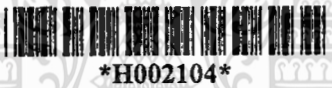
การพัฒนาเครื่องมือสำหรับไมนิ่งข้อมูลของ Proxy log โดยใช้เทคนิค Association rule

Development of Analysis Tool for Mining Proxy Log Using Association Rule

โดย

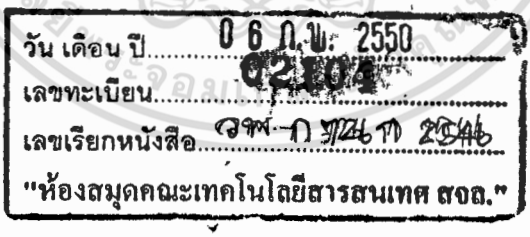
นางสาว กุลธิดา บุญโฉม

รหัส 45066072



อาจารย์ที่ปรึกษา

ผศ.ดร.วรพจน์ กรีสุระเดช



รายงานนี้เป็นส่วนหนึ่งของโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การพัฒนาเครื่องมือสำหรับไมนิ่งข้อมูลของ Proxy log โดยใช้เทคนิค Association rule
นักศึกษา	นางสาว กุลธิดา บุญโฉม
อาจารย์ที่ปรึกษา	ผศ.ดร.วรพจน์ กรีสุระเดช
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

จากการที่ผู้ใช้ท่องไปในเว็บไซต์ต่างๆ โดยข้อมูลจะมีการเก็บรายละเอียดการเข้าถึงเว็บนั้นๆ ไว้ในเว็บ log ของ Proxy Server หรือเรียกว่า Proxy log เนื่องจาก ผู้ใช้ทุกคนที่ร้องขอเว็บไซต์ต่างๆ ทั่วโลกจะถูกเก็บรายละเอียด ใน Proxy log จึงทำให้ Proxy log มีทรานแซ็กชั่นอยู่เป็นจำนวนมาก ซึ่งในโครงการพัฒนาระบบนี้จะวิเคราะห์หาความสัมพันธ์ของข้อมูลการเรียกดูเว็บของผู้ใช้ โดยจะแบ่งการทำงานเป็น 3 ขั้นตอน คือขั้นตอนแรกเป็นการนำข้อมูลดิบใน proxy log มาบันทึกลงฐานข้อมูล ขั้นตอนที่ 2 เป็นการหารูปแบบการเดินทางในการเรียกดูเว็บเพจของผู้ใช้ด้วยเทคนิคการหาเส้นทางไปข้างหน้าไกลที่สุด และขั้นตอนสุดท้ายเป็นการนำรูปแบบเส้นทางเหล่านั้นมาวิเคราะห์หาความสัมพันธ์ในการร้องขอเว็บของผู้ใช้โดยใช้เทคนิคของ Association rule เพื่อนำไปประยุกต์ในการวางแผนทางด้านสารสนเทศขององค์กรต่อไป เช่นการนำกฎที่ได้มาปรับปรุงระบบ cache ของ Proxy server ซึ่งทำให้สามารถพัฒนาประสิทธิภาพของ Proxy server ในการ Access เครือข่ายภายในองค์กรได้อย่างมีประสิทธิภาพสูงสุด

Title	Development of Analysis Tool for Mining Proxy Log Using Association Rule
Student	Miss Kunlathida Boonchome
Advisor	Asst. Prof. Dr. Worapoj Krisuradej
Level of Study	Master of Science in Information Techonlogy
Major	Information Science
Academic Year	2003

Abstract

Due to the huge data of users about navigating through the web site that will be collected in the log file of a proxy server called proxy log. In this paper , we explore data mining capability which involves mining path traversal patterns in a distributed information. Our solution procedure consist of three steps. First we import an the original log data to database ,second we derive an algorithm to convert original sequence of log data into a set of maximal forward reference, Third , we derive algorithms to determine large reference sequence using association rule technique to improve the performance of the proxy server in accessing to the network

กิตติกรรมประกาศ

โครงการพัฒนาเครื่องมือสำหรับไมนิ่งข้อมูลของ Proxy log โดยใช้เทคนิค Association rule จะไม่สามารถดำเนินการมาจนแล้วเสร็จได้ หากขาดความช่วยเหลือของบุคคลเหล่านี้ ข้าพเจ้ามีความรู้สึกขอบคุณทุกท่านที่มีส่วนช่วยเหลือในด้านต่างๆ ด้วยความจริงใจ หากขาดบุคคลที่จะกล่าวถึงดังต่อไปนี้ไป ก็จะส่งผลกระทบต่อความสำเร็จของโครงการศึกษากรณีพิเศษฉบับนี้

ขอขอบพระคุณ บิดา มารดา ที่ให้โอกาสและสนับสนุนทางการศึกษา

ขอขอบพระคุณ ผศ.ดร.วรพจน์ กรีสุระเดช ผู้เป็นอาจารย์ที่ปรึกษาโครงการพัฒนาระบบงานที่กรุณาให้คำปรึกษา แนะนำในด้านต่างๆ

ข้าพเจ้าหวังเป็นอย่างยิ่งว่าบทความนี้จะเป็นแนวคิดในการปฏิบัติงานเพื่อสามารถนำไปใช้ประยุกต์กับงานด้านอื่นๆ ได้เป็นอย่างดี

นางสาวกุลธิดา บุญโฉม

กุมภาพันธ์ 2547

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูปภาพ

หน้า

รูปที่

สารบัญรูปภาพ (ต่อ)

บทที่ 1
บทนำ

จากการที่ผู้ใช้ท่องไปในเว็บไซต์ต่างๆ โดยข้อมูลจะมีการเก็บรายละเอียดการเข้าถึงเว็บนั้นๆ ไว้ในเว็บ log ของ Proxy Server หรือเรียกว่า Proxy log เนื่องจาก ผู้ใช้ทุกคนที่ร้องขอเว็บไซต์ต่างๆ ทั่วโลกจะถูกเก็บรายละเอียด ใน Proxy log จึงทำให้ Proxy log มีทรานแซ็กชั่นอยู่เป็นจำนวนมาก ซึ่งในโครงการพัฒนาระบบนี้จะวิเคราะห์หาความสัมพันธ์ของข้อมูลการเรียกดูเว็บของผู้ใช้ โดยจะแบ่งการทำงานเป็น 3 ขั้นตอน คือขั้นตอนแรกเป็นการนำข้อมูลดิบใน proxy log มาบันทึกลงฐานข้อมูล ขั้นตอนที่ 2 เป็นการหารูปแบบการเดินทางในการเรียกดูเว็บเพจของผู้ใช้ด้วยเทคนิคการหาเส้นทางไปข้างหน้าไกลที่สุด และขั้นตอนสุดท้ายเป็นการนำรูปแบบเส้นทางเหล่านั้นมาวิเคราะห์หาความสัมพันธ์ในการร้องขอเว็บของผู้ใช้โดยใช้เทคนิคของ Association rule เพื่อนำไปประยุกต์ในการวางแผนทางด้านสารสนเทศขององค์กรต่อไป เช่นการนำกฎที่ได้มาปรับปรุงระบบ cache ของ proxy server ซึ่งทำให้สามารถพัฒนาประสิทธิภาพของ Proxy server ในการ Access เครือข่ายภายในองค์กรได้อย่างมีประสิทธิภาพสูงสุด

1.1. วัตถุประสงค์

(1) เพื่อพัฒนาเครื่องมือสำหรับการวิเคราะห์ข้อมูลใน Proxy Server ด้วยเทคนิค Association Rule

(2) เพื่อพัฒนาเครื่องมือสำหรับหารูปแบบเส้นทางการเดินทางการเรียกดูเว็บ (Path Traversal Patterns) ซึ่งเป็นพฤติกรรมการเรียกดูเว็บของผู้ใช้

(3) เพื่อเพิ่มความเข้าใจในการออกแบบและประยุกต์ใช้เทคโนโลยีเว็บไมนิ่ง

(4) เพื่อเป็นแนวทางในการออกแบบและประยุกต์ใช้เทคโนโลยีเว็บไมนิ่ง

(5) สามารถนำผลลัพธ์ที่ได้นำมาช่วยในการวิเคราะห์พฤติกรรมในการเข้าถึงเว็บของผู้ใช้เพื่อปรับปรุงประสิทธิภาพของ Proxy Server ต่อไป

1.2. แนวทางการศึกษา

(1) ศึกษาแนวทางการนำหลักการดาต้าไมนิ่งมาประยุกต์ใช้กับการวิเคราะห์ข้อมูลที่ได้จากการเก็บประวัติการเข้าถึงเว็บ

(2) กำหนดขอบเขตของการทำงาน

(3) ศึกษาอัลกอริทึมในการทำเว็บไมนิ่งวิเคราะห์ความเป็นไปได้และพิจารณาตัดสินใจเลือกอัลกอริทึมที่เหมาะสมในการทำงาน

(4) พัฒนาโปรแกรมโดยใช้เครื่องมือดังนี้

a. ระบบจัดการฐานข้อมูลใช้ MS SQL Server 2000 Enterprise Edition

b. เครื่องมือสำหรับพัฒนาโปรแกรมและสร้างรายงานใช้ MS Visual Studio .Net 2003

1.3 ขอบเขตของการพัฒนาระบบ

พัฒนาเครื่องมือสำหรับการวิเคราะห์ข้อมูลใน Proxy Log ด้วยเทคนิค Association Rule ซึ่งสามารถแบ่งออกเป็นระบบย่อยได้ดังต่อไปนี้

(1) ระบบย่อยรับข้อมูลจาก Proxy Log บันทึกลงฐานข้อมูลซึ่งต้องมีรูปแบบตามที่กำหนดไว้เท่านั้น

(2) ระบบย่อยสำหรับหารูปแบบเส้นทางการเดินทางภายในเว็บ

(3) ระบบย่อยสำหรับการวิเคราะห์ข้อมูลรูปแบบเส้นทางการเดินทางภายในเว็บของผู้ใช้ โดยใช้เทคนิค Association rule ซึ่งในระบบนี้จะใช้อัลกอริทึม Apriori

1.4. ประโยชน์ที่คาดว่าจะได้รับ

(1) ได้เครื่องมือสำหรับวิเคราะห์หารูปแบบเส้นทางการเดินทางภายในเว็บที่สามารถนำไปใช้งานได้จริง

(2) ได้เครื่องมือสำหรับวิเคราะห์ความสัมพันธ์ของพฤติกรรมการเข้าถึงเว็บของผู้ใช้ด้วยเทคนิค Association rule

(3) สามารถนำเอาข้อมูลสารสนเทศที่มีอยู่แล้วในองค์กรมาใช้ประโยชน์ได้มากขึ้น

(4) สามารถนำผลลัพธ์ที่ได้จากการวิเคราะห์มาเพิ่มประสิทธิภาพของ Proxy server ภายในองค์กร

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ดาต้าไมนิ่ง

2.1.1 ความสำคัญของดาต้าไมนิ่ง

ความสำคัญในการรวบรวมข้อมูลในการดำเนินธุรกิจเป็นสิ่งที่ควรให้ความสำคัญอย่างยิ่ง เนื่องจากเป็นองค์ประกอบสำคัญที่จะทำให้ธุรกิจบรรลุความสำเร็จ ซึ่งในการที่จะเปลี่ยนข้อมูลจำนวนมากมาเป็นกุญแจแห่งความสำเร็จ จำเป็นต้องหาวิธีที่จะดึงความรู้เหล่านั้นออกมาจากข้อมูลที่เก็บรวบรวมไว้ซึ่งมีอยู่จำนวนมาก ซึ่งวิธีการที่กล่าวถึงคือหลักการของดาต้าไมนิ่ง

ดาต้าไมนิ่งเป็นกระบวนการในการค้นหาสารสนเทศที่มีประโยชน์จากฐานข้อมูลที่มีอยู่ ซึ่งการนำเอาดาต้าไมนิ่งมาใช้ในการค้นหาสารสนเทศจากฐานข้อมูลนั้นเนื่องมาจากว่า

1) ปัจจุบันในการดำเนินการทางธุรกิจทั้งองค์กรขนาดเล็ก ขนาดกลาง และขนาดใหญ่ ซึ่งจะต้องใช้ฐานข้อมูลมาเก็บรวบรวมข้อมูลเกี่ยวกับกิจกรรมต่างๆที่เกิดขึ้น ซึ่งในแต่ละวันข้อมูลที่เก็บก็จะเพิ่มขึ้นเรื่อยๆ และจากข้อมูลทั้งหมดที่เก็บอยู่ในฐานข้อมูลอาจจะมีลักษณะบางอย่างแอบแฝงอยู่ เช่น ความสัมพันธ์ แนวโน้ม(Trend) หรือรูปแบบเฉพาะ(Pattern) ซึ่งสารสนเทศเหล่านี้มีประโยชน์อย่างมากในการตัดสินใจเชิงธุรกิจ เพื่อให้องค์กรมีโอกาสในการแข่งขันมากยิ่งขึ้น แต่เนื่องจากการที่จะวิเคราะห์หาประโยชน์จากฐานข้อมูลขนาดใหญ่นี้จำเป็นต้องใช้หลักการดาต้าไมนิ่งช่วยในการวิเคราะห์ให้ได้สารสนเทศที่ถูกต้อง และสามารถนำไปใช้ประโยชน์ได้จริงในการดำเนินธุรกิจ

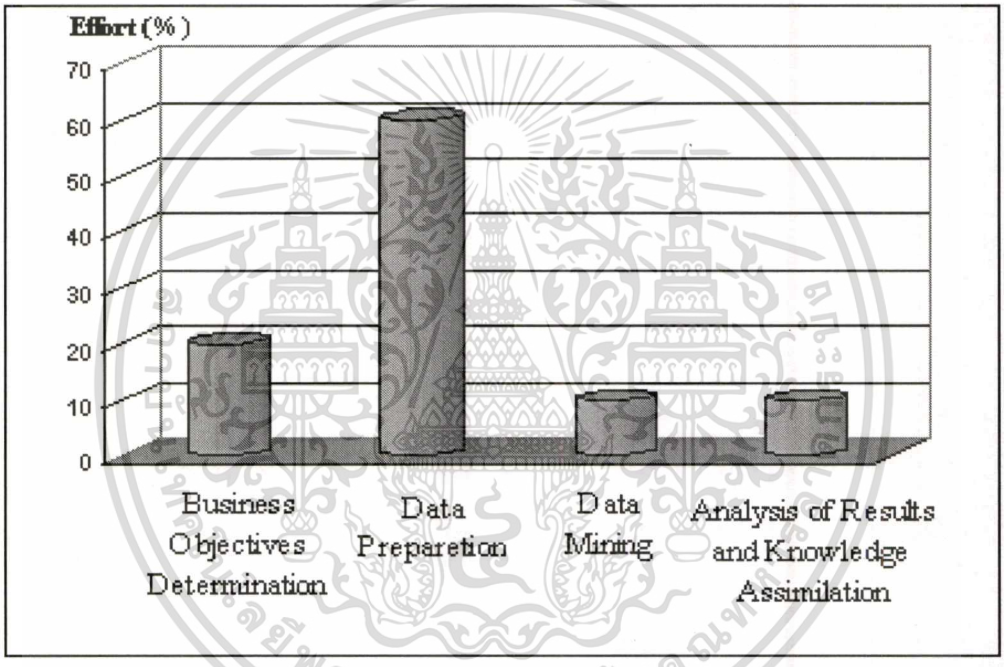
2) จากการที่มีข้อมูลจำนวนมากในฐานข้อมูลอีกทั้งข้อมูลเหล่านั้นเป็นข้อมูลที่ซับซ้อน ทำให้การวิเคราะห์ข้อมูลให้ถูกต้องนั้นยากเกินกว่าที่จะวิเคราะห์โดยมนุษย์ได้ ซึ่งมนุษย์อาจจะวิเคราะห์ข้อมูลเหล่านั้นได้ไม่ครอบคลุม หรือมองข้ามตัวแปรบางตัวที่มีผลกับผลลัพธ์ที่จะเกิดขึ้น แต่ถ้าใช้หลักการดาต้าไมนิ่งในการวิเคราะห์สามารถมั่นใจได้ว่าการวิเคราะห์นั้นถูกต้อง ครอบคลุมทุกตัวแปร

3) เมื่อเปรียบเทียบกันระหว่างองค์กรที่จ้างผู้เชี่ยวชาญหลายๆคนกับองค์กรที่พัฒนาเครื่องมือโดยใช้หลักการดาต้าไมนิ่งมาใช้วิเคราะห์ข้อมูลจำนวนมากและซับซ้อนนั้นจะเห็นว่าค่าใช้จ่ายในการจ้างผู้เชี่ยวชาญหลายๆคนอาจจะสูงกว่าการพัฒนาเครื่องมือดาต้าไมนิ่งก็ได้รวมทั้งการวิเคราะห์โดยมนุษย์จะใช้เวลานานกว่าการวิเคราะห์ของเครื่องมือดาต้าไมนิ่งมาก

2.1.2 กระบวนการของดาต้าไมนิ่ง

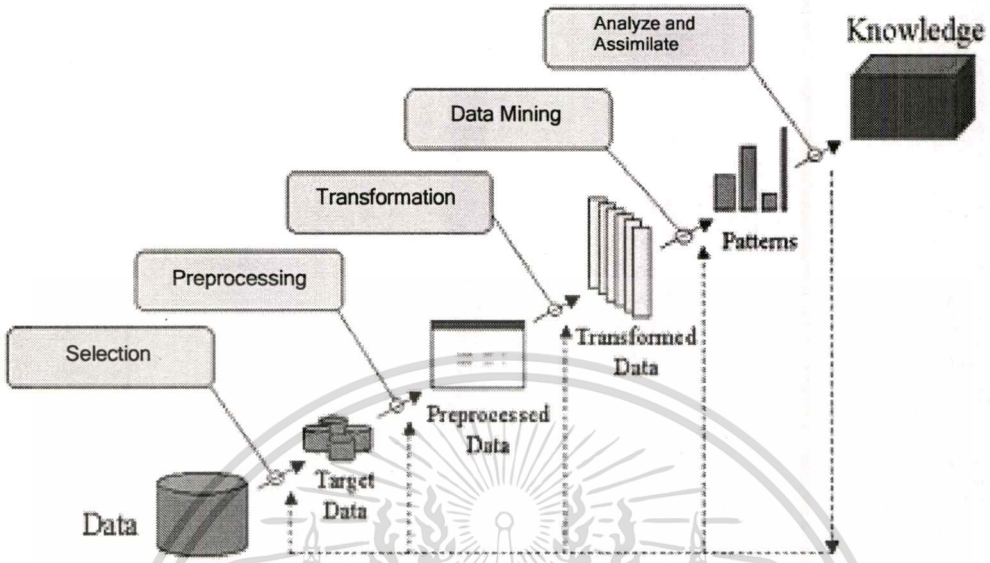
ถ้ากล่าวถึงดาต้าไมนิ่งส่วนใหญ่จะให้ความสำคัญกับการไมนิ่งข้อมูล แต่ที่จริงแล้วการไมนิ่งข้อมูลเป็นเพียงขั้นตอนหนึ่งของกระบวนการดาต้าไมนิ่งเท่านั้น

โดยขั้นตอนหลักในการทำดาต้าไมนิ่งมี 5 ขั้นตอน ดังรูปที่ 2.1 ซึ่งจะแสดงเปอร์เซ็นต์ในการทำงานในแต่ละขั้นตอนด้วยดังนี้



รูปที่ 2.1 แสดงขั้นตอนของกระบวนการทำดาต้าไมนิ่ง[1]

จากรูปที่ 1 มี 5 ขั้นตอนหลัก และสามารถแบ่งการทำงานออกเป็นขั้นตอนย่อยได้ดังรูปที่ 2.2 โดยสามารถอธิบายในแต่ละขั้นตอนได้ดังนี้



รูปที่ 2.2 แสดงขั้นตอนการทำงานของดาต้าไมนิ่ง[1]

ขั้นตอนที่ 1 : กำหนดวัตถุประสงค์ทางธุรกิจ (Business Objectives Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจจะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจ เพราะจะเป็นตัวกำหนดทิศทางในการทำดาต้าไมนิ่งและสามารถกำหนดได้ว่าเมื่อไหร่จะใช้ดาต้าไมนิ่งในการแก้ปัญหา เนื่องจากในทุกปัญหาไม่สามารถแก้ไขได้ด้วยหลักการดาต้าไมนิ่งทั้งหมด ซึ่งในส่วนนี้จะประกอบด้วยการวิเคราะห์ทางธุรกิจและวิเคราะห์ข้อมูลเบื้องต้นว่าเรามีข้อมูลอะไรอยู่และต้องการอะไรจากข้อมูล ปัญหามีขอบเขตแค่ไหน ซึ่งในขั้นตอนนี้จะสามารถมองถึงอัลกอริทึมและฐานข้อมูลที่จะใช้เบื้องต้นที่สัมพันธ์กับวัตถุประสงค์ทางธุรกิจได้ การกำหนดวัตถุประสงค์ทางธุรกิจนี้จะใช้เวลาประมาณ 20 % ของการทำดาต้าไมนิ่ง

ขั้นตอนที่ 2 : การเตรียมข้อมูล (Data preparation)

การเตรียมข้อมูลเป็นขั้นตอนที่ใช้เวลานานที่สุดประมาณ 60 % ของการทำดาต้าไมนิ่งเพราะเป็นส่วนสำคัญที่สุดในการทำดาต้าไมนิ่ง เนื่องจากบางครั้งอาจมีการนำข้อมูลจากหลายแหล่งมารวมกันเพื่อดูความสัมพันธ์ของข้อมูล ข้อมูลที่ได้จากขั้นตอนนี้จะต้องมีความชัดเจน ถูกต้อง เหมาะสม โดยขั้นตอนการเตรียมข้อมูลจะแบ่งการทำงานเป็น 3 ขั้นตอนย่อย ดังนี้

1. การเลือกข้อมูล (Data Selection)

ข้อมูลก่อนทำดาต้าไมนิ่งอาจจะมาจากข้อมูลหลายๆแหล่งรวมกันจึงต้องมีการเลือกข้อมูลที่สำคัญออกมา วัตถุประสงค์ของการเลือกข้อมูลคือ การระบุลักษณะข้อมูล , เลือกข้อมูลที่ต้องการ และนำข้อมูลที่ไม่ต้องการออกไป ซึ่งการเลือกข้อมูลจะขึ้นอยู่กับวัตถุประสงค์ของแต่ละธุรกิจ การเลือกข้อมูลจำเป็นต้องมีความเข้าใจกับชนิดของข้อมูล ค่าที่เป็นไปได้ แหล่งกำเนิดของข้อมูล รูปแบบของข้อมูลและลักษณะอื่นๆ ซึ่งเราสามารถแบ่งลักษณะและรูปแบบของข้อมูลเป็น 2 ลักษณะคือ

(1) ตัวแปรแบบ Categorical

- Nominal : เป็นตัวแปรที่ลำดับของข้อมูลไม่มีผลกับค่า เช่น สถานะการแต่งงาน (single , married , divorced)
- Ordinal : เป็นตัวแปรที่ลำดับของข้อมูลมีผลกับค่า เช่น เกรดของนักศึกษา (A , B , C , D , F) หรือลำดับคุณภาพของสินค้า (ดี , ปานกลาง , เลว)

(2) ตัวแปรแบบ Quantitative

- Continuous : ค่าที่เก็บจะเป็นจำนวนจริง (real number) หรือเป็นค่าต่อเนื่อง เช่น น้ำหนักของนักศึกษา
- Discrete : ค่าที่เก็บเป็นเลขจำนวนเต็ม (Integer) เช่น จำนวนนักศึกษา

2. การกลั่นกรองข้อมูล (Data preprocessing)

เมื่อเลือกข้อมูลที่ต้องการแล้วจะต้องกลั่นกรองข้อมูลเพื่อให้ข้อมูลที่เลือกมามีคุณภาพเหมาะกับจะนำไปทำดาต้าไมนิ่งเนื่องจากข้อมูลที่เลือกมาอาจมีค่าที่ไม่ถูกต้อง โดยประเด็นที่ต้องพิจารณาเพิ่มเติม 2 ประเด็น คือ

