

บทที่ 2

แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องที่ผู้วิจัยได้นำมาศึกษาตลอดจนประยุกต์ใช้ในการทำวิจัยในครั้งนี้ ซึ่งผู้วิจัยได้นำเสนอตามลำดับ ดังนี้

1. หลักการเว็บเบสเทคโนโลยี (Web Based Technology)
2. หลักการการจัดการระบบฐานข้อมูล (Database System)
3. หลักการเหมืองข้อมูล (Data Mining)
4. หลักสถิติและความน่าจะเป็น (Statistical Model and Probability)
5. หลักวิศวกรรมซอฟต์แวร์ (Software Engineering)
6. หลักการออนโทโลยีเชิงเวลา (Temporal Ontology)
7. การสร้างอัลกอริทึมในภาษาสออบตามเชิงโครงสร้าง
8. หลักสถิติในการวิเคราะห์ข้อมูลการวิจัย
9. งานวิจัยที่เกี่ยวข้อง

1. หลักการเว็บเบสเทคโนโลยี (Web Based Technology)

เว็บเบสเทคโนโลยี (Web Based Technology) ที่เกี่ยวข้องในงานวิจัยนี้ คือ

1.1 PHP (PHP Hyper Text Preprocessor)

PHP เป็นภาษาสคริปต์ที่ประมวลผลที่ฝั่งแม่ข่าย (Server) แล้วส่งผลลัพธ์ไปแสดงผลที่ฝั่งเครื่องลูกข่าย (Client) ผ่านบราวเซอร์เช่นเดียวกับ CGI (Common Gateway Interface) และ ASP (Active Server Page) ต่อมาเมื่อมีผู้ใช้งานมากขึ้นจึงมีการร้องขอให้มีการพัฒนาประสิทธิภาพของ PHP/FI ให้สูงขึ้น

PHP มีคุณสมบัติเด่น คือ สนับสนุนระบบปฏิบัติการทั้ง Window 95/98/ME/NT, Linux และเว็บเซิร์ฟเวอร์อย่าง IIS, PWS, Apache, OmniHTTPd สนับสนุนการรองรับระบบฐานข้อมูลได้หลายรูปแบบ เช่น SQL Server, MySQL, Ingres, mSQL, Oracle, Informix, ODBC เป็นต้น

ความสามารถของ PHP ด้านความสามารถพื้นฐานที่ภาษาสคริปต์ทั่ว ๆ ไปมีนั้น PHP มีความสามารถทำได้ทัดเทียมเช่นเดียวกัน เช่น การรับข้อมูลจากฟอร์ม, การสร้าง Content ในลักษณะ Dynamic, รับส่ง Cookies, สร้าง, เปิด, อ่าน และปิดไฟล์ในระบบ

1.1.1 ข้อดีของ PHP คือ (ไพศาล, 2543) เป็นภาษาแบบ Open Source ซึ่งมีข้อดีดังนี้

1.1.1.1 ไม่มีค่าลิขสิทธิ์

1.1.1.2 Open Source Software หลายตัวมีคุณภาพสูงและได้รับ

ความนิยม จนกลายเป็นมาตรฐานหรือมีอิทธิพลต่อการกำหนดมาตรฐาน

1.1.1.3 เรียนรู้และดัดแปลงได้ด้วยตัวเอง แก้ไขปัญหาได้ทันที และมีกลุ่มผู้ใช้ร่วมกันช่วยกันแก้ปัญหาโดยสามารถซักถามปัญหาผ่านทางอินเทอร์เน็ต ในกรณีที่ตัวเองไม่สามารถแก้ปัญหาเองได้

1.1.2 เป็นโปรแกรมที่สามารถ Download มาใช้งานได้ฟรี ทั้งตัว Source Code และเอกสารอ้างอิงมาใช้โดยไม่ผิดกฎหมาย

1.1.3 สามารถรองรับการทำงานได้หลาย Platform เช่น Windows 95 / me / NT / 2000, Unix, Linux

1.1.4 PHP สนับสนุนการติดต่อและจัดการกับฐานข้อมูลได้หลายประเภท เช่น MySQL, Oracle และ SQL Server โดยมีการเรียกใช้ผ่านตัวฐานข้อมูลโดยตรง

1.1.5 PHP เป็นภาษาที่มีความยืดหยุ่นสูง โดยสามารถแทรกในตำแหน่งใดก็ได้ในแท็กของ HTML โดยเฉพาะความสามารถในการรับส่งข้อมูลจากฟอร์มของ HTML

1.1.6 PHP มีประสิทธิภาพการทำงานสูง ทำงานได้รวดเร็ว

ซึ่งความสามารถในการจัดการกับตัวแปรหลาย ๆ ประเภท เช่น เลขจำนวนเต็ม (Integer) เลขทศนิยม (Float) สตริง (String) และอาร์เรย์ (Array) เป็นต้น และสามารถรองรับการเขียนโปรแกรมเชิงวัตถุ (Object Oriented Programming) และความสามารถในการสร้างกราฟิกและยังมีไลบรารีสำหรับติดต่อกับแอปพลิเคชันได้มากมาย มีความยืดหยุ่นสูงสามารถนำไปสร้างแอปพลิเคชันได้หลากหลาย และอีกข้อดีหนึ่งที่โดดเด่นคือก็สามารถแทรกลงในแท็ก HTML ในตำแหน่งใดก็ได้

2. หลักการการจัดการระบบฐานข้อมูล (Database System)

การจัดเก็บข้อมูลด้วยระบบคอมพิวเตอร์ เช่น ข้อมูลทั่วไปของบุคลากรของมหาวิทยาลัย หรือ ผู้ใช้อินเตอร์เน็ต ข้อมูลการปฏิบัติงานต่าง ๆ แม้ว่าจะจัดเก็บไฟล์ไว้อย่างเป็นระเบียบและแบ่งเป็นหมวดไว้อย่างชัดเจนเป็นฐานข้อมูล (Database) แต่เมื่อข้อมูลมีจำนวนมากขึ้นก็อาจเกิดปัญหาขึ้นได้ เช่น ความยุ่งยากในการค้นหาข้อมูลที่ต้องการเก็บข้อมูลที่ซ้ำซ้อนกัน การเก็บข้อมูลที่ขัดแย้งกัน ซึ่งปัญหาเหล่านี้สามารถแก้ไขได้โดยการใช้โปรแกรมที่เป็นระบบจัดการฐานข้อมูล (Database Management System)

2.1 การจัดองค์กรข้อมูล

ในการเก็บรวบรวมข้อมูลด้วยระบบคอมพิวเตอร์ข้อมูลจะจัดเก็บอยู่จำนวนมาก ซึ่งมีการจัดองค์กรข้อมูล (Data Organization) จากส่วนย่อยไปหาส่วนใหญ่สุด ซึ่งแต่ละส่วนมีรายละเอียด ดังต่อไปนี้

2.1.1 ฟیلด์หรือเขตข้อมูล (Field) คือ อักขระหรือกลุ่มของอักขระ (ตัวอักษร ตัวเลข ตัวอักษรพิเศษ) มารวมกันและก่อให้เกิดความหมาย

2.1.2 เรคอร์ดหรือระเบียบข้อมูล (Record) คือ ชุดของข้อมูลตั้งแต่ 1 ฟیلด์ที่มีความสัมพันธ์กัน เมื่อนำมารวมกันแล้วสามารถบ่งบอกได้ว่าเป็นรายละเอียดของสิ่งใด

2.1.3 ไฟล์หรือแฟ้มข้อมูล (File) คือ กลุ่มข้อมูลตั้งแต่ 1 เรคอร์ดที่มีโครงสร้างเหมือนกันสามารถจำแนกตามหน้าที่การใช้งานได้ ดังนี้

2.1.3.1 แฟ้มข้อมูลหลัก (Master File)

2.1.3.2 แฟ้มข้อมูลรายการปรับปรุง (Transaction File)

2.1.3.3 แฟ้มข้อมูลอ้างอิง (Reference File)

2.1.3.4 แฟ้มรายงาน (Report File)

2.1.3.5 แฟ้มอื่น ๆ เช่น แฟ้มรหัสผ่าน เป็นต้น

2.1.4 ฐานข้อมูล (Database) คือส่วนที่เก็บรวบรวมไฟล์ที่เกี่ยวข้องกับระบบและมีความสัมพันธ์กัน เพื่อลดความซ้ำซ้อนของข้อมูล สะดวกในการใช้ข้อมูลร่วมกันระหว่างหลาย ๆ งานและมีความสัมพันธ์เป็นอิสระต่อกันระหว่างข้อมูลกับงานที่เรียกใช้

ข้อมูลในฐานข้อมูลโดยทั่วไปจะถูกสร้างให้มีโครงสร้างที่ถ่ายทอดความเข้าใจและการใช้โดยทั่วไปแล้ว ฐานข้อมูลที่มีใช้อยู่ในปัจจุบันจะมีโครงสร้าง 3 แบบด้วยกัน คือ ฐานข้อมูลแบบลำดับขั้น (Hierarchical Database) ฐานข้อมูลแบบเครือข่าย (Network Database) และ ฐานข้อมูลแบบเชิงสัมพันธ์ (Relational Database) (เทคโนโลยีฐานข้อมูล, ม.ป.ป., เว็บไซต์)

1. ฐานข้อมูลแบบลำดับชั้น (Hierarchical Database)

เป็นลักษณะของฐานข้อมูลที่มีความสัมพันธ์ของข้อมูลเป็นหนึ่งต่อหนึ่ง หรือ หนึ่งต่อกลุ่มแต่จะไม่มีความสัมพันธ์แบบกลุ่มต่อกลุ่มในฐานข้อมูลแบบนี้

ลักษณะโครงสร้างของฐานข้อมูลแบบลำดับชั้นนี้ จะมีลักษณะคล้ายต้นไม้ที่คว่ำ หัวลงจึงอาจเรียกโครงสร้างข้อมูลนี้ได้อีกแบบว่าเป็น "โครงสร้างแบบต้นไม้" โดยจะมีระเบียบพ่อแม่ ระเบียบในแถวถัดลงมาจะเรียกว่า ระเบียบลูก ซึ่งระเบียบพ่อแม่จะสามารถมีระเบียบเพียงหนึ่งระเบียบเท่านั้น ตัวอย่างของบางข้อมูลแบบนี้ เช่น การขายสินค้าของพนักงานให้แก่ลูกค้าแต่ละคนจะพบว่าพนักงานขายแต่ละคน จะมีลูกค้าได้หลายคน และลูกค้าก็ได้หลายอย่างขึ้นไป จะสามารถแสดงความสัมพันธ์ของระเบียบของพนักงานขายและระเบียบลูกค้า และความสัมพันธ์ระหว่างระเบียบลูกค้าและระเบียบสินค้าของฐานข้อมูลแบบลำดับชั้น ดังต่อไปนี้

2. ฐานข้อมูลแบบเครือข่าย (Network Database)

ข้อมูลภายในฐานข้อมูลแบบนี้สามารถมีความสัมพันธ์กันแบบใดก็ได้ เช่น อาจเป็นแบบหนึ่งต่อกลุ่ม หรือกลุ่มต่อกลุ่ม ตัวอย่างของฐานข้อมูลแบบนี้ เช่น การสั่งซื้อสินค้าจากร้านผู้ผลิตสินค้า และการนำสินค้าไปเก็บในคลังสินค้า ซึ่งจะแสดงความสัมพันธ์ระหว่างระเบียบร้านผู้ผลิตสินค้าและระเบียบลูกค้า และความสัมพันธ์ระหว่างระเบียบสินค้าและระเบียบที่เก็บสินค้า

3. ฐานข้อมูลเชิงสัมพันธ์ (Relational Database)

ข้อมูลเชิงสัมพันธ์เป็นฐานข้อมูลที่มีความนิยมใช้กันมากในปัจจุบัน ซึ่งสามารถใช้งานได้กับเครื่องคอมพิวเตอร์ทุกระดับตั้งแต่ไมโครคอมพิวเตอร์ จนกระทั่งถึงเมนเฟรม - คอมพิวเตอร์ ฐานข้อมูลแบบนี้จะมีโครงสร้างข้อมูลต่างจากฐานข้อมูลสองแบบแรก กล่าวคือ ข้อมูลจะถูกเก็บอยู่ในรูปแบบของตาราง (Table) ภายในตารางก็จะแบ่งเป็นแถว (Row) และคอลัมน์ (Column) แต่ละตารางจะมีจำนวนแถว ได้หลายแถว และจำนวนคอลัมน์ ได้หลายคอลัมน์ แถวแต่ละแถวจะสามารถเรียกได้อีกอย่างว่าระเบียบ (Record) คอลัมน์แต่ละคอลัมน์สามารถเรียกได้อีกอย่างหนึ่งว่าเขตข้อมูลหรือฟิลด์ (Field) นอกจากนี้ตารางแต่ละตารางยังสามารถเรียกได้อีกอย่างว่ารีเลชัน (Relation) แถวแต่ละแถวภายในตารางยังอาจเรียกว่า ทัพเพิล (Tuple) และคอลัมน์แต่ละคอลัมน์อาจจะเรียกว่า แอททริบิวต์ (Attribute)

2.2 วิธีการออกแบบฐานข้อมูล

วิธีการออกแบบฐานข้อมูลมี 2 แบบด้วยกันคือ

1. การออกแบบจากล่างขึ้นบน (Bottom-up Database Design) เป็นวิธีการนำเอาระบบงานเดิมที่มีอยู่แล้วมารวบรวมกันเข้าเป็นระบบงานใหม่ที่สมบูรณ์กว่าเดิม ข้อมูลและโปรแกรมเดิมที่นำมารวบรวมกันเข้านั้นเป็นข้อมูลและโปรแกรมที่ดีของระบบงานแต่ละส่วน การรวบรวมงานเดิมเหล่านี้เข้าด้วยกันเป็นงานที่ยุ่งยากมากพอสมควร และเสียเวลามาก ในการที่จะออกแบบระบบและสร้างระบบฐานข้อมูลที่สมบูรณ์ได้

