การทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร

โดย

นางสาวพิจิตรา จอมศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

ภาควิชาคอมพิวเตอร์

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร

ปีการศึกษา 2549

WEB CONTENT PREDICTION USING DATA MINING TECHNIQUE : A CASE STUDY OF SILPAKORN UNIVERSITY

By

Pijitra Jomsri

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

Department of Computing

Graduate School

SILPAKORN UNIVERSITY

2006

บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร อนุมัติให้วิทยานิพนธ์เรื่อง “ การทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร ” เสนอโดย นางสาวพิจิตรา จอมศรี เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

…….....

(รองศาสตราจารย์ ดร.ศิริชัย ชินะตังกูร)

คณบดีบัณฑิตวิทยาลัย

วันที่.....เดือน..... พ.ศ.....

ผู้ควบคุมวิทยานิพนธ์

ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์

คณะกรรมการตรวจสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์นันท์นภัส โตอดิเทพย์)

...../...../.....

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์)

...../...../.....

..... กรรมการ

(อาจารย์ ดร.กุศยา ปลั่งพงษ์พันธ์)

...../...../.....

46307310 : สาขาวิชาวิทยาการคอมพิวเตอร์
คำสำคัญ : พร๊อกซี / เหมืองข้อมูล / การค้นหากฎความสัมพันธ์

พิจิตรา จอมศรี : การทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร. อาจารย์ผู้ควบคุมวิทยานิพนธ์ : ผศ.ดร.ปานใจ ธารทัศนวงศ์. 106 หน้า.

งานวิจัยนี้ได้นำเทคนิคกฎการค้นหาความสัมพันธ์ ซึ่งเป็นเทคนิคหนึ่งในเทคนิคเหมืองข้อมูล มาประยุกต์ใช้สร้างตัวแบบเพื่อทำนายข้อมูลการเรียกใช้เว็บในอนาคต โดยผู้วิจัยได้จัดเก็บข้อมูล 2 ส่วนคือการเรียกใช้เว็บภายในมหาวิทยาลัยศิลปากรจากระบบพร๊อกซี เซิร์ฟเวอร์ และจัดทำฐานข้อมูลเว็บเพื่อจัดหมวดหมู่ของเว็บ แล้วนำข้อมูลทั้ง 2 ส่วนมาสร้างความสัมพันธ์ โดยงานวิจัยนี้นำข้อมูลวัน เวลา หมวดเว็บ และเว็บ มาค้นหาความสัมพันธ์เพื่อสร้างตัวแบบ โดยพิจารณาตัวแบบจากค่าความเชื่อมั่นและค่าสนับสนุน ผู้วิจัยได้ทำการทดลองศึกษาตัวแบบ โดยผลของการศึกษาตัวแบบพบว่าโมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ในวันถัดมาได้ ผลของการใช้เทคนิคเหมืองข้อมูลพบว่าตัวแบบที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ได้โดยมีความถูกต้องร้อยละ 66.67 % นอกจากนี้ผู้วิจัยยังนำเสนอระบบการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูลซึ่งสร้างอยู่บนขั้นตอนวิธีที่นำเสนอและอธิบายถึงผลการทดลองบนข้อมูลจริงอีกด้วย

งานวิจัยนี้สามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ในอนาคตได้ จึงสามารถเพิ่มประสิทธิภาพการทำงานของระบบพร๊อกซี เซิร์ฟเวอร์ได้ ซึ่งถ้าประสิทธิภาพการทำงานของระบบพร๊อกซี เซิร์ฟเวอร์เพิ่มขึ้น อาจทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและสามารถลดปริมาณข้อมูลในระบบเครือข่ายได้ อย่างไรก็ตามเทคนิคนี้ยังไม่สามารถครอบคลุมการทำงานในช่วงเหตุการณ์ที่ไม่เป็นปกติ เช่น อุบัติภัย และเทศกาลต่างๆ เป็นต้น

ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร ปีการศึกษา 2549

ลายมือชื่อนักศึกษา.....

ลายมือชื่ออาจารย์ผู้ควบคุมวิทยานิพนธ์ .....

46307310 : MAJOR : COMPUTER SCIENCE
KEY WORD : PROXY / DATA MINING / ASSOCIATION RULE DISCOVERY
PIJITRA JOMSRI : WEB CONTENT PREDICTION USING DATA MINING TECHNIQUE : A CASE STUDY OF SILPAKORN UNIVERSITY. THESIS ADVISOR : ASST. PROF. PANJAI TANTATSANAWONG,Ph.D. 106 pp.

This research using the association rule discovery is one technique of applied data mining technique to improve model for predicted web content. The two Data sets had been collected. First data set is Proxy servers of Silpakorn University. Second data set collected database web page for classify group web and they were created data relationship. This research select data such as: Date, Time, Group web and Web for created model .The model is considering by confident and support values. The results of test train model showed that model can predict web content for next day, also by using this technique, the future accesses of websites predict corrected 66.67%. We also present a web content prediction system which has been implemented based upon the proposed method and discuss our experimental results on real web data.

The results can predict web page access. It can increase efficient of Proxy servers by using this technique. If efficient of Proxy server increases, the performance of internet access will be improved and this can reduce traffic in networks. However, this technique could not be used for abnormal phenomenal such as protest, disaster or festival, etc.

Department of Computing Graduate School, Silpakorn University Academic Year 2006
Student's signature .....
Thesis Advisor's signature .....

กิตติกรรมประกาศ

ในการวิจัยครั้งนี้สำเร็จลุล่วงไปด้วยดีนั้น ผู้วิจัยต้องขอขอบพระคุณอาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.ปานใจ ธารทัศนวงศ์ ประธานกรรมการ ผู้ช่วยศาสตราจารย์นันท์นภัส โตอดิเทพย์ และกรรมการผู้ทรงคุณวุฒิ ดร.กุศยา ปลั่งพงษ์พันธ์ ที่กรุณาให้คำปรึกษา และตรวจสอบความถูกต้องของงานวิจัย ขอขอบคุณ คุณบุญมา เพ่งซ่วน และเพื่อนๆ ทุกคนที่ให้ความช่วยเหลือเป็นกำลังใจซึ่งกันและกัน และสุดท้ายนี้ต้องขอบพระคุณคุณพ่อและคุณแม่ที่สนับสนุนทุนการศึกษา คอยให้กำลังใจ และเป็นแรงผลักดันให้ผู้วิจัยได้ศึกษาต่อจนสำเร็จการศึกษา

สารบัญ

สารบัญตาราง

สารบัญภาพ

บทที่ 1
บทนำ

ความเป็นมาและความสำคัญของปัญหา

ในยุคปัจจุบันเครือข่ายอินเทอร์เน็ตถือเป็นเทคโนโลยีที่มีความสำคัญอย่างมากในการศึกษาค้นคว้าข้อมูล รวมทั้งการบริการต่างๆ เช่น เวิลด์ไวด์เว็บ (World Wide Web) หรือเรียกโดยย่อว่าเว็บ ซึ่งความก้าวหน้าทางเทคโนโลยีของเครือข่ายอินเทอร์เน็ต ทำให้ลดข้อจำกัดของเวลาและสถานที่ เช่น การศึกษาข้อมูลทางวิชาการ หรือการประกอบธุรกรรม และมีการใช้ในสถาบันการศึกษาที่มีการค้นคว้าวิจัยเป็นอย่างมาก

การเรียกใช้เว็บ นั้นในองค์กรส่วนใหญ่มีการนำระบบพร๊อกซี่ (Proxy) มาใช้เพื่อจัดเก็บข้อมูลรายละเอียดเว็บไซต์ที่เคยมีการเรียกใช้ไว้ที่ระบบพร๊อกซี่ หากมีผู้ใช้เรียกใช้เว็บเดิมนี้อีกก็จะทำการดึงข้อมูลเว็บที่อยู่ในพร๊อกซี่ มาให้กับผู้ใช้ซึ่งทำให้ผู้ใช้ได้รับข้อมูลเร็วกว่าการเรียกใช้งานตรงจากเว็บต้นฉบับ (Original Web) และยังเป็นการลดความหนาแน่นของระบบเครือข่ายได้เป็นอย่างดี แต่เนื่องจากพร๊อกซี่มีการทำงานแบบ Cache คือมีปริมาณของเว็บที่เก็บในหน่วยความจำของเครื่อง แม่ข่ายพร๊อกซี่จำนวนจำกัดไม่สามารถจะจัดเก็บเว็บทั้งหมดซึ่งมีจำนวนมากได้ จึงจำเป็นต้องมีการลบออก (page out) ซึ่งบางครั้งเว็บที่ลบออกไปมีการเรียกใช้อีก ทำให้ต้องเสียเวลาในการไปดึงจากเว็บต้นฉบับมาใหม่ (page in) ทำให้ประสิทธิภาพการทำงานของพร๊อกซี่ลดลง ซึ่งสามารถวัดได้จากอัตราการพบในพร๊อกซี่(Hit Rate)

ด้วยปัญหาดังกล่าวผู้วิจัยจึงได้จัดทำโครงการวิจัยขึ้นเพื่อเพิ่มอัตราการพบ โดยใช้เทคนิคเหมืองข้อมูลเพื่อทำนายเว็บเพจที่จะมีการเรียกใช้ในอนาคต

1.1 วัตถุประสงค์การวิจัย

1.1.1 เพื่อศึกษาการทำงานของเหมืองข้อมูล

1.1.2 พัฒนาโมเดลเพื่อใช้ในการทำนายแนวโน้มการใช้งานเว็บในอนาคต โดยใช้เทคนิคการ ทำเหมืองข้อมูล

1.1.3 ทดสอบความถูกต้องโมเดลที่พัฒนาขึ้น

1.1.4 เพื่อนำโมเดลที่พัฒนาขึ้นมาประยุกต์ใช้

1.2 ประโยชน์ที่คาดว่าจะได้รับ

1.2.1 เพื่อทราบโมเดลในการทำนายแนวโน้มการใช้งานเว็บในอนาคตซึ่งใช้เทคนิคการทำเหมืองข้อมูล

1.2.2 สามารถนำอัลกอริทึมที่ได้ศึกษามาไปประยุกต์ใช้ในการปรับปรุงการทำงานของระบบพร็อกซี เซิร์ฟเวอร์ ได้อย่างมีประสิทธิภาพ รวมทั้งทำการดึงข้อมูลมาเก็บไว้ล่วงหน้าก่อนที่ผู้ใช้จะเรียกใช้ (Prefetch) เพื่อช่วยให้การใช้งานเว็บรวดเร็วขึ้น

1.3 ขอบเขตการวิจัย

การพัฒนาโมเดลเพื่อใช้ในการทำนายแนวโน้มการใช้งานเว็บในอนาคตโดยใช้เทคนิคการทำเหมืองข้อมูล มีขอบเขตการวิจัยภายในมหาวิทยาลัยศิลปากร โดยเก็บข้อมูลการใช้เว็บภายในมหาวิทยาลัย เป็นเวลา 3 เดือน ตั้งแต่วันที่ 1 กันยายน 2548 ถึง วัน 30 พฤศจิกายน 2548

1.4 ขั้นตอนการดำเนินการวิจัย

ในงานวิจัยนี้สามารถแบ่งขั้นตอนในการดำเนินการวิจัยออกเป็น 7 ขั้นตอน ดังนี้

1. เก็บรวบรวมข้อมูลจากแหล่งข้อมูลที่เกี่ยวข้อง
2. วิเคราะห์และเลือกใช้ทฤษฎีและอัลกอริทึมที่เหมาะสม
3. สร้างโมเดล
4. ทำการทดลองทำนายเนื้อหาของเว็บ
5. วิเคราะห์ผลการทดลองโดยการเปรียบเทียบประสิทธิภาพ
6. สรุปผลการทดลอง
7. รวบรวมข้อเสนอแนะ

1.5 เครื่องมือที่ใช้ในการวิจัย

1.5.1 ฮาร์ดแวร์

- Intel Pentium M 1.73 GHz
- RAM 512 MB
- Hard disk 80 GB

1.5.2 ซอฟต์แวร์

- ระบบปฏิบัติการ : Window XP Professional
- ฐานข้อมูล : SAS 9.1

บทที่ 2
ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

การศึกษาทฤษฎีและผลงานวิจัยที่เกี่ยวข้องในบทนี้ผู้วิจัยได้ศึกษาจากเอกสาร แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง ประกอบด้วย

1) การทำเหมืองข้อมูล (Data mining)
2) ระบบพร๊อกซี เซิร์ฟเวอร์ (Proxy server)
3) วิวัฒนาการของเหมืองข้อมูล
4) รูปแบบในการทำเหมืองข้อมูล
5) ขั้นตอนการทำเหมืองข้อมูล
6) ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

1. การทำเหมืองข้อมูล(data mining)

มีผู้ให้คำจำกัดความของการทำเหมืองข้อมูลไว้หลายคำจำกัดความ ดังนี้

การทำเหมืองข้อมูล หมายถึง การสำรวจและวิเคราะห์ข้อมูลที่มีขนาดใหญ่เพื่อค้นหารูปแบบหรือกฎที่ซ่อนอยู่ในข้อมูลนั้น และนำความรู้ที่ค้นพบไปใช้ประโยชน์ในการพัฒนาองค์กร (Berry and Linnoff 2004 : 7)

การทำเหมืองข้อมูล หมายถึง กระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูล (บุญเสริม กิจศิริกุล 2545 : 2)

การทำเหมืองข้อมูล หมายถึง กระบวนการที่นำข้อมูลที่มีอยู่ในฐานข้อมูลขนาดใหญ่มาทำการศึกษา วิเคราะห์ ทำความเข้าใจ และนำผลลัพท์ที่ได้จากการศึกษามาใช้ในการตัดสินใจทางธุรกิจ (Connolly and E.Begg 2002 : 1115)

การทำเหมืองข้อมูล หมายถึง ขบวนการทำงาน(process) ที่สกัดข้อมูล (Extract data) จากฐานข้อมูลขนาดใหญ่ (Large Information) เพื่อให้ได้สารสนเทศ (Usefull Information) ที่เรายังไม่รู้ (Unknown data) โดยเป็นสารสนเทศที่มีเหตุผล (Valid) และสามารถนำไปใช้ได้ (Actionable) ซึ่ง

เป็นสิ่งสำคัญในการที่จะช่วยการตัดสินใจในการทำธุรกิจ (Data Mining & Data Exploration Lab คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง:2005)

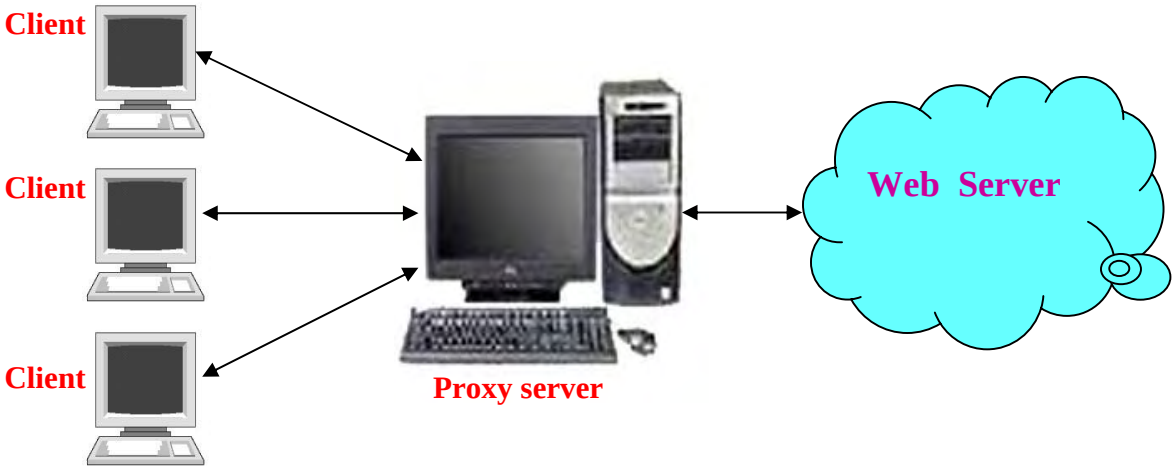
2. ระบบพร๊อกซี เซิร์ฟเวอร์ (proxy server)

พร๊อกซี เซิร์ฟเวอร์ หรือเรียกว่า แคช (Cache) คือการนำเครื่องคอมพิวเตอร์ที่ให้บริการแก่กลุ่มผู้ใช้ที่อยู่ในบริเวณเดียวกัน และกำหนดให้ผู้ใช้ทุกคนเรียกใช้ข้อมูลเว็บ ผ่านเครื่องคอมพิวเตอร์นี้ โดยเครื่องดังกล่าวจะมีการติดตั้งโปรแกรมเพื่อทำหน้าที่เรียกข้อมูลเว็บมาให้บริการแก่ผู้ใช้ และจัดเก็บข้อมูลที่เคยถูกเรียกนั้นไว้ ในเครื่อง เพื่อให้บริการแก่ผู้ใช้ข้อมูลนั้นซ้ำได้ทันทีโดยไม่ต้องเสียเวลาไปเรียกข้อมูลมาจากแหล่งข้อมูลใหม่ ซึ่งวิธีนี้ทำให้ผู้ใช้สามารถเรียกใช้ข้อมูลที่เคยมีผู้เรียกใช้มาก่อนได้รวดเร็วขึ้น เนื่องจากไม่ต้องเสียเวลาไปเรียกข้อมูลจากแหล่งข้อมูลใหม่ ทำให้ประสิทธิภาพในการใช้งานระบบเครือข่ายอินเทอร์เน็ต เพิ่มขึ้น (มหาวิทยาลัยธรรมศาสตร์:2005) ดังรูปที่ 1

หลักการทำงานของพร๊อกซี เซิร์ฟเวอร์

เมื่อมีผู้ใช้บริการทำการเรียกข้อมูลของเว็บไซต์ (Web Site) โดยผ่านพร๊อกซี เซิร์ฟเวอร์ในครั้งแรก พร๊อกซี เซิร์ฟเวอร์จะทำการตรวจสอบว่า มีข้อมูลของเว็บไซต์นั้นมีอยู่หรือไม่ หากพบว่าไม่มีข้อมูลพร๊อกซี เซิร์ฟเวอร์จะทำการเรียกข้อมูลนั้นจากเว็บไซต์แล้วเก็บไว้ในเครื่อง และเมื่อมีผู้ใช้บริการทำการเรียกเว็บไซต์นี้อีกครั้งพร๊อกซี เซิร์ฟเวอร์จะทำการส่งข้อมูลไปยังเครื่อง ของผู้ใช้บริการทันที ในกรณีที่ เว็บไซต์มีการปรับปรุงข้อมูลพร๊อกซี เซิร์ฟเวอร์ จะทำการตรวจสอบข้อมูลที่มีอยู่ว่า ปรับปรุงหรือไม่ และจะทำการปรับปรุงข้อมูลใหม่ทันที ในกรณีที่มีผู้เรียกใช้บริการก็จะได้ข้อมูลที่ปรับปรุงอยู่เสมอ (ศูนย์คอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี:2005)

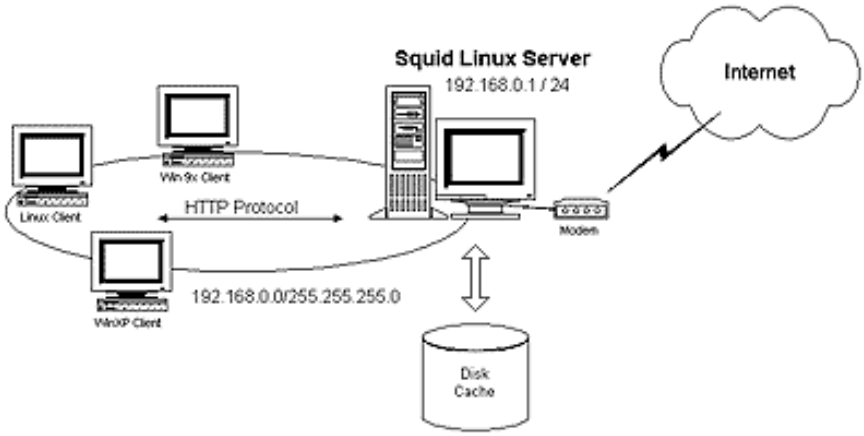
ซึ่งกรณีที่ข้อมูลบนเว็บไซต์มีการปรับปรุงอยู่ตลอดเวลานั้น พร๊อกซี เซิร์ฟเวอร์ จะต้องไปปรับปรุงข้อมูลจากเว็บเซิร์ฟเวอร์(Web Server) เนื่องจากไม่ได้เป็นข้อมูลที่ได้มาจากเว็บเซิร์ฟเวอร์โดยตรงจึงอาจทำให้ผู้ใช้บริการได้รับข้อมูลที่ไม่เป็นปัจจุบัน หรือ เสียเวลาในการรอรับข้อมูลจากพร๊อกซี เซิร์ฟเวอร์ ซึ่งต้องไปปรับปรุงข้อมูลจากเว็บเซิร์ฟเวอร์ ก่อนถึงจะนำข้อมูลมาให้บริการแก่ผู้ใช้ได้



รูปที่ 1 การทำงานของ Proxy server

ระบบพร็อกซี เซิร์ฟเวอร์ที่นิยมใช้มากที่สุด และมีความสามารถสูง คือ squid (www.squid-cache.org) ซึ่งจะมีมาพร้อมกับลินุกซ์เซิร์ฟเวอร์ทุกตัว โปรแกรม Squid เป็นพร็อกซีเซิร์ฟเวอร์ ที่มีคุณสมบัติในการจำกัด ควบคุมการแอกเซสเข้าสู่เว็บไซต์ภายนอกองค์กรได้เป็นอย่างดี และมีประสิทธิภาพ ที่เรียกว่า Access Control List (ACL) ซึ่งเป็นการนิยามชื่อลิสต์ขึ้นแทนคุณสมบัติของสิ่งที่ต้องการอ้างอิง จากนั้นจึงตั้งข้อกำหนดลงไปว่าต้องการให้ลิสต์นั้นสามารถแอกเซสผ่านพร็อกซี ได้หรือไม่ ดังนั้นการที่เสริมการทำงานของอินเตอร์เน็ตเซิร์ฟเวอร์ด้วย Squid Proxy Server จึงเป็นการควบคุมการเข้าสู่อินเตอร์เน็ตของผู้ใช้งานในองค์กรได้ตามต้องการ และยังช่วยเพิ่มประสิทธิภาพให้แก่ระบบอีกด้วยเพราะ Squid จะมีคุณสมบัติเป็น HTTP Object cache ที่ช่วยเก็บข้อมูลจากเว็บไซต์ภายนอกไว้ในหน่วยความจำ (RAM และฮาร์ดดิสก์) ของตัวเซิร์ฟเวอร์เองอีกด้วย ช่วยให้การเรียกเว็บไซต์ที่เคยเข้าถึงมาก่อนทำได้รวดเร็วยิ่งขึ้น เนื่องจากมีข้อมูลบางส่วนของเว็บเพจที่ ยังคงอยู่ในแคช

สำหรับระบบเครือข่ายที่จะใช้งาน squid อาจจะใช้การแชร์อินเทอร์เน็ตผ่านโมเด็มที่ติดตั้งไว้ที่เครื่องคอมพิวเตอร์ที่รันระบบปฏิบัติการลีนุกซ์ และ Squid ดังรูปที่ 2



รูปที่ 2 ระบบเครือข่ายที่ใช้งาน Squid

ที่มา : ธีรภัทร มนตรีศาสตร์, Squid Proxy Caching Server [Online], Accessed 5 November 2005. Available from http://micro.se-ed.com/content/mc205/MC205_181.asp

Access Log ของ Squid

เมื่อเครื่องลูกข่ายเรียกใช้งานเว็บไซต์ภายนอก จะทำการติดต่อผ่านพอร์ต 8080 มายังเครื่องเซิร์ฟเวอร์ลีนุกซ์ที่รัน Squid ไว้ เราสามารถมอนิเตอร์ดูความเคลื่อนไหวของการเรียกใช้งานดังกล่าวได้จาก LOG ไฟล์ของ squid จะเห็นข้อความแสดงเหตุการณ์ต่าง ๆ ที่เกิดขึ้นเกี่ยวกับ Squid ได้ ดังรูป ซึ่งเราสามารถตีความหมายได้ เช่น

TCP_MISS หมายถึง การร้องขอนั้นไม่มีการแคชข้อมูลไว้ กรณีนี้จะต้องไป GET มาจากเว็ปไซต์ปลายทางมาจริง ซึ่งจะต้องสิ้นเปลืองแบนด์วิธของโมเด็ม

TCP_HIT หรือ TCP_MEM_HIT หมายถึง มีข้อมูลที่เครื่องลูกข่ายร้องขอมาอยู่แล้วในแคช กรณีนี้ไม่มีการใช้แบนด์วิธ

```
1025162259.349 6 192.168.0.254 TCP_MEM_HIT/200 440 GET
http://www.thairath.co.th/picuse/rmore_it.gif - NONE/- image/gif
1025162259.454 11 192.168.0.254 TCP_MEM_HIT/200 653 GET
http://www.thairath.co.th/picuse/foot_it.gif - NONE/- image/gif
1025162259.758 3 192.168.0.254 TCP_HIT/200 9413 GET
http://www.thairath.co.th/link/schedule.jpg - NONE/- image/jpeg
1025162259.970 17 192.168.0.254 TCP_HIT/200 8550 GET
http://www.thairath.co.th/link/IT.banner.gif - NONE/- image/gif
1025162263.053 3321 192.168.0.254 TCP_MISS/200 400 GET
http://adserve.inet.co.th/jnserver/SITE=thairath.co.th/AREA=thairath.index/AAMSZ=195x
39 - DIRECT/203.150.14.102 application/x-javascript
1025162263.165 2992 192.168.0.254 TCP_MISS/200 430 GET
http://www.thairath.co.th/pichead/dotrd.gif - DIRECT/203.151.217.25 image/gif
```

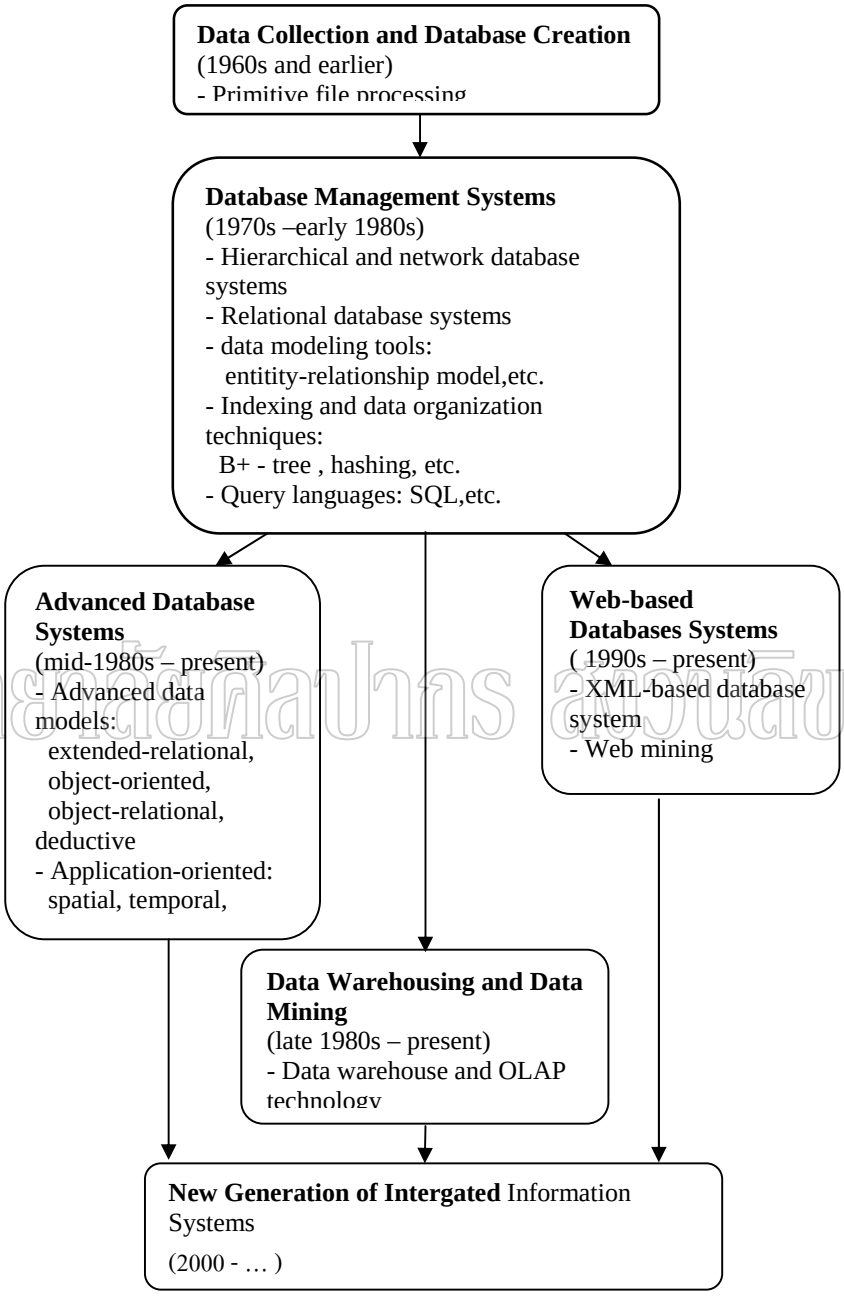
รูปที่ 3 ตัวอย่างไฟล์ access.log

ที่มา :ธีรภัทร มนตรีศาสตร์, Squid Proxy Caching Server [Online], Accessed 5 November 2005. Available from http://micro.se-ed.com/content/mc205/MC205_181.asp

ไฟล์ access.log ของ squid นี้จะมีขนาดเพิ่มขึ้นเรื่อย ๆ จนอาจทำให้เนื้อที่ดิสก์เต็ม และสร้างปัญหาให้กับระบบได้โดยอาจจะเขียนเป็นเชลล์สคริปต์เพื่อทำให้ Log ไฟล์เกิดการหมุนเวียนกันไป ในแต่ละสัปดาห์ก็ได้

ปัจจุบันมีการวิจัยเพื่อพัฒนาการทำงานของระบบพร็อกซี เซิร์ฟเวอร์มากมาย ซึ่ง พรีเฟช (Prefetching) หรือการกำหนดเงื่อนไขการจัดเตรียมข้อมูลล่วงหน้าของโปรแกรม เพื่อให้สามารถแสดงหน้าเว็บเพจของเว็บไซต์ที่เราไปเรียกใช้รวดเร็วขึ้นนั้น ถือเป็นการพัฒนาระบบ พร็อกซีเซิร์ฟเวอร์วิธีหนึ่ง (อุดมทรัพย์ กรรดิพณิชกุล: 2000)

3. วิวัฒนาการของเหมืองข้อมูล



รูปที่ 4 วิวัฒนาการของเหมืองข้อมูล

ที่มา : Han, Jiawei and Micheline Kamber, Data Mining Concepts and Techniques. (USA : Morgan Kaufman,2001), 2.