1) Noisy Data เป็นข้อมูลที่มีลักษณะต่างจากข้อมูลที่คาดการณ์ไว้ ซึ่งอาจจะเป็นข้อมูลในแง่ดี คือ สิ่งที่เรามองหาอยู่ หรือเป็นข้อมูลในแง่ร้าย คือเป็นข้อมูลที่ไม่สมบูรณ์ อาจเกิดจากการบันทึกข้อมูลผิดพลาด หรือความผิดพลาดในการรับส่งข้อมูลผ่าน network ดังนั้นจะต้องจัดการกับข้อมูลเหล่านี้ก่อน โดยตัดข้อมูลเหล่านี้ทิ้งถ้ามีจำนวนน้อย หรือใช้เทคนิคทางสถิติแก้ไขให้เหมาะสมก่อนนำมาวิเคาะห์

2) Missing Data เป็นข้อมูลที่มีลักษณะบางส่วนของข้อมูลขาดหายไป สาเหตุอาจเกิดคล้ายกับ Noisy Data แก้ไขโดยการตัดข้อมูลเหล่านี้ทิ้งถ้ามีจำนวนน้อย หรือ ถ้าข้อมูลที่ขาดหายมีมากก็ให้แก้ไขโดยถ้าเป็นข้อมูล Quantitative ก็ให้แทนข้อมูลเหล่านี้ด้วยค่าเฉลี่ยหรือค่าที่ปรากฏบ่อย ถ้าเป็นข้อมูลแบบ Categorical อาจแทนด้วยค่าที่ปรากฏบ่อยหรือแทนด้วย Unknown

3. การแปลงข้อมูล (Data Transformation)

เป็นการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการทำดาต้าไมนิ่งเช่น การแปลง Quantitative เป็น Categorical เช่นแปลงเกรดของนักศึกษาจาก A , B เป็น high แปลง C เป็น medium แปลง D , F เป็น low หรือการแปลงจาก Categorical เป็น Quantitative เช่น รถยนต์ แปลงเป็น 01 รถจักรยานยนต์ แปลงเป็น 02

ขั้นตอนที่ 3 : การทำดาต้าไมนิ่ง(Data Mining)

เป็นขั้นตอนการประมวลผลข้อมูลตามอัลกอริทึมที่กำหนดไว้ โดยถ้าผลลัพธ์ที่ได้จากขั้นตอนนี้ไม่เป็นไปอย่างที่คาดหวังก็อาจย้อนกลับไปทำขั้นตอนกลั่นกรองข้อมูลใหม่ได้ ในขั้นตอนนี้จะเกี่ยวข้องกับการเลือกอัลกอริทึมในการทำดาต้าไมนิ่งซึ่งจะต้องพิจารณาลักษณะของปัญหาเป็นหลัก เพราะในแต่ละปัญหาต้องเลือกใช้อัลกอริทึมที่เหมาะสมจึงจะได้ผลการวิเคราะห์ที่ถูกต้อง ซึ่งอาจจะใช้หลายอัลกอริทึมเพื่อเปรียบเทียบผลลัพธ์ได้

ขั้นตอนที่ 4 : การวิเคราะห์ผลลัพธ์ที่ได้จากการทำดาต้าไมนิ่ง(Analysis of Results)

เป็นการวิเคราะห์ผลของการประมวลผล ซึ่งจะทำการแปลความหมายและประเมินผลที่ได้จากการทำไมนิ่ง โดยที่เมื่อเราได้รูปแบบของความสัมพันธ์หรือโมเดลแล้ว ขั้นตอนนี้จะเป็นการทดสอบโมเดลที่ได้กับชุดข้อมูลอีกชุดหนึ่งที่เราทราบผลลัพธ์อยู่แล้ว ว่าเมื่อทดสอบกับโมเดลแล้วได้ผลลัพธ์ถูกต้องหรือยอมรับได้หรือไม่ หากยอมรับไม่ได้ก็อาจจะแก้ไขโดยเพิ่มจำนวนข้อมูลให้มากขึ้น หรือเปลี่ยนอัลกอริทึมใหม่

ขั้นตอนที่ 5 : การปรับความรู้ที่ได้เข้ากับธุรกิจ (Assimilation of Knowledge)

เป็นการรวบรวมความเข้าใจทางธุรกิจที่มีผลจากขั้นตอนการวิเคราะห์ผลลัพธ์เพื่อนำไปใช้ในโอกาสต่อไป โดยขั้นตอนนี้จะมีหลักอยู่ 2 ประการ คือ

1) การนำเสนอแนวคิดที่ค้นพบใหม่
2) การหาแนวทางที่จะนำความรู้ที่ค้นพบใหม่ไปใช้ให้เกิดประโยชน์สูงสุดต่อไป

2.1.3 โอเปอเรชั่นของดาต้าไมนิ่ง(Data Mining Operation)

ในการประยุกต์ใช้ดาต้าไมนิ่งในงานทางธุรกิจจะประกอบด้วย 4 โอเปอเรชั่น ดังนี้

1. Predictive Modeling

เป็นโมเดลที่คล้ายกับการเรียนรู้ของมนุษย์คือต้องเข้าใจลักษณะของสิ่งที่จะศึกษาอย่างแท้จริง ซึ่งจะนำโมเดลนี้มาใช้การวิเคราะห์ฐานข้อมูลที่มีอยู่ โดยการสร้างโมเดลจะนำข้อมูลในอดีตมาสร้างแบบจำลองพยากรณ์ มี 2 ขั้นตอน คือ

1) Training Phase เพื่อสร้างโมเดลขึ้นมาใหม่โดยใช้ข้อมูลในอดีต

2) Testing Phase เพื่อทดสอบโมเดลที่สร้างว่าเหมาะสมหรือไม่ โดยใช้ข้อมูลอีกชุดหนึ่งที่ทราบผลลัพธ์อยู่แล้ว

การสร้างแบบจำลองพยากรณ์ประกอบด้วย 2 ลักษณะ คือ

1) **Classification** เป็นลักษณะการสร้างโมเดลเพื่อทำนายกลุ่มของข้อมูลที่สนใจ เช่นทำนายว่าลูกค้าใดควรส่งจดหมายแนะนำสินค้าให้โดยดูจากพฤติกรรมการซื้อ

2) Value Prediction เป็นการทำนายค่าตัวเลขที่สัมพันธ์กับข้อมูลที่มีอยู่ ใช้เพื่อทำนายค่าของเหตุการณ์ในอนาคต เช่น ช่วงเวลาที่ลูกค้าจะซื้อรถ โดยดูจากรายได้ อายุ เพศ เป็นต้น

2. Database Segmentation

เป็นการแบ่งข้อมูลที่มีลักษณะเหมือนกันในฐานข้อมูลออกเป็นกลุ่มที่มีคุณสมบัติเหมือนกัน เพื่อให้ง่ายต่อการวิเคราะห์ เช่นแบ่งกลุ่มนักศึกษาตามเพศ อายุ

3. Link Analysis

เป็นการวิเคราะห์ว่าข้อมูลแต่ละรายการมีความสัมพันธ์กันหรือไม่ อย่างไรประกอบด้วย 3 ลักษณะ คือ

1) Association Discovery เป็นการค้นหาความสัมพันธ์ของข้อมูล เช่นพบว่าลูกค้าที่ซื้อขนมปังจะซื้อนมข้นด้วย

2) Sequential Pattern Discovery เป็นความสัมพันธ์ของข้อมูลที่เกี่ยวข้องกับลำดับเหตุการณ์ เพื่อทำความเข้าใจพฤติกรรมในระยะยาว เช่น ลูกค้าที่ซื้อโทรทัศน์มีแนวโน้มที่จะซื้อวีดีโอในเวลาต่อมา

3) Similar Time Sequence Discovery เป็นความสัมพันธ์ระหว่างข้อมูล 2 กลุ่มซึ่งขึ้นต่อกันทางด้านเวลา โดยนำรูปแบบความสัมพันธ์นั้นที่เวลาเดียวกันมาเปรียบเทียบเพื่อหารูปแบบหลักๆไว้ใช้ในการทำนายในอนาคตต่อไป

4. Deviation Detection

เป็นโมเดลที่พยายามหาค่าที่แตกต่างไปจากมาตรฐาน ซึ่งจะเห็นถึงข้อผิดพลาดหรือส่วนที่ไม่เกี่ยวข้องประกอบด้วย 2 ลักษณะ คือ

1) Statistics ใช้หลักการทางสถิติในการวิเคราะห์หาความแตกต่างนั้น

2) Visualization ใช้ภาพแสดงให้เห็นของความแตกต่าง เช่นการวาดกราฟ

โอเปอเรชั่นนี้สามารถใช้ได้กับการตรวจหาจุดบกพร่องของชิ้นงาน การตรวจสอบลายเซ็นปลอม หรือบัตรเครดิตปลอม

โมเดลนี้นำไปประยุกต์ใช้ในงานด้านธุรกิจได้ดังตารางที่ 2.1 แต่จะไม่เจาะจงได้ว่าธุรกิจประเภทใดต้องใช้โมเดลไหน เพียงแต่บอกว่าลักษณะงานทางธุรกิจใดมีความเกี่ยวข้องกันและลักษณะงานแบบไหนควรใช้โมเดลแบบใด

ตารางที่ 2.1 แสดงการประยุกต์ใช้โมเดลกับงานทางธุรกิจ

Market Management	Risk Management	Fraud Management
Target Marketing	Forecasting	Fraud detection
Customer Relationship	Customer retention	
Market basket analysis	Improved underwriting	
Cross selling	Quality control	
Market segmentation	Competitive analysis	

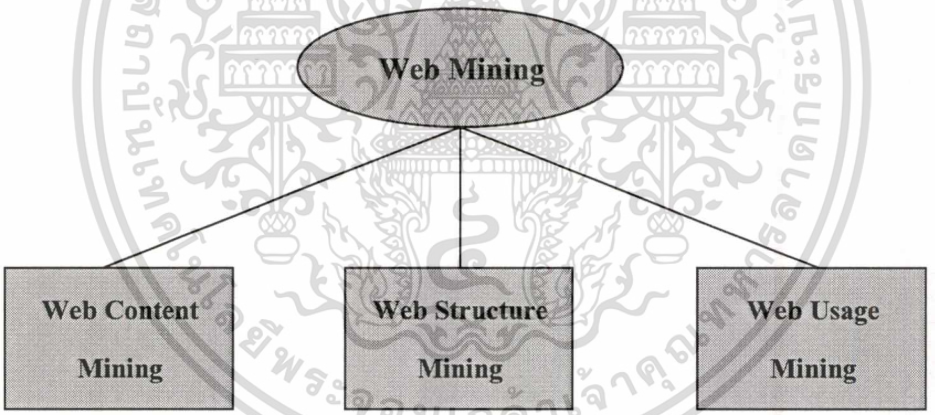
Predictive Modeling	Database Segmentation	Link Analysis	Deviation Detection
Classification	Demographic	Association discovery	Visualization
Value prediction	clustering	Sequential pattern discovery	Statistics
	Neural clustering	Similar time sequence discovery	

จากที่กล่าวมาแล้วข้างต้นว่าโอเปอเรชั่นของดาต้าไมนิ่งมีมากมาย สำหรับโครงการที่นำมาเสนอนี้ จะนำเสนอโอเปอเรชั่นของ Link Analysis โดยการใช้ Association Rule มาใช้เพื่อหาความสัมพันธ์ของข้อมูลการเข้าถึงเว็บเพจของผู้ใช้ในองค์กร โดยจะกล่าวถึงรายละเอียดในบทถัดไป

2.2 เว็บไมนิ่ง (Web Mining)

2.2.1 ความสำคัญของเว็บไมนิ่ง

ปัจจุบันข้อมูลบนเวิลด์ไวด์เว็บมีจำนวนเพิ่มขึ้นอย่างมากหลากหลายชนิด มีทั้งข่าว โฆษณาบันเทิง รัฐบาลและบริการทางข้อมูลอีกมากมาย ซึ่งเว็บไมนิ่งนี้เป็นเทคนิคในการค้นหาและวิเคราะห์ข้อมูลสารสนเทศที่มีประโยชน์จากเวิลด์ไวด์เว็บสามารถแบ่งลักษณะของเว็บไมนิ่งได้ 3 ประเภทด้วยกันดังรูปที่2.3



รูปที่ 2.3 แสดงประเภทของเว็บไมนิ่ง

ประเภทของเว็บไมนิ่งมี 3 ประเภท

1) Web Content Mining

คือการไมนิ่งเนื้อหาในเว็บ เนื่องจากข้อมูลสารสนเทศที่ได้จากเวิลด์ไวด์เว็บไม่มีโครงสร้างที่แน่นอนทำให้การค้นหาข้อมูลสารสนเทศทำได้ยาก เครื่องมือค้นหาต่างๆ เช่น google , lycos , alta vista และอื่นๆนั้นได้อำนวยความสะดวกให้กับผู้ใช้ในการค้นหาเท่านั้น แต่ก็ไม่ได้เตรียมข้อมูลสารสนเทศที่

เป็นโครงสร้างไปมากกว่าการแบ่งกลุ่ม (Categorize) การกรอง (Filter) หรือการตีความเอกสาร ทำให้ประสิทธิภาพของการค้นหาไม่ดีเท่าที่ ผลลัพธ์ที่ได้ออกมาจากการค้นหายังมีมากเกินไป

ในหลายปีที่ผ่านมานี้ได้มีนักพัฒนาสร้างเครื่องมือที่ฉลาดขึ้นเพื่อให้ได้มาซึ่งข้อมูลสารสนเทศที่มีประโยชน์ เครื่องมือดังกล่างเช่น ตัวแทนเว็บที่ฉลาด (Intelligent web agent) ซึ่งจะมีอยู่ 2 วิธีที่จะเพิ่มความสามารถในการค้นหาข้อมูลกึ่งโครงสร้างในเว็บดังนี้

1.1) Agent-based approach

วิธีที่ใช้ตัวแทน แบ่งเป็น 3 กลุ่ม ดังนี้

1.1.1) Intelligent search agent เป็นตัวแทนการค้นหาที่จะใช้คุณลักษณะของโดเมนและรูปแบบของผู้ใช้ (User profile) เพื่อจัดการและตีความสารสนเทศที่ต้องการ

1.1.2) Information filtering/categorization เป็นตัวแทนการค้นหาที่ใช้ลักษณะของการกรองและจัดกลุ่มของข้อมูลอย่างอัตโนมัติ

1.1.3) Personalized web agent เป็นตัวแทนการค้นหาโดยจะเรียนรู้ความสนใจของผู้ใช้ก่อนและค้นหาแหล่งข้อมูลสารสนเทศตามความสนใจเหล่านั้น

1.2) Database approach

วิธีการทางฐานข้อมูลจะจัดการข้อมูลที่มีลักษณะกึ่งโครงสร้างโดยใช้เทคนิคในการถาม (Query) จากฐานข้อมูลและใช้เทคนิคดาต้าไมนิ่งในการวิเคราะห์ผลลัพธ์ที่ได้มา

2) Web Structure Mining

คือการไมนิ่งลิงค์ของเว็บไซต์ต่างๆ เพื่อที่จะหาความเกี่ยวข้องกันของเนื้อหา

3) Web Usage Mining

เนื่องจากเวิลด์ไวด์เว็บมีการเติบโตขึ้นอย่างต่อเนื่องทั้งในด้านปริมาณของการขนส่งข้อมูล ขนาดและความซับซ้อนของเว็บไซต์ การออกแบบเว็บไซต์จึงเป็นงานที่สำคัญที่จะทำให้ผู้ที่ต้องการค้นหาสารสนเทศทำได้ง่ายขึ้น ดังนั้นการออกแบบนี้จะต้องวิเคราะห์ว่าเว็บไซต์ถูกใช้อย่างไร ซึ่งสามารถวิเคราะห์ได้โดยการใช้หลักการทางสถิติ เช่นความถี่ในการเข้าถึงไซต์ ข้อมูลสารสนเทศของการใช้งานสามารถนำมาปรับโครงสร้างเว็บไซต์ใหม่เพื่อให้รองรับความต้องการของผู้ใช้ได้ดียิ่งขึ้น เส้นทางการเดินทางที่ผิดไปจากที่กำหนดไว้หรือเพจที่มีความสำคัญถูกเรียกใช้งานน้อยแสดงให้เห็นว่าลิงค์ของไซต์และข้อมูลสารสนเทศไม่ได้ถูกออกแบบให้ดีเท่าที่ควร แต่เมื่อเราทราบรูปแบบการเดินทางของการค้นหาข้อมูลของผู้ใช้ ก็สามารถนำมาออกแบบเว็บไซต์ใหม่ให้ดียิ่งขึ้น

บัจจุบันระบบเว็บและเทคนิคที่ใช้ในการค้นหาและวิเคราะห์รูปแบบมีความซับซ้อนมากยิ่งขึ้น จึงได้เกิดเครื่องมือที่ช่วยในการวิเคราะห์ข้อมูลสารสนเทศขึ้นแบ่งเป็น 2 กลุ่มคือ

1.1) Pattern discovery tool

เป็นเครื่องมือที่ใช้ในการค้นหารูปแบบการเข้าถึงเว็บของผู้ใช้โดยการใช้หลักการของ AI ,ดาต้าไมนิ่ง,จิตวิทยา (psychorogy),และทฤษฎีข้อมูล (information theory)

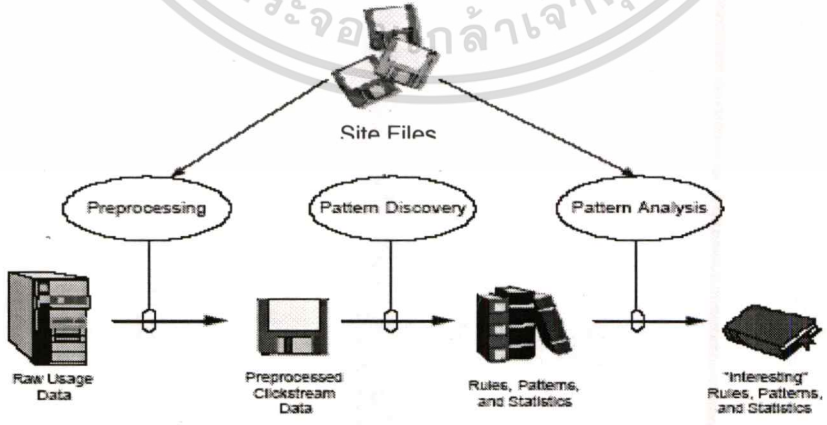
1.2) Pattern analysis tool

เป็นเครื่องมือที่ใช้ในการวิเคราะห์รูปแบบการเข้าถึงของผู้ใช้ โดยจะแปลรูปแบบที่ได้เป็นลักษณะที่สามารถเข้าใจได้ง่าย เห็นภาพชัดเจน

2.2.2 Web Usage Mining

Web Usage Mining เป็นเครื่องมือที่มีประโยชน์มากสำหรับผู้เป็นเจ้าของเว็บ เพราะทำให้ได้ทราบถึงพฤติกรรมการเข้าถึงเว็บของผู้ใช้ ทำให้ทราบยุทธศาสตร์ทางการค้าโดยเฉพาะเว็บไซด์ที่ดำเนินธุรกรรมแบบพาณิชย์อิเล็กทรอนิกส์ จะทำให้ทราบถึงกลุ่มลูกค้าที่มีศักยภาพดีพอที่จะซื้อสินค้าของเรา ในปัจจุบันสภาพการแข่งขันในตลาดเพิ่มขึ้นอย่างมากมาย ดังนั้นการดำเนินพาณิชย์อิเล็กทรอนิกส์จะต้องมีการวางแผน มีการวางยุทธศาสตร์การจัดการที่ถูกต้องจึงจะประสบความสำเร็จ Web Usage Mining เป็นเครื่องมืออีกชนิดหนึ่งที่ให้ข้อมูลช่วยในการตัดสินใจการวางแผนที่ดีต่อไป

ขั้นตอนของ Web usage mining แบ่งได้เป็น 3 ขั้นส่วนใหญ่ดังรูปที่ 2.4



รูปที่ 2.4 แสดงขั้นตอนของ Web usage mining

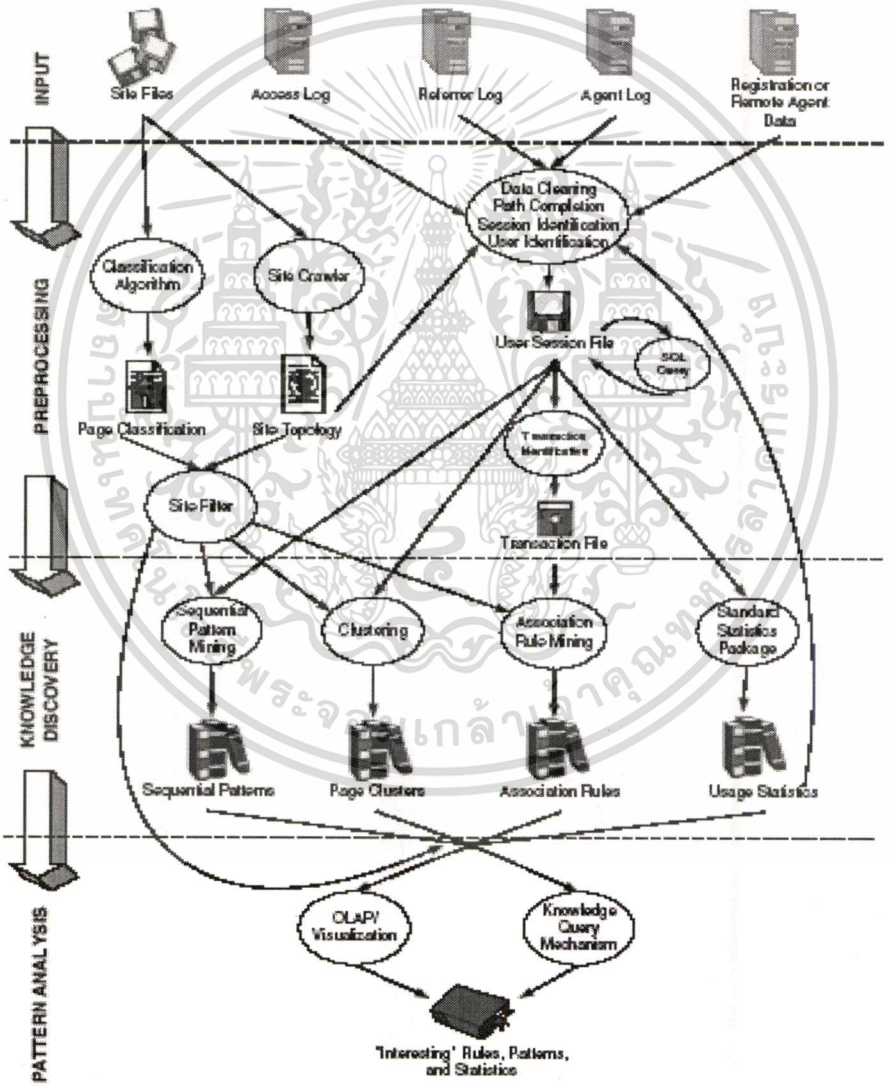
คือ 3 ขั้นตอนดังนี้

1) Preprocessing เป็นขั้นตอนของการเตรียมข้อมูล

2) Pattern Discovery เป็นขั้นตอนการค้นหาความสัมพันธ์

3) Pattern Analysis เป็นขั้นตอนการวิเคราะห์ความสัมพันธ์ที่ได้

โดยจากขั้นตอนหลักทั้ง 3 ขั้นตอนของ Web usage mining สามารถอธิบายขั้นตอนอย่างละเอียดได้ดังรูปที่ 2.5



รูปที่ 2.5 แสดง Architecture ของ Web usage mining

2.2.2.1 ส่วนของการนำเข้าของข้อมูล

จะมาจากส่วนของ server log 3 แหล่ง คือ access log , referrer log และ agent log ส่วนของไฟล์ HTML ที่มาจากไซด์ และส่วนของข้อมูลอื่นๆเช่นข้อมูลการลงทะเบียนหรือ remote agent log ข้อมูลที่จะนำมาวิเคราะห์หารูปแบบสามารถนำมาจากหลายแหล่ง ดังนี้