2. การออกแบบจากบนลงล่าง (Top-down Database Design) เป็นวิธีการที่นิยมใช้กันในการออกแบบระบบ มีขั้นตอนคือ เลือกเอาผู้ที่เข้าใจระบบที่สุด อาจจะเป็นหนึ่งคนหรือหลายคนก็ได้มาศึกษาถึงความต้องการขององค์กร แล้วจึงนำข้อมูลที่ได้นั้นมาออกแบบเป็นโครงสร้างทั้งหมดของระบบฐานข้อมูล เนื่องจากมีความหลากหลายของข้อมูลแต่ละฝ่าย ข้อเสียของการออกแบบวิธีนี้ คือ ผู้ออกแบบจำเป็นต้องเป็นผู้ที่ศึกษาและเข้าใจระบบจริง ๆ จึงจะสามารถออกแบบระบบฐานข้อมูลได้อย่างครบถ้วน

2.3 หลักการพิจารณาการออกแบบฐานข้อมูล

หลักการพิจารณาการออกแบบฐานข้อมูล หมายถึง การเลือกรูปแบบข้อมูลที่เกี่ยวข้องในฐานข้อมูล เช่น การกำหนดเอนทิตี (Entity) การกำหนดแอททริบิวต์ (Attributes) การกำหนดคีย์หลัก (Primary Key) คีย์รอง (Secondary Key) หรือการกำหนดคีย์นอก (Foreign Key) รวมทั้งการกำหนดความสัมพันธ์ระหว่างเอนทิตีโดยสามารถอธิบายได้ ดังนี้

1. การกำหนดเอนทิตีที่เกี่ยวข้อง หมายถึง การสร้างเพิ่มข้อมูลที่เกี่ยวข้องในฐานข้อมูลว่าจะทำการสร้างตารางข้อมูลอะไรบ้างในฐานข้อมูลนี้เพื่อประโยชน์สูงสุดในการประมวลผลข้อมูลในระบบฐานข้อมูล

2. การกำหนดชื่อเอนทิตีต่าง ๆ นั้นจะต้องมีความสัมพันธ์กับค่าของข้อมูลที่จะใส่เข้าไปในเอนทิตีนั้น ๆ คำอธิบายสำหรับแต่ละเอนทิตี นั้นจะต้องชัดเจน

3. การกำหนดแอททริบิวต์ หมายถึง การกำหนดฟิลด์ต่าง ๆ ในตารางข้อมูล ซึ่งจะเป็นตัวให้รายละเอียดของข้อมูลได้ชัดเจนมากขึ้นเพียงใด แอททริบิวต์ที่กำหนดไว้ในเอนทิตีเดียวกันนั้นจะต้องมีความถี่ในการใช้งานใกล้เคียงกัน เพื่อลดเวลาการเข้าถึงและขนาดของเอนทิตี

4. การกำหนดคีย์หลัก และคีย์รอง จะต้องกำหนดอย่างชัดเจน แอททริบิวต์ที่เป็นคีย์หลักจะต้องมีข้อมูลที่ไม่ซ้ำกัน (Unique Key) ในกรณีที่เอนทิตีมีแอททริบิวต์ที่มีคุณสมบัติเป็นคีย์หลักมากกว่าหนึ่ง ให้เลือกใช้เพียงแอททริบิวต์หนึ่งก็พอส่วนที่เหลือให้ใช้เป็นคีย์รอง

5. การกำหนดความสัมพันธ์ระหว่างเอนทิตี (Entity) เพื่อเป็นการลด ความซ้ำซ้อนของข้อมูลที่กระจายอยู่ตามเอนทิตีต่าง ๆ ให้สามารถใช้ข้อมูลเดียวกันได้ การกำหนดความสัมพันธ์ระหว่างเอนทิตีนี้ จะต้องระบุให้ชัดเจนว่าใช้แอททริบิวต์ใด อาจมีหนึ่งแอททริบิวต์หรือมากกว่าสำหรับการกำหนดให้เป็นคีย์นอกที่สามารถอ้างอิงไปถึงแอททริบิวต์ที่เป็นคีย์หลักในอีกเอนทิตีได้รวมทั้งเงื่อนไขในการเปลี่ยนแปลงข้อมูลของคีย์หลักในอีกเอนทิตีหนึ่งที่ถูกอ้างอิงถึงนั้น จะต้องพิจารณาโดยละเอียด

2.4 ระบบการจัดการฐานข้อมูล (Database Management System: DBMS)

เทคโนโลยีที่ใช้ในการสร้าง คือ สร้าง Web Application โดยใช้ภาษา PHP ที่ทำงานในรูปแบบ Server Side Script กล่าวคือ โปรแกรมจะทำงานผ่าน Web Browser เมื่อผู้ใช้เข้าไปในระบบและมีการร้องขอการแสดงผล รายละเอียดที่ร้องขอจะถูกส่งไปประมวลผลที่เครื่องฝั่ง Server ภายหลังจากที่มีการประมวลผลเสร็จ ก็จะส่งผลลัพธ์ไปให้เครื่องลูกข่ายที่ร้องขอข้อมูลนั้น

ในงานวิจัยนี้ผู้วิจัยได้ใช้ MySQL (Structural Query Language) มาใช้ในการจัดการระบบฐานข้อมูล

MySQL เป็นฐานข้อมูลเชิงสัมพันธ์ (Relational Database Management System: RDBMS) ซึ่งสามารถทำงานกับตารางข้อมูลหลายตารางพร้อม ๆ กัน โดยสามารถแสดงความสัมพันธ์ของตารางเหล่านั้นด้วยฟิลด์ (Field) ที่ใช้ร่วมกันเป็น Software ประเภท Open Source

สถาปัตยกรรมของ MySQL คือการออกแบบการทำงานในลักษณะของ Client/Server ซึ่งประกอบด้วยส่วนหลัก ๆ 2 ส่วน คือ ส่วนของผู้ให้บริการแม่ข่าย (Server) และส่วนของเครื่องลูกข่าย (Client) โดยในแต่ละส่วนก็จะมีโปรแกรมสำหรับการทำงานตามหน้าที่ของตน ดังนี้

ส่วนของผู้ให้บริการ หรือ Server จะเป็นส่วนที่ทำหน้าที่บริหารจัดการระบบฐานข้อมูลซึ่งในที่นี้หมายถึงตัว MySQL Server นั้นเอง และเป็นที่ยึดเก็บข้อมูลทั้งหมด ข้อมูลที่เก็บไว้มีทั้งข้อมูลที่เป็นสำหรับการทำงานกับระบบฐานข้อมูลและข้อมูลที่เกิดจากการที่ผู้ใช้แต่ละคนสร้างขึ้นมา

ส่วนของผู้ใช้บริการ หรือ ไคลเอนท์ (Client) คือผู้ใช้โดยโปรแกรมสำหรับผู้ใช้งานในส่วนนี้ ได้แก่ MySQL Client, Access, Web Development Platform ต่าง ๆ เช่น Java, Perl, PHP, ASP เป็นต้น

แต่ตัวจัดการฐานข้อมูลที่เลือกมาใช้ในงานวิจัยนี้คือ MySQL เหตุที่เลือกตัวนี้ คือ เป็นที่นิยมอย่างกว้างขวางและประเด็นหนึ่งที่จะต้องพิจารณา คือ ไม่มีค่าใช้จ่าย เพราะ MySQL จัดเป็น Software ประเภท Freeware รองรับระบบปฏิบัติการได้หลายระบบ

ข้อดีของ MySQL มีดังนี้

MySQL จัดเป็นระบบฐานข้อมูลประเภท SQL-Based ผู้ใช้หรือผู้พัฒนาสามารถ ใช้คำสั่ง SQL ในการส่ง หรือใช้งาน MySQL Server ได้

สนับสนุนการใช้งานสำหรับตัวประมวลผลกลาง (CPU: Central Processing Unit) หลาย ๆ ตัว

สนับสนุน API เพื่อใช้งานกับการพัฒนา Platform ต่าง ๆ มากมาย ไม่ว่าจะเป็น C, C++, Java, Perl, PHP, ASP, Python ซึ่งทำให้สามารถใช้งานได้กับเครื่องมืออื่น ๆ บน Windows Platform เช่น Access เป็นต้น

MySQL สามารถรันได้บนระบบปฏิบัติการหลายตัว เช่น Unix, Linux และ ระบบปฏิบัติการอื่น ๆ

การกำหนดสิทธิและรหัสผ่านให้มีความปลอดภัย ความยืดหยุ่นสูงในการเข้าถึง ข้อมูลได้มีการเข้ารหัสข้อมูล (Encryption) ทำให้ผู้ใช้มีความมั่นใจว่าข้อมูลจะมีความปลอดภัย ไม่มีใครสามารถทำการเข้าถึงข้อมูลได้หากไม่ได้รับอนุญาต

สามารถทำดัชนี (Index) ได้สูงสุดถึง 32 ดัชนี ในแต่ละตารางข้อมูล โดยที่ในแต่ละดัชนีสามารถใช้ฟิลด์ได้ตั้งแต่ 1 ถึง 16 ฟิลด์

สามารถรองรับข้อมูลขนาดใหญ่

สนับสนุนรูปแบบภาษา (Character Set) หลายชนิด ทำให้สามารถจัดเรียง ลำดับ ของข้อมูล (Sort) หรือกำหนดการแสดงผลข้อมูล

3. หลักการเหมืองข้อมูล (Data Mining)

เทคโนโลยีเหมืองข้อมูล (Data Mining) คือ การดึงข้อมูลจากฐานข้อมูลที่มีขนาดใหญ่ หรือคลังข้อมูล (Data Warehouse) เพื่อนำข้อมูลนั้นมาใช้งาน เพื่อการวิเคราะห์ข้อมูลให้เกิด ประโยชน์สูงสุด การทำเหมืองข้อมูลมีการแบ่งขั้นตอนการทำงานที่แตกต่างกันไปตามทฤษฎีที่ ผู้วิจัยเหมืองข้อมูลแต่ละคนใช้อ้างอิง หรือนำไปประยุกต์ใช้ ดังนั้นจึงไม่มีขั้นตอนการทำงาน ของเหมืองข้อมูลที่แน่นอน แต่สามารถแบ่งขั้นตอนการทำงานของเหมืองข้อมูลคร่าว ๆ ได้ ดังนี้ (การจัดการความรู้, 2549. เว็บไซต์)

1. เรียนรู้ขอบเขตของการประยุกต์ใช้เหมืองข้อมูล (Learning)
2. สร้างชุดของข้อมูลจากกลุ่มข้อมูลที่เลือกไว้ (Data Selection)
3. ตัดข้อมูลที่ไม่มีสำคัญออกไป (Data Cleaning)
4. แปลงข้อมูลให้อยู่ในรูปแบบที่สามารถนำไปใช้งานได้ (Data Reduction and Transformation)
5. เลือกวิธีการจัดการเหมืองข้อมูล (Choosing Function of Data Mining)
6. เลือกอัลกอริทึมที่ใช้จัดการ (Choosing the Mining Algorithm)
7. ผลของการจัดกลุ่มในแต่ละวิธีจะถูกนำมาคำนวณในรูปแบบเชิงสถิติอย่างง่าย เทคโนโลยีเหมืองข้อมูล (Data Mining) เป็นการค้นหาความสัมพันธ์และรูปแบบทั้งหมดที่มีอยู่จริงในฐานข้อมูล ซึ่งความสัมพันธ์และรูปแบบเหล่านั้นได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมากที่มีอยู่ โดยเทคโนโลยีเหมืองข้อมูลสามารถดึงความรู้และความสัมพันธ์และทำการสำรวจและวิเคราะห์ข้อมูลให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ โดยความสัมพันธ์หน่วยนี้แสดงให้เห็นถึงความรู้ต่าง ๆ (Knowledge) ที่มีประโยชน์ในฐานข้อมูลในปัจจุบัน (Daras et al., 2002)

ตัวอย่างการใช้เทคนิคเหมืองข้อมูล เช่น สร้างความสัมพันธ์กับลูกค้าในธุรกิจ โดยสังเกตจากความต้องการ ความชอบและความสนใจของลูกค้า และอาจมีการเรียนรู้ได้จากผลสะท้อนในอดีตว่าจะทำอย่างไรให้การบริการลูกค้ามีประสิทธิภาพดีขึ้นในอนาคต เช่น บริษัทที่เป็นผู้ออกบัตรเครดิตและธนาคารต่าง ๆ จะมีขบวนการที่ใช้เทคโนโลยีเหมืองข้อมูลให้เป็นประโยชน์ ในการตัดสินใจว่าลูกค้ากลุ่มใดเป็นกลุ่มที่ดี ทำความเข้าใจลูกค้า ช่วยในการจำแนกประเภทของลูกค้าและจะทำนายกลุ่มของประชากรที่คาดว่าจะมาเป็นลูกค้าในอนาคต เป็นต้น อย่างไรก็ตามการเรียนรู้และการวิเคราะห์นั้นต้องมาจากการเก็บสะสมข้อมูลอย่างตรงไปตรงมา ซึ่งจะทำให้การทำงานเป็นไป อย่างมีประสิทธิภาพ (Viveros et al., 1996)

3.1 การจัดหมวดหมู่ (Classification)

การจัดหมวดหมู่ถือเป็นงานธรรมดาทั่วไปของเทคโนโลยีเหมืองข้อมูล เพราะการทำ ความเข้าใจและการติดต่อสื่อสารต่าง ๆ ก็เกี่ยวข้องกับการแบ่งเป็นหมวดหมู่การจัดแยกประเภท และการแบ่งแยกชนิดโดยการจัดหมวดหมู่ประกอบด้วยการสำรวจจุดเด่นของวัตถุที่ปรากฏออกมา และทำการกำหนด จุดเด่นนั้นๆ เป็นตัวที่ใช้แบ่งหมวดหมู่ งานในการแบ่งหมวดหมู่คือการบ่งบอก ลักษณะ โดยการอธิบายจุดเด่นที่เป็นที่รู้จักดีในหมวดหมู่นั้น และเทรนนิ่งเซต (Training Set)