จาก รูปที่ 4 แสดงให้เห็นถึงวิวัฒนาการของการจัดเก็บและวิเคราะห์ข้อมูล ตั้งแต่ ปี ค.ศ. 1960 มีการนำข้อมูลมาจัดเก็บอย่างเหมาะสมในอุปกรณ์ที่น่าเชื่อถือ และป้องกันการสูญหาย(Data Collection and Database Creation) ต่อมามีการพัฒนาระบบฐานข้อมูลให้มีความสามารถในการจัดการข้อมูลด้านต่าง ๆ เช่น การกำหนดความสัมพันธ์ระหว่างฟิลด์ต่าง ๆ ในเรคอร์ด , การจัดการประมวลผลปรับเปลี่ยน แก้ไขข้อมูล และจัดการกำหนดควบคุมการใช้ข้อมูลที่มีอยู่ได้อย่างเป็นระบบ(Database Management Systems) ในปี ค.ศ. 1970 และในปีช่วงกลางของ ค.ศ. 1980 มีการเพิ่มประสิทธิภาพของระบบฐานข้อมูลโดยการนำข้อมูลที่จัดเก็บมาสร้างความสัมพันธ์เชิงวัตถุเพื่อใช้ในการวิเคราะห์และตัดสินใจที่มีคุณภาพ (Advanced Database Systems) ในช่วงหลังปี ค.ศ. 1980 จนถึงปัจจุบัน มีการรวมรวมข้อมูลมาจัดเก็บลงในฐานข้อมูลขนาดใหญ่ เพื่อช่วยสนับสนุนการตัดสินใจ และนำข้อมูลจากฐานข้อมูลมาวิเคราะห์และประมวลผลโดยการสร้างแบบจำลองและความสัมพันธ์ทางสถิติ ปี ค.ศ. 1990 ได้เริ่มมีการจัดทำระบบฐานข้อมูลบนเว็บ (Web-base Databases Systems)

4. รูปแบบการทำเหมืองข้อมูล

เราสามารถแบ่งรูปแบบการสร้างแบบจำลองในการทำเหมืองข้อมูล ออกเป็น 2 ประเภท คือ

4.1 การสร้างแบบจำลองเพื่อการทำนาย (Predictive Modelingหรือ Supervised Learning)

คือ การนำข้อมูลที่มีอยู่มาใช้ในการทำนายผลข้อมูลในอนาคต ซึ่งการสร้างแบบจำลองรูปแบบนี้จะเน้นการแบ่งข้อมูลออกเป็นกลุ่มตามคุณสมบัติของข้อมูล ในกรณีที่ข้อมูลไม่ต่อเนื่อง จะใช้เทคนิค การจำแนกประเภทข้อมูล (Classification) และในกรณีที่ข้อมูลมีความต่อเนื่องจะใช้เทคนิค การถดถอย (regression)

4.2 การสร้างแบบจำลองในการบรรยาย (Descriptive Modelingหรือ Unsupervised Learning)

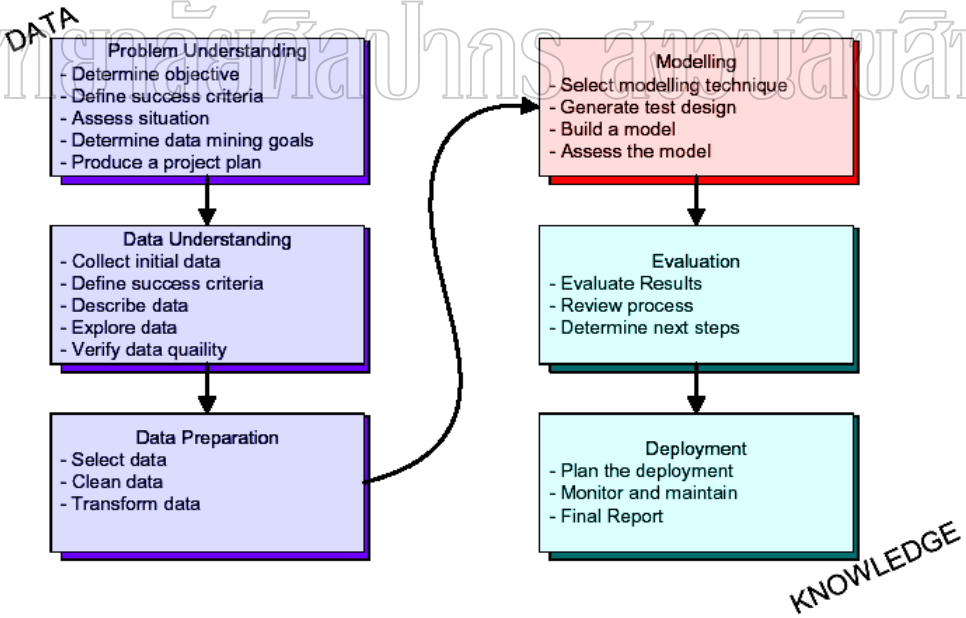
คือ การนำข้อมูลที่มีอยู่มาศึกษา เป็นการเรียนรู้จากข้อมูลที่มีอยู่และอธิบายให้เห็นภาพชัดเจน ซึ่งอาจใช้เทคนิคการหาความสัมพันธ์ (Association) หรือ เทคนิคการจัดกลุ่ม (Clustering)

5. ขั้นตอนการทำเหมืองข้อมูล

ในการทำเหมืองข้อมูลนั้นต้องใช้ความรู้จากศาสตร์หลายแขนง ซึ่งประกอบด้วยความรู้ทางด้านต่างๆ ต่อไปนี้

- ฐานข้อมูล (Database technology) เป็นแหล่งในการจัดเก็บ และรวบรวมข้อมูล
- สถิติ (Statistics) ใช้สำหรับวิเคราะห์ข้อมูลทางสถิติเบื้องต้น
- การเรียนรู้ของเครื่อง (Machine learning) เป็นอัลกอริทึมที่ใช้ในการค้นหารูปแบบและความสัมพันธ์ของข้อมูล
- การมองเห็น (Visualization) เป็นการแสดงผลลัพธ์ และ ความสัมพันธ์เพื่อให้ผู้ใช้เข้าใจได้ง่าย
- ความรู้ด้านวิทยาศาสตร์ (Information science) เป็นความรู้ทางด้านวิทยาศาตร์อื่นๆ ที่เกี่ยวข้อง

ขั้นตอนการทำเหมืองข้อมูลประกอบด้วยขั้นตอนหลัก 6 ขั้นตอน ดังนี้



รูปที่ 5 ขั้นตอนการทำเหมืองข้อมูล

ที่มา : บุญเสริม กิจศิริกุล, “รายงานวิจัยฉบับสมบูรณ์,” โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2546.

การทำความเข้าใจกับปัญหา (Problem Understanding)

ขั้นตอนการทำความเข้าใจปัญหาประกอบด้วย การตั้งวัตถุประสงค์ของการทำเหมืองข้อมูล (Determine objective) ตั้งเกณฑ์วัดความสำเร็จ (Define success criteria) ประเมินสถานการณ์ในด้านต่างๆ (Assess situation) ระบุเป้าหมายที่ใช้ในการตัดสินใจทำเหมืองข้อมูล (Determine data mining goals) วางแผนการทำเหมืองข้อมูล (Produce a project plan) ว่าจะเก็บข้อมูลด้วยวิธีใด และใช้อัลกอริทึมไหน

5.1 การทำความเข้าใจข้อมูล (Data Understanding)

ขั้นตอนการทำความเข้าใจข้อมูล ประกอบด้วยรวบรวมข้อมูล (Collect initial data) กำหนดคุณสมบัติของข้อมูล (Define success criteria) อธิบายรายละเอียดของข้อมูล (Describe data) การสำรวจข้อมูล (Explore data) การตรวจสอบความถูกต้องและความสมบูรณ์ของข้อมูล (Verify data quality)

5.2 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการเตรียมข้อมูล ประกอบด้วย การคัดเลือกข้อมูลที่จะนำมาใช้ (Select data),

- การทำความสะอาดข้อมูล (Clean data) ซึ่งเป็นกระบวนการเตรียมข้อมูลให้เหมาะสมที่สุดเพื่อนำไปใช้ในขั้นตอนต่อไป เช่น
- แก้ไขข้อมูลให้ถูกต้องสมบูรณ์ เช่น การแก้ไขค่าว่างของข้อมูลโดยใส่ค่า 9 (เก้า)
- ปรับเปลี่ยนข้อมูลให้มีค่าที่เหมาะสมในการตัดสินใจ เช่น ข้อมูลที่มีค่า “มาม่า” และ “ไวไว” อาจเปลี่ยนค่าเป็น “บะหมี่กึ่งสำเร็จรูป”
- เลือกข้อมูลเฉพาะที่สนใจ เช่น ต้องการหาลักษณะลูกค้าที่ซื้อรถเก๋ง ไม่ควรนำรายชื่อพนักงานขายเข้ามาเกี่ยวข้อง
- คอลัมน์ที่มีค่าสำหรับทุกแถวเป็นค่าเดียว เช่น “สัญชาติไทย” หรือ คอลัมน์ที่มีค่าไม่ซ้ำกันเลย เช่น “หมายเลขสมาชิก” ไม่ควรนำมาใช้ เนื่องจากไม่สามารถบอกรูปแบบของข้อมูลได้

การปรับเปลี่ยนรูปแบบข้อมูล(Transform data) เช่น นำตารางในฐานข้อมูลมาเชื่อมต่อกัน ขั้นตอนนี้เป็นขั้นตอนที่สำคัญมากเนื่องจากความถูกต้อง และสมบูรณ์ของผลลัพธ์สุดท้ายซึ่งขึ้นอยู่กับว่านักวิเคราะห์ข้อมูลนั้นตัดสินใจกำหนดโครงสร้างและเสนอลักษณะของข้อมูลที่จะใช้ในการประมวลผลอย่างไรอย่างไร กรรมวิธีนี้ รวมไปถึงการทำ Data Recording (การจัดเก็บข้อมูล) และ Data

Format Conversion(รูปแบบในการแปลงข้อมูล) เช่นการแปลงเวลา unix timestamp เป็นเวลาปัจจุบัน เป็นต้น

5.3 การสร้างแบบจำลอง (Modeling)

ขั้นตอนการสร้างแบบจำลอง ประกอบด้วย การเลือกเทคนิคที่เหมาะสมในการทำเหมืองข้อมูล (Select modeling technique) กำหนดรูปแบบการทดสอบผลลัพธ์(Generate test design) สร้างแบบจำลองตามเทคนิคที่เลือก (Model Building) ทดสอบความถูกต้องและความน่าเชื่อถือของแบบจำลองที่สร้างขึ้น(Model Assessnt)

5.4 การประเมินผล (Evaluation)

ขั้นตอนการประเมิน ประกอบด้วย การประเมินผลที่ได้จากการทดลอง(evaluate Results) อาจจะเป็นการประเมินแบบจำลองที่สร้างขึ้นด้วยการลองนำไปใช้กับสถานการณ์จริงเพื่อตรวจสอบประสิทธิภาพของแบบจำลอง การทบทวนกระบวนการ (review process) ใช้เป็นขั้นตอนถัดไปในการตัดสินใจ (Determine next steps)

5.5 การนำไปใช้ (Deployment)

ขั้นตอนการนำไปใช้ ประกอบด้วย แผนการในการนำไปใช้ (Plan the deployment) และ สรุปผลของการทดลอง

6. ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

การทำเหมืองข้อมูลมีทฤษฎีที่เกี่ยวข้องในการวิเคราะห์ความสัมพันธ์ทั้งด้านคอมพิวเตอร์และทางสถิติหลายทฤษฎี โดยแต่ละทฤษฎีนั้นได้มีการศึกษาอัลกอริทึมที่เกี่ยวข้องกับทฤษฎีต่างๆ ซึ่งในที่นี้ผู้วิจัยจะกล่าวถึงทฤษฎีที่เกี่ยวข้องกับงานวิจัยที่จัดทำขึ้น ดังนี้

6.1.1 การจำแนกประเภทข้อมูล (Data Classification)

เป็นการแบ่งหมวดหมู่โดยทำการกำหนดสิ่งที่เป็นลักษณะเด่นในแต่ละหมวดหมู่ซึ่งจะแบ่งข้อมูลตามความคล้ายคลึงกันจากตัวอย่างข้อมูลที่มีอยู่(supervised learning)

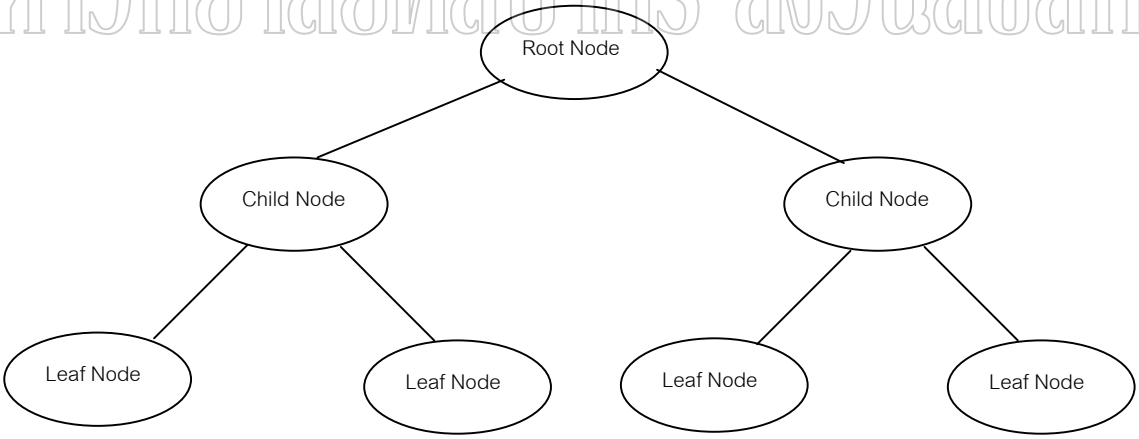
มีผู้วิจัยนำเทคนิคการจำแนกประเภทข้อมูลมาสร้างตัวจำแนกข้อมูล ในการวิจัยได้นำเทคนิคการจำแนกประเภท และเทคนิคการค้นหากฎความสัมพันธ์ มาประยุกต์ใช้ในการจัดสรรกฎหมายที่เหมาะสมกับคดีความ โดยนำเทคนิคเทคนิคการจำแนกประเภทข้อมูลมาสร้างตัวจำแนกข้อมูลจากกฎเกณฑ์ที่ได้จากเทคนิคการค้นหากฎความสัมพันธ์อีกทีหนึ่ง ซึ่งตัวจำแนกข้อมูลที่ได้นี้จะสามารถนำไปใช้ทำนายคดีความแต่ละคดีว่าควรใช้กฎหมายฉบับใดในการพิจารณา ผลการวิจัยที่ได้

แสดงให้เห็นว่าการสร้างตัวจำแนกตามวิธีที่เสนอนี้ได้ประสิทธิภาพดีกว่าการสร้างตัวจำแนกตามเทคนิค Data classification แบบปกติ นอกจากนี้ วิธีดังกล่าวยังช่วยแก้ปัญหาต่างๆที่เกิดขึ้นจากการใช้เทคนิคการจำแนกประเภทข้อมูลโดยทั่วไปได้อีกด้วย(กฤษณะ ไวยมัย และ ธีระวัฒน์ พงษ์ศิริปรีดา 2003:143)

6.1.2 ต้นไม้ตัดสินใจ (Decsion Trees)

เป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ในรูปของต้นไม้ตัดสินใจซึ่งต้นไม้ตัดสินใจนั้นมีการ สร้างแบบจำลองการจัดหมวดหมู่(Supervised Learning)ได้จากกลุ่มตัวอย่างของข้อมูลที่กำหนดได้ก่อนล่วงหน้า(Training Set) โดยอัตโนมัติและสามารถพยากรณ์กลุ่มของข้อมูลที่ยังไม่เคยนำมาจัดหมวดหมู่ได้

รูปแบบของต้นไม้จะประกอบด้วยโหนด (Node) แรกสุดที่เรียกว่าโหนดราก(Root Node) จาก โนดรากก็จะแตกออกเป็น โหนดลูก(Child Node) และที่ โนดลูกก็จะมีลูกของตัวเองซึ่งโนดที่ระดับสุดท้ายจะเรียก โหนดใบ (Leaf Node) ดังรูปที่ 6



รูปที่ 6 โครงสร้างต้นไม้

อัลกอรึทึมพื้นฐานของต้นไม้ตัดสินใจ คือ กรีดี อัลกอริทึม (greedy algorithm) ซึ่งเป็นการค้นหาจากบนลงล่าง (top-down) อัลกอริทึมนี้ ส่วนใหญ่ใช้กับข้อมูลที่เป็นแบบต่อเนื่อง (continuous data) ซึ่งจะจำกัดข้อมูลที่เป็นตัวแปรตาม (dependent variable) 1 ตัวแปรต่อ แบบจำลอง

ในกรณีที่มีตัวแปรหลายตัว จะต้องสร้างแบบจำลองของตัวแปรตามแต่ละตัว ซึ่งได้มีผู้พัฒนารูปแบบของอัลกอรึทึมนี้มากมาย ซึ่งแต่ละอัลกอริทึมก็มีความแตกต่างกัน ได้แก่

Classification and Regression Trees (CART) เป็นอัลกอริทึมในการทำการสำรวจข้อมูล และทำนายข้อมูล โดยใช้ค่าสัมประสิทธิ์จีนี (GINI)ในการตัดสินใจเลือกคุณสมบัติของโหนดแต่ละตัว [Breiman,et)]

Chi-squared Automatic Interaction Detection (CHAID) เป็นอัลกอริทึมที่คล้ายกับ CART แต่มีความแตกต่างกันที่วิธีการแยกข้อมูล

อัลกอริทึมที่แพร่หลายมากที่สุดคืออัลกอริทึม ID3 (Quinlan, 1986) และ C 4.5 (Quinlan , 1993) โดยใช้ค่ามาตรฐานเกน (Gain criterion) ในการตัดสินใจเลือกคุณสมบัติของโนดแต่ละตัว

มีผู้วิจัยได้ทำการศึกษาพฤติกรรมการใช้เว็บโดยการเรียงลำดับหน้าเว็บที่มีการเรียกใช้งาน จากข้อมูลจริงใน access log ที่ พร๊อกซี เซิร์ฟเวอร์โดยมี วัตถุประสงค์เพื่อทำนายความต้องการของผู้ใช้ และ โครงสร้างของพร๊อกซี สำหรับทำการดึงหน้าเว็บ ซึ่งงานวิจัยนี้ใช้ข้อมูลจริงในการทดลอง นี้แบ่งขั้นตอนในการทดลองเป็น 4 ขั้น ตอน คือ ขั้นตอนการเลือกข้อมูล ขั้นตอนการจัดการกับข้อมูลเป็นจัดทำข้อมูลให้เหมาะสมโดยใช้รูปแบบของต้นไม้ (tree) ก่อนการทำเหมืองข้อมูล , ขั้นตอนการทำเหมืองข้อมูลซึ่งใช้อัลกอริทึมต้นไม้ตัดสินใจ และขั้นตอนสุดท้ายคือ ขั้นตอนของการทำนาย ซึ่งผลของการทำนายที่ได้จะเก็บรวบรวมไว้เพื่อใช้ในการ prefetching การทดลองนี้มีข้อจำกัดคือพิจารณาเฉพาะกรณีที่มีผู้ใช้เพียงคนเดียว ผลการทดลองพบว่าการทำนายมีความถูกต้องสูง โดย hit ratio เหมาะสมที่สมที่สุดในการทำนายคือ 75.69% ซึ่งใช้เวลาในการทำนายเพียง 1.9 ms ซึ่งถือว่าใช้เวลาในการบริการโดยเฉลี่ยน้อยมาก (Yi-Hung and P.Chen 2002)

6.1.3 เครือข่ายใยประสาท (Neural Network)

เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence:AI) ซึ่งมีพื้นฐานมาจากการจำลองการทำงานของเซลล์สมองของมนุษย์ เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล ซึ่งเป็นวิธีการที่ซับซ้อนกว่าวิธีอื่น ผลลัพธ์ที่ได้จากวิธีนี้ จะต้องนำไปคำนวณค่าผิดพลาดก่อน เพื่อปรับค่าค่าถ่วงน้ำหนัก (Weight) ก่อน

ลักษณะของปัญหาที่เหมาะกับการใช้เครือข่ายใยประสาท คือ

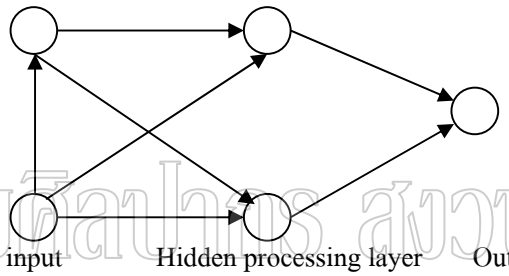
- สามารถระบุ คุณลักษณะ และ ความสำคัญ ของอินพุต (input)ชัดเจน

• สามารถกำหนด เอาต์พุต (output) ที่ชัดเจน

• มีข้อมูลตัวอย่างเพียงพอในการเรียนรู้

คุณลักษณะของเครือข่ายใยประสาท คือจะทำงานได้ดีที่สุดเมื่อมีการกำหนดช่วงอินพุต(input) และ เอาต์พุต (output) ระหว่าง 0 และ 1 จึงต้องมีการปรับค่าของข้อมูลโดย Neural Network จะต้องเรียนรู้ผ่านทางการฝึก (train) ซึ่งเป็นการปรับค่าถ่วงน้ำหนัก

โครงสร้างของเครือข่ายใยประสาทจะประกอบด้วยโหนด (Node) สำหรับ Input – Output และการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ input layer , output layer และ hidden layers การประมวลผลของเครือข่ายใยประสาทจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ใน layer เหล่านี้ ดังรูปที่ 7



รูปที่ 7 โครงสร้างเครือข่ายใยประสาท

ขั้นตอนในการนำเครือข่ายใยประสาทมาใช้กับการทำเหมืองข้อมูลในการสร้าง Model สำหรับการแยกประเภท และการทำนายค่า มีดังนี้

1. กำหนด input และ output
2. จัดการกับ input และ output ให้มีค่าอยู่ระหว่าง 0 และ 1
3. สร้าง Neural Network ที่มี topology ที่เหมาะสมกับงานนั้น
4. Train Neural Network โดยใช้ตัวอย่างจากชุดเทรน
5. ทดสอบ Neural Network โดยใช้ชุดทดสอบซึ่งเป็นอิสระจากชุดเทรน (ข้อมูลในชุดทดสอบจะต้องไม่ซ้ำกับข้อมูลที่อยู่ในชุดเทรน) ถ้าได้ผลไม่เป็นที่น่าพอใจอาจจะ ต้องเทรนใหม่ หรือเปลี่ยนชุดเทรน Topology และค่า Parameter ต่างๆ)
6. นำ Model ที่ได้ไปใช้ในการทำนาย หรือแยกประเภท

6.1.4 การแบ่งกลุ่มข้อมูล (Cluster Analysis)

การแบ่งกลุ่มเป็นเทคนิคที่ใช้จำแนกหรือแบ่งเป็นกรณี (คน สัตว์ สิ่งของ หรือองค์กร ฯลฯ) หรือแบ่งตัวแปรออกเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไป โดยกรณีที่อยู่กลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วนกรณีที่อยู่ต่างกลุ่มกันจะมีลักษณะที่แตกต่างกัน ดังนั้นการพิจารณาเลือกลักษณะตัวแปรที่จะนำมาใช้แบ่งกลุ่ม Caseจึงมีความสำคัญ นอกจากนั้น Case ใด Case หนึ่งจะอยู่ในกลุ่มหนึ่งเพียงกลุ่มเดียว (ดร. กัลยา วินิชย์บัญชา., 2001 : 125)

เทคนิคCluster Analysisแบ่งเป็นหลายประเภทโดยเทคนิคที่ใช้กันมากมี 2 เทคนิค คือ

1. Hierachical Cluster Analysis
2. K-Means Cluster Analysis

เทคนิค Hierachical Cluster Analysisเป็นเทคนิคที่นิยมใช้กันมากในการแบ่งกลุ่ม Case หรือแบ่งกลุ่มตัวแปรที่มีจำนวนไม่มาก สามารถใช้ได้กับตัวแปรเชิงคุณภาพและตัวแปรเชิงปริมาณ ซึ่งไม่จำเป็นต้องทราบว่าตัวแปรนั้นอยู่กลุ่มใด และมีจำนวนกลุ่มมากเท่าใด

เทคนิค K-Mean Clustering เป็นเทคนิคการจำแนก Case ออกเป็นกลุ่มย่อย โดยจะใช้กับ Case ที่มีจำนวนมากและต้องกำหนดจำนวนกลุ่ม ตัวแปรที่ใช้จะต้องเป็นตัวแปรเชิงปริมาณคือเป็นสเกลอันตรภาค (Interval Scale) หรือ สเกลอัตราส่วน (Ratio Scale) เท่านั้น ผู้วิเคราะห์จะต้องทำข้อมูลให้เป็นมาตรฐานก่อน

มีผู้วิจัยได้ทำการประยุกต์เทคนิคการทำเหมืองข้อมูลในการวิเคราะห์ ข้อมูล access log จาก Publico On-line ซึ่งเป็นเว็บหนังสือพิมพ์ โดยใช้โปรแกรมทางด้านเหมืองข้อมูล คือ SPSS Clementine และ IBM Intelligent Miner ผู้วิจัยมีการกำหนดช่องทางในการเข้า เว็บ Publico On-line ในเดือน เมษายน 1999 จำนวน 2 ช่องทาง คือ ช่องทางสั้น และช่องทางยาว ผู้วิจัยได้ทำการปรับปรุงข้อมูลโดยการแปลงข้อมูลจากตาราง access log ไปเป็นรูปแบบของตัวเลขและ ตรรกศาสตร์ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลโดยใช้สถิติพื้นฐาน โดยข้อมูลที่เป็นตัวแปรชนิดตัวเลข สามารถ หา ค่าต่ำสุด สูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐานได้ สำหรับข้อมูลที่เป็นตัวแปรตรรกศาสตร์ สามารถวิเคราะห์ได้โดยการหาความถี่ และนำเทคนิคการแบ่งกลุ่มข้อมูล clustering มาใช้ในการจัดกลุ่มของผู้ใช้ซึ่งแบ่งเป็น 2 กลุ่มคือ demographic clustering และ neural clustering ซึ่งผลของการวิเคราะห์ของการประยุกต์ โดยกลุ่มของข้อมูลที่เป็นตรรกศาสตร์ พบว่าลักษณะของผู้อ่านมีรูปแบบที่เหมือนกัน โดยส่วนใหญ่จะเข้าดูในส่วนของ กีฬาและ ข่าวต่างประเทศ (Paulo and Mario J. 1999)

6.1.5 ฟัซซีลอจิก (Fuzzy Logic)

ฟัซซีลอจิกเป็นเทคนิคในทางตรรกศาสตร์หรือการคิดหาเหตุผลแบบหนึ่ง ซึ่งจะเป็นวิธีการในการจัดการกับกฎเกณฑ์ความรู้ที่มีความคลุมเครือของข้อมูลแฝงอยู่ด้วย ฟัซซีลอจิก เป็นศาสตร์ที่พยายามเลียนแบบวิธีการคิดหาเหตุผลของมนุษย์ เป็นวิธีการที่ทำให้คอมพิวเตอร์สามารถทำงานหรือคิดหาเหตุผลได้ภายใต้เงื่อนไขหรือกฎเกณฑ์ที่มีความไม่แน่นอนของข้อมูลระดับหนึ่ง (มหาวิทยาลัยสุโขทัยธรรมาธิราช 2545:439)

หลักการสำคัญของ ฟัซซีลอจิก คือการแทนรูปของกฎเกณฑ์ความรู้ในลักษณะ “เงื่อนไข ถ้า……..” ให้ โดยที่ความจริงของเหตุการณ์ที่เกี่ยวข้อง ไม่จำเป็นต้องมีเพียงค่า “จริง” หรือ “เท็จ” เหมือนที่ใช้กันในตรรกศาสตร์แบบปกติ (Classical Logic หรือ Boolean Logic) แต่ค่าความจริงของประเด็นที่สนใจใน ฟัซซีลอจิก นี้ อาจจะมีได้หลายระดับหรือมีขั้นกว่าในระดับต่างๆ เช่นจะมีวิธีการในการเก็บความรู้ที่แทนความหมายของคำว่า “น้อย” “ค่อนข้างน้อย” “ปานกลาง” “ค่อนข้างมาก” “มากที่สุด” เป็นต้น

ได้เสนอวิธีใหม่เพื่อใช้ทำนายการเข้าเว็บ โดยใช้เทคนิคของกฎความสัมพันธ์ฟัซซี ซึ่งเป็นการนำเอากฎฟัซซี และ เหตุผลในการเข้าใช้ (case-base reasoning:CBR) มาเป็นความคิดใหม่ในการทำนายโดยใช้ข้อมูลจาก web log ซึ่งกฎความสัมพันธ์ฟัซซี(fuzzy association rule) ในการพัฒนาประสิทธิภาพการทำนายและนำดัชนีต้นไม้ของฟัซซี (fuzzy index tree) มาช่วยในการเพิ่มความเร็วในการทำงานของกฎความสัมพันธ์ที่พัฒนาขึ้น โดยจะพิจารณาถึงช่องทางที่ผู้ใช้แต่ละคนเข้าใช้ในแต่ละช่วงเวลา (Wong, Shiu, and Pal 2001)

6.1.6 กฎการวิเคราะห์ความสัมพันธ์ (Association rule)

เป็นการหาความสัมพันธ์ระหว่างข้อมูลด้วยกันเองซึ่งมีพื้นฐานมาจากการเกิดขึ้นร่วมกันหรือพร้อมกันในฐานข้อมูล

รูปแบบของการค้นหากฎความสัมพันธ์

รูปแบบทั่วไปของการค้นหากฎความสัมพันธ์ คือ $A \rightarrow B$

โดยที่ A : เป็นเงื่อนไข หรือ LHS (Left - Hand Side)

และ B : เป็นผลลัพธ์ที่เกิดขึ้น หรือ RHS (Right – Hand Side)

หรืออยู่ในรูปของ “ถ้า…….แล้ว” (If……Then….) เช่น

$A \rightarrow B$; if A Then B เป็นกฎที่ 1

B→A ; if B Then A เป็นกฎที่ 2

การประเมินค่าของกฎจะใช้ค่าสนับสนุน(Support) และค่าความเชื่อมั่น (Confidence)

โดยที่

ค่าสนับสนุน คือ เปอร์เซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

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

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการหาผลลัพธ์

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์

Transaction(A,B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

ค่าความเชื่อมั่น คือเปอร์เซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

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

โดยที่ Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว

ในการเลือกว่าจะกฎใดนั้นจะต้องพิจารณาค่าสนับสนุน และค่าความเชื่อมั่นที่มีค่าสูงกว่าค่า Threshold ที่ตั้งไว้ นอกจากนี้จะต้องกำหนดค่าสนับสนุนต่ำสุด (Minimum Support) และค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) โดยทั่วไปจะกำหนดค่าสนับสนุนต่ำสุดเป็น 5-10 % และค่าความเชื่อมั่นต่ำสุดเป็น 50-100%

ประเภทของการค้นหากฎความสัมพันธ์

1. Boolean Association Rule : เป็นการแสดงความสัมพันธ์ระหว่างสิ่งที่มีอยู่หรือไม่อยู่

   buys(X,"computer") → buys(X, "financial_management_software")

   หมายความว่า ถ้า X ซื้อ หนังสือ computer แล้ว X จะซื้อซอฟต์แวร์การจัดการการเงิน

2. Quantitative Association Rule : เป็นการแสดงความสัมพันธ์ของข้อมูลที่เป็นตัวเลข (Interval) เช่น

   age(X,"30…39") ∧ income(X,"42K…48K")→ buys(X, high resolution TV)

   หมายความว่า ถ้า X อายุระหว่าง 30-39 และมีรายได้ระหว่าง 42K-48K บาท แล้ว X จะซื้อทีวีที่มีความละเอียดสูง

3. Single-Dimension Association Rule : ซึ่งจะอยู่ในรูปแบบดังนี้

   buys(X,"computer") → buys(X, "financial_management_software")

4. Multi dimension : ซึ่งจะอยู่ในรูปแบบดังนี้

   age(X,"30…39") ∧ income(X,"42K…48K")→ buys(X, high resolution TV)

5. Multilevel Association Rule : เป็นกฎที่สร้างจากแอทริบิวต์ ที่มีระดับต่างกัน

   Age(X,"30-39") → (X,"laptop_computer")

   Age(X,"30-39") → (X,"Computer")

อัลกอริทึมกฎการวิเคราะห์ความสัมพันธ์

มีหลายอัลกอริทึมที่ใช้ในการค้นหาความสัมพันธ์เช่น

- อัลกอริทึม Apriori
- อัลกอริทึม AprioriTid
- อัลกอริทึม AIS
- อัลกอริทึม SETM

อัลกอริทึม Apriori (agrawal and Srikant 1994) เป็นประเภทของการค้นหาความสัมพันธ์แบบ Boolean Association Rule ซึ่งเป็นอัลกอริทึมที่ใช้กันอย่างแพร่หลาย คือเทคนิควิธีที่ใช้สำหรับค้ นหาสิ่งที่ปรากฏเด่นชัด (Frequent Itemsets)จากฐานข้อมูลที่ กำหนดโดยมีหลักการทำงานคืออัลกอริทึม Apriori ทำหน้าที่สร้างเซ็ตไอเท็มหรือกลุ่มข้อมูลที่ต้องการวิเคราะห์ที่เป็นไปได้ทั้งหมดที่มีค่าสนับสนุน มากกว่าค่าสนับสนุนขั้นต่ำโดยอัลกอริทึม Apriori เป็นการทำงานในแบบล่างขึ้นบน (bottom up) โดยมีขั้นตอนดังนี้

ขั้ นตอนที่1 อัลกอริทึม Apriori อ่านฐานข้อมูลทั้งหมดและสร้างไอเท็มเซ็ตที่ผ่านค่าสนับสนุนขั้นต่ำความยาว 1 ไอเท็ม (Frequent 1-itemset)

ขั้นตอนที่2 อัลกอริทึม Apriori สร้างเซ็ตไอเท็มทดสอบ (Candidate Itemset) ที่มีความยาว 2 ไอเท็มจากเซ็ตไอเท็มที่ปรากฏเด่นชัดความยาว 1 ไอเท็มในขั้นตอนแรกและนำไปหาค่าสนับสนุน เพื่อค้นหาไอเท็มเซ็ตที่ ปรากฏเด่นชัดความยาว 2 ไอเท็ม กรรมวิธีอพิออรีจะวนรอบทำงานจนกระทั่งไม่พบไอเท็มเซ็ตที่ผ่านค่าสนับสนุนขั้นต่ำจึงจบการทำงานไอเท็มเซ็ตที่ผ่านค่าสนับสนุน ขั้นต่ำในแต่ละรอบคือสิ่งที่ปรากฏเด่นชัดจากฐานข้อมูล

ในปี 1994 มีผู้วิจัยเสนอแนวคิดใหม่ในการหากฎความสัมพันธ์ของฐานข้อมูล ซึ่งเป็นเทคนิคหนึ่งที่สำคัญของเหมืองข้อมูล (Data Mining) งานวิจัยได้อธิบายถึง วิธีการ และจุดเด่นของอัลกอริทึมใหม่ 2 อัลกอริทึม คือ อัลกอริทึม AprioriTid และอัลกอริทึม AprioriHybrid ที่ใช้ในการหากฎความสัมพันธ์ในฐานข้อมูลการขายสินค้าขนาดใหญ่ อัลกอริทึมใหม่ทั้งสองอัลกอริทึมนี้อ้างอิงมาจากอัลกอริทึม Apriori ซึ่งผู้วิจัยได้แสดงการพัฒนาอัลกอริทึม AprioriHybrid จากการรวมอัลกอริทึม Apriori และ AprioriTid มีการเปรียบเทียบประสิทธิภาพของอัลกอริทึมใหม่กับอัลกอริทึมแบบเดิมที่ใช้ในการหากฎความสัมพันธ์ เช่น อัลกอริทึม AIS, อัลกอริทึม SETM และอัลกอริทึม Apriori การทดสอบประสิทธิภาพของอัลกอริทึมต่าง ๆ นี้ได้ทดสอบกับฐานข้อมูลที่มีขนาดและรูปแบบต่าง ๆ กัน จากการทดสอบประสิทธิภาพของอัลกอริทึมใหม่พบว่ามีประสิทธิภาพดีกว่าอัลกอริทึมที่เคยมีการนำเสนอมาแล้วมากกว่า 3 เท่าสำหรับฐานข้อมูลที่มีข้อมูลเป็นจำนวนมาก(Agrawal and Srikant 1994)