1) ข้อมูลที่นำมาจาก Server หรือ Web server

ข้อมูลที่ได้จาก Web server จะมีลักษณะเป็น log file ซึ่งเป็นข้อมูลที่มีความสำคัญมากในการนำไปทำเว็บไมนิ่ง ข้อมูลภายใน server log จะเก็บลักษณะการเข้าใช้เว็บไซด์แบบต่างๆเอาไว้ไม่ว่าจะเป็นการเข้าใช้ทีละคน หรือหลายๆคนเข้าใช้พร้อมกัน รูปแบบในการเก็บ log file มีหลายแบบไม่ว่าจะเป็นแบบธรรมดาหรือ common log หรือแบบ extended log log file ใน web server จะมีฟิลด์ดังนี้ ลำดับของ entry , ip address , userid (จะปรากฏเมื่อใช้ cookies) ,เวลาการเรียกใช้ , เพจที่ถูกเรียกใช้ โดยมีการอ้างอิงมาจากเพจไหน , สถานะของการเรียกใช้ (ปกติหรือผิดปกติ) , ขนาดของเพจที่เรียกใช้ และ agent ที่ใช้งาน ซึ่งรายละเอียดจะอธิบายในส่วนการเตรียมข้อมูล ข้อมูลที่ได้จาก log file ของ server ก็ยังมีข้อมูลบางอย่างที่ขาดไป และ ทำให้ไม่น่าเชื่อถือ เช่น กรณีที่มีผู้ใช้เรียกดูข้อมูลที่มาจาก caching page ของ proxy server

2) ข้อมูลที่เก็บมาจาก client

การจะเก็บข้อมูลจากทางฝั่ง client ส่วนใหญ่ต้องอาศัยความร่วมมือจากผู้ใช้ โดยการเก็บข้อมูลทำได้โดยใช้ remote agent เช่น Java script และ Java applet โดยถ้าพิจารณาในเทอมของการกำหนดเวลาในการดู ข้อมูลที่ได้จาก Java applet ไม่ได้ดีไปกว่าข้อมูลจาก server log และยังมี overhead เมื่อ load ขึ้นมาครั้งแรกด้วย ส่วน Java script จะกินเวลาในการแปลและไม่สามารถรองรับการ click ทั้งหมดของผู้ใช้ได้ เช่นเมื่อมีการ reload หรือ กดปุ่ม back วิธีเก็บแบบนี้จะได้เฉพาะข้อมูลแบบ single-user, single-site คือผู้ใช้หนึ่งคนเข้าใช้เว็บไซด์หนึ่งเว็บไซด์ ส่วนวิธีที่ใช้การปรับปรุง browser ที่มีอยู่แล้วเพื่อให้มีความสามารถในการเก็บรวบรวมได้ วิธีนี้จะต้องให้ผู้ใช้ใช้ browser ที่ถูกปรับปรุงนี้เป็นประจำ โดย browser ตัวนี้จะสามารถเก็บข้อมูลประเภทที่มีผู้ใช้คนเดียวแต่เข้าใช้ เว็บไซด์หลายเว็บไซด์ได้ ซึ่งข้อมูลที่ได้จากฝั่ง client นี้จะช่วยปรับปรุงข้อมูลที่เป็นประเภท caching page และการทำ session identification ได้

3) ข้อมูลที่เก็บมาจาก proxy server

proxy server เป็น cache ระหว่าง browser ของ client กับ web server โดย proxy จะเก็บเพจที่มีการเรียกดูบ่อย ๆ หรือเพจล่าสุดที่ถูกเรียกดูไว้จำนวนหนึ่ง เมื่อผู้ใช้ต้องการดูเพจเหล่านี้อีก มันก็จะ

ส่งไปให้ client ซึ่งทำให้ช่วยลดการทำงานของ web server ดังนั้นข้อมูลที่ได้จาก log file ของ server จะเป็นข้อมูลที่บอกรายละเอียดการร้องขอของ HTTP จาก client หลายตัวไปหา web server ต่าง ๆ โดยสามารถนำข้อมูลที่ได้มาแบ่งประเภทของผู้ใช้ที่ใช้ proxy ร่วมกันได้

4) ข้อมูลที่นำมาจากฐานข้อมูล

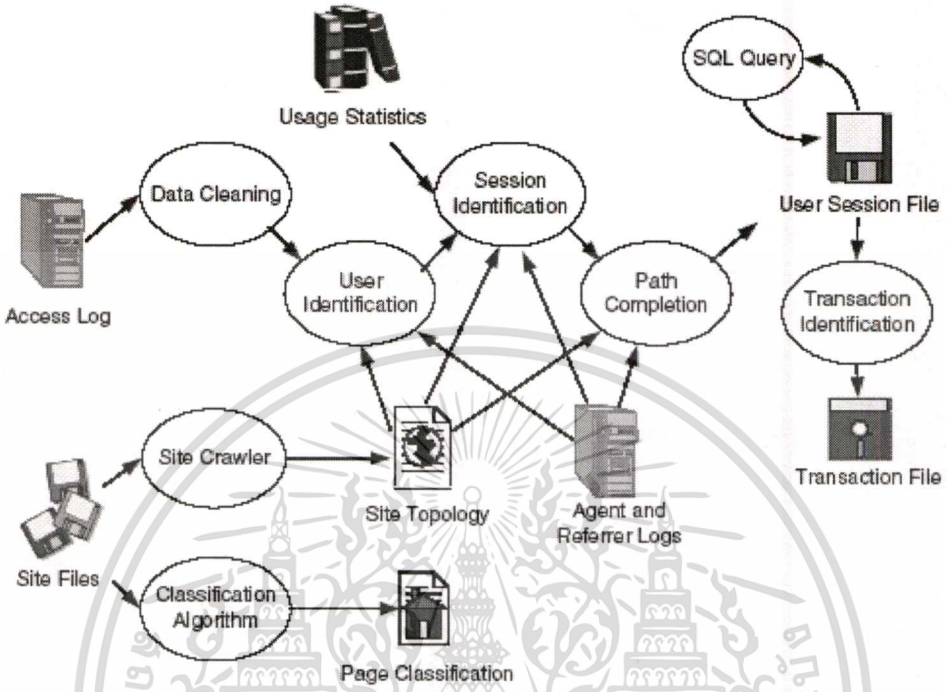
ข้อมูลที่นำมาจากฐานข้อมูล เป็นข้อมูลที่เก็บอยู่ภายในฐานข้อมูลขององค์กร เช่นอาจเป็นข้อมูลเกี่ยวกับธุรกิจขององค์กร ข้อมูลการตลาด ข้อมูลที่รวบรวมมาจากที่ต่าง ๆ

ประเภทของเพจมี 5 ประเภท ดังนี้

- เพจหลัก (Head Page) เป็นเพจที่มีจุดประสงค์ให้เป็นเพจแรกที่ผู้ใช้จะเข้ามาเยี่ยมชม เช่น โฮมเพจ
- เพจเนื้อหา (Content Page) เป็นที่บรรจุส่วนของเนื้อหาข้อมูลสารสนเทศที่เว็บไซต์มีการจัดเตรียมไว้
- เพจนำทาง (Navigation Page) เป็นเพจที่มีจุดประสงค์คือเตรียมลิงค์เพื่อนำทางผู้ใช้ไปยังเพจเนื้อหา
- เพจค้นดู (Look-up Page) เป็นเพจที่ใช้ในการจัดเตรียมคำนิยามหรือการขยายของคำย่อ
- เพจส่วนตัว (Personal Page) เป็นเพจที่ใช้ในการแสดงข้อมูลสารสนเทศของประวัติบุคคลหรือลักษณะของบุคคลที่เกี่ยวข้องกับองค์กรที่พัฒนาเว็บไซต์นั้น

2.2.2.2 ส่วนของการเตรียมข้อมูล (Data preprocessing)

รายละเอียดของการเตรียมข้อมูลของ Web usage mining แสดงได้ดังรูปที่ 2.6 ซึ่งข้อมูลอินพุตสำหรับขั้นตอนการเตรียมข้อมูลจะมี Server log , Site file หรืออาจจะมีสถิติการใช้งานที่ได้มาจากการวิเคราะห์ก่อนหน้านี้หรือไม่ก็ได้ ส่วนข้อมูลเอาพุตจะได้ไฟล์เซสชันของผู้ใช้ , ไฟล์ทรานแซ็กชั่น , Site topology และ page classification สิ่งกีดขวางที่สำคัญในการสร้างไฟล์เซสชั่นของผู้ใช้คือเรื่องของ browser และ proxy server caching ซึ่งปัจจุบันการเก็บข้อมูลสารสนเทศเกี่ยวกับการอ้างอิง cache คือการใช้ cookies และ cache busting ซึ่ง cache busing เป็นการปฏิบัติเพื่อป้องการ browser จากการเรียกใช้เพจเวอร์ชั่นที่ถูกเก็บไว้ในเครื่องของตน โดยการบังคับให้เรียกเพจใหม่จาก server ทุกๆครั้งที่มีการเรียกดูเพจ แต่ผู้ใช้อาจจะลบ cookie หรือยกเลิกการใช้ cache busting ได้ ทำให้ไม่ได้ใช้ความสามารถของ cache อย่างเต็มที่ ยังมีวิธีการอื่นที่ใช้ในการระบุผู้ใช้คือการลงทะเบียนผู้ใช้ก่อนที่จะมีการเรียกดูเพจ ซึ่งการลงทะเบียนนี้ก็มีข้อดีคือสามารถเก็บข้อมูลสารสนเทศเพิ่มเติมได้มากกว่าที่ server log เก็บ ทำให้ระบุผู้ใช้และเซสชั่นได้ง่ายกว่า แต่ก็มีข้อเสียคือผู้ใช้ส่วนใหญ่จะไม่เข้าไปเรียกใช้เพจที่ต้องมีการลงทะเบียนหรือบางทีผู้ใช้อาจจะใส่ข้อมูลที่ไม่ถูกต้องลงไป



รูปที่ 2.6 แสดงรายละเอียดของการเตรียมข้อมูลของ Web usage mining

สิ่งที่ต้องทำในส่วนของการเตรียมข้อมูล มีดังนี้

1) การทำความสะอาดข้อมูล (Data Cleaning)

การทำความสะอาด server log เพื่อกำจัดรายการที่ไม่เกี่ยวข้องออกไปนั้นถือว่าเป็นส่วนสำคัญสำหรับการวิเคราะห์ข้อมูล ความสัมพันธ์ที่ค้นพบหรือรายงานทางสถิติจะเป็นประโยชน์สูงสุดก็ต่อเมื่อข้อมูลใน server log ให้ภาพที่ถูกต้องแม่นยำของการเข้าถึงเว็บไซต์ของผู้ใช้เท่านั้น ซึ่งในการร้องขอเพจหนึ่งๆจาก web server โปรโตคอล HTTP จะแยกการเชื่อมต่อสำหรับทุกๆไฟล์ที่อยู่ในเว็บเพจ ดังนั้นการร้องขอของผู้ใช้เพื่อนที่จะเรียกดูเพจหนึ่งๆนั้น จะทำให้เกิดหลายๆ log entry เนื่องจากไฟล์กราฟฟิกและสคริปต์ไฟล์ต่างๆจะถูกโหลดมาพร้อมๆกับไฟล์ html อย่างอัตโนมัติ แต่ในการวิเคราะห์การใช้งานเว็บเรานำไฟล์ html เท่านั้นที่นำมาวิเคราะห์ เนื่องจากจุดประสงค์หลักของการไมนิ่งการใช้งานเว็บคือการหาพฤติกรรมของผู้ใช้ เพราะฉะนั้นจึงไม่จำเป็นที่จะนำไฟล์ที่ไม่ได้ร้องขอโดยตรงของผู้ใช้เข้าไปด้วย การกำจัดรายการที่ไม่เกี่ยวข้องสามารถทำได้โดยการตรวจสอบส่วนหลัง (suffix) ของชื่อ URL เช่น จะกำจัด log entry ที่มีชื่อไฟล์ส่วนหลังเป็น GIF , JPEG , jpg , JPG รวมทั้งพวกไฟล์สคริปต์ต่างๆ เช่น “count.cgi” แต่การลบไฟล์พวกนี้อาจจะมีการเปลี่ยนแปลงได้ขึ้นอยู่กับประเภทของไซต์ที่จะทำการ

วิเคราะห์เช่น เว็บไซต์ที่เก็บรูปภาพอาจจะไม่ต้องลบไฟล์เหล่านี้เพราะเป็นสิ่งที่แสดงพฤติกรรมของผู้ใช้ในการโหลดข้อมูลรูปภาพได้

2) การระบุผู้ใช้ (User Identification)

ต่อจากนั้นจะต้องมีการระบุผู้ใช้แต่ละคน ซึ่งมีความยุ่งยากมากขึ้นถ้ามีการใช้ cache , firewall และ proxy server โดยที่การจะแยกผู้ใช้แต่ละคนในกรณีที่มี IP Address เดียวกัน นั้นสามารถแบ่งผู้ใช้โดยพิจารณาจากชนิดของ browser software หรือชนิดของ OS คือ ถ้า IP Address เดียวกันแต่มี browser ในการร้องขอเว็บเพจต่างกันจะสามารถสรุปได้ว่าเป็นผู้ใช้คนละคนกัน เช่นรูปที่ 2.7 เป็น log ของ web server

#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
9	123.456.78.9	-	[25/Apr/1998:03:10:02 -0500]	"GET O.html HTTP/1.0"	200	2270	F.html	Mozilla/3.04 (Win95, I)
10	123.456.78.9	-	[25/Apr/1998:03:10:45 -0500]	"GET J.html HTTP/1.0"	200	9430	C.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
11	123.456.78.9	-	[25/Apr/1998:03:12:23 -0500]	"GET G.html HTTP/1.0"	200	7220	B.html	Mozilla/3.04 (Win95, I)
12	209.456.78.2	-	[25/Apr/1998:05:05:22 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
13	209.456.78.3	-	[25/Apr/1998:05:06:03 -0500]	"GET D.html HTTP/1.0"	200	1680	A.html	Mozilla/3.04 (Win95, I)

รูปที่ 2.7 ตัวอย่าง log file ในWeb server

จะเห็นว่าทุกๆ log entry จะมี IP Address เดียวกัน UserID ไม่ได้ถูกบันทึกไว้ เมื่อพิจารณา entry ที่ 5 , 6, 8 และ 10 จะเป็นการเข้าถึงเพจโดยการใช้ตัวแทนที่ต่างจาก entry อื่นๆ จึงอาจจะสรุปได้ว่า log นี้แสดงอย่างน้อย 2 ผู้ใช้งาน อีกหลักการหนึ่งที่สามารถระบุผู้ใช้ได้คือการพิจารณาจากฟิลด์ของเพจที่ร้องขอ (URL field) และเพจที่อ้างอิงก่อนหน้า (Referrer field) โดยถ้าเพจที่ร้องขอไม่ได้เข้าถึงโดยตรงจากเพจใดๆแล้วก็สรุปได้ว่าอาจจะเป็นผู้ใช้คนละคนกัน เช่น entry ที่3 (เพจ L) ไม่ได้เข้าถึงโดยตรงจากเพจ A,B และ entry ที่ 7 (เพจ R) อ้างอิงมาจากเพจ L เท่านั้น ด้วยเหตุนี้จึงสามารถสรุปได้ว่ามี 3 ผู้ใช้ใน IP Address เดียวกัน มีเส้นทางการเดินทางแต่ละคนดังนี้ A-B-F-O-G-A-D , A-B-C-J และ L-R แต่หลักการนี้จะทำให้เกิดความสับสนได้ในกรณีที่ผู้ใช้ 2 คนที่มี IP Address เดียวกัน และยังใช้

browser เดียวกันบนเครื่องชนิดเดียวกัน หรือ กรณีผู้ใช้คนเดียวที่ใช้ browser 2 ตัวทำงาน หรือกรณีที่ผู้ใช้พิมพ์ URL โดยตรงโดยไม่ใช้โครงสร้างของลิงค์ซึ่งทำให้เข้าใจผิดว่าเป็นผู้ใช้หลายคนได้

3) การระบุเซสชั่น (Session Identification)

สำหรับ log ที่เก็บเป็นระยะเวลานานจะมีความเป็นไปได้อย่างมากที่ผู้ใช้จะกลับเข้ามาเยี่ยมชมเว็บไซด์มากกว่า 1 ครั้ง จุดมุ่งหมายของการระบุเซสชั่นคือการแบ่งการเข้าถึงเพจของผู้ใช้แต่ละคนเป็นเซสชั่น วิธีที่ง่ายที่สุดคือใช้วิธีการกำหนดขอบเขตเวลา (timeout) โดยถ้าเวลาระหว่างการร้องขอเพจเกินขอบเขตที่กำหนด ก็สามารถสมมติฐานได้ว่าผู้ใช้มีการเริ่มเซสชั่นใหม่ ซึ่งหลายๆผลิตภัณฑ์จะมี default ค่าขอบเขตเวลาไว้ที่ 30 นาที อีกทั้งเมื่อ log ที่เคยถูกวิเคราะห์และได้หาสถิติการใช้งานค่าขอบเขตเวลาที่เหมาะสมของแต่ละไซด์นั้นแล้วสามารถนำกลับเข้าไปเป็นข้อมูลอินพุตในอัลกอริทึมที่ทำการระบุเซสชั่นได้อีก ตัวอย่างจากรูปที่ 8 จากเส้นทางของผู้ใช้คนที่ 1 คือ A-B-F-O-G-A-D ถ้ากำหนดค่าขอบเขตเวลาเท่ากับ 30 นาที จะแตกเซสชั่นได้เป็น 2 เซสชั่นเนื่องจาก 2 เพจที่อ้างอิงสุดท้ายห่างจาก 5 การอ้างอิงแรกกว่าชั่วโมง ดังนั้นในขั้นตอนของการระบุเซสชั่นทำให้ได้ผลลัพธ์ทั้งสิ้น 4 เซสชั่น ดังนี้ A-B-F-O-G , A-D , A-B-C-J และ L-R

4) การทำเส้นทางให้สมบูรณ์ (Path Completing)

ในกรณีที่การร้องขอเพจไม่ได้ถูกอ้างอิงโดยตรงจากเพจสุดท้ายที่ผู้ใช้ร้องขอแล้ว referrer log สามารถใช้ในการตรวจสอบเพื่อดูการร้องขอเพจว่ามาจากไหน ถ้าเพจอยู่ในการร้องขอของผู้ใช้ที่เก็บประวัติก่อนหน้านี้แล้วสามารถตั้งสมมติฐานได้ว่าผู้ใช้ทำการย้อนกลับโดยใช้ปุ่ม “BACK” โดยจะทำการเรียกเพจที่อยู่ใน cache ขึ้นมาแสดง ถ้ามีมากกว่า 1 เพจในประวัติของผู้ใช้ที่บรรจุลิงค์ไปยังเพจที่ร้องขอ จะสมมติฐานว่าเพจที่ใกล้ที่สุดก่อนหน้าเพจที่ร้องขอคือเพจต้นทางของเพจนั้น ดังนั้นจึงต้องเพิ่มเพจอ้างอิงก่อนหน้าเข้าไปในไฟล์เซสชั่นของผู้ใช้ด้วย ตัวอย่างจากรูปที่ 8 entry ที่ 11 เพจ G ไม่ได้มีการเข้าถึงมาจากเพจ O โดยตรง ในส่วนของ referrer log ของเพจ G ได้ร้องขอเพจ B จากเหตุการณ์นี้สามารถบอกได้ว่าผู้ใช้ได้ย้อนกลับไปยังเพจ B โดยใช้ปุ่ม “BACK” ก่อนที่จะร้องขอเพจ G ดังนั้นจะต้องเพิ่มเพจ F และ B เข้าไปในเซสชั่นไฟล์ แต่ก็เป็นไปได้ที่ผู้ใช้รู้ URL ของเพจ G และร้องขอโดยตรงซึ่งเป็นสิ่งที่ไม่ต้องการให้เกิดขึ้น แต่ก็ถือว่าเป็นเหตุการณ์ที่เกิดขึ้นน้อยไม่มีผลกระทบต่อการไมนิ่งข้อมูล เพราะฉะนั้นผลลัพธ์ของขั้นตอนการทำเส้นทางให้สมบูรณ์ จะได้เส้นทางของผู้ใช้ดังนี้ A-B-F-O-F-B-G , A-D , A-B-A-C-J และ L-R ซึ่งผลลัพธ์ของการเตรียมข้อมูลสรุปได้ดังตารางที่ 2.2

ตารางที่ 2.2 สรุปผลลัพธ์ของตัวอย่างการเตรียมข้อมูล log

Task	Result
Clean Log	• A-B-L-F-A-B-R-C-O-J-G-A-D
User Identification	• A-B-F-O-G-A-D • A-B-C-J • L-R
Session Identification	• A-B-F-O-G • A-D • A-B-C-J • L-R
Path Completion	• A-B-F-O-F-B-G • A-D • A-B-A-C-J • L-R

5) การจัดรูปแบบ (Formatting)

ส่วนสุดท้ายของขั้นตอนการเตรียมข้อมูลคือการเตรียมรูปแบบที่เหมาะสมของเซสชั่นหรือทรานแซ็กชั่นสำหรับการทำดาต้าไมนิ่งแต่ละชนิด เช่นถ้าเป็นการเตรียมข้อมูลสำหรับการหากฎของสิ่งที่สัมพันธ์กัน (Association rule) จะมีการตัดเวลาในการร้องขอเพจทิ้งไป และจัดรูปแบบให้เหมาะสมกับอัลกอริทึมที่เราเลือกใช้

6) การระบุทรานแซ็กชั่น (Transaction Identification)

เมื่อเราระบุเซสชั่นของผู้ใช้จากขั้นตอนการเตรียมข้อมูลแล้ว จะนำมาวิเคราะห์หาทรานแซ็กชั่นของผู้ใช้ จุดมุ่งหมายของการระบุทรานแซ็กชั่นคือการสร้างกลุ่มของการอ้างอิงเพจที่เหมาะสมของผู้ใช้แต่ละคน เพราะฉะนั้นงานของการระบุทรานแซ็กชั่นคือเป็นได้ทั้งการแบ่ง (Deviding) ทรานแซ็กชั่นขนาดใหญ่ให้เป็นทรานแซ็กชั่นขนาดเล็กลงหลายๆทรานแซ็กชั่น หรือการรวม (merging) ทรานแซ็กชั่นขนาดเล็กให้เป็นทรานแซ็กชั่นที่ใหญ่ขึ้น เพื่อสร้างทรานแซ็กชั่นที่เหมาะสมสำหรับการวิเคราะห์ด้วยดาต้าไมนิ่ง โดยการระบุทรานแซ็กชั่นมี 3 วิธี ดังนี้

6.1) ระบุทรานแซ็กชั่นโดยใช้การอ้างอิงแบบยาว (Reference Length)

การระบุทรานแซ็กชั่นแบ่งโดยใช้ความยาวของแต่ละการอ้างอิง ประมาณหาผลต่างของเวลาระหว่างการอ้างอิงครั้งถัดไปและการอ้างอิงปัจจุบัน จะเห็นว่าการอ้างอิงครั้งสุดท้ายในแต่ละทรานแซ็กชั่นจะไม่มีเวลาครั้งถัดไปที่จะใช้ในการประมาณความยาวของการอ้างอิง วิธีการอ้างอิงความยาวนี้จะสมมติว่าทุกๆการอ้างอิงเพจสุดท้ายคือการอ้างอิงเพจเนื้อหา และจะละทิ้งเพจนี้ในขณะที่คำนวณค่า

เวลาแบ่งแยก จากรูปที่ 8 ทำให้ได้ทรานแซ็กชั่นของเพจเนื้อหาคือ F-G , D , L-R และ J ส่วนทรานแซ็กชั่นของเพจสนับสนุน-เพจเนื้อหาเป็น A-B-F , O-F-B-G ,A-D , L , R และ A-B-A-C-J

6.2) ระบุทรานแซ็กชั่นโดยใช้การอ้างอิงไปข้างหน้าไกลที่สุด(Maximal Forward Reference)