ของตัวอย่างในแต่ละหมวดหมู่ ซึ่งมีภาระหน้าที่ในการสร้างโมเดลของบางชนิดที่ไม่สามารถจัดหมวดหมู่ของข้อมูลได้ ให้สามารถจัดเป็น หมวดหมู่ได้ ตัวอย่างของการจัดหมวดหมู่ เช่น การจัดหมวดหมู่ของผู้ยื่นขอเครดิต (Credits) เป็นระดับต่ำ ระดับกลาง และระดับสูงของ ความเสี่ยงที่จะได้รับ เป็นต้น

3.1.1 การจัดกลุ่มโดยอาศัยความใกล้ชิดกัน

การจัดกลุ่มโดยอาศัยความใกล้ชิดกัน หรือการวิเคราะห์ของตลาด งาน ในการจัดกลุ่มหรือการวิเคราะห์ตลาด คือ การตัดสินใจรวมสิ่งที่สามารถไปด้วยกันเข้าไว้ใน กลุ่มเดียวกันตัวอย่างของการจัดกลุ่มโดยอาศัยความใกล้ชิดกันหรือการวิเคราะห์ตลาด เช่น การตัดสินใจว่าสิ่งใดบ้างที่จะไปอยู่ด้วยกันอย่างสม่ำเสมอในรถเข็นในซูเปอร์มาร์เก็ต

3.1.2 การรวมตัว

การรวมตัวคืองานที่ทำการรวมส่วนต่าง ๆ ในแต่ละส่วนที่ต่างชนิดกันให้อยู่ ในรวมกันเป็นกลุ่มย่อย หรือคลัสเตอร์ (Clusters) โดยในแต่ละคลัสเตอร์อาจจะประกอบด้วย ส่วนต่าง ๆ ที่ต่างชนิดกัน ซึ่งความแตกต่างของการรวมตัวจากการจัดหมวดหมู่ คือ การรวมตัว จะไม่พึ่งพาอาศัยการกำหนดหมวดหมู่ล่วงหน้าและไม่ใช้ตัวอย่าง ข้อมูลจะรวมตัวกันบนพื้นฐาน ของความคล้ายในตัวเอง

วิธีคลัสเตอร์จริงนี้ เป็นวิธีที่อาจจะเรียกว่าเป็นการทำเทคโนโลยีเหมืองข้อมูล แบบอ้อม ๆ เนื่องจากการหาผลลัพธ์ในแต่ละครั้งนั้น แม้กระทั่งผู้หายังไม่อาจจะทราบว่าการที่ ต้องการจะหานั้นคืออะไร จำเป็นต้องรอจนกว่าการค้นหาจะทำเสร็จสมบูรณ์จึงจะทราบข้อมูลที่ ซ่อนอยู่ เปรียบเสมือนกับการมีข้อมูลจำนวนมากมาอยู่ในระบบฐานข้อมูล แล้วจากนั้นก็ใช้ เทคนิคหรือวิธีการมาจัดเรียงข้อมูลหน่วยนั้นให้อยู่เป็นกลุ่ม ซึ่งเราสามารถวิเคราะห์ข้อมูลของ แต่ละกลุ่มตามความต้องการได้

Zaiane (1999) ได้แบ่งวิธีการของการจัดกลุ่ม (Clustering) ออกเป็น 5 ประเภทหลัก ๆ ดังนี้ การแบ่งส่วน (Partitioning Algorithms) การทำลำดับชั้น (Hierarchy Algorithms) พื้นฐานของความหนาแน่น (Density-based) พื้นฐานของกริด (Grid-based) พื้นฐานของโมเดล (รูปแบบการจัดกลุ่มที่ดีที่สุด) (Model-based) ซึ่งเทคนิคการจัดกลุ่มแบบ การแบ่งส่วนเป็นวิธีที่เหมาะสมที่สุดในงานวิจัยนี้ โดยแนวคิดพื้นฐานของการแบ่งส่วนนี้จะทำได้ โดยการสร้างและใช้อัลกอริทึมในการแบ่งส่วนเป็นหลาย ๆ ส่วน และประเมินส่วนต่าง ๆ นั้น โดยใช้เกณฑ์ที่เรากำหนดไว้ ตัวอย่าง อัลกอริทึมแบบการแบ่งส่วน

แบ่งวัตถุหรือสิ่งของจำนวน n วัตถุ ออกเป็นกลุ่มของคลัสเตอร์จำนวน k กลุ่ม จาก ฐานข้อมูล D จากนั้นค้นหาส่วนต่างๆ ในจำนวน k กลุ่ม ที่เหมาะสมกันที่สุด ตามเกณฑ์ที่เราได้ตั้งไว้ โดยอาจใช้การระบุจำนวนการแบ่งส่วนทั้งหมด (Global Optimal)

ข้อดีของ Clustering

1. ลดขอบเขตในการหาข้อมูลที่มีจำนวนมาก โดยจัดกลุ่มของระเบียบ (Record) ที่คล้ายคลึงกันให้มาอยู่ในกลุ่มเดียวกัน (Christen and Crunches, 2002)
2. การที่นำระเบียบที่คล้ายคลึงกันให้มาอยู่ในกลุ่มเดียวกัน ทำให้ง่ายต่อการเปรียบเทียบและวิเคราะห์ข้อมูล
3. ช่วยเพิ่มความรวดเร็วในการเข้าถึงข้อมูล ในการพัฒนาและทดลองระบบ ฐานข้อมูลที่มีขนาดใหญ่

อย่างไรก็ตามระบบในปัจจุบันนี้ สามารถแก้ปัญหาโดยใช้หลักการทำดัชนี (Index) ของระบบการจัดการฐานข้อมูล (Data Base Management System: DBMS) เนื่องจากมีความสะดวกและง่ายต่อการดำเนินการและง่ายต่อการจัดกลุ่มใหม่

วัตถุประสงค์ของการจัดกลุ่ม คือ การนำเสนอข้อมูลที่ไม่เหมือนกันไปสู่ในรูปแบบที่สามารถจัดการได้ง่าย โดยใช้หลักการแบ่งแยกให้เป็นหมวดหมู่ (Van Ooyen, 2001) ตัวอย่างเช่น Reaney and Wilson (1997) ได้จัดหมวดหมู่ชื่อสกุลภาษาอังกฤษ และชื่อที่มีความเกี่ยวพันคล้ายคลึงกันในรูปแบบพจนานุกรม

การจัดกลุ่ม Grouping หรือ Clustering เป็นการทำงานที่มีประโยชน์มากที่สุด ในกระบวนการทำเทคนิคเหมืองข้อมูล (Data Mining) สำหรับการค้นหากลุ่มและระบุความสนใจที่แตกต่างมาให้มาอยู่ในรูปแบบข้อมูลเดียวกัน (Halkidi et al., 2001) ตัวอย่างเช่น การจัดหมวดหมู่ข้อมูลการขายสินค้าในห้างสรรพสินค้า เพื่อที่จะทำให้มีประสิทธิภาพในการขายที่ดีขึ้น (Ramkuma and Swami, 1998)

จักรกฤษณ์ (Snae, 2006) ได้อธิบายการนำแนวความคิดของการจัดกลุ่มไปประยุกต์ใช้กับการเชื่อมโยงระเบียบข้อมูล ดังต่อไปนี้

ใช้ในการจัดเรียงข้อมูลที่มีความสัมพันธ์และเกี่ยวข้องกัน โดยการนำหลักการ จัดหมู่อย่างเป็นลำดับขั้นตอน ซึ่งการจัดกลุ่มและการทำดัชนีจะเก็บชนิดของการเชื่อมโยงที่มีข้อมูลเกี่ยวข้องกันไว้ด้วยกัน จะช่วยในการประหยัดเวลาและง่ายต่อการบำรุงรักษา ข้อมูลและสามารถค้นหาได้ง่าย ถ้าข้อมูลมีการจัดเรียงกัน เช่น เรียงตามชื่อหรือเวลา ของข้อมูลหรือการทำดัชนีของชื่อหรือประเภทของข้อมูลที่คล้ายคลึงกันอยู่ด้วยกัน ใช้ในการกรองข้อมูลที่ไม่เกี่ยวข้องกัน

จากเอกสารตามที่เราสนใจมีความจำเป็นที่จะ สามารถสร้างการเชื่อมโยงข้อมูลที่มีประสิทธิภาพ ได้ดีและป้องกันการเชื่อมโยงที่ไม่ต้องการในระหว่างการสืบค้นข้อมูลจากฐานข้อมูล การจัดกลุ่ม และการสร้างดัชนี สามารถถูกนำไปใช้ในการแบ่งประเภทของข้อมูลที่พบออกเป็น ข้อมูลที่มีความเกี่ยวข้องกันและข้อมูลที่ไม่มีความเกี่ยวข้องกัน จะถูกกำจัดออกซึ่งจะช่วยให้ประหยัดเวลา ในการสืบค้นและการเข้าถึงข้อมูลมีประสิทธิภาพและทำได้อย่างรวดเร็วยิ่งขึ้น

3.2 การประเมินค่า (Estimation) และการทำนายล่วงหน้า

การประเมินค่าทางธุรกิจอย่างต่อเนื่องจะก่อให้เกิดผลลัพธ์ที่มีประโยชน์กับธุรกิจ การป้อนข้อมูลที่เราได้อยู่เข้าไป เพื่อใช้ในการประเมินสิ่งต่าง ๆ ที่จะก่อให้เกิดประโยชน์ หรือ สำหรับตัวแปรที่เราไม่รู้ค่า แน่แน่นอนเช่น รายได้จากการค้า จุดสูงสุดทางธุรกิจ หรือดุลยภาพของ บัตรเครดิต ในทางปฏิบัติการประเมิน ค่าจะถูกใช้ในการทำงานการจัดหมวดหมู่ ตัวอย่างของ การประเมินค่า เช่น การประเมินรายได้รวมของ ครอบครัว หรือการประเมินจำนวนบุตรใน ครอบครัว

การทำนายล่วงหน้าก็เป็นงานที่มีลักษณะคล้ายกับการจัดหมวดหมู่หรือการประเมิน ค่า ยกเว้น เพียงแต่จะใช้สถิติการบันทึกของการจัดหมวดหมู่ในการทำนายอนาคตของพฤติกรรม หรือการประเมิน ค่าที่จะเกิดขึ้นในอนาคต ตัวอย่างของงานการทำนายล่วงหน้า เช่น การทำนาย การเปลี่ยนแปลงพฤติกรรม ของตลาด หรือการทำนายจำนวนลูกค้าที่จะออกจากธุรกิจของเรา ใน 6 เดือนข้างหน้า เป็นต้น

3.3 การบรรยาย (Description)

ในบางครั้งวัตถุประสงค์ของเทคโนโลยีเหมืองข้อมูล คือ ต้องการอธิบาย ความลับของฐานข้อมูลในทางที่จะเพิ่มความเข้าใจในส่วนของผู้บริโภค ผลิตภัณฑ์ หรือขบวนการให้มากขึ้น เทคโนโลยีเหมืองข้อมูลส่วนใหญ่ต้องการทราบถึงข้อมูลจำนวนมากที่ ประกอบด้วยหลาย ๆ ตัวอย่างเพื่อจะสร้างกฎที่ใช้ในการจัดหมวดหมู่ กฎของความสัมพันธ์ คลัสเตอร์ การทำนายล่วงหน้า ดังนั้น ชุดของข้อมูลขนาดเล็กจะนำไปสู่ความไม่เข้าใจของผลสรุปที่ดีได้ ไม่มีเทคนิคใดเลยที่จะสามารถแก้ปัญหาของเทคโนโลยีเหมืองข้อมูลได้ทุก ปัญหา ดังนั้นความหลากหลายของเทคนิคจึงเป็นสิ่งจำเป็นในการไปสู่วิธีการแก้ปัญหาของ เทคโนโลยีเหมืองข้อมูลได้ดีที่สุด

4. หลักสถิติและความน่าจะเป็น (Statistical Model and Probability)

สถิติศาสตร์ เป็นคณิตศาสตร์ประยุกต์ซึ่งจัดเป็นแขนงหนึ่งของวิทยาศาสตร์ที่ใช้ศึกษาคุณสมบัติเชิงปริมาณและคุณภาพของข้อมูลที่สังเกตได้ (จากตัวแปรในการวิจัย) ซึ่งสามารถนำมาใช้ทำความเข้าใจปรากฏการณ์ธรรมชาติหรือเหตุการณ์ต่าง ๆ ทางสังคม สถิติศาสตร์มีความเป็นมาที่ยาวนาน เริ่มตั้งแต่การใช้เป็นเครื่องมือแสวงหาข้อมูลสำหรับการบริหารงานจนถึงการวิเคราะห์ข้อมูลที่ได้จากกลุ่มตัวอย่างเพื่อสรุปอ้างอิง (สร้างนัยทั่วไป) เกี่ยวกับประชากร เป้าหมาย อันช่วยเป็นแนวทางสู่การตัดสินใจภายใต้สถานการณ์ที่ไม่แน่นอน (ศิริชัย กาญจนวาสี , ทวีวัฒน์ ปิตยานนท์ และดิเรก ศรีสุข, 2544. หน้า 33)

สถิติศาสตร์มีความสำคัญต่อการวิจัย ในแง่ของการช่วยให้แนวทางจัดประเภทระเบียบของข้อมูล การเสนอข้อมูลในรูปแบบที่เหมาะสม ให้เทคนิคของการเปรียบเทียบการประมาณค่า การทดสอบสมมติฐาน และการสรุปอ้างอิงที่ถูกต้อง

