดับวันที่ ชื่อวัน วันที่ เดือน ไตรมาส ปี



ภาพที่ 3.3 ตัวอย่างข้อมูลล็อกไฟล์รูปแบบเท็กซ์ (Text)

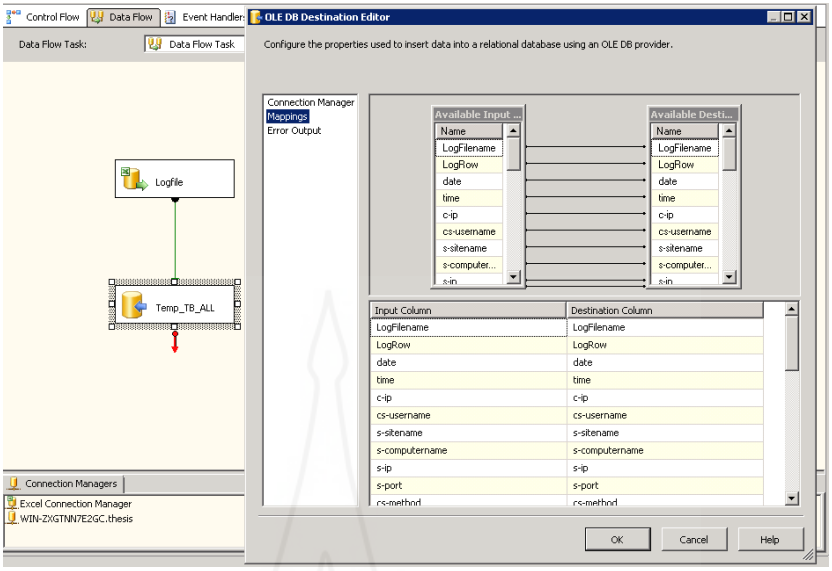


	A	B	C	D	E	F	G	H	I
1	date	time	s-ip	cs-method	cs-uri-stem	cs-uri-query	s-port	cs-username	c-ip
2	5/8/2013	0:05:30	203.158.166.2	GET	/slidebox-it/	-	80	-	10.70.23.17
3	5/8/2013	0:05:30	203.158.166.2	GET	/slidebox-info/jquery-slidebox.easing.1.3.js	-	80	-	10.70.23.17
4	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/page1.html	-	80	-	10.70.23.17
5	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/images-info/footer1.jpg	-	80	-	10.70.23.17
6	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/images-info/head4.jpg	-	80	-	10.70.23.17
7	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/images/check-T1.png	-	80	-	10.70.23.17
8	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/images/check-f1.png	-	80	-	10.70.23.17
9	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/slidebox/slidebox_previous.png	-	80	-	10.70.23.17
10	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/slidebox/slidebox_next.png	-	80	-	10.70.23.17
11	5/8/2013	0:05:53	203.158.166.2	GET	/slidebox-it/slidebox_thumb.png	-	80	-	10.70.23.17
12	5/8/2013	0:07:42	203.158.166.2	GET	/slidebox-info/	-	80	-	10.70.200.3
13	5/8/2013	0:07:43	203.158.166.2	GET	/slidebox-info/jquery-slidebox.easing.1.3.js	-	80	-	10.70.200.3
14	5/8/2013	0:08:13	203.158.166.2	GET	/slidebox-info/	-	80	-	10.70.200.3
15	5/8/2013	0:08:13	203.158.166.2	GET	/slidebox-info/img60.jpg	-	80	-	10.70.200.3
16	5/8/2013	0:08:15	203.158.166.2	GET	/slidebox-info/images-info/head4.jpg	-	80	-	10.70.200.3
17	5/8/2013	0:08:15	203.158.166.2	GET	/slidebox-info/slidebox/slidebox_next.png	-	80	-	10.70.200.3
18	5/8/2013	0:08:16	203.158.166.2	GET	/slidebox-info/slidebox/slidebox_previous.png	-	80	-	10.70.200.3
19	5/8/2013	0:08:16	203.158.166.2	GET	/slidebox-info/slidebox_thumb.png	-	80	-	10.70.200.3
20	5/8/2013	0:08:17	203.158.166.2	GET	/slidebox-info/images-info/footer1.jpg	-	80	-	10.70.200.3
21	5/8/2013	0:16:35	203.158.166.2	GET	/slidebox-it/	-	80	-	10.70.23.17
22	5/8/2013	0:16:35	203.158.166.2	GET	/slidebox-info/jquery-slidebox.easing.1.3.js	-	80	-	10.70.23.17
23	5/8/2013	0:16:59	203.158.166.2	GET	/slidebox-it/page1.html	-	80	-	10.70.23.17

u_ex130805

ภาพที่ 3.4 ตัวอย่างข้อมูลล็อกไฟล์รูปแบบไฟล์เอ็กเซล

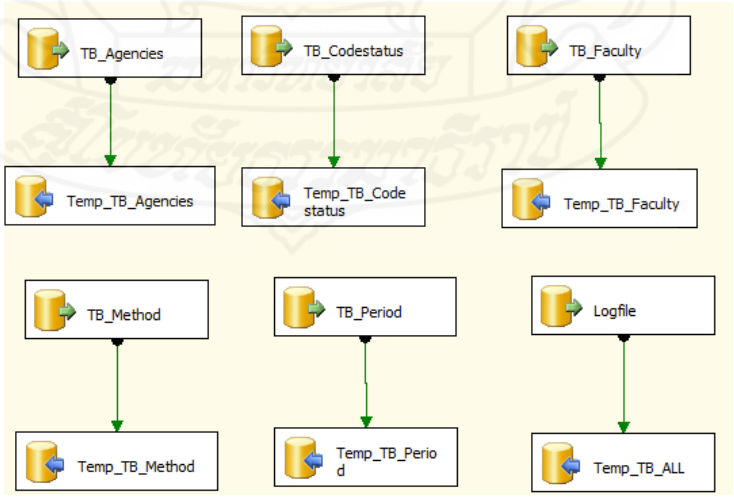
2) *นำข้อมูลจากไฟล์เอ็กเซล* และข้อมูลจากฐานข้อมูล เข้าสู่คลังข้อมูลที่ผู้วิจัยได้สร้างไว้ เพื่อเป็นแหล่งข้อมูลในการทำงานวิจัย โดยการนำเข้าข้อมูล ต้องกำหนดให้อยู่ในรูปแบบที่ผู้วิจัยได้ออกแบบไว้ และดำเนินการใช้เครื่องมือจัดการกับข้อมูล ซึ่งงานวิจัยนี้ผู้วิจัยได้ใช้ซอฟต์แวร์ SQL Server Integration Service (SSIS) เพื่อจัดการข้อมูลต่างๆ เข้าสู่ที่พักข้อมูลดังภาพที่ 3.5



ภาพที่ 3.5 การนำข้อมูลจากไฟล์เอ็กเซลเข้าตาราง

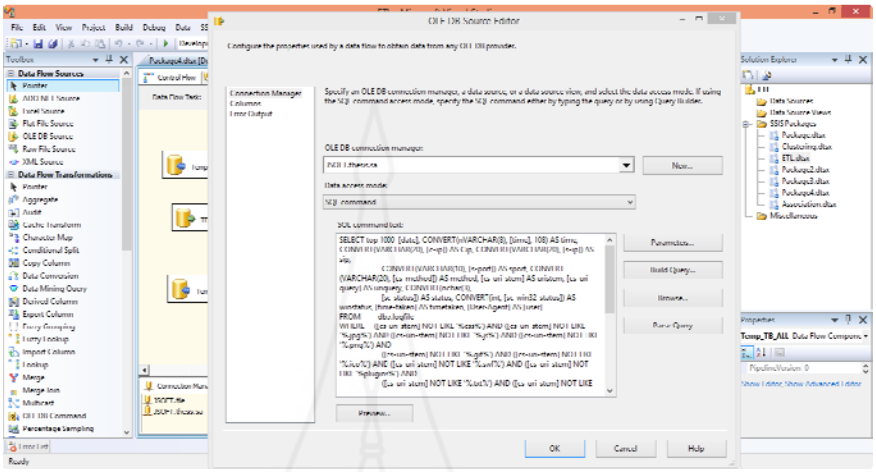
3.1.2 ขั้นตอนการแปลงข้อมูล (Transform - T) หรือการทำความสะอาดข้อมูล (Data cleansing) คือการนำข้อมูลจากกระบวนการแรกมาทำการตรวจสอบความถูกต้องของข้อมูล แล้วดำเนินการแก้ไขข้อมูลให้ถูกต้องและสอดคล้องกับข้อมูลที่ต้องการนำไปใช้ โดยกำจัดข้อมูลที่ผิดพลาดออกไป การแปลงข้อมูลยังรวมถึงการปรับปรุงรูปแบบของข้อมูลให้สามารถนำไปวิเคราะห์ได้ ข้อมูลในกระบวนการนี้จะดำเนินการโดยใช้ซอฟต์แวร์ SQL Server Integration Service (SSIS) มาดำเนินการในส่วนนี้และมีขั้นตอนการดำเนินงานดังนี้

*1) คัดเลือกข้อมูลที่ต้องการนำไปใช้*ในการทำงานวิจัย แล้วดำเนินการพักข้อมูลไว้ในที่พักข้อมูล (staging area) อีกครั้ง เพื่อดำเนินการในกระบวนการต่อไป ดังภาพที่3.6



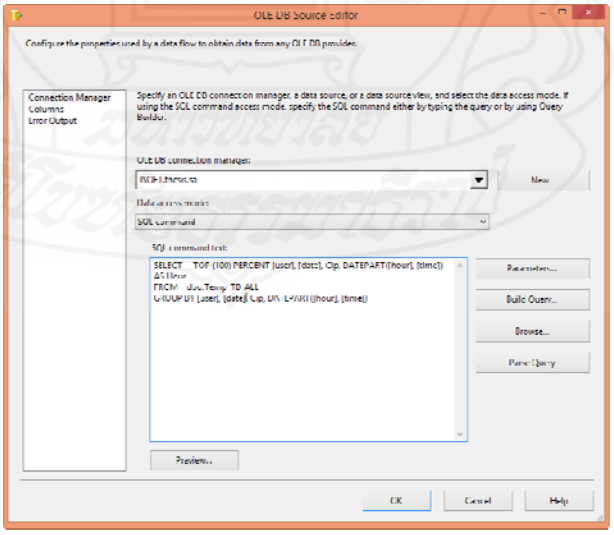
ภาพที่ 3.6 การคัดเลือกข้อมูลเข้าที่พักข้อมูล

*2) สกัดข้อมูลที่ไม่ต้องการ*ออกจากตารางพักข้อมูลล็อกไฟล์ เช่น ข้อมูลภาพ ข้อมูลไฟล์ จาวาสคริป (JavaScript) สไตล์ชีต (Style sheet) และไฟล์บ็อต (Bot) โดยใช้คำสั่ง SQL ดังภาพที่ 3.7



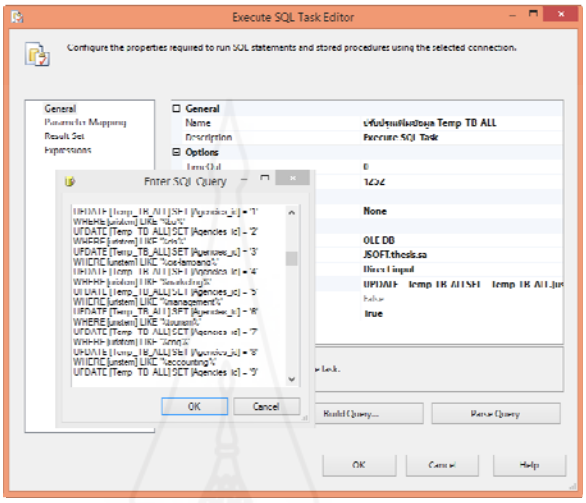
ภาพที่ 3.7 การใช้คำสั่ง SQL สกัดข้อมูลที่ไม่ต้องการ

*3) กำหนดรูปแบบข้อมูลชื่อผู้ใช้*เนื่องจากข้อมูลในล็อกไฟล์ไม่ได้ระบุชื่อผู้ใช้งานเว็บไซต์ ผู้วิจัยจึงต้องนำข้อมูลดังกล่าวมากำหนดชื่อผู้ใช้โดยใช้ชื่อที่อยู่ไอพี (IP Address) วันที่เข้าใช้ และเครื่องมือในการเข้าใช้ มาเป็นตัวแปรในการกำหนดชื่อผู้ใช้ดังกล่าว โดยส่วนนี้ใช้คำสั่ง SQL เพื่อหาชื่อผู้ใช้ และสร้างตาราง TB_Temp_user เพื่อเก็บข้อมูลแล้วดำเนินการปรับปรุงชื่อผู้ใช้ในตารางล็อกไฟล์ดังภาพที่ 3.8



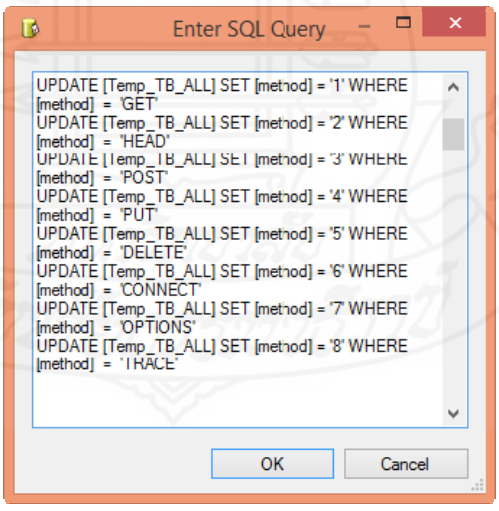
ภาพที่ 3.8 การใช้คำสั่ง SQL ค้นหาข้อมูลผู้ใช้

4) กำหนดรูปแบบหมวดหมู่ของหน้าเว็บ เพื่อให้สอดคล้องกับชื่อหน่วยงานของเว็บนั้นทางผู้วิจัยจึงได้กำหนดตัวแปร Agencies_id จัดเก็บรหัสของหน่วยงาน โดยใช้คำสั่ง SQL ในการปรับปรุงข้อมูลดังกล่าวดังภาพที่ 3.9



ภาพที่ 3.9 การใช้คำสั่ง SQL ปรับปรุงรูปแบบของข้อมูลหน่วยงาน

5) กำหนดรูปแบบข้อมูลรหัสสถานะของการรับส่งข้อมูล HTTP เพื่อให้สอดคล้องกับตารางที่ผู้วิจัยได้ออกแบบไว้โดยใช้คำสั่ง SQL ในการปรับปรุงตารางดังภาพที่ 3.10

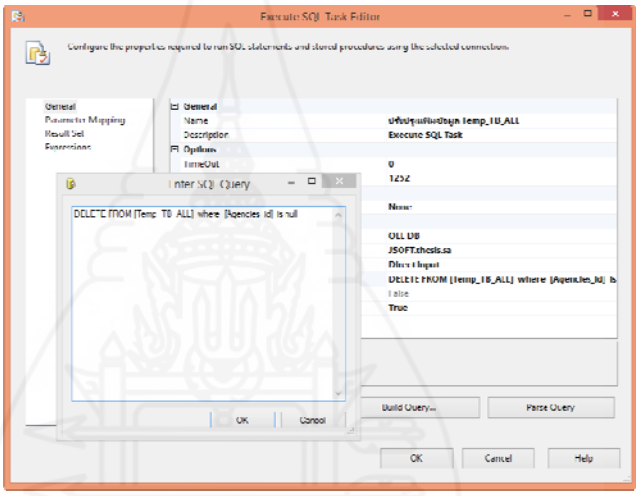


ภาพที่ 3.10 การใช้คำสั่ง SQL ปรับปรุงรูปแบบของรหัสสถานการณ์รับส่งข้อมูล

6) เมื่อกำหนดรูปแบบของข้อมูลต้องตามความต้องการ และสอดคล้องกับตารางข้อมูลในการทำวิจัยแล้วจะพักข้อมูลไว้ในที่พักข้อมูลจากนั้นจะนำข้อมูลมาดำเนินการตรวจสอบความถูกต้องของข้อมูล หรือหาข้อมูลที่ไม่ถูกต้อง หรือเรียกอีกอย่างหนึ่งว่าการทำความสะอาดข้อมูล

โดยจะคัดแยกข้อมูลไม่ถูกต้อง และไม่สอดคล้องกับการทำงานวิจัยโดยมีกระบวนการในการแก้ไขข้อมูลให้มีความถูกต้อง ได้แก่ การแก้ไขข้อมูลที่ไม่ต้องตามแหล่งอ้างอิง และข้อมูลที่เป็นค่าว่าง โดยวิธีการค้นหาข้อมูลและเปรียบเทียบกับแหล่งข้อมูลข้อมูลที่ต้องการ การกำจัดข้อมูลเหล่านั้นอาจจะด้วยวิธีการแก้ไข หรือลบรายการดังกล่าว เพื่อกำจัดข้อมูลที่มีปัญหาออกไป และไม่ให้เกิดความผิดพลาดในการนำข้อมูลเข้าสู่ระบบ

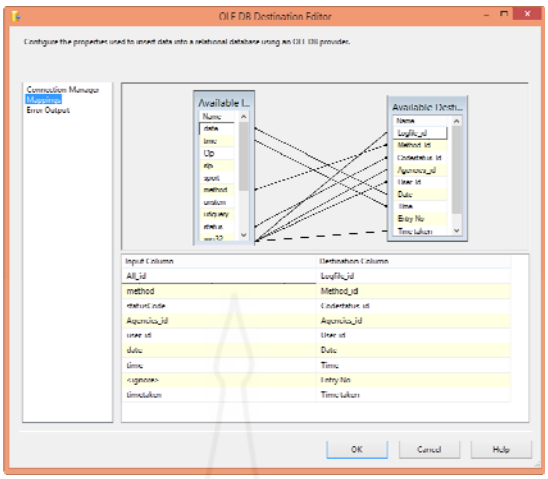
7) *งานวิจัยครั้งนี้ได้พบปัญหา*ในส่วนของข้อมูลที่ไม่สมบูรณ์อันเนื่องมาจากการกำหนดรหัสหน่วยงานในตาราง Temp_TB_ALL (ล็อกไฟล์) ซึ่งข้อมูล cs-uri-stem ไม่ตรงกับแหล่งอ้างอิงทำให้การปรับปรุง Agencies_id จึงเป็นค่าว่าง (NULL) ผู้วิจัยจึงได้กำจัดข้อมูลที่ผิดพลาดโดยวิธีการลบข้อมูลเหล่านั้น โดยใช้คำสั่ง SQL “DELETE FROM [Temp_TB_ALL] where [Agencies_id] is null” ดังภาพที่ 3.11



ภาพที่ 3.11 การใช้คำสั่ง SQL กำจัดข้อมูลที่ผิดพลาด

3.1.3 ขั้นตอนการโหลด (Load – L) เป็นการนำข้อมูลที่ผ่านการแปลงข้อมูลที่ตรงกับความต้องการ และตรวจสอบความถูกต้องข้อมูลเรียบร้อยแล้ว เข้าสู่คลังข้อมูล (data warehouse) เป็นการดำเนินการนำข้อมูลจากที่พักข้อมูล โดยข้อมูลเป็นข้อมูลที่ถูกต้องตามโครงสร้างที่ได้กำหนดไว้ และมีการทำความสะอาดข้อมูลที่ไม่ถูกต้องสมบูรณ์แล้ว โดยขั้นตอนในการโหลดข้อมูลประกอบด้วย

1) *การพิจารณาข้อมูลจากที่พักข้อมูล* โดยใช้เครื่องมือ “Data Flow Task” ในการนำข้อมูลเข้าคลังข้อมูล โดยเลือกข้อมูลจากที่พักข้อมูลแล้วกำหนดข้อมูลที่ต้องการจะนำลงคลังข้อมูล ดังภาพที่ 3.12



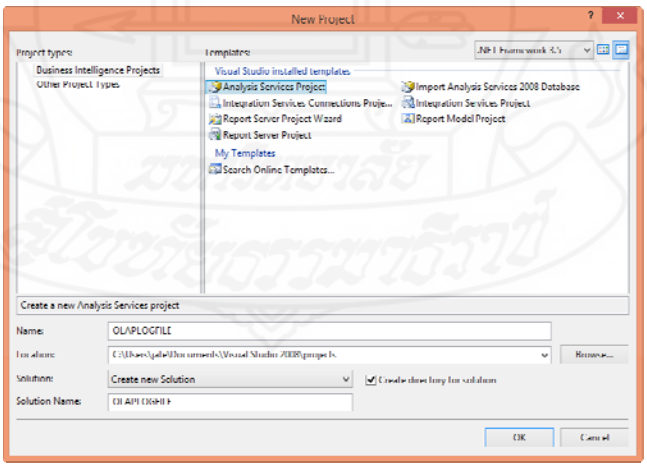
ภาพที่ 3.12 แสดงการโหลดข้อมูลเข้าตารางโดยการแม็ปข้อมูล

2) การดำเนินการโหลดข้อมูล หรือการนำข้อมูลจากที่พักข้อมูลภายหลังการกำหนดค่าที่ต้องการแล้ว ลงในคลังข้อมูลที่กำหนดไว้ชื่อ “DWLOGFILE”

3.2 การสร้างคิวบ์ (cube)

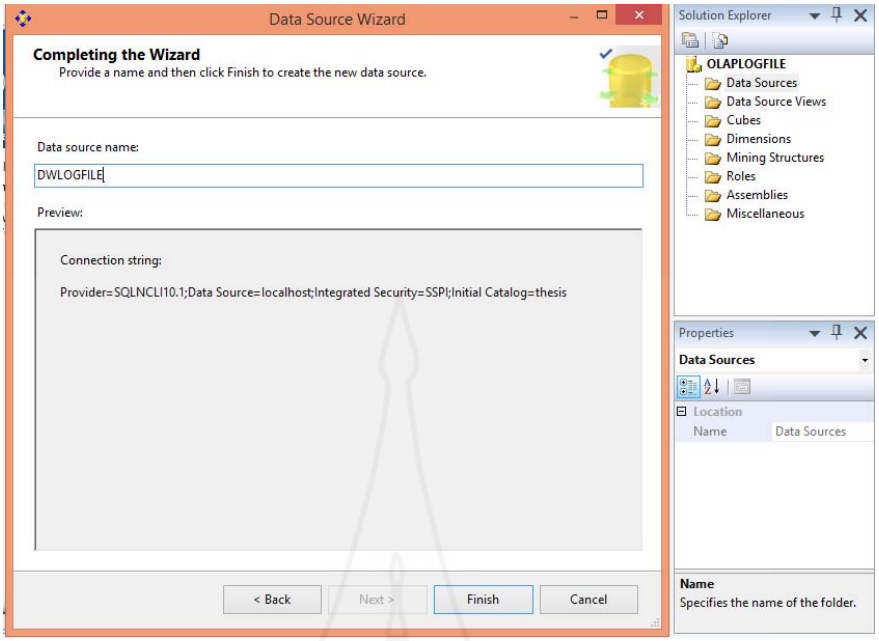
หลังจากได้ดำเนินการในส่วนของกระบวนการ ETL ด้วย Microsoft SQL Server Management Studio แล้ว ทางผู้วิจัยได้ใช้ Microsoft SQL Server 2008 Analysis Services (SSAS) ซึ่งเป็นเครื่องมือสำหรับการสร้างคิวบ์ (cube) โดยมีขั้นตอนการดำเนินการดังนี้

3.2.1 สร้างโปรเจคต์ (Project) ชื่อ “OLAPLOGFILE” โดยเลือกเท็มเพลต (Templates) เป็น Analysis Services Project ดังภาพที่ 3.13



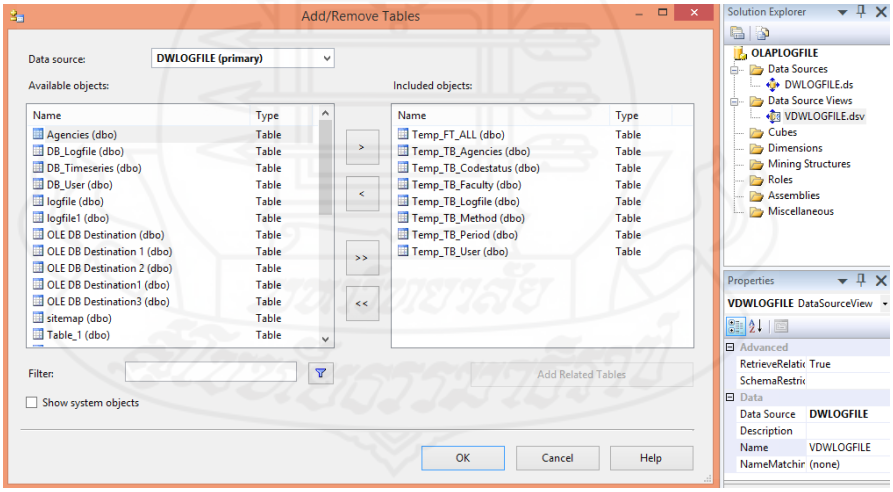
ภาพที่ 3.13 แสดงการสร้างโปรเจคต์ ด้วย Microsoft SQL Server Management Studio

3.2.2 สร้าง Data Source ชื่อว่า DWLOGFILE โดยระบุ Name Server และ Database ตามที่ต้องการดังภาพที่ 3.14



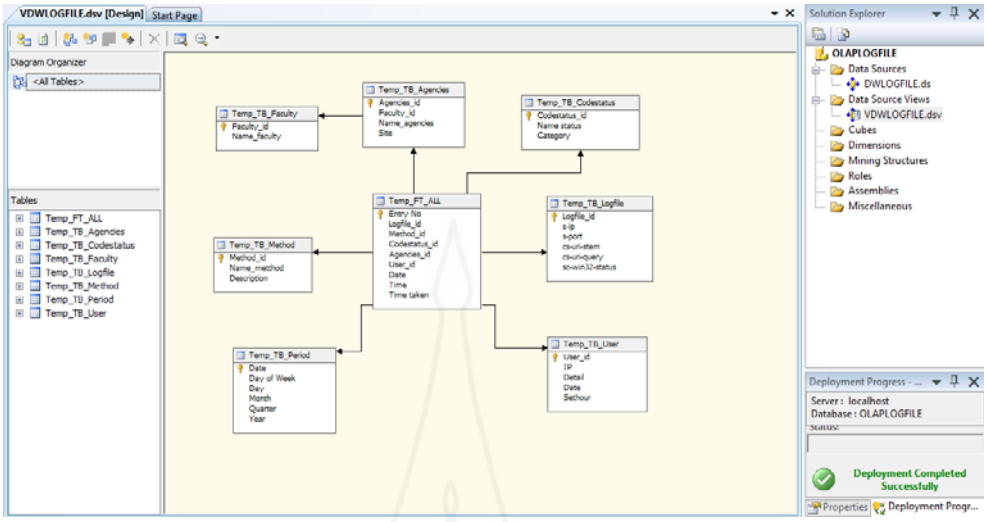
ภาพที่ 3.14 แสดงการสร้าง Data Source

3.2.3 สร้าง Data Source View ชื่อว่า VDWLOGFILE โดยระบุตารางตามที่ต้องการ จากฐานข้อมูลที่เลือกไว้ดังภาพที่ 3.15