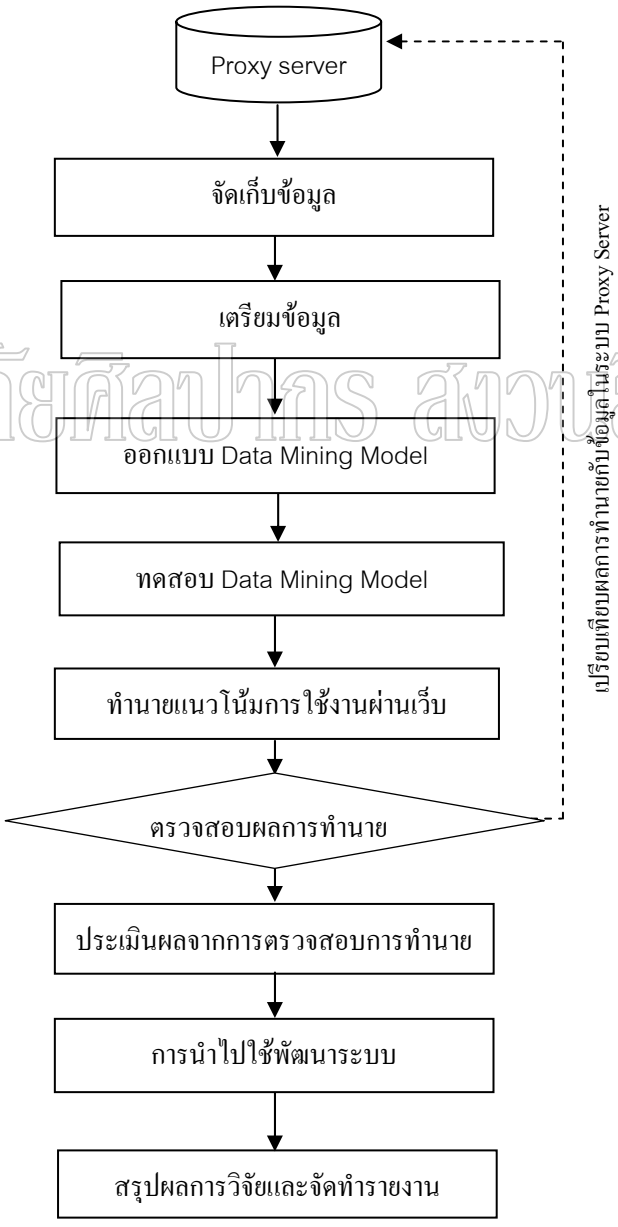
ในปี 1999 มีผู้วิจัยได้ทำการประยุกต์เทคนิคการทำเหมืองข้อมูลในการวิเคราะห์ ข้อมูล access log จาก Publico On-line ซึ่งเป็นเว็บหนังสือพิมพ์ โดยใช้โปรแกรมทางด้านเหมืองข้อมูล คือ SPSS Clementine และ IBM Intelligent Miner ผู้วิจัยมีการกำหนดช่องทางในการเข้าเว็บ Publico On-line ในเดือน เมษายน 1999 จำนวน 2 ช่องทางคือ ช่องทางสั้น(shot session)ซึ่งเป็นจำนวนกลุ่มของผู้ใช้ที่เข้าใช้จาก IP เดียวกัน และช่องทางยาว (long session) ซึ่งเป็นกลุ่มของผู้ใช้ที่ใช้งานติดต่อกันโดยเอาข้อมูลมาจากcookie เป็นเวลา 1 เดือน ผู้วิจัยได้ทำการปรับปรุงข้อมูลโดยการแปลงข้อมูลจากตาราง access log ไปเป็นรูปแบบของตัวเลขและ ตรรกศาสตร์ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลโดยใช้สถิติพื้นฐาน โดยข้อมูลที่เป็นตัวแปรชนิดตัวเลข สามารถ หา ค่าต่ำสุด สูงสุด ค่าเฉลี่ย ค่าส่วนเบี่ยงเบนมาตรฐานได้ สำหรับข้อมูลที่เป็นตัวแปรตรรกศาสตร์ สามารถวิเคราะห์ได้โดยการหาความถี่ การวิเคราะห์ความสัมพันธ์โดยใช้ อัลกอริทึมApriori ในการวิเคราะห์ความสัมพันธ์ โดยกำหนดค่าสนับสนุนที่ต่ำที่สุด 10% และค่าความเชื่อมั่นต่ำสุด 60% ซึ่งพบว่า กฎความสัมพันธ์ในช่องทางที่สั้นเกือบทั้งหมดอยู่ในกลุ่มของกฎความสัมพันธ์ของช่วงทางที่ยาว (Batista and Mario J. 1999)

ในปี 2002 มีผู้วิจัยได้เสนอรูปแบบสำหรับการทำนายเหตุการณ์ที่เกิดขึ้นบนเว็บ ซึ่งเป็นการทำนายช่วงเวลาไหนที่จะมีการเรียกใช้เว็บที่เคยถูกเรียกใช้แล้วอีกครั้ง เช่น เมื่อสำรวจพบว่ามีผู้เรียกใช้เว็บ A และ B น้อยกว่า เว็บ C แล้ว ระบบอาจทำนายว่า เว็บ C ควรที่จะเป็นเพจต่อไปที่จะถูกเรียกใช้ภายใน 10-20 วินาที จากเวลาปัจจุบัน ซึ่งเป็นวิธีการจำแนกประเภทโดยกฎความสัมพันธ์ (association rule of classification method) ผู้วิจัยเก็บข้อมูลจริงจากเว็บเซิร์ฟเวอร์ ของ NASA เป็นเวลา 1 เดือน โดยนำ อัลกอริทึมClassification มาใช้ในขั้นตอนของการเตรียมข้อมูลและใช้ประยุกต์สร้างรูปแบบการทำนายเวลาที่จะเกิดขึ้นที่เรียกว่า อัลกอริทึม moving-window ที่พัฒนามาจากจากกฎความสัมพันธ์ (LHS $\rightarrow$ RHS) ซึ่งการวิจัยนี้ ได้ศึกษาความแตกต่างของวิธี temporal regian prediction ซึ่งเป็นเทคนิคที่พัฒนามาจากกฎการค้นหาความสัมพันธ์ โดยทำเปรียบเทียบความแตกต่างของวิธี temporal regian prediction 3 วิธี คือ 1) วิธี naïve 2) วิธีcomfidence interval bases 3) วิธี minimal temperal region โดยผลการทดลองปรากฏว่า วิธี comfidence interval bases และ วิธี minimal temperal region ได้ผลการทดลองที่ใกล้เคียงกัน ซึ่งวิธี minimal temperal region จะมีความถูกต้องแม่นยำกว่าเพียงเล็กน้อย (Yang, Wang,and Zhang 2002)

ในปี 2005 มีผู้วิจัยได้ทำการศึกษาวิธีที่ทำให้แคชสามารถใช้ข้อมูลร่วมกันได้ ซึ่งพร็อกซีทุกตัวยังคงทำงานอย่างมีประสิทธิภาพที่วัดจากระดับความน่าเชื่อถือของ hit ratio งานวิจัยนี้กำหนดให้แคชเป็นเหมือนแคชส่วนตัวโดยขึ้นอยู่กับการรูปแบบการเข้าใช้ของผู้ใช้ ซึ่งผู้วิจัยได้ออกแบบสถาปัตยกรรมของแคช โดยมี web cache server ทำหน้าที่เป็นตัวจัดการระบบ web proxy cache และทำการจัดเก็บข้อมูลที่ได้จาก proxy cache แต่ละตัว โดย proxy cacheจะแชร์ข้อมูลและกำหนดรูปแบบการใช้งานของ client โดยการสุ่ม งานวิจัยนี้ผู้วิจัยได้จัดกลุ่มตัวแทน(agent) ของผู้ใช้เป็น 4 กลุ่มโดยวิธีอัลกอริทึม Classification และใช้อัลกอริทึมApriori ของ Association rule เพื่อบอกลำดับความถี่และเวลาในการใช้เว็บจาก cache log ซึ่งการวิจัยครั้งนี้ถือเป็นการศึกษารูปแบบและจัดเตรียมแคชที่มีประสิทธิภาพให้กับผู้ใช้ในการค้นหาข้อมูลโดยซอร์ฟแวร์นี้สามารถทำนายเวลาที่ใช้โปรแกรมกับการใช้ทรัพยากรคอมพิวเตอร์อย่างเหมาะสม (Mohan, E.K. ,and Han 2005)

บทที่ 3
วิธีการดำเนินการวิจัย

งานวิจัยนี้ได้นำเทคนิคการทำเหมืองข้อมูลเพื่อทำนายการใช้งานของผู้ใช้เว็บโดยข้อมูลที่ใช้ในงานวิจัยนี้คือข้อมูลจริงของการใช้เว็บจากระบบพร็อกซี เซิร์ฟเวอร์ โดยได้กำหนดขั้นตอนการดำเนินการวิจัยดังนี้



รูปที่ 8 แผนผังขั้นตอนการดำเนินการวิจัย

1. การจัดเก็บข้อมูล

งานวิจัยนี้ได้ดำเนินการจัดเก็บข้อมูลการเข้าใช้เว็บจริง ของผู้ใช้ภายในมหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ จากพร๊อกซี เซิร์ฟเวอร์ จำนวน 3 ตัว เป็นเวลา 3 เดือน ตั้งแต่วันที่ 1 กันยายน 2548 ถึงวันที่ 30 พฤศจิกายน 2548 และฐานข้อมูลเว็บ ดังรายละเอียดต่อไปนี้

1.1 ข้อมูลจากพร๊อกซี เซิร์ฟเวอร์

```
1128531640.658    148 202.44.135.35 TCP_REFRESH_HIT/200 504 GET
http://www.sanook.com/menu/images/sm2nbg.gif -
DIRECT/203.107.136.7 image/gif
1128531640.704    211 202.44.135.35 TCP_MISS/200 4809 GET
http://www.sanook.com/menu/nav.php -
TIMEOUT_DIRECT/203.107.136.7 text/html
1128531647.627    116 172.27.7.35 TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php? - DIRECT/164.115.2.146
image/jpeg
1128531650.101    119 172.27.7.35 TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php? - DIRECT/164.115.2.146
image/jpeg
```

รูปที่ 9 ตัวอย่างข้อมูล Access log มหาวิทยาลัยศิลปากร

ข้อมูลที่ทำการจัดเก็บจากพร๊อกซี เซิร์ฟเวอร์ เป็นข้อมูลของ access log ซึ่งข้อมูลเป็นชนิด text ไฟล์ ดังรูปที่ 3-1 โดยประกอบด้วยข้อมูล 10 ข้อมูล ดังนี้

Time : เวลาที่ใช้ในการใช้งาน ซึ่งเป็น Unix time stamp

Duration : ระยะเวลาที่ใช้ในการทำงานของโปรแกรมต่อการร้องขอนั้น มีหน่วยเป็นมิลลิวินาที

Client address : IP address ของผู้ที่ร้องขอ

Result Code :เก็บผลการทำงานของการร้องแต่ละครั้ง

Byte :เก็บขนาดของข้อมูลทั้งหมดที่ถูกส่งไปยังผู้รับบริการ แต่ข้อมูลส่วนนี้ไม่ใช่ขนาดของข้อมูลที่แท้จริงเพราะจะนับรวมถึงส่วนของ header ด้วย

Request Method : เก็บข้อมูลรูปแบบของการร้องขอที่เกิดขึ้นกับข้อมูล

URL : เก็บที่อยู่ของข้อมูลที่ถูกร้องขอ

Rfc931 : ทำหน้าที่เก็บลักษณะของไคลเอนต์ที่ร้องขอข้อมูล

Hieraechy Code : เก็บข้อมูลที่เกี่ยวข้องกับการทำงานเป็นลำดับชั้นของโปรแกรม

Type :ประเภทของออบเจ็กต์ที่ถูกส่งมาให้ผู้ใช้

1.2 ฐานข้อมูลเว็บ

รวบรวมรายละเอียดของหมวดเว็บต่างๆ จากเว็บที่มีการดำเนินการแบ่งหมวดเว็บไว้แล้ว เช่น google , sanook , meesook เป็นต้น หรือจากเนื้อหาการให้บริการของแต่ละเว็บ ซึ่งสามารถแบ่งหมวดเว็บได้ดังรูปที่ 10

	Group_Id	Group_Name
+	1	การศึกษา
+	2	กิจกรรมและเหตุการณ์สำคัญ
+	3	ข่าว และสื่อ
+	4	คอมพิวเตอร์
+	5	ท่องเที่ยว
+	6	ธุรกิจ
+	7	บุคคล สังคม และวัฒนธรรม
+	8	หน่วยงานราชการและองค์กร
+	9	อินเทอร์เน็ต
+	10	การแพทย์และสุขภาพ
+	11	กีฬา
+	12	ความรู้และข้อมูลสำคัญ
+	13	ชอปปิ้ง
+	14	ธนาคาร และสถานบันเทิง
+	15	บันเทิงและนันทนาการ
+	16	ยานยนต์
+	17	อสังหาริมทรัพย์
+	18	เกมส์
+	19	อื่นๆ
+	20	ไม่ถูกจัดหมวดหมู่ไว้

รูปที่ 10 ตัวอย่างแสดงตัวอย่างหมวดเว็บ

ซึ่งจะได้หมวดเว็บทั้งหมด จำนวน 20 หมวดเว็บ ในแต่ละหมวดเว็บจะมีหมวดเว็บย่อย เว็บที่รวบรวมได้ทั้งหมดจำนวน 8148 เว็บ โดยเว็บที่ไม่ได้ถูกจัดหมู่ไว้จะถูกจัดอยู่ในรหัสหมวดที่ 20 คือไม่ถูกจัดหมวดหมู่ไว้ แต่ละเว็บจะมีหมวดเว็บที่สังกัดเพียงหมวดเดียวเท่านั้น

2. การเตรียมข้อมูล

2.1 คัดเลือกข้อมูล

งานวิจัยนี้นำข้อมูลการใช้เว็บภายในมหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์มาใช้ การเตรียมข้อมูลเพื่อนำไปสร้างโมเดลคัดเลือกเฉพาะข้อมูลที่มีความสัมพันธ์กันคือ

- Date วัน/เดือน/ปี ในการใช้งาน
- Times เวลา ในการใช้งาน
- Result Code เก็บผลการทำงานของการร้องแต่ละครั้ง
- URL เก็บที่อยู่ของออบเจ็กต์ที่ถูกร้องขอ

2.2 แปลงข้อมูลจาก text ไฟล์ ให้เป็นข้อมูลที่สามารถประมวลผลได้ โดยแปลงข้อมูลลงใน SAS ดังรูปที่ 11

	Date	Times	newURL	newResultCode
1	31/08/2005	5:00:11 PM	www.mthai.com	TCP_HIT
2	31/08/2005	5:00:11 PM	soi13.com	TCP_IMS_HIT
3	31/08/2005	5:00:11 PM	www.yimsiam.com	TCP_HIT
4	31/08/2005	5:00:11 PM	www.yimsiam.com	TCP_HIT
5	31/08/2005	5:00:11 PM	www.mthai.com	TCP_HIT
6	31/08/2005	5:00:11 PM	img364.imageshack.us	TCP_REFRESH_HIT
7	31/08/2005	5:00:11 PM	www.mthai.com	TCP_HIT
8	31/08/2005	5:00:11 PM	www.thaigraph.com	TCP_REFRESH_HIT
9	31/08/2005	5:00:11 PM	www.thaigraph.com	TCP_NEGATIVE_HIT
10	31/08/2005	5:00:11 PM	www.thaigraph.com	TCP_REFRESH_HIT
11	31/08/2005	5:00:11 PM	www.mthai.com	TCP_HIT
12	31/08/2005	5:00:11 PM	gamecenter.kapook.com	TCP_HIT
13	31/08/2005	5:00:12 PM	saosex.porkyhost.com	TCP_IMS_HIT
14	31/08/2005	5:00:12 PM	www.mthai.com	TCP_HIT
15	31/08/2005	5:00:12 PM	www.mthai.com	TCP_HIT
16	31/08/2005	5:00:13 PM	www.yimsiam.com	TCP_HIT
17	31/08/2005	5:00:13 PM	www.yimsiam.com	TCP_HIT
18	31/08/2005	5:00:14 PM	www.mthai.com	TCP_HIT
19	31/08/2005	5:00:15 PM	soi13.com	TCP_IMS_HIT
20	31/08/2005	5:00:15 PM	soi13.com	TCP_IMS_HIT

รูปที่ 11 ตัวอย่างข้อมูลที่แปลงจาก text ไฟล์

2.3 ปรับปรุงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมในการสร้างโมเดล เช่น แปลงเวลาจาก Unix timestamp เป็น เวลาปัจจุบัน

2.4 จัดเก็บข้อมูลไว้ในระบบเหมืองข้อมูล

3. การสร้างโมเดล (Modeling)

งานวิจัยนี้เลือกใช้เทคนิคการค้นหากฎความสัมพันธ์ (Association Rule) เพื่อใช้ในการทำนายเนื้อหาเว็บโดยพิจารณาจากร้อยละของค่าความเชื่อมั่น และค่าสนับสนุนที่เกิดขึ้น

สามารถหาค่าความเชื่อมั่น และค่าสนับสนุนจากสมการดังนี้

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

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

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการทำนาย

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์ที่ได้จากการทำนาย

Transaction (A, B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว

การหากฎความสัมพันธ์ทั้งหมดจะต้องมีค่าสนับสนุนมากกว่าค่าสนับสนุนต่ำสุดที่กำหนดไว้ และมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นต่ำสุดที่กำหนดไว้

4. การทดสอบโมเดล

การทดสอบและตรวจสอบความถูกต้องของโมเดลที่สร้างขึ้น จะเปรียบเทียบกับผลลัพธ์ที่ได้จากการใช้เครื่องมือของ SAS

5. ทำนายแนวโน้มการใช้งานผ่านเว็บ

การทำนายการใช้งานผ่านเว็บเป็นการนำผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยโปรแกรม SAS ของการใช้เว็บที่มีการใช้งานมากในแต่ละวัน โดยการเรียงลำดับจากมากไปน้อย นำไปทำนายการใช้งานวันถัดไป ซึ่งใช้เวลาในการทำนายเป็นเวลา 3 เดือน

6. ตรวจสอบผลการทำนาย

นำผลที่ได้จากการทำนาย ไปเปรียบกับการใช้เว็บจริงในแต่ละวัน หากเว็บที่ใช้ในการทำนายตรงกับเว็บที่ถูกใช้งานจริงในแต่ละวัน ถือว่าผลของการทำนายถูกต้อง

7. ประเมินผลจากการตรวจสอบการทำนาย

นำผลที่ได้จากการตรวจสอบมาประเมินผล โดยวัดจากประสิทธิภาพและความน่าเชื่อถือที่ตั้งไป

8. การนำไปใช้การพัฒนาระบบ

นำระบบที่พัฒนาขึ้นไปช่วยสนับสนุนการตัดสินใจในการจัดเก็บเว็บในระบบพร็อกซี เซิร์ฟเวอร์

9. สรุปผลการวิจัยและจัดทำรายงานวิทยานิพนธ์

เมื่อการทำวิจัยสำเร็จบรรลุตามวัตถุประสงค์แล้ว หลังจากนั้นก็สรุปผลการวิจัยและจัดทำรายงานวิทยานิพนธ์

บทที่ 4
ผลการดำเนินงานวิจัย

ในการดำเนินงานวิจัย ผู้วิจัยได้พัฒนาโปรแกรมขึ้นด้วย Microsoft Visual Basic 6 เพื่อให้งานวิจัยบรรลุตามวัตถุประสงค์ คือศึกษาและพัฒนาตัวแบบเพื่อใช้ในการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล กรณีศึกษามหาวิทยาลัยศิลปากร โดยใช้ตัวแบบที่สร้างขึ้น
ขั้นตอนการดำเนินวิจัยแบ่งขั้นตอนออกเป็นดังนี้

1. การเตรียมข้อมูล เพื่อใช้สำหรับสร้างตัวแบบ
2. ศึกษาตัวแบบเพื่อใช้ในการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล
3. พัฒนาตัวแบบเพื่อใช้ในการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล
4. ทดสอบตัวแบบที่พัฒนาขึ้น
5. พัฒนาระบบพร๊อกซี เซิร์ฟเวอร์โดยใช้ตัวแบบที่สร้างขึ้น

ขั้นตอนต่างๆ สามารถอธิบายขั้นตอนการทำงานได้ดังนี้

1. การเตรียมข้อมูล

1.1 อธิบายข้อมูล

ข้อมูลที่นำมาใช้ในงานวิจัยในครั้งนี้ มีการจัดเก็บข้อมูล 2 ประเภท คือ

1.1.1 จัดเก็บข้อมูลจริงจากการเรียกใช้งานเว็บในระบบพร๊อกซี เซิร์ฟเวอร์ของมหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ ระหว่างเดือนกันยายน – พฤศจิกายน พ.ศ. 2548 เป็นข้อมูลของ access log ซึ่งข้อมูลเป็นชนิด text ไฟล์ โดยประกอบด้วยข้อมูล 10 เขตข้อมูล (Field) ดังนี้

- **Time** : เวลาที่ใช้ในการใช้งาน ซึ่งเป็น Unix time stamp
- **Duration** : ระยะเวลาที่ใช้ในการทำงานของโปรแกรมต่อการร้องขอนั้น มีหน่วยเป็นมิลลิวินาที
- **Client address** : IP address ของผู้ที่ร้องขอ
- **Result Code** : เก็บผลการทำงานของการร้องแต่ละครั้ง
- **Byte** : เก็บขนาดของข้อมูลทั้งหมดที่ถูกส่งไปยังผู้รับบริการ แต่ข้อมูลส่วนนี้ไม่ใช่ขนาดของข้อมูลที่แท้จริงเพราะจะนับรวมถึงส่วนของ header ด้วย

- **Request Mothod** : เก็บข้อมูลรูปแบบของการร้องขอที่เกิดขึ้นกับข้อมูล
- **URL** : เก็บที่อยู่ของข้อมูลที่ถูกร้องขอ
- **RFC931** : ทำหน้าที่เก็บลักษณะของเครื่องลูกข่ายที่ร้องขอข้อมูล
- **Hierarchy Code** : เก็บข้อมูลที่เกี่ยวข้องกับการทำงานเป็นลำดับชั้นของโปรแกรม
- **Type** :ประเภทของออบเจ็กต์ที่ถูกส่งมาให้ผู้ใช้

1.1.2 จัดทำฐานข้อมูลเว็บ โดยแยกจัดเก็บตามหมวดหมู่ โดยประกอบด้วยข้อมูลดังนี้

- IDsite : รหัสเว็บไซต์
- Title : ชื่อเว็บไซต์
- Description : คำอธิบายของเว็บ
- Keywords : คำสำคัญของเว็บ
- URL : ชื่อเรียกที่อยู่ของเว็บไซต์
- SubGroup_Name : ชื่อหมวดเว็บย่อย
- SubGroup_ID : รหัสหมวดเว็บย่อย
- Group_ID : รหัสหมวดเว็บ
- Group_Name : ชื่อหมวดเว็บ

1.2 การแปลงข้อมูล

ในขั้นตอนการแปลงข้อมูลผู้วิจัยได้พัฒนาโปรแกรมสำหรับการแปลงข้อมูลเพื่อให้ได้ข้อมูลที่พร้อมจะนำไปใช้ต่อไป โดยมีรายละเอียดดังนี้

1.2.1 ข้อมูลการเรียกใช้งานเว็บในระบบพร็อกซี เซิร์ฟเวอร์ข้อมูลของ access log ต้องมีการแปลงข้อมูลได้แก่

- แปลงข้อมูล Access log ที่เป็นข้อมูลชนิด text file ลงในฐานข้อมูล SAS ทั้งหมด
- แปลงข้อมูลเวลา Unix timestamp ให้อยู่ในรูปแบบของวันและเวลา เช่น 01/09/2005 , 12.00.00 AM
- คัดเลือกข้อมูล URL เฉพาะชื่อURL เท่านั้นเช่น www.mthai.com

เมื่อมีการแปลงข้อมูลเสร็จเรียบร้อยจึงคัดเลือกข้อมูล Access log ที่เกี่ยวข้องกับงานวิจัยดังนี้

- วัน
- เวลา
- URL
- Resultcode ที่คัดเลือกเฉพาะในส่วนที่พบ(HIT)เท่านั้น

1.2.2 ฐานข้อมูลเว็บที่จัดเก็บไว้ในระบบ Microsoft Access ต้องแปลงข้อมูลลงใน SAS โดยมีการแปลงข้อมูลต่อไปนี้

- แปลงข้อมูล URL ให้เหลือแต่เฉพาะชื่อ เช่น แปลง http://www.mthai.com เป็น www.mthai.com

เมื่อมีการแปลงข้อมูลเสร็จเรียบร้อยจึงคัดเลือกข้อมูลเว็บที่เกี่ยวข้องกับงานวิจัยครั้งนี้

- Title : ชื่อเว็บไซต์
- Description : คำอธิบายของเว็บ
- URL : ชื่อเรียกที่อยู่ของเว็บไซต์
- SubGroup_Name : ชื่อหมวดเว็บย่อย
- SubGroup_ID : รหัสหมวดเว็บย่อย
- Group_ID : รหัสหมวดเว็บ
- Group_Name : ชื่อหมวดเว็บ

เมื่อเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมแล้วนำก็นำข้อมูลนั้นไปใช้ในขั้นตอนการสร้างตัวแบบต่อไป

2. ศึกษาตัวแบบเพื่อใช้ในการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูล

ในการศึกษาตัวแบบนั้นผู้วิจัยได้เลือกเทคนิคการค้นหากฎความสัมพันธ์ และศึกษาทฤษฎีการค้นหากฎความสัมพันธ์ ดังนี้

ศึกษาการค้นหากฎความสัมพันธ์

รูปแบบของการค้นหากฎความสัมพันธ์สามารถเขียนได้ดังนี้

A → B

โดยที่ A เป็นเงื่อนไข และ B เป็นผลลัพธ์ที่เกิดขึ้น

ในงานวิจัยนี้ผู้วิจัยต้องการค้นหาความสัมพันธ์ระหว่างวัน เวลา หมวดเว็บ และเว็บซึ่งกำหนดให้เงื่อนไข คือ วัน เวลา และหมวดเว็บ ผลลัพธ์ที่เกิดขึ้น คือ เว็บซึ่งสามารถเขียนกฎความสัมพันธ์ได้ดังนี้

วัน , เวลา และเว็บ → เว็บ

เช่น 01/09/2005 , 12.00.00 AM , การศึกษา → www.su.ac.th

จากความสัมพันธ์นี้สามารถอธิบายได้ว่าวันที่ 01/09/2005 เวลา12.00.00 AM ผู้ที่เรียกใช้เว็บในหมวดการศึกษา มีแนวโน้มที่จะเรียกใช้เว็บ www.su.ac.th

การประเมินค่าของกฎจะใช้ค่าสนับสนุน(Support) และค่าความเชื่อมั่น (Confidence) โดยที่

ค่าสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

$$\frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลทั้งหมด}} \quad \text{.....สมการที่ 1}$$

ดังนั้นงานวิจัยนี้สามารถหาค่าสนับสนุนจากสมการดังนี้

ค่าสนับสนุน(วัน, เวลา,หมวดเว็บ → เว็บ) เท่ากับ

$$\frac{\text{จำนวนรายการข้อมูลที่มีวัน,เวลา ,หมวดเว็บและเว็บตามกฎ}}{\text{จำนวนรายการข้อมูลทั้งหมด}} \quad \text{.....สมการที่ 2}$$

ค่าความเชื่อมั่น คือร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนรายการข้อมูลที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

$$\frac{\text{จำนวนรายการข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องกันตามกฎ}}{\text{จำนวนรายการข้อมูลที่เป็นเงื่อนไข}} \quad \text{.....สมการที่ 3}$$

ดังนั้นงานวิจัยนี้สามารถหาค่าสนับสนุนจากสมการดังนี้

ค่าความเชื่อมั่น (วัน,เวลา และหมวดเว็บ → เว็บ) เท่ากับ

$$\frac{\text{จำนวนรายการข้อมูลที่มีวัน เวลา หมวดเว็บ และเว็บตามกฎ}}{\text{จำนวนรายการข้อมูลที่มีวัน เวลาและหมวดเว็บตามกฎ}} \quad \text{.....สมการที่ 4}$$

จากสมการผู้วิจัยพบว่าในการสร้างตัวแบบจะต้องคำนวณหาค่าพารามิเตอร์ดังต่อไปนี้

1. TOTAL = จำนวนรายการข้อมูลทั้งหมด
2. CNTCOND = จำนวนรายการข้อมูลที่มีวันเวลาและหมวดเว็บ เดียวกัน
3. CNTASSO = จำนวนรายการข้อมูลที่มีวัน เวลา หมวดเว็บและเว็บเดียวกัน
4. คำนวณหาค่าสนับสนุน (SUPPORT)= CNTASSO / TOTAL
5. คำนวณหาค่าความเชื่อมั่น (CONFIDENCE) = CNTASSO / CNTCOND

ตารางที่ 1 ตัวอย่างข้อมูลAccess log วันที่ 1/9/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.

ลำดับ	วัน	เวลา	หมวดเว็บ	เว็บ
1	1/9/2005	0:18:28	กีฬา	www.dserver.org
2	1/9/2005	0:03:13	ธนาคาร และสถานบันเทิง	www.bangkokbank.com
3	1/9/2005	0:54:53	บันเทิงและนันทนาการ	www.dailynews.co.th
4	1/9/2005	0:06:08	เกมส์	www.all-final.com
5	1/9/2005	0:39:52	อินเทอร์เน็ต	www.buildboard.com
6	1/9/2005	0:58:27	ความรู้และข้อมูลสำคัญ	www.academic.chula.ac.th
7	1/9/2005	0:00:09	หน่วยงานราชการและองค์กร	www.ch7.com
8	1/9/2005	0:58:09	การศึกษา	www.chula.ac.th
9	1/9/2005	0:18:25	บุคคล สังคม และวัฒนธรรม	board.dserver.org
10	1/9/2005	0:14:33	คอมพิวเตอร์	www.geocities.com
11	1/9/2005	0:58:05	คอมพิวเตอร์	www.bcoms.net
12	1/9/2005	0:17:57	ไม่ถูกจัดหมวดหมู่ไว้	community.ubcaf.com
13	1/9/2005	0:16:02	บุคคล สังคม และวัฒนธรรม	geocities.com
14	1/9/2005	0:03:16	การศึกษา	members.thai.net
15	1/9/2005	0:16:20	ไม่ถูกจัดหมวดหมู่ไว้	www.ubcaf.com
16	1/9/2005	0:08:08	หน่วยงานราชการและองค์กร	www.childthai.org
17	1/9/2005	0:22:07	ไม่ถูกจัดหมวดหมู่ไว้	diary.sanook.com
18	1/9/2005	0:13:39	ความรู้และข้อมูลสำคัญ	longdo.ex.nii.ac.jp
19	1/9/2005	0:54:53	ไม่ถูกจัดหมวดหมู่ไว้	db.dailynews.co.th
20	1/9/2005	0:49:09	ไม่ถูกจัดหมวดหมู่ไว้	kapook.com

จากตารางที่ 1 สามารถคำนวณหาค่าพารามิเตอร์ กฎความสัมพันธ์ ค่าสนับสนุนและค่าความเชื่อมั่นได้ดังนี้

1. คำนวณจำนวนรายการข้อมูลทั้งหมด (TOTAL) = 20
2. คำนวณจำนวนรายการข้อมูลที่มีวัน เวลาและหมวดเว็บเดียวกัน (CNTCOND) ดังตาราง

ตารางที่ 2 ผลการคำนวณจำนวนรายการข้อมูลของแต่ละวัน เวลาและหมวดเว็บจากข้อมูลตารางที่ 1

หมวดเว็บ	จำนวนรายการข้อมูล (CNTCOND)
กีฬา	2
ธนาคาร และสถานบันเทิง	5
บันเทิงและนันทนาการ	12
เกมส์	496
อินเทอร์เน็ต	136
ความรู้และข้อมูลสำคัญ	22
หน่วยงานราชการและองค์กร	115
การศึกษา	20
บุคคล สังคม และวัฒนธรรม	4
คอมพิวเตอร์	26
ไม่ถูกจัดหมวดหมู่ไว้	1,210

3. คำนวณจำนวนรายการข้อมูลที่มีวัน เวลา หมวดเว็บ และเว็บเดียวกัน (CNTASSO)
4. คำนวณค่าสนับสนุน (SUPPORT) และคำนวณค่าความเชื่อมั่น (CONFIDENCE) ได้ดังตาราง3

ตารางที่ 3 ผลการค้นหากฎความสัมพันธ์และคำนวณหาค่าพารามิเตอร์ที่มีวันเวลาเดียวกันคือ วันที่1/9/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.