ผลของการจัดกลุ่มในแต่ละวิธีจะถูกนำมาคำนวณในรูปแบบเชิงสถิติอย่างง่ายโดยใช้สมการ

$$p = \frac{f}{n} \times 100 \text{ -----(1)}$$

โดยที่ p หมายถึง ค่าร้อยละ

f หมายถึง จำนวนเหตุการณ์ที่ใช้เป็นเงื่อนไขในการหาผลลัพธ์

f สามารถมีหลายเหตุการณ์ได้ เช่น f ประกอบไปด้วยเหตุการณ์ a และเหตุการณ์ b

a หมายถึง จำนวนเหตุการณ์ที่มีเวลาเข้ามาเกี่ยวข้อง

b หมายถึง จำนวนเหตุการณ์ที่ผู้ใช้เข้าถึงเว็บต่าง ๆ ต่อเหตุการณ์
ณ เวลาใดเวลาหนึ่ง

ดังนั้น เหตุการณ์ $(f) = \text{เหตุการณ์ } \{ a, b \}$

n หมายถึง จำนวนเหตุการณ์ทั้งหมด

ซึ่งการแสดงผลลัพธ์ในรูปแบบสถิตินี้จะช่วยให้ผู้ใช้สามารถเข้าใจผลลัพธ์ของการประมวลผลได้ง่ายยิ่งขึ้น

5. หลักวิศวกรรมซอฟต์แวร์ (Software Engineering)

การออกแบบเป็นกระบวนการค้นพบและเปลี่ยน "สิ่งที่จะต้องทำ" ให้กลายเป็น "สิ่งที่ทำให้ได้คำตอบที่ต้องการ" จรณิต แก้วกังวาน (2540) โดยเริ่มจากขั้นตอนการวิเคราะห์ และกำหนดคำจำกัดความของระบบ (Requirement Definition) มีเอกสารรายงานผลการวิจัยความเป็นไปได้ (Feasibility Study) และรายละเอียดพื้นฐานของระบบ (Requirement Specification) ในขั้นตอนการกำหนดรายละเอียดของระบบ (System Specification) นอกจากนี้มีเอกสารระบุรายละเอียดการทำงานของระบบแล้ว ยังถือเป็นจุดเริ่มต้นลงมือร่างคู่มือการใช้ระบบสำหรับผู้ใช้งาน (Draft User Manual)

ในขั้นตอนการออกแบบระบบงานจริงไม่ว่าจะเป็นส่วนของโครงสร้างส่วนเชื่อมต่อกับผู้ใช้ระบบ หรือส่วนอื่น ๆ ในระบบ จะมีเอกสารระบุรายละเอียดแต่ละส่วนควบคู่ไปกับเอกสารระบุวิธีการตรวจสอบแต่ละส่วนของระบบ (Test Specification) เมื่อลงมือเขียนโปรแกรม (Coding) เอกสารที่ปรากฏเห็นได้ชัด คือ ตัวโปรแกรมที่ได้เป็นผลลัพธ์ออกมา

ในการพัฒนาระบบงานใด ๆ ก็ตาม ผู้ออกแบบสร้างระบบต้องมองเห็นภาพรวมทั้งหมดของระบบอย่างชัดเจน และต้องมีเป้าหมายที่แน่นอนว่าระบบดังกล่าวจะสร้างขึ้นมาจากจุดมุ่งหมายใดเพื่อใคร และควรมีขั้นตอนในการดำเนินการอย่างไร โดยปกติแล้วเราอาจกำหนดลำดับความสำคัญของเป้าหมายในการออกแบบพัฒนาระบบได้เป็น 3 ประเด็น ตามลำดับ ดังนี้

1. โครงสร้างการทำงานของระบบงาน (Functionality) คือ การกำหนดโครงสร้างของชิ้นงาน โดยจัดสรรแบ่งงานออกเป็นส่วน ๆ อย่างเหมาะสม ซึ่งเป้าหมายแรกสุดของการพัฒนาระบบจึงอยู่ที่การวิเคราะห์งาน (Task Analysis)

2. คุณภาพของระบบงาน (Quality) ซึ่งจะครอบคลุมด้านความเชื่อถือได้ (Reliability) ความพร้อม (Availability) ความมั่นคงปลอดภัย (Security) และความถูกต้องสมบูรณ์ (Integrity) ทั้งนี้หลังจากที่ได้กำหนดเป้าหมายในด้านการวิเคราะห์โครงสร้างของระบบแล้ว เป้าหมายในลำดับที่สองที่ผู้ออกแบบจะต้องคำนึงถึง คือ การกำหนดให้ระบบงานมีความเชื่อถือได้ และได้ผลตรงตามที่ผู้ใช้ต้องการเสมอ ระบบที่ออกแบบขึ้นมานั้นต้องมีความพร้อมที่จะทำให้อุ้สามารถใช้สามารถติดตั้ง และดำเนินการประยุกต์ใช้ได้อย่างง่ายดายและมีประสิทธิภาพ อีกทั้งผู้พัฒนาระบบต้องตั้งเป้าหมายของการออกแบบคุณสมบัติของระบบให้มีความมั่นคงปลอดภัยและความสมบูรณ์ของข้อมูล ผู้ออกแบบควรจะได้กำหนดขั้นตอนการเข้าออกและการเรียกใช้ข้อมูลต่าง ๆ ในระบบอย่างระมัดระวัง ป้องกันการสูญหายและเสียหายของข้อมูล ในขณะเดียวกันก็ปกป้องข้อมูลในส่วนที่เป็นเรื่องส่วนตัวหรือลับเฉพาะขององค์กร เป็นต้น

3. กำหนดการ (Schedules) ในขั้นตอนสุดท้ายผู้ออกแบบจะต้องวางแผนดำเนินการอย่างเป็นระบบ ขั้นตอนการดำเนินการควรจะมีกำหนดการที่ชัดเจน

5.1 กระบวนการออกแบบและพัฒนาซอฟต์แวร์

กระบวนการพัฒนาซอฟต์แวร์ (Software Process) หรือระบบงานคอมพิวเตอร์นั้นมีความคล้ายคลึงกับกระบวนการที่ใช้กันในทางวิศวกรรม (Engineering Process) การใช้กระบวนการจัดกิจกรรมของระบบงานทางวิศวกรรมในการดำเนินงานเรื่องต่าง ๆ ทำให้เราสามารถมองเห็นขั้นตอนการทำงานได้อย่างเด่นชัด แต่อย่างไรก็ตามกระบวนการออกแบบและพัฒนาซอฟต์แวร์ด้วยกระบวนการทางวิศวกรรมนี้ก็ยังคงมีลักษณะเป็นการศึกษาหาวิธีดำเนินการที่เหมาะสมที่สุด

กระบวนการออกแบบและพัฒนาระบบงานคอมพิวเตอร์หรือซอฟต์แวร์ที่ใช้กันอย่างแพร่หลาย สามารถจำแนกตามรูปแบบหรือโครงสร้างของวิธีดำเนินการ (Models/Paradigms) ได้ 3 รูปแบบด้วยกัน ได้แก่

1. วอเตอร์ฟอล แอพโพรช (Waterfall Approach)
2. เอ็กซ์โพลราโตรีโปรแกรมมิง (Exploratory Programming)
3. โปรโตไทป์มิง (Prototyping)

5.1.1 วอเตอร์ฟอล แอพโพรช (Waterfall Approach)

วิธีการนี้เป็นการแบ่งกระบวนการดำเนินการออกเป็นขั้นตอนต่าง ๆ เช่น การกำหนดลักษณะจำเป็นพื้นฐาน การออกแบบระบบงาน การสร้างและทดสอบระบบงาน เป็นต้น ขั้นตอนในแต่ละช่วงสืบเนื่องกันไปจากขั้นหนึ่งสู่อีกขั้นหนึ่งตามลำดับเหมือนกับสายน้ำตก จึงเรียกว่าเป็นกระบวนการพัฒนาแบบ Waterfall ทำให้เกิดเป็นวงจรชีวิตของซอฟต์แวร์ (Software Life Cycle) ซึ่งอาจแบ่งได้เป็น 5 ขั้นตอน ดังนี้

5.1.1.1 การวิเคราะห์และให้คำจำกัดความของระบบงาน

(Requirements Analysis and Definition) ในขั้นตอนแรกนี้ผู้ออกแบบระบบและผู้ใช้ระบบจะต้องพูดคุยให้เป็นที่เข้าใจกันว่า เป้าหมายและโครงสร้างระบบงานที่ผู้ใช้ต้องการคืออะไร

5.1.1.2 การออกแบบระบบและซอฟต์แวร์ (System and Software Design)

ในการออกแบบระบบงาน ผู้ออกแบบจะต้องคำนึงถึงโครงสร้างของฮาร์ดแวร์และซอฟต์แวร์ที่จำเป็นต้องพัฒนาหรือจัดหาใช้งาน ส่วนการออกแบบซอฟต์แวร์เป็นการกำหนดโครงสร้างหลักของโปรแกรมที่จะถูกพัฒนาขึ้นมา

5.1.1.3 การกำหนดสร้างและทดสอบหน่วยย่อย (Implementation and Unit Testing) ในขั้นตอนนี้เป็นการแบ่งส่วนของซอฟต์แวร์ออกเป็นหน่วยของโปรแกรมย่อย ๆ (Program Units) และเมื่อเขียนโปรแกรมแต่ละหน่วยย่อยเสร็จแล้ว ก็จะต้องมีการตรวจสอบว่าแต่ละหน่วยมีความถูกต้องตรงตามรูปแบบที่กำหนดไว้

5.1.1.4 การเชื่อมโยงและการทดสอบทั้งระบบ (Integration and System Testing) โปรแกรมย่อยแต่ละส่วนจะถูกนำมาประสานรวมกันเป็นระบบงาน หลังจากนั้นทั้งระบบซอฟต์แวร์ที่พัฒนาขึ้นต้องได้รับการตรวจสอบ เพื่อให้แน่ใจว่าตรงตามแผนการที่ได้ออกแบบไว้และสามารถทำงานได้อย่างถูกต้อง เมื่อทดสอบเป็นที่พอใจแล้วจึงนำระบบนั้นไปเสนอต่อผู้ให้ระบบ

5.1.1.5 การติดตั้งใช้งานและการบำรุงรักษา (Operation and Maintenance) ขั้นตอนนี้มักเป็นขั้นตอนที่ใช้เวลานานที่สุด เมื่อระบบถูกนำมาติดตั้งและใช้งานจริงแล้ว การบำรุงรักษาที่มักปรากฏคือ การแก้ไขข้อผิดพลาดที่อาจไม่เคยพบมาก่อนในช่วงพัฒนาระบบหรืออาจเป็นการปรับแต่งระบบให้สวยงาม มีประสิทธิภาพในการทำงานดียิ่งขึ้น ขั้นตอนนี้จึงมักจะเป็นการย้อนกลับไปพิจารณาปรับแต่งแก้ไขสิ่งที่ได้กำหนดไว้ใน 3 ขั้นตอนที่ผ่านมาแล้วตามที่เห็นเหมาะสมในภายหลัง

ในกระบวนการออกแบบและพัฒนาระบบงานดังกล่าวนี้ ผู้ออกแบบระบบไม่สามารถแบ่งแต่ละขั้นตอนจากขั้นหนึ่งสู่อีกขั้นหนึ่งออกจากกันเด็ดขาดได้ เพราะในการทำงานจริงเราพบว่า มีบางขั้นตอนที่เกี่ยวข้องกันอยู่แบ่งเด็ดขาดไม่ได้ กระบวนการพัฒนาที่เกิดขึ้นจึงมักไม่ได้อยู่ในรูปแบบที่เดินไปข้างหน้าแบบเส้นตรง (Linear Model) เสมอไป แต่มักเป็นรูปแบบที่มีการกระทำย้อนกลับไปมา หมุนวนในระหว่างช่วงขั้นตอนต่างๆ (Iteration Model) เช่น ในขณะที่อยู่ในขั้นตอนการออกแบบอาจต้องย้อนกลับไปกำหนดคำจำกัดความบางอย่างของระบบใหม่ หรือในขณะที่กำลังเขียนโปรแกรมอยู่อาจต้องย้อนกลับ ไปแก้แบบแผนที่ได้วางไว้ตั้งแต่ต้น ในบางเรื่อง เป็นต้น แต่อย่างไรก็ตามผู้ออกแบบจึงต้องเป็นผู้ที่มีความสามารถในการชั่งน้ำหนักระหว่างข้อดีกับข้อเสียของการย้อนกลับไปในแต่ละขั้นตอน

กระบวนการที่สำคัญในช่วงชีวิตของระบบซอฟต์แวร์ดังกล่าวข้างต้นนี้ คือ การทดสอบความถูกต้องสมบูรณ์ของระบบงานก่อนที่จะนำไปเสนอต่อผู้ให้ระบบ กระบวนการทดสอบ อาจแบ่งได้เป็น 2 ประเด็น คือ การยืนยันผลลัพธ์และการยืนยันความเที่ยงตรง (Verification and Validation – V & V) จำแนกข้อแตกต่างระหว่าง V & V ไว้ว่า Verification เป็นการตรวจสอบว่าผลผลิตที่ได้จากการพัฒนาครั้งนี้ตรงกับข้อกำหนดเบื้องต้นที่ได้ออกแบบไว้

หรือไม่ ส่วน Validation เป็นการตรวจสอบยืนยันผลว่าส่วนต่าง ๆ ของผลผลิตที่ได้นี้ทำงานได้ตรงตามกับความต้องการของผู้ใช้ระบบหรือไม่

ในกระบวนการสุดท้าย ซึ่งเป็นการติดตั้งใช้ระบบและการบำรุงรักษาระบบงาน ผู้ใช้ระบบมักพบว่ามีปัญหาที่ผู้ออกแบบไม่เคยคาดคิดมาก่อน หรือผู้ใช้ระบบอาจต้องการเปลี่ยนรูปแบบบางอย่างที่ผู้ออกแบบไม่ได้จัดไว้ให้ ดังนั้น กระบวนการกำหนดทดสอบรายละเอียดและแก้ไข จึงย้อนกลับไปเริ่มต้นยังจุดต่าง ๆ ในช่วงชีวิตของซอฟต์แวร์อีกครั้งหนึ่ง