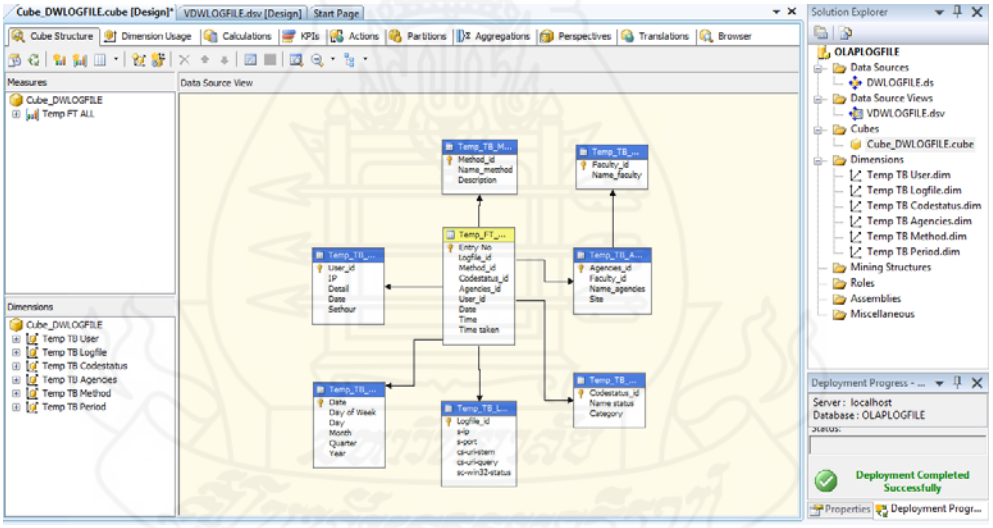
ภาพที่ 3.15 แสดงการเลือกตาราง ใน Data Source View

3.2.4 ตรวจสอบ และกำหนดความสัมพันธ์ของตารางให้ถูกต้องดังภาพที่ 3.16



ภาพที่ 3.16 แสดงโครงสร้าง และความสัมพันธ์ของตารางข้อมูล

3.2.5 สร้าง Cube เพื่อกำหนด Measures และ Dimensions ดังภาพที่ 3.17



ภาพที่ 3.17 แสดงการสร้าง Cube

3.2.6 Deploy โพรเจ็กต์

3.3 การทำเหมืองข้อมูล

การทำเหมืองข้อมูล คือกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล สำหรับงานวิจัยนี้การทำเหมืองข้อมูลดำเนินการตามแบบจำลอง คริสพ์-ดีเอ็ม (CRISP-DM Model) ประกอบด้วย 6 ขั้นตอนดังนี้

3.3.1 ทำความเข้าใจปัญหา ผู้วิจัยได้ศึกษางานวิจัยที่เกี่ยวข้องและศึกษาข้อมูลเกี่ยวกับเว็บไซต์ของมหาวิทยาลัย และวิเคราะห์ปัญหาที่เกิดขึ้นดังนี้

1) การออกแบบหน้าเว็บไซต์ และกลุ่มการจัดวางเมนูของการให้บริการข้อมูลของทางมหาวิทยาลัยไม่ตรงกับความต้องการของผู้ใช้งาน ทำให้การค้นหาบริการต่างๆ เกิดความล่าช้าและค้นหาได้ยาก

2) การเชื่อมโยงลิงค์ของหน้าเว็บ ไม่ตรงตามความต้องการของผู้เข้าใช้ เพราะมีจำนวนลิงค์อยู่เป็นจำนวนมากเพราะไม่สามารถทราบถึงความต้องการของผู้ใช้

*3) ไม่สามารถทราบถึงจำนวนผู้เข้าใช้ในช่วงเหตุการณ์*เช่น แต่ละเดือนมีผู้เข้าบริการมากน้อยเท่าไร เพื่อเตรียมความพร้อมให้การให้บริการ และเฝ้าระวังการทำงานของเครื่องเซิร์ฟเวอร์

4) ไม่สามารถทราบถึงพฤติกรรมการใช้งานเว็บไซต์ เพื่อจะนำข้อมูลเหล่านั้นมาดำเนินการปรับปรุงเพื่อให้เกิดความเหมาะสม

3.3.2 ทำความเข้าใจข้อมูล ผู้วิจัยได้รวบรวมข้อมูลจากการศึกษาถึงปัญหาดังกล่าวเพื่อนำไปวิเคราะห์ด้วยเทคนิคการทำเหมืองข้อมูลซึ่งการทำเหมืองข้อมูลมีอัลกอริทึมอยู่หลายอัลกอริทึม เพื่อใช้ในการวิเคราะห์ข้อมูลต่างๆ แต่ละอัลกอริทึมมีวัตถุประสงค์ที่แตกต่างกัน ผู้วิจัยจึงต้องศึกษาข้อมูลที่ได้จากการรวบรวมที่เกี่ยวข้อง ซึ่งเก็บไว้อยู่ในรูปแบบของล็อกไฟล์ว่ามีข้อมูลส่วนไหนบ้างที่สามารถนำไปวิเคราะห์ด้วยอัลกอริทึมต่างๆ ของการทำเหมืองข้อมูลได้ จากการศึกษาพบว่าข้อมูลที่จัดเก็บอยู่ในรูปแบบของล็อกไฟล์นั้น เป็นข้อมูลของการเข้าใช้เว็บไซต์ของทางมหาวิทยาลัยซึ่งประกอบด้วยแอ็ตทริบิวต์ต่างๆ สามารถนำไปวิเคราะห์โดยใช้เทคนิคการทำเหมืองข้อมูลได้

3.3.3 เตรียมข้อมูล ขั้นตอนนี้ผู้วิจัยได้ใช้กระบวนการอีทีแอล (Extract Transform and Load: ETL) โดยเรียกใช้เครื่องมือ Integration Service ของ Microsoft SQL Server มาใช้ในการจัดการข้อมูลเพื่อปรับโครงสร้างให้เหมาะสม สามารถนำไปใช้กับแบบจำลองการทำเหมืองข้อมูลของแต่ละอัลกอริทึมได้ ซึ่งการทำวิจัยครั้งนี้ผู้วิจัยได้เรียกใช้อัลกอริทึมในการวิเคราะห์ข้อมูล จำนวนประกอบด้วย 4 อัลกอริทึม คือ1) แอสโซซิเอชันรูลส์ (Association Rules) 2) ไทม์ซีรีส์ (Time Series) 3) ซีเควันซ์ คลัสเตอริง (Sequence Clustering) และ 4) คลัสเตอริง (Clustering) ซึ่งมีรายละเอียดต่อไปนี้

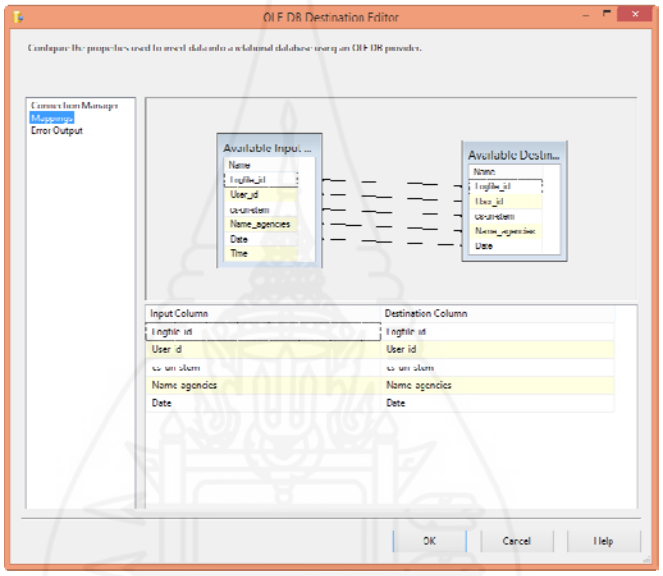
1) แอสโซซิเอชันรูลส์ (Association Rules) เป็นอัลกอริทึม เพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลเว็บว่ามีหน้าเว็บไหนที่ถูกเข้าถึงข้อมูลด้วยกัน ซึ่งประกอบด้วยแอ็ตทริบิวต์ต่างๆ ที่ผู้วิจัยนำมาใช้ในการวิเคราะห์ ดังข้อมูลตารางที่ 3.1, 3.2 และภาพที่ 3.18, 3.19

ตารางที่ 3.1 แสดงแอ็ตทริบิวต์ของตารางผู้ใช้

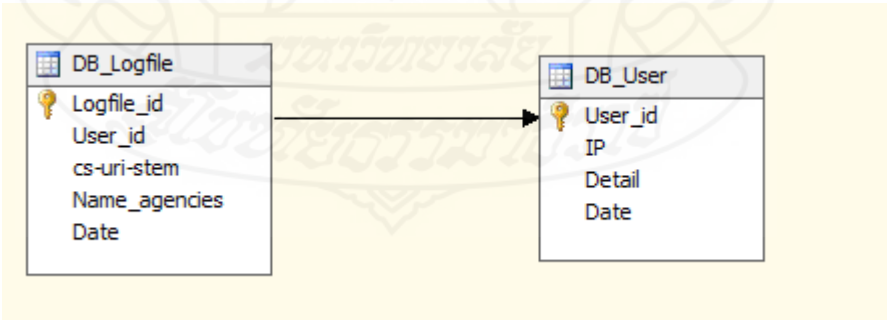
Col.	Name	Description
1	User_id	รหัสผู้ใช้
2	IP	ที่อยู่ไอพี
3	Detail	รายละเอียดเบราเซอร์
4	Date	วันที่เข้าใช้

ตารางที่ 3.2 แสดงแอ็ตทริบิวต์ของตารางล็อกไฟล์

Col.	Name	Description
1	Logfile_id	รหัสล็อกไฟล์
2	User_id	รหัสผู้ใช้
3	cs-uri-stem	ที่อยู่เว็บ
4	Name_agencies	ชื่อหน่วยงาน
5	Date	วันที่เข้าใช้



ภาพที่ 3.18 แสดงการนำเข้าข้อมูลตารางล็อกไฟล์โดยการแม็ปข้อมูล



ภาพที่ 3.19 แสดงความสัมพันธ์ของข้อมูลตารางในการวิเคราะห์ด้วยเทคนิค Association Rules

2) *ไทม์ซีรีส์ (Time Series)* เป็นการพยากรณ์ค่าตัวเลขของจำนวนผู้เข้าใช้ในอนาคต ตามช่วงเวลาหรือเหตุการณ์ โดยแบ่งตามหมวดหมู่ของหน่วยงานที่ผู้วิจัยได้กำหนดไว้ และในอัลกอริทึมนี้ ผู้วิจัยได้กำหนดแอ็ตทริบิวต์ต่างๆ ในการนำมาใช้วิเคราะห์ ดังตารางที่ 3.3 และภาพที่ 3.20

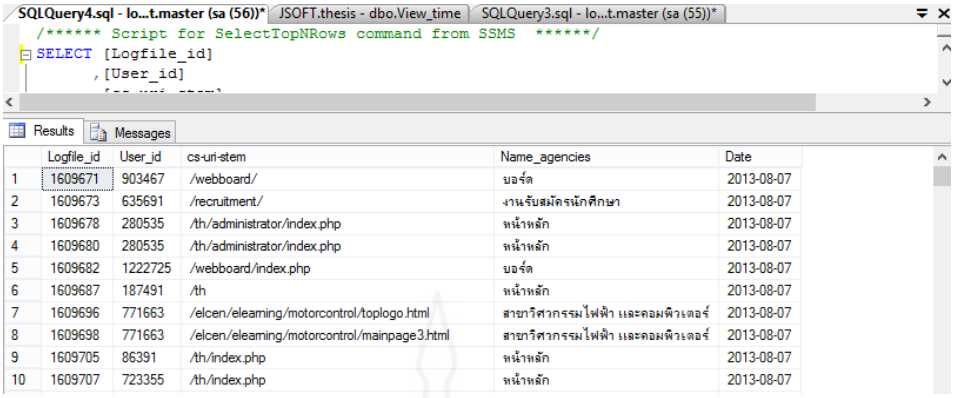
ตารางที่ 3.3 แสดงแอ็ตทริบิวต์ของตาราง ในการวิเคราะห์ด้วยเทคนิคไทม์ซีรีส์

Col.	Name	Description
1	Date	วันที่
2	Count Faculty A	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “กองการศึกษา”
3	Count Faculty B	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “กองบริหารทรัพยากร”
4	Count Faculty C	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “บริหารธุรกิจและศิลปศาสตร์”
5	Count Faculty D	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “วิทยาศาสตร์และเทคโนโลยีการเกษตร”
6	Count Faculty E	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “วิศวกรรมศาสตร์”
7	Count Faculty F	จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “หน่วยงานอื่นๆ”
8	Count	รวมจำนวนผู้เข้าใช้ทั้งหมด

	Date	Count Faculty A	Count Faculty B	Count Faculty C	Count Faculty D	Count Faculty E	Count Faculty F	Count Faculty G	Count
1	2013-01-01	236	9	394	14	7	190	4	854
2	2013-01-02	489	43	776	22	16	376	11	1733
3	2013-01-03	545	52	1010	25	33	395	14	2074
4	2013-01-04	763	43	949	32	25	408	2	2222
5	2013-01-05	399	47	544	21	17	329	4	1361
6	2013-01-06	393	27	467	13	23	404	8	1335
7	2013-01-07	490	47	871	31	22	497	13	1971
8	2013-01-08	507	54	925	38	18	578	14	2134
9	2013-01-09	433	22	719	36	21	404	19	1654
10	2013-01-10	498	39	910	18	30	541	9	2045
11	2013-01-11	377	50	688	32	32	443	8	1630

ภาพที่ 3.20 ตัวอย่างข้อมูลตาราง DB_Timeseries

3) ซีเควนซ์ คลัสเตอริง (Sequence Clustering) เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูลที่มีลักษณะเรียงลำดับเหตุการณ์ เช่น ลำดับการเข้าชมหน้าเพจของผู้เข้าใช้แต่ละครั้ง ผู้วิจัยได้นำตาราง DB_Logfile มีข้อมูลจำนวน 3,534,712 เรคอร์ด และตาราง DB_User มีข้อมูลจำนวน 650,645 เรคอร์ด มาใช้วิเคราะห์ในอัลกอริทึมนี้ดังภาพที่ 3.21, 3.22



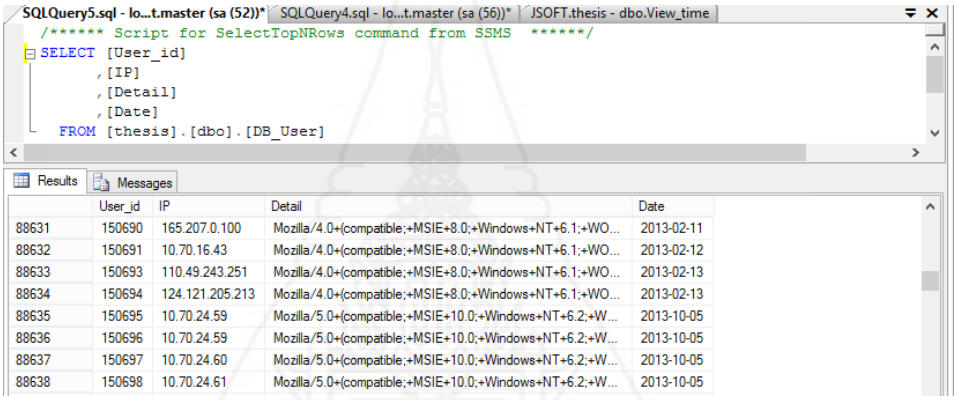
SQLQuery4.sql - lo...t.master (sa (56))* | JSOFT.thesis - dbo.View_time | SQLQuery3.sql - lo...t.master (sa (55))*

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [Logfile_id]
      ,[User_id]
```

Results | Messages

	Logfile_id	User_id	cs-uri-stem	Name_agencies	Date
1	1609671	903467	/webboard/	บอร์ด	2013-08-07
2	1609673	635691	/recruitment/	งานรับสมัครนักศึกษา	2013-08-07
3	1609678	280535	/th/administrator/index.php	หน้าหลัก	2013-08-07
4	1609680	280535	/th/administrator/index.php	หน้าหลัก	2013-08-07
5	1609682	1222725	/webboard/index.php	บอร์ด	2013-08-07
6	1609687	187491	/th	หน้าหลัก	2013-08-07
7	1609696	771663	/elcen/eleaming/motorcontrol/toplogo.html	สาขาวิศวกรรมไฟฟ้า และคอมพิวเตอร์	2013-08-07
8	1609698	771663	/elcen/eleaming/motorcontrol/mainpage3.html	สาขาวิศวกรรมไฟฟ้า และคอมพิวเตอร์	2013-08-07
9	1609705	86391	/th/index.php	หน้าหลัก	2013-08-07
10	1609707	723355	/th/index.php	หน้าหลัก	2013-08-07

ภาพที่ 3.21 ตัวอย่างข้อมูลตาราง DB_Logfile



SQLQuery5.sql - lo...t.master (sa (52))* | SQLQuery4.sql - lo...t.master (sa (56))* | JSOFT.thesis - dbo.View_time