กฎ	กฎ	CNTASSO	ค่าความเชื่อมั่น	ค่าสนับสนุน
1	กีฬา→ www.dserver.org	2	100%	0.10%
2	ธนาคาร และสถานบันเทิง→ www.bangkokbank.com	5	100%	0.24%
3	บันเทิงและนันทนาการ→ www.dailynews.co.th	12	100%	0.58%
4	เกมส์→ www.all-final.com	496	100%	24.18%
5	อินเทอร์เน็ต→ www.buildboard.com	139	100%	6.77%
6	ความรู้และข้อมูลสำคัญ→ www.academic.chula.ac.th	21	95.45%	1.02%
7	ความรู้และข้อมูลสำคัญ→ longdo.ex.nii.ac.jp	1	4.5%	0.05%
8	หน่วยงานราชการและองค์กร→ www.ch7.com	109	97.78%	5.31%
9	หน่วยงานราชการและองค์กร→ www.childthai.org	6	5.22%	0.29%
10	การศึกษา→ www.chula.ac.th	16	80%	0.78%
11	การศึกษา→ members.thai.net	4	20%	0.20%
12	บุคคล สังคม และวัฒนธรรม→ board.dserver.org	3	75%	0.14%
13	บุคคล สังคม และวัฒนธรรม → geocities.com	1	25%	0.05%
14	คอมพิวเตอร์→ www.geocities.com	16	61.53%	0.78%
15	คอมพิวเตอร์→ www.bcoms.net	10	38.46%	0.49%
16	ไม่ถูกจัดหมวดหมู่ไว้→ community.ubcaf.com	657	54.30%	32.03%
17	ไม่ถูกจัดหมวดหมู่ไว้→ www.ubcaf.com	250	20.66%	12.19%
18	ไม่ถูกจัดหมวดหมู่ไว้→ diary.sanook.com	115	9.50%	5.61%
19	ไม่ถูกจัดหมวดหมู่ไว้→ db.dailynews.co.th	94	7.76%	4.58%
20	ไม่ถูกจัดหมวดหมู่ไว้→ kapook.com	94	7.76%	4.58%

จากกฎความสัมพันธ์ที่แสดงในตารางที่ 3 สามารถอธิบายได้ดังนี้

กฎที่ 1 สามารถอธิบายได้ว่า วันที่ 01/09/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.ผู้ที่เรียกใช้เว็บในหมวดกีฬาทั้งหมด มีแนวโน้มที่จะเรียกใช้เว็บ www.dserver.org โดยที่คิดเป็น 0.10 % ของจำนวนรายการทั้งหมด

กฎที่ 2 สามารถอธิบายได้ว่า วันที่ 01/09/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.ผู้ที่เรียกใช้เว็บในหมวดธนาคารและสถานบันเทิงทั้งหมด มีแนวโน้มที่จะเรียกใช้เว็บ www.bangkokbank.com โดยที่คิดเป็น 0.24 % ของจำนวนรายการทั้งหมด

กฎที่ 3 สามารถอธิบายได้ว่า วันที่ 01/09/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.ผู้ที่เรียกใช้เว็บในหมวดบันเทิงและนันทนาการทั้งหมด มีแนวโน้มที่จะเรียกใช้เว็บ www.dailynews.co.th โดยที่คิดเป็น 0.58 % ของจำนวนรายการทั้งหมด

กฎที่ 4 สามารถอธิบายได้ว่า วันที่ 01/09/2005 ระหว่างเวลา 0.00.00 – 1.00.00 น.ผู้ที่เรียกใช้เว็บในหมวดเกมส์ทั้งหมด มีแนวโน้มที่จะเรียกใช้เว็บ www.all-final.com โดยที่คิด 24.18 % ของจำนวนรายการทั้งหมด

กฎที่ 5 สามารถอธิบายได้ว่า วันที่ 01/09/2005 เวลา12.00.00 AM ผู้ที่เรียกใช้เว็บในหมวดอินเทอร์เน็ตทั้งหมด มีแนวโน้มที่จะเรียกใช้เว็บ www.buildboard.com โดยที่คิด 6.77 % ของจำนวนรายการทั้งหมด

กฎที่ 6 สามารถอธิบายได้ว่าผู้ที่เรียกใช้เว็บหมวดความรู้และข้อมูลสำคัญ 95.45 % ที่เรียกใช้เว็บวันที่01/09/2005 ระหว่างเวลา 0.00.00–1.00.00น. มีแนวโน้มที่จะเรียกใช้เว็บ www.academic.chula.ac.th โดยที่คิดเป็น 1.02 % ของจำนวนรายการทั้งหมด

จากตารางที่ 3 ต้องนำตัวแบบที่ได้มาจำแนกกฎความสัมพันธ์โดยพิจารณาจากหลักเกณฑ์ดังนี้

1. พิจารณาจากค่าความเชื่อมั่นที่สูงสุดของแต่ละเงื่อนไข
2. ถ้าค่าความเชื่อมั่นเท่ากัน ให้พิจารณาค่าสนับสนุนที่สูงสุดของแต่ละเงื่อนไข
3. ถ้าค่าความเชื่อมั่นและค่าสนับสนุนมีค่าเท่ากันให้พิจารณากฎที่มาก่อนให้มีค่าความสำคัญมากกว่า

เมื่อพิจารณาตามหลักเกณฑ์จะได้ตัวแบบที่จำแนกกฎความสัมพันธ์แล้วดังตารางที่4

ตารางที่ 4 ผลการจำแนกกฎความสัมพันธ์จากข้อมูลตารางที่ 3

กฎ	กฎ	CNTASSO	ค่าความเชื่อมั่น	ค่าสนับสนุน
1	กีฬา→ www.dserver.org	2	100%	0.10%
2	ธนาคาร และสถานบันเทิง→ www.bangkokbank.com	5	100%	0.24%
3	บันเทิงและนันทนาการ→ www.dailynews.co.th	12	100%	0.58%
4	เกมส์→ www.all-final.com	496	100%	24.18%
5	อินเทอร์เน็ต→ www.buildboard.com	139	100%	6.77%
6	ความรู้และข้อมูลสำคัญ→ www.academic.chula.ac.th	21	95.45%	1.02%
8	หน่วยงานราชการและองค์กร→ www.ch7.com	109	97.78%	5.31%
10	การศึกษา→ www.chula.ac.th	16	80%	0.78%
12	บุคคล สังคม และวัฒนธรรม→ board.dserver.org	3	75%	0.14%
14	คอมพิวเตอร์→ www.geocities.com	16	61.53%	0.78%
16	ไม่ถูกจัดหมวดหมู่ไว้→ community.ubcaf.com	657	54.30%	32.03%

จากตารางที่ 4 เมื่อพิจารณาเกณฑ์การจำแนกกฎความสัมพันธ์แล้วจะเหลือกฎความสัมพันธ์อยู่ 11 กฎ คือกฎที่ 1-6, 8, 4 ,10,12 ,14 และ 16

เมื่อได้ศึกษาขั้นตอนการสร้างตัวแบบโดยการค้นหากฎความสัมพันธ์แล้ว จึงดำเนินการสร้างตัวแบบ และทดสอบตัวแบบในลำดับต่อไป ซึ่งในการทดสอบตัวแบบจะต้องแบ่งข้อมูลออกเป็น 2 ส่วนโดยใช้วิธีการเลือกตัวอย่างแบบมีระบบ

ศึกษาการเลือกตัวอย่างแบบมีระบบ

วีรานันท์ พงศาภักดี (2536 : 59-61) กล่าวว่า การเลือกตัวอย่างแบบมีระบบ เป็นเทคนิคที่ได้รับความนิยมกว้างขวางนอกจากจะใช้ได้กับวิธีการเลือกตัวอย่างด้วยความน่าจะเป็นแบบเท่ากันแล้วยังใช้ได้สะดวกกับวิธีการเลือกตัวอย่างด้วยความน่าจะเป็นที่เป็นสัดส่วนกับขนาด หรือความน่าจะเป็นแบบไม่เท่ากันอีกด้วย

การเลือกตัวอย่างแบบมีระบบจะแบ่งเป็น 2 วิธีคือ

- การเลือกตัวอย่างแบบระบบเส้นตรง

การเลือกตัวอย่างแบบระบบเส้นตรง มีวิธีการเลือกดังนี้

1. ให้หมายเลขแก่หน่วยประชากรจาก 1-N
2. หาช่วงการเลือกตัวอย่าง (Sampling interval) คือ I = N/n โดยที่ n คือ จำนวนตัวอย่างที่ต้องการเลือก
3. เลือกเลขสุ่ม (Random number) R ใดๆ ขึ้นมาโดยที่ 1< R < I

ดังนั้นการเลือกตัวอย่างข้อมูล จะได้ผลดังนี้

หน่วยหมายเลขที่ R	คือหน่วยตัวอย่างที่ 1
หน่วยหมายเลขที่ R+I	คือหน่วยตัวอย่างที่ 2
……..	……
……..	……
หน่วยหมายเลขที่ R + (n-1)I	คือหน่วยตัวอย่างที่ n

- การเลือกตัวอย่างแบบระบบวงกลม

การเลือกตัวอย่างแบบมีระบบวงกลม มีวิธีการเลือกตัวอย่าง เช่นเดียวกับการเลือกตัวอย่างแบบมีระบบเส้นตรง แต่เป็นวิธีที่เหมาะกับการคำนวณค่า I ซึ่งมีค่าไม่ลงตัวและใช้การปัดเศษให้เป็นตัวเลขจำนวนเต็ม และการเลือกเลขสุ่มจะใช้ R ที่อยู่ระหว่าง 1 – N (0 < R < N)

ตัวอย่าง การเลือกตัวอย่างขนาด 10 หน่วย จากประชากร 500 หน่วย

3. ให้หมายเลขแก่หน่วยประชากรจาก 1 – 500

3. หาช่วงการเลือกตัวอย่าง I = N/n = 500/10 = 50

3. เลือกเลขสุ่ม R = 100 โดยที่ 0 < R < 500

ดังนั้นการเลือกตัวอย่างข้อมูล จะได้ผลดังนี้

หน่วยตัวอย่างที่ 1	คือหน่วยหมายเลขที่ 100
หน่วยตัวอย่างที่ 2	คือหน่วยหมายเลขที่ 150
……	……..
……	……..
หน่วยตัวอย่างที่ 9	คือหน่วยหมายเลขที่ 500
หน่วยตัวอย่างที่ 10	คือหน่วยหมายเลขที่ 50 (วนกลับไปเริ่มต้น)

ในงานวิจัยนี้เลือกวิธีการเลือกตัวอย่างแบบมีระบบวงกลม เพราะการสุ่มเลขเริ่มต้นมีช่วงที่กว้างกว่าวิธีการเลือกตัวอย่างแบบมีระบบเส้นตรง ทำให้หน่วยตัวอย่างมีโอกาสที่จะถูกเลือกเท่าๆ กัน

3. ทดสอบตัวแบบเพื่อใช้ในการทำนายเนื้อหาเว็บโดยใช้เทคนิคเหมืองข้อมูล

ในการสร้างตัวแบบผู้วิจัยจะแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลการเรียนรู้ และข้อมูลตรวจสอบซึ่งการแบ่งข้อมูลจะใช้การเลือกตัวอย่างแบบมีระบบบวงกลม

โดยขั้นตอนการเลือกตัวอย่างแบบวงกลมจากนั้นจะทำการแบ่งข้อมูลการเรียกใช้งานเว็บที่ต้องการนำมาทดสอบตัวแบบ ออกเป็นข้อมูลการเรียนรู้ และข้อมูลตรวจสอบ ตามสัดส่วนที่กำหนดโดยใช้วิธีการเลือกตัวอย่างแบบมีระบบบวงกลม

ขั้นตอนก่อนเลือกตัวอย่างแบบวงกลมนั้นจะต้องพิจารณาว่าต้องการนำเว็บที่เป็น Search Engine มาพิจารณาในการทดสอบตัวแบบหรือไม่ ซึ่งเว็บ Search Engine คือ เว็บที่ให้บริการค้นหาข้อมูลเว็บไซต์ ที่ผู้ใช้งานอินเทอร์เน็ตเรียกใช้เพื่อค้นหาเว็บที่มีข้อมูลที่ต้องการ เมื่อนำเว็บที่เป็นSearch Engine มาพิจารณาจะให้ผลการคำนวณที่แตกต่างกัน

สัดส่วนที่ใช้ในการแบ่งข้อมูล เป็นสัดส่วนที่ต้องใช้แบ่งข้อมูลของแต่ละกลุ่มชั้นข้อมูล และการเลือกตัวอย่างแบบมีระบบบวงกลมก็จะเลือกตัวอย่างของแต่ละชั้นข้อมูลด้วย โดยมีขั้นตอนดังนี้

3.1 กำหนดรหัสลำดับที่ให้กับรายการข้อมูล

กำหนดรหัสลำดับให้กับข้อมูลโดยแบ่งเป็นกลุ่มตามหมวดเว็บและ URL

3.2 คำนวณหาจำนวนรายการข้อมูลทั้งหมดของแต่ละกลุ่ม และคำนวณรายการข้อมูลที่ต้องการตามสัดส่วน

โดยที่

รายการข้อมูลตามสัดส่วน = (จำนวนรายการข้อมูลของแต่ละกลุ่ม * สัดส่วน) / 100

ถ้าตัวเลขที่ได้เป็นทศนิยมให้ปัดเป็นตัวเลขจำนวนเต็ม เช่น 1.66 เมื่อทำการปัดเศษขึ้นเป็นตัวเลขเต็มจะมีค่าเท่ากับ 2

3.3 หาช่วงการเลือกตัวอย่าง

3.4 เลือกเลขสุ่มโดยการสุ่มตัวเลขเริ่มต้น R ของข้อมูลแต่ละกลุ่ม ซึ่งค่า R จะอยู่ในช่วงตั้งแต่ 1 ถึง N

ดังนั้นผลของการเลือกตัวอย่างจะได้รายการข้อมูลการเรียนรู้ (Train) ที่เลือกได้ในแต่ละกลุ่มจะได้ผลดังตารางและสำหรับข้อมูลตรวจสอบ (Validation) คือรายการข้อมูลที่ไม่ได้ถูกเลือก ซึ่งจะต้องทำการค้นหากฎความสัมพันธ์จากข้อมูลการเรียนรู้และข้อมูลตรวจสอบที่ได้เลือกไว้

4. การทดสอบความถูกต้องตัวแบบ

นำข้อมูลตรวจสอบที่เลือกได้มาตรวจสอบความถูกต้องของตัวแบบ

ตารางที่ 5 ตัวอย่างผลการทดสอบความถูกต้องตัวแบบ วันที่ 1 กันยายน 2005 ระหว่างเวลา 0.00.00 น.– 1.00.00 น.

กฎที่	Rule	%ค่าความเชื่อมั่นข้อมูลเรียนรู้	%ค่าความเชื่อมั่นข้อมูลตรวจสอบ	ความถูกต้อง
1	การศึกษา→ www.chula.ac.th	80.00	100.00	T
2	กีฬา→ www.dserver.org	100.00	100.00	T
3	ความรู้และข้อมูลสำคัญ→ www.academic.chula.ac.th	95.45	100.00	T
4	คอมพิวเตอร์→ www.bcoms.net	38.46	45.45	T
5	คอมพิวเตอร์→www.geocities.com	61.54	18.18	F
6	ธนาคาร และสถานบันเทิง→ www.bangkokbank.com	100.00	100.00	T
7	บันเทิงและนันทนาการ→ www.dailynews.co.th	100.00	80.00	F
8	บุคคล สังคม และวัฒนธรรม→ board.dserver.org	75.00	100.00	T
9	หน่วยงานราชการและองค์กร→ www.ch7.com	93.97	100.00	T
10	อินเทอร์เน็ต→ www.buildboard.com	95.86	100.00	T
11	เกมส์→www.all-final.com	100.00	100.00	T
12	ไม่ถูกจัดหมวดหมู่ไว้→ us.i1.yimg.com	1.75	0.38	F
13	ไม่ถูกจัดหมวดหมู่ไว้→ www.trekkerhut.com	0.20	6.46	T

จากตารางที่ 5 สามารถอธิบายได้ว่า

ลำดับที่ 1 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้เรียกใช้เว็บ 80 % ที่เรียกใช้หมวดเว็บการศึกษา ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.chula.ac.th เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า ผู้เรียกใช้เว็บทั้งหมดที่เรียกใช้หมวดเว็บการศึกษา ในวันที่ 01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.chula.ac.th ดังนั้นลำดับที่ 1 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 2 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้เรียกใช้เว็บทั้งหมด ที่เรียกใช้หมวดเว็บกีฬา ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.dserver.org เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า ผู้เรียกใช้เว็บทั้งหมดที่เรียกใช้หมวดเว็บกีฬา ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.dserver.org ดังนั้นลำดับที่ 2 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 3 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้เรียกใช้เว็บ 95.45% ที่เรียกใช้หมวดเว็บความรู้ความสำคัญ ในวันที่01/09/2005 เวลา 0.00.00น. –1.00.00น. จะเรียกใช้เว็บ www.academic.chula.ac.th เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า ผู้เรียกใช้เว็บทั้งหมดที่เรียกใช้หมวดเว็บความรู้ความสำคัญ ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.academic.chula.ac.th ดังนั้นลำดับที่ 3 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 4 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้เรียกใช้เว็บ 38.46% ที่เรียกใช้หมวดเว็บคอมพิวเตอร์ ในวันที่01/09/2005 เวลา0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.bcoms.net เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า ผู้เรียกใช้เว็บ 45.45% ที่เรียกใช้หมวดเว็บคอมพิวเตอร์ ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.bcoms.net ดังนั้นลำดับที่ 4 จึงมีความถูกต้องเท่ากับ T

ลำดับที่ 5 สามารถอธิบายได้ว่าจากตัวแบบเรียนรู้มีผู้เรียกใช้เว็บ 61.54 % ที่เรียกใช้หมวดเว็บคอมพิวเตอร์ ในวันที่01/09/2005 เวลา0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.geocities.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า ผู้เรียกใช้เว็บ 18.18% ที่เรียกใช้หมวดเว็บคอมพิวเตอร์ ในวันที่01/09/2005 เวลา 0.00.00 น. – 1.00.00 น. จะเรียกใช้เว็บ www.geocities.com ดังนั้นลำดับที่ 5 จึงมีความถูกต้องเท่ากับ F เป็นต้น

ตารางที่ 6 แสดงร้อยละความถูกต้องของการทดสอบตัวแบบ

การสุ่มครั้งที่	**จำนวนตัวอย่างสุ่ม**	**ร้อยละความถูกต้อง**
1	32	78.13
2	34	67.65
3	33	70.16
ค่าเฉลี่ยรวม		**72.13**

จากตารางที่ 6 เมื่อนำความถูกต้องของตัวแบบข้อมูลการเรียนรู้จากการสุ่มตัวอย่างครั้ง 3 ครั้ง มาคำนวณร้อยละของความถูกต้องพบว่ามีความถูกต้องโดยเฉลี่ยคิดเป็นร้อยละ 72.13 ของจำนวนรายการทั้งหมด โดยร้อยละของความถูกต้องครั้งที่ 1 เท่ากับ 78.13 , ครั้งที่ 2 เท่ากับ 67.65 และครั้งที่ 3 เท่ากับ 70.16 ดังนั้นสามารถนำตัวแบบที่ได้ศึกษานี้ไปใช้ในการทำนายเนื้อหาเว็บได้ต่อไป

5. การสร้างตัวแบบ

ผู้วิจัยสร้างตัวแบบขึ้นเพื่อทำนายแนวโน้มการเรียกใช้เว็บโดยใช้วิธีและเทคนิคที่ได้นำเสนอไปแล้ว แต่ข้อมูลที่นำมาใช้ในการสร้างตัวแบบไม่ต้องแบ่งชุดข้อมูล และมีการรับค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด เพื่อเลือกเฉพาะกฎค้นหาความสัมพันธ์ที่ค่าความเชื่อมั่นต่ำสุด และค่าสนับสนุนที่เราสนใจเท่านั้น โดยผู้ใช้ระบบสามารถเลือกได้ว่าต้องการนำเว็บที่เป็น Search Engine มาพิจารณาหรือไม่ ซึ่งตัวแบบที่ได้จากการคำนวณนั้นจะเป็นตัวแบบที่ระบบพร็อกซี่ เซอร์ฟเวอร์ควรที่จะจัดเก็บไว้เพื่อรอให้ผู้ใช้บริการเรียกใช้ในวันถัดไป

ตารางแสดงความสัมพันธ์
จำนวนกฎความสัมพันธ์ = 9
จำนวนหมวดเว็บ = 9

กฎที่	Date	Times	Rule	ค่าความเชื่อมั่น(%)	ค่าสนับสนุน(%)
1	01/09/2005	12:49:29 AM	การศึกษา==> www.obec.go.th	81.53	2.15
2	01/09/2005	12:14:50 AM	กีฬา==> www.pixiart.com	94.12	0.70
3	01/09/2005	12:54:40 AM	ข่าว และสื่อ==> www.manager.co.th	99.72	4.91
4	01/09/2005	12:11:01 AM	ความรู้และข้อมูลสำคัญ==> www.kapook.com	92.64	6.32
5	01/09/2005	12:33:20 AM	ช้อปปิ้ง==> www.marketathome.com	100.00	0.01
6	01/09/2005	12:47:01 AM	ท่องเที่ยว==> www.trekkingthai.com	87.21	1.18
7	01/09/2005	12:38:09 AM	ธนาคาร และสถานบันเทิง==> www.pantip.com	83.22	1.96
8	01/09/2005	12:46:29 AM	ธุรกิจ==> www.tarad.com	94.23	0.71
9	01/09/2005	12:06:08 AM	เกมส์==> www.all-final.com	100.00	9.69

รูปที่ 12 ตัวอย่างตัวแบบที่สร้างได้ วันที่01/09/2005 ระหว่างเวลา 0.00.00 น. – 1.00.00 น.

6. การตรวจสอบผลการทำนายกับข้อมูลจริง

การตรวจสอบผลการทำนายกับข้อมูลจริงคือการนำตัวแบบที่สร้างขึ้น มาตรวจสอบความถูกต้องโดยวิธีการตรวจสอบดังนี้

1. ระบุวัน เวลา ข้อมูลจริงที่นำมาตรวจสอบ
2. ทำการเปรียบเทียบเว็บจากตัวแบบที่สร้างขึ้นกับข้อมูลที่นำมาตรวจสอบ
3. วัดร้อยละของความถูกต้องที่พบเว็บที่ได้จากการสร้างตัวแบบในข้อมูลของวันถัดมาที่นำมาตรวจสอบ

โดยจากผลการตรวจสอบผลการทำนายกับข้อมูลจริงวันที่01/09/2005 ระหว่างเวลา 0.00.00 น. – 1.00.00 น. พบว่าตัวแบบมีความถูกต้องร้อยละ 66.67 %

7. การจำแนกกฎความสัมพันธ์

การจำแนกกฎความสัมพันธ์คือการเลือกกฎความสัมพันธ์เพียงกฎเดียวในแต่ละเงื่อนไข ซึ่งเงื่อนไขคือวัน เวลา และหมวดเว็บ โดยพิจารณาจากหลักเกณฑ์ดังนี้

4. พิจารณาจากค่าความเชื่อมั่นที่สูงสุดของแต่ละเงื่อนไข
5. ถ้าค่าความเชื่อมั่นเท่ากัน ให้พิจารณาค่าสนับสนุนที่สูงสุดของแต่ละเงื่อนไข
6. ถ้าค่าความเชื่อมั่นและค่าสนับสนุนมีค่าเท่ากันให้พิจารณากฎที่มาก่อนให้มีค่าความสำคัญมากกว่า

ถ้าพิจารณาตามหลักเกณฑ์การจำแนกกฎความสัมพันธ์จะได้ผลลัพธ์ดังรูปที่ 14

แสดงการจำแนกกฎความสัมพันธ์
จำนวนกฎความสัมพันธ์ = 5
จำนวนหมวดเว็บ = 5

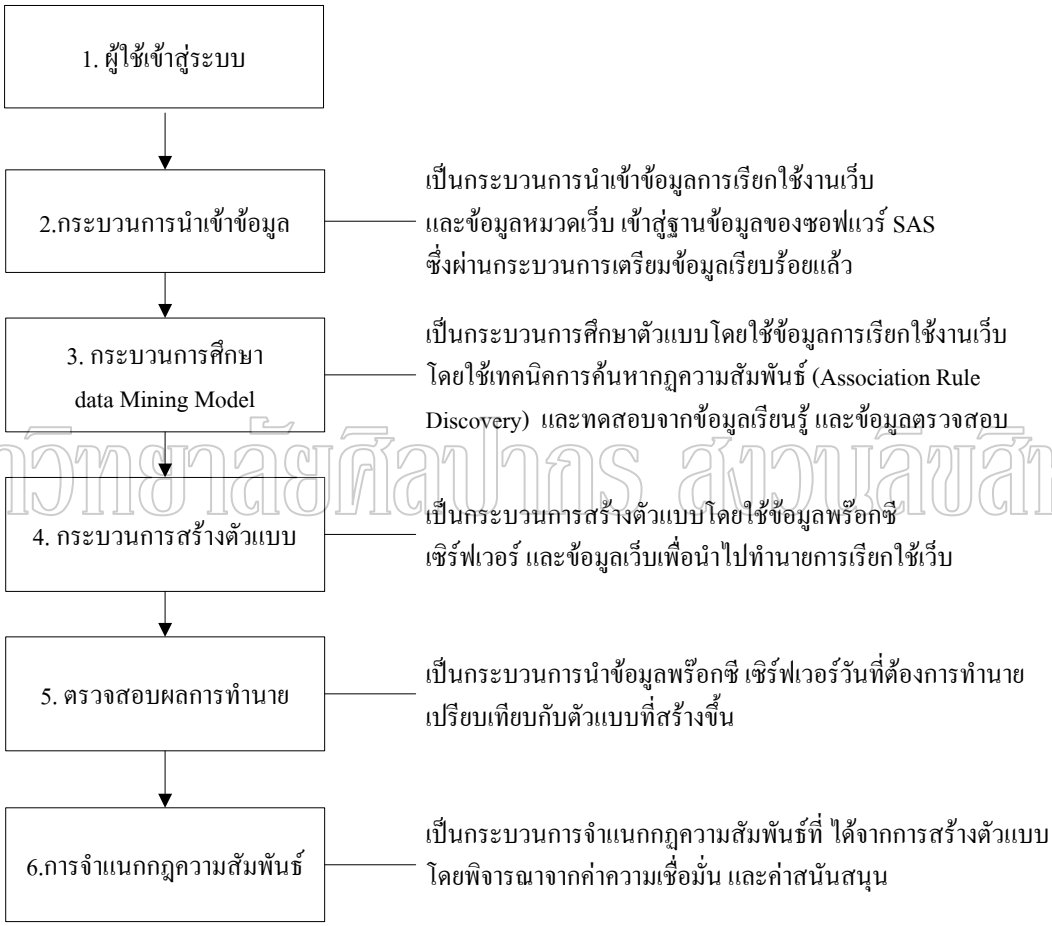
ruleno	กฎ	ค่าความเชื่อมั่น (%)	ค่าสนับสนุน (%)
1	ข่าว และสื่อ==> www.bangkokbiznews.com	100.00	0.18
2	ท่องเที่ยว==> travel.sanook.com	100.00	0.18
3	ธนาคาร และสถานบันเทิง==> www.bangkokbank.com	100.00	0.18
4	ยานยนต์==> www.bkknight.com	100.00	0.18
5	อสังหาริมทรัพย์==> www.easyhorpak.com	100.00	0.18

รูปที่ 13 ผลการจำแนกกฎความสัมพันธ์

การดำเนินงานวิจัยดังที่ได้กล่าวมาสามารถเรียกใช้โปรแกรมที่ผู้วิจัยได้พัฒนาขึ้นมา ซึ่งต่อไปนี้จะอธิบายรายละเอียดขั้นตอนการพัฒนาโปรแกรม ดังนี้

1. ขั้นตอนการทำงานระบบ

จากการที่ได้กล่าวมาแล้วนั้นสามารถแสดงให้เห็นภาพรวมของระบบโดยประกอบด้วยกระบวนการหลักๆ ดังนี้



รูปที่ 14 ขั้นตอนการทำงานของระบบ

2. โครงสร้างข้อมูล

รายละเอียดโครงสร้างข้อมูล

ตารางที่ 7 โครงสร้างตาราง addnew

ชื่อตาราง : addnew				
รายละเอียดตาราง : เก็บข้อมูลการเรียกใช้งานเว็บจากระบบพร๊อกซีที่ผ่านกระบวนการแปลงวัน และเวลาเรียบร้อยแล้ว				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	newResultcode	ผลการทำงานของการร้องแต่ละครั้ง	ตัวอักษร	50

ตารางที่ 8 โครงสร้างตาราง Web

ชื่อตาราง : Web				
รายละเอียดตาราง : เก็บข้อมูลเว็บและรายละเอียดของเว็บ				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Title	ชื่อเว็บ	ตัวอักษร	255
2	Description	คำอธิบายเว็บ	ตัวอักษร	255
3	SubGroup_ID	รหัสหมวดเว็บย่อย	ตัวเลข	3
4	SubGroup_Name	ชื่อหมวดเว็บย่อย	ตัวอักษร	255
5	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
6	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
7	newURL	ชื่อเรียกที่อยู่ของเว็บไซต์	ตัวอักษร	255

ตารางที่ 9 โครงสร้างตาราง MergeComplete

ชื่อตาราง : MergeComplete				
รายละเอียดตาราง : นำตาราง Import และตาราง Web มารวมกันด้วยข้อมูล newURL				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	SubGroup_ID	รหัสหมวดเว็บย่อย	ตัวเลข	3
7	SubGroup_Name	ชื่อหมวดเว็บย่อย	ตัวอักษร	255
8	CNTTOTAL	จำนวนรายการข้อมูลทั้งหมด	ตัวเลข	8
9	CNTCOND	จำนวนรายการข้อมูลของแต่ละหมวดเว็บ	ตัวเลข	8
10	CNTASSO	จำนวนรายการข้อมูลที่มีความสัมพันธ์กัน	ตัวเลข	8
11	CONFIDENCE	ค่าความเชื่อมั่น	ตัวเลข	8
12	SUPPORT	ค่าสนับสนุน	ตัวเลข	8
13	RULE	แสดงกฎ A→ B (หมวดเว็บ → ชื่อเว็บ)	ตัวอักษร	255

ตารางที่ 10 โครงสร้างตาราง TRAINDATA

ชื่อตาราง : TRAINDATA				
รายละเอียดตาราง : ข้อมูลเรียนรู้ที่ได้จากการสุ่มแบบวงกลม				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	SubGroup_ID	รหัสหมวดเว็บย่อย	ตัวเลข	3
7	SubGroup_Name	ชื่อหมวดเว็บย่อย	ตัวอักษร	255
8	CNTTOTAL	จำนวนรายการข้อมูลทั้งหมด	ตัวเลข	8
9	CNTCOND	จำนวนรายการข้อมูลของแต่ละหมวดเว็บ	ตัวเลข	8
10	CNTASSO	จำนวนรายการข้อมูลที่มีความสัมพันธ์กัน	ตัวเลข	8
11	CONFIDENCE	ค่าความเชื่อมั่น	ตัวเลข	8
12	SUPPORT	ค่าสนับสนุน	ตัวเลข	8
13	RULE	แสดงกฎ A→ B (สาขาวิชา,เพศ → อาชีพ)	ตัวอักษร	255

ตารางที่ 11 โครงสร้างตาราง VALIDDATA

ชื่อตาราง : VALIDDATA				
รายละเอียดตาราง :ข้อมูลตรวจสอบที่ได้จากการสุ่มตัวอย่างแบบวงกลม				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	SubGroup_ID	รหัสหมวดเว็บย่อย	ตัวเลข	3
7	SubGroup_Name	ชื่อหมวดเว็บย่อย	ตัวอักษร	255
8	CNTTOTAL	จำนวนรายการข้อมูลทั้งหมด	ตัวเลข	8
9	CNTCOND	จำนวนรายการข้อมูลของแต่ละหมวดเว็บ	ตัวเลข	8
10	CNTASSO	จำนวนรายการข้อมูลที่มีความสัมพันธ์กัน	ตัวเลข	8
11	CONFIDENCE	ค่าความเชื่อมั่น	ตัวเลข	8
12	SUPPORT	ค่าสนับสนุน	ตัวเลข	8

ตารางที่ 12 โครงสร้างตาราง ASSESS_MODEL

ชื่อตาราง : ASSESS_MODEL				
รายละเอียดตาราง : ข้อมูลที่ได้จากการศึกษาตัวแบบ				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	RULE	แสดงกฎ A→ B (หมวดเว็บ → เว็บ)	ตัวอักษร	255
7	RULENO	ลำดับกฎ	ตัวอักษร	5
8	CONF_TRAIN	ค่าความเชื่อมั่นตัวแบบเรียนรู้	ตัวเลข	8
9	CONF_VALID	ค่าความเชื่อมั่นตัวแบบทดสอบ	ตัวเลข	8
10	ACCURACY	ความถูกต้อง T = ถูกต้อง F = ไม่ถูกต้อง	ตัวอักษร	1