วิธีการระบุทรานแซ็กชั่นโดยการอ้างอิงไปข้างหน้าไกลที่สุดจะใช้หลักการที่ว่าทรานแซ็กชั่นจะถูกกำหนดให้เป็นชุดของเพจจากเพจแรกในเซสชั่นจนถึงเพจก่อนที่จะทำการอ้างอิงย้อนกลับ (Backward reference) การอ้างอิงไปข้างหน้า (Forward reference) จะเป็นเพจที่ยังไม่เคยอยู่ในชุดของทรานแซ็กชั่นปัจจุบัน ทรานแซ็กชั่นใหม่จะเริ่มต้นอีกครั้งเมื่อการอ้างอิงไปข้างหน้าครั้งถัดไปเกิดขึ้น เพจการอ้างอิงไปข้างหน้าที่ไกลที่สุดคือเพจเนื้อหาและเพจที่นำไปสู่แต่ละเพจการอ้างอิงไปข้างหน้าที่ไกลที่สุดคือเพจสนับสนุน ดังนั้นวิธีนี้เหมือนกับวิธีการอ้างอิงความยาวทรานแซ็กชั่นมี 2 ชุด คือ ทรานแซ็กชั่นเพจสนันสนุน-เพจเนื้อหา หรือทรานแซ็กชั่นเฉพาะเพจเนื้อหา จากรูปที่ 8 และหัวข้อการทำเส้นทางให้สมบูรณ์ จะได้ทรานแซ็กชั่นเพจสนันสนุน-เพจเนื้อหาเป็น A-B-F-O , A-B-G , L-R , A-B , A-C-J และ A-D ส่วนทรานแซ็กชั่นเฉพาะเพจเนื้อหาคือ O-G , R , B-J และ D ข้อดีของวิธีการอ้างอิงไปข้างหน้าไกลที่สุดคือไม่ต้องการพารามิเตอร์เป็นข้อมูลเข้า

6.3) ระบุทรานแซ็กชั่นโดยใช้ช่วงเวลา (Time Window)

วิธีการระบุทรานแซ็กชั่นโดยใช้ช่วงเวลามีการแบ่งทรานแซ็กชั่นของผู้ใช้ตามคาบเวลา วิธีการนี้ไม่ได้อยู่บนพื้นฐานของเพจสนับสนุนและเพจเนื้อหาเหมือน 2 วิธีแรก แต่อยู่บนสมมติฐานที่ว่าทรานแซ็กชั่นที่มีค่าเฉลี่ยของความยาวทั้งหมดที่สัมพันธ์กัน สำหรับค่าของช่วงเวลาที่กว้างพอทำให้แต่ละทรานแซ็กชั่นบรรจุทั้งเซสชั่นของผู้ใช้ ดังนั้นทำให้ไม่สามารถแยกออกเป็น 2 ทรานแซ็กชั่นได้ จากรูปที่ 8 จะได้ทรานแซ็กชั่นดังนี้ A-B-F , O-F-B-G , A-D , L-R และ A-B-A-C-J

2.2.2.3 ส่วนของการค้นหาความสัมพันธ์ (Pattern Discovery)

ขั้นตอนการค้นหาความสัมพันธ์ทำได้หลาหลายวิธีขึ้นอยู่กับจุดประสงค์ของการทำ Web usage mining ซึ่งแต่ละวิธีจะให้ผลลัพธ์ที่สามารถนำมาประยุกต์ใช้ได้แตกต่างกัน นอกจากนั้นเรายังสามารถใช้ขั้นตอนต่างๆ มารวมกันเพื่อประยุกต์ใช้ในทิศทางใหม่ๆได้ ตัวอย่างเทคนิคที่ใช้ค้นหาความสัมพันธ์มีดังนี้

1) การวิเคราะห์เส้นทาง (Pattern analysis)

เนื่องจากกราฟเป็นสิ่งที่แทนความสัมพันธ์ระหว่างเว็บเพจ จึงนิยมใช้กราฟมาทำการวิเคราะห์หาเส้นทาง เพราะกราฟเป็นการแสดงโครงสร้างทางกายภาพของเว็บไซต์โดยมีโหนดแทนเว็บเพจและ

กิ่งที่เชื่อมต่อระหว่างเพจคือไฮเปอร์เทกซ์ลิงค์ ในปัจจุบันงานวิจัยต่างจัยต่างๆให้ความสนใจที่จะค้นหารูปแบบการเดินทางที่เกิดขึ้นบ่อย (Frequent traversal pattern) หรือลำดับการอ้างอิงที่มากที่สุด (Large reference sequence) จากโครงสร้างทางกายภาพ การวิเคราะห์หาเส้นทางใช้ในการหาเส้นทางที่มีการเข้าเยี่ยมชมมากที่สุดในเว็บไซด์ ตัวอย่างของข้อมูลสารสนเทศที่พบโดยการใช้การวิเคราะห์เส้นทางเป็นดังนี้

- 70% ของ client ที่เข้าถึง /company/product2 จะเริ่มต้นที่ /company ก่อนจากนั้นจะผ่าน /company/new , /company/products และ /company/product1

  กฎนี้จะบอกได้ว่าผู้ใช้จะเดินทางอ้อมเพื่อไปยังเพจ /company/product2 แสดงว่าเพจนี้ไม่ได้ถูกออกแบบให้เข้าถึงอย่างเด่นชัด

- 80% ของ client ที่เข้าถึงไซด์จะเริ่มต้นจาก /company/products

  กฎนี้จะบอกได้ว่าผู้ใช้มีการเข้าถึงไซด์โดยผ่านเพจอื่นที่ไม่ใช่เพจหลัก (เช่น สมมติว่าเป็นเพจ /company) ซึ่งทำให้มีแนวคิดที่ว่าจะทำการรวมสารสนเทศลงไปในเพจนี้เลย

- 65% ของ client ออกจากไซด์หลังจากที่มีการอ้างอิงเพจ 4 เพจหรือน้อยกว่านั้น

  กฎนี้แสดงให้เห็นถึงอัตราการใช้งานของไซด์ถ้าผู้ใช้มีการค้นหามากกว่า 4 เพจในไซด์ผู้ใช้จะออกจากไซด์ ดังนั้นจึงต้องบรรจุข้อมูลสารสนเทศให้อยู่ภายใน 4 เพจที่ผู้ใช้มักเข้าถึง

2) การวิเคราะห์หาสิ่งที่สัมพันธ์กัน (Association rule)

เทคนิคการไมนิ่งสิ่งที่สัมพันธ์กันจะค้นหาสิ่งที่ไม่เรียงลำดับระหว่างรายการสิ่งของที่พบในฐานข้อมูล จะใช้หลักการเดียวกับ market basket โดยการค้นหาหน้าเว็บที่สัมพันธ์กัน โดยเทียบกับจำนวนความสัมพันธ์กับค่า support ค่าหนึ่ง จำนวนความสัมพันธ์ที่เกินค่า support จะถูกเลือกเป็นความสัมพันธ์ การค้นหาความสัมพันธ์ระหว่างแต่ละ session ช่วยให้เว็บดีไซเนอร์สามารถ restructure โครงสร้างเว็บเพจได้ เว็บเพจที่สัมพันธ์กันจะถูกนำไปอยู่กลุ่มเดียวกันและเราสามารถสร้างเว็บเพจให้เกิดความสัมพันธ์ตามใจผู้ออกแบบได้ ตัวอย่างของกฎของสิ่งที่ค้นพบดังนี้

- 40% ของ client ที่เข้าถึงเว็บเพจด้วย URL /company/product1 ยังมีการเข้าถึง /company/product2 ด้วย
- 30% ของ client ที่เข้าถึง /company/special จะมีการสั่งสินค้าแบบออนไลน์ใน /company/product1 ด้วย

ซึ่ง % หมายถึงความเชื่อมั่น (Confidence) ซึ่งคือจำนวนของทรานแซ็กชั่นที่บรรจุทุกๆรายการในกฎหารด้วยจำนวนของทรานแซ็กชั่นที่เป็น rule antecedent (/company/product1 ในตัวอย่างแรก)

3) การวิเคราะห์หารูปแบบตามลำดับ (Sequential pattern)

เทคนิคการค้นหารูปแบบตามลำดับ คือการค้นกาลำดับที่ปรากฏมากที่สุดโดยพิจารณาจากค่า support ค่าหนึ่ง โดยลำดับรูปแบบนั้นคือ เหตุการณ์ที่เกิดขึ้นโดยมีกาลเวลาเป็นตัวกำหนดลำดับ ใน Web usage mining การค้นกาลำดับรูปแบบคือการค้นกาลำดับของการ browse เข้าไปในเว็บเพจ ลำดับการ browse ที่ได้จากการค้นหาสามารถใช้ในการทำนายรูปแบบการ browse ในอนาคตได้ซึ่งทำให้ผู้ออกแบบสามารถติดตั้งโฆษณาได้ตรงกับการเข้าชมของผู้ใช้เว็บ ตัวอย่างลำดับของรูปแบบการเข้าชมแสดงได้ดังนี้

- 30% ของ client ที่เข้าเยี่ยมชม /company/products ได้มาจากการที่ผู้ใช้ทำการค้นหาใน Yahoo เมื่อสัปดาห์ที่ผ่านมาด้วยคำค้น w
- 60% ของ client ผู้ที่สั่งสินค้าแบบออนไลน์ใน /company/product1 ยังมีการสั่งสินค้าออนไลน์ใน /company/product4 ภายใน 15 วัน

ซึ่ง % หมายถึงค่าสนับสนุน (Support) ซึ่งคือ % ของทรานแซ็กชั่นที่บรรจุรูปแบบที่ให้ไว้ ทั้งค่าของความเชื่อมั่นและค่าสนับสนุนถูกใช้เป็นเหมือนเทรสโฮล์ดเพื่อที่จะจำกัดจำนวนของกฎที่ค้นพบและรายงานออกมา ตัวอย่างเช่นถ้าให้เทรสโฮลด์ค่าสนับสนุนเท่ากับ 50 % แล้วตัวอย่างที่ 1 จะไม่ถูกรายงานออกมา

4) การวิเคราะห์เพื่อการแบ่งกลุ่ม (Clustering)

การแบ่งกลุ่มของข้อมูลคือเทคนิคในการแบ่งกลุ่มข้อมูลที่มีคุณสมบัติเหมือนกัน ใน Web usage mining ซึ่งคือการแบ่งกลุ่มรูปแบบการ browse เว็บเพจใน Web server

การแบ่งกลุ่มรูปแบบการ browse ทำให้เราสามารถแบ่งกลุ่มผู้ใช้เว็บออกเป็นกลุ่มๆ ได้ซึ่งส่งผลให้เราสามารถทำนานแนวโน้มของผู้เข้าชมเว็บได้ อีกทั้งยังสามารถติดตั้งโฆษณาได้ตรงกลุ่มเป้าหมายด้วย

2.2.2.4 ส่วนของการวิเคราะห์ความสัมพันธ์ที่ได้(Pattern Analysis)

ขั้นตอนการวิเคราะห์ความสัมพันธ์ที่ได้คือการตัดกฎหรือความสัมพันธ์มีไม่น่าสนใจหรือไม่เด่นชัดออกไป ความสัมพันธ์ที่ได้จากการค้นกาอาจจะมีมากจนไม่สามารถนำมาประยุกต์ใช้ได้อย่างมีประสิทธิภาพจึงต้องมีกรรมวิธีในการตัดความสัมพันธ์ที่ไม่น่าสนใจออกไป เช่นการเปรียบเทียบกับค่า support เป็นต้น

นอกจากการตัดกฎหรือความสัมพันธ์ที่ไม่น่าสนใจแล้วการแสดงผลและการจัดเก็บกฎและความสัมพันธ์ยังส่งผลต่อความสะดวกในการวิเคราะห์ความสัมพันธ์ด้วย การแสดงผลด้วยภาพหรือ chart ที่เข้าใจได้ง่าย ทำให้การวิเคราะห์ความสัมพันธ์เป็นไปได้ง่ายขึ้น การจัดเก็บความสัมพันธ์ไว้เป็นฐานข้อมูลจะช่วยให้การค้นหากฎความสัมพันธ์เป็นไปได้สะดวกยิ่งขึ้น

บทที่ 3
ทฤษฎีที่นำมาใช้

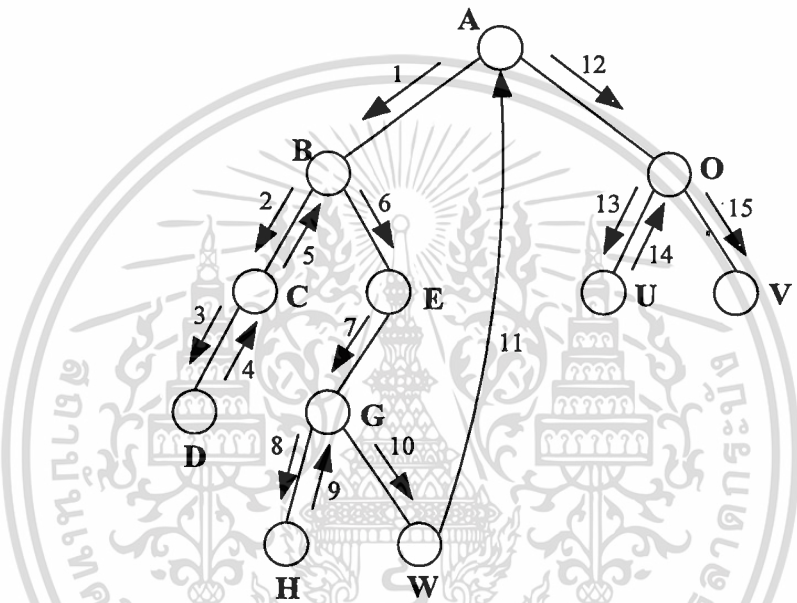
ในการดำเนินการเกี่ยวกับปัญหาในเรื่องการไมนิ่งข้อมูลของ proxy log จะมีขั้นตอนหลักอยู่ 3 ขั้นตอน คือ ขั้นตอนแรกจะต้องนำไฟล์ proxy log ซึ่งเป็นไฟล์แบบ text มาทำขั้นตอนของ preprocessing ก่อน คือต้องทำความสะอาดข้อมูลก่อนโดยเลือกเฉพาะข้อมูลที่เป็นประโยชน์ในการวิเคราะห์เท่านั้น จากนั้นจะต้องมีการระบุผู้ใช้ ระบุเซสชั่น และระบุทรานแซ็กชั่นของผู้ใช้ จากนั้นขั้นตอนที่สอง จะหารูปแบบลำดับการเดินทางของผู้ใช้ในการเข้าถึงเว็บเพจต่างๆ โดยในที่นี้จะเลือกใช้วิธีการอ้างอิงไปข้างหน้าไกลที่สุด วิธีการนี้จะใช้อัลกอริทึม MF (maximal forward reference) ในการแปลงข้อมูลไปเป็นชุดลำดับการเดินทาง (Traversal subsequence) ขั้นตอนสุดท้ายจะเป็นขั้นตอนการวิเคราะห์หาความสัมพันธ์ของชุดลำดับการเดินทางที่ได้มาจากอัลกอริทึม MF โดยในที่นี้จะใช้อัลกอริทึมของ Apriori ในการหาความสัมพันธ์นั้น

3.1 การหารูปแบบอ้างอิงไปข้างหน้าไกลที่สุด (Maximal forward reference)

3.1.1 ที่มาของปัญหา

ในสภาวะแวดล้อมที่ออปเจ็คของข้อมูลสารสนเทศถูกเชื่องโยงเข้าด้วยกัน ผู้ใช้มักจะเดินทางไปยังออปเจ็คก่อนหน้าหรือถัดไปด้วยลิงค์หรือไอคอนที่ได้จัดเตรียมไว้ ดังนั้นทำให้มีผลว่าบางโหนดจะถูกเข้าถึงซ้ำ เช่นในเวิลด์ไวด์เว็บผู้ใช้มักจะย้อนกลับไปยังโหนดก่อนหน้าแล้วจึงจะเดินทางต่อไปยังโหนดอื่นแทนที่จะเปิด URL ใหม่ ดังนั้นรูปแบบการเข้าถึงของผู้ใช้จากฐานข้อมูล log จึงทำได้ยากขึ้น ในมุมมองแบบนี้เราตั้งสมมติฐานได้ว่าการอ้างอิงย้อนกลับถูกทำเพื่อให้สะดวกในการเดินทาง แต่ไม่ใช่สำหรับการค้นหา (browse)การเข้าถึงย้อนกลับหมายถึงการเข้าถึงออปเจ็คซ้ำอีกครั้งโดยผู้ใช้คนเดียวกัน ซึ่งเราจะมุ่งความสนใจไปที่การค้นหาของรูปแบบการอ้างอิงไปข้างหน้า และการอ้างอิงไปข้างหน้าจะสิ้นสุดลงเมื่อเกิดการอ้างอิงย้อนกลับ ผลลัพธ์ของเส้นทางการเข้าถึงไปข้างหน้านี้จะอยู่ในรูปของ Maximal forward reference เมื่อได้รับ Maximal forward reference แล้วเราจะกลับไปที่จุดเริ่มต้นของการเข้าถึงไปข้างหน้าอีกครั้งและสืบค้นเส้นทางของการเข้าถึงไปข้างหน้าเส้นทางอื่นใหม่ และถ้ามีโหนดที่มีจุดเริ่มต้นเป็น null เกิดขึ้นจะสรุปได้ว่าเป็นการสิ้นสุดเส้นทางของการเข้าถึงไปข้างหน้าที่กำลังทำอยู่และเป็นจุดเริ่มต้นของเส้นทางอื่น

การนำเสนออัลกอริทึม Maximal forward reference (MF) จะกล่าวถึงในหัวข้อถัดไป ซึ่งในที่นี้จะอธิบายการหา Maximal forward reference โดยสมมติ log การเดินทางของผู้ใช้คนหนึ่งดังต่อไปนี้ {A, B, C, D, C, B, E, G, H, G, W, A, O, U, O, V} ดังแสดงได้ในรูปที่ 3.1 จากนั้นใช้อัลกอริทึม MF หาชุดของ Maximal forward reference สำหรับผู้ใช้รายนี้ คือ {ABCD, ABEGH, ABEGW, AOU, AOV}



รูปที่ 3.1 แสดงตัวอย่างของเส้นทางการเดินทางของผู้ใช้

3.1.2 อัลกอริทึมสำหรับการหารูปแบบอ้างอิงไปข้างหน้าไกลที่สุด (MF Algorithm)

อัลกอริทึม MF จะแปลงลำดับการเดินทางให้เป็นชุดของ Maximal forward reference ซึ่งการหา Maximal forward reference โดยหลังจากเตรียมข้อมูลแล้วฐานข้อมูล log ของการเดินทางจะบรรจุคู่ของ (ต้นทาง,ปลายทาง) ของแต่ละการเดินทางที่เชื่อมต่อกัน สำหรับการเริ่มต้นการเดินทางโหนดต้นทางจะเป็น null โดยให้ลำดับของการเดินทางของผู้ใช้เป็น $\{(s_1, d_1), (s_2, d_2), \ldots, (s_n, d_n)\}$ เราจะจับลำดับนี้ไปเป็นชุดลำดับหลายๆชุดลำดับ โดยแต่ละชุดลำดับแสดงถึง Maximal forward reference อัลกอริทึมสำหรับการหา Maximal forward referenceในทุกๆลำดับการเดินทางมีการทำงานดังนี้ คือ ขั้นแรกต้องทำการกรองข้อมูลที่ไม่เกี่ยวข้องออก (Data cleaning) จากฐานข้อมูล log , ระบุผู้ใช้ (User identification) , ระบุเซสชั่น (Session identification) , ทำเส้นทางการเดินทางให้สมบูรณ์ (Path

Completion) และการแปลงรูปแบบข้อมูลให้เหมาะสม (Formatting) ผลลัพธ์ที่ได้จะได้เส้นทางการเดินทาง $\{(s_1, d_1), (s_2, d_2), \ldots, (s_n, d_n)\}$ สำหรับแต่ละผู้ใช้ โดยที่คู่ของ (s_i, d_i) ถูกเรียงลำดับตามเวลา จากนั้นจะหา Maximal forward reference โดยการประยุกต์ใช้กับอัลกอริทึม MF ผลลัพธ์ที่ได้จะนำไปเก็บไว้ใน D_F ซึ่งเป็นฐานข้อมูลที่ใช้เก็บทุกๆ Maximal forward reference

โดยการทำงานของอัลกอริทึม MF จะแสดงได้ดังรูปที่ 3.2

Step 1: Set $i = 1$ and string Y to null for initialization, where string Y is used to store the current forward reference path. Also, set the flag $F = 1$ to indicate a forward traversal.

Step 2: Let $A = s_i$ and $B = d_i$.
If A is equal to null then
/* this is the beginning of a new traversal */
begin
Write out the current string Y (if not null) to the database D_F;
Set string $Y = B$;
Go to Step 5.
end

Step 3: If B is equal to some reference (say the j-th reference) in string Y then
/* this is a cross-referencing back to a previous reference */
begin
If F is equal to 1 then write out string Y to database D_F;
Discard all the references after the j-th one in string Y;
$F = 0$;
Go to Step 5.
end

Step 4: Otherwise, append B to the end of string Y.
/* we are continuing a forward traversal */
If F is equal to 0, set $F = 1$.

Step 5: Set $i = i + 1$. If the sequence is not completed scanned then go to Step 2.