```
/****** Script for SelectTopNRows command from SSMS  ******/
SELECT [User_id]
      ,[IP]
      ,[Detail]
      ,[Date]
  FROM [thesis].[dbo].[DB_User]
```

Results | Messages

	User_id	IP	Detail	Date
88631	150690	165.207.0.100	Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+6.1;+WO...	2013-02-11
88632	150691	10.70.16.43	Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+6.1;+WO...	2013-02-12
88633	150693	110.49.243.251	Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+6.1;+WO...	2013-02-13
88634	150694	124.121.205.213	Mozilla/4.0+(compatible;+MSIE+8.0;+Windows+NT+6.1;+WO...	2013-02-13
88635	150695	10.70.24.59	Mozilla/5.0+(compatible;+MSIE+10.0;+Windows+NT+6.2;+W...	2013-10-05
88636	150696	10.70.24.59	Mozilla/5.0+(compatible;+MSIE+10.0;+Windows+NT+6.2;+W...	2013-10-05
88637	150697	10.70.24.60	Mozilla/5.0+(compatible;+MSIE+10.0;+Windows+NT+6.2;+W...	2013-10-05
88638	150698	10.70.24.61	Mozilla/5.0+(compatible;+MSIE+10.0;+Windows+NT+6.2;+W...	2013-10-05

ภาพที่ 3.22 ตัวอย่างข้อมูลตาราง DB_User

4) *คลัสเตอริง (Clustering)* เป็นอัลกอริทึมที่ใช้ในการจำแนกหรือจัดกลุ่มการใช้จากข้อมูลการเข้าใช้งานเว็บไซต์ เช่น ข้อมูลความถี่ในการเข้าชมเว็บของหน่วยงานต่างๆ ภายในมหาวิทยาลัย โดยผู้วิจัยได้เลือกตาราง DB_Logfile ซึ่งสามารถนำข้อมูลมาวิเคราะห์อัลกอริทึมนี้ได้

3.3.4 สร้างแบบจำลองเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคData Mining ซึ่งจากการศึกษาและทำความเข้าใจกับข้อมูลแล้ว ผู้วิจัยได้เลือกอัลกอริทึมที่จะนำมาประมวลผล 4 อัลกอริทึม คือ1) แอสโซซิเอชันรูลส์ (Association Rules) 2) ไทม์ซีรีส์ (Time Series) 3) ซีเควันซ์คลัสเตอริง (Sequence Clustering) และ 4) คลัสเตอริง (Clustering) ในขั้นตอนการประมวลผลโดยการใช้เทคนิคดาต้าไมนิ่ง ผู้วิจัยได้ใช้ซอฟต์แวร์ SQL Server Business Intelligence Development Studio ดังรายละเอียดต่อไปนี้

1) *สร้างโปรเจ็กต์*และกำหนดค่าดังนี้

(1) เลือกเท็มเพลต (Templates) เป็น Analysis Service Project

(2) กำหนดชื่อโปรเจ็กต์

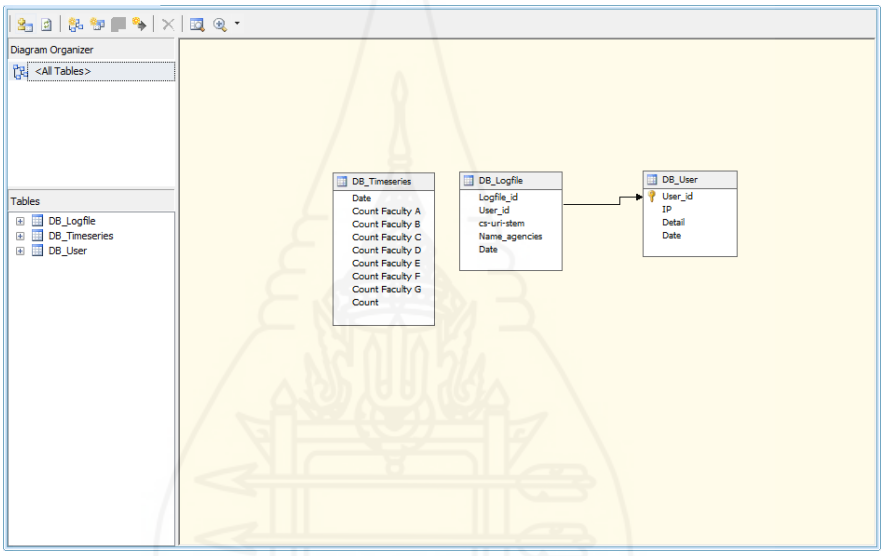
*(3) กำหนดโฟลเดอร์*ที่ใช้เก็บข้อมูลโปรเจ็กต์

*(4) กำหนดไดเรกทอรี*สำหรับเก็บโซลูชัน (solution) ของโปรเจ็กต์

2) กำหนดที่ตั้งของแหล่งข้อมูล โดยคลิกขวาที่ชื่อ Data Sources เลือก New Data Sources แล้วกำหนดค่าแหล่งข้อมูลที่นำมาใช้ในงานวิจัย โดยผู้วิจัยตั้งชื่อว่า DWLOGFILE.ds

3) ทำการกำหนด Data Sources View คือข้อมูลตารางหรือวิวของตารางที่ต้องการนำเข้าเพื่อนำมาวิเคราะห์โดยคลิกขวาที่ Data Sources View เลือก New Data Sources View แล้วทำการเลือก Data Sources ที่ต้องการ จากนั้นเลือกตารางหรือวิวของตารางที่ต้องการ ผู้วิจัยได้เลือกตาราง DB_Logfile, ตาราง DB_Timeserie, ตาราง DB_User และทำการตั้งชื่อ Data Sources Viewว่า VDWAnalysis.dsv

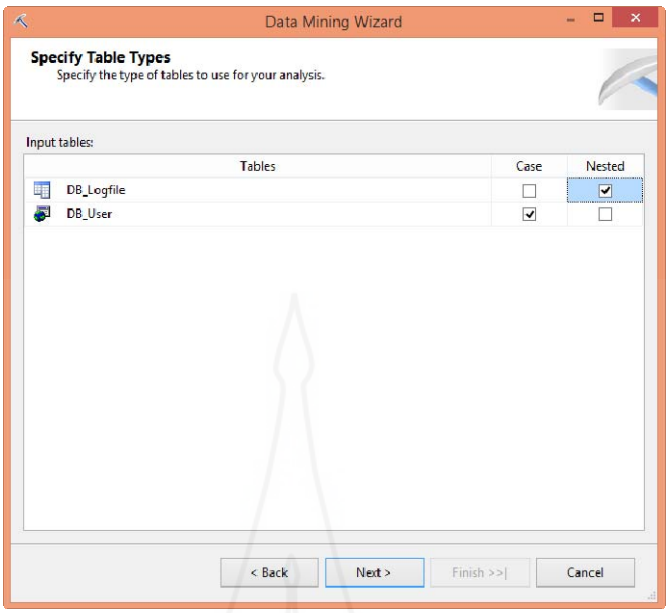
*4) จากนั้นดับเบิลคลิก*ที่ชื่อ VDWAnalysis.dsv เพื่อตรวจสอบความถูกต้องของตารางดังกล่าวดังภาพที่ 3.23



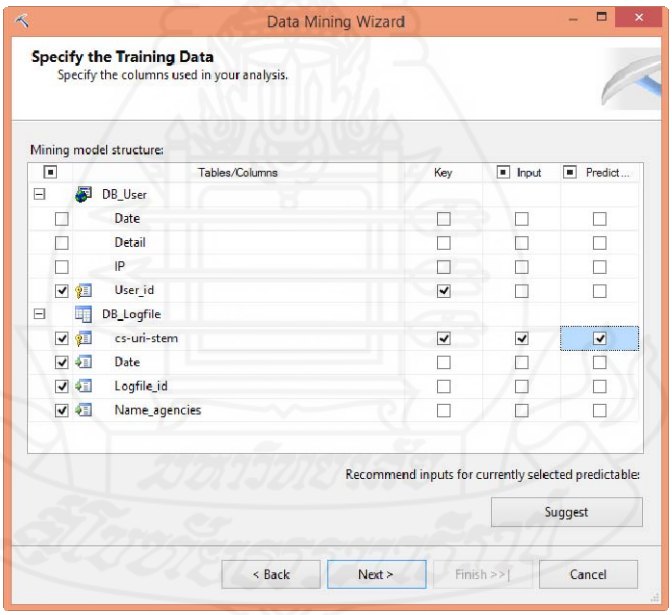
ภาพที่ 3.23 แสดง Data Sources View ตารางที่นำมาวิเคราะห์

5) ขั้นตอนการวิเคราะห์ข้อมูลโดยใช้เทคนิคการทำเหมืองข้อมูล โดยคลิกขวาที่ Mining Structures เลือก New Mining Structures ในแต่ละขั้นตอน ต้องทำการเลือกอัลกอริทึมที่ต้องการประมวลผล และเลือกตารางที่ต้องการจาก Data Sources View แล้วทำการกำหนดตารางที่เป็น Case หรือ Nested และกำหนดค่าแอตทริบิวต์ในส่วนคีย์ (Key), อินพุต (Input), พยากรณ์ (predict) ที่เกี่ยวข้อง ซึ่งผู้วิจัยได้กำหนดค่าในอัลกอริทึม โดยมีรายละเอียดต่อไปนี้

(1) แอสโซซิเอชันรูลส์ (Association Rules) เพื่อหาความสัมพันธ์ของหน้าเว็บที่ผู้ใช้เข้าชม โดยผู้วิจัยได้เลือกใช้ตาราง DB_Logfile และตาราง DB_User ดังภาพที่ 3.22 และกำหนดให้ตาราง DB_User เป็น Case ตาราง DB_User เป็น Nested และกำหนดค่าแอ็ตทริบิวต์ User_id เป็นคีย์ และกำหนดค่าแอ็ตทริบิวต์ Cs-uri-stem เป็นคีย์ อินพุต และพยากรณ์ ดังภาพที่ 3.24, 3.25



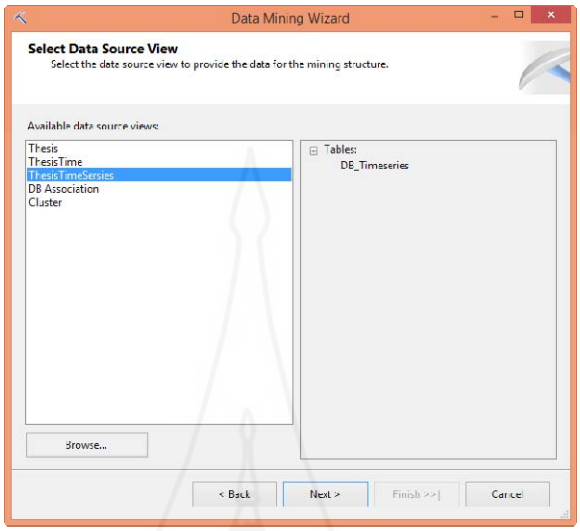
ภาพที่ 3.24 แสดงการนำเข้าตารางในอัลกอริทึม Association Rules



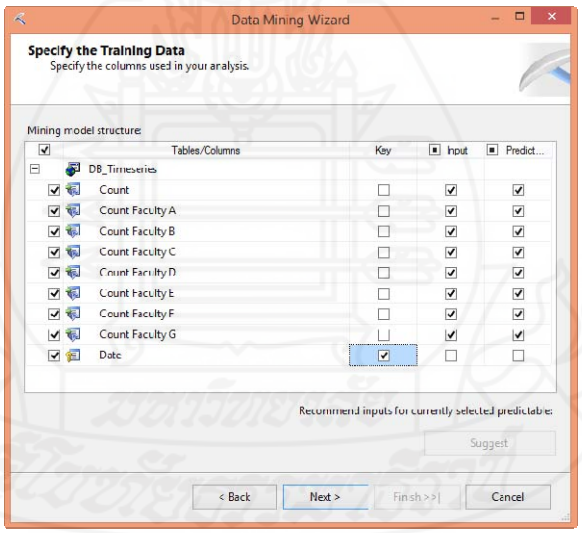
ภาพที่ 3.25 แสดงการกำหนดแอ็ตทริบิวต์ในอัลกอริทึม Association Rules

*(2) ไทม์ซีรีส์(Time Series)*เป็นการพยากรณ์จำนวนของผู้เข้าใช้ในอนาคตตามช่วงเวลาหรือเหตุการณ์ โดยผู้วิจัยได้เลือกตาราง DB_Timeseries นำมาประมวลผลในอัลกอริทึมไทม์ซีรีส์ และได้กำหนดแอตทริบิวต์ Date เป็นคีย์ซึ่งเป็นข้อมูลของวันที่ ระหว่างวันที่ 1 ม.ค. 2556 ถึง วันที่ 31 ธ.ค. 2556 และกำหนดแอตทริบิวต์ Count, Count Faculty A, Count Faculty B, Count Faculty C, Count Faculty D, Count Faculty E, Count Faculty F, Count

Faculty G ซึ่งเป็นข้อมูลจำนวนผู้เข้าชมเว็บในแต่ละหน่วยงานหลักกำหนดเป็นอินพุต และพยากรณ์ดังภาพที่ 3.26, 3.27

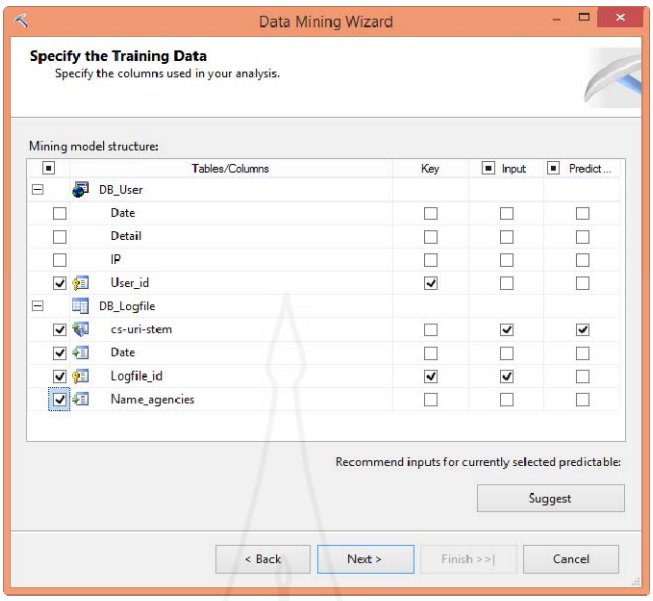


ภาพที่ 3.26 การเลือกตารางในการนำมาวิเคราะห์ด้วยอัลกอริทึมไทม์ซีรีส์

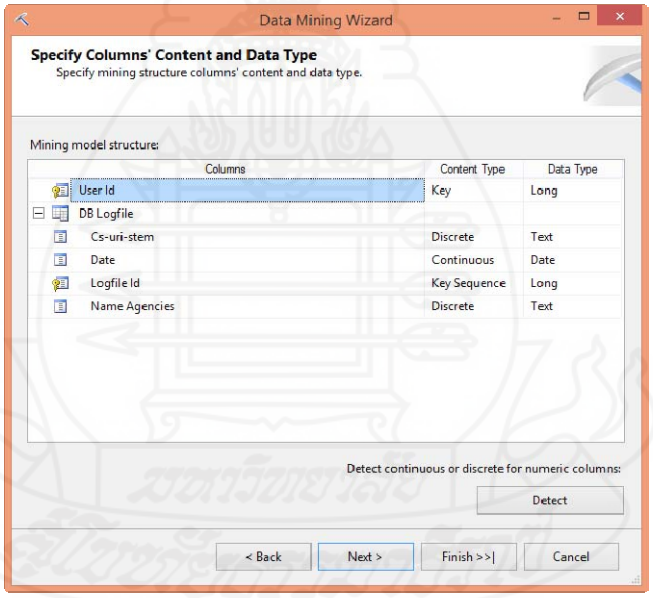


ภาพที่ 3.27 แสดงการกำหนดแอ็ตทริบิวต์ในอัลกอริทึมไทม์ซีรีส์

(3) ซีเควนซ์ คลัสเตอริง (Sequence Clustering) เป็นการจัดกลุ่มโดยการเรียงลำดับการเข้าใช้งานเว็บไซต์ซึ่งผู้วิจัยได้เลือกตาราง DB_Logfile และตาราง DB_User ในการนำมาวิเคราะห์ด้วยอัลกอลิธึมนี้ โดยกำหนดแอ็ตทริบิวต์ Date เป็นคีย์ กำหนดแอ็ตทริบิวต์ Logfile_id เป็นคีย์ซีเควนซ์ (Key Sequence) และกำหนดแอ็ตทริบิวต์ Cs-uri-stem เป็นอินพุต และพยากรณ์ดังภาพที่ 3.28, 3.29

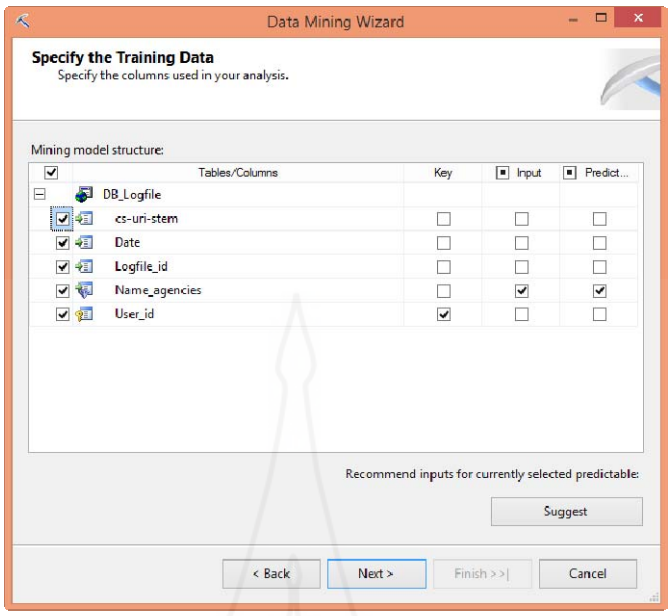


ภาพที่ 3.28 แสดงการกำหนดแอ็ตทริบิวต์ในอัลกอริทึมซีเคว็นซ์ คลัสเตอริง



ภาพที่ 3.29 แสดงรายละเอียดข้อมูลแอ็ตทริบิวต์ในอัลกอริทึมซีเคว็นซ์ คลัสเตอริง

(4) คลัสเตอริง (Clustering) เป็นการจัดกลุ่มของข้อมูลความถี่ในการเข้าชมเว็บของหน่วยงานต่างๆ ภายในมหาวิทยาลัย โดยผู้วิจัยได้เลือกตาราง DB_Logfile แล้วกำหนดแอ็ตทริบิวต์ User_id เป็นคีย์ และกำหนดแอ็ตทริบิวต์ Name Agencies เป็นอินพุต และพยากรณ์ ดังภาพที่ 3.30



ภาพที่ 3.30 แสดงการกำหนดแอ็ตทริบิวต์ในอัลกอริทึมคลัสเตอริง

3.3.5 ประเมินผลลัพธ์ หลังจากที่ได้ผลลัพธ์จากการประมวลผลข้อมูลตามอัลกอริทึมแต่ละอัลกอริทึมแล้ว ขั้นตอนต่อมาคือนำผลลัพธ์มาประเมินว่า ผลลัพธ์ที่ได้ทำให้ทราบความรู้ใหม่ๆ ในเรื่องใด ผลลัพธ์มีความสมเหตุสมผลไหม ซึ่งรายละเอียดจะกล่าวต่อไปในบทที่ 4 ผลการวิจัย

3.3.6 การนำผลลัพธ์ไปใช้ประโยชน์ กรณีที่ผลการประเมินผลลัพธ์พบว่า ผลลัพธ์ที่ได้ทำให้ทราบความรู้ใหม่ที่ไม่เคยทราบมาก่อน ขั้นตอนต่อไปคือนำผลลัพธ์นั้นไปประยุกต์หรือใช้เพื่อให้เกิดประโยชน์ต่อการพัฒนาเว็บไซต์ของมหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ซึ่งรายละเอียดจะกล่าวต่อไปในบทที่ 4 และบทที่ 5

บทที่ 4
ผลการวิจัย

ผลการศึกษาวิจัยเกี่ยวกับการทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์ กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปางประกอบด้วย 3 ส่วน ได้แก่

1. คลังข้อมูลล็อกไฟล์ของการเข้าใช้งานเว็บไซต์
2. รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป
3. ผลลัพธ์จากการทำเหมืองข้อมูล

1. คลังข้อมูลล็อกไฟล์ของการเข้าใช้งานเว็บไซต์

ผู้วิจัยได้พัฒนาคลังข้อมูลล็อกไฟล์ของการเข้าใช้งานเว็บไซต์ มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ซึ่งนำข้อมูลจากล็อกไฟล์ และฐานข้อมูลที่เกี่ยวข้องมาผ่านกระบวนการอีทีแอล ซึ่งประกอบด้วยขั้นตอนการเตรียมข้อมูล การกำหนดรูปแบบข้อมูลเพื่อให้สอดคล้องกัน การตรวจสอบความถูกต้องสมบูรณ์ของข้อมูล และรวมถึงการโหลดข้อมูลเข้าสู่คลังข้อมูล ผู้วิจัยได้ออกแบบโครงสร้างคลังข้อมูลแบบสโนว์เฟลคมีทั้งหมด 8 ตาราง ได้แก่ ตารางข้อเท็จจริง Temp_FT_ALL 1 ตาราง และมีตารางมิติ 7 ตาราง คือ 1) ตาราง Temp_TB_Agencies 2) ตาราง Temp_TB_Codestatus 3) ตาราง Temp_TB_Faculty 4) ตาราง Temp_TB_Logfile 5) ตาราง Temp_TB_Method6) ตาราง Temp_TB_Period7) ตาราง Temp_TB_User

ตารางที่ 4.1 ตาราง Temp_TB_Logfile เก็บข้อมูลล็อกไฟล์

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Logfile_id	int	ลำดับ	1609666	PK
s-ip	nvarchar(50)	ไอพีเครื่องเซิฟเวอร์	203.158.166.2	
s-port	nvarchar(50)	พอตร์	80	
cs-uri-stem	nvarchar(255)	URL	/th/administrator/index.php	
cs-uri-query	nvarchar(MAX)	URL Query	action=register2	
sc-win32-status	nvarchar(10)	สถานะ	64	

ตารางที่ 4.2 ตาราง Temp_TB_Agencies เก็บข้อมูลหน่วยงานย่อย

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Agencies_id	int	รหัสหน่ายงาน	1	PK
Faculty_id	int	รหัสหมวด	1	FK
Name_agencies	varchar(150)	ชื่อหน่วยงาน	สาขาบริหารธุรกิจ	
Site	varchar(50)	ชื่อเว็บหน่ายงาน	bu	

ตารางที่ 4.3 ตาราง Temp_TB_Faculty เก็บข้อมูลคณะหรือหน่วยงานหลัก

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Faculty_id	int	รหัสหมวด	1	PK
Name_faculty	varchar(50)	ชื่อหมวดคณะ	บริหารธุรกิจและศิลปศาสตร์	

ตารางที่ 4.4 ตาราง Temp_TB_Codestatus เก็บสถานะการรับส่งข้อมูล HTTP

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Codestatus_id	int	รหัสสถานะ	100	PK
Name status	varchar(150)	ชื่อสถานะ	ดำเนินการต่อไป	
Category	varchar(150)	หมวดสถานะ	ข้อมูลทั่วไป	

ตารางที่ 4.5 ตาราง Temp_TB_Method เก็บเมธอด

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Method_id	int	รหัสเมธอด	3	PK
Name_metthod	varchar(50)	ชื่อเมธอด	POST	
Description	text	รายละเอียด	A POST request is used to send data to the server, for example customer information, file upload etc using HTML forms.	

ตารางที่ 4.6 ตาราง Temp_TB_Periodเก็บข้อมูลวันเวลา

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Date	date	ลำดับวันที่	2013-01-01 00:00:00Z	PK
Day of Week	varchar(20)	วัน	2:อังคาร	
Day	varchar(10)	วันที่	01	
Month	varchar(20)	เดือน	01:มกราคม	
Quarter	varchar(20)	ไตรมาส	ไตรมาส 1	
Year	varchar(4)	ปี	2013	

ตารางที่ 4.7 ตาราง Temp_TB_User เก็บรายละเอียดผู้ใช้

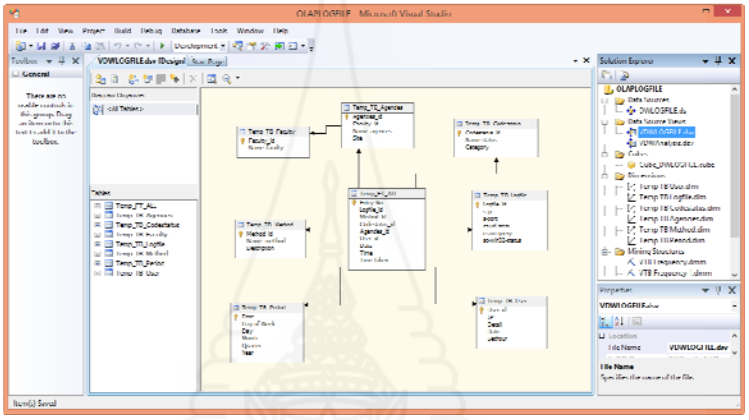
ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
User_id	int	รหัสผู้ใช้	12	PK
IP	varchar(50)	ไอพี	84.95.88.100	
Detail	nvarchar(MAX)	รายละเอียด	android	
Date	date	วันที่เข้าชม	2013-04-27 00:00:00Z	
Sethour	varchar(50)	ค่าชั่วโมงที่เข้าชม	14	

ตารางที่ 4.8 ตารางข้อเท็จจริง

ชื่อฟิลด์	ประเภทข้อมูล	คำอธิบาย	ข้อมูลตัวอย่าง	Key
Entry No	int	ลำดับ	1	PK
Logfile_id	int	รหัสล็อกไฟล์	1	FK
Method_id	int	รหัสเมธอด	1	FK
Codestatus_id	int	รหัสสถานะ	200	FK
Agencies_id	int	รหัสหน่วยงาน	59	FK
User_id	int	รหัสผู้ใช้	1244258	FK
Date	date	วันที่เข้าชม	2013-06-23 00:00:00Z	FK
Time	time	เวลาเข้าชม ณ ขณะนั้น	00:47:28.0000000	
Time taken	int	เวลาที่ใช้ในการเข้าชม (Milliseconds)	374	

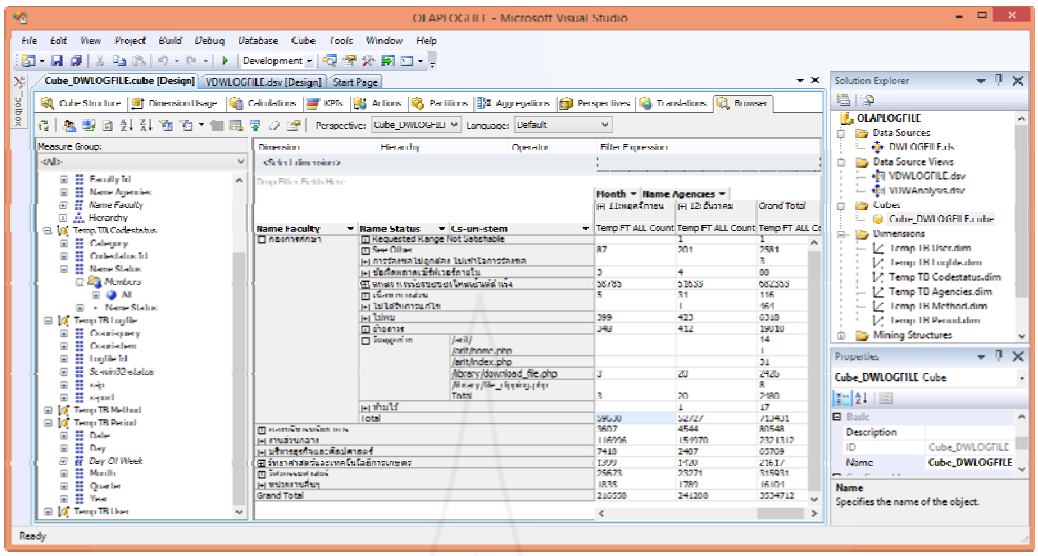
2. รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป

ในการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป นั้น ต้องสร้างคิวบ์ขึ้นมาก่อนเพื่อจะได้ออกรายงานหลายมิติได้ โดยผู้วิจัยได้สร้าง Data Source ชื่อ DWLOGFILE.ds ใน SSAS เชื่อมต่อกับคลังข้อมูล จากนั้นสร้าง Data Source View ชื่อ VDWLOGFILE.dsv ดังแสดงในภาพที่ 4.1 เพื่อเลือกตารางจากคลังข้อมูลทั้งที่เป็น Fact Tables และ Dimension Tables ที่ต้องการออกรายงานตามมิตินั้นๆ



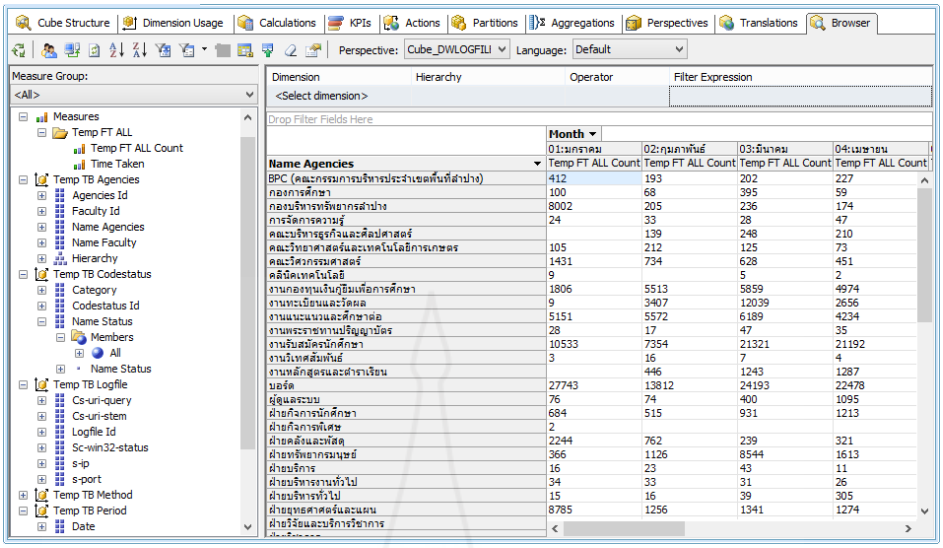
ภาพที่ 4.1 แสดงผลที่ได้จากการสร้าง Data Source View ชื่อ DWLIBRARY.dsv

จากนั้นผู้วิจัยได้สร้าง Cube ชื่อ DWLOGFILE.cube สำหรับใช้ประมวลผลเพื่อแสดงข้อมูลในลักษณะหลายมิติ หรือที่เรียกว่า การประมวลผลเชิงวิเคราะห์แบบออนไลน์โดยเลือก Tab Browser ดังแสดงในภาพที่ 4.2 โดยผู้ใช้สามารถเปลี่ยน Measures และ Dimension ตามที่ต้องการวิเคราะห์โดยลากชื่อ Measures หรือ Dimension ที่ต้องการทางด้านซ้ายมือไปวางทางด้านขวามือเพื่อดูข้อมูลต่างๆ เกี่ยวกับการเข้าใช้เว็บไซต์ของทางมหาวิทยาลัย



ภาพที่ 4.2 แสดงผลตัวอย่างการสร้างคิวบ์