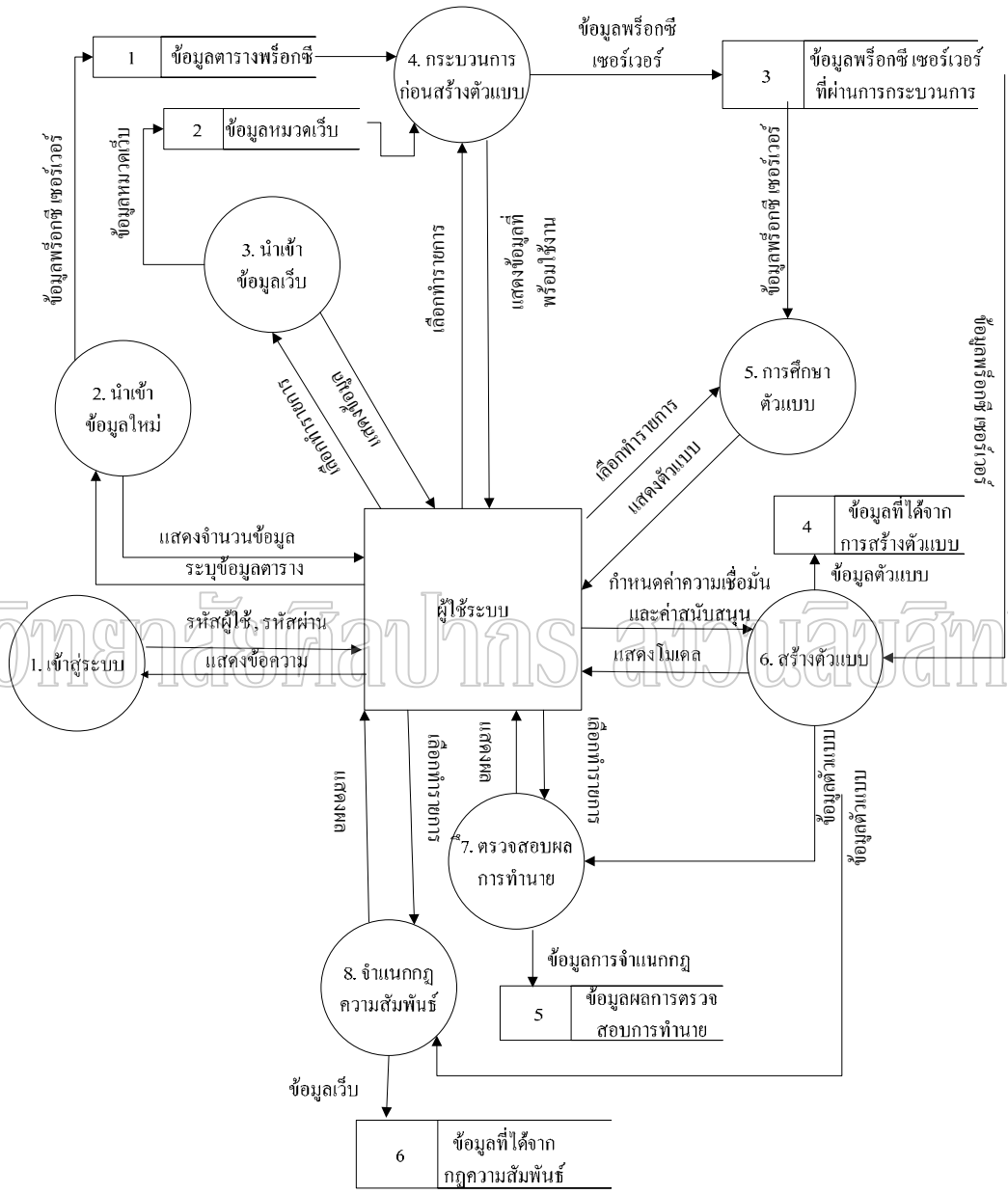
ตารางที่ 13 โครงสร้างตาราง TestModel

ชื่อตาราง : TestModel				
รายละเอียดตาราง : ข้อมูลที่ได้จากการตรวจสอบผลการทำนาย				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	RULE	แสดงกฎ A→ B (หมวดเว็บ → เว็บ)	ตัวอักษร	255
7	RULENO	ลำดับกฎ	ตัวอักษร	5
8	ACCURACY	ความถูกต้อง T = ถูกต้อง F = ไม่ถูกต้อง	ตัวอักษร	1

ตารางที่ 14 โครงสร้างตาราง Model_Classify

ชื่อตาราง : Model_Classify				
รายละเอียดตาราง : ข้อมูลที่ได้จากการจำแนกกฎความสัมพันธ์ของโมเดล				
ลำดับที่	ชื่อรายการข้อมูล	คำอธิบาย	ประเภท	ขนาด
1	Date	วัน เดือน ปี	ddmmyy10	10
2	Times	เวลา	timeampm11	11
3	newURL	ชื่อเว็บ	ตัวอักษร	255
4	Group_ID	รหัสหมวดเว็บ	ตัวเลข	3
5	Group_NAME	ชื่อหมวดเว็บ	ตัวอักษร	255
6	RULE	แสดงกฎ A→ B (หมวดเว็บ → เว็บ)	ตัวอักษร	255
7	RULENO	ลำดับกฎ	ตัวอักษร	5
8	CONF_TRAIN	ค่าความเชื่อมั่นตัวแบบเรียนรู้	ตัวเลข	8
9	CONF_VALID	ค่าความเชื่อมั่นตัวแบบทดสอบ	ตัวเลข	8

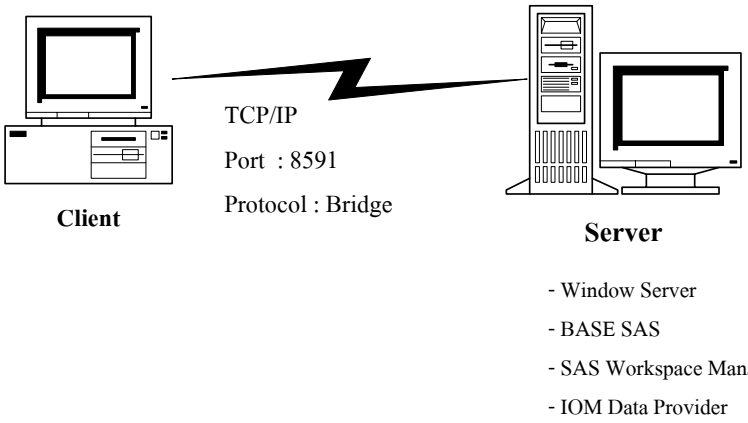
3. แผนภาพกระแสข้อมูล (Data Flow Diagram)



รูปที่ 15 แผนภาพกระแสข้อมูล

4. การออกแบบระบบ

4.1 สถาปัตยกรรมระบบ



รูปที่ 16 สถาปัตยกรรมระบบ

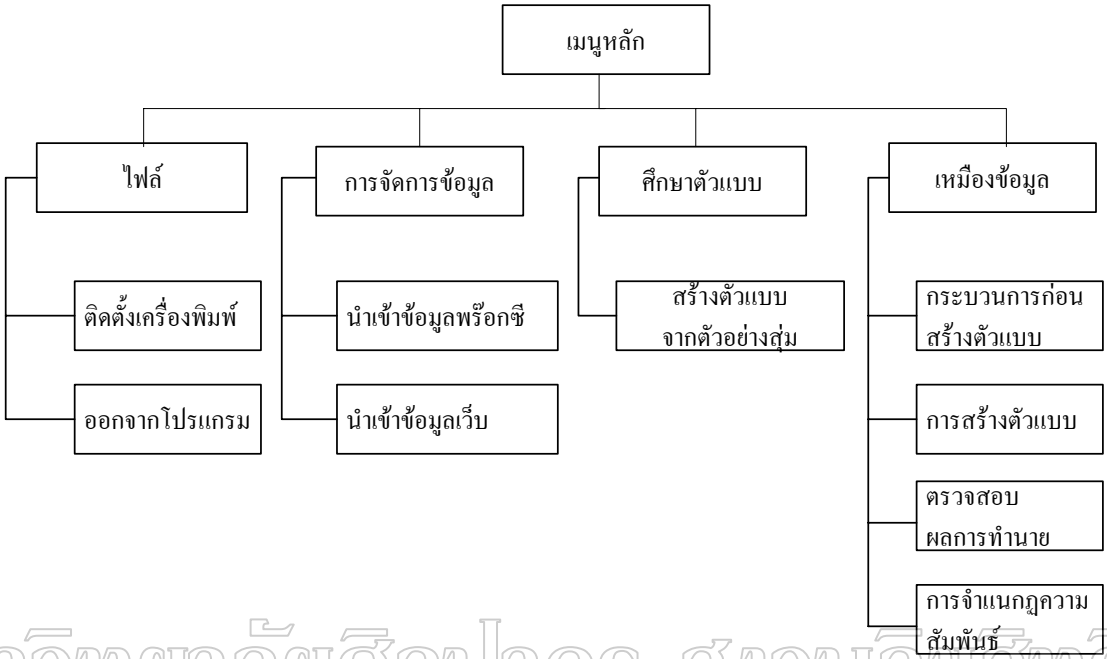
จากรูปสามารถอธิบายได้ว่า สถาปัตยกรรมระบบจะเป็นลักษณะ Client/Server ซึ่ง Server จะเป็นระบบปฏิบัติ Windows 2000 ในการติดต่อเครื่อง Server ต้องใช้ Port 8591 และโปรโตคอล Bridge หลังจากนั้นจะเข้าถึงฐานข้อมูล SAS โดยใช้ IOM Data Provider ในการพัฒนาโปรแกรม ด้วย Microsoft Visual Basic นั้นผู้วิจัยใช้ Microsoft ActiveX Data Objects (ADO) ในการอ่านเขียนข้อมูลและใช้ SAS Workspace Manager ในการสร้าง SAS Workspaces

ตัวอย่างการสร้าง Workspace และการติดต่อฐานข้อมูลด้วย ADO

```
Dim obConnection As New ADODB.Connection
obServer.Port = 8591
obServer.Protocol = “Bridge”
obConnection.Provider = "SAS.IOMProvider.1"
Set obWs = obWSMgr.Workspaces.CreateWorkspaceByServer("Thesis_WS",
VisibilityProcess, obServer, userLogin, UserPWD, errString)
 obConnection.Open "SAS Workspace ID = " & obWs.UniqueIdentifier
Dim obRs as New ADODB.Recordset
obRs.Open “sashelp.class”, obConnection, adOpenForwardOnly, adLockReadOnly, _
adcmdTableDirect
```

การออกแบบหน้าจอ (User Interface Design)

หน้าจอการใช้งานจะประกอบด้วยเมนูการทำงาน และแถบเครื่องมือดังนี้



รูปที่ 17 เมนูการใช้งานโปรแกรม

ในส่วนของการใช้งานโปรแกรมซึ่งจะกล่าวโดยละเอียดในภาคผนวก ก

5. การพัฒนาระบบ

การพัฒนาโปรแกรม (Application Development)

โปรแกรมที่พัฒนาขึ้นเป็นการพัฒนาด้วยภาษา Microsoft Visual Basic โปรแกรมที่พัฒนาขึ้นจะประกอบด้วยโมดูลต่างๆ ซึ่งเป็นที่เก็บโปรแกรม แต่ละโมดูลจะประกอบด้วย Event Procedure เป็นส่วนที่เก็บโปรแกรมที่ทำงานเมื่อมี Event ต่างๆเป็น Form และ Procedure ที่ไม่เกี่ยวข้องกับ Event ของ Object และนอกจากนั้นยังมีการเรียกใช้สคริปต์ไฟล์ที่เขียนด้วยภาษา SAS เพื่อให้การจัดการข้อมูลทำได้ง่ายและสามารถแก้ไขได้ง่ายขึ้น

- โมดูลการเข้าสู่ระบบ

โมดูลการเข้าสู่ระบบเป็นการตรวจสอบผู้มีสิทธิเข้าใช้ระบบโดยรหัสผู้ใช้และรหัสผ่าน ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 15โมดูลการเข้าสู่ระบบ

ชื่อฟอร์ม : FrmConnect		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdIOMConn_Click	open_baseSAS	
	Assign_Library	

- โมดูลการนำเข้าข้อมูล

โมดูลการเข้านำเข้าข้อมูลเป็นการนำข้อมูลที่อยู่ในระบบพร๊อกซี เซิร์ฟเวอร์ เข้าสู่ฐานข้อมูล SAS ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 16 โมดูลการนำเข้าข้อมูลใหม่

ชื่อฟอร์ม : FrmImnew		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdBrowse_Click		
cmdImport_Click		import
		cleantime

- โมดูลการนำเข้าข้อมูลเว็บ

โมดูลการเข้านำเข้าข้อมูลเป็นการนำข้อมูลเว็บที่อยู่ใน Microsoft Access เข้าสู่ฐานข้อมูล SAS ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 17 โมดูลการนำเข้าข้อมูลเว็บ

ชื่อฟอร์ม : FrmImweb		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdBrowse_Click		
cmdImpweb_Click		web

- โมดูลกระบวนการก่อนสร้างตัวแบบ/ศึกษาตัวแบบ

โมดูลการกระบวนการก่อนสร้างตัวแบบเป็นการนำข้อมูลพร๊อกซี เซิร์ฟเวอร์ มาดำเนินการเลือกวัน เลือกเวลา และสร้างความสัมพันธ์กับข้อมูลเว็บ ก่อนนำไปสร้างตัวแบบ ดังตาราง

ตารางที่ 18โมดูลกระบวนการก่อนสร้างตัวแบบ/ ศึกษาตัวแบบ

ชื่อฟอร์ม : FrmImport		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdSelectdate_Click		Selectdate
cmdSelecttime_Click		Selecttime
Cmdrelate_click		mergedata

- โมดูลการศึกษาตัวแบบ

โมดูลการศึกษาตัวแบบเป็นการนำข้อมูลการพร๊อกซี เซิร์ฟเวอร์มาแบ่งออกเป็น 2 ส่วนสำหรับเป็นข้อมูลเรียนรู้ และข้อมูลตรวจสอบ หลังจากนั้นก็นำข้อมูลเรียนรู้มาสร้างตัวแบบโดยการค้นหากฎความสัมพันธ์ของวัน เวลา หมวดเว็บ และเว็บ แล้วนำข้อมูลตรวจสอบมาทดสอบตัวแบบที่สร้างขึ้น ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 19โมดูลการศึกษาตัวแบบ

ชื่อฟอร์ม : FrmTrainMode		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdSampling_Click		sampledata
cmdCreateModel_Click		model_nosearch
		model_cnttotal
		model_condA
		model_condAB
		model_conf_supp
		model_sym
		model_html
cmdCorrect_Click		model_cnttotal

ตารางที่ 19โมดูลการศึกษาตัวแบบ (ต่อ)

ชื่อฟอร์ม : FrmTrainMode		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
		model_condA
		model_condAB
		model_conf_supp
		assess_model
		assess_html

- โมดูลการสร้างตัวแบบ

โมดูลการสร้างตัวแบบเป็นการนำข้อมูลพร๊อกซี เซิร์ฟเวอร์ที่ได้จากกระบวนการก่อนสร้างตัวแบบทั้งหมดมาค้นหากฎความสัมพันธ์ของวัน เวลา หมวดเว็บ และเว็บ โดยมีการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 20 โมดูลการสร้างตัวแบบ

ชื่อฟอร์ม : FrmModel		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdCreateModel		model_nosearch
		model_cnttotal
		model_condA
		model_condAb
		model_conf_supp
		model_sym
		model_html
cmdQuery_Click		query_model

- โมดูลการตรวจสอบตัวแบบกับข้อมูลจริง

โมดูลการตรวจสอบตัวแบบกับข้อมูลจริง เป็นการนำข้อมูลที่ได้จากการสร้างตัวแบบมาเปรียบเทียบกับข้อมูลที่มีการถูกเรียกใช้จริง ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

ตารางที่ 21 โมดูลการตรวจสอบตัวแบบกับข้อมูลจริง

ชื่อฟอร์ม : frmTestmodel		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
cmdSelectdate1_Click		DateTestModel
cmdSelecttime_Click		TimeTestModel
CmdrelateAuto_click		MergeTestModel
cmdchemodel_Click		testmodel
		percentcorrect

- โมดูลการจำแนกกฎความสัมพันธ์

โมดูลการจำแนกกฎความสัมพันธ์ เป็นการเลือกกฎความสัมพันธ์โดยใช้เกณฑ์พิจารณาที่ได้กล่าวมาแล้ว ซึ่งประกอบด้วยส่วนต่างๆ ดังตาราง

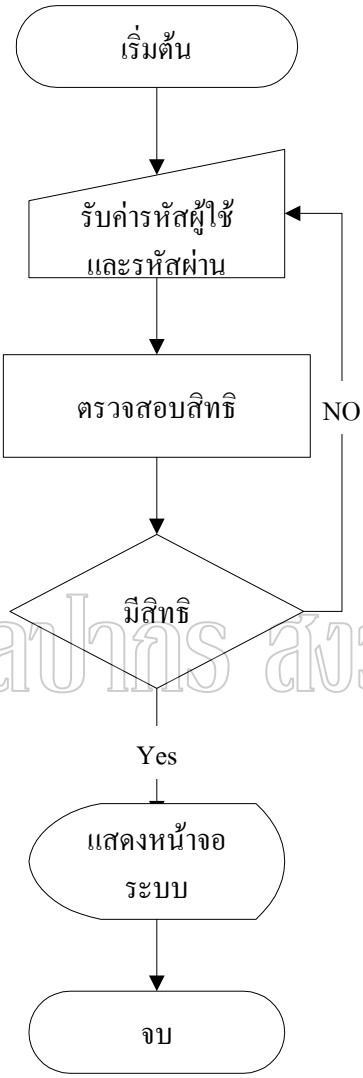
ตารางที่ 22 โมดูลการจำแนกกฎความสัมพันธ์

ชื่อฟอร์ม : MdiMenu		
Event Procedure	Sub / Function	สคริปต์ไฟล์(.SAS)
mnuClassify_Click	classify_process	model_classify

ผังงาน (Flowchart) ของแต่ละกระบวนการ

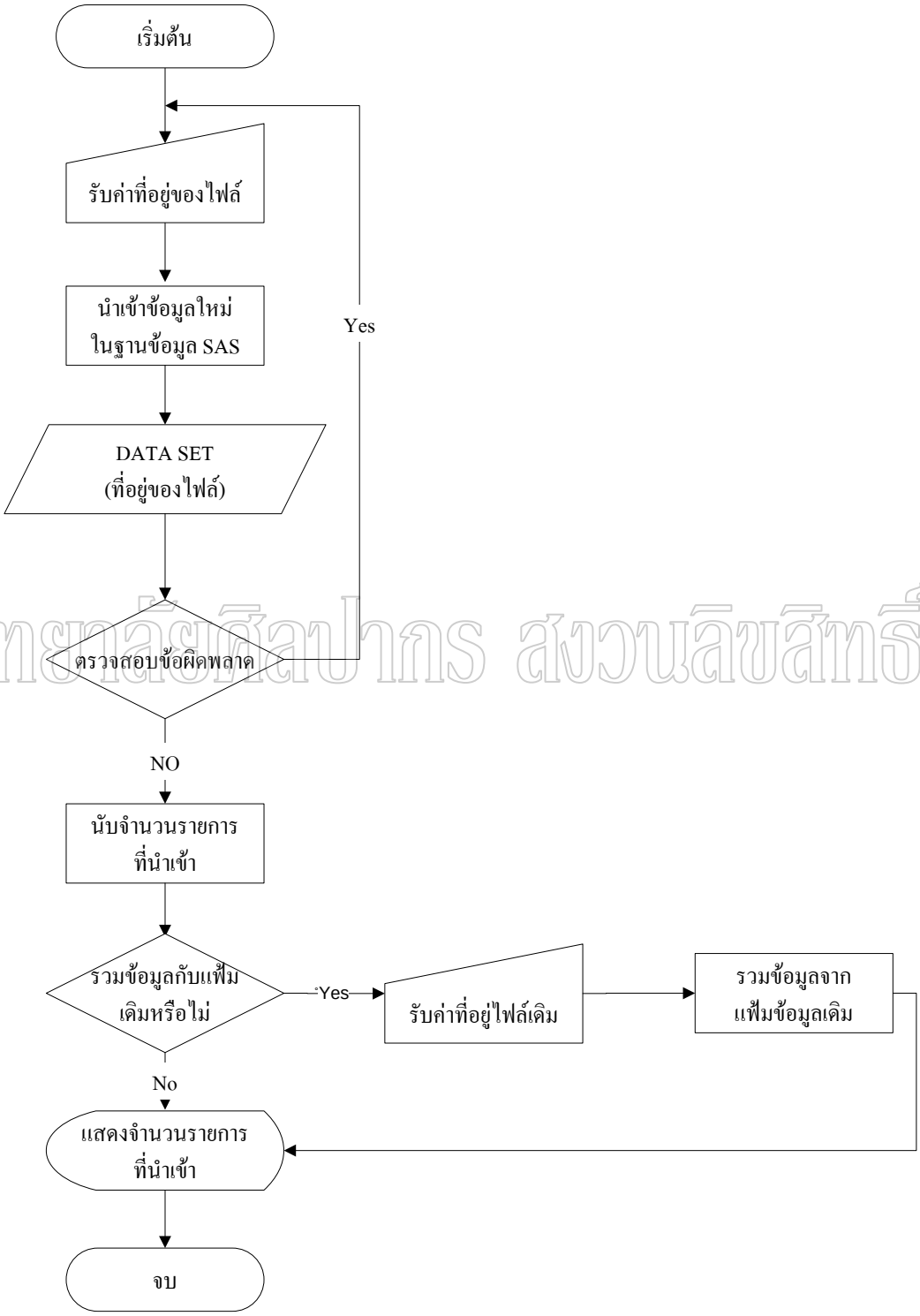
จากอัลกอริทึมสามารถแสดงให้เห็นในลักษณะผังงานได้ดังนี้

- **โมดูลการเข้าสู่ระบบ**



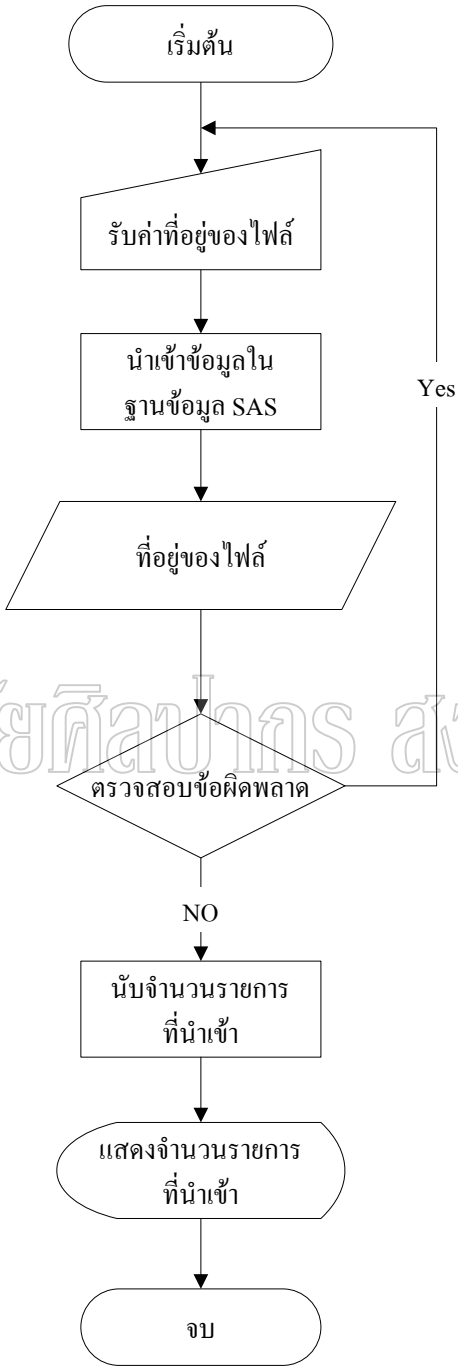
รูปที่ 18 ผังงานแสดงขั้นตอนการเข้าสู่ระบบ

- **โมดูลการนำเข้าข้อมูลใหม่**



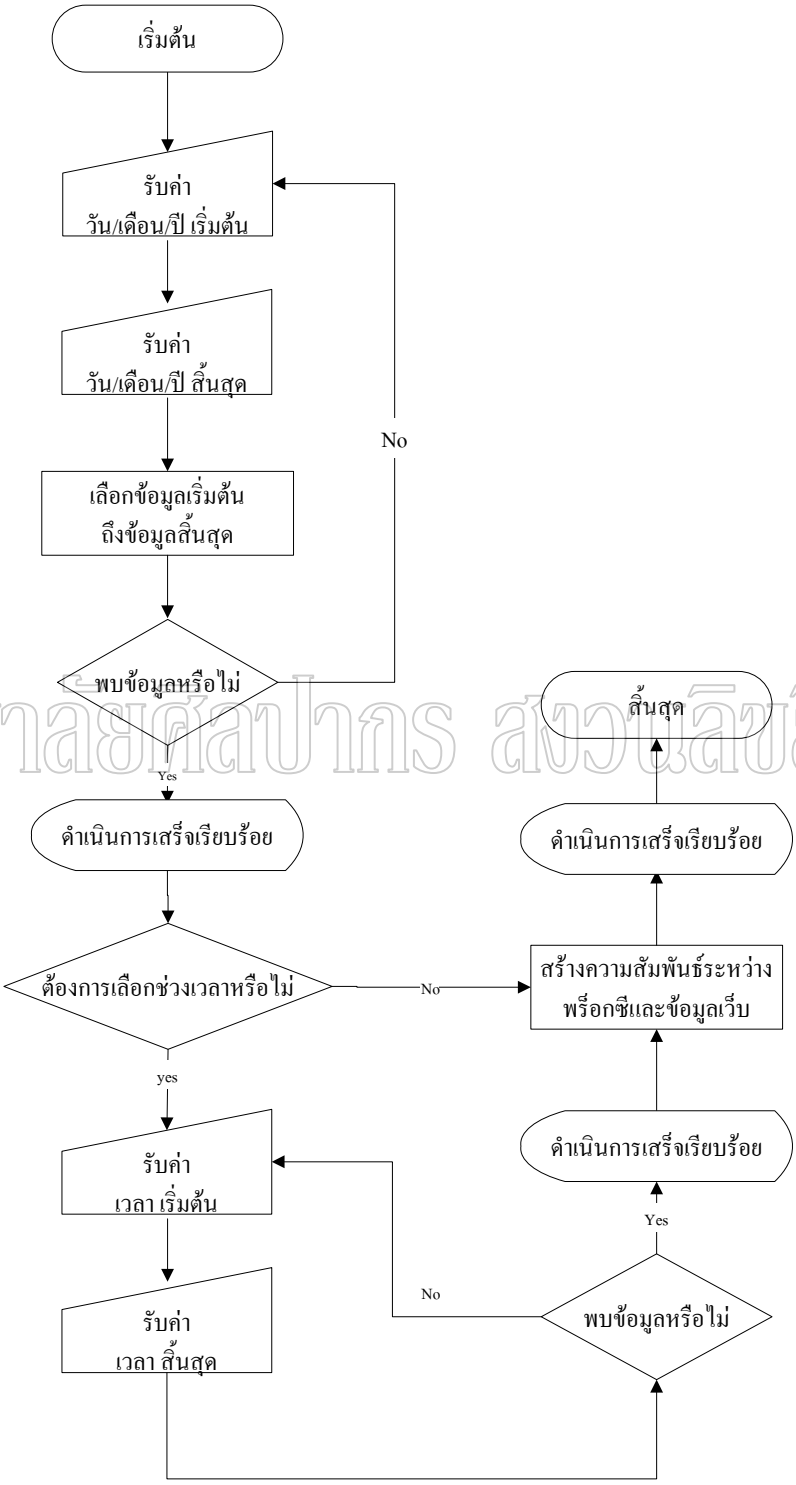
รูปที่ 19 ผังงานแสดงขั้นตอนการนำเข้าข้อมูลใหม่

- **โมดูลการนำเข้าข้อมูลเว็บ**



รูปที่ 20 ผังงานแสดงขั้นตอนการนำเข้าข้อมูลเว็บ

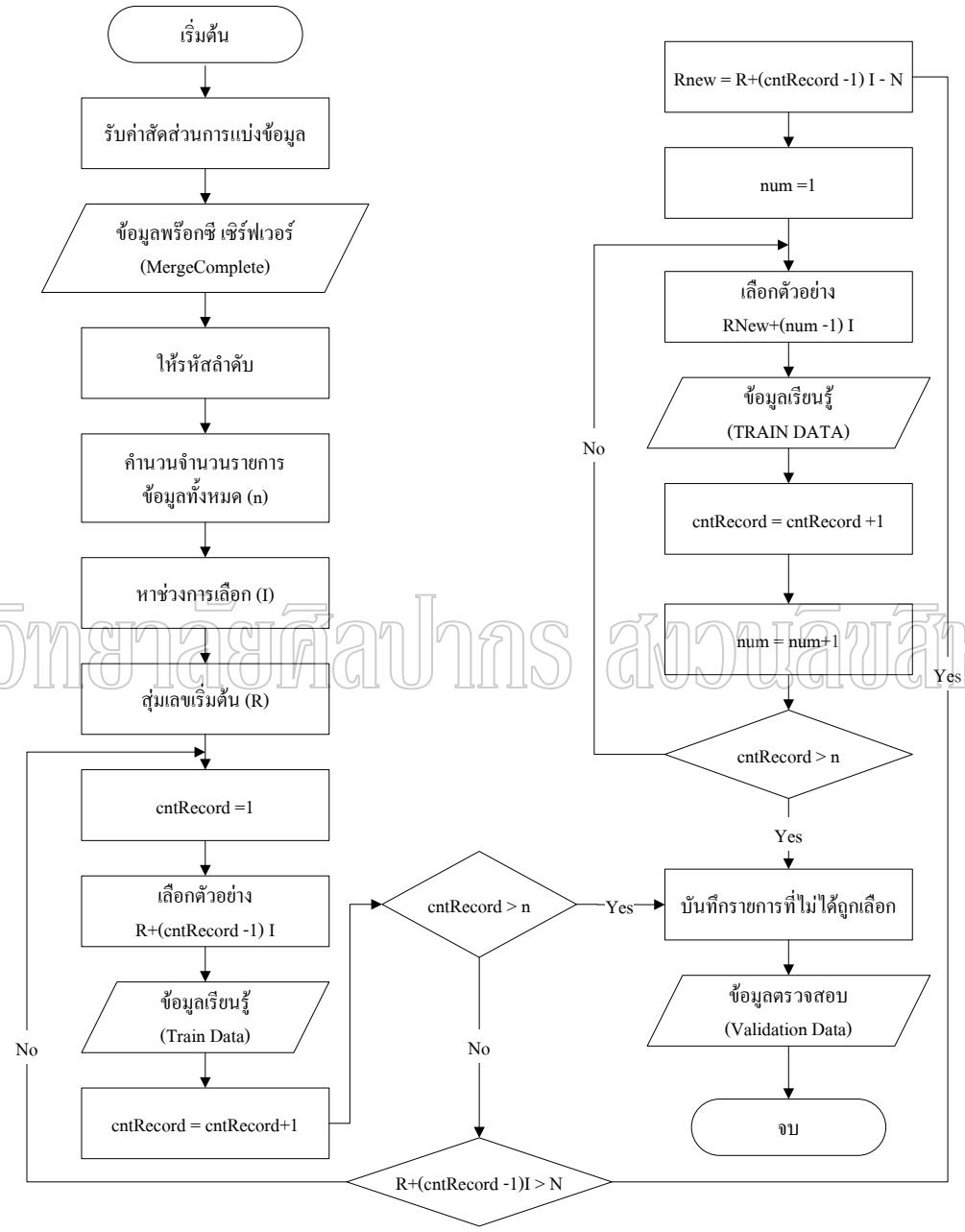
- **โมดูลกระบวนการก่อนศึกษาตัวแบบ/สร้างตัวแบบ**