รูปที่ 3.2 แสดงอัลกอริทึมของ MF

ตัวอย่างการเดินทางในรูปที่ 3.2 พบได้ว่าการเข้าถึงย้อนกลับที่เกิดขึ้นครั้งแรก พบในการเคลื่อนที่ครั้งที่ 4 (จาก D ไปยัง C) ณ จุดนี้ Maximal forward reference ABCD จะถูกเขียนลงใน D_F (โดยขั้นตอนที่ 3) ในการเคลื่อนที่ถัดมา (จาก C ไป B) แม้ว่าเงื่อนไขแรกในขั้นตอนที่ 3 จะเป็นจริง แต่จะไม่มีการเขียนลง D_F เนื่องจาก flag = 0 ซึ่งหมายถึงการเดินทางย้อนกลับ การเข้าถึงไปข้างหน้าถัดไป

คือ ABEGH จะเก็บลงใน string Y และจากนั้นจะถูกเขียนลงใน D_F เมื่อมีการพบการเข้าถึงย้อนกลับ (จาก H ไป G) รูปแบบการทำงานของอัลกอริทึม MF โดยมีข้อมูลในรูปที่ 3.1 เป็นอินพุต แสดงไว้ในตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างการทำงานโดยใช้อัลกอริทึม MF

move	string Y	output to D_F
1	AB	–
2	ABC	–
3	$ABCD$	–
4	ABC	$ABCD$
5	AB	–
6	ABE	–
7	$ABEG$	–
8	$ABEGH$	–
9	$ABEG$	$ABEGH$
10	$ABEGW$	–
11	A	$ABEGW$
12	AO	–
13	AOU	–
14	AO	AOU
15	AOV	AOV (end)

3.2 การวิเคราะห์หาความสัมพันธ์ของชุดลำดับการเดินทาง

เมื่อฐานข้อมูลที่บรรจุทุกๆ Maximal forward reference ของผู้ใช้ทุกคน (D_F) ถูกสร้างขึ้นเราสามารถหารูปแบบการเดินทางที่เกิดขึ้นบ่อยโดยการกำหนดความถี่ของการปรากฏลำดับการเข้าถึงใน D_F ลำดับ $s_1,\ldots,s_n$ เราสามารถบอกได้ว่าบรรจุลำดับที่ตามกันมา $r_1,\ldots,r_n$ ถ้ามีการเกิดของ i ที่ทำให้ $s_{i+j} = r_j$ สำหรับ $1 \leq j \leq k$ ถ้ามีจำนวนของ Maximal forward reference ใน D_F ที่บรรจุ $r_1,\ldots,r_n$ มีจำนวนเพียงพอแล้วลำดับของ k-reference $(r_1,\ldots,r_n)$ เรียกได้ว่าเป็น Large reference sequence

ในที่นี่จะใช้อัลกอริทึม Apriori ในการวิเคราะห์หาความสัมพันธ์ของ Maximal forward reference ที่ได้มา ซึ่งอัลกอริทึม Apriori นี้ เป็นอัลกอริทึมของหลักการ Association Discovery ซึ่งเป็นหลักการที่วิเคราะห์หาความสัมพันธ์ของข้อมูลที่เกิดขึ้นในรายการเดียวกัน

3.2.1 Association Rule

หลักการของ Association rule เป็นการวิเคราะห์หาความสัมพันธ์ของข้อมูลที่เกิดขึ้นในรายการเดียวกัน ซึ่งสินค้าเหล่านี้มักมีแนวโน้มที่จะถูกซื้อควบคู่กันไป การวิเคราะห์แบบนี้เรียกว่า "Market Basket Analysis" (MBA) แนวคิดนี้นำไปใช้ในร้าน super market เพื่อกำหนดว่าสิ่งของอะไรจะถูกซื้อควบคู่กันไปในการซื้อแต่ละครั้ง ทำให้ทางร้านสามารถกำหนดได้ว่าควรจัดเรียงสินค้าอย่างไร หรือควรเตรียมแคตตาล็อกเพื่อขายสินค้าอย่างไร รวมถึงการวางแผนเพื่อจัดโปรโมชั่นอย่างไร

Association rule มีรูปแบบอยู่ในลักษณะ "IF A THEN B" หรือ A→ B โดยที่ A และ B เกิดขึ้นพร้อมกันในทรานแซ็กชั่นเดียวกัน เรียก A ว่า "เหตุ" (หรือ Rule body หรือ Antecedent หรือ Left-hand side) เรียก B ว่า "ผล" (หรือ Rule head หรือ Consequent หรือ Right-hand side)

Association rule จะมีตัววัดหลักๆ อยู่ 2 ตัว คือ Support factor และ Confidence factor ตัวอย่าง เช่น ถ้าลูกค้าซื้อขนมปัง จะซื้อนมข้นด้วย โดยที่มีค่า Confidence = 75% และ Support = 35%

- Confidence คือ ค่าที่แสดงสัดส่วนระหว่างจำนวนชุดข้อมูลที่มีทั้งข้อมูล "เหตุ" และ "ผล" เทียบกับจำนวนข้อมูลที่มีเฉพาะเหตุการณ์ที่เป็น "เหตุ"
  ดังนั้น ค่า Confidence = 50% ได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการขนมปังและนมข้นคู่กัน}}{\text{จำนวนชุดข้อมูลที่มีรายการขนมปัง}}$$

- Support คือค่าที่แสดงสัดส่วนระหว่างชุดของข้อมูลที่มีทั้งข้อมูล "เหตุ" และ "ผล" เทียบกับจำนวนข้อมูลเหตุการณ์ทั้งหมด
  ดังนั้น ค่า Support = 35% ได้มาจาก

$$\frac{\text{จำนวนชุดข้อมูลที่มีรายการขนมปังและนมข้นคู่กัน}}{\text{จำนวนชุดข้อมูลทั้งหมด}}$$

เนื่องจากการทำงานของ Association rule มีแนวคิดมาจากการนับจำนวนครั้งของรายการที่เกิดขึ้น และรวมเหตุการณ์ที่เป็นไปได้เข้าด้วยกัน ทำให้กฎที่ได้มีจำนวนมาก เทคนิค "Pruning" จะช่วย

ลดจำนวนกฎที่ไม่ตรงกับเงื่อนไขลงได้ อัลกอริทึมทั่วไปมักจะให้ผู้ใช้ระบุค่า "Minimum Support" และ "Minimum Confidence" เพื่อให้เวลาที่ใช้ในการคำนวณการกฎไม่มากเกินไป

เทคนิค "Pruning" ที่ใช้กันมากที่สุดคือ "Minimum Support Pruning" จะเป็นตัวกำหนดว่ากฎที่ได้จะต้องมาจากรายการที่มีการเกิดอย่างน้อยเท่ากับ Minimum Support เช่น มี 100 รายการ กำหนดค่า Minimum Support เท่ากับ 2% ดังนั้น กฎที่สนใจจะต้องมีค่า Support อย่างน้อย 2 รายการ

ข้อดีของ Association Rule

1) ทำงานได้ดีกับข้อมูลขนาดใหญ่ ขณะที่เทคนิคอื่นๆจะมีปัญหาการทำงานกับข้อมูลปริมาณมาก
2) ผู้ใช้สามารถระบุค่า Minimum Support และ Minimum Confidence ได้ทำให้ควบคุมผลลัพธ์ได้
3) สามารถทำการไมนิ่งกับข้อมูลบางส่วนได้เพื่อลดปัญหากรณีที่มีข้อมูลไม่สมบูรณ์ได้
4) สามารถจัดการกับข้อมูลที่รูปแบบต่างกันได้โดยไม่เสียสารสนเทศ
5) ง่ายต่อการทำความเข้าใจเพราะใช้สัญลักษณ์ในการแสดงผล

ข้อเสียของ Association Rule

1) ถ้าใช้กับข้อมูลที่เกิดขึ้นไม่บ่อยในทรานแซ็กชั่นจะทำให้ข้อมูลนี้แยกออกมาจากกลุ่มข้อมูลอย่างชัดเจนทำให้ประสิทธิภาพลดน้อยลง
2) กฎที่ได้จากอัลกอริทึมนี้มีมากเกินไป ถ้าผู้ใช้กำหนด Minimum Support และ Minimum Confidence สูงหรือต่ำเกินไปจะทำให้กฎที่ได้ผิดเพี้ยน
3) บอกความแตกต่างของกฎที่ได้มายากว่าเป็นกฎจริงหรือกฎที่บังเอิญข้อมูลมาพ้องกัน
4) กฎบอกได้เพียงอะไรมีแนวโน้มที่จะเกิดขึ้นด้วยกันไม่ได้ให้สารสนเทศที่เป็นเหตุเป็นผล

3.2.2 อัลกอริทึม Apriori

การวิเคราะห์ข้อมูลด้วยเทคนิค Association Rule โดยอาศัยอัลกอริทึม Apriori มีหลักการทำงาน 2 ขั้นตอนคือ

1) การหาชุดของข้อมูล (Itemset) ในรายการขายที่มีค่าความถี่ในการเกิดมากกว่าหรือเท่ากับค่า Support ที่น้อยที่สุด (Minimum Support) โดยจะเรียก Itemset นี้ว่า "Frequent Itemset" หรือ "Large Itemset"

2) การนำ "Large Itemset" มาสร้างเป็นกฎบนพื้นฐาน ถ้า ABCD และ AB เป็น "Frequent Itemset" สามารถสร้างกฎ AB → CD โดยคำนวณค่า Confidence จาก Support(ABCD) / Support(AB) กฎที่ได้จะถูกต้องเมื่อมีค่า Confidence มากกว่าหรือเท่ากับค่า Confidence ที่น้อยที่สุด (Minimum Confidence)

ตัวแปรที่เกี่ยวข้องมีดังนี้

D	คือ ฐานข้อมูล แต่ละทรานแซ็กชั่นเก็บ <TID,Items>
TID	คือ ตัวเลขระบุรายการทรานแซ็กชั่น
Size	คือ จำนวน Item ในเซ็ตของข้อมูล
k-Itemset	คือ เซ็ตของข้อมูลที่แต่ละเซ็ตประกอบด้วยสมาชิกจำนวน k ตัว
L_k	คือ เซ็ตของ Frequent k Itemset ซึ่งทุกเซ็ตมีความถี่ในการเกิดมากกว่าหรือเท่ากับค่า Minimum Support
C_k	คือ เซ็ตของ Candidate k Itemset ที่ถูกเลือกมาจาก L_k

ขั้นตอนการทำงานของอัลกอริทึม Apriori

1) ขั้นตอนการเชื่อม (Join Step) เป็นขั้นตอนการสร้าง C_k จากการเชื่อมกันของ L_{k-1} กับเซตของ L_{k-1} เอง
2) ขั้นตอนการตัดทิ้ง (Prune Step) เป็นขั้นตอนของการตัดเซตสมาชิกใน C_k ที่มีความถี่น้อยกว่าค่า Minimum Support ออก เพื่อสร้างเป็น L_k

จะทำเช่นนี้ไปเรื่อยๆจนกระทั่งไม่พบเซตที่มีมากกว่า k-itemset เมื่อจบการทำงานจะได้เป็น L_k จึงนำเซตข้อมูล L_k ที่ได้มาสร้างเป็นกฎความสัมพันธ์

Pseudo code ของอัลกอริทึม Apriori

จะแสดงได้ดังรูปที่ 3.3

Main Apriori Algorithm

L_1 = {large 1-itemset}

For (k = 1 ; $L_k \neq \varnothing$; k++) do begin

C_{k+1} = Generate candidate (L_k) ;

For all transactions t in database do

Increment the count of all candidates

In C_{k+1} that are contained in t ;

L_{k+1} = {candidates in C_{k+1} with minimum support}

End

Return $\cup_k L_k$;

Generate candidate function

- **Join Step**

Insert into C_k

Select $p.item_1$, $p.item_2$,..., $p.item_{k-1}$,$q.item_{k-1}$

From L_{k+1} p , L_{k+1} q

Where $p.item_1 = q.item_1$,...,$p.item_{k-2} = q.item_{k-2}$,

$p.item_{k-1} < q.item_{k-1}$

- **Prune Step**

For all itemset c in C_k do

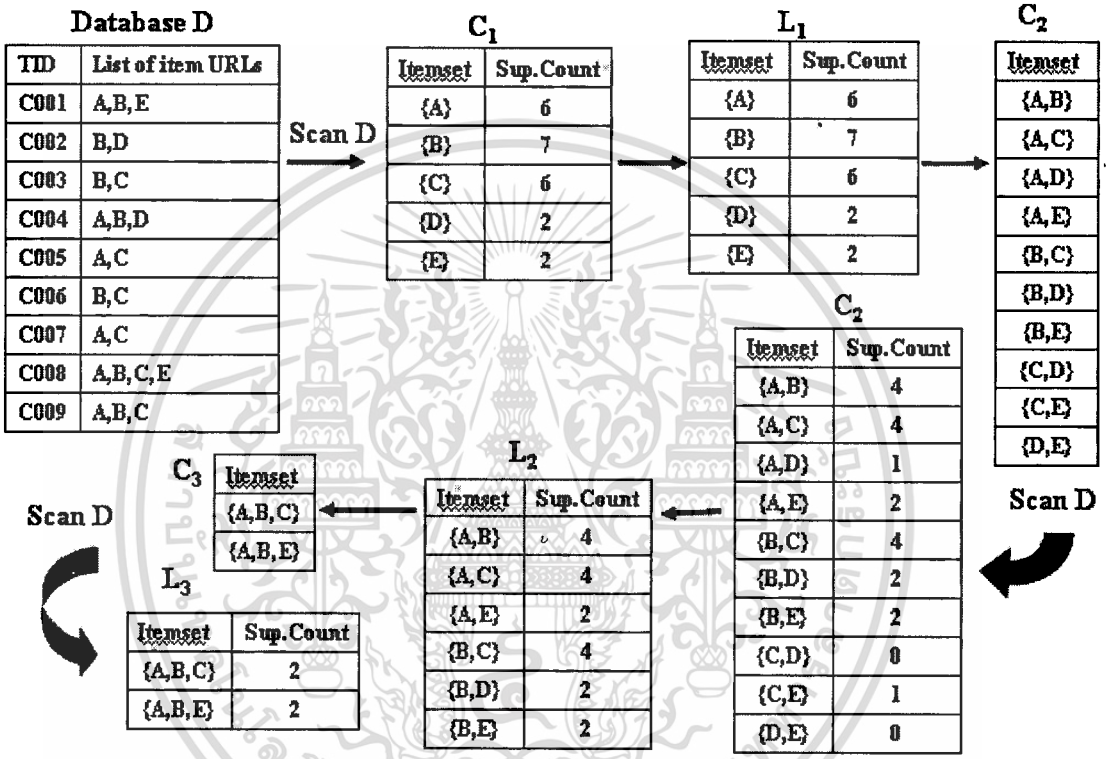
For all (k-1)-subsets s of c do

If (s is not in L_{k-1}) then delete c from C_k ;

รูปที่ 3.3 Pseudo code ของอัลกอริทึม Apriori

ตัวอย่างข้อมูลที่ผ่านการคำนวณจากอัลกอริทึม Apriori เป็นดังนี้

กำหนดค่า Minimum Support เท่ากับ 2 ได้ดังรูปที่ 3.4



รูปที่ 3.4 แสดงตัวอย่างข้อมูลที่ผ่านการคำนวณจากอัลกอริทึม Apriori

3.2.3 สร้างกฎความสัมพันธ์ (Generating rule)

นำ Frequent Itemset ที่ได้มาสร้างเป็นกฎ โดยนำค่า Frequent Itemset ตั้งแต่ L2 เป็นต้นไป มาคำนวณหา Subset และนำแต่ละ Subset ที่ได้มาสร้างเป็นกฎ อัลกอริทึมการทำงานแสดงได้ดังรูปที่ 3.5 และแสดงตัวอย่างข้อมูลที่ได้หลังจากสร้างกฎได้ดังรูปที่ 3.6

```
//Fastest Algorithm
forall frequent k-itemset l_k,k ≥ 2 do begin
        H_1 = {consequence of rules derived from l_k with one item in the consequent};
        Call ap-genrules(l_k,H_1);
End

//genrules generates all valid rules
procedure ap-genrules (l_k : frequent k-itemset , H_m : set of m-item consequence)
        if (k > m+1) then begin
                H_m+1 = Apriori-gen(H_m);
                Forall h_m+1 ∈ H_m+1 do begin
                        Conf = support(l_k) / support(l_k - h_m+1);
                        If (conf ≥ minconf) then begin
                                Output the rule (l_k – h_m+11 ⇒ h_m+1) , with confidence = conf
                                                And support = support(l_k);
                        Else
                                Delete h_m+11 from H_m+1;
                End
                Call ap-genrules(l_k,H_m+1);
        end
```

รูปที่ 3.5 แสดงอัลกอริทึมการทำงานของ genrule

ตัวอย่างข้อมูลที่ได้หลังจากสร้างกฎ

จาก L2

Itemset	Sup.Count
{A,B}	4
{A,C}	4
{A,E}	2
{B,C}	2
{B,D}	2

Gen Rules →

Rules	Confidence	Support
A → B	0.67	4
B → A	0.57	4
A → C	0.67	4
C → A	0.67	4
A → E	0.33	4
E → A	1	4
B → C	0.28	2
C → B	0.33	2
B → D	0.28	2
D → B	1	2

จาก L3

Itemset	Sup.Count
{A,B,C}	2
{A,B,E}	2

Gen Rules →

Rules	Confidence	Support
A → B & C	0.33	2
B → A & C	0.28	2
C → A & B	0.33	2
A & B → C	0.5	2
A & C → B	0.5	2
B & C → A	1	2
A → B & E	0.33	2
B → A & E	0.28	2
E → A & B	1	2
A & B → E	0.5	2
A & E → B	1	2
B & E → A	1	2

รูปที่ 3.6 แสดงผลลัพธ์กฎที่ถูก generate ออกมา

บทที่ 4

การวิเคราะห์และออกแบบระบบงานสำหรับไมนิ่งข้อมูลของ Proxy log

การพัฒนาเครื่องมือสำหรับไมนิ่งข้อมูลของ proxy log โดยใช้เทคนิค Association Rule เป็นระบบที่พัฒนาขึ้นเพื่อหารูปแบบการเดินทางของการเรียกใช้เว็บเพจของผู้ใช้ทำให้ทราบพฤติกรรมการเรียกดูเพจของผู้ใช้ภายในองค์กร และเพื่อหาความสัมพันธ์ของเว็บเพจที่ผู้ใช้เรียกดูทำให้สามารถนำผลที่ได้ไปประยุกต์ใช้กับการเก็บ cache ใน Proxy server ได้

4.1 ขอบเขตการทำงานของเครื่องมือสำหรับการวิเคราะห์ข้อมูลใน Proxy แบ่งได้เป็น 3 ส่วนย่อยดังนี้

1) ส่วนการรับข้อมูลจาก Proxy Log

- เป็นส่วนที่เลือกไฟล์ proxy log ซึ่งเป็นไฟล์ text และอิมพอร์ต proxy log ลงฐานข้อมูลที่เตรียมไว้
- โดยก่อนที่จะบันทึกลงฐานข้อมูล ไฟล์ text จะต้องมีรูปแบบตามที่กำหนดไว้คือรูปแบบเช่นเดียวกับไฟล์ access.log ของ Squid server (Linux) โดยในวิชาโครงการศึกษากรณีพิเศษนี้ จะใช้ ไฟล์ access.log ของ squid2.4.STABLE6
- การบันทึกลงฐานข้อมูลจะบันทึกเฉพาะข้อมูลที่มีประโยชน์ในการวิเคราะห์หาความสัมพันธ์ของข้อมูลเท่านั้น
- มีการเอ็กซ์พอร์ตข้อมูล proxy log ที่ Cleaning แล้ว ออกมาเป็นไฟล์ text

2) ส่วนการหารูปแบบเส้นทางการเดินทางของการเรียกดูเว็บเพจของผู้ใช้

- เป็นส่วนที่นำ log การเรียกดูเว็บเพจของผู้ใช้ มาแยกเป็นทรานแซ็กชั่นโดยการระบุผู้ใช้ โดยแยกผู้ใช้ตาม IP Address และระบุเซสชั่นตามขอบเขตเวลาที่ผู้ใช้กำหนด (เป็นนาที)
- นำทรานแซ็กชั่นที่ได้มาหารูปแบบเส้นทางการเดินทางของการเรียกดูเว็บเพจ โดยใช้อัลกอริทึม Maximal forward reference
- เมื่อได้รูปแบบเส้นทางการเดินทางแล้วจะบันทึกผลลัพธ์ที่ได้ลงฐานข้อมูลเพื่อจะนำไปวิเคราะห์หาความสัมพันธ์ของเว็บเพจที่ผู้ใช้เรียกดู
- มีการเอ็กซ์พอร์ตข้อมูลรูปแบบเส้นทางการเดินทางออกมาเป็นไฟล์ text

3) ส่วนการวิเคราะห์หาความสัมพันธ์ของเว็บเพจที่ผู้ใช้เรียกดู

- การวิเคราะห์หาความสัมพันธ์ของเว็บเพจที่ผู้ใช้เรียกดูจะนำข้อมูลรูปแบบการเดินทางของผู้ใช้มาวิเคราะห์โดยใช้เทคนิค Association rule ซึ่งจะใช้อัลกอริทึม Apriori โดยผู้ใช้สามารถกำหนด Minimum Support และ Minimum Confidence เอง
- จะบันทึกผลลัพธ์ที่เป็นกฎความสัมพันธ์ลงฐานข้อมูล
- มีการเอ็กซ์พอร์ตข้อมูลกฎความสัมพันธ์ออกมาเป็นไฟล์ text

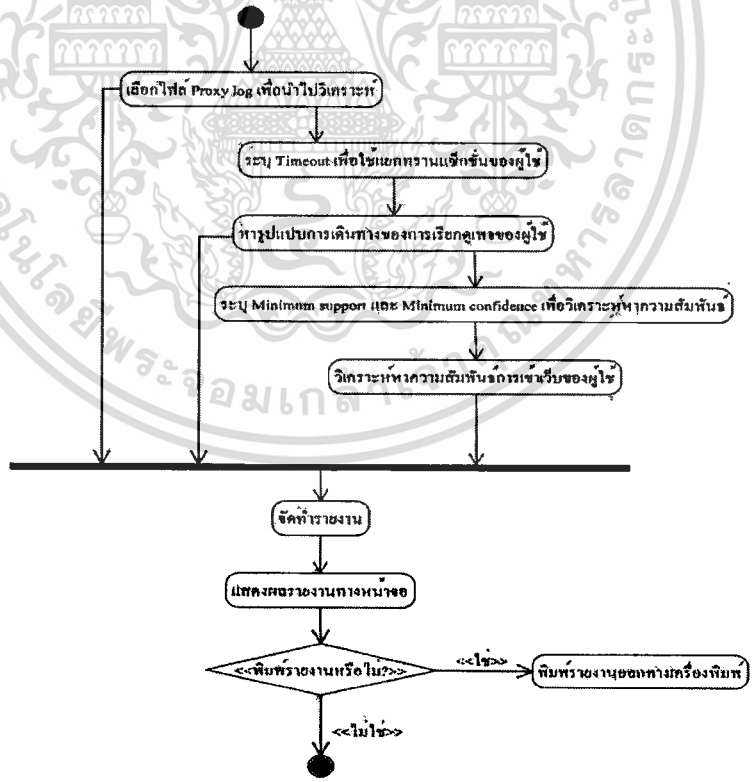
4.2 การวิเคราะห์ระบบ

4.2.1 การกำหนด Actor

- ผู้ใช้ระบบ เป็นผู้ใช้ระบบงานไมนิ่งข้อมูลของ Proxy log

4.2.2 การกำหนด Activity Diagram

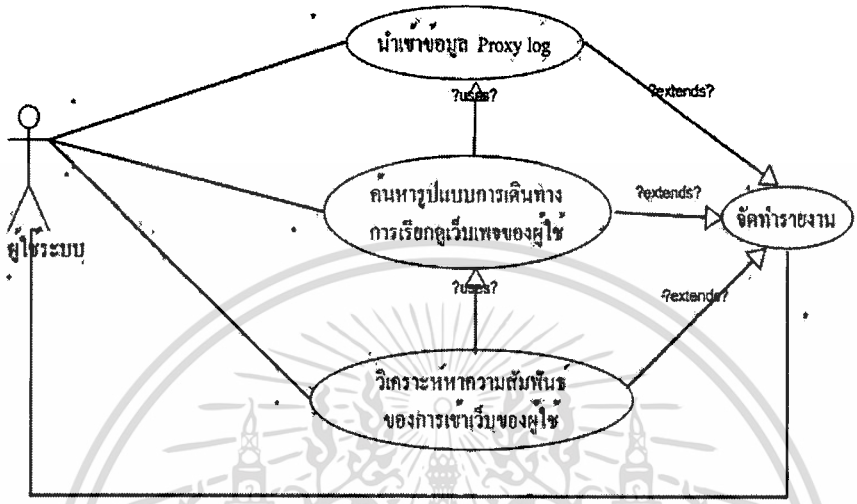
สามารถแสดงได้ดังรูปที่ 4.1



รูปที่ 4.1 แสดง Activity Diagram ของระบบไมนิ่งข้อมูลของ proxy log

4.2.3 การสร้าง Use-Case Diagram

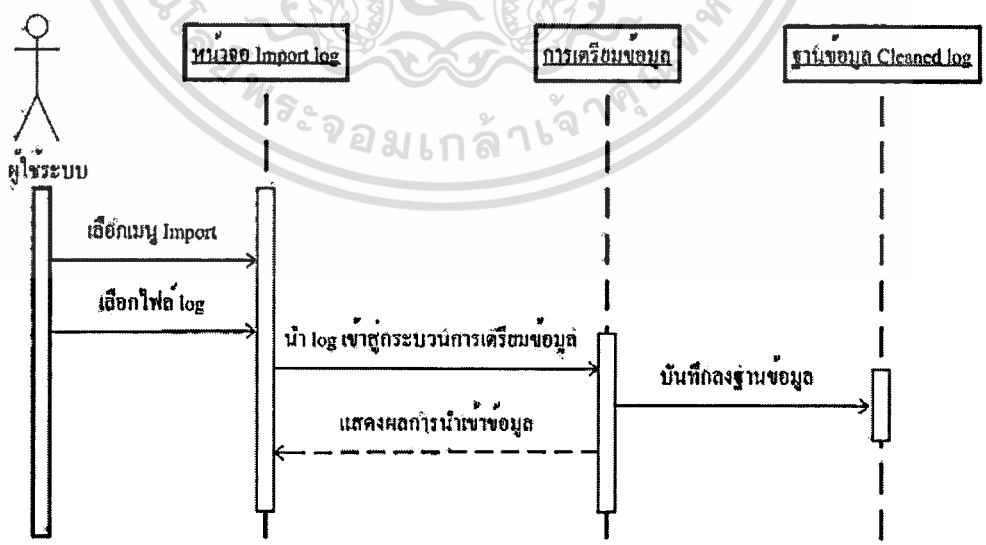
สามารถแสดงได้ดังรูปที่ 4.2



รูปที่ 4.2 แสดง Use case Diagram ของระบบไมนิ่งข้อมูลของ proxy log

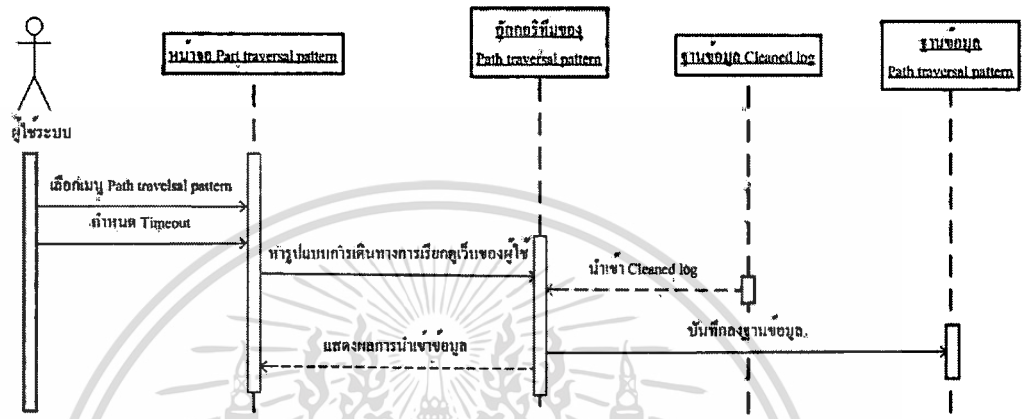
4.2.4 การสร้าง Sequence Diagram สำหรับแต่ละ Use case