เทคนิคการออกแบบและพัฒนาซอฟต์แวร์แบบ Waterfall Approach นี้ เป็นเพียงการประยุกต์ใช้หลักการวิเคราะห์และออกแบบระบบงานธรรมดาทั่ว ๆ ไปเท่านั้น ข้อด้อย คือ อาจมีความไม่สมดุลเกิดขึ้นระหว่างการย้อนกลับไปทบทวนแก้ไขรายละเอียดในขั้นตอนต่าง ๆ กับการที่จะทำได้ผลผลิตเสร็จสมบูรณ์ตรงตามความต้องการของผู้ใช้ระบบ และทันเวลาที่ผู้ใช้ต้องการ ดังนั้น อีก 2 เทคนิคคือ Exploratory Programming และ Prototyping จึงได้รับการพัฒนาขึ้นมาเพื่อแก้ไขข้อด้อยดังกล่าว

5.1.2 เอ็กซ์โพลราโตรีโปรแกรมมิง (Exploratory Programming)

เทคนิคการออกแบบและพัฒนาซอฟต์แวร์แบบ Exploratory Programming จะคำนึงถึงความรวดเร็วของการออกแบบพัฒนาระบบมากกว่าความถูกต้องและแม่นยำ แล้วค่อยปรับแต่งให้ตรงตามความต้องการของระบบภายหลัง

กระบวนการออกแบบนี้เริ่มจากกำหนดโครงร่างเพียงคร่าว ๆ (Outline Specification) โดยที่ไม่สามารถกำหนดรายละเอียดเบื้องต้นของระบบ (Detailed System Specification) หลังจากนั้นจึงสร้างและทดลองใช้ระบบจนกว่าจะได้รูปแบบที่เหมาะสมที่สุดที่จะนำไปใช้งานจริงในขั้นสุดท้าย

ข้อแตกต่างที่สำคัญระหว่างการออกแบบ Exploratory Programming และ แบบ Waterfall Approach (ซึ่งเน้นการออกแบบรายละเอียดของระบบ) อยู่ตรงที่ขั้นตอนการตรวจสอบและยืนยันความถูกต้อง (V&V) ทั้งนี้เพราะ V แรก (Verification) เป็นการตรวจสอบว่าระบบที่พัฒนาขึ้นมาตรงตามรายละเอียดที่ได้ออกแบบไว้หรือไม่อย่างไร ดังนั้น เมื่อเราไม่สามารถกำหนดรายละเอียดเบื้องต้นได้ ขั้นตอนการตรวจสอบนี้ก็จะเป็นไปไม่ได้ กระบวนการตรวจสอบความถูกต้อง V ที่สอง (Validation) จึงเป็นการตรวจสอบความถูกต้องเหมาะสมของการทำงานของโปรแกรมมากกว่าเป็นการตรวจสอบความเที่ยงตรงตามโครงสร้างที่ได้ออกแบบเอาไว้ความถูกต้องเหมาะสมเป็นสิ่งที่วัดได้ยากและค่อนข้างเป็นการตัดสินใจตามใจ

ผู้ตรวจสอบ (Subjective) เราไม่สามารถรับรองได้ว่าพฤติกรรมของมนุษย์เป็นความถูกต้องแต่เราสามารถบอกได้ว่า เราพอใจกับพฤติกรรมที่ทำงานได้สอดคล้องกับงานที่ต้องการจะให้ทำ

การออกแบบพัฒนาระบบซอฟต์แวร์โดยใช้เทคนิค Exploratory Technique อาจเหมาะสำหรับระบบงานบางอย่างโดยเฉพาะ แม้ว่าอาจใช้ไม่ได้ในการพัฒนาระบบงานทั่วไป แต่อย่างไรก็ตามเทคนิคนี้ก็อาจนำมาประยุกต์ใช้ได้ดีในการบำรุงรักษาระบบงานทั่วไป ที่ยังไม่มีความชัดเจนว่าจะใช้วิธีการใดในการปรับแต่งแก้ไขปัญหาที่เกิดขึ้น โดยสรุปแล้วเทคนิค Exploratory Programming เหมาะที่จะถูกเลือกใช้ในกรณีต่าง ๆ ต่อไปนี้

5.1.2.1 โครงสร้างของระบบงานที่กำลังพัฒนาเป็นไปในลักษณะที่ต้องมีการนำไปทดลองใช้อย่างสม่ำเสมอ การปรับเปลี่ยนแก้ไขไปมาเพื่อความเหมาะสมจึงถือได้ว่าเป็นการประเมินความก้าวหน้าของโปรแกรมไปตามลำดับ

5.1.2.2 ระบบงานที่แม้จะมีโครงสร้างที่ไม่ชัดเจน แต่ผู้พัฒนาระบบและผู้แก้ไขระบบในภายหลังเป็นบุคคลเดียวกัน (แต่อย่างไรก็ตามในกรณีที่ระบบงานขนาดใหญ่ และต้องใช้งานในหน่วยงานอยู่เป็นเวลานาน เทคนิคนี้จะไม่เหมาะสม เพราะคนสร้างระบบและคนปรับแต่งแก้ไขระบบในภายหลังมักจะไม่ใช้คนเดียวกัน จึงอาจมีผลทำให้โครงสร้างและรูปแบบของระบบบิดเบือนไปจากจุดเดิมจนเสียรูปไปได้)

5.1.2.3 แม้วางยังไม่มีงานวิจัยยืนยันลักษณะของทีมงานที่เหมาะสมสำหรับงานออกแบบโดยใช้เทคนิคนี้ แต่มีผู้แนะนำไว้กว้าง ๆ ว่าทีมงานที่จะใช้วิธีการออกแบบชนิดนี้ได้ดี น่าจะเป็นทีมงานขนาดเล็กที่ประกอบด้วยผู้ที่มีทักษะเชี่ยวชาญ และเป็นผู้ที่มีแรงจูงใจในการทำงานสูง

5.1.3 โปรโตไทป์ (Prototyping)

เทคนิคการออกแบบลักษณะ Exploratory Programming และลักษณะ Prototyping จะมีส่วนที่เหมือนกัน คือ ได้พัฒนาระบบไปเป็นเฟส แล้วส่งให้ผู้ใช้ได้ทดลองใช้ก่อนแล้วค่อยแก้ไขในภายหลัง แต่วิธีการทั้งสอง ก็มีข้อแตกต่างที่สำคัญคือ Exploratory Programming จะเริ่มต้นจากความเข้าใจของระบบที่ยังไม่ชัดเจน หลังจากทดลองใช้แล้ว จึงมีการเก็บรายละเอียดย่อยของงานจนเป็นระบบที่สมบูรณ์ ส่วนวิธีแบบ Prototyping จะเริ่มต้นจากโครงสร้างของระบบ แบบคร่าว ๆ (Outline Requirement) แต่การพัฒนาระบบขึ้นมาใช้งาน มีจุดประสงค์ที่จะค้นหารายละเอียดของระบบ บางทีอาจจะเรียกว่าเป็นต้นแบบ (Prototype) เมื่อได้ระบบต้นแบบมาทดลองใช้และต่อเติมตามลำดับจนเห็นว่าใช้งานได้ (Executable) ก็จะหยุด

พัฒนาระบบต้นแบบ แล้วจะสร้างระบบงานจริงใหม่อีกครั้ง เพื่อให้ได้ระบบงาน ที่มีคุณภาพตรงตามต้องการของผู้ใช้ระบบ

5.2 การทดสอบซอฟต์แวร์ (Software Testing)

การทดสอบซอฟต์แวร์ เป็นขั้นตอนสุดท้ายของกระบวนการออกแบบและพัฒนา ระบบงาน คือ การตรวจสอบระบบงาน เริ่มตั้งแต่การตรวจสอบแต่ละหน่วยย่อย (Unit Testing) การตรวจสอบหน่วยใหญ่ (Module Testing) การตรวจสอบแบบเชื่อมโยงหน่วยต่าง ๆ เข้าด้วยกัน (Integration Testing) การตรวจสอบระบบทั้งหมด (System Testing) และสุดท้าย คือ การตรวจสอบเพื่อการยอมรับระบบ (Acceptance Testing) ในการตรวจสอบทุกขั้นตอน ต้องมีการเขียนรายงานสรุปผลการตรวจสอบ เพื่อใช้เป็นหลักฐานในการปรับปรุงแก้ไขปัญหาที่เกิดขึ้นในแต่ละขั้นตอน

5.3 การออกแบบส่วนเชื่อมต่อกับผู้ใช้ระบบ (User interface Design)

การออกแบบระบบงานโดยคำนึงถึงตัวผู้ใช้ (User Interface Design)

เป็นการออกแบบที่ต้องผสมผสานศาสตร์ 2 ด้าน ที่ไม่น่าจะอยู่รวมกันเข้าด้วยกัน คือ ด้านจิตวิทยาและด้านคอมพิวเตอร์ศาสตร์ แนวความคิดพื้นฐานของทั้งสองศาสตร์นี้แตกต่างกัน กล่าวคือ จิตวิทยาจะเน้นในเรื่องของคน ในขณะที่คอมพิวเตอร์ศาสตร์เกี่ยวข้องกับเครื่องมือจักรกลอิเล็กทรอนิกส์ โปรแกรม ภาพลักษณ์นักจิตวิทยา คือ ผู้มีความเข้าใจและพยายามให้ความช่วยเหลือผู้คน ส่วนภาพลักษณ์ของนักคอมพิวเตอร์ คือ ผู้เชี่ยวชาญทางด้านคณิตศาสตร์และความถูกต้องแม่นยำ

หลักการสำคัญของการสร้างความสัมพันธ์กับผู้ใช้ คือ การที่ระบบยอมให้ผู้ใช้สามารถใช้ระบบได้อย่างเต็มที่เท่าที่ระบบจะสามารถเอื้อให้ได้ ระบบงานที่ถูกพัฒนาขึ้นมาโดยไม่คำนึงถึงตัวผู้ใช้จึงถือได้ว่าเป็นระบบงานที่ไร้ประโยชน์ แต่โดยปกติแล้วเรามักพบว่าการคำนึงถึงตัวผู้ใช้ระบบมักเป็นเรื่องที่ผู้ออกแบบทั่วไปให้ความสนใจน้อยในระยะแรก ๆ ของการออกแบบ แต่จะเป็นประเด็นที่เกิดขึ้นภายหลัง (After Thought) หลังจากที่ระบบถูกนำไปใช้แล้วและผู้ใช้เกิดมีปัญหามา ทำให้ผู้ออกแบบระบบ หรือนักเขียนโปรแกรมต้องตามมาแก้ไข ความจริงแล้วถ้าในช่วงแรก ๆ ของการออกแบบระบบหรือในขณะที่เขียนโปรแกรมได้มีการคำนึงถึงตัวผู้ใช้ระบบ ประกอบกันไปด้วย จะทำให้ระบบงานที่ดีขึ้น มีข้อผิดพลาดที่ต้องแก้ไขในภายหลังน้อยลง

โดยสรุปแล้ว จุดมุ่งหมายหลักของระบบงานคอมพิวเตอร์ คือ การพัฒนาประโยชน์ของการดำเนินการเรื่องต่าง ๆ ในด้านการสื่อสาร (Communication) และการควบคุม (Control)

ในระดับตัวผู้ใช้ระบบ บุคคลต้องการที่จะสื่อสารกับบุคคลอื่น ๆ อย่างมีประสิทธิภาพ เชื่อถือได้ และไม่มีข้อผิดพลาด และในขณะเดียวกันก็ต้องการที่จะรู้สึกรู้ว่าตนเองสามารถควบคุมกระบวนการทำงานของตนเองและของระบบงานได้

หลักการออกแบบส่วนเชื่อมต่อกับผู้ใช้ระบบสามารถอธิบายเป็นข้อ ๆ ตามแนวคิดของ จรณิต แก้วกังวาน (2540) ได้ ดังต่อไปนี้

1. เขียนโปรแกรมโดยคำนึงถึงทุกทางเลือกที่เป็นไปได้ที่ผู้ใช้ระบบจะสั่งดำเนินการ
2. ลดขั้นตอนที่ทำให้ผู้ใช้ระบบต้องศึกษาระบบงานอย่างลำบาก โดยการนำเสนอระบบให้เรียบง่ายและสะดวก ไม่ต้องเสียเวลาเรียนรู้นานเกินไป

3. มีการตอบโต้กับผู้ใช้อย่างสม่ำเสมอ และชัดเจน
4. นำเสนอโดยคำนึงถึงพื้นฐานความรู้ของผู้ใช้ระบบ
5. เปิดโอกาสให้ผู้ใช้แก้ไขข้อผิดพลาดของตนเองได้
6. คำนึงถึงประสบการณ์ของผู้ใช้
7. มีคำตอบโต้ (Feed Back) ในทันทีที่ได้รับคำสั่ง
8. เรียบง่าย (Simple) ระบบต้องดูเป็นธรรมชาติ ไม่ซับซ้อน
9. ตอบโต้ (Responsive) การตอบโต้รวดเร็ว สื่อความหมายตรงตามคำสั่งของผู้ใช้ระบบ

ผู้ใช้ระบบ

10. อธิบาย (Self-documenting) คำสั่งในระบบควรมีลักษณะอธิบายในตนเอง (ผู้ใช้เข้าใจคำสั่งได้โดยอัตโนมัติ) มีคู่มือและบทเรียนจัดไว้เป็นส่วนหนึ่งของระบบ

11. เชื่อมั่น (Reliable) ไม่มีปัญหาในด้านการสื่อสารระหว่างคนกับเครื่อง
12. ขยายแก้ไข (User-modifiable) ผู้ใช้ที่มีความสามารถสูงอาจปรับแต่งแก้ไขสภาพแวดล้อมของระบบให้ตรงกับความต้องการของตนหรือหน่วยงานได้