รูปที่ 21 ผังงานแสดงกระบวนการก่อนศึกษาตัวแบบ/สร้างตัวแบบ

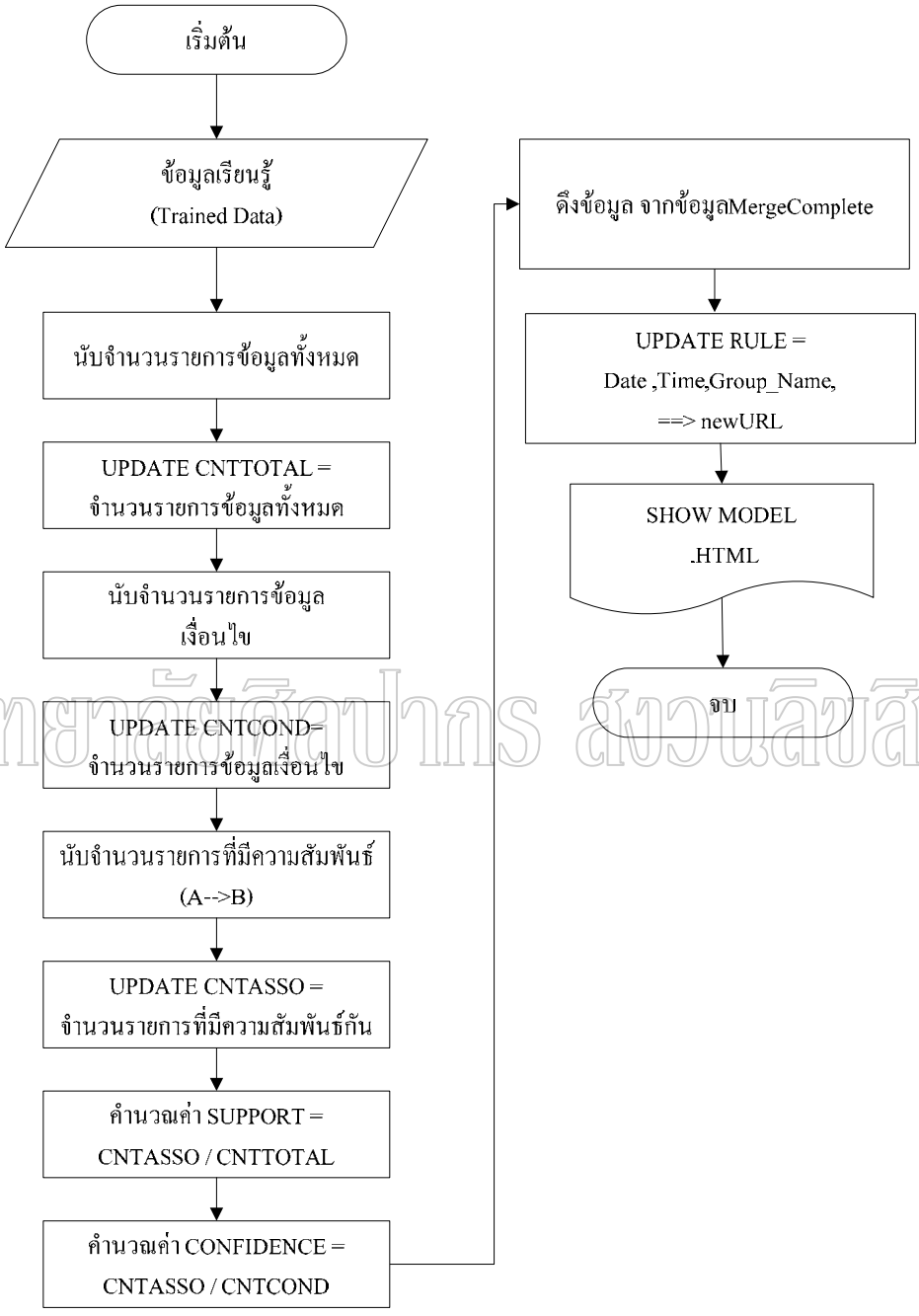
- **โมดูล การศึกษาตัวแบบ**

  <u>การเลือกตัวอย่างแบบมีระบบบวงกลม</u>



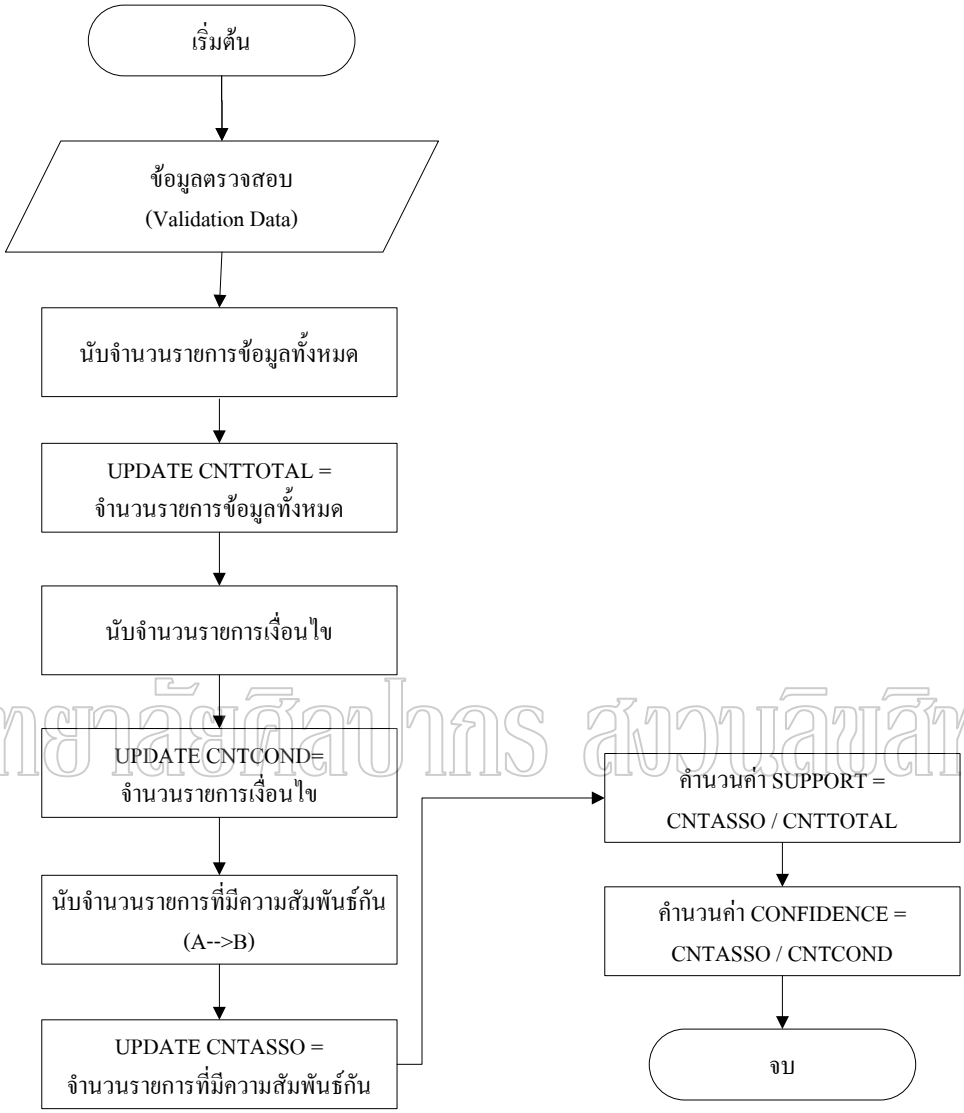
รูปที่ 22 ผังงานแสดงขั้นตอนการเลือกตัวอย่างแบบมีระบบบวงกลม

การสร้างตัวแบบข้อมูลเรียนรู้



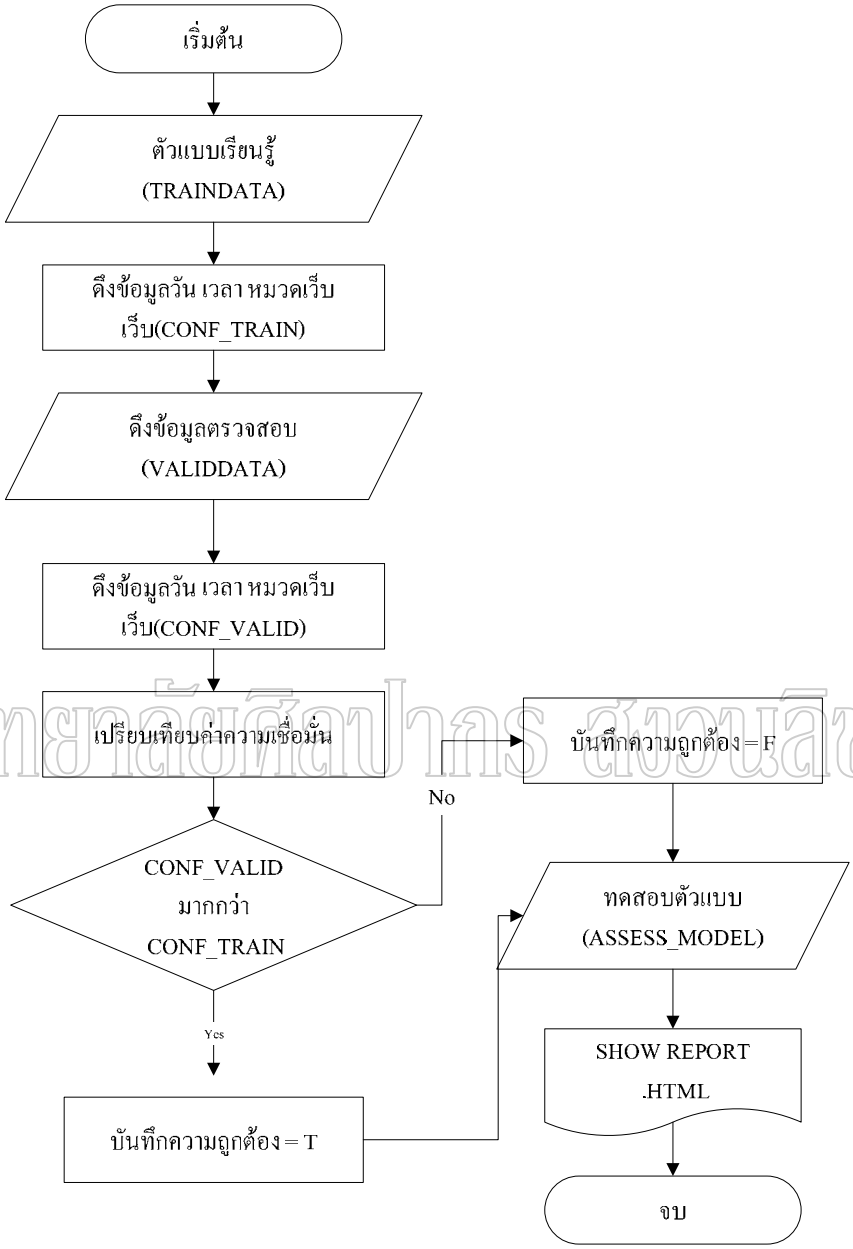
รูปที่ 23 ผังงานแสดงขั้นตอนการสร้างตัวแบบข้อมูลเรียนรู้

การสร้างตัวแบบข้อมูลตรวจสอบ



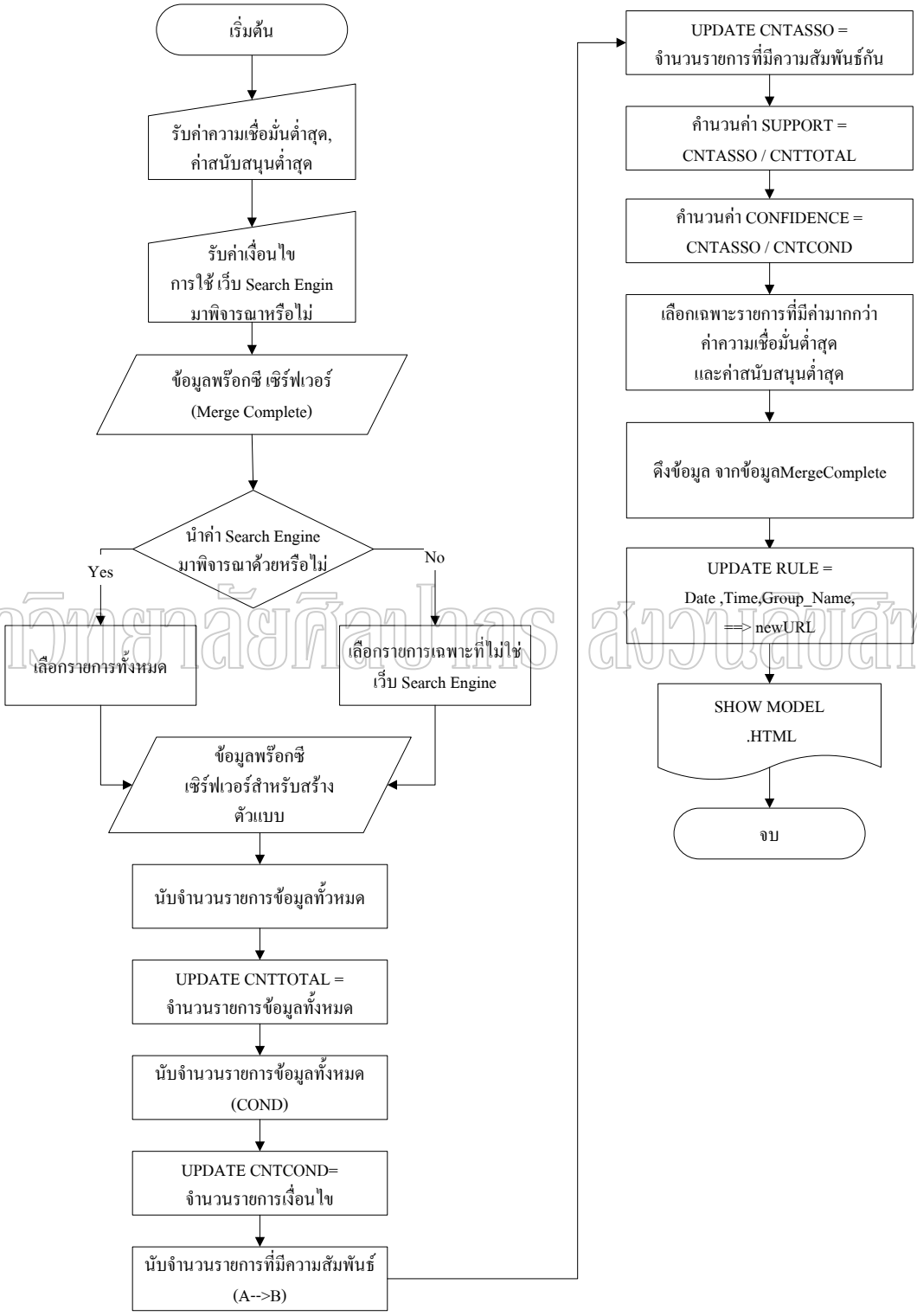
รูปที่ 24 ผังงานแสดงขั้นตอนการสร้างตัวแบบข้อมูลตรวจสอบ

<u>ทดสอบความถูกต้องตัวแบบ</u>



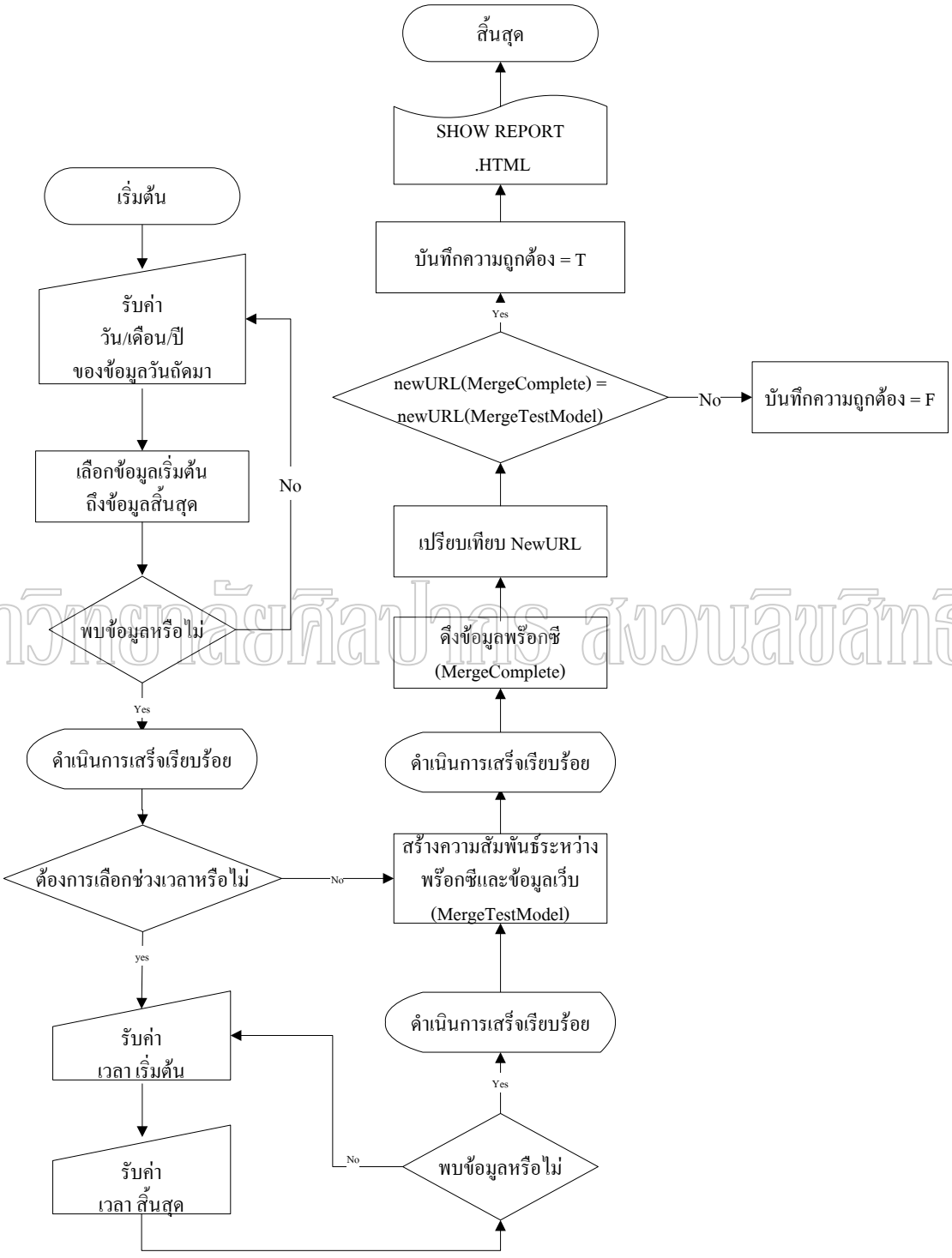
รูปที่ 25 ผังงานแสดงขั้นตอนการทดสอบความถูกต้องตัวแบบ

- **โมดูลการสร้างตัวแบบ**



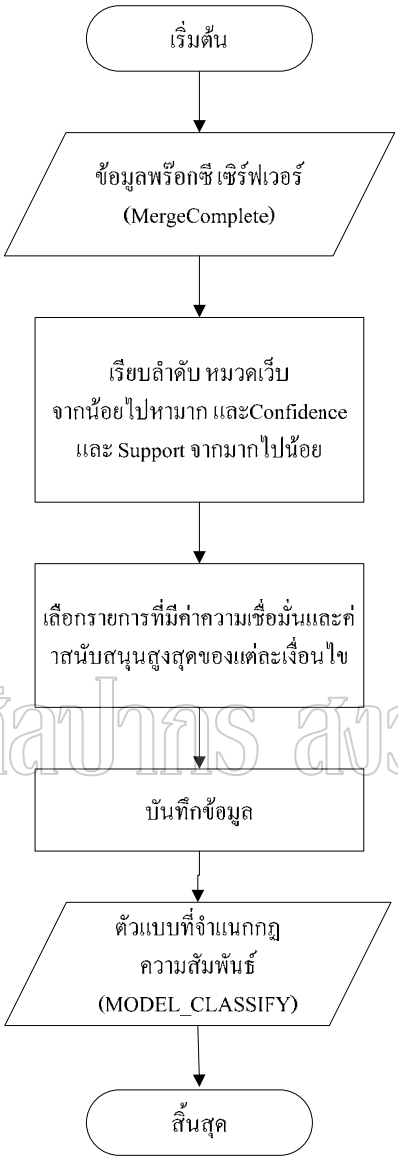
รูปที่ 26 ผังงานแสดงขั้นตอนการสร้างตัวแบบ

- **โมดูลการตรวจสอบตัวแบบกับข้อมูลจริง**



รูปที่ 27 ผังงานแสดงขั้นตอนการตรวจสอบตัวแบบกับข้อมูลจริง

- **โมดูลการจำแนกกฎความสัมพันธ์**



รูปที่ 28 ผังงานแสดงขั้นตอนการจำแนกกฎความสัมพันธ์

6. การทดสอบระบบ

การทดสอบระบบผู้วิจัยได้ทดสอบระบบโดยดูจากผลลัพธ์ที่ได้ในแต่ละกระบวนการโดยการตรวจสอบจากความถูกต้องของข้อมูล ดังตาราง

ตารางที่ 23 การทดสอบระบบ

กระบวนการ	วิธีการทดสอบ
การเข้าใช้ระบบ	สมมุติรหัสผู้ใช้ หรือรหัสผ่านที่ไม่อยู่ในระบบ
การนำเข้าข้อมูลใหม่	ตรวจสอบจำนวนรายการข้อมูลที่นำเข้าก่อนและหลังนำเข้า ตรวจสอบชนิด
การนำเข้าข้อมูลเว็บ	ตรวจสอบจำนวนรายการข้อมูลที่นำเข้าก่อนและหลังนำเข้า ตรวจสอบชนิด
กระบวนการก่อนสร้างตัวแบบ - เลือกวัน - เลือกเวลา - สร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร๊อกซีเซิร์ฟเวอร์	 ตรวจสอบจำนวนรายการทั้งหมด และจำนวนรายการหลังจากประมวลผลต้องเท่ากัน ตรวจสอบจำนวนรายการทั้งหมด และจำนวนรายการหลังจากประมวลผลต้องเท่ากัน โปรแกรมสามารถทำการรวมตารางระหว่างหมวดเว็บและข้อมูลพร๊อกซีเซิร์ฟเวอร์ได้ถูกต้อง
การศึกษาตัวแบบ - การเลือกตัวอย่างข้อมูลแบบมีระบบวงกลม - การสร้างตัวแบบข้อมูลเรียนรู้ - การทดสอบความถูกต้องตัวแบบ	 ตรวจสอบการทำงานของโปรแกรม โดยการตรวจสอบรายการข้อมูลที่เลือกได้ต้องถูกต้องตามวิธีการเลือกตัวอย่างแบบมีระบบวงกลม ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวนรายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่าความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้ ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวน

ตารางที่ 23 (ต่อ)

กระบวนการ	วิธีการทดสอบ
- การสร้างตัวแบบข้อมูลตรวจสอบ	รายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่าความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้
- คำนวณร้อยละความถูกต้อง	ตรวจสอบการเปรียบเทียบค่าความเชื่อมั่นมีความถูกต้องตามโปรแกรม และคำนวณร้อยละความถูกต้องของตัวแบบ
การสร้างตัวแบบ	ตรวจสอบค่าที่ได้จากการคำนวณต่าง ได้แก่จำนวนรายการข้อมูลทั้งหมด, จำนวนรายการที่เป็นเงื่อนไข (A), จำนวนรายการที่มีความสัมพันธ์กัน (A→B), การคำนวณค่าความเชื่อมั่น และค่าสนับสนุน และตรวจสอบกับข้อมูลที่นำมาใช้
การจำแนกกฎความสัมพันธ์	ตรวจสอบความถูกต้องในแต่ละเงื่อนไขจะต้องมีกฎเดียว และเป็นกฎที่มีค่าความเชื่อมั่นและค่าสนับสนุนสูงกว่าค่าความเชื่อมั่นและค่าสนับสนุนที่ต่ำสุด
การจัดทำรายงาน	ตรวจสอบรายงานแบบตาราง ต้องมีความถูกต้องตามข้อมูลที่มีจริง

7. การประเมินผลระบบ

ผู้วิจัยได้พัฒนาโปรแกรมและมีการทดสอบการใช้โปรแกรม พบว่าโปรแกรมที่พัฒนาขึ้นสามารถสร้างตัวแบบในการทำนายเนื้อหาของเว็บโดยใช้เทคนิคเหมืองข้อมูลกรณีศึกษามหาวิทยาลัยศิลปากร และนำตัวแบบที่สร้างไปประยุกต์ใช้ได้

ในการทดสอบระบบในแต่ละขั้นตอนสามารถสรุปผลได้ดังนี้

สรุปผลการทดสอบความถูกต้องตัวแบบ

ผู้วิจัยได้สร้างตัวแบบข้อมูลเรียนรู้ และสร้างตัวแบบข้อมูลตรวจสอบ โดยใช้ข้อมูลที่แบ่งได้ตามสัดส่วนต่างๆ และตรวจสอบความสอดคล้องกันของตัวแบบได้โดยผลจากการทดสอบที่ได้สามารถสรุปได้ว่า ในการแบ่งสัดส่วนแต่ละครั้งจะได้ผลลัพธ์ความถูกต้องไม่เท่ากัน เพราะในแต่

ละครั้งของการเลือกตัวอย่างข้อมูลจะได้ข้อมูลไม่เหมือนกัน ดังนั้นความถูกต้องของตัวแบบจะขึ้นอยู่กับข้อมูลที่เลือกมาได้ด้วย แต่จะพบว่าการแบ่งสัดส่วนข้อมูลจะมีผลกับความถูกต้องด้วย นั่นคือ ถ้าสัดส่วนการแบ่งข้อมูลมากตัวแบบจะมีร้อยละความถูกต้องมากขึ้นด้วย

สรุปผลการสร้างตัวแบบ

เมื่อตัวแบบที่ผู้วิจัยศึกษาสามารถนำมาใช้ได้ จึงใช้ข้อมูลพร๊อกซี เซิร์ฟเวอร์ที่ผ่านกระบวนการสร้างความสัมพันธ์เรียบร้อยแล้วมาหาความสัมพันธ์ ซึ่งจำนวนรายการข้อมูลจะขึ้นอยู่กับข้อมูลที่ผู้ใช้ระบบเลือกใช้ โดยสามารถเลือกได้ว่าต้องการนำเว็บ Search Engine มาพิจารณาสร้างตัวแบบหรือไม่ และจากการสร้างตัวแบบพบว่ากฎความสัมพันธ์ที่ค้นหาได้มีจำนวนมาก จึงต้องมีการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด เพื่อเลือกเฉพาะกฎความสัมพันธ์ที่มีความเชื่อมั่นสูงและจำนวนรายการข้อมูลหมวดเว็บที่มีจำนวนมาก

สรุปผลการจำแนกกฎความสัมพันธ์

การจำแนกกฎความสัมพันธ์เป็นการเลือกกฎความสัมพันธ์ที่มีค่าความเชื่อมั่นและค่าสนับสนุนสูงสุดของแต่ละกฎความสัมพันธ์ จะมีความสอดคล้องกับการกำหนดค่าความเชื่อมั่นและค่าสนับสนุนในการสร้างตัวแบบ นั่นคือเมื่อจำแนกกฎแล้วในแต่ละกฎจะมีค่าความเชื่อมั่นและค่าสนับสนุนน้อยที่สุดจะเท่ากับค่าความเชื่อมั่นและค่าสนับสนุนต่ำสุดตามที่กำหนดตอนสร้างตัวแบบ

สรุปผลการตรวจสอบความถูกต้องกับข้อมูลจริง

โดยทำการนำข้อมูลที่ทำการสร้างโมเดลมาเปรียบเทียบกับข้อมูลจริงของข้อมูลซึ่งพบว่าโมลเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บได้และสามารถเพิ่มอัตราการพบในระบบพร็อกซีได้โดยข้อมูลที่ใช้ในการตรวจสอบเป็นข้อมูลที่ถูกเรียกใช้จริงในวันถัดมา ซึ่งผลของการตรวจสอบถ้าอัตราการพบในระบบพร๊อกซีเพิ่มขึ้น อาจทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและสามารถลดปริมาณขอมูลในระบบเครือข่ายได้

บทที่ 5
สรุป อภิปรายผลและข้อเสนอแนะ

สรุปผลการวิจัย

ผู้วิจัยได้เสนอขั้นตอนวิธีการศึกษาตัวแบบและพัฒนาตัวแบบ เพื่อนำตัวแบบที่ได้นำไปทำนายเนื้อหาของเว็บประยุกต์ใช้ในการทำงานของระบบพร็อกซี เซิร์ฟเวอร์ โดยขั้นตอนวิธีที่นำเสนอสามารถแบ่งได้ดังนี้

ส่วนแรกเป็นการดำเนินการก่อนสร้างตัวแบบโดยทำการเลือกช่วงวันที่ต้องการทำนาย และเลือกเวลาที่ต้องการทำนาย และทำการความสัมพันธ์ของข้อมูลพร็อกซีและหมวดเว็บ เพื่อเป็นการเตรียมข้อมูลก่อนการศึกษาตัวแบบหรือก่อนการสร้างตัวแบบ

ส่วนที่ 2 เป็นการศึกษาตัวแปรเพื่อใช้ในการทำนายเนื้อหาของเว็บ โดยใช้เทคนิคการค้นหากฎความสัมพันธ์ ผู้วิจัยได้ศึกษาตัวแปรที่จะนำมาใช้สร้างเป็นเงื่อนไขและผลลัพธ์ ซึ่งในการวิจัยนี้ได้ใช้วัน เวลา และหมวดเว็บเป็นเงื่อนไข ส่วนผลลัพธ์คือเว็บ นั่นคือการสร้างตัวแบบนี้จะต้องหากฎความสัมพันธ์ระหว่าง วัน เวลา หมวดเว็บ และเว็บ โดยที่กฎความสัมพันธ์ที่ได้จะต้องคำนวณค่าความเชื่อมั่นและค่าสนับสนุน

ส่วนที่ 3 เป็นการศึกษาตัวแบบและทดสอบความถูกต้องของตัวแบบโดยนำข้อมูลพร็อกซีเซิร์ฟเวอร์ ที่ผ่านการกระบวนการก่อนสร้างตัวแบบ เรียบร้อยแล้วมาสร้างตัวแบบตามที่ได้ศึกษามา โดยแบ่งข้อมูลออกเป็น 2 ชุด คือข้อมูลเรียนรู้ และข้อมูลตรวจสอบ ในการแบ่งข้อมูลได้ใช้วิธีการเลือกตัวอย่างแบบมีระบบวงกลม และนำข้อมูลแต่ละชุดมาสร้างตัวแบบเป็นตัวแบบข้อมูลเรียนรู้และตัวแบบข้อมูลตรวจสอบโดยจะนำตัวแบบข้อมูลเรียนรู้และตัวแบบข้อมูลตรวจสอบมาเปรียบเทียบเพื่อหาความสอดคล้องกันของตัวแบบ ถ้าผลการเปรียบเทียบได้ร้อยละความสอดคล้องเกินร้อยละ 50 นั่นหมายความว่าตัวแบบที่ศึกษาและพัฒนามานั้นสามารถนำไปใช้ในการทำนายการเรียกใช้เว็บได้

ส่วนที่ 4 เป็นการสร้างตัวแบบเพื่อใช้ทำนายแนวโน้มการเรียกใช้งานเว็บ โดยนำข้อมูลพร็อกซี เซิร์ฟเวอร์ ที่ผ่านการกระบวนการก่อนสร้างตัวแบบเรียบร้อยแล้วมาสร้างตัวแบบ โดยสร้างความสัมพันธ์ระหว่าง วัน เวลา หมวดเว็บ และเว็บ ซึ่งกฎความสัมพันธ์ที่ได้จะคำนวณค่าความเชื่อมั่นและค่าสนับสนุนตามที่ผู้ใช้กำหนด

ส่วนที่ 5 เป็นตรวจสอบผลการทำนายจากข้อมูลจริงที่มีการเรียกใช้เว็บ โดยทำการนำข้อมูลที่ทำการสร้างตัวแบบมาเปรียบเทียบกับข้อมูลจริงของข้อมูลซึ่งพบว่าโมลเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บได้และสามารถเพิ่มประสิทธิภาพการทำงานของระบบพร็อกซี เซอร์เวอร์ ได้ซึ่งถ้าประสิทธิภาพการทำงานในระบบพร็อกซี เซอร์เวอร์เพิ่มขึ้น อาจทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและสามารถลดปริมาณข้อมูลในระบบเครือข่ายได้

ข้อเสนอแนะ

ผู้วิจัยมีข้อเสนอแนะบางประการเพื่อให้ตัวแบบที่สร้างขึ้นและโปรแกรมที่พัฒนาขึ้นมีประสิทธิภาพดีขึ้น ดังนี้

- การทำงานโดยใช้เทคนิคเหมืองข้อมูลในงานวิจัยนี้ความถูกต้องจะขึ้นอยู่กับการสร้างตัวแบบและข้อมูลของพร็อกซี เซิร์ฟเวอร์ ที่นำมาเสร้างตัวแบบ
- งานวิจัยนี้เป็นการวิจัยเชิงวิเคราะห์และการจำลองสถานการณ์ (Simulation) หากจะนำวิธีการนี้ไปใช้จริงจำเป็นจะต้องมีการพัฒนาซอร์ฟแวร์เพิ่มเติมและเสริมเข้าไปในระบบพร็อกซี
- พัฒนาโปรแกรมให้มีดำเนินการเร็วขึ้น
- พัฒนาโปรแกรมให้ใช้งานจริงได้ในระบบพร็อกซี เซิร์ฟเวอร์
- การวิจัยนี้เป็นการทำนายเนื้อหาของเว็บไทยเท่านั้น หากต้องการนำเทคนิคนี้ไปใช้จริงจำเป็นต้องจัดเก็บเนื้อหาเว็บต่างประเทศเพิ่มเติม

บรรณานุกรม

ภาษาไทย

กฤษณะ ไวยมัย และ ธีระวัฒน์ พงษ์ศิริปรีดา. “การใช้เทคนิค Association Rule Discovery เพื่อการจัดสรรกฎหมายในการพิจารณาคดีความ.” NECTEC Technical Journal (2003) :143-152.

กัลยา วินิชย์บัญชา. การวิเคราะห์สถิติขั้นสูงด้วย SPSS for Windows. กรุงเทพฯ: ธรรมสาร จำกัด, 2546.

ธีรภัทร มนตรีศาสตร์. Squid Proxy Caching Server [Online]. Accessed 5 November 2005.Available from http://micro.se-ed.com/content/mc205/MC205_181.asp

บุญเสริม กิจศิริกุล. “รายงานวิจัยฉบับสมบูรณ์.” โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2546.

มหาวิทยาลัยธรรมศาสตร์ . แนะนำในการ set Proxy Server ของธรรมศาสตร์[Online]. Accessed 5 October 2005. Available from http://www.tu.ac.th/internet/proxy/

มหาวิทยาลัยสุโขทัยธรรมมาธิราช.สาขาวิชาวิทยาศาสตร์และเทคโนโลยี. เอกสารการสอนชุดวิชาระบบสนับสนุนการตัดสินใจทางธุรกิจ(Business decisition support systems) , หน่วยที่ 9-15 . นนทบุรี : สำนักพิมพ์ มหาวิทยาลัยสุโขทัยธรรมมาธิราช, 2545.