4.2.4.1 Sequence Diagram ของ Use case นำเข้าข้อมูล proxy log แสดงได้ดังรูปที่ 4.3



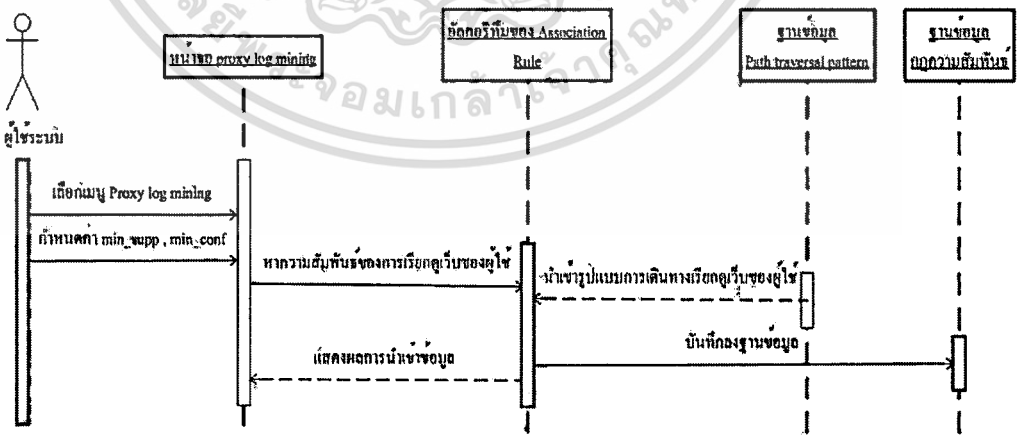
รูปที่ 4.3 แสดง Sequence Diagram ของ Use case นำเข้าข้อมูล proxy log

4.2.4.2 Sequence Diagram ของ Use case ค้นหารูปแบบการเดินทางของการเรียกดูเว็บเพจของผู้ใช้ แสดงได้ดังรูปที่ 4.4



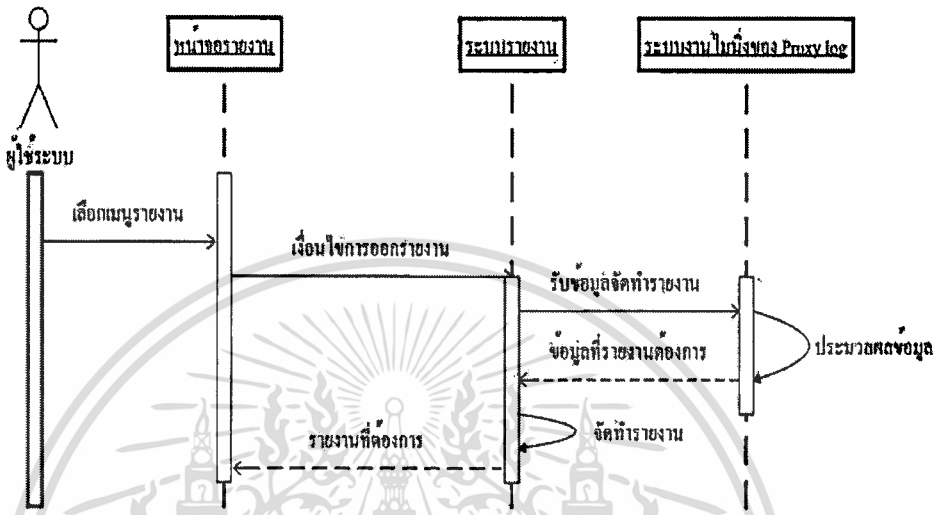
รูปที่ 4.4 แสดง Sequence Diagram ของ Use case ค้นหารูปแบบการเดินทางของการเรียกดูเว็บเพจของผู้ใช้

4.2.4.3 Sequence Diagram ของ Use case การวิเคราะห์หาความสัมพันธ์ของการเข้าเว็บของผู้ใช้ แสดงได้ดังรูปที่ 4.5



รูปที่ 4.5 แสดง Sequence Diagram ของ Use case การวิเคราะห์หาความสัมพันธ์ของการเข้าเว็บของผู้ใช้

4.2.4.4 Sequence Diagram ของ Use case จัดทำรายงาน แสดงได้ดังรูปที่ 4.6

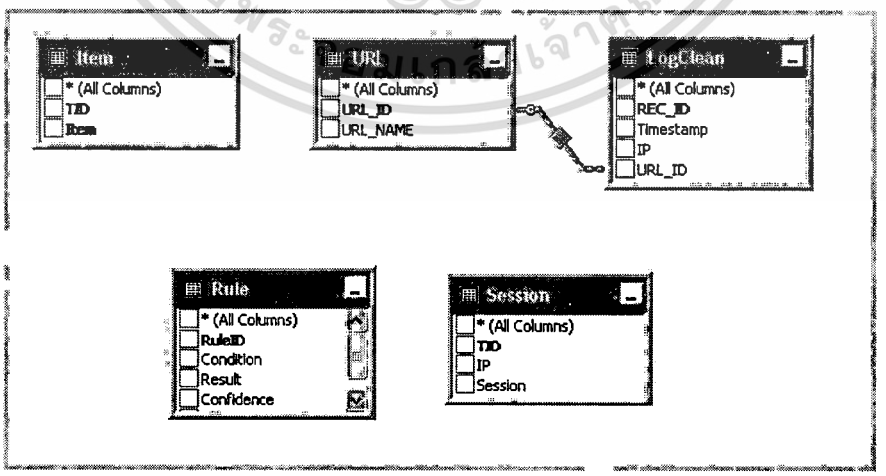


รูปที่ 4.6 แสดง Sequence Diagram ของ Use case จัดทำรายงาน

4.3 การออกแบบฐานข้อมูล

4.3.1 ความสัมพันธ์ของฐานข้อมูล

ความสัมพันธ์ของฐานข้อมูลระบบไมนิ่งข้อมูลของ Proxy log สามารถแสดง ER Diagram ได้ดังรูปที่ 4.7



รูปที่ 4.7 แสดง ER Diagram

4.3.2 ตารางในฐานข้อมูล

ความหมายของสัญลักษณ์ที่ใช้ในตาราง

Name คือ คำอธิบายแอตทริบิวต์

Code คือ ชื่อแอตทริบิวต์

Type คือ ชนิดของข้อมูล

Length คือ ขนาดของข้อมูล

P คือ Primary Key

F คือ Foreign Key

1) **ตารางข้อมูลเว็บเพจ** : เก็บข้อมูลเกี่ยวกับรหัสและชื่อเว็บเพจที่ผู้ใช้เรียกดูใน proxy log

ตารางที่ 4.1 ข้อมูลเว็บเพจ

Table Name : ข้อมูลเว็บเพจ **Table Code :** URL					
Name	**Code**	**Type**	**Length**	**P**	**F**
รหัสเว็บเพจ	URL_ID	INTEGER	4	Yes	Yes
ชื่อเว็บเพจ	URL_NAME	VARCHAR(900)	900	No	No

2) **ตารางข้อมูล LogCLean** : เก็บข้อมูล log entry ที่ผ่านการทำความสะอาดจาก proxy log แล้ว

ตารางที่ 4.2 ข้อมูล LogClean

Table Name : ข้อมูล LogClean **Table Code :** LogClean					
Name	**Code**	**Type**	**Length**	**P**	**F**
หมายเลข Record	REC_ID	INTEGER	4	Yes	No
เวลาที่ผู้ใช้ร้องขอเว็บเพจ	Timestamp	VARCHAR(25)	25	No	No
หมายเลข IP ที่ร้องขอเว็บเพจ	IP	VARCHAR(15)	15	No	No
รหัสเว็บเพจที่ร้องขอ	URL_ID	INTEGER	4	No	Yes

3) **ตารางข้อมูล Session** : เก็บข้อมูลเส้นทางการเดินทางที่สมบูรณ์ของการเรียกดูเว็บเพจของผู้ใช้แต่ละคน

ตารางที่ 4.3 ข้อมูล Session

Table Name : ข้อมูล Session **Table Code :** Session					
Name	**Code**	**Type**	**Length**	**P**	**F**
หมายเลขทรานแซ็กชั่น	TID	INTEGER	4	Yes	Yes
หมายเลข IP ที่ร้องขอเว็บเพจ	IP	VARCHAR(15)	15	No	No
เส้นทางการเดินทางเว็บเพจของผู้ใช้	Session	VARCHAR(8000)	8000	No	No

4) **ตารางข้อมูล Item** : เก็บข้อมูลรูปแบบการเดินทางการเรียกดูเว็บเพจหลังจากใช้อัลกอริทึมของ Path forward reference

ตารางที่ 4.4 ข้อมูล Item

Table Name : ข้อมูล Item **Table Code :** Item					
Name	**Code**	**Type**	**Length**	**P**	**F**
หมายเลขทรานแซ็กชั่น	TID	INTEGER	4	Yes	Yes
รหัสเว็บเพจที่ผู้ใช้เรียก	Item	INTEGER	4	Yes	Yes

5) **ตารางข้อมูล Rule** : เก็บกฎความสัมพันธ์ของการเรียกดูเว็บของผู้ใช้

ตารางที่ 4.5 ข้อมูล Rule

Table Name : ข้อมูล Rule **Table Code :** Rule					
Name	**Code**	**Type**	**Length**	**P**	**F**
หมายเลขกฎ	RuleID	INTEGER	4	Yes	No
ส่วนเงื่อนไขของกฎ	Condition	VARCHAR(4000)	4000	No	No
ส่วนผลลัพธ์ของกฎ	Result	VARCHAR(4000)	4000	No	No
ค่า Confidence ของกฎ	Confidence	FLOAT	8	No	No
ค่า Support ของกฎ	Support	FLOAT	8	No	No

4.4 การเตรียมข้อมูล

การทำงานของระบบที่มี proxy server ติดตั้งอยู่จะใช้หลักการของ client/server ที่มีเครื่อง proxy server เป็นผู้ให้บริการแก่ client ซึ่งในที่นี้คือ HTTP client โดยบริการในที่นี้หมายถึงการเป็นตัวแทนของ client เหล่านั้นในการเรียกข้อมูลจากเครื่อง Web server มาให้แก่ client ตามที่ได้ร้องขอมา ทั้งนี้ proxy server ยังสามารถทำการจัดเก็บข้อมูลที่ได้มาเหล่านั้นไว้ชั่วระยะเวลาหนึ่ง (ใน cache) ซึ่งถ้ามีการร้องขอข้อมูล (object) เดียวกันเข้ามา proxy server ก็สามารถนำ object ที่จัดเก็บไว้ส่งให้ client ได้เลยโดยไม่ต้องไปดึงมาจาก Web serer ภายนอกอีก ในกรณีนี้จะเรียกว่าพบ (HIT) object ใน cache แต่ถ้า object ที่ client ร้องขอมาไม่มีอยู่ใน cache หรือเป็น object ใหม่ (เกิดกรณีที่ object มีการเปลี่ยนแปลงขนาดหรือวันที่สร้าง) ก็จะเกิดกรณีไม่พบ (MISS) และ proxy server ก็จะต้องไปดึงข้อมูลภายนอกเข้ามาให้กับ client ใหม่

เมื่อ proxy server ทำงานตามการร้องขอจาก client แต่ละรายการเสร็จก็จะทำการบันทึกผลการทำงานของการร้องขอนั้นลงสู่ไฟล์ log เพื่อแสดงรายละเอียดต่างๆ ซึ่งในโครงการพัฒนาระบบงานนี้จะใช้ไฟล์ชื่อ access.log ของ Squid เวอร์ชั่น 2.4.STABLE6 ซึ่งทำงานเป็น proxy server ของบริษัทสามารถคอมเทค โดยในไฟล์ access.log 1 บรรทัดจะประกอบด้วย 10 ฟิลด์ดังนี้

Timestamp Elapsed ClientIP Action/HTTP-Code

Size Req-Method URL Ident Hierarchy/Hostname Content-Type

โดยฟิลด์ต่างๆจะอธิบายตามลำดับดังนี้

Timestamp คือ เวลาที่การ request นั้นเสร็จสิ้นสมบูรณ์ (วินาที จาก 1 มกราคม 1970)

Elapsed คือ ช่วงเวลาตั้งแต่การ accept และ close ของ socket ในฝั่ง client

ClientIP คือ IP Address ของเครื่อง client ที่ทำการ request

Action คือ ลักษณะของการจัดการ request

HTTP-Code คือ code การ reply ของ HTTP

Size คือ ขนาดของข้อมูล (Byte)

Req-Method คือ วิธีการ request ของ HTTP

URL คือ ตำแหน่งของ object ที่ client ทำการ request

Ident คือ Username ของ client ที่ connect ถ้าทำการ disable ident จะแสดงเป็น "-" แทน

Hierarchy คือ อธิบายวิธีการได้รับ Object มา เช่น direct

Hostname คือ Hostname ที่ client ทำการ request ขอ object

Content-Type คือ ชนิดของเนื้อหาของ object เช่น jpg , gif , text

ตัวอย่างของ access.log เป็นดังตารางที่ 4.6

ตารางที่ 4.6 แสดงตัวอย่างไฟล์ access.log

1054659715.450	10	10.0.0.62	TCP_IMS_HIT/304	221	GET	http://www.pantip.com/cafe/toy/image/songkarn11.gif	-	NONE/-	image/gif
1054659715.565	5	10.0.0.62	TCP_IMS_HIT/304	221	GET	http://www.pantip.com/cafe/toy/image/songkarn12.gif	-	NONE/-	image/gif
1054659715.648	27	10.0.0.62	TCP_CLIENT_REFRESH_MISS/304	209	GET	http://www.pantip.com/cafe/toy/image/songkarn14.gif	-	NONE/-	image/gif
1054685529.169	31	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	209	GET	http://www.eric.chula.ac.th/gisthai/article/interview3.htm	-	DIRECT/161.200.192.1	-
1054685530.157	31	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	209	GET	http://www.eric.chula.ac.th/gisthai/bookshelf/g_gover.html	-	DIRECT/161.200.192.1	-
1054685534.200	30	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	221	GET	http://www.eric.chula.ac.th/gisthai/research/index.html	-	DIRECT/161.200.192.1	-
1054687590.935	5	10.0.2.52	TCP_IMS_HIT/304	221	GET	http://chat.sanook.com/5/header.html	-	NONE/-	text/html
1054687591.524	4	10.0.2.52	TCP_IMS_HIT/304	221	GET	http://chat.sanook.com/5/chatform.html	-	NONE/-	text/html
1054687591.834	24	10.0.2.52	TCP_IMS_HIT/304	222	GET	http://chat.sanook.com/images/sanookchat.jpg	-	NONE/-	image/gif

การเตรียมข้อมูลก่อนนำข้อมูลไปวิเคราะห์สามารถแบ่งได้ 3 ส่วน

4.4.1 ส่วนที่นำเข้าข้อมูลดิบจากไฟล์ access.log มาบันทึกลงฐานข้อมูล

ก่อนที่จะบันทึกข้อมูลดิบลงฐานข้อมูลจะต้องนำไฟล์ access.log มาทำขั้นตอนดังนี้

1) การทำความสะอาดข้อมูล (Data Cleaning)

การทำความสะอาดข้อมูลจะทำการกำจัดไอเท็มที่ไม่เกี่ยวข้องออกไป เพื่อให้ได้ภาพที่ถูกต้องแม่นยำของการเข้าถึงของผู้ใช้ที่แท้จริง เนื่องจากโปรโตคอล HTTP มีความต้องการการเชื่อมต่อที่แยกจากกันของแต่ละไฟล์ที่ถูกร้องขอจาก server ดังนั้นการร้องขอของผู้ใช้เพื่อที่จะดูเพจหนึ่งๆ จะทำให้เกิดผลลัพธ์ใน log หลายๆ entry เนื่องจากมีการดาวน์โหลดกราฟฟิกและสคริปต์ต่างๆ เข้ามาเพิ่มเติมจากเพจที่ผู้ใช้ต้องการจริงๆ เนื่องจากในที่นี้เราต้องการที่จะค้นหาพฤติกรรมการเดินทางของผู้ใช้จริงๆ ซึ่งก็คือการร้องขอเพจที่เกิดจากผู้ใช้ไม่ใช่เกิดจากเทรกในเพจ ดังนั้นเราจึงกำจัดไอเท็มที่ไม่เกี่ยวข้องออกไปโดยการตรวจสอบที่ฟิลด์ Content-Type และ Req-Method ของไฟล์ access.log ร่วมกับการตรวจสอบส่วนต่อท้าย (suffix) ของชื่อ URL ซึ่งในการทำงานของโปรแกรมนี้จะเลือกเฉพาะไอเท็มที่ Content-Type เป็นชนิด /text/… , Req-Method เป็น “GET” และมีส่วนต่อท้ายที่ไม่ใช่พวกกราฟฟิกและสคริปต์เท่านั้นมาใช้งาน ผลลัพธ์จากการทำความสะอาดข้อมูลดิบในตารางที่ 4.7 จะได้ดังรูปที่ 4.8

ตารางที่ 4.8 แสดงข้อมูล log หลังจากการทำความสะอาดแล้ว

1054685529.169	31	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	209	GET	http://www.eric.chula.ac.th/gisthai/article/interview3.htm	-	DIRECT/161.200.192.1	-
1054685530.157	31	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	209	GET	http://www.eric.chula.ac.th/gisthai/bookshelf/g_gover.html	-	DIRECT/161.200.192.1	-
1054685534.200	30	10.0.3.114	TCP_CLIENT_REFRESH_MISS/304	221	GET	http://www.eric.chula.ac.th/gisthai/research/index.html	-	DIRECT/161.200.192.1	-
1054687590.935	5	10.0.2.52	TCP_IMS_HIT/304	221	GET	http://chat.sanook.com/5/header.html	-	NONE/-	text/html
1054687591.524	4	10.0.2.52	TCP_IMS_HIT/304	221	GET	http://chat.sanook.com/5/chatform.html	-	NONE/-	text/html

2) การนำ Cleaned log มาบันทึกลงฐานข้อมูล

การบันทึกข้อมูลลงฐานข้อมูลจะทำการตัดฟิลด์ที่ไม่จำเป็นในการวิเคราะห์หาความสัมพันธ์ทิ้งไปเช่น Elapsed , Action , Size ฯลฯ ซึ่งจะเหลือฟิลด์ที่ต้องบันทึกลงฐานข้อมูล ดังนี้ Timestamp , ClientIP และ URL โดยจะเก็บ URL เป็น URL_ID และบันทึกข้อมูลลงฐานข้อมูล “URL” และ “LogClean” ดังรูปที่ 4.8 และ 4.9 ตามลำดับ

Data in Table 'URL' in 'ProxyMining' on '(local)'

URL_ID	URL_NAME
1	http://www.pantip.com/ads/cafe/banner_jatujak.shtml
2	http://www.pantip.com/cafe/jatujak/topic/J2272194/J2272194.html
3	http://www.pantip.com/cafe/jatujak/topic/J2271883/J2271883.html
4	http://spweather.whenu.com/summary/TH?0/0002.html
5	http://akapp.whenu.com/ClockDB
6	http://wguts.weatherbug.com/DeskWx/GetStations.asp?
7	http://command.weatherbug.com/command/Commandv5.02.asp?
8	http://command.weatherbug.com/DeskWx/Announce.asp?
9	http://live.weatherbug.com/deskwx/guts/livepagetracking.asp?
10	http://wguts.weatherbug.com/ControlPanel/controlpanel.asp?
11	http://ww2.weatherbug.com/deskwx/multibutton/MultiButton.asp?
12	http://a.tribalfusion.com/f.ad?
13	http://ms101cfg.mysearch.com/ms101cfg.jsp?
14	http://daily.webshots.com/?
15	http://a1964.g.akamai.net/f/1964/2730/1h/app.whenu.com/OffersData?
16	http://www.pantip.com/
17	http://www.pantip.com/home.php
18	http://download.macromedia.com/pub/shockwave/cabs/flash/swflash.cab
19	http://adserve.inet.co.th/accipiter/hserver/SITE=pantip.com/AREA=pantip.index/aamsz=468x60
20	http://www.pantip.com/=
21	http://www.pantip.com/cafe/chalermthai/
22	http://www.pantip.com/cafe/chalermthai/main.php
23	http://www.pantip.com/cafe/menu.html
24	http://www.pantip.com/cafe/menu_tools.html
25	http://www.pantip.com/cafe/chalermthai/menu.shtml
26	http://www.pantip.com/cafe/chalermthai/home.shtml
27	http://www.pantip.com/cafe/chalermthai/Atopic.php
28	http://www.pantip.com/ads/cafe/banner_chalermthai.shtml
29	http://www.pantip.com/cafe/chalermthai/listA.php
30	http://www.pantip.com/cafe/chalermthai/listA.php?
31	http://www.pantip.com/cafe/chalermthai/topic/A2301102/A2301102.html
32	http://www.pantip.com/cafe/chalermthai/topic/A2301095/A2301095.html
33	http://www.pantip.com/cafe/chalermthai/topic/A2301088/A2301088.html

รูปที่ 4.8 แสดงตัวอย่างข้อมูลในตาราง "URL"

Data in Table 'LogClean' in 'ProxyMining' on '(local)'

REC_ID	Timestamp	IP	URL_ID
1	1054659621.287	10.0.0.62	1
2	1054659776.502	10.0.0.62	1
3	1054659837.493	10.0.0.62	1
4	1054659901.807	10.0.0.62	1
5	1054659969.576	10.0.0.62	1
6	1054660030.609	10.0.0.62	1
7	1054660085.280	10.0.0.62	2
8	1054660095.067	10.0.0.62	1
9	1054660124.450	10.0.0.62	3
10	1054660162.311	10.0.0.62	1
11	1054661086.777	10.0.8.123	4
12	1054662886.421	10.0.8.123	4
13	1054664686.079	10.0.8.123	4
14	1054666485.859	10.0.8.123	4
15	1054668285.501	10.0.8.123	4
16	1054670085.158	10.0.8.123	4
17	1054671884.814	10.0.8.123	4
18	1054673684.519	10.0.8.123	4
19	1054675484.210	10.0.8.123	4
20	1054677283.837	10.0.8.123	4
21	1054679083.488	10.0.8.123	4
22	1054680883.245	10.0.8.123	4
23	1054682238.693	10.0.7.65	5
24	1054682240.897	10.0.7.65	6
25	1054682244.180	10.0.7.65	7
26	1054682244.224	10.0.7.65	8
27	1054682249.092	10.0.7.65	9
28	1054682249.373	10.0.7.65	10
29	1054682250.103	10.0.7.65	11
30	1054682251.889	10.0.7.65	12
31	1054682255.870	10.0.7.65	12