จากภาพที่ 4.2 แสดงข้อมูลของการเข้าใช้เว็บไซต์ของมหาวิทยาลัย จะเห็นว่าข้อมูลในมิติด้านบนสุดหรือคอลัมน์เป็นข้อมูลของเดือนและหน่วยงานย่อย และด้านซ้ายมือเป็นข้อมูลหน่วยงานหลัก ชื่อสถานะการรับส่งและที่อยู่หน้าเว็บ (Path file) ซึ่งข้อมูลแสดงให้เห็นจำนวนของการเข้าชมหน้าเว็บในแต่ละเดือนตามข้อมูลมิติที่ผู้ใช้ต้องการนำเสนอ เช่น หน้าเว็บของหน่วยงานกองการศึกษา มีจำนวนผู้เข้าชมทั้งหมดในเดือนพฤศจิกายน จำนวน 59,630 หน้า โดยพบว่าสถานะตกลงการร้องขอของไคลเอ็นต์สำเร็จ 58,785 หน้า สถานะไม่พบ 399 หน้า สถานะเนื้อหาบางส่วน 5 หน้า สถานะอื่นๆ (See Other) 87 หน้าสถานะย้ายถาวร 348 หน้าสถานะข้อผิดพลาดเซิร์ฟเวอร์ภายใน 3 หน้า สถานะวัตถุถูกย้าย 3 หน้า และจะแสดงรายการหน้าเว็บซึ่งการแสดงผลเชิงวิเคราะห์หลายมิติสามารถกำหนดข้อมูลที่ต้องการ อาทิ ต้องการเรียดดูข้อมูลจำนวนการเข้าชมหน้าเว็บของหน่วยงานย่อย ในแต่ละเดือน ดังแสดงในภาพที่ 4.3

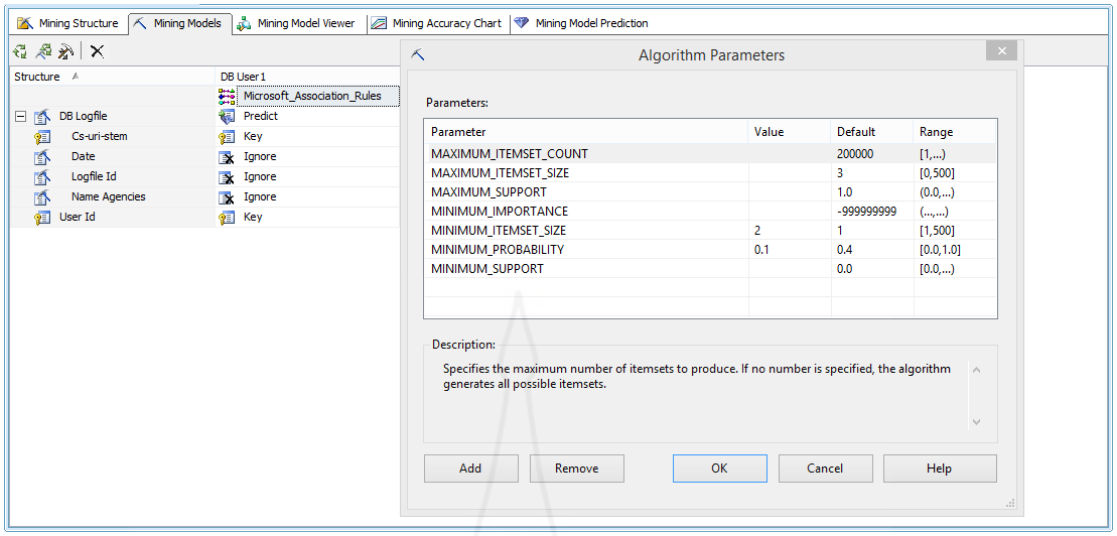


ภาพที่ 4.3 แสดงผลการเรียดดูข้อมูลจำนวนการเข้าชมหน้าเว็บของหน่วยงานย่อยในแต่ละเดือน

3. ผลลัพธ์จากการทำเหมืองข้อมูล

การทำเหมืองข้อมูลในงานวิจัยนี้ดำเนินการตามแบบจำลองคริสพ์-ดีเอ็ม (CRISP-DM Model) ส่วนอัลกอริทึมที่ผู้วิจัยใช้ในการทำเหมืองข้อมูลมีทั้งหมด 4 อัลกอริทึมประกอบด้วย 1) แอสโซซิเอชันรูลส์ 2) ไทม์ซีรีส์ 3) ซีเควินซ์ คลัสเตอริง และ 4) คลัสเตอริง

3.1 ผลการค้นหากฎความสัมพันธ์ หรือแอสโซซิเอชันรูลส์ เป็นอัลกอริทึมเพื่อวิเคราะห์ความสัมพันธ์ของข้อมูลหน้าเว็บว่า มีหน้าเว็บไหนที่ถูกเข้าถึงข้อมูลด้วยกัน ซึ่งผู้วิจัยได้ดึงข้อมูลจากตาราง DB_Logfile ประกอบด้วยข้อมูล ลำดับล็อกไฟล์, รหัสชื่อผู้ใช้, หน้าเว็บที่เข้าใช้ หน่วยงาน, วันที่เข้าใช้ จำนวน 3,534,712 เรคอร์ด และตาราง DB_User ประกอบด้วย รหัสชื่อผู้ใช้, ไอพีแอดเดรส, รายละเอียดเบราเซอร์, วันที่เข้าใช้ จำนวน 650,645 เรคอร์ดเพื่อหาความสัมพันธ์ของหน้าเว็บที่ผู้ใช้เข้าใช้เว็บไซต์ในแต่ละหน้า จากตาราง DB_Logfile ซึ่งมีข้อมูลของการเข้าใช้หน้าเว็บ 3,534,712 รายการและตาราง DB_User ซึ่งมีข้อมูลผู้ใช้จำนวน 650,645 รายการ โดยกำหนดค่าพารามิเตอร์ดังที่แสดงในภาพที่ 4.4



ภาพที่ 4.4 แสดงการกำหนดค่าพารามิเตอร์ของเทคนิคแอสโซซิเอชันรูลส์

MAXIMUM_ITEMSET_COUNT เป็นการกำหนดค่าสูงสุดของชุด item ที่จะให้อัลกอริทึมนับถ้าไม่กำหนดค่านี้อัลกอริทึมจะสร้างชุดของ item ทั้งหมดโดยพิจารณาจากค่า MINIMUM_SUPPORT การกำหนดค่านี้จะช่วยป้องกันไม่ให้อัลกอริทึมสร้างจำนวนของชุด item ที่มีจำนวนมากเกินไปในกรณีที่มีชุดของ item มากเกินไปอัลกอริทึมจะเก็บเฉพาะชุดของitemเพียง nชุดเท่านั้นโดยดูจากคะแนนของค่า IMPORTANCE เป็นพื้นฐาน

MAXIMUM_ITEMSET_SIZE เป็นการกำหนดจำนวนสูงสุดของ item ที่จะให้อัลกอริทึมสร้างขึ้นมาในชุดของItemในที่นี้กำหนดขนาดเท่ากับ 5 หมายความว่าอัลกอริทึมจะสร้างชุดของitemขึ้นมาตั้งแต่ 1 item 2 item 3 item4 item และสูงสุดไม่เกิน 5 item ในการกำหนดถ้าปรับเปลี่ยนค่านี้ลดลงจะช่วยลดเวลาในการประมวลผลลงค่าเริ่มต้นคือ 0 ซึ่งหมายความว่าไม่มีการจำกัดขนาดของชุด item

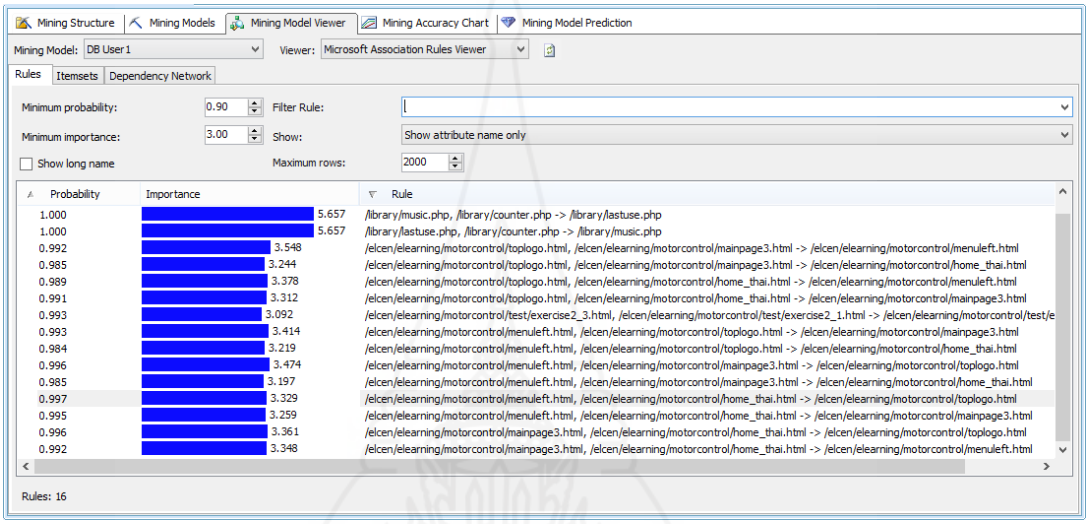
MAXIMUM_SUPPORT โดยปกติชุดของ item จะปรากฏในผลลัพธ์ได้ก็ต่อเมื่อมีค่าความถี่ไม่สูงกว่าค่าที่กำหนดใน MAXIMUM_SUPPORT ค่าที่กำหนดจะอยู่ระหว่าง 0 ถึง 1 ค่าเริ่มต้นคือ 1.00 หรือ 100% ของจำนวนเรคอร์ดทั้งหมดที่นำมาประมวลผล

MINIMUM_IMPORTANCE เป็นค่าพารามิเตอร์ที่ใช้ในการสร้างกฎความเกี่ยวข้องกฎที่มีค่า IMPORTANCE น้อยกว่าค่า MINIMUM_IMPORTANCE จะถูกกรองออกไป MINIMUM_ITEMSET_SIZE เป็นการกำหนดจำนวนต่ำสุดของ item ที่จะให้อัลกอริทึมสร้างขึ้นมาในชุดของ itemเช่นถ้ากำหนดไว้เท่ากับ 4 หมายความว่าถ้าชุดของ item ที่มี item น้อยกว่า 4 item ชุดของ item นั้นจะไม่ปรากฏในผลลัพธ์ค่าเริ่มต้นคือ 0 ในการกำหนดถ้าปรับเปลี่ยนค่านี้ลดลงจะไม่มีผลต่อการช่วยลดเวลาในการประมวลผลลงเพราะอัลกอริทึมต้องเริ่มด้วยค่าขนาดของชุด item เท่ากับ 1 และเพิ่มขนาดขึ้นทีละขั้นๆ

MINIMUM_PROBABILITY เป็นค่าความน่าจะเป็นต่ำสุดที่ใช้ในการสร้างกฎความเกี่ยวข้องค่าจะอยู่ระหว่าง 0 ถึง 1 และค่าเริ่มต้นคือ 0.4 ซึ่งหมายความว่าถ้าจำนวนเรคอร์ดของข้อมูลที่นำมาวิเคราะห์พบความเกี่ยวข้องกันระหว่าง item มีค่าน้อยกว่า 40% ของเรคอร์ดทั้งหมดที่

นำมาประมวลผลความเกี่ยวข้องของชุด item ดังกล่าวจะไม่นำมาสร้างเป็นกฎสำหรับในที่นี้กำหนดค่าไว้เท่ากับ 0.1 หมายความว่าถ้าจำนวนเรคอร์ดของการเรียกใช้หน้าเว็บที่มีความเกี่ยวข้องกันระหว่าง item มีค่าน้อยกว่า 10% ความเกี่ยวข้องกันระหว่าง item นั้นจะไม่ถูกนำมาสร้างเป็นกฎ

MINIMUM_SUPPORT โดยปกติชุดของitemจะปรากฏในผลลัพธ์ได้ก็ต่อเมื่อมีค่าความถี่ของเคสไม่น้อยกว่าค่าที่กำหนดใน MINIMUM _SUPPORT ค่าที่กำหนดจะอยู่ระหว่าง 0 ถึง 1 ค่าเริ่มต้นคือ 0.0 ของจำนวนเคสทั้งหมดที่นำมาประมวลผล



ภาพที่ 4.5 แสดงผลข้อมูลการหาความสัมพันธ์

จากภาพที่ 4.5 แสดงผลข้อมูลการหาความสัมพันธ์โดยกำหนดค่าความน่าจะเป็น (Probability) ไว้ที่ 0.90 หรือมีค่าความเชื่อมั่น (Confidence) ไม่น้อยกว่าร้อยละ 90 และกำหนดค่าความสำคัญ (Importance) ไว้ที่ 3.00 หรือมีค่าความน่าสนใจ (Lift) ไม่น้อยกว่า 3.00 พบว่ามีกฎความสัมพันธ์ที่เกิดขึ้นทั้งหมด 16 กฎดังนี้

กฎข้อที่ 1 ถ้าผู้ใช้เข้าหน้าเว็บ /library/music.php และ /library/counter.php แล้วต่อไปจะเข้าหน้าเว็บ /library/lastuse.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 5.65

กฎข้อที่ 2 ถ้าผู้ใช้เข้าหน้าเว็บ /library/lastuse.php และ /library/counter.php แล้วต่อไปจะเข้าหน้าเว็บ /library/music.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 5.65

กฎข้อที่ 3 ถ้าผู้ใช้เข้าหน้าเว็บ /library/music.php และ /library/lastuse.php แล้วต่อไปจะเข้าหน้าเว็บ /library/counter.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 3.36

กฎข้อที่ 4 ถ้าผู้ใช้เข้าหน้าเว็บ/elcen/elearning/motorcontrol/menuleft.html และ/elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/elcen/elearning/motorcontrol/toplogo.html ด้วยค่าความเชื่อมั่น 0.997 และค่าความน่าสนใจ 3.32

กฎข้อที่ 5 ถ้าผู้ใช้เข้าหน้าเว็บ/elcen/elearning/motorcontrol/mainpage3.html และ /elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/elcen/elearning/motorcontrol/toplogo.html ด้วยค่าความเชื่อมั่น 0.996 และค่าความน่าสนใจ 3.36

กฎข้อที่ 6 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html และ /elcen/elearning/motorcontrol/mainpage3.html แล้วต่อไปจะเข้าหน้าเว็บ/ elcen/elearning/motorcontrol/toplogo.html ด้วยค่าความเชื่อมั่น 0.996 และค่าความน่าสนใจ3.47

กฎข้อที่ 7 ถ้าผู้ใช้เข้าหน้าเว็บ/elcen/elearning/motorcontrol/menuleft.html และ /elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/elcen/elearning/motorcontrol/mainpage3.html ด้วยค่าความเชื่อมั่น 0.995 และค่าความน่าสนใจ 3.25

กฎข้อที่ 8 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html และ /elcen/elearning/motorcontrol/toplogo.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html ด้วยค่าความเชื่อมั่น 0.993 และค่าความน่าสนใจ 3.41

กฎข้อที่ 9 ถ้าผู้ใช้เข้าหน้าเว็บ/elcen/elearning/motorcontrol/test/exercise2_3.html และ /elcen/elearning/motorcontrol/test/exercise2_1.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/test/exercise2_2.html ด้วยค่าความเชื่อมั่น 0.993 และค่าความน่าสนใจ 3.09

กฎข้อที่ 10 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html และ /elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/ elcen/elearning/motorcontrol/menuleft.html ด้วยค่าความเชื่อมั่น 0.992 และค่าความน่าสนใจ3.34

กฎข้อที่ 11 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/mainpage3.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html ด้วยค่าความเชื่อมั่น 0.992 และค่าความน่าสนใจ 3.54

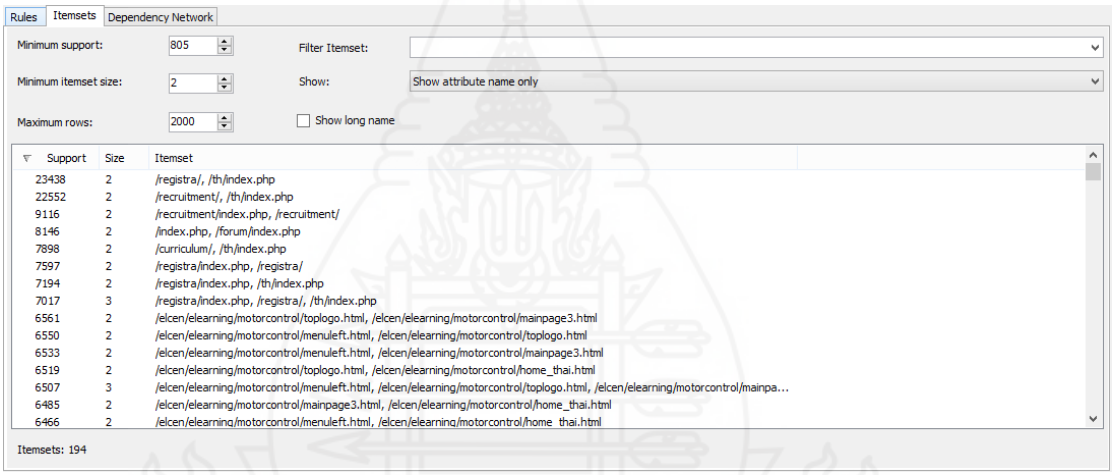
กฎข้อที่ 12 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/elcen/elearning/motorcontrol/mainpage3.html ด้วยค่าความเชื่อมั่น 0.991 และค่าความน่าสนใจ 3.31

กฎข้อที่ 13 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/home_thai.html แล้วต่อไปจะเข้าหน้าเว็บ/ elcen/elearning/motorcontrol/menuleft.html ด้วยค่าความเชื่อมั่น 0.989 และค่าความน่าสนใจ 3.37

กฎข้อที่ 14 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/mainpage3.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/home_thai.html ด้วยค่าความเชื่อมั่น 0.985 และค่าความน่าสนใจ 3.24

กฎข้อที่ 15 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html และ /elcen/elearning/motorcontrol/mainpage3.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/home_thai.html ด้วยค่าความเชื่อมั่น 0.985 และค่าความน่าสนใจ 3.19

กฎข้อที่ 16 ถ้าผู้ใช้เข้าหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html และ /elcen/elearning/motorcontrol/toplogo.html แล้วต่อไปจะเข้าหน้าเว็บ /elcen/elearning/motorcontrol/home_thai.htmlด้วยค่าความเชื่อมั่น 0.984 และค่าความน่าสนใจ 3.21



ภาพที่ 4.6 แสดงผลหน้าเว็บที่มีการคลิกไปพร้อมกัน

จากภาพที่ 4.6 แสดงผลหน้าเว็บที่มีการคลิกไปพร้อมกันโดยสามารถวิเคราะห์ผลลัพธ์จากกฎที่กำหนดความสัมพันธ์การเข้าใช้หน้าเว็บพร้อมกันจำนวน 2 item โดยเลือกข้อมูล item ที่มีค่าความสัมพันธ์ (Support) อันดับสูงสุด 10 อันดับแรกมีรายละเอียดดังนี้

ลำดับที่ 1 ถ้าเข้าดูหน้าเว็บ /registra/ แล้วจะเข้าดูหน้าเว็บ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 23,438 เรคอร์ด

ลำดับที่ 2 ถ้าเข้าดูหน้าเว็บ /recruitment/ แล้วจะเข้าดูหน้าเว็บ /th/index.php ต่อมีค่าความสัมพันธ์เท่ากับ 22,552 เรคอร์ด

ลำดับที่ 3 ถ้าเข้าดูหน้าเว็บ /recruitment/index.php แล้วจะเข้าดูหน้าเว็บ /recruitment/ ต่อมีค่าความสัมพันธ์เท่ากับ 9,116 เรคอร์ด

ลำดับที่ 4 ถ้าเข้าดูหน้าเว็บ /index.php แล้วจะเข้าดูหน้าเว็บ /forum/index.php ต่อมีค่าความสัมพันธ์เท่ากับ 8,146 เรคอร์ด

ลำดับที่ 5 ถ้าเข้าดูหน้าเว็บ /curriculum/ แล้วจะเข้าดูหน้าเว็บ /th/index.php ต่อมีค่าความสัมพันธ์เท่ากับ 7,898 เรคอร์ด

ลำดับที่ 6 ถ้าเข้าดูหน้าเว็บ /registra/index.php แล้วจะเข้าดูหน้าเว็บ /registra/ ต่อมีค่าความสัมพันธ์เท่ากับ 7,597 เรคอร์ด

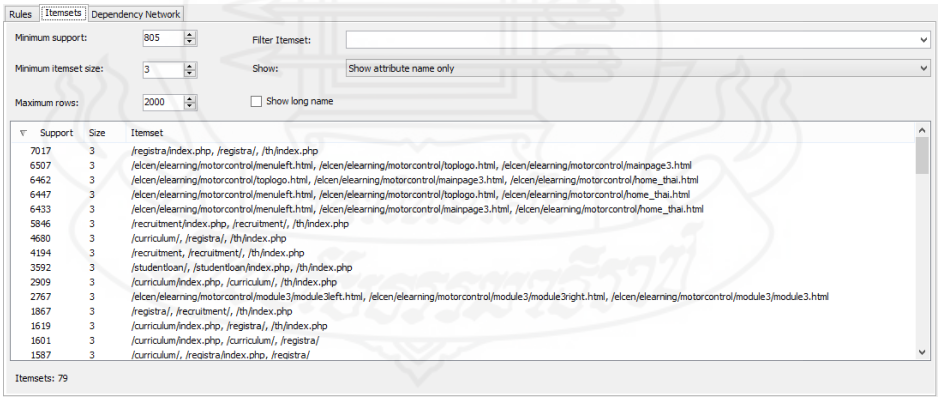
ลำดับที่ 7 ถ้าเข้าดูหน้าเว็บ /registra/index.php แล้วจะเข้าดูหน้าเว็บ /th/index.php ต่อมีค่าความสัมพันธ์เท่ากับ 7,194 เรคอร์ด

ลำดับที่ 8 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html ต่อมีค่าความสัมพันธ์เท่ากับ 6,561 เรคอร์ด

ลำดับที่ 9 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html ต่อมีค่าความสัมพันธ์เท่ากับ 6,550 เรคอร์ด

ลำดับที่ 10 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html ต่อมีค่าความสัมพันธ์เท่ากับ 6,533 เรคอร์ด

จากผลลัพธ์ที่ได้ทำให้ทราบว่าหน้าเว็บที่ผู้ใช้นิยมดูต่อจากกันคือหน้าเว็บ /registra/ แล้วจะเข้าดูหน้าเว็บ /th/index.php ซึ่งผลลัพธ์นี้ สามารถนำไปในการกำหนดเมนูลิงค์ หรือหน้าเว็บนำทางแนะนำให้กับผู้ใช้งาน ในการออกแบบเว็บไซต์ได้



ภาพที่ 4.7แสดงผลวิเคราะห์ผลลัพธ์จากกฎความสัมพันธ์การเข้าใช้หน้าเว็บพร้อมกันจำนวน 3 item

จากภาพที่ 4.7 การหาความสัมพันธ์สามารถวิเคราะห์ผลลัพธ์จากกฎความสัมพันธ์การเข้าใช้หน้าเว็บพร้อมกันจำนวน 3iem โดยเลือกข้อมูล item ที่มีค่าความสัมพันธ์อันดับสูงสุด 10 อันดับแรกมีรายละเอียดดังนี้

ลำดับที่ 1 ถ้าเข้าดูหน้าเว็บ /registra/index.php แล้วจะเข้าดูหน้าเว็บ /registra/ และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 7,017 เรคอร์ด

ลำดับที่ 2 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/mainpage3.html ต่อมีค่าความสัมพันธ์ เท่ากับ 6,507 เรคอร์ด

ลำดับที่ 3 ถ้าเข้าดูหน้าเว็บ/elcen/elearning/motorcontrol/toplogo.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html และ /elcen/elearning/motorcontrol/home_thai.html ต่อมีค่าความสัมพันธ์ เท่ากับ 6,462 เรคอร์ด

ลำดับที่ 4 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/toplogo.html และ /elcen/elearning/motorcontrol/home_thai.html ต่อมีค่าความสัมพันธ์ เท่ากับ 6,447 เรคอร์ด

ลำดับที่ 5 ถ้าเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/menuleft.html แล้วจะเข้าดูหน้าเว็บ /elcen/elearning/motorcontrol/mainpage3.html และ /elcen/elearning/motorcontrol/home_thai.html ต่อมีค่าความสัมพันธ์ เท่ากับ 6,433 เรคอร์ด

ลำดับที่ 6 ถ้าเข้าดูหน้าเว็บ /recruitment/index.php แล้วจะเข้าดูหน้าเว็บ /recruitment/ และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 5,846 เรคอร์ด

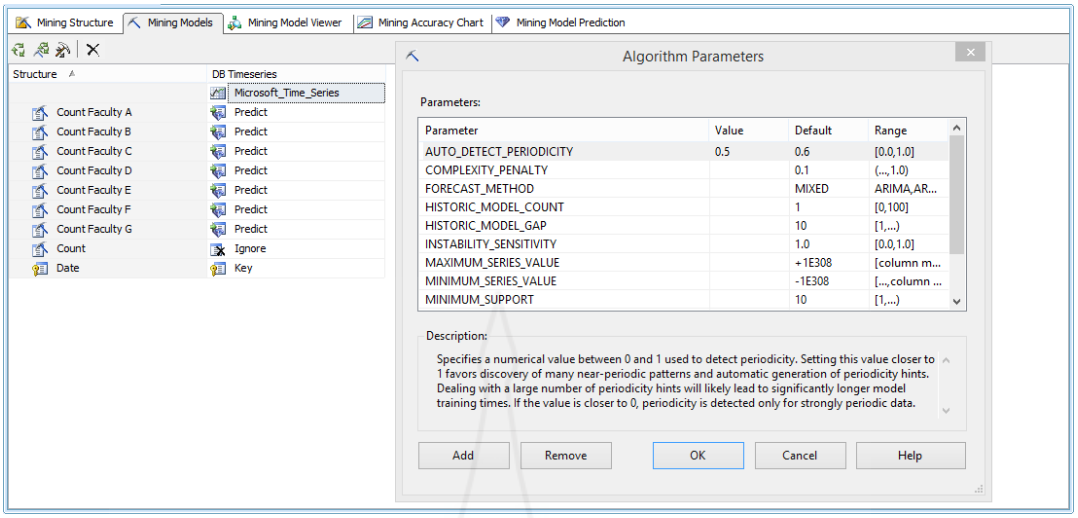
ลำดับที่ 7 ถ้าเข้าดูหน้าเว็บ /curriculum/ แล้วจะเข้าดูหน้าเว็บ /registra/ และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 4,680 เรคอร์ด