13. มีประสิทธิภาพ (Power) มีคำสั่งที่ครอบคลุมการทำงานทั้งหมดของผู้ใช้ระบบ และยังเปิดโอกาสให้ผู้ใช้สามารถปรับแต่งแก้ไขระบบได้ตามความต้องการของตนเอง และที่สำคัญระบบจะต้องมีความเชื่อมั่นได้ มีความถูกต้องสมบูรณ์ และมีการปกป้องไม่ให้เกิดข้อผิดพลาด

14. สร้างความพอใจ (User-satisfaction) ความพอใจเป็นคุณละประเด็นกับประสิทธิภาพ ระบบอาจจะทำงานได้ดีเยี่ยม ไม่ทำให้เกิดความสับสนและติดกับในระบบจนหาทางออกไม่ได้

6. หลักการออนโทโลยีของเวลา (Temporal Ontology)

คำว่า Ontology ได้กลายมาเป็นที่รู้จักและใช้กันอย่างแพร่หลายในปัจจุบันและหลาย ๆ ปีที่ผ่านมา เช่น ในสาขาวิชาปัญญาประดิษฐ์ (Artificial Intelligence: AI) สาขาวิชาคอมพิวเตอร์ (Computer) และสาขาวิชาวิทยาศาสตร์สารสนเทศ (Information Science) โดยเฉพาะอย่างยิ่งในขอบเขตสาขาทางด้านระบบสารสนเทศ (Information System) การรวบรวมสารสนเทศทางปัญญา (Intelligence Information Integration) การสืบค้นข้อมูล (Information Retrieval and Extraction) การแทนที่ความรู้ (Knowledge Representation) การออกแบบฐานข้อมูล (Database Design) การจัดการพื้นฐานองค์ความรู้ (Knowledge Base Management) และระบบการจัดการฐานข้อมูล (Database Management System) (Guarino, 1998, Andrade and Saltz, 1999) ในสาขาทั้งหมดที่กล่าวมานั้น ความหมายของ Ontology ก็ยังแตกต่างกันอีกด้วย และไม่มีนิยามเฉพาะตัว ดังนั้น ในหัวข้อนี้เราได้นำเสนอความหมายของ Ontology ที่แตกต่างกันและคำว่า Ontology ที่ได้ถูกกล่าวถึงในวรรณกรรมต่าง ๆ เช่น ในสาขาของ วิชาปรัชญา และสาขาวิชาปัญญาประดิษฐ์ เป็นต้น

ออนโทโลยี (Ontology) คือการเรียนรู้เกี่ยวกับชนิดของสิ่งใด ๆ ที่มีอยู่จริง (Backburn, 1996) ซึ่งอาจจะมีหลายรูปแบบ แต่ที่จำเป็นและเห็นกันอยู่ชัดเจน จะรวมถึงคำศัพท์ของคำหรือสิ่งของ และรายละเอียดที่เฉพาะเจาะจงความหมายของคำหรือสิ่งของนั้น ๆ ซึ่งจะรวมนิยามและระบุว่าแนวคิดมีความเกี่ยวข้องกันอย่างไร (Uschold et al., 1998)

Gruber (1993) ได้นิยามความหมายของ Ontology ซึ่งเป็นที่รู้จักกว้างขวางมากที่สุด คือ Ontology เป็นรายละเอียดที่ชัดเจนแน่นอนของแนวความคิด "An ontology is an explicit specification of a conceptualization" ดังนั้น เมื่อเราต้องการแสดงหรือระบุแนวความคิดของวัตถุหรือสิ่งของนั้น ๆ เราควรที่จะแสดงรายละเอียดที่ชัดเจนและแน่นอน เช่น การอธิบาย Ontology ของโปรแกรม โดยการนิยามกลุ่มของคำที่แสดงความสัมพันธ์ระหว่างกลุ่มคำนั้น ๆ การนิยามนี้อาจจะเกี่ยวข้องกับชื่อของสิ่งที่มีอยู่ในโปรแกรม ซึ่งอาจจะประกอบไปด้วย Classes, Relation, Function หรือ Object อื่น ๆ

ในงานวิจัยนี้หลักการ Ontology ได้ถูกนำมาใช้กับ เวลา (Time) โดยจะถูกเรียกว่า Temporal Ontology หรือ Ontology of Time ซึ่งส่วนใหญ่จะนำมาใช้ในการนิยามเวลา หรือใช้ในการจัดกลุ่มของข้อมูลที่มีความสัมพันธ์กัน โดยใช้เวลาเป็นเกณฑ์ในการวิเคราะห์

Knight (Knight, et al, 1998) ได้แบ่งและอธิบาย Ontology of Time เป็น 3 ประเภท จุด (Points) คือ เวลา ณ เวลาจุดใดจุดหนึ่ง เช่น 08.00 น.

TK
5/05.875
.IS
4/720
2550



19 พ.ย. 2550

ช่วงเวลา (Interval) คือ ช่วงระหว่างเวลาเวลาหนึ่งซึ่งจะไม่นับรวมจุดแรกและจุดสุดท้าย เช่น 08.00 น. ถึง 09.00 น. ความหมาย คือ ตั้งแต่เวลา 08.01 น. ถึง 08.59 น.
จุดและช่วงเวลา (Points and Interval) คือ ช่วงระหว่างเวลาเวลาหนึ่งซึ่งจะนับรวมจุดแรกและจุดสุดท้ายด้วย เช่น 08.00 น. ถึง 09.00 น. ความหมายคือ ตั้งแต่เวลา 08.00 น. ถึง 09.00 น.

Allen (1983) ได้อธิบายความสัมพันธ์ระหว่างช่วงเวลา และนิยามของกลุ่มของความสัมพัทธ์ของเวลาระหว่าง 2 เหตุการณ์ ซึ่งสามารถอธิบายได้โดยการใช้การนิยามขอบเขตของเวลาได้ดังต่อไปนี้ ก่อน (Before), ระหว่าง (During), คาบเกี่ยว (Overlap), เริ่ม (Start), สิ้นสุด (Ends), เท่ากับ (Equal) ตัวอย่างเช่น ก่อนเวลา 18.00 น.

ในงานวิจัยนี้เราได้ใช้ Temporal Ontology มาหาช่วงเวลาที่มีคนใช้อินเตอร์เน็ตมากที่สุด

จักรกฤษณ์ (Snae, 2006) ได้ใช้ Temporal Ontology ในการสร้างกฎความสัมพันธ์สำหรับการค้นหาเชื่อมโยงข้อมูลและระบุความสัมพันธ์ของความเป็นพ่อลูกเข้าด้วยกัน ตัวอย่างของกฎความสัมพันธ์ของ Temporal Ontology สามารถระบุได้ ดังนี้

คนที่จะเป็นพ่อลูกกันได้นั้นนอกจากจะมีนามสกุลเหมือนกันแล้ว จะต้องมีความแตกต่างของอายุ 18 ปี อีกทั้ง Temporal Ontology ยังสามารถใช้ในการระบุคู่สามีภรรยาได้อีกด้วย โดยการใช้กฎ Temporal Ontology ที่ว่าผู้ที่จะเป็นสามีภรรยากันได้นั้นจะต้องมีอายุไม่น้อยกว่า 18 ปี ซึ่งจากกฎความสัมพันธ์นี้เราสามารถสรุปได้ว่าผู้ที่มีอายุต่ำกว่า 18 ปี ไม่สามารถเป็นพ่อได้

7. หลักการสร้างอัลกอริทึมในภาษาสอบถามเชิงโครงสร้าง

การสร้างอัลกอริทึมในภาษาสอบถามเชิงโครงสร้างโดยหลักการ Semantic Query Optimisation (SQO) Andrade และ Saltz ได้นิยามหลักการของ (SQO) ไว้ด้วยกัน 2 ด้าน คือ Semantic และ Data Access Path (DAP) โดย Semantic นั้นจะทำการเข้าถึงข้อมูลโดยใช้ความรู้เกี่ยวกับ Semantic ในการประเมินค่าของเวลาในการเข้าถึงและประมวลผลต่ำกว่า ในส่วนของ Data Access Path (DAP) จะเป็นการใช้ความรู้เกี่ยวกับข้อมูล โดยอยู่ในรูปแบบดัชนี และการเข้าถึงข้อมูลที่มีผลต่อเวลาในการประมวลผล

จุดประสงค์การสร้างอัลกอริทึมในภาษาสอบถามเชิงโครงสร้างที่ดีที่สุดเป็นสิ่งสำคัญอย่างยิ่งในการลดการใช้ทรัพยากร ในที่นี้การสร้างจำนวนการจัดการเข้าออกของการทำงานในหน่วยความจำรอง เพราะการทำงานในส่วนนี้จะทำงานได้ช้ากว่าทรัพยากรตัวอื่น ด้วยเหตุนี้ Semantic Query Optimisation จะต้องมีการวางแผนที่ดีด้วยประสิทธิภาพที่ดี เพื่อจะช่วยให้

เวลาในการเข้าถึงและประมวลผลที่ดีที่สุด (น้อยที่สุด) ทำให้ตัวแบบสอบถามเชิงโครงสร้างทำงานได้ง่ายยิ่งขึ้น จะนำไปสู่การประมวลผลที่น้อยกว่า เพราะตัวแบบสอบถามเชิงโครงสร้างจะตัดส่วนที่ซ้ำซ้อนกันของนิพจน์ออกจากภาษาสอบถามเชิงโครงสร้าง ซึ่งจะช่วยให้ลดเวลาในการประมวลผลและการเข้าถึงข้อมูล

Andrade และ Saltz (1999) ได้เสนอเทคนิค Semantic Query Optimisation และวิเคราะห์ภาษาสอบถามเชิงโครงสร้างให้มีประสิทธิภาพและเกิดประโยชน์มากที่สุดในงานวิจัยนี้ ซึ่ง KESS (Knowledge Engineering SQL Server) จะใช้ Semantic ในรูปแบบข้อมูลของ Ontology โดยใช้เทคนิคการสร้างอัลกอริทึมในภาษาสอบถามเชิงโครงสร้างที่ดีที่สุด ซึ่งอาศัยนิยาม 2 ข้อ ดังต่อไปนี้

นิยามที่ 1 Short Circuit Evaluation (SCE) มีการนิยามเวลา ดังนี้
ตัวอย่าง กำหนดให้ $p(X, Y, w)$ หรือ $p(X, Z, w)$ ส่วน p จะมีความสัมพันธ์ซึ่งกันและกัน นิพจน์ย่อยสามารถลดเป็น $p(X, Y, w)$ Chen and Dunham(1998)

```
select web, user
from Table
where Time >= '9:00' and <='16:00' or
Time like '%11:00'
```

จากอัลกอริทึมแบบสอบถามเชิงโครงสร้างนี้ จะมีการส่งค่ากลับเป็นผู้ใช้และเว็บที่มีการเข้าถึงช่วงเวลา 9.00 น. ถึง 16.00 น. หรือ ณ เวลา 11.00 น. การประยุกต์ใช้เทคนิค Short Circuit Evaluation (SCE) จากนิยามข้างต้นจะพบว่าสามารถ Optimisation อัลกอริทึมของภาษาสอบถามเชิงโครงสร้าง (ตัดนิพจน์ที่ไม่จำเป็นออก) ได้ ดังนี้

```
select web, user
from Table
where Time >= '9:00' and <='16:00'
```

นิยามที่ 2 Constant Folding (CF) มีการนิยามเวลา ดังนี้
ตัวอย่าง กำหนดให้ $p(X, Y)$ หรือ $(X=a)$ หรือ $(Y=b)$ (โดยที่ a, b เป็นค่าคงที่)

```
select web, user
from Table
where Time >= '9:00' and <='16:00' and
Time >= '8:00' and <='17:00'
```

จากอัลกอริทึมแบบสอบถามเชิงโครงสร้างนี้ จะมีการส่งค่ากลับเป็นผู้ใช้และเว็บที่มีการเข้าถึงช่วงเวลา 9.00 น. ถึง 16.00 น. และ ช่วงเวลา 8.00 น. ถึง 17.00 น. การประยุกต์ใช้ Constant Folding (CF) จากนิยามข้างต้นจะพบว่าสามารถ Optimisation อัลกอริทึมของภาษาสอบถามเชิงโครงสร้าง (ตัดสินใจที่ไม่จำเป็นออก) ได้ ดังนี้

```
select web, user
from Table
where Time >= '8:00' and <='17:00'
```

8. หลักสถิติในการวิเคราะห์ข้อมูลการวิจัย

ในการวิเคราะห์ข้อมูลการวิจัยครั้งนี้ มีสถิติที่เกี่ยวข้อง ได้แก่ สถิติเชิงบรรยายและสถิติในการทดสอบสมมติฐาน ซึ่งสถิติเชิงบรรยายที่ใช้ในงานวิจัยนี้ มีสถิติที่เกี่ยวข้อง 3 ตัว ได้แก่ ค่าร้อยละ (Percentage) ค่าเฉลี่ย (Mean หรือ \bar{X}) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation หรือ S.D.) ซึ่งค่าร้อยละใช้บรรยายการแจกแจงข้อมูลอัตราส่วนร้อยละ ค่าเฉลี่ยใช้บรรยายแนวโน้มเข้าสู่ส่วนกลาง และส่วนเบี่ยงเบนมาตรฐานใช้บรรยายความแตกต่างของข้อมูล (เทียมจันทร์ พานิชผลินไชย, ม.ป.ป.) และสถิติในการทดสอบสมมติฐานที่ใช้ในงานวิจัยนี้มีสถิติที่เกี่ยวข้อง 2 ตัว ได้แก่ ค่าสถิติ t-test (t-test for independent samples) และค่าสถิติ F-test (One-way Analysis of Variance) ซึ่งหลักเกณฑ์ในการใช้ค่าสถิติ t-test (t-test for independent samples) นั้น ใช้เพื่อศึกษาความแตกต่างระหว่างกลุ่ม 2 กลุ่มที่เป็นอิสระจากกัน ส่วนหลักเกณฑ์ในการใช้ค่าสถิติ F-test (One-way Analysis of Variance) นั้น ใช้เพื่อศึกษาความแตกต่างระหว่างกลุ่ม มากกว่า 2 กลุ่มที่เป็นอิสระจากกัน