รูปที่ 4.9 แสดงตัวอย่างข้อมูลในตาราง "LogClean"

4.4.2 ส่วนที่นำข้อมูลจาก Cleaned log ไปทำการหารูปแบบการเดินทางการเรียกดูเว็บของผู้ใช้

ก่อนที่จะนำข้อมูลจาก Cleaned log ไปหารูปแบบการเดินทางการเรียกดูเว็บของผู้ใช้ ต้องทำการเตรียมข้อมูลดังนี้

1) การระบุผู้ใช้ (User Identification)

เนื่องจากในการศึกษาครั้งนี้ไม่มีการคิดผลกระทบที่เกิดจาก cache , firewall และ proxy server เพราะฉะนั้นเราจึงระบุผู้ใช้โดยการดูจากฟิลด์ ClientIP เท่านั้น

2) การระบุเซสชั่น (Session Identification)

เนื่องจาก access.log ของเรามีการเก็บไฟล์ละ 1 วันซึ่งเป็นคาบเวลาที่ยาว มีความเป็นไปได้อย่างมากที่ผู้ใช้จะกลับมาเยี่ยมชมเว็บไซต์มากกว่า 1 ครั้ง โดยจุดมุ่งหมายของการระบุเซสชั่นคือต้องการการแบ่งการเข้าถึงเพจของแต่ละผู้ใช้ไปเป็นเซสชั่น โดยการระบุเซสชั่นเราจะใช้วิธี timeout ซึ่งในการศึกษาครั้งนี้จะให้ผู้ใช้เป็นผู้ระบุ timeout ที่ใช้ในการแบ่งเซสชั่นเอง ตัวอย่างของเซสชั่นของผู้ใช้แสดงได้ดังรูปที่ 4.10 โดยการกำหนด timeout เท่ากับ 30 นาที

Data in Table 'Session' in 'ProxyMining' on '(local)'

TID	IP	Session
1	10.0.0.252	485,487,493,333,334,335,579,334,580,664,485,53,49,146,752,757,758,146,146,757,758,146,146,758,757,786,790,791,802,8
2	10.0.0.252	485,907,912,913,821,485,487,493,495,495,495
3	10.0.0.252	1156,1157,1158,1159,1162,1164,485,1550,1551,257,1552
4	10.0.0.252	1620,1698,1703,485,487,1703,1703
5	10.0.0.253	457,458
6	10.0.0.253	1044
7	10.0.0.253	1259,257,1257,1268,1270,1269,1271,1273
8	10.0.0.62	1,1,1,1,1,1,2,1,3,1
9	10.0.0.62	1531,1533,1583,1584,1585,1586,1587,1587,1586,1587,1589,1589,1589,1589,1589,1589,1600,1602,1603,1604,1605,1606,158
10	10.0.0.63	1194,1195,1203,274,1216,1220,1261,1272,1275,1276,1288,1290,1480,228,1483,1484,1499,1499,1499,1499,1499,1540,1540,:
11	10.0.0.63	1638
12	10.0.0.8	349,588,589,592,593,353,349,349
13	10.0.0.86	5,340,707,365,713,714,723,724,741,743,746,748,749,750,751,333,335,334,579,334,580,759,760,761,764,765,760,774,713,7:
14	10.0.0.86	664,1045,707,707,1229,713
15	10.0.0.86	777,1277,511,513,511,513,786,790,791,1431,1432,1433,1434,1435,1436,1437,1438,1439,1440,1441,1442,1443,1444,1445,14
16	10.0.0.86	1640,1641,1660,1660,1668,1668,759,1669,760,1675,1676,1675,1677,1680,1681,760,1684,1685,760
17	10.0.0.90	543,544,546,548,543,609,611,612,616,625,627,629,543,609,612,611,616,625,629,627,543,543
18	10.0.0.90	548,883,916,970,916,548,916,548,916,548,945,1125,1126,1127
19	10.0.0.90	1183,1446,1461,1464,1465,1468,1482,1486,1487,1488,1493,1493,1493,1496,1500,1446,1501,1502,1513,1514,1516,1517,151
20	10.0.0.99	49,146,146,779,146,146,799,800,803,805,810,822,822,823,824,825,826,827,828,829,830,831,832,826,836,837,850,853,853,{
21	10.0.0.99	973,973,974,978,146,916,978
22	10.0.2.21	245,51,50,51,234,52,269,269,270,269,269,269,269,269,269,269,269,269,269,269,269,269,271,275,276,277,269,269,269,
23	10.0.2.21	50,51,234
24	10.0.2.22	274,228,278,279,280,281,282,283,284,269,269,269,269,270,269,269,269,269,269,269,269,269,269,269,269,271,289,290
25	10.0.2.52	235,236,237,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238,238

รูปที่ 4.10 แสดงตัวอย่างของเซสชั่นของผู้ใช้โดยกำหนด timeout เท่ากับ 30 นาที

3) การทำเส้นทางให้สมบูรณ์ (Path Completing)

เนื่องจากเราไม่คิดผลกระทบที่เกิดจากการกดปุ่ม "Back" บน Browser ดังนั้นจึงตั้งสมมติฐานว่าในไฟล์ access.log จะเก็บทุกๆการร้องขอของผู้ใช้ เพราะฉะนั้นจึงไม่มีการทำงานในส่วนนี้

4) การจัดรูปแบบ (Formatting)

ส่วนสุดท้ายของการเตรียมข้อมูลคือการจัดรูปแบบให้เหมาะสมโดยในที่นี้เราจะใช้อัลกอริทึม MF ในการสร้างทรานแซ็กชั่นที่บรรจุเส้นทางการเดินทางไปข้างหน้าไกลที่สุด (Maximal forward reference) จากตัวอย่างข้อมูลรูปที่ 4.10 จะนำข้อมูลมาจัดรูปแบบด้วยอัลกอริทึม MF จะได้ผลของเส้นทางการเดินทางไปข้างหน้าไกลที่สุดดังรูปที่ 4.11 ซึ่งจะเห็นว่าเซสชั่นของ TID ที่1ของรูปที่ 4.110จะจัดรูปแบบการเดินทางไปข้างหน้าไกลที่สุดได้เป็น TID ที่ 1 ถึง TID ที่ 5 ของรูปที่ 4.11

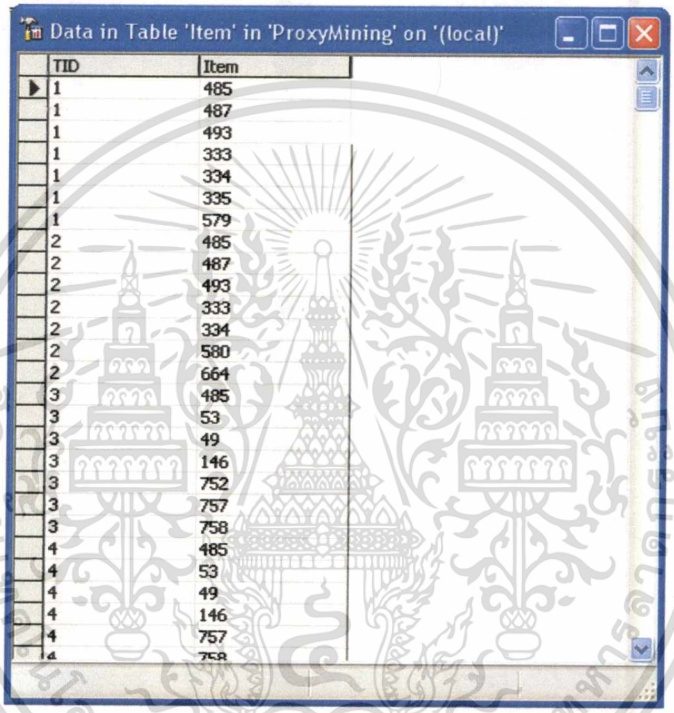
Data in Table 'Items' in 'ProxyMining' on '(local)'

TID	IP	Items
1	10.0.0.252	485,487,493,333,334,335,579
2	10.0.0.252	485,487,493,333,334,580,664
3	10.0.0.252	485,53,49,146,752,757,758
4	10.0.0.252	485,53,49,146,757,758
5	10.0.0.252	485,53,49,146,758,757,786,790,791,802,664,821
6	10.0.0.252	485,907,912,913,821
7	10.0.0.252	485,487,493,495
8	10.0.0.252	1156,1157,1158,1159,1162,1164,485,1550,1551,257,1552
9	10.0.0.252	1620,1698,1703,485,487
10	10.0.0.253	457,458
11	10.0.0.253	1044
12	10.0.0.253	1259,257,1257,1268,1270,1269,1271,1273
13	10.0.0.62	1,2
14	10.0.0.62	1,3
15	10.0.0.62	1531,1533,1583,1584,1585,1586,1587
16	10.0.0.62	1531,1533,1583,1584,1585,1586,1587,1589,1600,1602,1603,1604,1605,16
17	10.0.0.62	1531,1533,1583,1584,1585,1586,1587,1589
18	10.0.0.62	1531,1533,1583,1584,1585,1586,1587,1589,1602,1603,1604,1606,1642,16
19	10.0.0.63	1194,1195,1203,274,1216,1220,1261,1272,1275,1276,1288,1290,1480,228,
20	10.0.0.63	1194,1216,1220,1558,1560,48,44,45,46
21	10.0.0.63	1638
22	10.0.0.8	349,588,589,592,593,353
23	10.0.0.86	5,340,707,365,713,714,723,724,741,743,746,748,749,750,751,333,335,334
24	10.0.0.86	5,340,707,365,713,714,723,724,741,743,746,748,749,750,751,333,335,334
25	10.0.0.86	5,340,707,365,713,714,723,724,741,743,746,748,749,750,751,333,335,334

รูปที่ 4.11 แสดงตัวอย่างของเส้นทางการเดินทางไปข้างหน้าไกลที่สุด

5) การบันทึกข้อมูลลงฐานข้อมูล

หลังจากได้รูปแบบการเดินทางไปข้างหน้าไกลที่สุดจะบันทึกข้อมูลผลลัพธ์ที่ได้ลงฐานข้อมูลในตาราง "Item" เพื่อนำไปวิเคราะห์หาความสัมพันธ์ด้วยเทคนิค Association Rule ต่อไป ตัวอย่างข้อมูลที่เก็บลงฐานข้อมูลตาราง "Item" แสดงได้ดังรูปที่ 4.12


Data in Table 'Item' in 'ProxyMining' on '(local)'

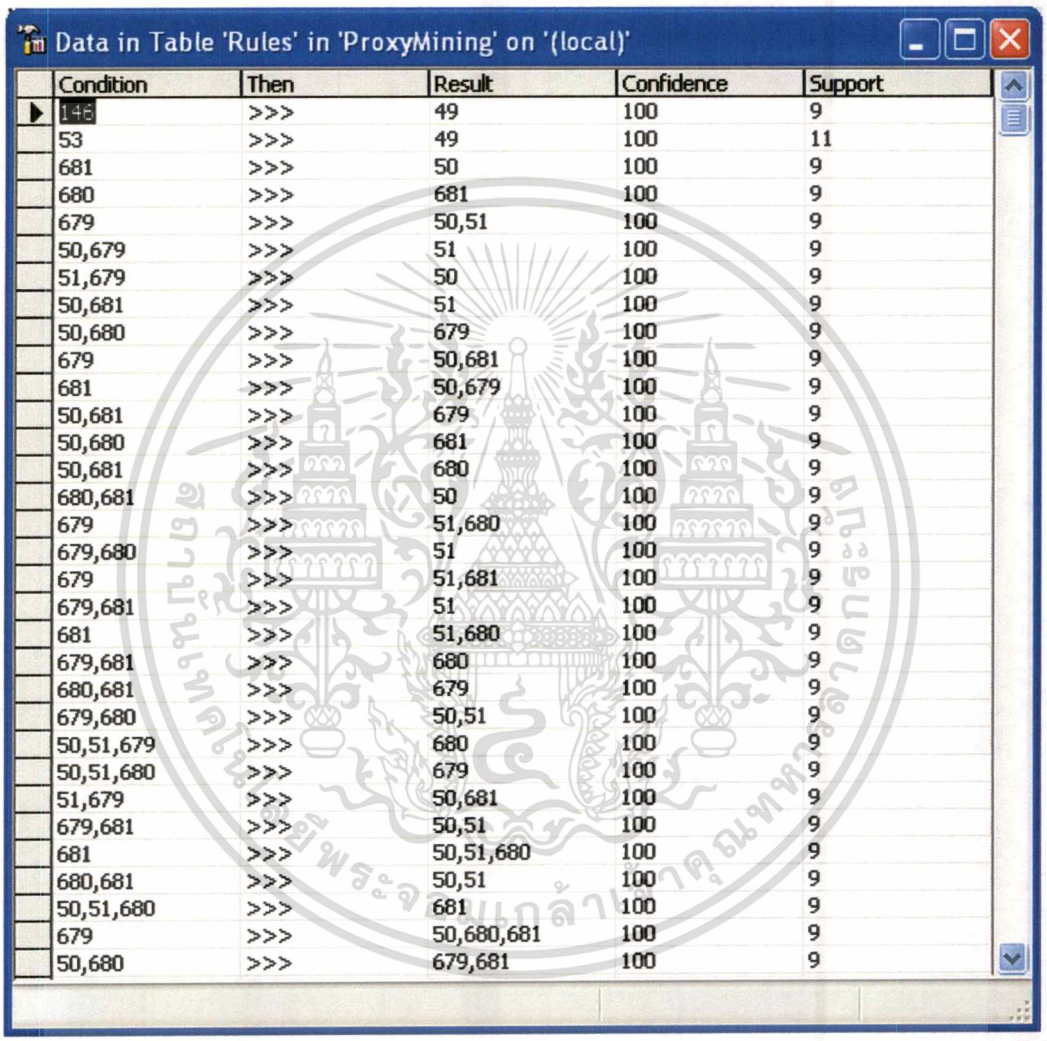
TID	Item
1	485
1	487
1	493
1	333
1	334
1	335
1	579
2	485
2	487
2	493
2	333
2	334
2	580
2	664
3	485
3	53
3	49
3	146
3	752
3	757
3	758
4	485
4	53
4	49
4	146
4	757

รูปที่ 4.12 แสดงตัวอย่างข้อมูลที่เก็บลงฐานข้อมูลตาราง "Item"

4.4.3 ส่วนที่นำรูปแบบการเดินทางการเรียกดูเว็บไปหากฎความสัมพันธ์การร้องขอเว็บของผู้ใช้

หลังจากได้รูปแบบการเดินทางไปข้างหน้าไกลที่สุดของผู้ใช้แล้วจะนำข้อมูลดังกล่าวไปหากฎความสัมพันธ์การร้องขอเว็บของผู้ใช้โดยรับค่า Minimum confidence , Minimum support และ จำนวน k-Itemset ที่ต้องการ เพื่อจำกัดจำนวนของรูปแบบเส้นทางการเดินทางที่ค้นพบและรายงานออกมา โดยผลลัพธ์ที่ได้ออกมาคือ Large reference sequence หรือเส้นทางการเดินทางที่มีจำนวนครั้งของที่ปรากฏมากกว่าค่าสนับสนุนที่ตั้งไว้ แล้วจะทำการบันทึกความสัมพันธ์ที่ได้จากการทำขั้นตอน generate rule

ลงฐานข้อมูลในตาราง “Rule” ตัวอย่างข้อมูลที่เก็บลงฐานข้อมูลตาราง “Rule” แสดงได้ดังรูปที่ 4.13 โดยใช้ค่า Minimum support = 20% และ Minimum confidence = 20%



Data in Table 'Rules' in 'ProxyMining' on '(local)'

Condition	Then	Result	Confidence	Support
146	>>>	49	100	9
53	>>>	49	100	11
681	>>>	50	100	9
680	>>>	681	100	9
679	>>>	50,51	100	9
50,679	>>>	51	100	9
51,679	>>>	50	100	9
50,681	>>>	51	100	9
50,680	>>>	679	100	9
679	>>>	50,681	100	9
681	>>>	50,679	100	9
50,681	>>>	679	100	9
50,680	>>>	681	100	9
50,681	>>>	680	100	9
680,681	>>>	50	100	9
679	>>>	51,680	100	9
679,680	>>>	51	100	9
679	>>>	51,681	100	9
679,681	>>>	51	100	9
681	>>>	51,680	100	9
679,681	>>>	680	100	9
680,681	>>>	679	100	9
679,680	>>>	50,51	100	9
50,51,679	>>>	680	100	9
50,51,680	>>>	679	100	9
51,679	>>>	50,681	100	9
679,681	>>>	50,51	100	9
681	>>>	50,51,680	100	9
680,681	>>>	50,51	100	9
50,51,680	>>>	681	100	9
679	>>>	50,680,681	100	9
50,680	>>>	679,681	100	9

รูปที่ 4.13 แสดงกฎความสัมพันธ์ในตาราง “Rule”

บทที่ 5
การพัฒนาระบบงาน

ในการพัฒนาเครื่องมือสำหรับไมนิ่งข้อมูลของ proxy log โดยใช้เทคนิค Association rule ในวิชาโครงการศึกษากรณีพิเศษนี้จะใช้การพัฒนาระบบในลักษณะของ Windows Application จึงได้มีการพัฒนาส่วนติดต่อผู้ใช้ทั้งหมดด้วย Microsoft Visual C# ของโปรแกรม Microsoft Visual Studio .net 2003 สาเหตุที่เลือกพัฒนาระบบด้วย Microsoft Visual C# มีสาเหตุเนื่องมาจาก มี Base Classes หรือ Class library พื้นฐาน ต่างๆเตรียมไว้ให้ใช้งานอย่างมากมาย ซึ่งผู้พัฒนาสามารถนำเอา Classes ต่างๆมาใช้งานได้ตามความเหมาะสม ซึ่งช่วยทำให้ลดระยะเวลาในการพัฒนาได้เป็นอย่างมาก อีกทั้ง Microsoft Visual C# เป็นเทคโนโลยีที่กำลังได้รับความนิยมในวงกว้าง ในการใช้งาน และเป็นภาษาที่ใช้ในการพัฒนาที่มีผู้ที่สามารถพัฒนาได้เป็นจำนวนมาก เพื่อไม่ให้เกิดปัญหาหากมีความจำเป็นต้องปรับปรุงแก้ไข เพิ่มเติม ระบบงานในภายหลัง อีกทั้งยังเป็นภาษาที่มีการพัฒนาอย่างต่อเนื่อง ตอบสนองได้กับการพัฒนาระบบตามวัตถุประสงค์ของโครงการศึกษากรณีพิเศษของผู้จัดทำ

5.1 องค์ประกอบต่างๆที่ใช้ในการพัฒนามีรายละเอียดดังนี้

- การพัฒนา Windows Application ใช้โปรแกรม Microsoft Visual Studio .net 2003
- ใช้ Microsoft SQL Server 2000 เป็นฐานข้อมูลหลักในการเก็บข้อมูล

5.2 การพัฒนาระบบไมนิ่งข้อมูลของ proxy log

จะมีหน้าจอหลักดังนี้

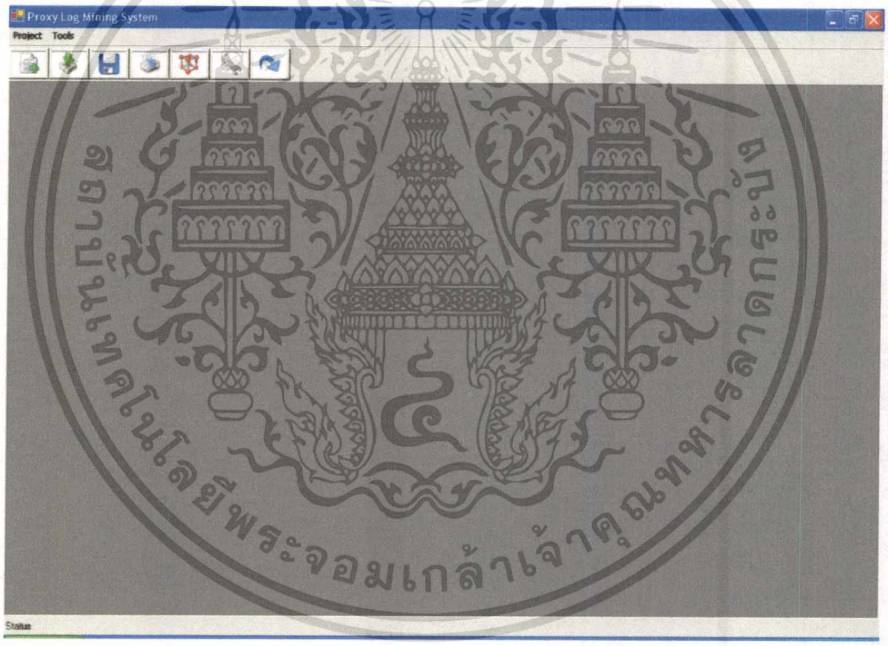
1) หน้าจอเมนเมนู (Main menu)

เป็นหน้าจอเริ่มต้นการทำงานของโปรแกรมมีเมนูดังรูปที่ 5.1

- Project
  - New เริ่มสร้างการไมนิ่งข้อมูล proxy log
  - Import log เป็นการเลือกข้อมูลดิบของ proxy log เข้ามาวิเคราะห์
  - Save บันทึกข้อมูลผลลัพธ์ที่ได้เป็นไฟล์ text

- Page setup ตั้งค่าการพิมพ์
- Print พิมพ์
- Exit ออกจากระบบ

- Tools
  - Path traversal pattern หารูปแบบการเดินทางการเรียกดูเว็บ
  - Proxy log mining วิเคราะห์หาความสัมพันธ์ของการเรียกดูเว็บ



รูปที่ 5.1 แสดงหน้าจอเมนเมนู

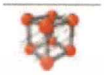
Toolbar ที่ใช้มีดังนี้

New

Import

Save

Print



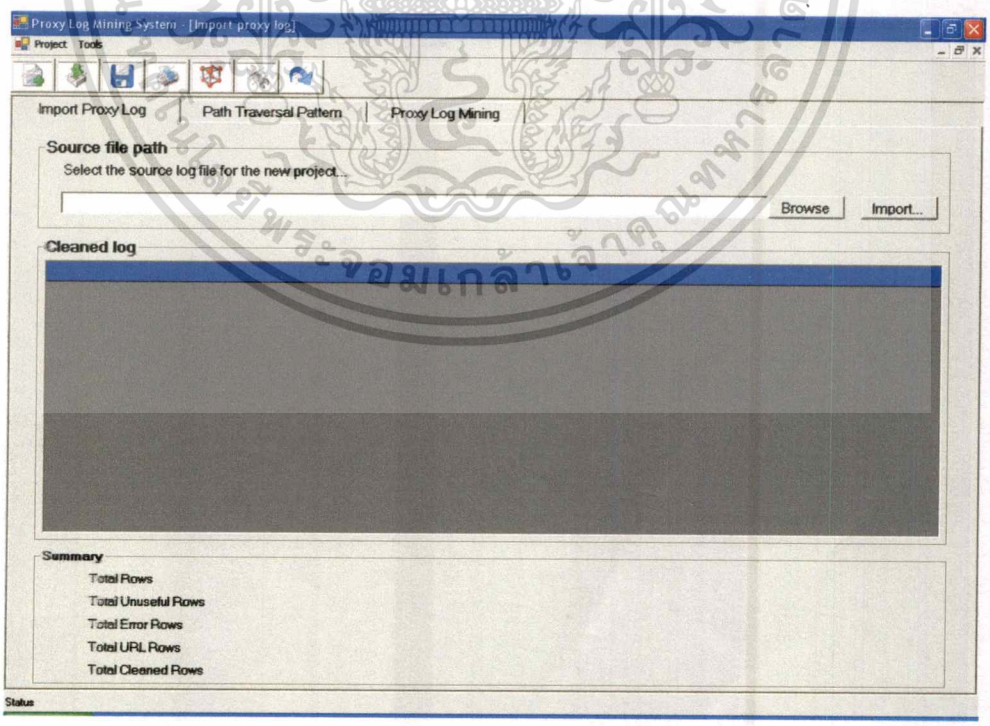
Path Traversal Pattern

Proxy Log Mining

Exit

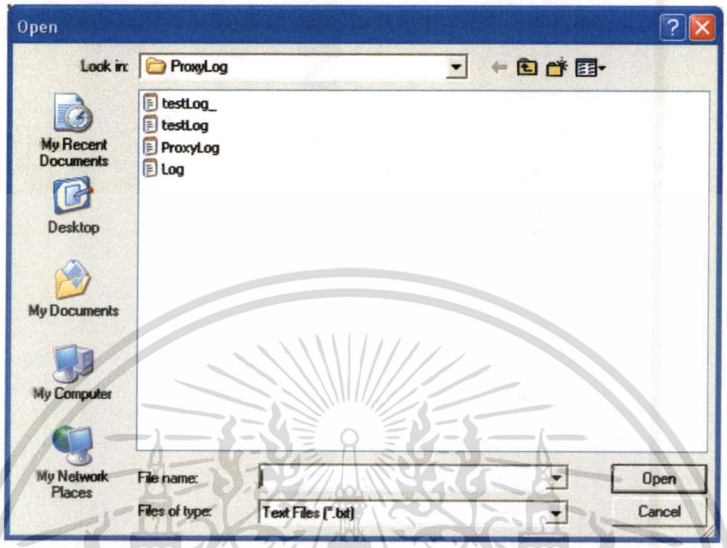
2) หน้าจอ New หรือ Import Log

เป็นหน้าจอในการนำเข้าข้อมูล proxy log มาบันทึกลงฐานข้อมูลเพื่อใช้ในการวิเคราะห์ แสดงดังรูปที่ 5.2