ลำดับที่ 8 ถ้าเข้าดูหน้าเว็บ /recruitment แล้วจะเข้าดูหน้าเว็บ /recruitment/ และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 4,194 เรคอร์ด

ลำดับที่ 9 ถ้าเข้าดูหน้าเว็บ /studentloan/ แล้วจะเข้าดูหน้าเว็บ /studentloan/index.php และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 3,592 เรคอร์ด

ลำดับที่ 10 ถ้าเข้าดูหน้าเว็บ /curriculum/index.php แล้วจะเข้าดูหน้าเว็บ /curriculum/ และ /th/index.php ต่อมีค่าความสัมพันธ์ เท่ากับ 2,909 เรคอร์ด

3.2 ผลการพยากรณ์ค่าตัวเลขการเข้าใช้ หรือไทม์ซีรีส์ ซึ่งไทม์ซีรีส์เป็นอัลกอริทึมเพื่อพยากรณ์ค่าตัวเลขของจำนวนผู้เข้าใช้ในอนาคตตามช่วงเวลาหรือเหตุการณ์ โดยแบ่งตามหมวดหมู่ของหน่วยงานที่ผู้วิจัยได้ดึงข้อมูลจากตาราง DB_Timeseries ประกอบด้วยข้อมูล วันที่, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “กองการศึกษา”, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “กองบริหารทรัพยากร”, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “บริหารธุรกิจและศิลปศาสตร์”, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “วิทยาศาสตร์และเทคโนโลยีการเกษตร”, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “วิศวกรรมศาสตร์”, จำนวนผู้เข้าใช้เว็บไซต์ในสังกัดหน่วยงาน “หน่วยงานอื่นๆ”, รวมจำนวนผู้เข้าใช้ทั้งหมดซึ่งเป็นข้อมูลระหว่างวันที่ 1 มกราคม 2556 – 31 ธันวาคม 2556 โดยกำหนดค่าพารามิเตอร์ดังที่แสดงในภาพที่ 4.8



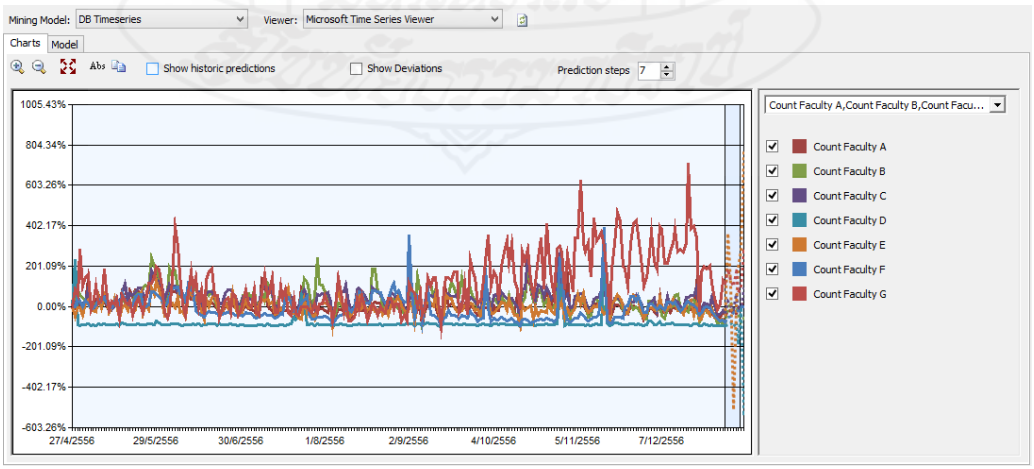
ภาพที่ 4.8 แสดงการกำหนดค่าพารามิเตอร์ของเทคนิค Time Series

AUTO_DETECT_PERIODICITY เป็นค่าที่ใช้ในตรวจหาช่วงระยะเวลา ถ้ากำหนดค่าเข้าใกล้ 1 จะค้นหาแพทเทิร์นของข้อมูลทั้งหมดที่มีช่วงระยะเวลาเข้ามาเกี่ยวข้อง แต่ถ้ากำหนดค่าเข้าใกล้ 0 การตรวจหาแพทเทิร์นระยะเวลาที่เกี่ยวข้องกับข้อมูลจะแสดงเฉพาะข้อมูลที่มีระยะเวลาเข้ามาเกี่ยวข้องมากๆ เท่านั้นในที่นี้กำหนด 0.5

COMPLEXITY_PENALTY เป็นค่าที่ใช้ในการควบคุมการขยายขนาดของ Auto regression trees ค่าที่กำหนดจะอยู่ในช่วง [0,1]

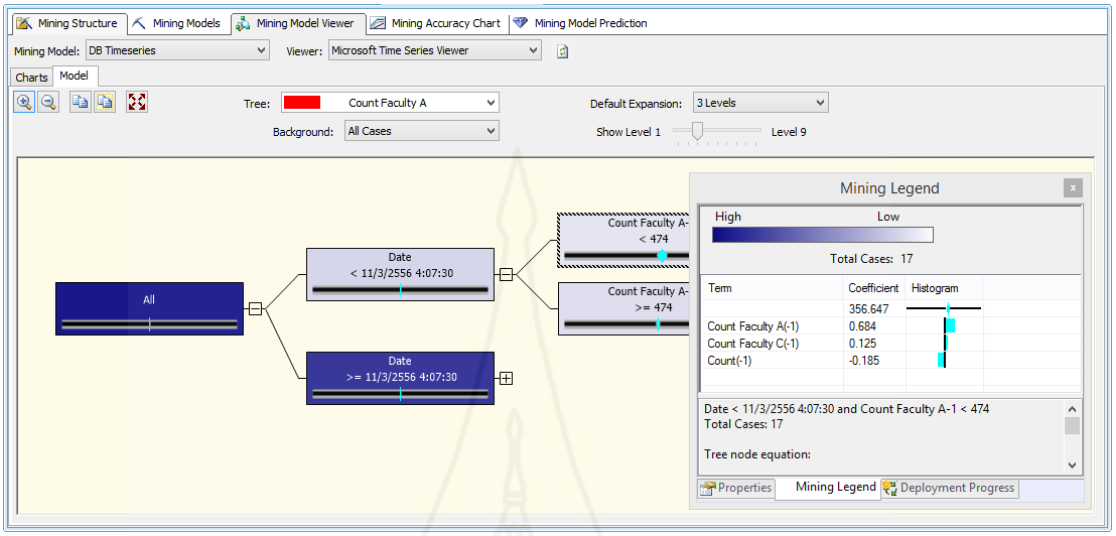
MISSING_ VALUE_ SUBSTITUTION กำหนดวิธีการที่จะใช้ในการเติมค่าที่สูญหายไป วิธีการที่มีให้เลือกได้แก่ การใช้ค่าของช่วงเวลาที่แล้ว (previous value) การใช้ค่าเฉลี่ย (meanvalue) การกำหนดค่าคงที่ (constant value)

PERIODICITY_HINT เป็นค่าที่บอกอัลกอริทึมเกี่ยวกับลักษณะของข้อมูลว่ามีเรื่องของฤดูกาลเข้ามาเกี่ยวข้องด้วย โดยรูปแบบที่กำหนดคือ {n [, n]} ซึ่งค่าnใน [ ] เป็นค่าที่ใส่หรือไม่ใส่ก็ได้ และค่า n คือ ค่าตัวเลขจำนวนเต็มหรือค่าตัวเลขทศนิยมใดๆ



ภาพที่ 4.9 แสดงผลการใช้เทคนิคในการพยากรณ์การเข้าใช้เว็บไซต์ของแต่ละหน่วยงาน

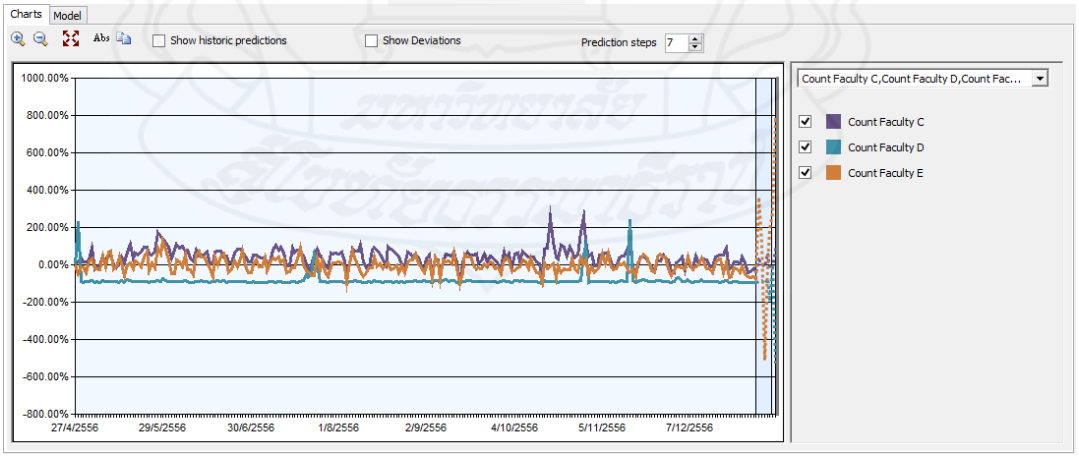
ผลที่ได้จากการใช้เทคนิคในการพยากรณ์การเข้าใช้เว็บไซต์ของแต่ละหน่วยงานแสดงผลในรูปแบบกราฟ ดังแสดงในรูปที่ 4.9 และสามารถดูในรูปแบบ Decision Trees ดังภาพที่ 4.10



ภาพที่ 4.10 รูปแบบ Decision Trees ของการเข้าใช้เว็บไซต์ของมหาวิทยาลัย

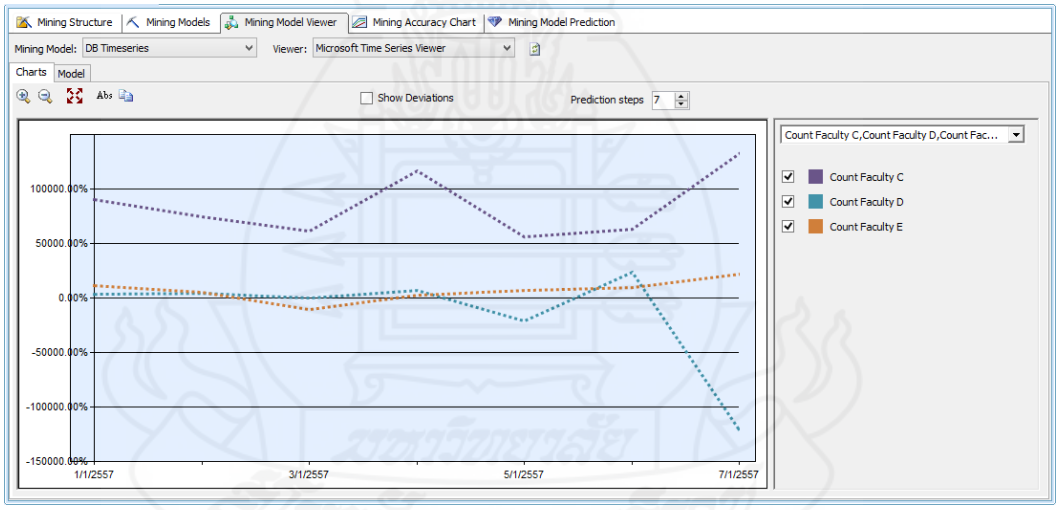
จากรูปภาพที่ 4.10 เป็นผลที่ได้จากการใช้เทคนิคการพยากรณ์ และสามารถแสดงเส้นกราฟที่แสดงผลเปรียบเทียบแนวโน้มของการเข้าใช้เว็บไซต์ของแต่ละหน่วยงาน ภายในมหาวิทยาลัย เป็นข้อมูลระหว่างวันที่ 1 มกราคม 2556 ถึง วันที่ 31 ธันวาคม 2556

นอกจากนี้ยังสามารถดูกราฟเปรียบเทียบจำนวนการเข้าใช้เว็บไซต์ของคณะต่างๆ ได้โดย การคลิกเลือกชื่อตัวแปรที่บริเวณด้านซ้ายของกราฟ ดังแสดงในภาพที่ 4.11



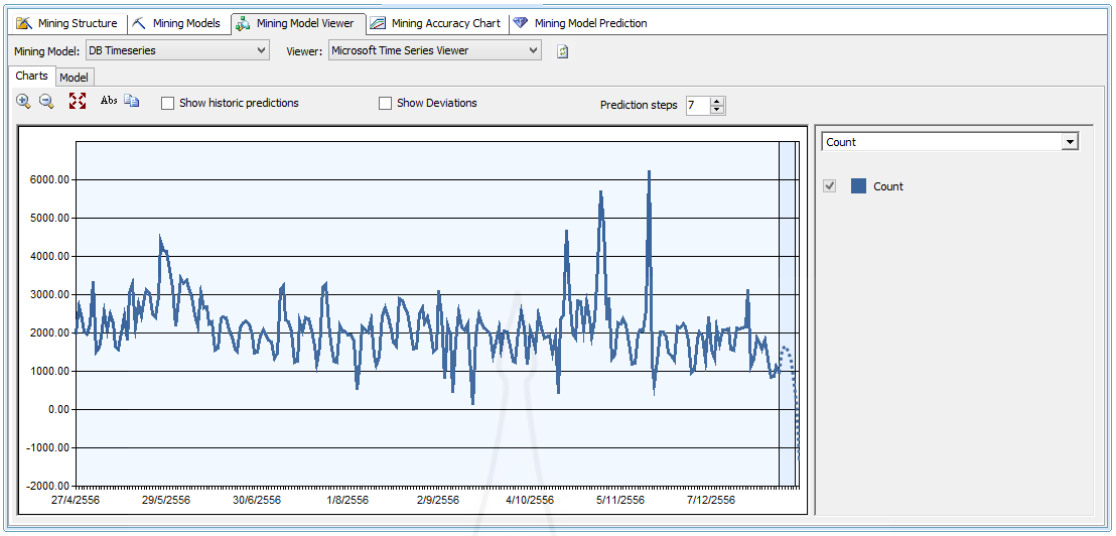
ภาพที่ 4.11 แสดงกราฟเปรียบเทียบจำนวนการเข้าใช้เว็บไซต์ของคณะทั้ง 3 คณะ

จากภาพที่ 4.11 จะเห็นได้ว่าการเข้าใช้เว็บไซต์ของมหาวิทยาลัยโดยแบ่งเป็นกลุ่มคณะ ซึ่งเป็นข้อมูลหน่วยงานย่อยที่สังกัดแต่ละคณะตลอดทั้งปี 2556 ซึ่งประกอบด้วย บริหารธุรกิจและศิลปศาสตร์ (Count Faculty C) วิทยาศาสตร์และเทคโนโลยีการเกษตร (Count Faculty D) วิศวกรรมศาสตร์ (Count Faculty E) จากกราฟ บริหารธุรกิจและศิลปศาสตร์มีจำนวนผู้เข้าใช้สูงกว่าคณะวิทยาศาสตร์และเทคโนโลยีการเกษตร และคณะวิศวกรรมศาสตร์ในช่างเวลาหรือช่วงเหตุการณ์และมีแน้วโน้มการเข้าใช้ใกล้เคียงกับคณะวิศวกรรมศาสตร์ ซึ่งมียอดการเข้าใช้รองลงมา และไม่มีการเปลี่ยนแปลงมากนักเมื่อเทียบกับจำนวนการเข้าใช้เว็บไซต์สังกัดคณะบริหารธุรกิจและศิลปศาสตร์ และจะเห็นได้ว่าการเข้าใช้เว็บไซต์สังกัดคณะวิทยาศาสตร์และเทคโนโลยีการเกษตรมียอดการเข้าใช้เว็บไซต์น้อยที่สุด เมื่อเปรียบเทียบทั้ง 3 คณะและได้มีแนวโน้มเพิ่มขึ้นในบางช่วงเวลา เช่น ในวันที่ 15/11/2556 มียอดการเข้าใช้มากที่สุด และหากดูแน้วโน้มการเข้าชมเว็บไซต์ต่อไปอีก 7 ช่วง ดังแสดงในภาพที่ 4.12 จะเห็นว่า แนวโน้นการเข้าใช้เว็บไซต์ของหน่วยงานที่สังกัดคณะบริหารธุรกิจและศิลปศาสตร์มีแนวโน้นการเข้าใช้อยู่ในทิศทางที่เพิ่มขึ้น และการเข้าใช้เว็บไซต์ของหน่วยงานที่สังกัดคณะวิทยาศาสตร์และเทคโนโลยีการเกษตร และคณะวิศวกรรมศาสตร์มีแนวโน้มการเข้าใช้ในช่วงระยะเวลาแรกที่ใกล้เคียงกันต่อมาคณะวิทยาศาสตร์และเทคโนโลยีการเกษตร มีแน้วโน้มไปทิศทางที่ลดลง



ภาพที่ 4.12 แสดงแนวโน้มการเข้าชมเว็บไซต์ของคณะต่อไปอีก 7 ช่วง

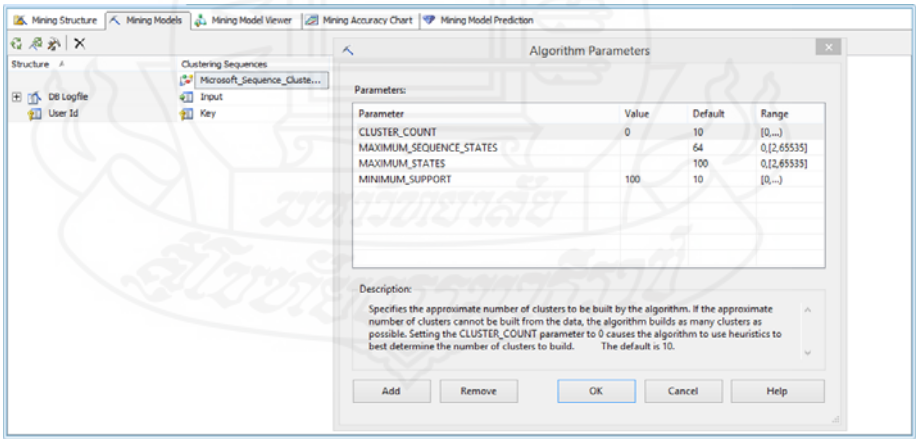
ส่วนแนวโน้มการเข้าใช้ในภาพรวมทั้งหมด พบว่า จำนวนการเข้าใช้ในแต่ละเดือนมีจำนวนการเข้าใช้ที่ใกล้เคียงกันและจะมีบางเดือนที่ยอดการเข้าใช้มีจำนวนสูงขึ้น เช่น ช่วงระหว่างเดือนพฤษภาคม ในวันที่ 25/05/2556 - 30/05/2556 มียอดการเข้าใช้เพิ่มขึ้น และช่วงระหว่างเดือนพฤศจิกายน ในวันที่ 10/11/2556 - 15/11/2556 มียอดการเข้าใช้เพิ่มขึ้น และหากดูแนวโน้มการเข้าชมเว็บไซต์ต่อไปอีก 100 ช่วง จะพบว่า แนวโน้มการเข้าใช้จะลดลงในช่วงเดือนมกราคม และจะปรับตัวเพิ่มขึ้นในเดือนกุมภาพันธ์ ดังแสดงภาพที่ 4.13



ภาพที่ 4.13 แสดงผลการพยากรณ์จำนวนผู้เข้าใช้ทั้งหมด

3.3 ผลจากการใช้เทคนิคการจัดกลุ่มโดยใช้ลำดับ หรือซีเควันซ์ คลัสเตอริง ของข้อมูลการเข้าใช้งานเว็บไซต์ทั้งหมด 360,902 เรคอร์ด เพื่อศึกษาลำดับการเข้าใช้หน้าเว็บในแต่ละหน้า ซึ่ง Sequence Clustering เป็นอัลกอริทึมที่ใช้ในการวิเคราะห์ข้อมูลที่มีลักษณะเรียงลำดับเหตุการณ์ เช่น ลำดับการเข้าชมหน้าเว็บของผู้เข้าใช้แต่ละครั้ง ผู้วิจัยได้ดึงข้อมูลจากตาราง DB_Logfile และตาราง DB_User นำมาวิเคราะห์ในอัลกอริทึมนี้

โดยกำหนดค่าพารามิเตอร์ดังแสดงในภาพที่ 4.14



ภาพที่ 4.14 แสดงการกำหนดค่าพารามิเตอร์ของเทคนิคลำดับการจัดกลุ่ม

จากภาพที่ 4.14 มีการกำหนดค่าพารามิเตอร์ดังต่อไปนี้

CLUSTER_COUNT เป็นค่าที่ใช้กำหนดจำนวนคลัสเตอร์ที่ต้องการ ถ้ากำหนดค่าเป็น 0 อัลกอริทึมจะกำหนดจำนวนคลัสเตอร์ที่ดีที่สุดให้ โดยปกติค่าดีฟอลท์คือ 0

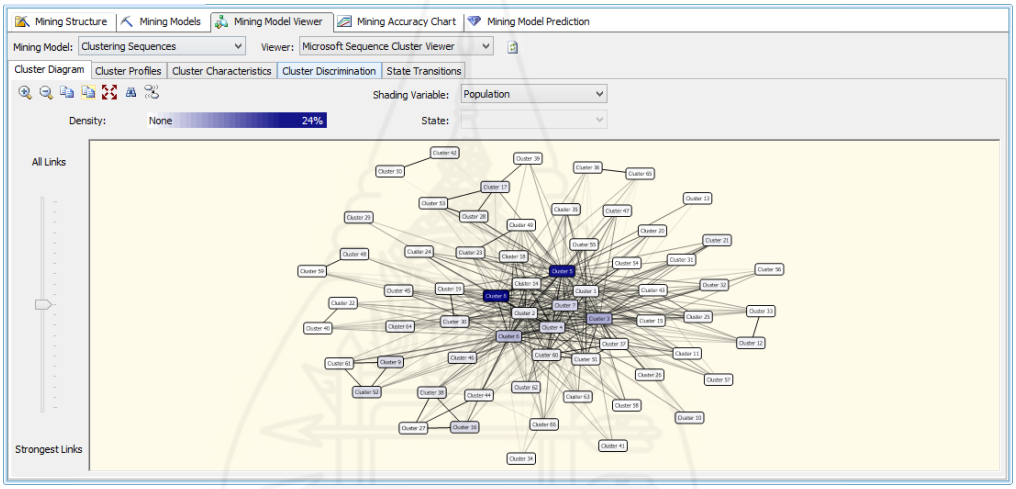
MAXIMUM_STATES เป็นค่าสูงสุดของสถานะของแอตทริบิวต์ในอัลกอริทึม โดยปกติค่าดีฟอลท์คือ 100

MAXIMUM_SEQUENCE_STATES เป็นค่าสูงสุดของจำนวนสถานะในการเรียงลำดับของแอตทริบิวต์ โดยปกติค่าดีฟอลท์คือ 64

MINIMUM_SUPPORT เป็นจำนวนเคสต่ำสุดที่ต้องมีในแต่ละคลัสเตอร์ทั้งนี้เพื่อป้องกันมิให้ในแต่ละคลัสเตอร์มีจำนวนเคสน้อยเกินไปในที่นี้กำหนดไว้เท่ากับ 100 ปกติค่าดีฟอลท์คือ 10

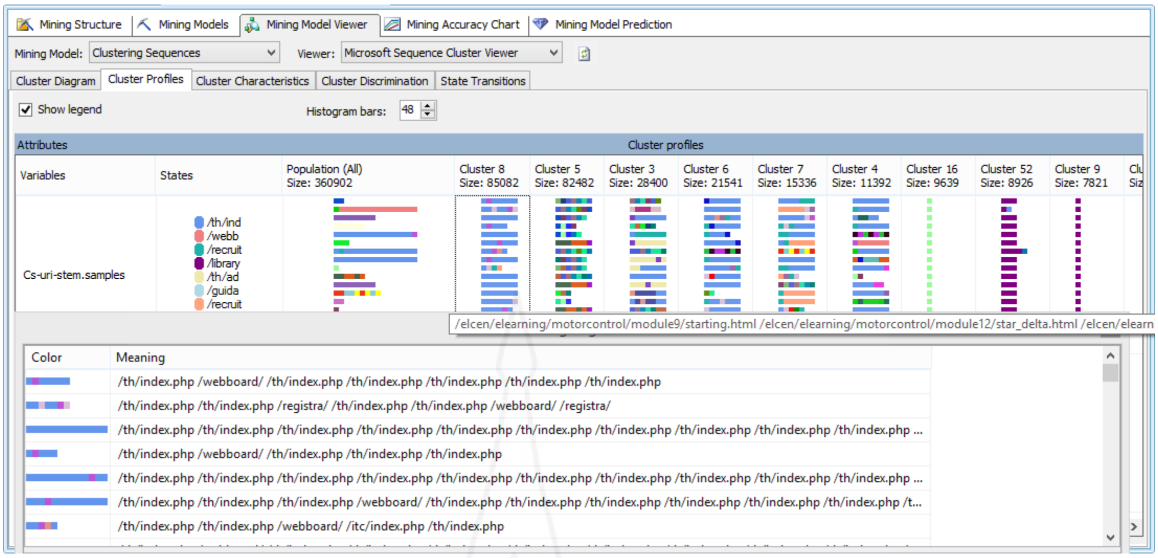
สำหรับรายละเอียดที่ค้นพบมีดังนี้

เมื่อคลิกที่แท็บ Cluster Diagram จะพบว่า อัลกอริทึมได้แบ่งคลัสเตอร์ออกเป็น 66 คลัสเตอร์คลัสเตอร์ที่ 8 เป็นคลัสเตอร์ที่ใหญ่ที่สุดมีจำนวน 85,082 เรคอร์ด และคลัสเตอร์ที่ 19 เป็นคลัสเตอร์ที่เล็กที่สุดมีจำนวน 24 เรคอร์ดดังแสดงในภาพที่ 4.15



ภาพที่ 4.15 แสดงคลัสเตอร์ที่ได้จากการใช้เทคนิคลำดับการจัดกลุ่ม

เมื่อคลิกที่แท็บ Cluster Profiles จะแสดงรายละเอียดของแต่ละคลัสเตอร์ดังแสดงในภาพที่ 4.16



ภาพที่ 4.16 แสดงรายละเอียดหน้าเว็บในแต่ละคลัสเตอร์

จากภาพที่ 4.16 แต่ละคอลัมน์จะแทนแต่ละคลัสเตอร์ และในแต่ละคลัสเตอร์ประกอบด้วยแท่งกราฟฮิสโตแกรมแนวนอนกราฟแต่ละแท่งจะแทนลำดับที่ของหน้าเว็บ โดยสามารถดูรายละเอียดได้โดยการเลือก Show Legend จะปรากฏหน้าต่าง Mining Legend ในภาพเป็นคลัสเตอร์ที่ 8 ซึ่งจะเห็นได้ว่า ผู้ใช้ในคลัสเตอร์นี้มีการเข้าใช้หน้าเว็บเป็นSequenceต่างๆ ดังต่อไปนี้

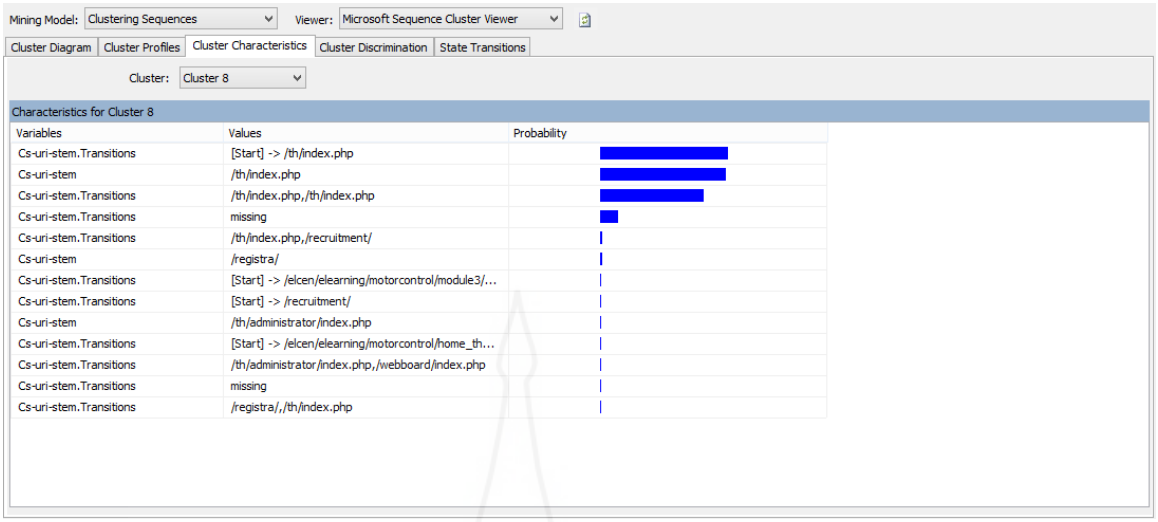
Sequenceที่ 1 ชุดหน้าเว็บ /th/index.php /webboard/ /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php

Sequence ที่ 2 ชุดหน้าเว็บ /th/index.php /th/index.php /registra/ /th/index.php /th/index.php /webboard/ /registra/

Sequenceที่ 3 ชุดหน้าเว็บ /th/index.php /webboard/ /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php /th/index.php

ฯลฯ

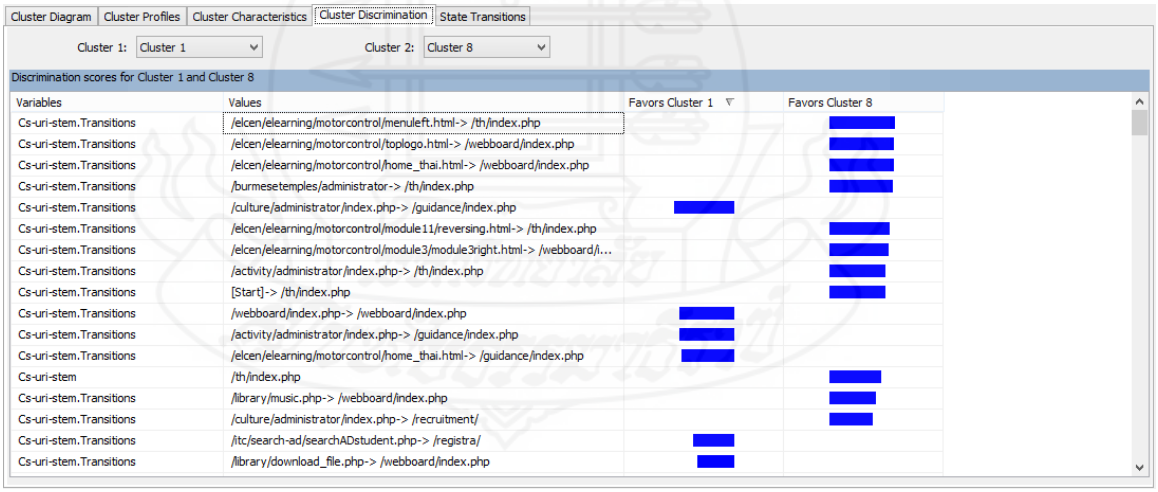
ในกรณีที่ต้องการดูรายละเอียดคุณลักษณะของแต่ละคลัสเตอร์สามารถคลิกที่แท็บ Cluster Characteristics จะมีหน้าต่างใหม่เปิดขึ้นมาสามารถเลือกคลัสเตอร์ที่ต้องการได้ ดังแสดงในภาพที่ 4.17



ภาพที่ 4.17 แสดงคุณลักษณะของคลัสเตอร์ที่ 8

จากภาพที่ 4.17 คลัสเตอร์การเข้าใช้เว็บไซต์ เริ่มจากการเข้าเรียกใช้หน้าเว็บ /th/index.php มากที่สุด ด้วยความน่าจะเป็น 0.95 และน้อยที่สุดคือ หน้าเว็บ /registra/,/th/index.php ด้วยความน่าจะเป็น 0.06

นอกจากนี้ยังสามารถเปรียบเทียบคุณลักษณะระหว่างคลัสเตอร์สองคลัสเตอร์ได้ โดยการคลิกที่แท็บ Cluster Discrimination ดังแสดงในภาพที่ 4.18

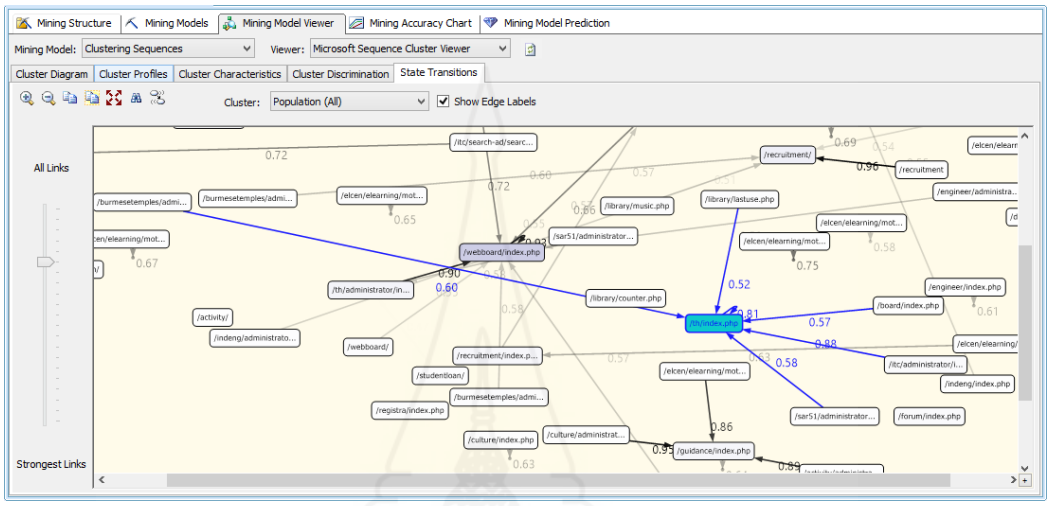


ภาพที่ 4.18 เปรียบเทียบลักษณะการเรียกใช้หน้าเว็บระหว่างคลัสเตอร์ที่ 1 กับคลัสเตอร์ที่ 8

จากภาพที่ 4.18 การเข้าใช้หน้าเว็บกลุ่มที่ 1 ส่วนมากจะเริ่มต้นเรียกใช้หน้าเว็บ /culture/administrator/index.php-> /guidance/index.php ร้อยละ 93.44 และการเข้าใช้หน้า

เว็บกลุ่มที่ 8 ส่วนมากจะเริ่มต้นเรียกใช้หน้าเว็บ/ elcen/elearning/motorcontrol/menuleft.html-> /th/index.php ร้อยละ 100

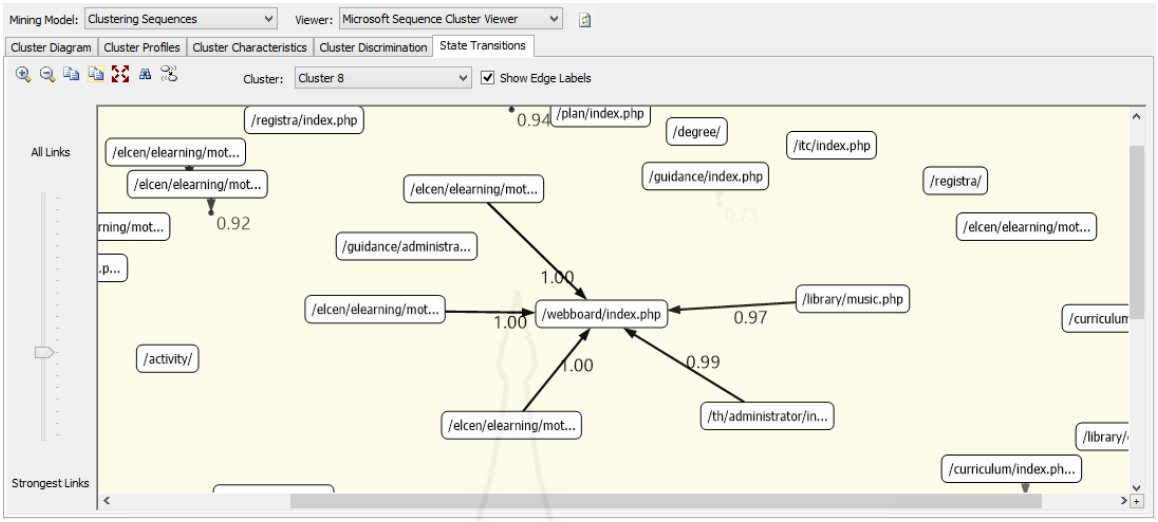
การเปลี่ยนแปลงสถานะในการเรียกใช้หน้าเว็บสามารถดูได้โดยการคลิกที่แท็บ State Transitions ดังแสดงในภาพที่ 4.19



ภาพที่ 4.19 ภาพรวมการเปลี่ยนแปลงสถานการณ์เข้าใช้หน้าเว็บของมหาวิทยาลัย

จากภาพที่ 4.19 พบว่า การเรียกใช้หน้าเว็บในกลุ่มตัวอย่างที่มีการเรียกใช้หน้าเว็บ /library/laseuse.php แล้วจะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.52 และการเรียกใช้หน้าเว็บ ict/administrator/index.php แล้วจะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.88 และการเรียกใช้หน้าเว็บ /sar51/administrator/ แล้วจะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.58 และการเรียกใช้หน้าเว็บ /borad/index.php/ แล้วจะเรียกใช้หน้าเว็บ/th/index.php ด้วยความน่าจะเป็น 0.57 และการเรียกใช้หน้าเว็บ /burmesetemples/ administrator/index.php แล้วจะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.60 และการเรียกเข้าใช้หน้าเว็บซ้ำกันในหน้า /th/index.php ด้วยความน่าจะเป็น 0.81

สำหรับการเรียกใช้หน้าเว็บในคลัสเตอร์ที่ 8 มีการเปลี่ยนแปลงสถานะการเข้าใช้หน้าเว็บ ดังแสดงในภาพที่ 4.20

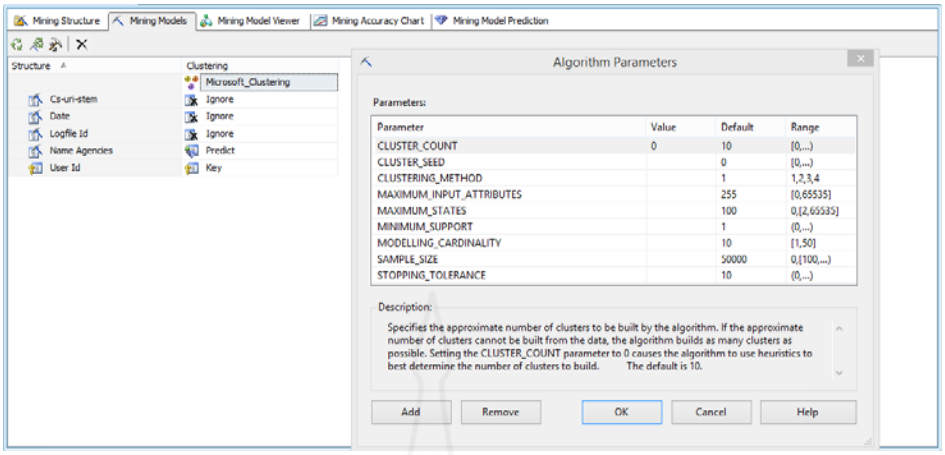


ภาพที่ 4.20 การเปลี่ยนแปลงสถานะการเข้าใช้หน้าเว็บของมหาวิทยาลัยในคลัสเตอร์ที่ 8

จากภาพที่ 4.20 พบว่า การเรียกใช้หน้าเว็บในคลัสเตอร์ที่ 8 มีการเรียกใช้หน้าเว็บ /elcen/elearning/motorcontrol/module2/symbol2.html แล้วจะเรียกใช้หน้าเว็บ /webboard/index.php ด้วยความน่าจะเป็น 1.00 และการเรียกใช้หน้าเว็บ /elcen/elearning/motorcontrol/module9/starting.html แล้วจะเรียกใช้หน้าเว็บ /webboard/index.php ด้วยความน่าจะเป็น 1.00 และการเรียกใช้หน้าเว็บ /elcen/elearning/motorcontrol/module3/module3right.htm แล้วจะเรียกใช้หน้าเว็บ webboard/index.php ด้วยความน่าจะเป็น 1.00 และการเรียกใช้หน้าเว็บ /library/music.php แล้วจะเรียกใช้หน้าเว็บ /webboard/index.php ด้วยความน่าจะเป็น 0.97 และการเรียกใช้หน้าเว็บ/th/administrator/ แล้วจะเรียกใช้หน้าเว็บ /webboard/index.php ด้วยความน่าจะเป็น 0.99

3.4 ผลจากการใช้เทคนิคการจัดกลุ่มหรือคลัสเตอริง เป็นอัลกอริทึมที่ใช้ในการจำแนกหรือจัดกลุ่มการใช้งานเว็บไซต์ เช่น ข้อมูลความถี่ในการเข้าชมหน้าเว็บของหน่วยงานต่างๆ ภายในมหาวิทยาลัย โดยผู้วิจัยได้ดึงข้อมูลตาราง DB_Logfile นำมาวิเคราะห์ในอัลกอริทึมนี้สำหรับรายละเอียดหรือผลที่ได้จากการใช้เทคนิคการจัดกลุ่มมีดังต่อไปนี้

ผู้วิจัยกำหนดค่าพารามิเตอร์ดังแสดงในภาพที่ 4.21



ภาพที่ 4.21 แสดงการกำหนดค่าพารามิเตอร์ของเทคนิคการจัดกลุ่ม

จากภาพที่ 4.21 มีการกำหนดค่าพารามิเตอร์ดังต่อไปนี้

CLUSTER_COUNT ซึ่งเป็นค่าของจำนวนคลัสเตอร์ที่ต้องการให้อัลกอริทึมจำแนก ให้ค่าดีฟอลท์คือ 10 แต่ถ้ากำหนดไว้เท่ากับ 0 หมายความว่าให้อัลกอริทึมกำหนดจำนวนคลัสเตอร์ที่ต้องการจำแนกเองด้วยการใช้หลักการของฮิวริสติกส์ในการจำแนก