9. งานวิจัยที่เกี่ยวข้อง

การใช้อินเทอร์เน็ตได้รับความนิยมอย่างแพร่หลายและการใช้อินเทอร์เน็ตอย่างต่อเนื่อง การเข้าถึง เครือข่ายอินเทอร์เน็ตถือเป็นเทคโนโลยีที่มีความสำคัญมากในการศึกษาค้นคว้าข้อมูลรวมทั้งบริการต่าง ๆ เช่น เวิลด์ ไรด์ เว็บ (World Wide Web) หรือ เรียกโดยย่อว่า เว็บ ซึ่งความก้าวหน้าทางเทคโนโลยีของเครือข่ายอินเทอร์เน็ต ทำให้ลดข้อจำกัดของเวลาและสถานที่ เช่น การศึกษาข้อมูลทางวิชาการ การใช้จดหมายอิเล็กทรอนิกส์ (E-mail) หรือการประกอบธุรกรรม และมีการใช้ในสถาบันการศึกษาที่มีการค้นคว้าวิจัยเป็นอย่างมาก

พิจิตรา และ ปานใจ (2006) ได้นำเอาเทคนิคการทำเหมืองข้อมูลมาใช้ในการทำนายข้อมูลการเรียกใช้เว็บในอนาคต โดยการนำข้อมูลการใช้เว็บภายในมหาวิทยาลัยศิลปากรมาค้นหา

ความสัมพันธ์ที่ได้จากข้อมูลการใช้เว็บเพื่อใช้ทำนายเว็บที่ถูกเรียกใช้ในอนาคต งานวิจัยนี้ได้นำเทคนิคการค้นหาคำความสัมพันธ์ ซึ่งเป็นเทคนิคหนึ่งในเหมืองข้อมูลมาประยุกต์ใช้ ผลของการใช้เทคนิคเหมืองข้อมูลพบว่า โมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ได้และสามารถเพิ่มอัตราการพบในระบบพรีอิกซ์ได้จากเดิม 38.15% เป็น 54.54% ซึ่งถ้าอัตราการพบในระบบพรีอิกซ์เพิ่มขึ้น อาจทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและสามารถลดปริมาณข้อมูลในระบบเครือข่ายได้ อย่างไรก็ตาม เทคนิคนี้ยังไม่สามารถครอบคลุมการทำงานในช่วงเหตุการณ์ที่ไม่เป็นปกติได้ เช่น อุบัติภัย และเทศกาลต่าง ๆ เป็นต้น

รูปแบบทั่วไปของการค้นหาคำความสัมพันธ์ คือ

A → B

โดยที่ A : เป็นเงื่อนไข

หรือ LHS (Left – Hand Side)

B : เป็นผลลัพธ์ที่เกิดขึ้น

หรือ RHS (Right – Hand Side)

หรืออยู่ในของ "ถ้า.....แล้ว" (If.....Then.....) เช่น

A → B; if A Then B เป็นกฎที่ 1

B → A; if B Then A เป็นกฎที่ 2

การประเมินค่าของกฎจะให้ค่านับสนับสนุน (Support) และค่าความเชื่อมั่น (Confidence) ค่านับสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการ ดังนี้

$$\text{ค่านับสนับสนุน (A,B)} = \frac{\text{จำนวน Transaction (A,B)}}{\text{จำนวน Transaction ทั้งหมด}}$$

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการหาผลลัพธ์

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์

Transaction (A,B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

ค่าความเชื่อมั่น คือ เปรอเซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นเงื่อนไข สามารถเขียนเป็นสมการ ดังนี้

$$\text{ค่าความเชื่อมั่น (A,B)} = \frac{\text{จำนวน Transaction (A,B)}}{\text{จำนวน Transaction (A)}}$$

โดยที่ Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว
 ในการเลือกจากกฎใดนั้นจะต้องพิจารณาค่าสนับสนุนและค่าความเชื่อมั่นที่มีค่าสูงกว่าค่า
 Threshold ที่ตั้งไว้ นอกจากนี้จะต้องกำหนดค่าสนับสนุนต่ำสุด (Minimum Support) และค่า
 ความเชื่อมั่นต่ำสุด (Minimum Confidence) เพื่อนำมาใช้ในการทำนายการใช้เว็บเพจ พบว่า
 สามารถนำมาทำนายการเรียกใช้เว็บเพจได้ การทำนายการเรียกใช้งานเว็บในอนาคตมีประโยชน์
 อย่างยิ่งในการเพิ่ม Hit Rate ของระบบพรีอิกซีได้เป็นอย่างมาก เมื่อ Hit Rate สูงขึ้น เช่น
 ช่วยให้การใช้งานเว็บรวดเร็วขึ้นเพราะเรียกใช้จากพรีอิกซีโดยตรง อันจะเป็นการเพิ่มประสิทธิภาพ
 ของระบบพรีอิกซีและลดปริมาณการใช้งานเครือข่าย ทำให้ภาพรวมของการใช้งานเครือข่าย
 อินเทอร์เน็ตดีขึ้น การทำงานโดยใช้เทคนิคเหมืองข้อมูลในงานวิจัยนี้ความถูกต้องจะขึ้นอยู่กับ
 การสร้างตัวแบบและข้อมูลของพรีอิกซีที่นำมาสร้างความสัมพันธ์

งานวิจัยนี้ได้แบ่งขั้นตอนการดำเนินการวิจัยออกเป็น 3 ขั้นตอน คือ

1. การจัดเตรียมข้อมูล งานวิจัยนี้ได้จัดเก็บข้อมูลการใช้งานเว็บไซต์ของผู้ใช้ภายใน
 มหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ จากเครื่องแม่ข่ายพรีอิกซี ในเดือน
 สิงหาคม กันยายน ตุลาคม พ.ศ.2548 เป็นเวลา 3 เดือนและจัดเก็บฐานข้อมูลการใช้งานเว็บไซต์ และ
 ฐานข้อมูลของหมวดเว็บต่าง ๆ มาสร้างเป็นตัวแบบ ได้แก่ เวลา เนื้อหาเว็บ หมวดของเว็บ

2. การสร้างตัวแบบในงานวิจัยนี้ จะนำข้อมูลมาแสดงความสัมพันธ์ระหว่างเวลา
 หมวดของเว็บและเนื้อหาเว็บ และเมื่อนำข้อมูลทั้งหมดมาหาความสัมพันธ์จะได้ค่าสนับสนุนและ
 ค่าความเชื่อมั่นที่แตกต่างออกต่างจากกันไป

3. การนำตัวแบบที่ได้มาทดสอบและประยุกต์ใช้ข้อมูลและการจำแนกข้อมูลจาก
 ความสัมพันธ์ งานวิจัยนี้ได้นำข้อมูลจากการเรียกใช้เว็บจากเครื่องแม่ข่ายพรีอิกซี มหาวิทยาลัย
 ศิลปากร มาเพื่อทำนายการเรียกใช้เว็บแล้วจำแนกข้อมูลจากความสัมพันธ์ โดยดูลำดับ
 ความสำคัญของเกณฑ์ ดังนี้

1. พิจารณาจากค่าความเชื่อมั่นสูงสุดของแต่ละเงื่อนไข
2. ถ้าความเชื่อมั่นเท่ากัน ให้พิจารณาค่าสนับสนุนที่สูงสุดของแต่ละเงื่อนไข
3. ถ้าค่าความเชื่อมั่นและค่าสนับสนุนมีค่าเท่ากัน ให้พิจารณาจากกฎที่มาก่อนให้มีค่า
 ความสำคัญมากกว่า

ค่าสนับสนุนและค่าความเชื่อมั่นสามารถหาจากสมการ ดังนี้

ค่าสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎ ต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการ ดังนี้

$$\text{ค่าสนับสนุน (เวลา, หมวดเว็บ} \rightarrow \text{เว็บ)} = \frac{\text{จำนวนรายการข้อมูลที่มีเวลา, หมวดเว็บ และเว็บ}}{\text{จำนวนรายการข้อมูลทั้งหมด}}$$

ค่าความเชื่อมั่น คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎ ต่อจำนวนรายการข้อมูลที่เป็นเงื่อนไข สามารถเขียนเป็นสมการ ดังนี้

$$\text{ค่าความเชื่อมั่น (เวลา, หมวดเว็บ} \rightarrow \text{เว็บ)} = \frac{\text{จำนวนรายการข้อมูลที่มีเวลา, หมวดเว็บ และเว็บ}}{\text{จำนวนรายการข้อมูลที่มีเวลา และหมวดเว็บ}}$$

กฎความสัมพันธ์ที่สร้างได้มีจำนวนมากดังนั้นจะต้องตัดความสัมพันธ์บางส่วนออก โดยการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด

ดุษฎีวิทย์ และ ปานใจ (2006) ได้ทำการศึกษาการทำนายปริมาณการจราจรในเครือข่าย โดยใช้เทคนิคเหมืองข้อมูลมาวิเคราะห์ปริมาณการจราจรในระบบเครือข่ายและทำการสร้างตัวแบบ ซึ่งสามารถทำนายแนวโน้มของปริมาณการจราจรเพื่อเป็นแนวทางในการตัดสินใจที่จะเลือกใช้เส้นทางในการจัดส่งข้อมูลต่อไปได้ เทคนิคเหมืองข้อมูลที่น่าสนใจในงานวิจัยนี้คือ เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule Discovery) ในงานวิจัยนี้ได้เก็บข้อมูลในการเข้าใช้อินเตอร์เน็ต จากเครื่องแม่ข่ายหรือที่ ปริมาณการจราจรจะถูกเก็บรวบรวมไว้ทุก ๆ 1 ชั่วโมง และมีระดับ (Rate) แสดงไว้ว่าการจราจรดังกล่าวอยู่ในระดับใด ระดับปริมาณการจราจรข้างต้นสามารถแสดงความสัมพันธ์ระหว่างช่วงเวลากับระดับปริมาณการจราจรได้ ในงานวิจัยนี้ได้แบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลการเรียนรู้และข้อมูลตรวจสอบ โดยที่การแบ่งข้อมูลนั้นใช้วิธีการแยกตัวอย่างแบบสุ่ม และนำข้อมูลที่ได้มาหาความสัมพันธ์ ประสิทธิภาพ และความเร็วของการใช้งานในระบบเครือข่ายขึ้นอยู่กับปัจจัยต่าง ๆ มากมาย ปริมาณการจราจรที่ใช้เส้นทางต่าง ๆ ก็เป็นปัจจัยหนึ่งที่จะส่งผลกระทบต่อการใช้งาน งานวิจัยนี้ทำการศึกษาหาความสัมพันธ์ของปริมาณการจราจรบนเส้นทางของระบบเครือข่าย พบว่า การใช้เทคนิคการค้นหากฎความสัมพันธ์สามารถทำนายแนวโน้มปริมาณการจราจรที่จะใช้ใน

เส้นทางต่าง ๆ ได้ การใช้เทคนิคการค้นหาคำค้นความสัมพันธ์เพื่อทำนายปริมาณการจราจรนี้สามารถนำไปเป็นแนวทางพิจารณาจัดการบริหารการเลือกใช้เส้นทางให้เกิดประสิทธิภาพสูงสุด และสามารถนำไปเป็นปัจจัยหนึ่งในการพัฒนาประสิทธิภาพการรับส่งข้อมูลให้มีประสิทธิภาพสูงขึ้นต่อไปได้

Mohan et al. (2005) ได้เสนอวิธีการใช้พร็อกซีแคช (Proxy Cache) ร่วมกันในการประมวลผลเพื่อให้การเข้าถึงของเว็บได้รวดเร็วขึ้น โดยใช้เทคนิคการค้นหาคำค้นความสัมพันธ์ ซึ่งใช้เทคนิคและวิธีการดังต่อไปนี้ อัลกอริทึมในการจัดกลุ่ม เทคนิคเฝ้าระวังข้อมูล Performance Monitor ในการตั้งค่าแคช โดยจะช่วยให้เพิ่มประสิทธิภาพ ศักยภาพในระบบพร็อกซี โดยจะวิเคราะห์ตามรูปแบบการเข้าถึงเว็บของผู้ใช้เว็บ โดยอัลกอริทึมในการจัดกลุ่มซึ่งอาศัยการค้นหาคำค้นความสัมพันธ์เป็นหลัก โดยอัลกอริทึมนี้ใช้ในการระบุการทำให้เป็นการจัดรูปแบบการเข้าเว็บที่เข้าบ่อย และสามารถอธิบายได้ ดังต่อไปนี้

1. กลุ่มในจำนวนของพร็อกซีแคช ($k+1$) ที่อยู่ในเร้าท์เตอร์เบคโบนของผู้ให้บริการสัญญาณอินเทอร์เน็ต (ISP)
2. จำนวนเครื่องลูกข่าย (Client) จำนวน n ต่อผู้ให้บริการสัญญาณอินเทอร์เน็ต (ISP)
3. ขั้นตอนแรกจะมอบหมายให้ $[n/k]$ ของแต่ละ Proxy Web Cache
4. คอยรอของแคช (Wait for Cache Cycle)
5. กรองข้อมูลเพิ่มบันทึกเข้าออก (Log File) ที่ไม่ต้องการออก
6. เริ่ม Monitor คุณสมบัติของแคช
7. เริ่มกำหนดค่าของ proxy โดยกำหนดค่าให้ตั้งแต่ 1 ถึง k
8. สำหรับ รอบใหม่ของแคชทุก ๆ รอบ
9. ...While Loop
10.ถ้าจำนวนของแคชต่ำกว่า threshold
11.ดั่งนั้น ถ้า $\text{classifying agent last_call_time} < x \text{ ms}$
12.ถ้าเป็นจริงให้เรียก $\text{cache_maintainance_agent}$
13.ถ้าไม่เป็นจริง
14.ให้เรียก $\text{Classification_agent}$
15.อ้างถึงบรรทัดที่ 10 ถ้าไม่เป็นจริง
16. ...ออกจากลูป