ศูนย์คอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี. เลือกใช้ Proxy อย่างไรให้มีประสิทธิภาพ[Online]. Accessed 5 November 2005. Available from http://www.sut.ac.th/ccs/news/tip_tech/tip002.asp

อุดมทรัพย์ กรรดิพณิชกุล.เพิ่มความเร็วให้กับการเล่นเน็ตของคุณด้วย NAVISCOPE [Online]. 15 September 2000. Available from http://www.pantip.com/tech/newscols/column/15-09-00/naviscope/naviscope.html

Data Mining & Data Exploration Lab คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. Data Mining [Online]. Accessed 5 October 2005. Available from http://www.it.kmitl.ac.th/dme/publication_th.htm

ภาษาต่างประเทศ

Agrawal, Rakesh and Ramakrishnan Srikant."Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, (Septemper 1994).

Berry, Michael J.A. and Gordon S Linnoff. Data Mining Techniques for Marketing, Sale and Customer Relationship Management. New York : Wiley Publishing, 2004.

Berson, Alex and Stephen J Smith. Data Warehousing, Data Mining and OLAP. Singapore:McGraw Hill, 1997.

Connolly ,Thomas M. and Carolyn E.Begg . Data Base systems A Pratical Approach to Design, Implementation and Management. 3rd ed. United Kingdom :London ,2002 .

Han, Jiawei and Micheline Kamber. Data Mining Concepts and Techniques. USA : Morgan Kaufman,2001.

L.Cherkasova. “Improving WWW proxies performance with greedy-dual-size-frequency-caching policy” , In HP Technical Report, Polo Alto ,(November 1998).

Mohan, Sujaa Rani, Park E.K. ,and Yijie Han. “ Association Rule Based Data Mining Agents for Personalized Web Caching.” roceeding of the 29th Annual International Computer Software and Applications Conference (COMPSAC’05) 29 (2005).

Wong , Cody , Simon Shiu,and Sankar Pal “ Mining Fuzzy association rules for web access case adaptation.” Proceedings of the Workshop Program at the Fourth International Conference on Case-Based Reasoning 4 (31 July 2001).

Wu, Yi-Hung and P.Chen ArbeeL. “ Prediction of Web Page Accesses by Proxy Server Log”. World Wide Web:Internet and Web Information System ,5(2002):67-68.

Yang,Qiang, Hui Wang, and Wei Zhang.“Web-log Mining for Quantitative Temporal-Event Prediction”. IEEE Computational Intelligence Bulletin,Vol.1 No.1, (Decenber 2002):10-18.

ภาคผนวก

ภาคผนวก ก

คู่มือการใช้โปรแกรม

คู่มือการใช้โปรแกรม

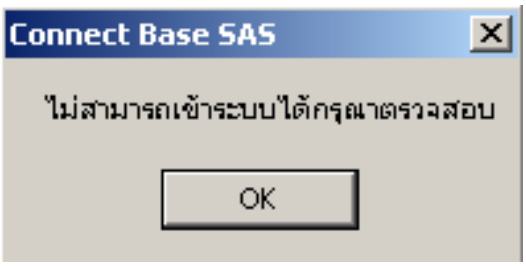
คู่มือการใช้โปรแกรมนี้จัดทำขึ้นเพื่อให้ผู้ใช้งานเข้าใจการทำงานของโปรแกรมและใช้โปรแกรมได้อย่างถูกต้อง ซึ่งโปรแกรมที่พัฒนาขึ้นเป็นโปรแกรมที่ใช้งานง่าย ในคู่มือการใช้โปรแกรมจะกล่าวเป็นขั้นตอนตามการทำงานดังนี้

ขั้นตอนการเข้าใช้ระบบ

1. เรียกใช้โปรแกรมจาก Shortcut จะปรากฏหน้าจอดังรูป

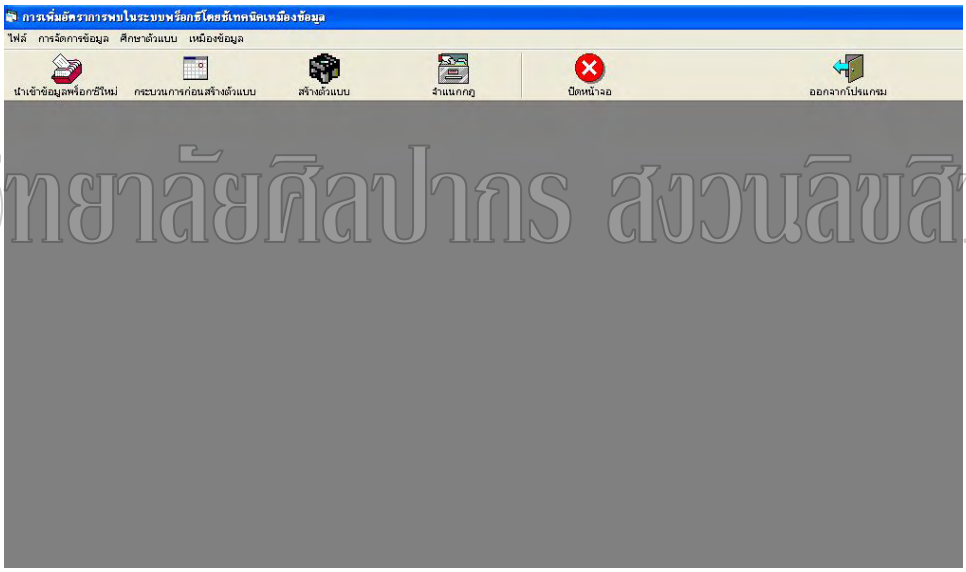
รูปที่ 29 หน้าจอการเข้าใช้ระบบ

2. ผู้ใช้กรอกข้อมูลตามช่องรายการดังนี้
   - Server name : ให้ใส่ชื่อของเครื่อง Server
   - Server port :ให้กำหนดเป็น 8591
   - Server protocol :ให้กำหนดเป็น Bridge
   - User name : ให้ใส่ชื่อรหัสผู้ใช้ที่มีสิทธิเข้าถึงข้อมูล
   - Password : ให้ใส่รหัสผ่านของผู้ใช้
3. เมื่อผู้ใช้กรอกข้อมูลครบแล้วให้กดปุ่ม **Connect**
4. ระบบจะทำการตรวจสอบสิทธิการเข้าถึงข้อมูล
5. ถ้าระบบตรวจสอบแล้วไม่สามารถเข้าใช้ระบบได้ ซึ่งสาเหตุอาจมาจากการกรอกข้อมูลตามช่องรายการผิดพลาด จะปรากฏหน้าจอแสดงข้อความดังรูป



รูปที่ 30 หน้าจอแสดงข้อความการเข้าใช้ระบบไม่ได้

6. ถ้าระบบตรวจสอบแล้วสามารถเข้าใช้ระบบได้ จะปรากฏหน้าจอเมนูการใช้งาน ดังรูป



รูปที่ 31 หน้าจอเมนูการใช้งาน

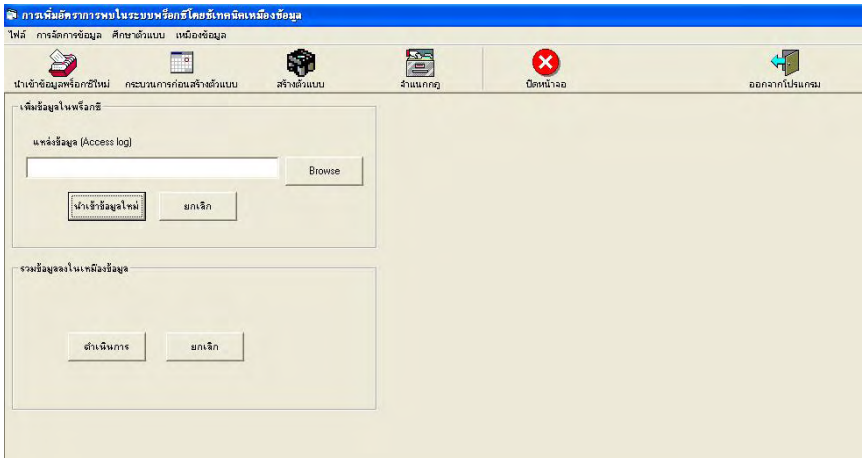
หลังจากที่ผู้ใช้สามารถเข้าใช้โปรแกรมได้แล้ว ต่อไปจะอธิบายขั้นตอนการเลือกใช้เมนูคำสั่งและแถบเครื่องมือโปรแกรมจะประกอบด้วยเมนูหลักอยู่ 5 เมนูหลักและเมนูย่อยต่างๆ ดังนี้

1. เมนูไฟล์ ประกอบด้วยเมนูย่อยดังนี้
   - ติดตั้งเครื่องพิมพ์
   - ออกจากโปรแกรม
2. เมนูการจัดการข้อมูล ประกอบด้วยเมนูย่อยดังนี้
   - การนำเข้าข้อมูลใหม่
   - นำเข้าข้อมูลเว็บ
3. เมนูการศึกษาตัวแบบ ประกอบด้วยเมนูย่อยดังนี้
   - สร้างตัวแบบจากการสุ่มตัวอย่าง
4. เมนูเหมืองข้อมูล ประกอบด้วยเมนูย่อยดังนี้
   - กระบวนการก่อนสร้างตัวแบบ
   - สร้างตัวแบบ
   - ตรวจสอบตัวแบบจากข้อมูลจริง
   - จำแนกกฎความสัมพันธ์

ขั้นตอนการนำข้อมูลใหม่จากพร็อกซีเซอร์เวอร์ สู่ฐานข้อมูล SAS

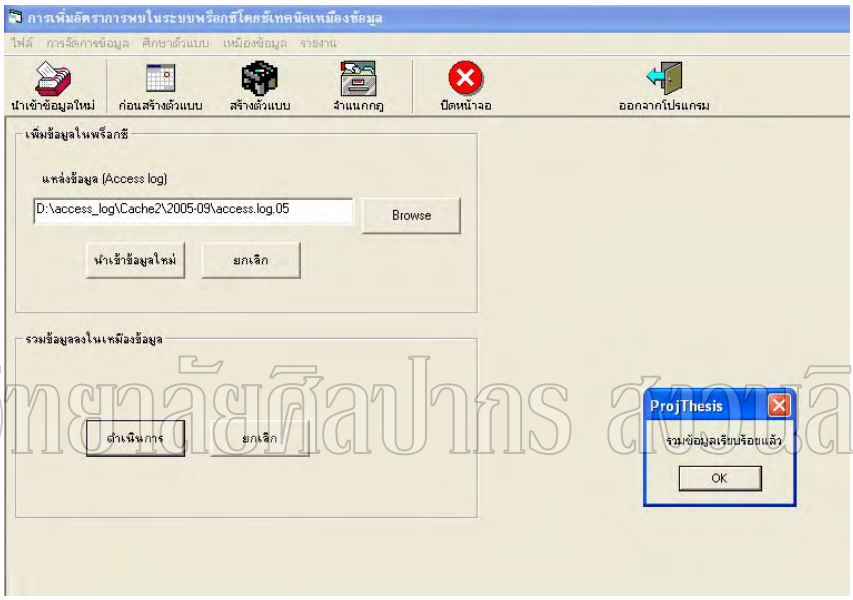
1. เลือกเมนูหลัก **<u>การจัดการข้อมูล</u>** → เมนูย่อย **<u>การนำเข้าข้อมูลพร็อกซีใหม่</u>** หรือกดปุ่ม 

จากแถบเครื่องมือจะปรากฏหน้าจอดังรูป



รูปที่ 32 หน้าจอการนำเข้าข้อมูลพร็อกซี

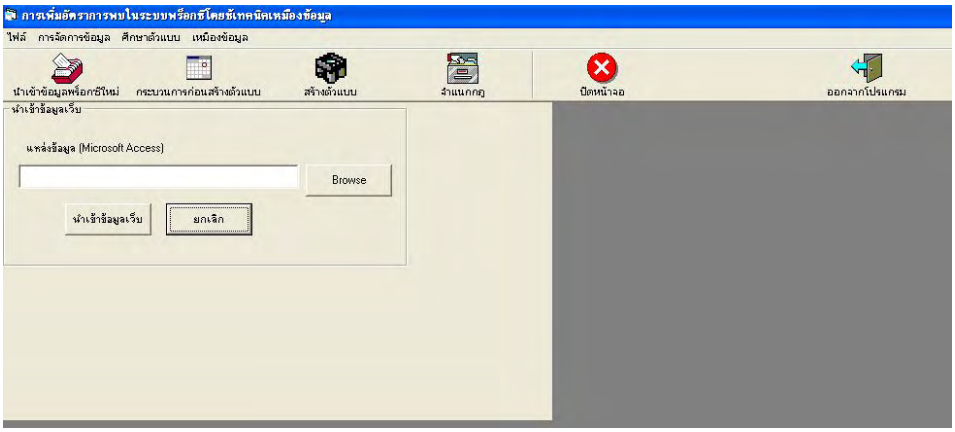
2. กดปุ่ม **Browse…**เพื่อเลือกไฟล์ข้อมูล
3. กดปุ่มนำเข้าข้อมูลใหม่
4. เมื่อเลือกไฟล์ข้อมูลครบถ้วนกดปุ่ม **นำเข้าข้อมูลใหม่** จะปรากฏข้อความ **ระบบกำลังประมวลผล** เมื่อนำข้อมูลเข้าสู่ฐานข้อมูล SAS แล้วระบบจะแสดงจำนวนข้อมูลที่ทำการนำเข้า
5. กดปุ่มดำเนินการ ในส่วนของการรวมข้อมูลลงในเหมืองข้อมูลดังรูป



รูปที่ 33 หน้าจอการรวมข้อมูลเรียบร้อย

ขั้นตอนการนำเข้าข้อมูลเว็บ

1. เลือกเมนูหลัก **การจัดการข้อมูล** → เมนูย่อย **การนำเข้าข้อมูลเว็บ** จะปรากฏหน้าจอดังรูป



รูปที่ 34 หน้าจอการนำเข้าข้อมูลเว็บ

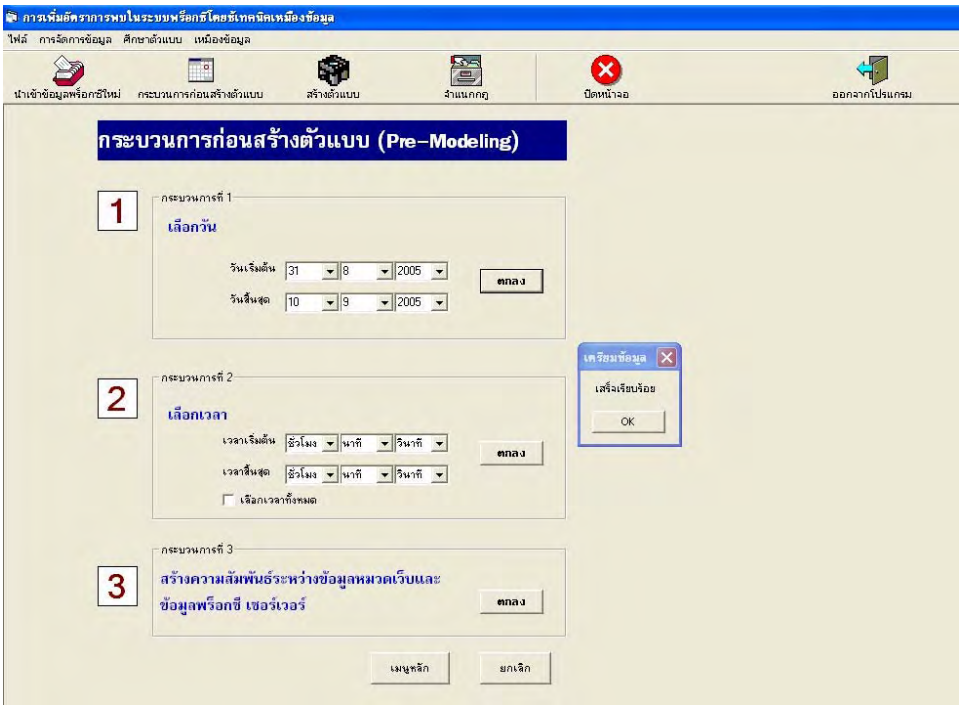
2. กดปุ่ม **Browse…**เพื่อเลือกไฟล์ข้อมูล
3. กดปุ่มนำเข้าข้อมูลเว็บ

ขั้นตอนก่อนสร้างตัวแบบ

ก่อนที่จะทำการศึกษาตัวแบบหรือสร้างตัวแบบเพื่อนำไปใช้ประโยชน์ต้องทำกระบวนการก่อนสร้างตัวแบบเพื่อให้ข้อมูลอยู่ในรูปแบบที่เหมาะสมในการสร้างตัวแบบ ซึ่งมีขั้นตอนดังนี้

1. เลือกเมนูหลัก **เหมืองข้อมูล** → เมนูย่อย **กระบวนการก่อนสร้างตัวแบบ** หรือกดปุ่ม 

บนแถบเครื่องมือจะปรากฏหน้าจอดังรูป

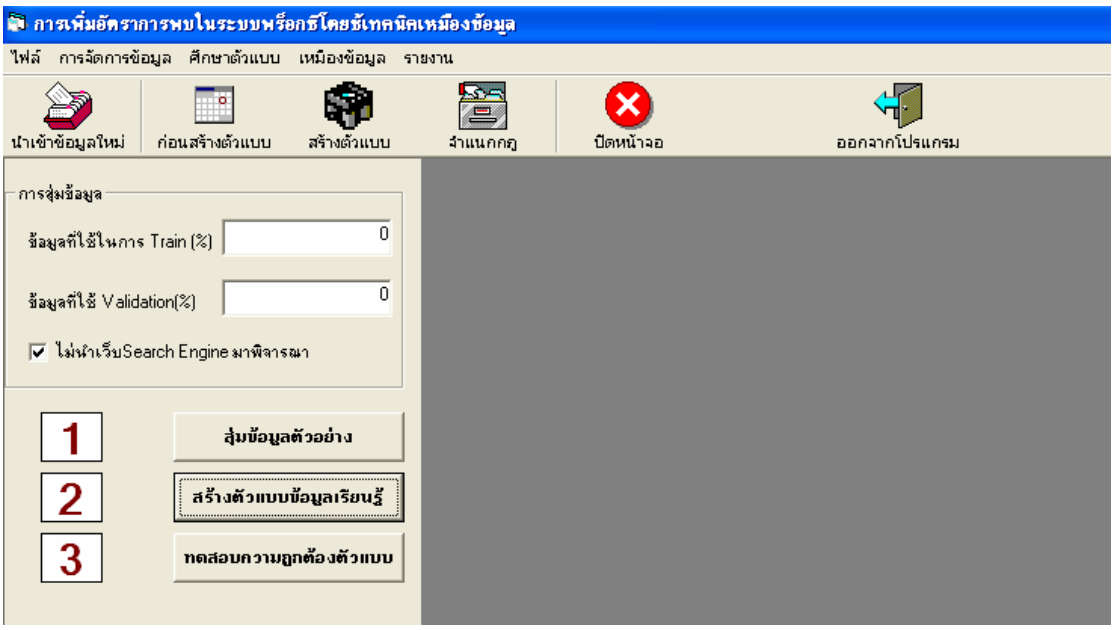


รูปที่ 35 หน้าจอกระบวนการก่อนการสร้างตัวแบบ

2. กดปุ่ม **ตกลง** ในกระบวนการที่ 1 เพื่อทำการเลือกวันที่ เมื่อประมวลในขั้นตอนที่ 1 เสร็จให้ทำขั้นตอนที่ 2 ต่อไป
3. กดปุ่ม **ตกลง** ในกระบวนการที่ 2 เพื่อทำการเลือกช่วงเวลา เมื่อประมวลในขั้นตอนที่ 2 เสร็จให้ทำขั้นตอนที่ 3 ต่อไป
4. กดปุ่ม **ตกลง** ในกระบวนการที่ 3 เพื่อทำการสร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลเว็บ

ขั้นตอนการศึกษาตัวแบบ

1. เลือกเมนูหลัก **ศึกษาตัวแบบ** → เมนูย่อย **การสร้างตัวแบบจากตัวอย่างสุ่ม** จะปรากฏหน้าจอดังรูป

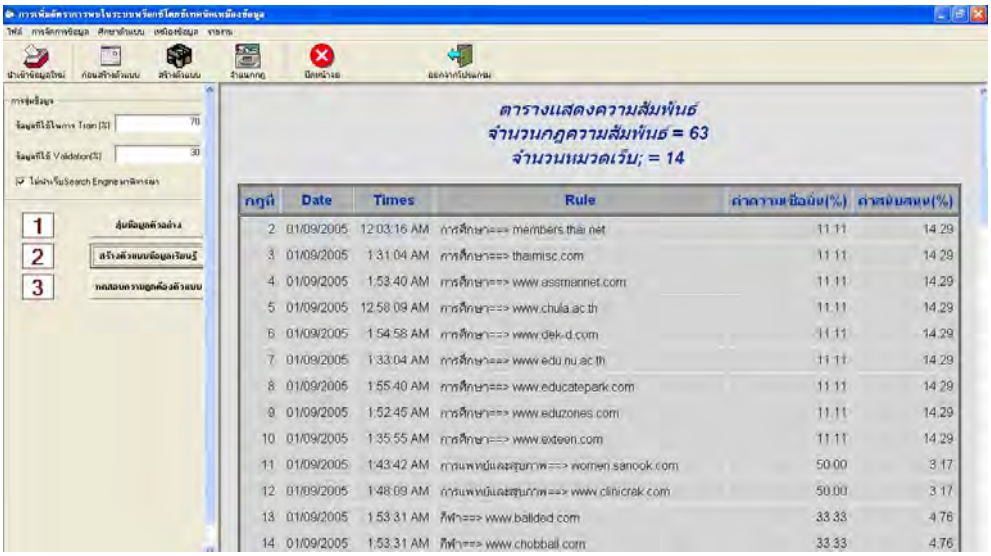


รูปที่ 36 หน้าจอการศึกษาตัวแบบ

2. กรอกสัดส่วนการแบ่งข้อมูลที่รายการ ข้อมูลที่ใช้ในการ Train (%) และช่องรายการ ข้อมูลที่ใช้ Validation (%)
3. เลือกช่องไม่นำเว็บที่เป็น Search Engine มาพิจารณา
4. กดปุ่ม**สุ่มข้อมูลตัวอย่าง** จะปรากฏหน้าจอแสดงข้อความการทำงานของกระบวนการย่อยต่างๆ ดังรูป

รูปที่ 37 หน้าจอแสดงข้อความการทำงานการสุ่มตัวอย่าง

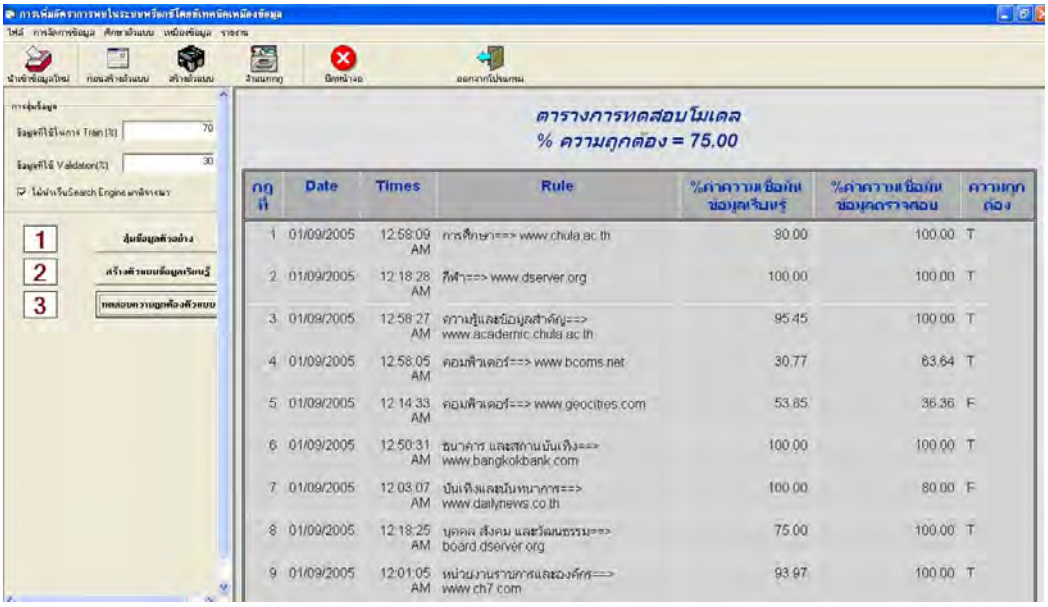
5. เมื่อประมวลผลเสร็จจะปรากฏหน้าจอแสดงข้อความ “สุ่มข้อมูลตัวอย่างเรียบร้อย”
6. กดปุ่ม**สร้างตัวแบบข้อมูลเรียนรู้** จะปรากฏหน้าจอแสดงข้อความการทำงานของกระบวนการย่อยต่างๆ เมื่อประมวลผลเสร็จเรียบร้อยจะแสดงตัวแบบที่สร้างได้ดังรูป



รูปที่ 38 หน้าจอแสดงตัวแบบที่สร้างได้ในรูปแบบกฎความสัมพันธ์

7. ในการแสดงผลจะแสดงในรูปของ HTML ผู้ใช้สามารถส่งข้อมูลไปยัง Microsoft Excel ได้ โดยคลิกขวาที่ตัวแบบ

8. ในการทดสอบความถูกต้องให้ผู้ใช้กดปุ่ม**ทดสอบความถูกต้องตัวแบบ** จะปรากฏหน้าจอแสดงข้อความการทำงานของกระบวนการย่อยต่างๆ เมื่อประมวลผลเสร็จเรียบร้อยจะประเมินผลร้อยละความถูกต้องของตัวแบบ ดังรูป



รูปที่ 39 หน้าจอแสดงการทดสอบความถูกต้อง

ขั้นตอนการทำเหมืองข้อมูล

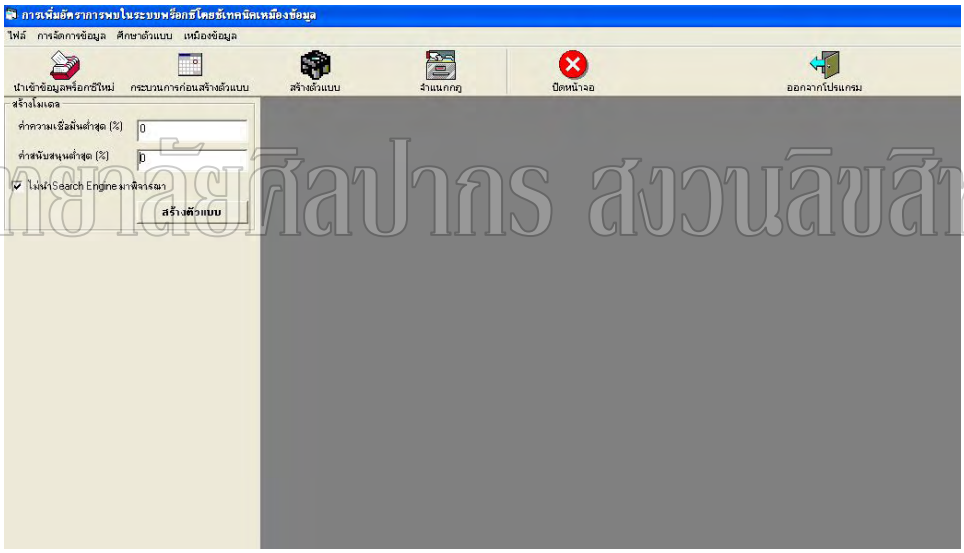
เมื่อผู้ใช้ได้เตรียมข้อมูลเรียบร้อยแล้วและได้ศึกษาตัวแบบว่ามีร้อยละความถูกต้องเป็นที่พอใจ ก็ทำการสร้างตัวแบบที่จะนำไปใช้ในการสนับสนุนการตัดสินใจต่อไป โดยมีขั้นตอนดังนี้

- **การสร้างตัวแบบ**

  การสร้างตัวแบบมีขั้นตอนดังนี้

1. เลือกเมนูหลัก **<u>เหมืองข้อมูล</u>** → เมนูย่อย **<u>การสร้างตัวแบบ</u>** หรือกดปุ่ม 

บนแถบเครื่องมือจะปรากฏหน้าจอดังรูป



รูปที่ 40 หน้าจอการสร้างตัวแบบ

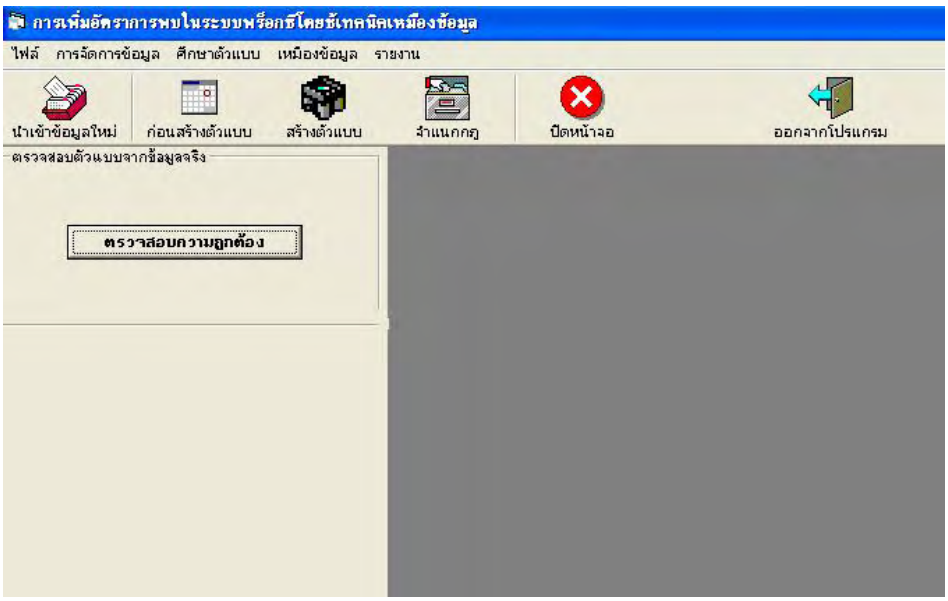
2. กำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด
3. เลือกไม่นำเว็บที่เป็น Search Engine มาพิจารณา
4. กดปุ่ม**<u>สร้างตัวแบบ</u>** จะปรากฏหน้าจอแสดงข้อความการทำงานของกระบวนการย่อยต่างๆ เมื่อประมวลผลเสร็จเรียบร้อยจะแสดงตัวแบบที่สร้างได้

- **กระบวนการตรวจสอบตัวแบบกับข้อมูลจริง**

  มีขั้นตอนดังนี้

1. เลือกเมนูหลัก **เหมืองข้อมูล** → เมนูย่อย **การตรวจสอบตัวแบบกับข้อมูลวันถัดมา**
2. ระบุวันที่ต้องการเปรียบเทียบ
3. ระบุช่วงเวลาที่ต้องการเปรียบเทียบ
4. สร้างความสัมพันธ์ระหว่างข้อมูลเว็บกับข้อมูลพร็อกซี
5. กดปุ่ม **ตรวจสอบ** เพื่อเข้าสู่กระบวนการตรวจสอบ

รูปที่ 41 หน้าจอการตรวจสอบตัวแบบกับข้อมูลจริง



รูปที่ 42 หน้าจอการกระบวนการตรวจสอบ

จากการกำหนดเงื่อนไขต้องการดูกฎความสัมพันธ์ที่ค้นหาได้มีอยู่หลายกฎ ดังนั้นถ้าจะนำตัวแบบไปใช้ต้องจำแนกกฎให้เหลือเพียงกฎเดียวในแต่ละเงื่อนไข ซึ่งมีขั้นตอนดังนี้

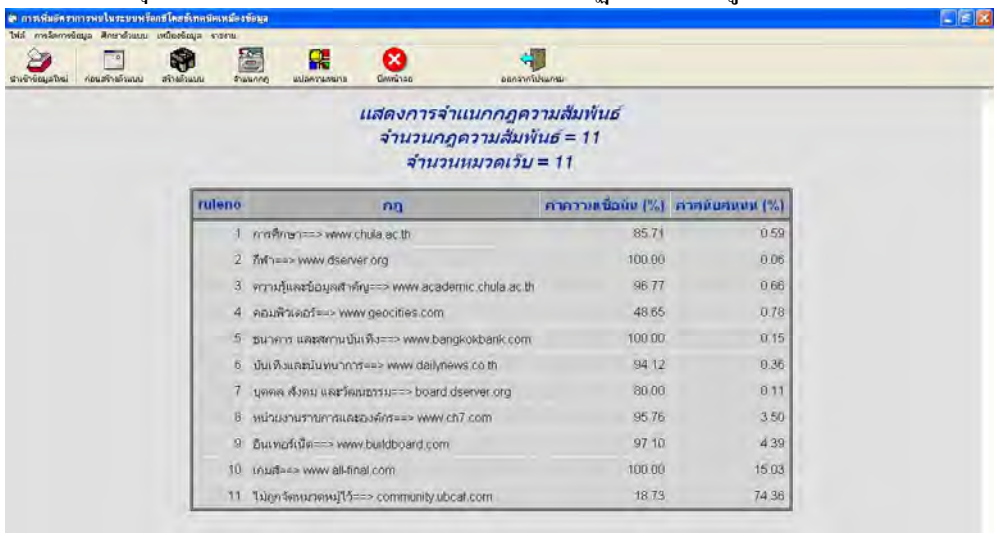
- **การจำแนกกฎความสัมพันธ์**

การจำแนกกำความสัมพันธ์ มีขั้นตอนดังนี้

1. เลือกเมนูหลัก **<u>เหมืองข้อมูล</u>** → เมนูย่อย **<u>การจำแนกกฎความสัมพันธ์</u>**