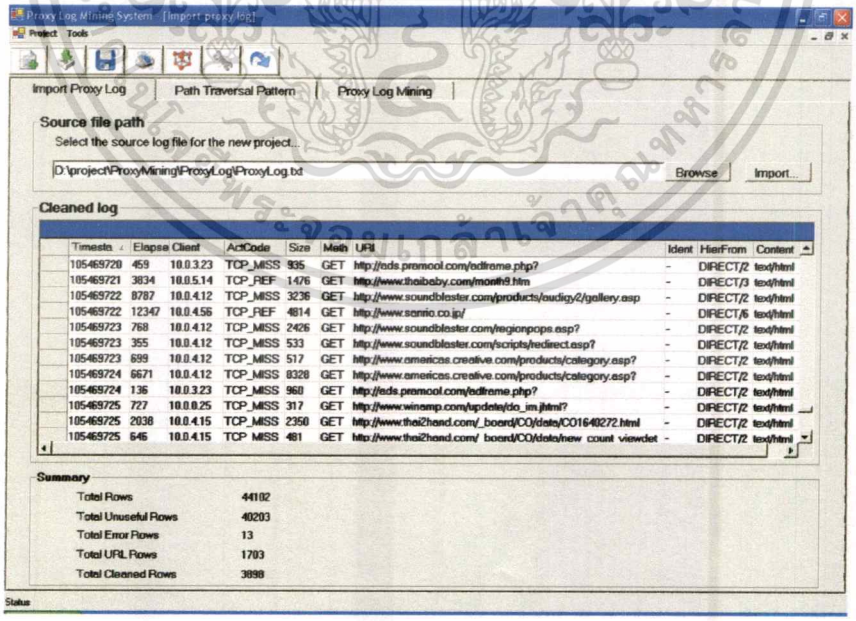
รูปที่ 5.2 แสดงหน้าจอของ New หรือ Import Log

กดปุ่ม Browse เพื่อทำการเลือกข้อมูลดิบ proxy จะแสดงได้ดังรูปที่ 5.3



รูปที่ 5.3 แสดงหน้าจอการเลือกไฟล์ Proxy log

จากนั้นกดปุ่ม Import... เพื่อเริ่มกดนำเข้าข้อมูลดิบ เมื่อเสร็จสิ้นแล้วจะได้ดังรูปที่ 5.4



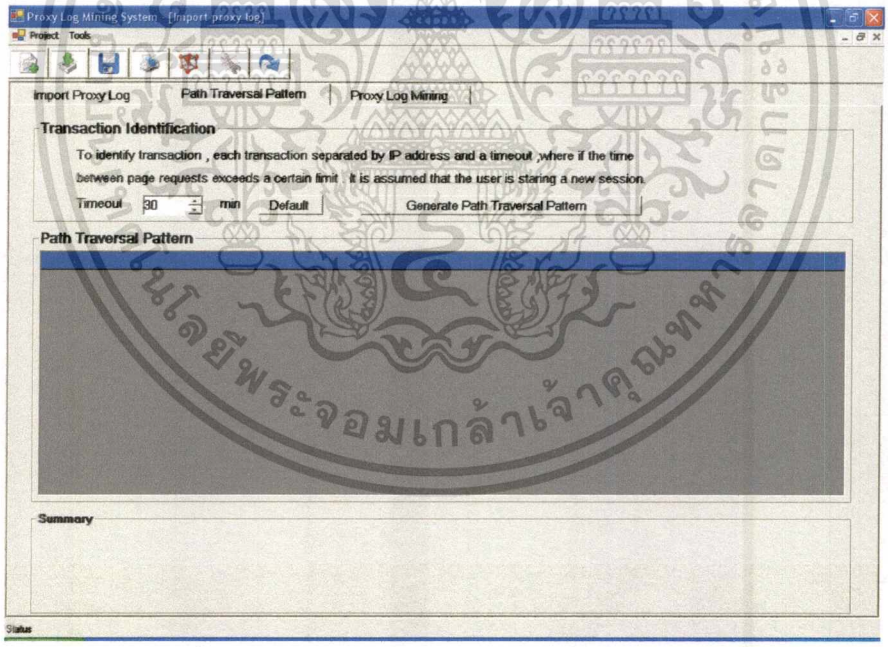
รูปที่ 5.4 แสดงหน้าจอผลลัพธ์การนำเข้าข้อมูลดิบที่ทำความสะอาดข้อมูลแล้ว

มีการแสดงผลการนำเข้าข้อมูลดังนี้

Total Rows	จำนวนบรรทัดข้อมูลดิบทั้งหมด
Total Unuseful Rows	จำนวนบรรทัดข้อมูลที่ไม่มีประโยชน์ในการวิเคราะห์จะตัดทิ้งไป
Total Error Rows	จำนวนบรรทัดของข้อมูลที่ไม่สามารถนำเข้าได้
Total URL Rows	จำนวนเว็บเพจที่ผู้ใช้เรียกดูของ proxy log นี้
Total Cleaned Rows	จำนวนบรรทัดข้อมูลที่ทำความสะอาดข้อมูลเรียบร้อย

3) หน้าจอ Path Traversal Pattern

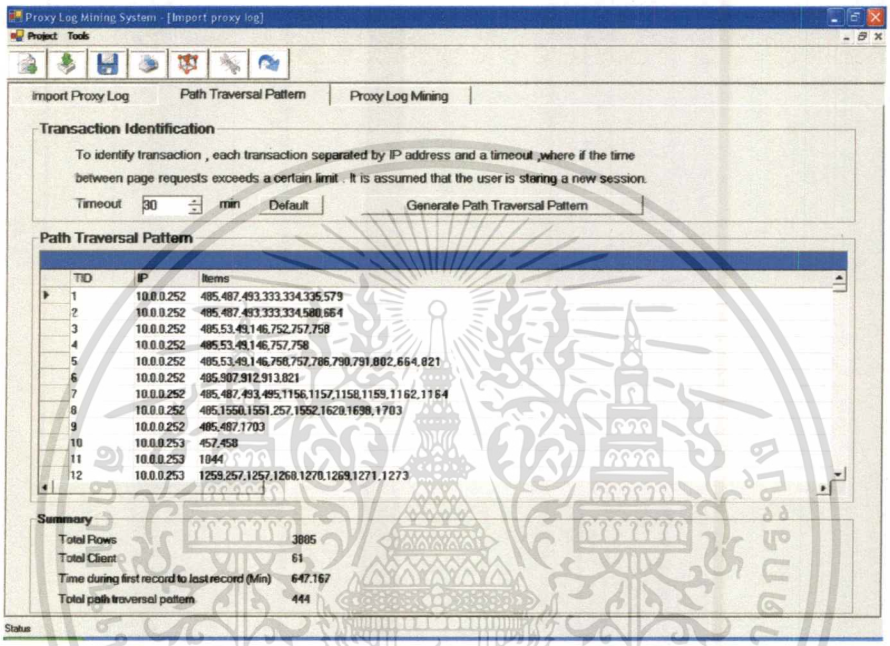
หน้าจอการหารูปแบบการเดินทางการเรียกดูเว็บของผู้ใช้ แสดงได้ดังรูปที่ 5.5



รูปที่ 5.5 แสดงหน้าจอการหารูปแบบการเดินทางการเรียกดูเว็บของผู้ใช้

ต้องกำหนดเวลา Timeout เป็นนาทีหรือกดปุ่ม Default เพื่อใช้เวลาดีฟอลต์ (30 นาที)

กดปุ่ม Generate Path Traversal Pattern เพื่อเริ่มต้นการหารูปแบบ จะได้ผลลัพธ์ดังรูปที่ 5.6



รูปที่ 5.6 แสดงหน้าจอการหารูปแบบการเดินทางที่ผู้ใช้เรียกดูเว็บเพจ

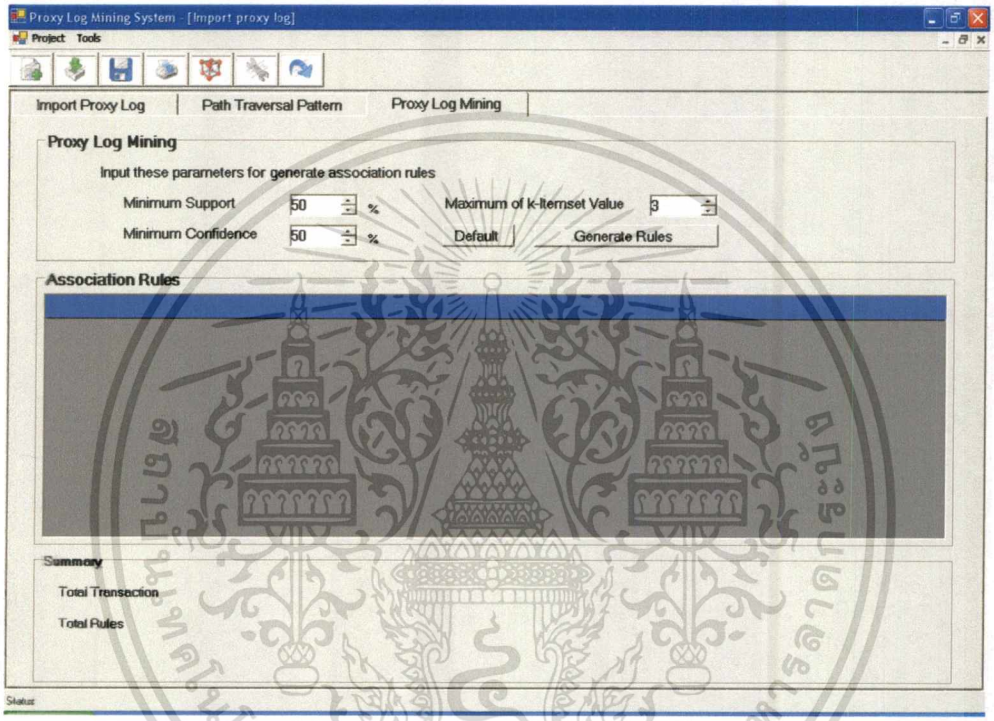
มีการแสดงผลการนำเข้าข้อมูลดังนี้

Total Rows จำนวนบรรทัดข้อมูลที่นำมาวิเคราะห์

Total Path traversal pattern จำนวนรูปแบบการเดินทางทั้งหมด

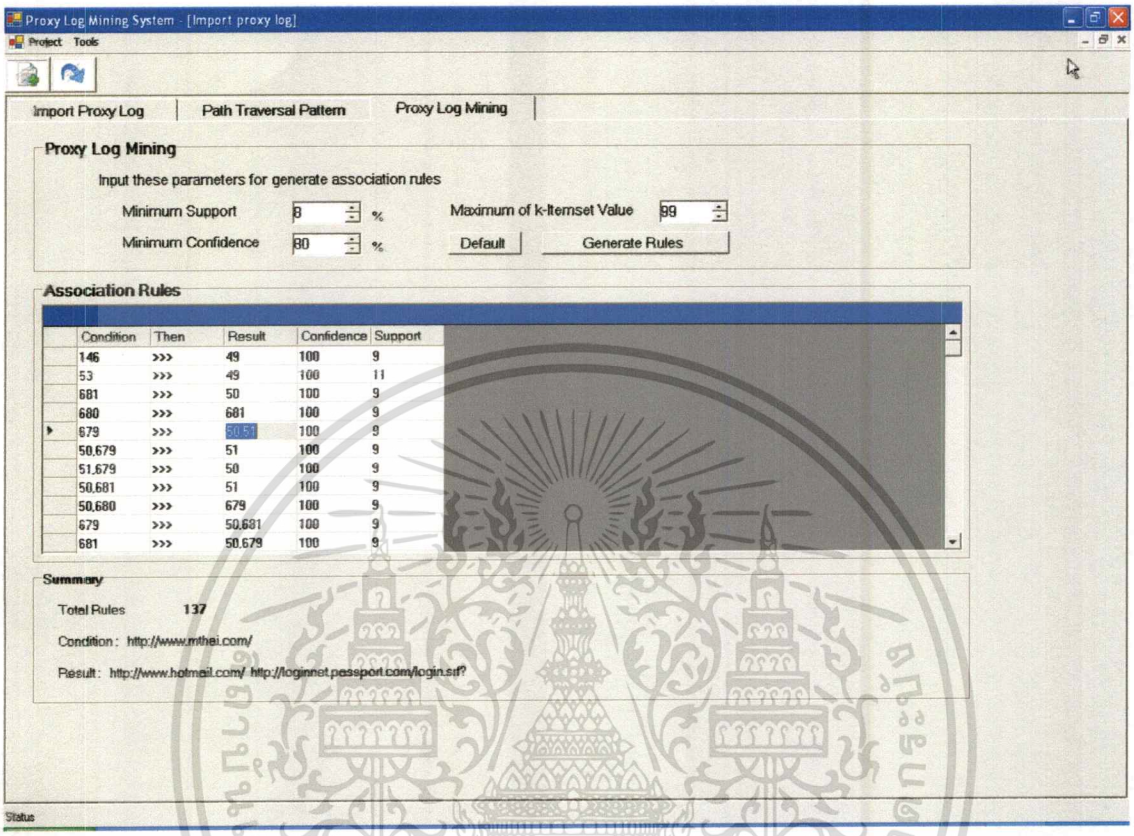
4) หน้าจอ Proxy Log Mining

เป็นหน้าจอในการวิเคราะห์หาความสัมพันธ์ของการเรียกดูเว็บเพจ แสดงได้ดังรูปที่ 5.7



รูปที่ 5.7 แสดงหน้าจอการวิเคราะห์หาความสัมพันธ์ของการเรียกดูเว็บเพจ

โดยจะมีพารามิเตอร์ให้กำหนด 3 ตัว คือ Minimum support , Minimum confidence และ Maximum of k-itemset Value หมายถึงจำนวน Item ที่ต้องการในการ generate กฎความสัมพันธ์ หรือกดปุ่ม Default (Minimum support = 50% , Minimum confidence = 50% และ Maximum of k-itemset Value = 3) จากนั้นกดปุ่ม Generate Rules ในที่นี้จะใส่ Minimum support = 8% ,Minimum Confidence = 80% และ k-itemset Value = 99 แล้วจึงเริ่มต้นหากฎความสัมพันธ์ ผลการ generate จะแสดงได้ดังรูปที่ 5.8



รูปที่ 5.8 แสดงผลการหาวิเคราะห์ความสัมพันธ์ของ proxy log

มีการแสดงผลการนำเข้าข้อมูลดังนี้

Total transaction	จำนวนบรรทัดข้อมูลที่นำมาวิเคราะห์
Total Rules	จำนวนกฎความสัมพันธ์ทั้งหมด
Rule	กฎความสัมพันธ์ของบรรทัดที่ mouse clicked

วิเคราะห์ผลการดำเนินงาน

ผลลัพธ์ที่ได้จากการทดสอบ มีรายการทรานแซ็กชั่นจำนวน 365 ทรานแซ็กชั่น และกำหนดค่า Minimum Support เท่ากับ 80% และค่า Minimum Confidence เท่ากับ 80% พบความสัมพันธ์การเข้าเว็บดังตัวอย่าง

ผู้ใช้ที่เข้า http://www.mthai.com แล้วจะเข้าเว็บ http://www.hotmail.com และ http://www.loginnet.passport.com/login.srf? พร้อมกันด้วยค่า Support 9% และ Confidence 100%

จากการทดลองดังกล่าวพบความสัมพันธ์ที่น่าสนใจมากมาย ซึ่งต้องอาศัยกระบวนการตัดสินใจว่าความสัมพันธ์ไหนจะเป็นประโยชน์ จากตัวอย่างนี้สามารถที่จะเพิ่มค่า Minimum Support หรือค่า Minimum Confidence เพื่อให้ได้กฎที่น้อยที่สุดได้ แล้วทำให้เราทราบพฤติกรรมของผู้ใช้เพื่อนำไปประยุกต์ใช้กับสารสนเทศภายในองค์กรต่อไปเช่นการปรับปรุงประสิทธิภาพของ cache ใน proxy server

บทที่ 6
บทสรุป

โครงการพัฒนาระบบการไมนิ่งข้อมูลของ Proxy log ด้วยเทคนิค Association rule นี้เป็นโครงการที่จัดทำเพื่อนำเสนอให้เห็นถึงประโยชน์ของการนำทฤษฎีดาต้าไมนิ่งมาใช้เพิ่มประสิทธิภาพในการวิเคราะห์หารูปแบบความสัมพันธ์ของข้อมูลการเรียกดูเว็บเพจของผู้ใช้ภายในองค์กร เพื่อนำผลที่ได้ไปวางแผนทางด้านสารสนเทศขององค์กรต่อไป

6.1. สรุปผลการดำเนินงาน

ดาต้าไมนิ่งเป็นกระบวนการที่ใช้เพื่อค้นหาข้อมูลที่มีประโยชน์ออกจากฐานข้อมูลเพื่อนำมาช่วยในการตัดสินใจ วิธีการแก้ปัญหาด้วยดาต้าไมนิ่งมีอยู่หลายรูปแบบขึ้นอยู่กับวัตถุประสงค์ของการทำงาน การที่จะนำเทคนิคของดาต้าไมนิ่งเข้ามาช่วยในการทำงานนั้น จำเป็นที่จะต้องเข้าใจลักษณะปัญหาที่แท้จริงก่อน เพื่อที่จะได้มีความสามารถในการกำหนดปัญหาและขอบเขตปัญหาได้ ถ้ากำหนดปัญหาได้ถูกต้องก็จะนำไปสู่ขั้นตอนการทำดาต้าไมนิ่งที่ถูกต้องและนำไปสู่ผลลัพธ์ที่ต้องการได้จากปัญหาขององค์กร

ระบบการไมนิ่งข้อมูลของ Proxy log ด้วยเทคนิค Association rule ที่พัฒนาขึ้นเป็นโปรแกรมที่พยายามนำเอาข้อมูลที่ทางองค์กรได้จัดเก็บไว้นำมาใช้ประโยชน์โดยการค้นหาความรู้ที่ได้จาก log ของการเข้าถึงเว็บเพจเหล่านี้ ซึ่งระบบแบ่งได้เป็น 3 ขั้นตอนหลัก คือขั้นตอนการเตรียมข้อมูลจะทำการแปลงข้อมูลและเก็บเฉพาะข้อมูลที่มีประโยชน์ในการวิเคราะห์ จากนั้นขั้นตอนที่สองจะทำการแปลงข้อมูลเหล่านั้นให้อยู่ในรูปแบบการเดินทางไปข้างหน้าไกลที่สุด (Maximal forward reference) ซึ่งจะใช้อัลกอริทึม MF ในการหารูปแบบ และขั้นตอนสุดท้ายจะเป็นการนำผลจากขั้นตอนที่ 2 มาหากฎความสัมพันธ์ของการเรียกดูเว็บของผู้ใช้ ซึ่งจะใช้อัลกอริทึม Apriori ของเทคนิค Association Rule เข้ามาวิเคราะห์ข้อมูลให้ได้ประสิทธิภาพในการทำงาน

ผลจากการพัฒนาระบบทำให้ได้เครื่องมือสำหรับการค้นหารูปแบบเส้นทางการเรียกดูเว็บเพจและกฎความสัมพันธ์การเรียกดูเว็บเพจของผู้ใช้ภายในองค์กร ซึ่งผลที่ได้นั้นผู้ใช้ต้องนำมาตีความหมายว่าจะนำไปประยุกต์ในการวางแผนทางด้านสารสนเทศขององค์กรต่อไปได้อย่างไรบ้าง เช่นการนำกฎที่ได้มาปรับปรุงระบบ cache ของ proxy server เป็นต้น

6.2. ข้อเสนอแนะ

ระบบการไมนิ่งข้อมูลของ Proxy log ด้วยเทคนิค Association rule นี้มีข้อจำกัดอยู่หลายประการที่ต้องปรับปรุง เพื่อให้ระบบมีความยืดหยุ่นเหมาะสมกับองค์กร สิ่งที่จะเสนอแนะมีดังนี้

- การระบุผู้ใช้ในระบบนี้แยกแยะโดยใช้ IP address เท่านั้น ซึ่งไม่เพียงพอที่จะนำมาแยกแยะผู้ใช้ เนื่องจากผู้ใช้หลายคนอาจจะใช้เครื่องเดียวกันในการเรียกดูเว็บเพจก็ได้ ซึ่งจะทำให้การระบุผู้ใช้ไม่ถูกต้อง ดังนั้นจึงต้องหาวิธีระบุผู้ใช้ที่แน่นอนให้ได้ เช่นการใส่รหัสผ่านก่อนการเรียกเว็บเพจ
- การนำเสนอรูปแบบเส้นทางการเรียกดูเว็บเพจของผู้ใช้ยังไม่มีความหลายหลาย ควรนำเสนอให้มีหลายรูปแบบเพื่อให้ผู้ใช้ทำความเข้าใจได้ง่ายขึ้น
- การค้นหาความสัมพันธ์ของการเรียกดูเว็บเพจใช้เวลาค่อนข้างมากในการวิเคราะห์ข้อมูลที่มีปริมาณมาก ซึ่งเป็นข้อเสียของอัลกอริทึม Apriori จึงควรเปลี่ยนแปลงอัลกอริทึมในการหาความสัมพันธ์เพื่อให้มีความเร็วในการค้นหามากขึ้น

บรรณานุกรม

Chen,M.S. et.al.1996. **"Data Mining for Path Traversal Pattern in a Web Environment."** 385-392.In *Proceedings of the 16th International Conference on Distributed Computing Systems.*

Cooley,R. et.al.1999. **"Data Preparation for Mining World Wide Web Browsing Patterns."**

Jan Kerkhofs , et.al., **"Web Usage Mining on Proxy Servers:A Case Study."**,Limburg University Centre

Rakesh Agrawal , Ramakrisnan Srikant , **"Fast Algorithms for Mining Association Rules."** ,487-499.In *proc.of the20th VLDB Conference,Santiago,Chile.*

Wenwu Lou, et.al.,**" Cut-and-Pick Transactions for Proxy Log Mining."** ,Hong Kong University

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวกุลธิดา บุญโฉม
สถานที่เกิด	จังหวัดกรุงเทพมหานคร
วุฒิการศึกษา	วิทยาศาสตรบัณฑิต คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ตำแหน่งหน้าที่	Engineer
สถานที่ทำงาน	บริษัทสามารถคอมเทค จำกัด