Chen et al. (2004) ได้แบ่ง Web Mining ออกเป็น 3 ประเภท คือ

Web Content Mining อ้างถึงการค้นหาข้อมูลที่มีประโยชน์จากสารบรรณเว็บรวมทั้งข้อความ รูป วีดิโอ และสื่อมัลติมีเดียในรูปแบบต่าง ๆ การวิจัยประเภทนี้จะรวมถึงการค้นหาแหล่งข้อมูลจากเว็บ การแยกประเภทและจัดกลุ่มของเอกสารและการดึงข้อมูลของเว็บจากเว็บ

Web Structure Mining จะเป็นการเรียนรู้รูปแบบการเชื่อมโยงโครงสร้างของเว็บ ซึ่งส่วนใหญ่จะเป็นการวิเคราะห์ของการเชื่อมโยงข้อมูลทั้งในและนอกเว็บ โดยวิธีการนี้จะถูกใช้สำหรับจัดลำดับผลลัพธ์ของระบบสืบค้น

Web Usage Mining จะมุ่งเน้นการใช้เทคนิคเหมืองข้อมูลในการวิเคราะห์ข้อมูลที่มี การค้นหาหรือมีการกระทำกิจกรรมบนเว็บ เพื่อที่จะค้นหารูปแบบหรือสิ่งที่น่าสนใจ ซึ่งข้อดีของวิธีการนี้จะช่วยให้ระบบเรียนรู้ประวัติของผู้ใช้

งานวิจัยนี้ได้อธิบายถึงเทคนิค ในการที่จะพัฒนาความสามารถของเครื่องมือ ที่จะช่วย ในการจัดการข้อมูลที่มีอยู่มากมายหลากหลายบนเว็บ ซึ่งข้อมูลดังกล่าวบางครั้งก็ไม่ได้จัดอยู่แบบ เป็นโครงสร้าง ทำให้ยากต่อการดึงนำมาใช้ประโยชน์ วิธีการหนึ่งที่สามารถแก้ปัญหาดังกล่าวได้

Daras et al. (2002) ใช้โมบายอินเทอร์เน็ตเคลื่อนที่ (Mobile Internet) ในการพัฒนา เว็บ ระบบเว็บเบสไร้สายทางการแพทย์ โดยใช้เทคนิคเหมืองข้อมูลเป็นตัวสนับสนุนระบบ ซึ่ง ระบบนี้จะช่วยในการขยายการบริการทางการแพทย์ที่มีอยู่อย่างจำกัดให้เพิ่มมากขึ้น อีกทั้งยังช่วย แก้ปัญหาในการติดต่อระหว่างหมอและคนไข้ให้มีประสิทธิภาพยิ่งขึ้น โดยใช้ภาพและสัญญาณ วีดิโอเป็นสื่อกลางในการตอบโต้และสามารถแลกเปลี่ยนข้อมูลทางการแพทย์ให้มีความยืดหยุ่น มากยิ่งขึ้น ระหว่างหมอและเจ้าหน้าที่ของโรงพยาบาล หรือหมอกับคนไข้

ในงานวิจัยนี้เทคนิคเหมืองข้อมูลจะทำหน้าที่เป็นตัวอัลกอริทึมที่จะค้นหารูปแบบอัตโนมัติ และวิเคราะห์เชิงปฏิสัมพันธ์ ซึ่งเทคนิคที่ใช้มีอยู่ 2 เป้าหมายใหญ่ ๆ ก็คือ การทำนายและ การบรรยายลักษณะ การทำนายนั้นจะเป็นการใช้ระเบียบของฐานข้อมูลเพื่อที่จะคาดเดาค่าของ ตัวแปรที่เราไม่รู้ ส่วนการบรรยายลักษณะจะเน้นการค้นหาแบบที่สามารถแยกประเภทข้อมูล ผ่านเทคนิคการจัดกลุ่มซึ่งในแต่ละระเบียบจะถูกจัดลงในกลุ่มเฉพาะตามเทคนิคการฝึกฝนและ กลุ่มข้อมูลของการฝึกงาน โดยจะใช้การพิจารณาจากขอบเขตค่าความคล้ายคลึงเท่านั้น

Data Mining เทคนิคเหมืองข้อมูลนี้จะช่วยในการวินิจฉัยอาการเจ็บป่วยได้อย่าง ใกล้เคียงและแม่นยำ โดยที่จะนำข้อมูลที่ได้มาเลือกดูว่าคนไข้มีอาการอย่างนี้จะจัดเข้ากลุ่มผู้ป่วย ไต แล้วก็จะจัดกลุ่มให้เล็กลงไปเรื่อย ๆ ก็จะทำให้คำวินิจฉัยใกล้เคียงมากขึ้นเรื่อย ๆ

ส่วนติดต่อ Web/Wap คนใช้สามารถเข้าถึงโปรแกรมโดยใช้คอมพิวเตอร์หรือโทรศัพท์ และคนใช้ก็สามารถส่งข้อมูลของตนเองผ่านทาง WWW หรือผ่านทางโทรศัพท์ และยิ่งกว่านั้นก็จะได้รับคำวินิจฉัยจากแพทย์ผู้เชี่ยวชาญด้วย

Elovici et al. (2004) ได้นำเสนอวัฒนธรรมขององค์ความรู้กับวิธีการในการสืบค้น ผู้ก่ออาชญากรรมบนอินเทอร์เน็ต จากการเรียนรู้ของประวัติและพฤติกรรมของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตโดยใช้อัลกอริทึมเหมืองข้อมูลในการจับข้อความสารบัญจากเว็บที่บุคคลเหล่านี้เข้าถึง ผลลัพธ์ของประวัติผู้ก่ออาชญากรรมบนอินเทอร์เน็ตจะถูกใช้ในการสืบค้นแบบระบบเวลาจริงของกิจกรรมที่ผู้ก่ออาชญากรรมบนอินเทอร์เน็ตได้ทำบนเว็บ

การหาความสัมพันธ์ของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตโดยใช้เทคนิคเหมืองข้อมูลสืบค้นพฤติกรรมของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตตามกิจกรรมบนเว็บ โดยพื้นฐานแล้วกลุ่มผู้ใช้อินเทอร์เน็ตจะเข้ามาใช้ข้อมูลหรือเข้ามาทำกิจกรรมต่าง ๆ บนเว็บไซต์เพื่อเป็นการป้องกันและสืบค้นความเสี่ยงต่าง ๆ บนเว็บไซต์ จึงเป็นเหตุผลสำคัญให้ผู้ให้บริการและตัวแทนจากทั่วโลกมาประชุมและหารือแนวทางในการป้องกันจากผู้ก่ออาชญากรรมบนอินเทอร์เน็ต ทางเดียวที่จะค้นหาการทำกิจกรรมต่าง ๆ บนเว็บไซต์ คือ ยึดถือหมายเลข IP Address เป็นสำคัญ แต่วิธีดังกล่าวเป็นการยากที่จะเข้าถึง IP Address ทั้งหมด ดังนั้น จึงใช้ ISPs Traffic เพื่อระบุ Terrorist บนเว็บไซต์และข้อมูลของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตสนใจเข้าถึงข้อมูล โดยมีตัวพารามิเตอร์เป็นตัวควบคุมเป้าหมายของงานวิจัยนี้ คือหากมีผู้ใช้อินเทอร์เน็ต ที่มีพฤติกรรมของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตที่พยายามจะเจาะข้อมูลสำคัญ หรือมีพฤติกรรมอันไม่สมควรแล้ว ผู้ให้บริการจะส่งข้อความเตือนถึงพฤติกรรมอันไม่เหมาะสม หากผู้ใช้งานดังกล่าวยังทำพฤติกรรมเดิมอยู่อีก ระบบจะทำการป้องกันผู้ใช้งานคนนั้นออกจากระบบทันที และเมื่อผู้ใช้คนใดที่ติดบัญชีดำอยู่ในฐานข้อมูลของผู้ก่ออาชญากรรมบนอินเทอร์เน็ตพยายามจะล็อกอินเข้าสู่ระบบก็จะส่งข้อความเตือนและไม่สามารถล็อกอินได้

Viveros et al. (1996) ได้ประยุกต์เทคนิคเหมืองข้อมูลไปใช้กับระบบสารสนเทศประกันสุขภาพ โดยใช้กฎความสัมพันธ์ (Association Rule) ในการประยุกต์ใช้ในการหาความสัมพันธ์ของฐานข้อมูลจำแนกกลุ่มของผู้ฝึกหัดเพื่อจะทำให้การตรวจสอบรูปแบบการฝึกงานของบริษัทและการบริการของผู้ฝึกงานให้มีประสิทธิภาพยิ่งขึ้น งานวิจัยนี้ได้นำเทคนิคเหมืองข้อมูลสองวิธีมาใช้ โดยรวมให้อยู่ในรูปของฐานข้อมูลขนาดใหญ่ ซึ่งอัลกอริทึมนี้สามารถจัดการกับข้อมูลขนาดใหญ่ ข้อมูลดิบ ที่มากกว่านั้นสามารถแสดงปริมาณของผลกำไร และสิ่งที่น่าสนใจภายในองค์กรได้

Yang et al. (2002) ได้พัฒนาระบบที่สามารถทำนาย เวลาที่ผู้ใช้จะเข้าถึงเว็บได้ โดยวิธีการนี้จะใช้หลักการของกฎความสัมพันธ์ของการจัดกลุ่มข้อมูลด้านเวลาจากแฟ้มบันทึกเข้าออกข้อมูล ซึ่งจะเก็บบันทึกข้อมูลและพฤติกรรมของผู้ใช้อินเทอร์เน็ตไว้มากมาย ในงานวิจัยนี้จะนำเรื่องเวลา (Time) มาใช้ในการทำนายเรื่องการเข้าออกเว็บไซต์โดยใช้เทคนิค กฎความสัมพันธ์ (Association Rule) มาช่วยในการทำนายว่าจะมีผู้ใช้อินเทอร์เน็ตเข้าใช้เว็บเป้าหมาย ณ เวลาใด การสร้างอัลกอริทึมในภาษาสอบถามเชิงโครงสร้าง ในวรรณกรรมที่เกี่ยวข้องกับเหมืองข้อมูลและการจัดกลุ่ม ยังมีการดำเนินงานวิจัยอยู่น้อยมากซึ่งส่วนใหญ่จะใช้กฎความสัมพันธ์หรือพัฒนาอัลกอริทึมในการจัดกลุ่มโดยตรงโดยไม่ผ่านแบบสอบถามเชิงโครงสร้าง

พิจิตรา และ ปานใจ (2006) ได้ นำเทคนิคกฎการค้นหาค่าความสัมพันธ์ ซึ่งเป็นเทคนิคหนึ่งในเหมืองข้อมูลมาประยุกต์ใช้ในการทำนายข้อมูลการเรียกใช้เว็บในอนาคต ดุลยวิทย์ และ ปานใจ (2006) ได้ใช้เทคนิคการค้นหาค่าความสัมพันธ์ (Association Rule Discovery) มาทำการศึกษาการทำนายปริมาณการจราจรในเครือข่ายโดยใช้เทคนิคเหมืองข้อมูลมาวิเคราะห์ปริมาณการจราจรในระบบเครือข่ายและทำการสร้างตัวแบบ ซึ่งสามารถทำนายแนวโน้มของปริมาณการจราจร เพื่อเป็นแนวทางในการตัดสินใจที่จะเลือกใช้เส้นทางในการจัดส่งข้อมูลต่อไป

ซึ่งการใช้เทคนิคของกฎความสัมพันธ์ได้ถูกนำมาพัฒนาในงานวิจัยของ Mohan et al. (2005) ในการจัดกลุ่มข้อมูลของผู้ใช้อินเทอร์เน็ต ซึ่งอัลกอริทึมในการจัดกลุ่มโดยอ้างอิงกฎความสัมพันธ์เป็นหลักการใช้ร่วมกันของพรีอิกซีแคชในการประมวลผลการเข้าถึงของเว็บได้รวดเร็วขึ้น อัลกอริทึมนี้ใช้ในการระบุการทำให้เป็นการจัดรูปแบบการเข้าถึงเว็บที่เข้าบ่อย ในลักษณะเดียวกันนี้ Yang et al. (2002) ได้พัฒนาระบบที่สามารถทำนายเวลาที่ผู้ใช้จะเข้าถึงเว็บ โดยใช้อ้างอิงหลักการใช้โครงสร้างของเวลาในข้อมูลที่อธิบายโดย Zhang (1999, 2000) วิธีนี้เป็นหลักการอย่างง่าย ๆ ในการลำดับเหตุการณ์จัดกลุ่มตามเวลาของพื้นที่นั้น ๆ วิธีการนี้จะใช้หลักการของกฎความสัมพันธ์มาช่วยในการจัดกลุ่มข้อมูลด้านเวลาจากแฟ้มบันทึกเข้าออกข้อมูลอีกด้วย แต่อย่างไรก็ตาม ไม่มีการระบุถึงอนโทโลยีของเวลาเข้ามาเกี่ยวข้อง Viveros et al. (1996) ได้ประยุกต์เทคนิคเหมืองข้อมูลไปใช้กับระบบสารสนเทศประกันสุขภาพ โดยใช้กฎความสัมพันธ์ (Association Rule) ในการประยุกต์ใช้ในการหาความสัมพันธ์ของฐานข้อมูลจำแนกกลุ่มของผู้ฝึกหัดเพื่อจะทำการตรวจสอบรูปแบบการฝึกงานของบริษัทและการบริการของผู้ฝึกงานให้มีประสิทธิภาพยิ่งขึ้น