CLUSTER_SEED เป็นค่าตัวเลขสุ่มเริ่มต้นที่จะใช้ในการจำแนกคลัสเตอร์ค่าดีฟอลท์คือ 0

CLUSTERING_METHOD เป็นการกำหนดวิธีการจำแนกคลัสเตอร์ที่จะใช้ในอัลกอริทึม ค่าดีฟอลท์คือค่า 1 หมายถึงใช้วิธี Scalable EM (Expected Mean) ค่า 2 หมายถึง Non-scalable EM ค่า 3 หมายถึง Scalable K-means และค่า 4 หมายถึง Non-scalable K-means

MAXIMUM_INPUT_ATTRIBUTES คือ ค่าของจำนวนแอตทริบิวต์สูงสุดที่สามารถนำมาใช้ในการจำแนกคลัสเตอร์ค่าดีฟอลท์คือ 255

MAXIMUM_STATES คือค่าที่ใช้ควบคุมจำนวนสถานะทั้งหมดที่แอ็ตทริบิวต์แต่ละแอ็ตทริบิวต์สามารถเปลี่ยนสถานะได้ค่านี้ไม่ควรกำหนดไว้สูงเพราะจะมีผลต่อหน่วยความจำที่ต้องใช้ในการทำงานของอัลกอริทึมอาจไม่พอค่าดีฟอลท์คือ 100

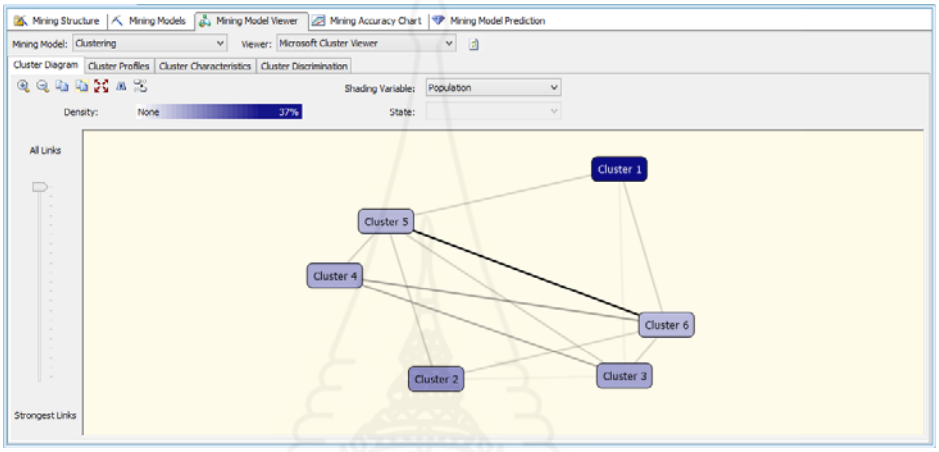
MINIMUM_SUPPORT คือค่าที่กำหนดว่าในแต่ละคลัสเตอร์เมื่อจำแนกแล้วควรมีจำนวนข้อมูลไม่น้อยกว่าค่าที่กำหนดในพารามิเตอร์ตัวนี้ดังนั้นค่านี้จึงใช้ในการพิจารณาว่าคลัสเตอร์ใดควรตัดทิ้งหรือจำแนกใหม่การกำหนดค่านี้สูงเกินไปอาจทำให้ได้ผลลัพธ์ที่ไม่ดีนักค่าดีฟอลท์คือ 1

MODELLING CARDINALITY เป็นค่าที่ควบคุมจำนวนแบบจำลองที่อัลกอริทึมจะสร้างขึ้นในระหว่างที่มีการจำแนกคลัสเตอร์หากกำหนดค่านี้ลดลงก็จะเพิ่มสมรรถนะในการทำงานของอับกอริธึมค่าดีฟอลท์คือ 10

SAMPLE_SIZE เป็นค่าแสดงจำนวนของเรคอร์ดข้อมูลที่จะใช้ในแต่ละขั้นตอนของกรอบงานซึ่งปรับเปลี่ยนได้ ถ้ากำหนดค่านี้ลดลงจะทำให้อัลกอริทึมลดการทำงานลงเพราะอัลกอริทึมจะไม่ประมวลผลเรคอร์ดข้อมูลทั้งหมด โดยเฉพาะอย่างยิ่งหากกำหนดร่วมกับค่า STOPPING_TOLERANCE ที่กำหนดให้มีค่ามาก ค่าดีฟอลท์คือ 50,000 เรคอร์ด

STOPPING_TOLERANCE เป็นค่าที่ใช้กำหนดว่าเมื่อใดอัลกอริทึมควรจะหยุดทำงาน โดยค่านี้จะแทนจำนวนเรคอร์ดสูงสุดที่สามารถเปลี่ยนไปเป็นสมาชิกในอีกคลัสเตอร์ได้ก่อนที่จะพิจารณาว่าแบบจำลองนั้นต้องหยุดทำงานแล้ว ค่าดีฟอลท์คือ 10

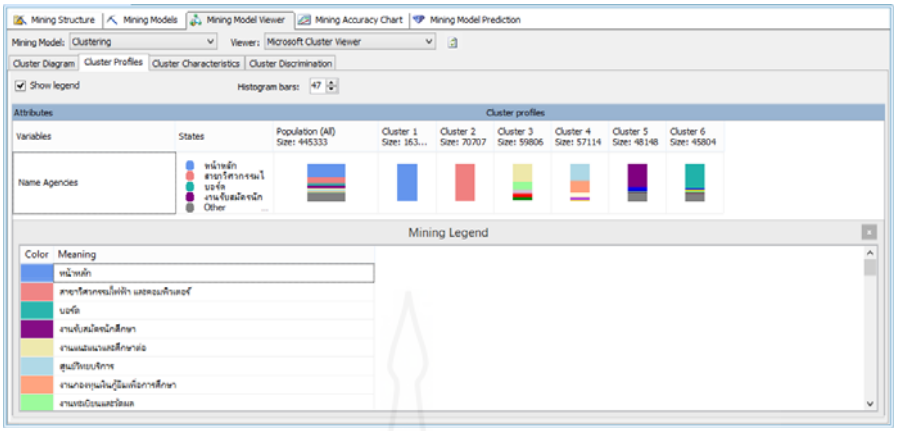
เมื่อคลิกที่แท็บ Model Mining Viewer พบว่า อัลกอริธึมได้จำแนกออกเป็น 6 คลัสเตอร์ และเมื่อเลื่อนสไลด์บาร์ด้านซ้ายมือซึ่งเป็นสไลด์บาร์ที่แสดงความสัมพันธ์ระหว่างคลัสเตอร์โดยเลื่อนลงจะพบว่าความสัมพันธ์ดังแสดงในภาพที่ 4.22



ภาพที่ 4.22 แสดงจำนวนคลัสเตอร์และความสัมพันธ์ระหว่างคลัสเตอร์

จากภาพที่ 4.22 พบว่า จำนวนการเข้าเว็บไซต์ของผู้เข้าใช้แบ่งเป็น 6 คลัสเตอร์ ได้แก่ คลัสเตอร์ที่ 1 มีจำนวนมากที่สุด คือ 163,754 เรคอร์ด คลัสเตอร์ที่ 2 มีจำนวน 70,707 เรคอร์ด คลัสเตอร์ที่ 3 มีจำนวน 59,806 เรคอร์ด คลัสเตอร์ที่ 4 มีจำนวน 57,114 เรคอร์ด คลัสเตอร์ที่ 5 มีจำนวน 48,148 เรคอร์ด คลัสเตอร์ที่ 6 มีจำนวนน้อยที่สุด 45,804 เรคอร์ด ประชากรหรือ Population ที่นำมาประมวลผลโดยใช้เทคนิคการจัดกลุ่มมีข้อมูลการเข้าใช้เว็บไซต์ที่สุ่มมาวิเคราะห์ทั้งหมด 445,333 เรคอร์ด

หากต้องการศึกษารายละเอียดคุณลักษณะของคลัสเตอร์ทั้งหมดสามารถดูได้โดยการคลิกที่แท็บ Cluster Profiles จะปรากฏหน้าต่างแสดงคุณลักษณะของแต่ละคลัสเตอร์ทั้งหมดเรียงในลักษณะของคอลัมน์ดังแสดงในภาพที่ 4.23



ภาพที่ 4.23 แสดงคุณลักษณะของการเข้าชมเว็บไซต์โดยจำแนกเป็น 6 คลัสเตอร์

จากภาพที่ 4.23 พบว่า ประชากรหรือ Population ที่นำมาประมวลผลโดยใช้เทคนิคการจัดกลุ่มที่สุ่มมาวิเคราะห์ทั้งหมด 445,333 เรคอร์ด ของข้อมูลการเข้าใช้งานเว็บไซต์โดยแบ่งเป็นหมวดงานหรือหน่วยงานย่อยของมหาวิทยาลัย จะเห็นได้ว่า จำนวนหน้าเว็บที่เข้าใช้มากที่สุดคือ หน้าหลักจำนวน 163,754 เรคอร์ด, สาขาวิศวกรรมไฟฟ้าและคอมพิวเตอร์จำนวน 70,707 เรคอร์ด,บอร์ดจำนวน 29,631 เรคอร์ด ฯลฯ เรียงลงมาตามลำดับและแสดงผลการเข้าใช้ของประชากรหรือ Population แต่ละคลัสเตอร์ ดังผลตารางที่ 4.9

ตารางที่ 4.9 ตารางแสดงผลการเข้าใช้ของประชากรหรือ Population แต่ละคลัสเตอร์

States	Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
	445,333	163,754	70,707	59,806	57,114	48,148	45,804
หน้าหลัก	163,754	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
สาขาวิศวกรรมไฟฟ้าและคอมพิวเตอร์	70,707	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%
บอร์ด	29,631	0.00%	0.00%	0.00%	0.00%	0.00%	64.30%
งานรับสมัครนักศึกษา	28,658	0.00%	0.00%	0.00%	0.00%	62.00%	0.00%
งานแนะแนวและศึกษาต่อ	28,204	0.00%	0.00%	46.90%	0.00%	0.00%	0.00%
ศูนย์วิทยบริการ	26,398	0.00%	0.00%	0.00%	45.10%	0.00%	0.00%
งานกองทุนเงินกู้ยืมเพื่อการศึกษา	18,885	0.00%	0.00%	0.00%	32.30%	0.00%	0.00%
งานทะเบียนและวัดผล	13,509	0.00%	0.00%	22.50%	0.00%	0.00%	0.00%
ฝ่ายยุทธศาศตร์และแผน	6,745	0.00%	0.00%	11.20%	0.00%	0.00%	0.00%
ฝ่ายทรัพยากรมนุษย์	6,671	0.00%	0.00%	0.50%	10.90%	0.00%	0.00%
งานหลักสูตรและตำราเรียน	6,081	0.00%	0.00%	0.00%	0.00%	8.20%	5.00%

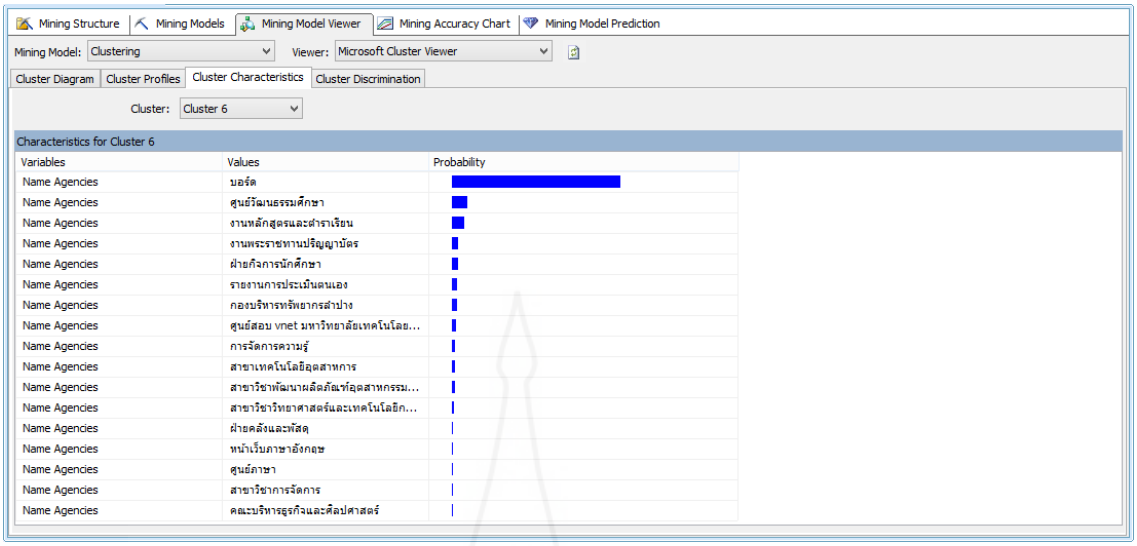
ตารางที่ 4.9 (ต่อ)

States	Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
	445,333	163,754	70,707	59,806	57,114	48,148	45,804
สาขาบริหารธุรกิจ	5,968	0.00%	0.00%	9.90%	0.00%	0.00%	0.00%
ศูนย์เทคโนโลยีสำรสนเทศ	5,380	0.00%	0.00%	8.90%	0.00%	0.00%	0.00%
คณะวิศวกรรมศาสตร์	5,146	0.00%	0.00%	0.00%	8.80%	0.00%	0.00%
ศูนย์วัฒนธรรมศึกษา	4,493	0.00%	0.00%	0.00%	2.90%	0.00%	6.10%
สาขาเทคโนโลยีอุตสำหการ	3,587	0.00%	0.00%	0.00%	0.00%	6.30%	1.50%
ฝ่ายกิจการนักศึกษา	2,089	0.00%	0.00%	0.00%	0.00%	2.20%	2.40%
รายงานการประเมินตนเอง	1,994	0.00%	0.00%	0.00%	0.00%	2.10%	2.20%
งานพระราชทานปริญญาบัตร	1,988	0.00%	0.00%	0.00%	0.00%	1.80%	2.50%
สาขาวิชาพัฒนาผลิตภัณฑ์อุตสำหกรรมเกษตร	1,907	0.00%	0.00%	0.00%	0.00%	2.90%	1.20%
กองบริหารทรัพยากรลำปาง	1,880	0.00%	0.00%	0.00%	0.00%	2.10%	2.00%
การจัดการความรู้	958	0.00%	0.00%	0.00%	0.00%	0.60%	1.50%
ศูนย์สอบ vnet มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง	927	0.00%	0.00%	0.00%	0.00%	0.30%	1.70%
หน้าเว็บภาษาอังกฤษ	785	0.00%	0.00%	0.00%	0.00%	1.00%	0.70%
สาขาวิชาวิทยาศาสตร์และเทคโนโลยีการอาหาร	653	0.00%	0.00%	0.00%	0.00%	0.50%	0.90%
สาขาการบัญชี	579	0.00%	0.00%	0.00%	0.00%	0.80%	0.40%
สาขาอุตสำหกรรมเกษตร	571	0.00%	0.00%	0.00%	0.00%	0.80%	0.40%
ฝ่ายคลังและพัสดุ	538	0.00%	0.00%	0.00%	0.00%	0.40%	0.80%
สาขาวิชาสัตวศาสตร์	526	0.00%	0.00%	0.00%	0.00%	0.90%	0.20%
สาขาวิชาการจัดการ	449	0.00%	0.00%	0.00%	0.00%	0.40%	0.60%
ฝ่ายวิทยบริการและเทคโนโลยีสำรสนเทศ	426	0.00%	0.00%	0.00%	0.00%	0.60%	0.30%
กองการศึกษา	409	0.00%	0.00%	0.00%	0.00%	0.60%	0.30%
สาขาไฟฟ้าและเทคโนโลยีคอมพิวเตอร์	400	0.00%	0.00%	0.00%	0.00%	0.40%	0.40%
สาขาวิชาการตลาด	387	0.00%	0.00%	0.00%	0.00%	0.60%	0.20%
สาขาวิชาเทคโนโลยีภูมิทัศน์	336	0.00%	0.00%	0.00%	0.00%	0.40%	0.40%
ศูนย์ภาษา	324	0.00%	0.00%	0.00%	0.00%	0.00%	0.70%
คณะบริหารธูรกิจและศิลปศาสตร์	319	0.00%	0.00%	0.00%	0.00%	0.10%	0.50%
สาขาวิชาคอมพิวเตอร์ธุรกิจ	309	0.00%	0.00%	0.00%	0.00%	0.30%	0.30%

ตารางที่ 4.9 (ต่อ)

States	Population	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
	445,333	163,754	70,707	59,806	57,114	48,148	45,804
สาขาเทคโนโลยีเครื่องกล	309	0.00%	0.00%	0.00%	0.00%	0.50%	0.20%
สำนักงานรองอธิการบดี	300	0.00%	0.00%	0.00%	0.00%	0.50%	0.20%
สาขาวิทยาศาสตร์	237	0.00%	0.00%	0.00%	0.00%	0.40%	0.10%
คณะวิทยาศาสตร์และเทคโนโลยีการเกษตร	218	0.00%	0.00%	0.00%	0.00%	0.20%	0.20%
สาขาเทคโนโลยีเครื่องกล	309	0.00%	0.00%	0.00%	0.00%	0.50%	0.20%
สาขาวิชาการท่องเที่ยว	184	0.00%	0.00%	0.00%	0.00%	0.20%	0.20%
ผู้ดูแลระบบ	183	0.00%	0.00%	0.00%	0.00%	0.20%	0.20%
สาขาวิชาระบบสำรสนเทศทางคอมพิวเตอร์	180	0.00%	0.00%	0.00%	0.00%	0.30%	0.10%
ฝ่ายบริหารทั่วไป	180	0.00%	0.00%	0.00%	0.00%	0.20%	0.20%
สำนักประกันคุณภาพการศึกษา	176	0.00%	0.00%	0.00%	0.00%	0.10%	0.30%
ฝ่ายวิชาการ	138	0.00%	0.00%	0.00%	0.00%	0.30%	0.00%
BPC (คณะกรรมการบริหารประจำเขตพื้นที่ลำปาง)	127	0.00%	0.00%	0.00%	0.00%	0.10%	0.20%
ฝ่ายวิจัยและบริการวิชาการ	96	0.00%	0.00%	0.00%	0.00%	0.10%	0.10%
สาขาวิชาประมง	94	0.00%	0.00%	0.00%	0.00%	0.20%	0.00%
คลีนิคเทคโนโลยี	92	0.00%	0.00%	0.00%	0.00%	0.10%	0.10%
สาขาวิชาภาษาอังกฤษเพื่อการสื่อสารสำกล	91	0.00%	0.00%	0.00%	0.00%	0.10%	0.10%
รูปกิจกรรม	62	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%
ฝ่ายบริหารงานทั่วไป	46	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%
ฝ่ายบริการ	21	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
งานวิเทศสัมพันธ์	14	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
สาขาวิชาการท่องเที่ยว	184	0.00%	0.00%	0.00%	0.00%	0.20%	0.20%

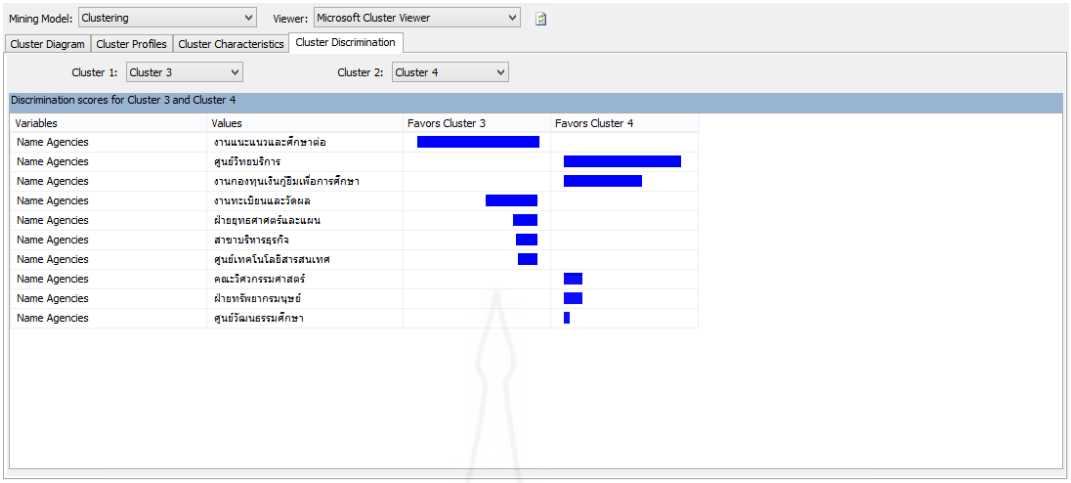
ในกรณีที่ต้องการศึกษารายละเอียดเฉพาะคลัสเตอร์ใดคลัสเตอร์หนึ่งให้คลิกที่แท็บ Cluster Characteristics จะมีหน้าต่างใหม่เปิดขึ้นมาสามารถเลือคลัสเตอร์ที่ต้องการศึกษาได้ โดยเลือกในช่อง Cluster ที่อยู่ด้านบนซ้ายสำหรับในที่นี้เลือกคลัสเตอร์ที่ 6 ดังแสดงในภาพที่ 4.24



ภาพที่ 4.24 แสดงคุณลักษณะของคลัสเตอร์ที่ 6

จากภาพที่ 4.24 เป็นข้อมูลการเข้าใช้เว็บไซต์ในหน้าเว็บของหมวดงานหรือหน่วยงานต่างๆ คลัสเตอร์ที่ 6 พบว่า จำนวนประชากรส่วนใหญ่เข้าใช้บอร์ดมากที่สุด ร้อยละ 64.34 เข้าใช้เว็บไซต์ศูนย์วัฒนธรรมศึกษา ร้อยละ 6.07 เข้าใช้เว็บไซต์งานพระราชทานปริญญาบัตร ร้อยละ 2.52 เข้าใช้เว็บไซต์ฝ่ายกิจการนักศึกษา ร้อยละ 2.36 เข้าใช้เว็บไซต์รายงานการประเมินตนเอง ร้อยละ 2.24 เข้าใช้เว็บไซต์กองบริหารทรัพยากรลำปาง ร้อยละ 2.01 เข้าใช้เว็บไซต์ศูนย์สอบ vnet มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง ร้อยละ 1.67 เข้าใช้เว็บไซต์ฝ่ายกิจการนักศึกษา ร้อยละ 2.36 เข้าใช้เว็บไซต์การจัดการความรู้ ร้อยละ 1.51 เข้าใช้เว็บไซต์สาขาเทคโนโลยีอุตสาหการ ร้อยละ 1.46 เข้าใช้เว็บไซต์สาขาวิชาพัฒนาผลิตภัณฑ์อุตสาหกรรมเกษตร ร้อยละ 1.23 เข้าใช้เว็บไซต์สาขาวิชาวิทยาศาสตร์และเทคโนโลยีการอาหาร ร้อยละ 0.87 เข้าใช้เว็บไซต์ฝ่ายคลังและพัสดุ ร้อยละ 0.77 เข้าใช้เว็บไซต์ หน้าเว็บภาษาอังกฤษ ร้อยละ 0.71 เข้าใช้เว็บไซต์ศูนย์ภาษา ร้อยละ 0.67 เข้าใช้เว็บไซต์ สาขาวิชาการจัดการ ร้อยละ 0.56 และน้อยที่สุดเข้าใช้เว็บไซต์คณะบริหารธุรกิจและศิลปศาสตร์ ร้อยละ 0.54

นอกจากนี้ยังสามารถเปรียบเทียบคุณลักษณะของคลัสเตอร์เป็นคู่ๆได้ด้วยโดยการคลิกที่แท็บ Cluster Discrimination จะมีหน้าต่างเปิดขึ้นมาใหม่สามารถเลือกคลัสเตอร์ที่ต้องการเปรียบเทียบคุณลักษณะได้โดยคลิกที่ช่อง Cluster1 และ Cluster2 ที่อยู่ด้านบนในแต่ละช่องจะมี drop down list ให้เลือกสำหรับในที่นี้จะเปรียบเทียบคุณลักษณะของการเข้าใช้เว็บไซต์ในคลัสเตอร์ที่ 3 กับคลัสเตอร์ที่ 4 ดังแสดงในภาพที่ 4.25

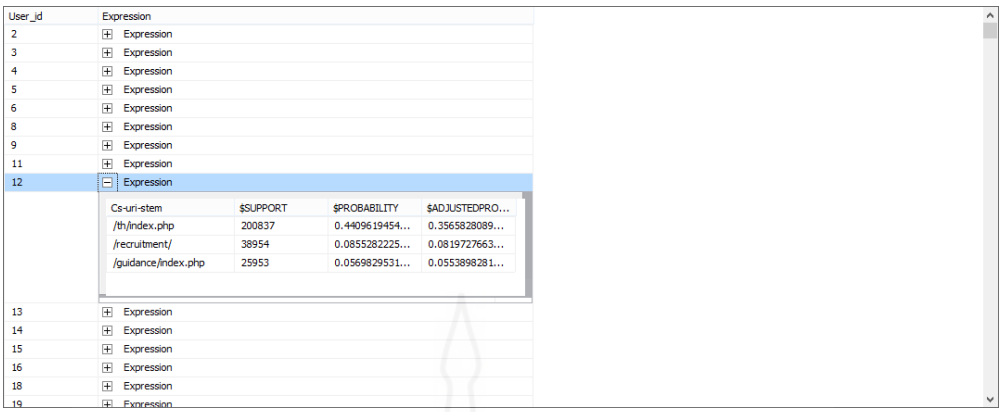


ภาพที่ 4.25 แสดงการเปรียบเทียบคุณลักษณะของคลัสเตอร์ที่ 3 กับคลัสเตอร์ที่ 4

จากภาพที่ 4.25 พบว่าจำนวนประชากรที่เข้าชมเว็บไซต์คลัสเตอร์ที่ 3 ร้อยละ 100 เข้าใช้เว็บไซต์งานแนะแนวนักศึกษา และรองลงมาเข้าใช้เว็บไซต์งานทะเบียนและวัดผล ร้อยละ 42.63 เว็บไซต์ฝ่ายยุทธศาสตร์และแผน ร้อยละ 20.31 เว็บไซต์สาขาบริหารธุรกิจ ร้อยละ 17.87 เว็บไซต์ศูนย์เทคโนโลยีสารสนเทศ ร้อยละ 16.05 แต่จำนวนประชากรที่เข้าชมเว็บไซต์คลัสเตอร์ที่ 4 ร้อยละ 96.24 เข้าใช้เว็บไซต์ศูนย์วิทยบริการ และรองลงมาเว็บไซต์งานกองทุนเงินกู้ยืมเพื่อการศึกษา ร้อยละ 64.69 เว็บไซต์คณะวิศวกรรมศาสตร์ ร้อยละ 15.96 เว็บไซต์ฝ่ายทรัพยากรมนุษย์ ร้อยละ 15.26 และน้อยที่สุดเว็บไซต์ศูนย์วัฒนธรรมศึกษาร้อยละ 5.12

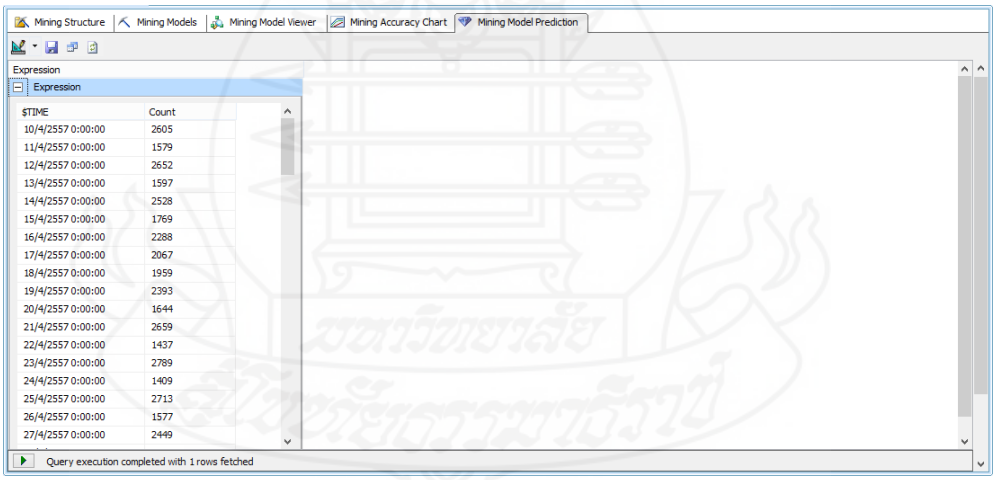
3.5 ประเมิน เป็นการประเมินผลลัพธ์จากแบบจำลองและอัลกอริทึมที่ใช้วิเคราะห์ข้อมูลว่าครอบคลุมและสามารถตอบวัตถุประสงค์ที่กำหนดไว้หรือไม่จากนั้นจึงนำเสนอประเมินแบบจำลองที่ได้จากการทำเหมืองข้อมูลเพื่อพิจารณาถึงความเหมาะสมและนำไปประยุกต์ใช้ว่าแบบจำลองที่ได้นั้นมีความแม่นยำในการทำนายมากน้อยเพียงใดโดยในการประเมินแบบจำลองนั้นจะแบ่งตามลักษณะของการทำเหมืองข้อมูลดังนี้

3.5.1 แอสโซซิเอชันรูลส์ สามารถประเมินโมเดลการทำเหมืองข้อมูลโดยการนำโมเดลนี้ไปพยากรณ์ด้วยการสร้างแบบสอบถามหรือคิวรีซึ่งใช้ฟังก์ชันในการพยากรณ์ชื่อว่าการพยากรณ์ความสัมพันธ์ (Predict Association) และกำหนดเงื่อนไขที่ต้องการการแสดงผล ตัวอย่างนี้ผู้วิจัยได้กำหนดเงื่อนไขให้นำข้อมูลมาแสดงผลคือ ข้อมูลรหัสผู้ใช้ และชื่อไฟล์หน้าเว็บ 3 ชื่อที่ผู้ใช้เรียกใช้พร้อมทั้งแสดงค่าสถิติด้วย ซึ่งจากการคิวรีพบว่าหน้าเว็บที่มีความสัมพันธ์กันดังภาพที่ 4.26



ภาพที่ 4.26 ผลการทดสอบโมเดลการทำเหมืองข้อมูลด้วยเทคนิค AssociationRule

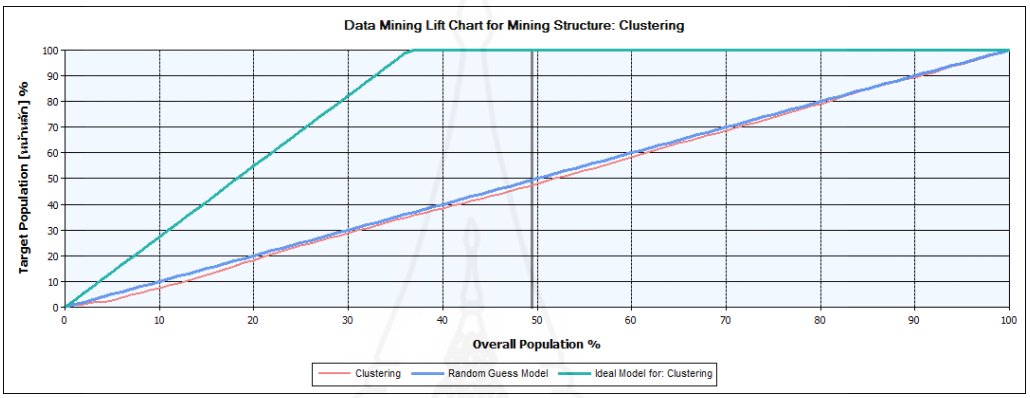
3.5.2 ไทม์ซีรีส์ สามารถประเมินโมเดลการทำเหมืองข้อมูลโดยการนำโมเดลนี้ไปพยากรณ์ด้วยการสร้างแบบสอบถามหรือคิวรีซึ่งใช้ฟังก์ชันในการพยากรณ์ชื่อว่าการพยากรณ์ไทม์ซีรีส์ (PredictTimeSeries) และกำหนดเงื่อนไขที่ต้องการแสดงผลพบว่าจำนวนผู้เข้าใช้ใกล้เคียงกับข้อมูลเดิมในแต่ละเดือน และไม่แตกต่างกันมาก เมื่อทำการเปรียบเทียบข้อมูลในแต่ละเดือน ซึ่งทำให้มั่นใจได้ว่าการใช้ไทม์ซีรีส์ในการพยากรณ์นั้นให้ความแม่นยำและสามารถนำไปใช้ในการพยากรณ์แนวโน้มการของการเข้าเยี่ยมชมเว็บไซต์ของทางมหาวิทยาลัยได้ดังภาพที่ 4.27



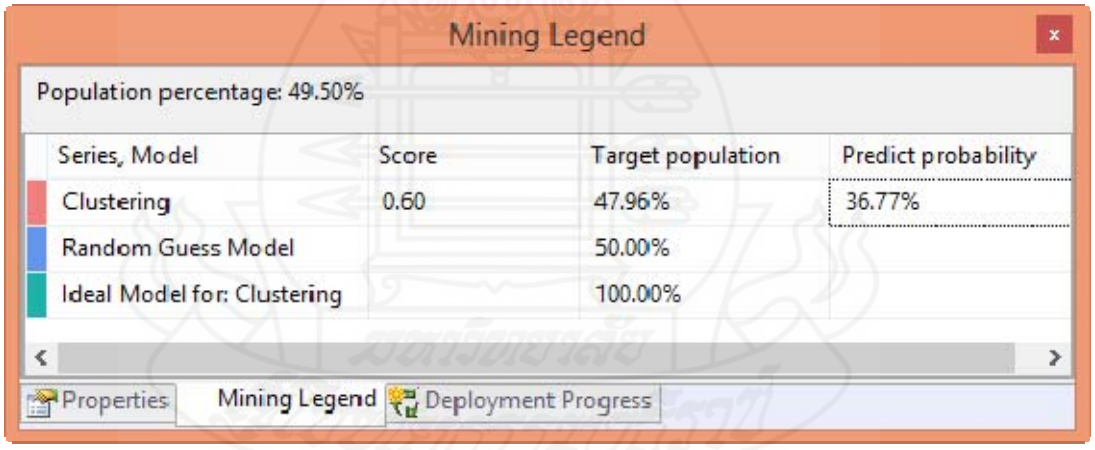
ภาพที่ 4.27 ผลการทดสอบโมเดลการทำเหมืองข้อมูลด้วยเทคนิคไทม์ซีรีส์

3.5.3 ซีเคว็นซ์ คลัสเตอริง สามารถประเมินโมเดลการทำเหมืองข้อมูล โดยการนำโมเดลนี้ไปพยากรณ์ด้วยการสร้างแบบสอบถามหรือคิวรี (Query) ซึ่งใช้ฟังก์ชันการจัดกลุ่มและใช้ฟังก์ชันการพยากรณ์ชื่อว่า Predict Clustering พบว่าจำนวนกลุ่มที่ได้ตรงกับกลุ่มที่ทำการประมวลผลด้วยเทคนิคทำเหมืองข้อมูลและการพยากรณ์ลำดับมีการเข้าหน้าเว็บ /th/index.php ซ้ำกัน 3 ลำดับ ซึ่งสอดคล้องกับผลลัพธ์ที่ได้

3.5.4 คลัสเตอริง สามารถประเมินผลโดยการนำโมเดลการทำเหมืองข้อมูลนั้นไปทำการทดสอบด้วยชุดข้อมูลสำหรับการทดสอบซึ่งได้กำหนดเอาไว้ 50% เพื่อดูว่าแพตเทิร์นที่ได้จากการสร้างโมเดลการทำเหมืองข้อมูลนั้นมีความถูกต้องเพียงใดการทดสอบความถูกต้องของการทำเหมืองข้อมูลด้วยการจัดกลุ่มใน SQL Server นั้นสามารถดูจากลิฟต์ชาร์ต (Lift Chart) ผลการทดสอบพบว่าโมเดลการทำเหมืองข้อมูลด้วยการจัดกลุ่มมีความน่าจะเป็นของการพยากรณ์ 36.70% และมีค่า Score ที่ใช้วัดประสิทธิภาพของโมเดลเท่ากับ 0.60 ดังภาพที่ 4.28 และภาพที่ 4.29



ภาพที่ 4.28 ลิฟต์ชาร์ตผลการทดสอบพบว่าโมเดลการทำเหมืองข้อมูลด้วยการจัดกลุ่ม



ภาพที่ 4.29 แสดงการผลการทดสอบโมเดลการจัดกลุ่ม

3.6 นำไปใช้งาน เป็นการนำองค์ความรู้ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลโดยอัลกอริทึมต่างๆ ไปใช้งานจริง ซึ่งผู้วิจัยได้นำเสนอผลของการทำเหมืองข้อมูลมาประยุกต์ใช้ เพื่อสนับสนุนการให้บริการทางด้านข้อมูลข่าวสารในเว็บไซต์ของมหาวิทยาลัย ดังต่อไปนี้

3.6.1 ผลการหากฎความสัมพันธ์ ของการเข้าใช้เว็บไซต์สามารถนำไปใช้เป็นแนวทางการกำหนดเชื่อมโยงลิงค์ของหน้าเว็บไซต์และออกแบบลิงค์เมนู เพื่อเป็นแนวทางการ

ตัดสินใจให้ผู้ใช้บริการได้เลือกใช้และอำนวยสะดวกในการเข้าถึงข้อมูลที่ต้องการต่อไปแล้วนำผลตอบกลับในการปรับปรุง นำมาพิจารณาปรับแก้การเชื่อมโยงลิงค์เพื่อให้สะดวกต่อผู้เข้าใช้บริการต่อไป

3.6.2 ผลการพยากรณ์จำนวนผู้เข้าใช้บริการ นำไปใช้เป็นแนวทางในการเตรียมความพร้อมในการบริการข้อมูลข่าวสาร และเฝ้าระวังการทำงานของเครื่องแม่ข่ายโดยนำมาพิจารณาในการจัดทำแผนการดำเนินงานของแผนกไอทีในการปรับปรุงแก้ไขเว็บไซต์และเฝ้าระวังการผิดพลาดจากเครื่องแม่ข่ายที่อาจเกิดขึ้นจากผู้ใช้เข้าใช้บริการพร้อมๆ กันมากเกินไปในช่วงการประกาศผลสอบและช่วงลงทะเบียนเรียนซึ่งอาจทำให้เครื่องแม่ข่ายทำงานหนักจึงทำให้ระบบล่มในที่สุด และยังสามารถนำไปใช้ในการเตรียมการเพื่อขยายโครงสร้างพื้นฐานด้านระบบเครือข่ายให้สามารถรองรับปริมาณผู้ใช้ที่จะเพิ่มมากขึ้นในอนาคตอีกด้วย

3.6.3 ผลการวิเคราะห์ด้วยเทคนิคการจัดกลุ่มโดยใช้ลำดับ การเข้าชมหน้าเว็บไซต์สามารถนำไปใช้เป็นแนวทางในการปรับปรุงหน้าเว็บและการให้ข้อมูลข่าวสารในหน่วยงานต่างๆ ของมหาวิทยาลัย โดยนำผลการจัดกลุ่มในการเข้าชมเว็บไซต์ของแต่ละหน่วยงาน ว่าผู้ใช้มีความต้องการในการรับข้อมูลข่าวสารด้านไหนมากที่สุด แล้วนำผลลัพธ์นั้นมากำหนดความสำคัญในการปรับปรุงเว็บไซต์ เพื่อให้เกิดประโยชน์ต่อผู้เข้าใช้บริการมากที่สุด และอาจทำแบบประเมินด้านการให้ข้อมูลของหน่วยงาน เพื่อนำมาปรับปรุงให้เหมาะสมกับความต้องการของผู้ใช้บริการให้มากที่สุด

3.6.4 ผลการวิเคราะห์ด้วยเทคนิคการจัดกลุ่ม สามารถนำไปใช้เป็นแนวทางในการออกแบบหน้าเว็บไซต์ และกลุ่มการจัดวางเมนูของการให้บริการข้อมูลข่าวสารต่างๆ เพื่อให้ตรงกับกลุ่มเป้าหมาย และสามารถนำมาช่วยพิจาราณาในการวางแผนการพัฒนาเว็บไซต์ เช่น เลือกพัฒนาเว็บไซต์กลุ่มที่มีผู้เข้าใช้มากที่สุดเป็นลำดับแรก และเลือกพัฒนาเว็บไซต์ต่างๆ ตามลำดับการเข้าชม

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

การวิจัยเรื่อง “การทำเหมืองข้อมูลสำหรับการพัฒนาเว็บไซต์ กรณีศึกษาเว็บไซต์มหาวิทยาลัยเทคโนโลยีราชมงคลล้านนา ลำปาง” ได้ข้อสรุปผลการวิจัย การอภิปรายผลและข้อเสนอแนะ ดังนี้

1. สรุปผลการวิจัย

สรุปผลการวิจัยแบ่งออกเป็น 3 หัวข้อ ดังนี้ 1) การสร้างคลังข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์ 2) รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป 3) และการทำเหมืองข้อมูล

1.1 การสร้างคลังข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์ ผู้วิจัยได้ออกแบบโครงสร้างของคลังข้อมูลโดยใช้แผนภาพแบบจำลองอี-อาร์ ซึ่งมีการกำหนดความสัมพันธ์ระหว่างเอนทิตี โดยใช้คีย์หลักและคีย์รอง จากนั้นได้กำหนดโครงสร้างคลังข้อมูลในรูปแบบสโนว์เฟลค ซึ่งประกอบด้วย 8 ตาราง แบ่งออกเป็น 2 กลุ่ม ดังนี้

1.1.1 ตารางข้อเท็จจริง ประกอบด้วย

1) ตารางการเข้าใช้งานเว็บไซต์

1.1.2 ตารางเก็บรหัสข้อมูล ประกอบด้วย

1) ตารางข้อมูลล็อกไฟล์
2) ตารางข้อมูลหน่วยงานย่อย
3) ตารางข้อมูลคณะหรือหน่วยงานหลัก
4) ตารางสถานะการรับส่งข้อมูล HTTP
5) ตารางเมธอด
6) ตารางข้อมูลวันเวลา
7) ตารางรายละเอียดผู้ใช้

ในส่วนของอีทีแอลผู้วิจัยได้ออกแบบในส่วนของการดึงข้อมูลจากแหล่งข้อมูล การแปลงข้อมูล การทำความสะอาดข้อมูลและการโหลดข้อมูล โดยการออกแบบเป็นแพ็กเกจในโปรแกรมสำหรับนำเข้าข้อมูลสู่คลังข้อมูล และจากคลังข้อมูลสู่ดาต้ามาร์ท ซึ่งในส่วนของการดึงข้อมูลผู้วิจัยได้ดึงข้อมูลจากฐานข้อมูลที่ต้องการใช้ แล้วเลือกข้อมูลที่ต้องการมาเก็บไว้ในตารางที่ได้ออกแบบไว้ แล้วนำเข้าสู่คลังข้อมูล และทำการศึกษาเพื่อให้เข้าใจกับข้อมูลอีกครั้ง เพื่อทำความสะอาดข้อมูลและกำหนดรูปแบบของข้อมูล เพื่อให้ข้อมูลมีความสอดคล้องกัน และมีความถูกต้องก่อนนำข้อมูลเข้าสู่คลังข้อมูล เมื่อดำเนินการตามขั้นตอนนี้แล้ว จึงออกแบบวิธีการนำเข้าข้อมูล โดยทำการสร้างระบบอีทีแอลด้วยเครื่องมือ Integration Service ของ Microsoft SQL Server 2008

ผู้วิจัยได้ใช้ระบบอีทีแอลที่สร้างนำเข้าข้อมูลจากแหล่งข้อมูลสู่ระบบคลังข้อมูลที่ได้สร้างไว้ เพื่อนำไปใช้ประโยชน์ในการสร้างคิวบ์ และวิเคราะห์ด้วยเทคนิคการทำเหมืองข้อมูล

1.2 รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป ผู้วิจัยได้ใช้เครื่องมือ Microsoft SQL Server 2008 Analysis Services เพื่อใช้ในการสร้าง Analysis Service Project สำหรับวิเคราะห์และสร้างคิวบ์ ซึ่งกำหนดให้ค้นคืนข้อมูลจากตารางข้อมูลในคลังข้อมูลที่ได้ออกแบบไว้ซึ่งเป็นโครงสร้างแบบสโนว์เฟลค ทำให้สามารถสร้างคิวบ์ในลักษณะของฐานข้อมูลหลายมิติ และสามารถเรียกดูข้อมูลการเข้าใช้งานเว็บไซต์ของมหาวิทยาลัยในคุณลักษณะที่ต้องการ

1.3 การทำเหมืองข้อมูลเป็นการได้มาซึ่งองค์ความรู้ที่ซ่อนอยู่ในคลังข้อมูล โดยใช้เครื่องมือ Microsoft SQL Server 2008 Analysis Services ซึ่งอัลกอริทึมที่ใช้ในงานวิจัยนี้มีด้วยกัน 4 อัลกอริทึม ดังนี้

1.3.1 กฎความสัมพันธ์ ในอัลกอริทึมนี้มีข้อมูลการเข้าใช้หน้าเว็บจำนวน 3,534,712 รายการ และมีข้อมูลของผู้ใช้ จำนวน 650,645 รายการ ในการนำมาวิเคราะห์พบว่า มีกฎความสัมพันธ์ที่เกิดขึ้นทั้งหมด 16 กฎ โดยกฎที่มีความความเชื่อมั่นเท่ากับ 1.0 มีทั้งหมด 3 กฎ คือ 1) ถ้าผู้ใช้เข้าหน้าเว็บ /library/music.php และ /library/counter.php แล้วต่อไปจะเข้าหน้าเว็บ /library/lastuse.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 5.65 2) ถ้าผู้ใช้เข้าหน้าเว็บ /library/lastuse.phpและ/library/counter.php แล้วต่อไปจะเข้าหน้าเว็บ /library/music.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 5.65 3) ถ้าผู้ใช้เข้าหน้าเว็บ /library/music.phpและ /library/lastuse.php แล้วต่อไปจะเข้าหน้าเว็บ /library/counter.php ด้วยค่าความเชื่อมั่น 1.0 และค่าความน่าสนใจ 3.36 และเมื่อวิเคราะห์ผลลัพธ์จากกฎความสัมพันธ์การเข้าใช้หน้าเว็บพร้อมกัน จำนวน 2 item พบว่า การเข้าดูหน้าเว็บ /registra/ และ /th/index.php พร้อมกันมีค่าความสัมพันธ์มากที่สุด เท่ากับ 23,438 และเมื่อความสัมพันธ์การเข้าใช้หน้าเว็บพร้อมกันจำนวน 3 item พบว่าการเข้าดูหน้าเว็บ /registra/index.php, /registra/ และ /th/index.php พร้อมกันมีค่าความสัมพันธ์มากที่สุดเท่ากับ 7,017

1.3.2 การพยากรณ์ค่าตัวเลขการเข้าใช้ เป็นการใช้เทคนิคอนุกรมเวลาในการทำนายจำนวนผู้ใช้ที่เข้าใช้เว็บไซต์ของมหาวิทยาลัยในอนาคต ซึ่งจากกราฟแสดงให้เห็นแนวโน้มของการเข้าใช้เว็บไซต์ของทางมหาวิทยาลัยในแต่ละเดือน เส้นกราฟแสดงให้เห็นว่าในช่วงเดือนมกราคมถึงเดือนกุมภาพันธ์ 2556 จะมีปริมาณการเข้าใช้ที่ลดลงและจะเพิ่มขึ้นในระดับเดิมภายในเดือนมีนาคม และจะเห็นว่าแน้วโน้มการเข้าใช้เว็บไซต์ของหน่วยงานที่สังกัดคณะบริหารธุรกิจและศิลปศาสตร์มีแน้วโน้นการเข้าใช้อยู่ในทิศทางที่เพิ่มขึ้นและการเข้าใช้เว็บไซต์ของหน่วยงานที่สังกัดคณะวิทยาศาสตร์และเทคโนโลยีการเกษตร และคณะวิศวกรรมศาสตร์มีแนวโน้มการเข้าใช้ในช่วงระยะเวลาแรกที่ใกล้เคียงกันต่อมาคณะวิทยาศาสตร์และเทคโนโลยีการเกษตร มีแน้วโน้มไปทิศทางที่ลดลง

1.3.3 การวิเคราะห์การจัดกลุ่มโดยใช้ลำดับในอัลกอริทึมนี้มีการสุ่มตัวอย่างมาทั้งหมด 566,660 เรคอร์ดอัลกอริทึม ได้แบ่งคลัสเตอร์ ออกเป็น 66 คลัสเตอร์ คลัสเตอร์ที่ 8 เป็นคลัสเตอร์ที่ใหญ่ที่สุด มีจำนวน 85,082 เรคอร์ด และคลัสเตอร์ที่ 19 เป็นคลัสเตอร์ที่เล็กที่สุด มีจำนวน 24 เรคอร์ดและเมื่อดูสถานะในการเรียกใช้หน้าเว็บ พบว่า1) เมื่อมีการเรียกใช้หน้าเว็บ/

library/laseuse.php แล้ว จะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.52 2) การเรียกใช้หน้าเว็บ ict/administrator/index.php แล้ว จะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.88 3) การเรียกใช้หน้าเว็บ /sar51/administrator/ แล้ว จะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.584) การเรียกใช้หน้าเว็บ /borad/index.php/ แล้ว จะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.575) การเรียกใช้หน้าเว็บ /burmesetemples/administrator/index.php แล้ว จะเรียกใช้หน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.606) การเรียกเข้าใช้หน้าเว็บซ้ำกันในหน้าเว็บ /th/index.php ด้วยความน่าจะเป็น 0.81

1.3.4 การวิเคราะห์การจัดกลุ่ม เป็นอัลกอริทึมที่ใช้ในการจำแนกหรือจัดกลุ่มการใช้จากข้อมูลการเข้าใช้งานเว็บไซต์ซึ่งประชากรหรือ Population ที่นำมาประมวลผลโดยใช้เทคนิค Clustering มีข้อมูลการเข้าใช้เว็บไซต์ที่สุ่มมาวิเคราะห์ทั้งหมด 445,333 เรคอร์ด พบว่า การจัดกลุ่มของผู้เข้าใช้เว็บไซต์แบ่งเป็น 6 คลัสเตอร์ ได้แก่ คลัสเตอร์ที่ 1 มีจำนวนมากที่สุด คือ 163,754 เรคอร์ด คลัสเตอร์ที่ 2 มีจำนวน 70,707 เรคอร์ด คลัสเตอร์ที่ 3 มีจำนวน 59,806 เรคอร์ด คลัสเตอร์ที่ 4 มีจำนวน 57,114 เรคอร์ด คลัสเตอร์ที่ 5 มีจำนวน 48,148 เรคอร์ด คลัสเตอร์ที่ 6 มีจำนวนน้อยที่สุด 45,804 เรคอร์ด และจะเห็นได้ว่าจากกลุ่มตัวอย่างทั้งหมดจำนวนหน้าเว็บที่เข้าใช้มากที่สุด 3 ลำดับแรก คือ หน้าหลักจำนวน 163,754 เรคอร์ด, สาขาวิศวกรรมไฟฟ้าและคอมพิวเตอร์จำนวน 70,707 เรคอร์ด, บอร์ดจำนวน 29,631 เรคอร์ด

2. การอภิปรายผล

2.1 การสร้างคลังข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์

การพัฒนาคลังข้อมูลโดยใช้ ซอฟต์แวร์ Microsoft SQL Server 2008 สามารถพัฒนาคลังข้อมูลโดยใช้กระบวนการอีทีแอลได้ทุกขั้นตอน เริ่มตั้งแต่ขั้นตอนคัดแยกข้อมูล หรือการดึงข้อมูลจากแหล่งข้อมูลอื่นตามความต้องการเข้าสู่ที่พักข้อมูล เพื่อดำเนินการทำความสะอาดข้อมูลและปรับรูปแบบของข้อมูลเพื่อให้เกิดความสอดคล้องกับข้อมูลที่ต้องการนำไปใช้ และกำจัดข้อผิดพลาดของข้อมูลทำให้ข้อมูลมีความถูกต้องที่สุดก่อนจะนำข้อมูลโหลดเข้าสู่คลังข้อมูลในกระบวนการสุดท้าย โดยใช้เครื่องมือ Database Engine และ SQL Server Integration Services และยังสามารถพัฒนาแอพพลิเคชันสำหรับให้ผู้ใช้แต่ละกลุ่มในการค้นคืนข้อมูลหรือหาความรู้จากคลังข้อมูล ทั้งที่เป็นการประมวลผลเชิงวิเคราะห์แบบออนไลน์หรือโอแลป และการสร้างรายงานต่างๆ โดยใช้ SQL Server Analysis Services และ SQL Server Reporting Services

การพัฒนาคลังข้อมูล มีประโยชน์ในการจัดเก็บข้อมูลจำนวนมากๆ ซึ่งอยู่ในลักษณะฐานข้อมูลเชิงสัมพันธ์ เพื่อใช้ในการสนับสนุนการตัดสินใจ ซึ่งสอดคล้องกับผลการวิจัยเรื่อง “ระบบคลังข้อมูลจาก Log File ของการใช้อินเทอร์เน็ต” ที่ดำเนินการโดย พันธ์รัตน์ อักษรศรีกุล และศิฬาณี นุชิตประสิทธิ์ชัย ได้กล่าวไว้ว่า 1) การนำแนวความคิดในการนำเทคโนโลยีสารสนเทศระบบฐานข้อมูลและระบบอินทราเน็ตขององค์กรมาประยุกต์ใช้ให้เกิดประโยชน์ ซึ่งถือเป็นความก้าวหน้าอีกขั้นหนึ่ง

ของการมีระบบจัดเก็บข้อมูลที่ดี 2) สามารถนำผลลัพธ์ที่ได้ไปใช้ในการวิเคราะห์และ 3) สนับสนุนการตัดสินใจเกี่ยวกับการวางแผนกลยุทธ์ด้านการพัฒนาระบบสารสนเทศ

2.2 รายงานจากการประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลป

ผู้วิจัยใช้เครื่องมือ SQL Server Analysis Service เพื่อสร้าง AnalysisService Project สำหรับวิเคราะห์การประมวลผลข้อมูลเชิงวิเคราะห์หรือโอแลปและกำหนดข้อมูลจากตารางข้อมูลที่อยู่ในคลังข้อมูลที่ทำการออกแบบตามโครงสร้างคลังข้อมูลและผู้ใช้สามารถเรียกดูข้อมูลการเข้าใช้งานเว็บไซต์ในมุมมองต่างๆ ตามที่ต้องการ ดังนั้นโอแลปจึงเป็นเทคโนโลยีที่ประกอบด้วยเครื่องมือที่ช่วยค้นคืน และช่วยในการนำเสนอข้อมูลเชิงมิติได้ในหลากหลายมิติและในมุมมองต่างๆ โดยที่โอแลปได้ถูกออกแบบมาช่วยในการตัดสินใจของผู้ใช้ในระดับผู้บริหารที่ต้องการนำผลการวิเคราะห์ข้อมูลเพื่อใช้ในการประกอบการตัดสินใจในระดับสูงและเพื่อช่วยในการคาดคะเนหรือการทำนายผลที่จะตามมาในการประกอบกิจการทางธุรกิจ หรือการดำเนินกิจกรรมต่างๆ ของหน่วยงาน และง่ายในการจัดทำรูปแบบรายงาน เช่น รายงานสรุปการเข้าชมหน้าเว็บต่างๆ จำแนกตามหน่วยงานในแต่ละไตรมาส

2.3 การทำเหมืองข้อมูล

การทำเหมืองข้อมูลเป็นการวิเคราะห์ข้อมูลปริมาณมหาศาลเพื่อค้นหาความรู้ใหม่ๆ หรือแพทเทิร์นความสัมพันธ์ของของข้อมูลที่หน่วยงานไม่เคยทราบมาก่อน ซึ่งจากงานวิจัยนี้ได้นำอัลกอริทึมต่างๆ ของการทำเหมืองข้อมูลมาวิเคราะห์ข้อมูลจากล็อกไฟล์การเข้าใช้งานเว็บไซต์ ทำให้ได้ทราบความรู้ใหม่ๆ เกี่ยวกับการใช้เว็บไซต์ คือ จำแนกกลุ่มการใช้เว็บเพจ สามารถพยากรณ์จำนวนผู้เข้าใช้ในอนาคต ทราบลำดับการใช้งานเว็บไซต์โดยการเรียงลำดับการเข้าถึงเว็บเพจแต่ละหน้า และทราบถึงความสัมพันธ์ของการเข้าใช้เว็บเพจต่างๆ ซึ่งสอดคล้องกับงานของอดุลย์ ยิ้มงาม (2553) กล่าวว่า “การทำเหมืองข้อมูลเว็บ คือการใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาองค์ความรู้ และสกัดข้อมูลสารสนเทศจากเอกสารบนหน้าเว็บ และการให้บริการเว็บไซต์โดยอัตโนมัติ เพื่อนำองค์ความรู้ที่สกัดมาใช้ประโยชน์ หรือแก้ปัญหาทั้งทางตรงและทางอ้อม” และสอดล้องกับงานวิจัยของพิจิตรา จอมศรี (2549) ได้ศึกษาการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร และงานวิจัยนี้ได้ใช้อัลกอริทึมการวิเคราะห์ข้อมูลประกอบด้วย 4 อัลกอริทึม คือ 1) คลัสเตอริง 2) ไทม์ซีรีส์ 3) แอสโซซิเอชันรูลส์ และ 4) ซีเควันซ์ คลัสเตอริง

(1) คลัสเตอริง เป็นการจัดการข้อมูลของการเข้าใช้เว็บไซต์ โดยวิธีการจัดกลุ่มการเข้าใช้ในหน้าเว็บของผู้ใช้ในแต่ละกลุ่ม เพื่อให้ทราบถึงจำนวนและการเข้าใช้เว็บไซต์ของผู้เยี่ยมชมในแต่ละกลุ่ม และนำผลที่ได้นำไปพัฒนาปรับปรุงให้เว็บไซต์เพื่อให้สอดคล้องกับความต้องการของผู้ใช้ในแต่ละกลุ่ม ซึ่งงานวิจัยได้สอดคล้องกับ Karuna P. Joshi และคณะ (2003) ได้ทำวิจัยเรื่อง On Using a Warehouse to Analyze Web Logs ซึ่งได้ศึกษาการสร้างคลังข้อมูล เพื่อใช้ในการวิเคราะห์ข้อมูลจากเว็บล็อกของเว็บไซต์ ด้วยการทำเหมืองข้อมูลเว็บ พบว่าการใช้อัลกอริทึมการจัดกลุ่ม สามารถนำไปจัดกลุ่มการเข้าใช้งานเว็บไซต์ได้ เพื่อนำมาใช้ในการพัฒนาปรับปรุงเว็บไซต์ให้เหมาะสมกับความต้องการของผู้ใช้

(2) ไทม์ซีรีส์ หรืออนุกรมเวลาเป็นการพยากรณ์ค่าตัวเลขการในการเข้าใช้งานเว็บไซต์ของมหาวิทยาลัย โดยพยาการณ์การเข้าใช้เว็บไซต์ในแต่ละเดือน ทำให้เห็นแนวโน้นของการเข้าใช้

เว็บไซต์ตามหมวดหมู่หรือหน่วยงานต่างๆ และรวมถึงภาพรวมทั้งหมดในการเข้าใช้ตลอกทั้งปี และยังสามารถเปรียบเทียบการเข้าใช้ในแต่ละหมวดได้ ซึ่งได้สอดคล้องกับงานวิจัยของ พิจิตรา จอมศรี (2549) ซึ่งได้ศึกษาการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปกร เพื่อพัฒนาโมเดลเพื่อใช้ในการทำนายแนวโน้มการใช้งานเว็บในอนาคต ซึ่งผลการศึกษาดังกล่าวสามารถนำแบบจำลองมาประยุกต์ใช้ในการทำนายแนวโน้มการใช้งานเว็บในอนาคตได้ และสามารถเพิ่มประสิทธิภาพการทำงานของระบบพร็อกซีเซิร์ฟเวอร์ได้ และทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและลดปริมาณข้อมูลในระบบเครือข่ายได้

(3) แอสโซซิเอชันรูลส์ เป็นการจัดการข้อมูลเพื่อวิเคราะห์ข้อมูลการเข้าใช้งานเว็บไซต์ หรือค้นหารูปแบบความสัมพันธ์ของพฤติกรรมเข้าใช้เว็บไซต์ของผู้ใช้ เพื่อหาความน่าจะเป็นและกฎความสัมพันธ์ในการเข้าใช้หน้าเว็บของหน่วยงานภายในมหาวิทยาลัย ซึ่งสอดคล้องกับงานวิจัยของ Houqun Yang และคณะ (2007) ซึ่งได้ศึกษาวิจัยเรื่อง An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining โดยเสนอวิธีการจัดกลุ่มเพื่อวิเคราะห์โครงสร้างของเว็บไซต์ด้วยเทคนิคการทำเหมืองข้อมูลเว็บ ด้วยอัลกอริทึมแอสโซซิเอชันรูลส์ เพื่อนำความรู้ที่ได้ไปปรับปรุงโครงสร้างของเว็บไซต์เพื่อให้เหมาะสมกับการให้บริการ และสอดคล้องกับ AzizulAzhar bin Ramli (2005) ที่ได้ศึกษาเรื่อง Web Usage Mining Using Apriori Algorithm: UUM Learing Care Portal Case โดยเสนอการทำเหมืองข้อมูลโดยใช้กฎความสัมพันธ์วิธี Apriori เพื่อการสร้างหรือปรับปรุงรูปแบบเว็บไซต์ตามพฤติกรรมของผู้ใช้

(4) ซีเควนซ์ คลัสเตอริง เป็นการวิเคราะห์การจัดกลุ่มโดยใช้ลำดับ เพื่อศึกษาการเรียงลำดับในการเข้าใช้หน้าเว็บของผู้ใช้ในแต่ละกลุ่ม เพื่อนำผลที่ได้นำวิเคราะห์ศึกษาพฤติกรรมของการเชื่อมโยงหน้าเว็บในแต่ละกลุ่มผู้ใช้ ซึ่งงานวิจัยได้สอดคล้องกับงานวิจัยของ Chaofeng Li และคณะ (2007) ได้ศึกษาวิจัยเรื่อง Similarity Measurement of Web Sessions by Sequence Alignment โดยมีวัตถุประสงค์เพื่อศึกษาการจัดกลุ่มผู้เข้าใช้เว็บไซต์ที่คล้ายคลึงกันด้วยวิธีการลำดับการจัดเรียง

3. ข้อเสนอแนะ

3.1 ข้อเสนอแนะในการนำผลการวิจัยไปใช้

3.1.1 ผู้ที่เกี่ยวข้องในการจัดเก็บข้อมูลภายในองค์กร การจัดเก็บข้อมูลต่างๆ ที่มีในมหาวิทยาลัย ซึ่งมีปริมาณข้อมูลที่มาก ควรเก็บอยู่ในรูปแบบของฐานข้อมูลซึ่งสามารถลดการเก็บข้อมูลที่ซ้ำซ้อน และยังสามารถดูแลข้อมูลทำได้อย่างสะดวก เพื่อนำไปใช้ประโยชน์ให้มากที่สุด และควรนำข้อมูลมาสร้างคลังข้อมูลภายในองค์กร เพื่อใช้ในการประกอบการวิเคราะห์ เพื่อประโยชน์ในการบริหารการตัดสินใจของผู้บริหารได้อย่างมีประสิทธิภาพมากที่สุด

3.1.2 การจัดทำรายงานในรูปแบบหลายมิติ ซึ่งเป็นรายงานที่ผู้บริหารสามารถนำข้อมูลที่ได้มาประกอบการตัดสินใจและยังสามารถแสดงผลในรูปแบบเวลาจริง ซึ่งมีประโยชน์อย่างมากในการบริหารงาน ซึ่งสามารถรับรู้ข้อมูลข่าวสารได้ทันท่วงที และยังสามารถประยุกต์ใช้ในแผนก

อื่นๆ ที่มีความสำคัญ อาทิ เช่น แผนกทะเบียน แผนกการเงิน แผนกบุคลากร เป็นต้น เพื่อให้การบริหารงานเป็นไปอย่างมีประสิทธิภาพมากที่สุด

3.1.3 การศึกษาวิจัยครั้งนี้เป็นการนำเสนอแนวทางการประยุกต์ใช้แบบจำลองที่ได้จากการทำเหมืองข้อมูลดังกล่าว แต่ยังไม่ได้นำแบบจำลองไปใช้งานจริง หากนำไปใช้งานจริงควรนำผลการวิจัยไปประเมินอีกครั้งเพื่อนำผลการประเมินมาปรับปรุงแก้ไขหรือพัฒนาแนวทางการทำเหมืองข้อมูล และทำให้การประยุกต์ใช้ให้มีประสิทธิภาพถูกต้องแม่นยำต่อไป

3.2 ข้อเสนอแนะในการวิจัยครั้งต่อไป

3.2.1 การวิจัยครั้งนี้เป็นการศึกษาการข้อมูลเข้าใช้เว็บไซต์ของมหาวิทยาลัยเท่านั้น โดยใช้ข้อมูลจากล็อกไฟล์ของการเข้าใช้เว็บไซต์ของทางมหาวิทยาลัย ซึ่งไม่สามารถระบุคุณลักษณะต่างๆ ของผู้ใช้ได้ และเพื่อให้เกิดประโยชน์สูงสุดในการนำข้อมูลมาวิเคราะห์พฤติกรรมผู้เข้าใช้ในประเภทต่างๆ อาทิเช่น อาจารย์ นักศึกษา ซึ่งเป็นกลุ่มเป้าหมายหลักของการให้บริการข้อมูลข่าวสาร ควรสร้างระบบฐานข้อมูลจัดเก็บการเข้าใช้งานของผู้ใช้กลุ่มต่างๆ ซึ่งจะเป็นประโยชน์ในการนำข้อมูลมาใช้ในการวิเคราะห์ และสามารถตอบสนองความต้องการของกลุ่มเป้าหมายในการนำผลการศึกษามาพัฒนาออกแบบเว็บไซต์ของทางมหาวิทยาลัยให้ตรงตามพฤติกรรมของกลุ่มผู้เข้าใช้ในแต่ละกลุ่ม

3.2.2 การศึกษาวิจัยควรนำอัลกอริทึมอื่นๆ มาประยุกต์ใช้ เพื่อให้ได้องค์ความรู้ใหม่ๆ ที่มีประโยชน์ และมีประสิทธิภาพมากที่สุด และนำผลการวิจัยที่ได้ไปพัฒนาปรับปรุงแก้ไข หรือประยุกต์ใช้ในรูปแบบต่างๆ ตามความเหมาะสม

บรรณานุกรม

บรรณานุกรม

กิตติพงษ์ กลมกล่อม .(2552). การออกแบบและพัฒนาคลังข้อมูล = *Data warehouse*. กรุงเทพมหานคร: เคทีพีคอมพ์แอนด์คอนซัลท์.

ดาวพระศุกร์ ฤทธิบัณฑิตย์. (2554). “ระบบสนุบสนุนการวิเคราะห์ข้อมูลจาราจรคอมพิวเตอร์” สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศคณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

พิจิตราจอมศรี .(2549).“ทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปกร” กรุงเทพมหานคร: มหาวิทยาลัยศิลปากร.

พันธ์รัตน์ อักษรศรีกุล และศิษาณี นุชิตประสิทธิ์ชัย. (2552). “ระบบคลังข้อมูลจาก Log File ของการใช้อินเทอร์เน็ต”ภาควิชาเทคโนโลยีสารสนเทศคณะเทคโนโลยีสารสนเทศมหาวิทยาลยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ราชกิจจานุเบกษา. พระราชบัญญัติ วาด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ พ.ศ. 2550 เล่ม 124 ตอนที่ 1 27 ก หน้า 11, 18 มิถุนายน 2550.

เลิศ เลิศศิริโสภณ. (2541). “ถึงเวลาของดาต้าแวร์เฮ้าส์แล้วหรือยัง” BCM Magazine, vol9, no.115, หน้า 9 กันยายน 2541.

วิภา เจริญภัณฑารักษ์. (2555). “หลักการพื้นฐานของการทำเหมืองข้อมูล” ใน *ประมวลสาระชุดวิชาคลังข้อมูล เหมืองข้อมูล และธุรกิจอัจฉริยะ=Data warehouse data mining and business intelligence* หน่วยที่ 8 หน้า 1-33 นนทบุรี มหาวิทยาลัยสุโขทยธรรมาธิราช บัณฑิตศึกษา สาขาวิชาวิทยาศาสตร์และเทคโนโลยี.

วิกิพีเดีย. (2556). “การทำเหมืองข้อมูลออนไลน์” ค้นคืนวันที่ 15 พฤษภาคม 2557 จาก http://th.wikipedia.org/wiki/การทำเหมืองข้อมูล.

สำรวย กมลายุตต์ และคณะ. (2551). “การพัฒนาต้นแบบคลังข้อมูลเพื่อวิเคราะห์คุณลักษณะของนักศึกษาระดับปริญญาตรีที่ออกกลางคันของมหาวิทยาลัยสุโขทัยธรรมาธิราช” มหาวิทยาลัยสุโขทัยธรรมาธิราช.

สุวรรณี อัศวกุลชัย. (2555). “หลักการพื้นฐานของคลังข้อมูล” ใน *ประมวลสาระชุดวิชาคลังข้อมูล เหมืองข้อมูล และธุรกิจอัจฉริยะ=Data warehouse data mining and business intelligence* หน่วย 1-30 หน้าที่ 1 นนทบุรี มหาวิทยาลัยสุโขทัยธรรมาธิราช บัณฑิตศึกษา สาขาวิชาวิทยาศาสตร์และเทคโนโลยี.

อนงค์ หลอดแก้ว. (2557). “ส่วนประกอบของเว็บเพ็จสาระสังเขปออนไลน์” ค้นคืนวันที่ 9 พฤษภาคม 2557 จาก https://sites.google.com/site/class0223/components.

อดุลย์ ยิ้มงาม. (2553). “การทำเหมืองข้อมูล” ออนไลน์ ค้นคืนวันที่ 15 พฤษภาคม 2557 จาก http://compcenter.bu.ac.th/index.php?option=com_content&task=view&id=75&Itemid=172.

AzizulAzhar bin Ramli. (2005). WEB USAGE MINING USING APRIORI ALGORITHM: UUM LEARNING CARE PORTAL CASE. International Conference on Knowledge Management, Malaysia, 2005, pp.1-19.

Houqun Yang,Jingsheng Lei,FaFu. (2007). An Approach of Multi-path Segmentation Clustering Based on Web Usage Mining, Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on, pp.644 - 648

Chaofeng Li. (2007). Similarity Measurement of Web Sessions by Sequence Alignment, Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference on,2007, 18-21 Sept. 2007 , pp.716 – 720.

L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai. (2011). ANALYSIS OF WEB LOGS AND WEB USER IN WEB MINING, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.1, January 2011.

PriyankaPatil and UjwalaPatil. (2012). Preprocessing of web server log file for web mining World Journal of Science and Technology 2012, 2(3):14-18 ISSN: 2231 – 2587 Available Online: www.worldjournalofscience.com.

KARUNA P. JOSHI. (2013). On Using a Warehouse to Analyze Web Logs ,Distributed and Parallel Databases, 13, 161–180, 2003.

https://www.microsoft.com/technet/prodtechnol/WindowsServer2003/Library/IIS/ffdd7079-47be-4277-921f-7a3a6e610dcb.mspx?mfr=true Retrieved May 20, 2014

http://winintro.ru/mail.en/html/3581adb1-c526-4169-b2d8-1d46c1611c34.htm Retrieved May 20, 2014.

http://msdn.microsoft.com/en-us/library/ms525807(v=vs.90).aspx Retrieved May 20, 2014

http://siripornk.blogspot.com/2010/08/data - mining.html Retrieved May 20, 2014

http://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf Retrieved May 20, 2014.

http://e.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf Retrieved May 20, 2014.

ประวัติผู้วิจัย

ชื่อ	นายอภิชาติ ปัญญา
วัน เดือน ปี เกิด	4 ธันวาคม 2514
สถานที่เกิด	อำเภอสูงเม่น จังหวัดแพร่
ประวัติการศึกษา	เศรษฐศาสตรบัณฑิต มหาวิทยาลัยรามคำแหง พ.ศ. 2539
สถานที่ทำงาน	ที่ทำการองค์การบริหารส่วนตำบลวังดิน อำเภอเมืองอุตรดิตถ์ จังหวัดอุตรดิตถ์
ตำแหน่ง	นักบริหารงานทั่วไป (หัวหน้าสำนักปลัด)