หรือกดปุ่ม 

บนแถบเครื่องมือจะปรากฏหน้าจอดังรูป



รูปที่ 43 หน้าจอการจำแนกกฎความสัมพันธ์

ภาคผนวก ข

คำอธิบายโมดูล

คำอธิบายโมดูล

- โมดูลการเข้าสู่ระบบ

  โมดูลการเข้าสู่ระบบเป็นการตรวจสอบผู้มีสิทธิเข้าใช้ระบบโดยรหัสผู้ใช้และรหัสผ่าน สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 24 คำอธิบายโปรแกรมย่อยโมดูลการเข้าสู่ระบบ

ชื่อฟอร์ม : FrmConnect	
Event Procedure / Sub / Function	คำอธิบาย
cmdIOMConn_Click	เมื่อกดปุ่ม Connect จะเรียกใช้ฟังก์ชัน open_baseSAS
open_baseSAS	เป็นฟังก์ชันในการติดต่อฐานข้อมูลถ้าติดต่อฐานข้อมูลได้จะทำการสร้าง Library โดยการเรียกโปรแกรมย่อย Assign_Library และแสดงหน้าจอเมนูโปรแกรม MdiMenu
Assign_Library	เป็นโปรแกรมย่อยมีหน้าที่สร้าง Library สำหรับอ้างถึงข้อมูล SAS

- โมดูลการนำเข้าข้อมูลพร๊อกซีใหม่

  โมดูลการเข้านำเข้าข้อมูลเป็นการนำข้อมูลที่อยู่ในพร๊อกซี เซิร์ฟเวอร์ เข้าสู่ฐานข้อมูล SAS สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 25 คำอธิบายโปรแกรมย่อยโมดูลการนำเข้าข้อมูลพร๊อกซีใหม่

ชื่อฟอร์ม : FrmImnew	
Event Procedure / Sub / Function	คำอธิบาย
cmdBrowse_Click	เมื่อกดปุ่ม Browse จะแสดงหน้าจอ Open เพื่อเลือก File ฐานข้อมูลที่ต้องการ
cmdImnew_Click	เมื่อกดปุ่มนำเข้า โปรแกรมจะเรียกใช้สคริปต์ไฟล์ Import_complete.sas และ cleantime.sas
Import_complete.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่นำข้อมูลที่ถูกระบุเข้าสู่ฐานข้อมูล SAS โดยจะแบ่งข้อมูลออกเป็น 10 คอลัมภ์
cleantime.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่แปลงเวลาจาก Unix timeStamp เป็นรูปแบบของเวลาปกติ

ตารางที่ 25 คำอธิบายโปรแกรมย่อยโมดูลการนำเข้าข้อมูลพร๊อกซีใหม่ (ต่อ)

ชื่อฟอร์ม : FrmImnew	
Event Procedure / Sub / Function	คำอธิบาย
cmdSumsata_Click	เมื่อกดปุ่มดำเนินการรวมข้อมูลลงในเหมืองข้อมูลระบบระเรียกใช้สคริปต์ไฟล์ addnewdata.sas
addnewdata.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่รวมตารางที่ได้จากการนำเข้าข้อมูลใหม่เข้าด้วยกัน

- โมดูลการนำเข้าข้อมูลเว็บ

โมดูลการเข้านำเข้าข้อมูลเว็บเป็นการนำข้อมูลที่อยู่ในฐานข้อมูล Microsoft Access เข้าสู่ฐานข้อมูล SAS สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 26 คำอธิบายโปรแกรมย่อยโมดูลการนำเข้าข้อมูล

ชื่อฟอร์ม : FrmImweb	
Event Procedure / Sub / Function	คำอธิบาย
cmdBrowse_Click	เมื่อกดปุ่ม Browse จะแสดงหน้าจอ Open เพื่อเลือก File ฐานข้อมูลที่ต้องการ
cmdImweb_Click	เมื่อกดปุ่มนำเข้า โปรแกรมจะเรียกใช้ไฟล์ importGroupWeb.sas MatchGroupWeb.sas meesook.sas Clean_me.sas
importGroupWeb.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่นำข้อมูลตารางที่ระบุเข้าสู่ฐานข้อมูล SAS
MatchGroupWeb.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่รวมตาราง au.GroupWeb กับ au.subGroupWeb
meesook.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่รวมตาราง au.MatchGroupWeb กับ work.Imp_meesook เข้าด้วยกัน
Clean_me.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่ตรวจสอบชื่อเว็บไม่ให้เกิดความซ้ำซ้อนในฐานข้อมูลเว็บ

- โมดูลกระบวนการก่อนสร้างตัวแบบ

โมดูลกระบวนการก่อนสร้างตัวแบบเป็นกระบวนการเตรียมข้อมูลก่อนสร้างตัวแบบ โดยระบุ วัน เวลา และสร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร็อกซี่เซิร์ฟเวอร์ สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 27 คำอธิบายโปรแกรมย่อยโมดูลกระบวนการก่อนสร้างตัวแบบ

ชื่อฟอร์ม :	
Event Procedure / Sub / Function	คำอธิบาย
cmdSelectDate_click	เมื่อกดปุ่มตกลง เพื่อเลือก วัน ในกระบวนการที่1 โปรแกรมจะรับค่าวันที่เริ่มต้น วันที่สิ้นสุด และจะเรียกใช้สคริปต์ไฟล์ selectdate.sas
selectdate.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่เลือกวันตามที่ผู้ใช้ต้องการ
cmdselecttime_click	เมื่อกดปุ่มตกลง เพื่อเลือกเวลา ในกระบวนการที่2 โปรแกรมจะรับค่าเวลาเริ่มต้น เวลาสิ้นสุด และจะเรียกใช้สคริปต์ไฟล์ selecttime.sas
selecttime.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่เลือกเวลาตามที่ผู้ใช้ต้องการ
cmdrelateAuto_click	เมื่อกดปุ่มตกลง ในกระบวนการที่3 เพื่อสร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร็อกซี่ เซิร์ฟเวอร์ โปรแกรมจะเรียกสคริปต์ไฟล์ Mergedata.sas
mergedata.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่สร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร๊อกซีเซิร์ฟเวอร์

- โมดูลการศึกษาตัวแบบ

โมดูลการศึกษาตัวแบบเป็นการนำข้อมูลพร๊อกซี เซิร์ฟเวอร์ ที่ผ่านกระบวนการก่อนสร้างตัวแบบ มาแบ่งออกเป็น 2 ส่วนสำหรับเป็นข้อมูลเรียนรู้ และข้อมูลตรวจสอบ หลังจากนั้นก็นำข้อมูลเรียนรู้มาสร้างตัวแบบโดยการค้นหากฎความสัมพันธ์ของวัน เวลา หมวดเว็บ และเว็บ และนำข้อมูลตรวจสอบมาทดสอบตัวแบบที่สร้างขึ้น สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 28 คำอธิบายโปรแกรมย่อยโมดูลการศึกษาตัวแบบ

ชื่อฟอร์ม : FrmTrainModel	
Event Procedure / Sub / Function	คำอธิบาย
cmdSampling_Click	เมื่อกดปุ่มสุ่มข้อมูลตัวอย่าง โปรแกรมจะเรียกใช้สคริปต์ไฟล์ model_nosearch.sas sampledata.sas
cmdCreateModel_Click	เมื่อกดปุ่มสร้างตัวแบบข้อมูลเรียนรู้ โปรแกรมจะเรียกใช้สคริปต์ไฟล์ตามลำดับดังนี้ - model_cnttotal.sas - model_condA.sas - model_condAB.sas - model_conf_supp.sas - model_sym.sas - model_html.sas
cmdCorrect_Click	เมื่อกดปุ่มสร้างตัวแบบข้อมูลเรียนรู้ โปรแกรมจะเรียกใช้สคริปต์ไฟล์ตามลำดับดังนี้ - model_cnttotal.sas - model_condA.sas - model_condAB.sas - model_conf_supp.sas - assess_model.sas - assess_html.sas
model_nosearch.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่เลือกเว็บที่เป็น Search Engine ออกจากโปรแกรม
sampledata.sas	เป็นสคริปต์ไฟล์ที่ทำหน้าที่แบ่งข้อมูลออกเป็น 2 ส่วนคือ ข้อมูลเรียนรู้ และข้อมูลตรวจสอบตามสัดส่วนที่ผู้ใช้กำหนด โดยใช้การเลือกตัวอย่างแบบมีระบบวงกลม
model_cnttotal.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลทั้งหมด (Total)
model_condA.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลแต่ละวัน เวลา และหมวดเว็บ (CntCond)

ตารางที่ 28 คำอธิบายโปรแกรมย่อยโมดูลการศึกษาตัวแบบ (ต่อ)

ชื่อฟอร์ม : FrmTrainModel	
Event Procedure / Sub / Function	คำอธิบาย
model_condAB.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลแต่ละวัน เวลา หมวดเว็บ และเว็บ (CntAsso)
model_conf_supp.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณค่าความเชื่อมั่นและค่าสนับสนุน
model_sym.sas	เป็นสคริปต์ไฟล์ทำหน้าที่ดึงข้อมูลจากตารางอ้างอิงเพื่อเตรียมข้อมูลให้พร้อมสำหรับแสดงตัวแบบ
model_html.sas	เป็นสคริปต์ไฟล์ทำหน้าที่แสดงตัวแบบที่ค้นหากฎความสัมพันธ์ได้เป็นแบบ HTML
assess_model.sas	เป็นสคริปต์ไฟล์ทำหน้าที่นำข้อมูลตรวจสอบมาทดสอบตัวแบบข้อมูลเรียนรู้
assess_html.sas	เป็นสคริปต์ไฟล์ทำหน้าที่นำผลการทดสอบมาแสดงในรูปแบบ HTML

- โมดูลการสร้างตัวแบบ

โมดูลการสร้างตัวแบบเป็นการนำข้อมูลพร๊อกซี เซิร์ฟเวอร์ที่ผ่านกระบวนการก่อนสร้างตัวแบบทั้งหมดมาค้นหากฎความสัมพันธ์ของวัน เวลา หมวดเว็บและเว็บโดยมีการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 29 คำอธิบายโปรแกรมย่อยโมดูลการสร้างตัวแบบ

ชื่อฟอร์ม : FrmModel	
Event Procedure / Sub / Function	คำอธิบาย
cmdCreateModel_Click	เมื่อกดปุ่มสร้างตัวแบบ โปรแกรมจะเรียกใช้ สคริปต์ไฟล์ตามลำดับดังนี้ - model_nosearch - model_cnttotal.sas - model_condA.sas - model_condAB.sas

ตารางที่ 29 คำอธิบายโปรแกรมย่อยโมดูลการสร้างตัวแบบ(ต่อ)

ชื่อฟอร์ม : FrmModel	
Event Procedure / Sub / Function	คำอธิบาย
	- model_conf_supp.sas - model_sym.sas - model_html.sas
cmdQuery_Click	เมื่อกดปุ่มสร้างตัวแบบข้อมูลเรียนรู้ โปรแกรมจะเรียกใช้สคริปต์ไฟล์ query_model
model_noosearch	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่เลือกเว็บที่เป็น Search Engine ออกจากโปรแกรม
model_cnttotal.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลทั้งหมด
model_condA.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลแต่ละ วัน เวลา หมวดเว็บ (CntCond)
model_condAB.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณจำนวนรายการข้อมูลแต่ละวัน เวลา หมวดเว็บ และเว็บ(CntAsso)
model_conf_supp.sas	เป็นสคริปต์ไฟล์ทำหน้าที่คำนวณค่าความเชื่อมั่นและค่าสนับสนุน
model_sym.sas	เป็นสคริปต์ไฟล์ทำหน้าที่ดึงข้อมูลจากตารางอ้างอิงเพื่อเตรียมข้อมูลให้พร้อมสำหรับแสดงตัวแบบ
model_html.sas	เป็นสคริปต์ไฟล์ทำหน้าที่แสดงตัวแบบที่ค้นหากฎความสัมพันธ์ได้เป็นแบบ HTML

- โมดูลการจำแนกกฎความสัมพันธ์

โมดูลการจำแนกกฎความสัมพันธ์ เป็นการเลือกกฎความสัมพันธ์โดยใช้เกณฑ์พิจารณาที่ได้กล่าวมาแล้ว เพื่อนำไปใช้ทำนายการเรียกใช้เนื้อหาเว็บต่อไป สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 30 คำอธิบายโปรแกรมย่อยโมดูลการจำแนกกฎความสัมพันธ์

ชื่อฟอร์ม : MdiMenu	
Event Procedure / Sub / Function	คำอธิบาย
mnuClassify_Click หรือ Toolbar1_ButtonClick	เมื่อกดปุ่มเมนูคำสั่งการจำแนกกฎความสัมพันธ์ หรือกดปุ่มการจำแนกกฎความสัมพันธ์ โปรแกรมจะเรียกใช้โปรแกรมย่อย classify_process
classify_process	เป็นโปรแกรมย่อยที่เรียกใช้สคริปต์ไฟล์ model_classify.sas
model_classify.sas	เป็นสคริปต์ไฟล์ทำหน้าที่เลือกกฎความสัมพันธ์จากตัวแบบที่สร้างขึ้นให้เหลือเพียงกฎเดียวในแต่ละเงื่อนไขสาขาวิชาและเพศและแสดงตัวแบบที่จำแนกกฎเสร็จแล้วในรูปของ HTML

- โมดูลตรวจสอบตัวแบบจากข้อมูลจริง

โมดูลตรวจสอบตัวแบบจากข้อมูลจริง เป็นการตรวจสอบความถูกต้องของตัวแบบกับข้อมูลที่ถูกเรียกใช้จริงในวันถัดมา สามารถอธิบายโปรแกรมย่อยต่างๆได้ดังตาราง

ตารางที่ 31 คำอธิบายโปรแกรมย่อยโมดูลตรวจสอบตัวแบบจากข้อมูลจริง

ชื่อฟอร์ม : frmtestmodel	
Event Procedure / Sub / Function	คำอธิบาย
cmdselectdate_click	เมื่อกดปุ่มตกลง เพื่อเลือก วัน ในกระบวนการที่1 โปรแกรมจะรับค่าวัน เดือน ปี และจะเรียกใช้สคริปต์ไฟล์ DateTestModel.sas
DateTestModel.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่เลือกวันที่ต้องการตรวจสอบ
cmdselecttime_click	เมื่อกดปุ่มตกลง เพื่อเลือกเวลา ในกระบวนการที่2 โปรแกรมจะรับค่าเวลาเริ่มต้น เวลาสิ้นสุด และจะเรียกใช้สคริปต์ไฟล์ TimeTestModel.sas
TimeTestModel.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่ เลือกเวลาที่ต้องการตรวจสอบ

ตารางที่ 31 คำอธิบายโปรแกรมย่อยโมดูลตรวจสอบตัวแบบจากข้อมูลจริง (ต่อ)

ชื่อฟอร์ม : frmtestmodel	
Event Procedure / Sub / Function	คำอธิบาย
cmdrelateAuto_Click	เมื่อกดปุ่มตกลง ในกระบวนการที่3 เพื่อสร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร็อกซี่เซิร์ฟเวอร์ โปรแกรมจะเรียกสคริปต์ไฟล์ MergeTestModel.sas
MergeTestModel.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่สร้างความสัมพันธ์ระหว่างข้อมูลหมวดเว็บและข้อมูลพร็อกซี่เซิร์ฟเวอร์ ที่ต้องการตรวจสอบ

- โมดูลตรวจสอบความถูกต้อง

โมดูลตรวจสอบความถูกต้อง เป็นกระบวนการต่อเนื่องจากโมดูลตรวจสอบตัวแบบจากข้อมูลจริง โดยจะทำการเปรียบเทียบเว็บของตัวแบบที่สร้างขึ้นกับเว็บที่ถูกใช้จริงในวันถัดไป และคำนวณในรูปแบบของร้อยละ สามารถอธิบายโปรแกรมย่อยต่างๆ ได้ดังตาราง

ตารางที่ 32 คำอธิบายโปรแกรมย่อยโมดูลตรวจสอบความถูกต้อง

ชื่อฟอร์ม : frmCheModel	
Event Procedure / Sub / Function	คำอธิบาย
cmdchemodel_click	เมื่อกดปุ่มตรวจสอบโปรแกรมจะเรียบใช้สคริปต์ไฟล์ testmodel.sas
testmodel.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่ตรวจสอบเว็บที่ได้จากการสร้างตัวแบบกับเว็บที่ถูกเรียกใช้จริงในวันถัดมา
Percentcorrect.sas	เป็นสคริปต์ไฟล์ที่เขียนด้วยภาษา SAS มีหน้าที่คำนวณร้อยละของความถูกต้องของโมเดลที่สร้างขึ้น

ภาคผนวก ค

เอกสารการเผยแพร่ผลงานวิชาการ

ณ การประชุมวิชาการระดับชาติด้านเทคโนโลยีสารสนเทศ ครั้งที่ 1

วันที่ 2-3 พฤศจิกายน 2549

กรุงเทพมหานคร

การเพิ่มอัตราการพบในระบบพร็อกซีโดยใช้เทคนิคเหมืองข้อมูล

Hit Rate Improvement in Proxy System using Data Mining Technique

นางสาวพิจิตรา จอมศรี ผศ.ดร.ปานใจ ธารทัศนวงศ์
Pijitra Jomsri Asst.Prof.Dr.Panjai Tantasanawong
ภาควิชาคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร
noo_au9@hotmail.com panjai@su.ac.th

Abstract

This research applied data mining technique to improve hit rate in Proxy system by predicting the future web page access. Data sets from Silpakorn University Proxy servers were collected and investigated data relationship, which used to predict the future access websites. The association rule discovery is one of the data mining techniques and used in this research. The Result of research can increased hit Rate in Proxy servers By using this technique, the future accesses of websites were predicted and increased hit rate from 38.15% to 54.54%.If the hit rate in Proxy increases, the performance of Internet access will improve and reduce traffic in networks. However, this technique couldn't use for abnormal phenomenal such as protest, disaster or festival.

Keyword: *Proxy, Hit Rate Improvement, Data mining, Association rule discovery*

บทคัดย่อ

งานวิจัยนี้นำเทคนิคการทำเหมืองข้อมูล มาใช้ในการทำนายข้อมูลการเรียกใช้เว็บในอนาคต โดยการนำข้อมูลการใช้เว็บภายในมหาวิทยาลัยศิลปากร มาค้นหาความสัมพันธ์ที่ได้จากข้อมูลการใช้เว็บเพื่อใช้ทำนายเว็บที่ถูกเรียกใช้ในอนาคต งานวิจัยนี้ได้นำเทคนิคกฎการค้นหาความสัมพันธ์ ซึ่งเป็นเทคนิคหนึ่งในเหมืองข้อมูลมาประยุกต์ใช้ ผลของการใช้เทคนิคเหมืองข้อมูลพบว่าโมเดลที่สร้างขึ้นสามารถทำนายเนื้อหาเว็บที่จะถูกเรียกใช้ได้และสามารถเพิ่มอัตราการพบในระบบพร็อกซี่ ได้จากเดิม 38.15% เป็น 54.54 % ซึ่งถ้าอัตราการพบในระบบพร็อกซี่เพิ่มขึ้น อาจทำให้ประสิทธิภาพการเรียกใช้เว็บเพิ่มขึ้นและสามารถลดปริมาณข้อมูลในระบบเครือข่ายได้ อย่างไรก็ตามเทคนิคนี้ยังไม่สามารถครอบคลุมการทำงานในช่วงเหตุการณ์ที่ไม่เป็นปกติ เช่น อุบัติภัย และเทศกาลต่างๆ เป็นต้น

คำสำคัญ พร็อกซี่ การเพิ่มอัตราการพบ เหมืองข้อมูล การค้นหากฎความสัมพันธ์

1. ความเป็นมา

ในยุคปัจจุบันเครือข่ายอินเทอร์เน็ตถือเป็นเทคโนโลยีที่มีความสำคัญอย่างมากในการศึกษาค้นคว้าข้อมูล รวมทั้งการบริการต่างๆ เช่น เวิลด์ไวด์เว็บ (World Wide Web) หรือเรียกโดยย่อว่า เว็บ ซึ่งความก้าวหน้าทางเทคโนโลยีของเครือข่ายอินเทอร์เน็ต ทำให้ลดข้อจำกัดของเวลาและสถานที่ เช่น การศึกษาข้อมูลทางวิชาการ หรือการประกอบธุรกรรม และมีการใช้ในสถาบันการศึกษาที่มีการค้นคว้า วิจัยเป็นอย่างมาก การเรียกใช้เว็บ นั้นในองค์กรส่วนใหญ่มีการนำระบบพร็อกซี่ (Proxy) มาใช้เพื่อจัดเก็บข้อมูลรายละเอียดเว็บไซต์ที่เคยมีการเรียกใช้ไว้ที่ระบบ พร็อกซี่ หากมีผู้ใช้เรียกใช้เว็บเดิมนี้อีกก็จะทำการดึงข้อมูลเว็บที่อยู่ใน พร็อกซี่ มาให้กับผู้ใช้ซึ่งทำให้ผู้ใช้ได้รับข้อมูลเร็วกว่าการเรียกใช้งานตรงจากเว็บต้นฉบับ (Original Web) และยังเป็นการลดความหนาแน่นของระบบเครือข่ายได้เป็นอย่างดี แต่เนื่องจากพร็อกซี่มีการทำงานแบบ Cache คือมีประมาณของเว็บที่เก็บในหน่วยความจำของเครื่องแม่ข่ายพร็อกซี่จำนวนจำกัดไม่สามารถจะจัดเก็บเว็บทั้งหมดซึ่งมีจำนวนมากได้ จึงจำเป็นต้องมีการลบออก (page out) ซึ่งบางครั้งเว็บที่ลบ

ออกไปมีการเรียกใช้อีก ทำให้ต้องเสียเวลาในการไปดึงจากเว็บต้นฉบับมาใหม่ (page in) ทำให้ประสิทธิภาพการทำงานของพร็อกซี่ลดลง ซึ่งสามารถวัดได้จากอัตราการพบในพร็อกซี่ (Hit Rate)

ด้วยปัญหาดังกล่าวผู้วิจัยจึงได้จัดทำโครงการวิจัยขึ้นเพื่อเพิ่มอัตราการพบ โดยใช้เทคนิคเหมืองข้อมูลเพื่อทำนายเว็บเพจที่จะมีการเรียกใช้ในอนาคต ในบทความนี้จะประกอบด้วย 4 ส่วนคือ ส่วนที่ 1 จะเป็นความเป็นมาที่กล่าวถึงความสำคัญและปัญหา ส่วนที่ 2 วรรณกรรมที่เกี่ยวข้อง ส่วนที่ 3 ขั้นการดำเนินงานวิจัย ส่วนที่ 4 บทสรุปและข้อเสนอแนะ

2. วรรณกรรมที่เกี่ยวข้อง

พร็อกซี่ เซิร์ฟเวอร์(Proxy Server) หรือเรียกว่า แคช เซิร์ฟเวอร์ (Cache Server) คือการนำเครื่องคอมพิวเตอร์ที่ให้บริการแก่กลุ่มผู้ใช้ที่อยู่ในบริเวณเดียวกัน และกำหนดให้ผู้ใช้ทุกคนเรียกใช้ข้อมูลเว็บ ผ่านเครื่องคอมพิวเตอร์นี้ โดยเครื่องดังกล่าวจะมีการติดตั้งโปรแกรมเพื่อทำหน้าที่เรียกข้อมูลเว็บมาให้บริการแก่ผู้ใช้ และจัดเก็บข้อมูลที่เคยถูกเรียกนั้นไว้ ในเครื่อง เพื่อให้บริการแก่ผู้ใช้ข้อมูลนั้นซ้ำได้ทันทีโดยไม่ต้องเสียเวลาไปเรียกข้อมูลมาจากแหล่งข้อมูลใหม่ ซึ่งวิธีนี้ทำให้ผู้ใช้สามารถเรียกใช้ข้อมูลที่เคยมีผู้เรียกใช้มาก่อนได้รวดเร็วขึ้น เนื่องจากไม่ต้องเสียเวลาไปเรียกข้อมูลจากแหล่งข้อมูลใหม่ ทำให้ประสิทธิภาพในการใช้งานระบบเครือข่ายอินเทอร์เน็ต เพิ่มขึ้น

หลักการทำงานของพร็อกซี่ เซิร์ฟเวอร์ [3] คือ เมื่อมีผู้ใช้บริการทำการเรียกข้อมูลของเว็บไซต์ (Web Site) โดยผ่านพร็อกซี่ เซิร์ฟเวอร์ในครั้งแรก พร็อกซี่ เซิร์ฟเวอร์จะทำการตรวจสอบว่า มีข้อมูลของเว็บไซต์นั้นมีอยู่หรือไม่ หากพบว่าไม่มีข้อมูลพร็อกซี่ เซิร์ฟเวอร์จะทำการเรียกข้อมูลนั้นจากเว็บไซต์แล้วเก็บไว้ในเครื่อง และเมื่อมีผู้ใช้บริการทำการเรียกเว็บไซต์นี้อีกครั้งพร็อกซี่ เซิร์ฟเวอร์จะทำการส่งข้อมูลไปยังเครื่อง ของผู้ใช้บริการทันที ในกรณีที่ เว็บไซต์มีการปรับปรุงข้อมูลพร็อกซี่ เซิร์ฟเวอร์ จะทำการตรวจสอบข้อมูลที่มีอยู่ว่า ปรับปรุงหรือไม่ และจะทำการปรับปรุงข้อมูลใหม่ทันที ในกรณีที่มีผู้เรียกใช้บริการก็จะได้ข้อมูลที่ปรับปรุงอยู่เสมอ

ซึ่งกรณีที่ข้อมูลบนเว็บไซต์มีการปรับปรุงอยู่ตลอดเวลานั้น พร็อกซี่ เซิร์ฟเวอร์ จะต้องไปปรับปรุงข้อมูลจากเว็บเซิร์ฟเวอร์(Web Server) เนื่องจากไม่ได้เป็นข้อมูลที่ได้มาจากเว็บเซิร์ฟเวอร์ โดยตรงจึงอาจทำให้ผู้ใช้บริการได้รับข้อมูลที่ไม่เป็นปัจจุบัน หรือ เสียเวลาในการรอรับข้อมูลจากพร็อกซี่ เซิร์ฟเวอร์ ซึ่งต้องไปปรับปรุงข้อมูลจากเว็บเซิร์ฟเวอร์ ก่อนถึงจะนำข้อมูลมาให้บริการแก่ผู้ใช้ได้

ระบบเครื่องแม่ข่ายพร็อกซี่ที่นิยมใช้ [5] และมีความสามารถ คือ squid ซึ่งจะมีมาพร้อมกับลินุกซ์เซิร์ฟเวอร์ท โปรแกรม Squid เป็นพร็อกซี่แคช (Proxy Cache) ที่มีคุณสมบัติในการลดการเข้าถึงเว็บไซต์ภายนอกองค์กรได้เป็นอย่างดีและมีประสิทธิภาพ

งานวิจัยนี้ได้นำเทคนิคการทำเหมืองข้อมูล (Data Mining) [1, 2, 4, 5, 6, 8] ซึ่งเป็นการสำรวจและวิเคราะห์ข้อมูลที่มีขนาดใหญ่เพื่อค้นหาความสัมพันธ์และรูปแบบหรือกฎที่ซ่อนอยู่และนำความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆเพื่อนำมาสร้างระบบโดยสามารถหาแนวโน้มในอนาคต[9,10,11,12] เช่น การทำนายการใช้เว็บโดยใช้เทคนิคเหมืองข้อมูล เป็นต้น

เทคนิคเหมืองข้อมูลที่นำมาประยุกต์ใช้ในงานวิจัยนี้ คือ เทคนิคการค้นหากฎความสัมพันธ์ (Association rule discovery) ซึ่งเป็นการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่เพื่อช่วยในการวิเคราะห์และตัดสินใจ รูปแบบทั่วไปของการค้นหากฎความสัมพันธ์ คือ

$$A \rightarrow B$$

โดยที่ A : เป็นเงื่อนไข

หรือ LHS (Left - Hand Side)

B : เป็นผลลัพธ์ที่เกิดขึ้น

หรือ RHS (Right – Hand Side)

หรืออยู่ในรูปของ “ถ้า…….แล้ว” (If……Then….) เช่น

A→B ; if A Then B เป็นกฎที่ 1

B→A ; if B Then A เป็นกฎที่ 2

การประเมินค่าของกฎจะใช้ค่าสนับสนุน(Support) และค่าความเชื่อมั่น (Confidence) ค่าสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

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

โดยที่ A หมายถึง เหตุการณ์ที่ใช้เป็นเงื่อนไขในการหาผลลัพธ์

B หมายถึง เหตุการณ์ที่เป็นผลลัพธ์

Transaction(A,B) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A และ B

ค่าความเชื่อมั่น คือเปอร์เซ็นต์ของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมดที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

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

โดยที่ Transaction (A) หมายถึง เหตุการณ์ที่ประกอบด้วยเหตุการณ์ A อย่างเดียว

ในการเลือกว่าจะกฎใดนั้นจะต้องพิจารณาค่าสนับสนุน และค่าความเชื่อมั่นที่มีค่าสูงกว่าค่า Threshold ที่ตั้งไว้ นอกจากนี้จะต้องกำหนดค่าสนับสนุนต่ำสุด (Minimum Support) และค่าความเชื่อมั่นต่ำสุด (Minimum Confidence) โดยทั่วไปจะกำหนดค่าสนับสนุนต่ำสุดเป็น 5-10 % และค่าความเชื่อมั่นต่ำสุดเป็น 50-100%

3. ขั้นตอนการดำเนินการวิจัย

งานวิจัยนี้แบ่งขั้นตอนการวิจัยออกเป็น 3 ขั้นตอนคือ

1. การจัดเก็บข้อมูลและเตรียมข้อมูล
2. การสร้างตัวแบบ (Model)
3. นำตัวแบบที่ได้มาทดสอบและประยุกต์ใช้

3.1 การจัดเก็บและเตรียมข้อมูล

งานวิจัยนี้ได้ดำเนินการจัดเก็บข้อมูลการเข้าใช้เว็บของผู้ใช้ภายในมหาวิทยาลัยศิลปากร วิทยาเขตพระราชวังสนามจันทร์ จากเครื่องแม่ข่ายพร็อกซี่ ในเดือนสิงหาคม กันยายน ตุลาคม พ.ศ. 2548 เป็นเวลา 3 เดือนและจัดเก็บฐานข้อมูลการใช้เว็บและฐานข้อมูลหมวดของเว็บต่างๆ มาสร้างตัวแบบ ข้อมูลที่เตรียมสำหรับสร้างตัวแบบ ได้แก่ เวลา เนื้อหาเว็บ หมวดของเว็บ ซึ่งข้อมูลเวลาผู้วิจัยจะต้องทำการแปลงข้อมูลจาก Unix time stamp ให้เป็นรูปแบบปกติ (Date/Month/Year) และข้อมูลเนื้อหาเว็บจะใช้ในส่วนที่เป็นชื่อเว็บ (Universal Resource Locator, URL)

3.1.1 ข้อมูลจากพร็อกซี่ เซิร์ฟเวอร์

ตารางที่ 2 : แสดงตัวอย่างข้อมูลจากเครื่องแม่ข่ายพร็อกซี่

```
1128531640.658    148 202.44.135.35
TCP_REFRESH_HIT/200 504 GET
http://www.sanook.com/menu/images/sm2nbg.
gif - DIRECT/203.107.136.7 image/gif
1128531640.704    211 202.44.135.35
TCP_MISS/200 4809 GET
http://www.sanook.com/menu/nav.php -
TIMEOUT_DIRECT/203.107.136.7 text/html
1128531647.627    116 172.27.7.35
TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php?
- DIRECT/164.115.2.146 image/jpeg
1128531650.101    119 172.27.7.35
TCP_MISS/200 338 GET
http://truehits2.gits.net.th/biggen.php?
- DIRECT/164.115.2.146 image/jpeg
```

ข้อมูลที่ทำการจัดเก็บจากเครื่องแม่ข่ายพร็อกซี่เป็นข้อมูลของ access log ซึ่งข้อมูลเป็นชนิด text ไฟล์ ดังตารางที่ 3โดยประกอบด้วยข้อมูล 10 เขตข้อมูล (Field) ดังนี้

Time : เวลาที่ใช้ในการใช้งาน ซึ่งเป็น Unix time stamp

Duration : ระยะเวลาที่ใช้ในการทำงานของโปรแกรมต่อการร้องขอนั้น มีหน่วยเป็นมิลลิวินาที

Client address : IP address ของผู้ที่ร้องขอ

Result Code : เก็บผลการทำงานของการร้องแต่ละครั้ง

Byte : เก็บขนาดของข้อมูลทั้งหมดที่ถูกส่งไปยังผู้รับบริการ แต่ข้อมูลส่วนนี้ไม่ใช่ขนาดของข้อมูลที่แท้จริงเพราะจะนับรวมถึงส่วนของ header ด้วย

ตารางที่ 3 :แสดงตัวอย่างข้อมูลที่แปลงจาก text ไฟล์

time	duration	IP	Result_Code	byte	R_M	URL	rfc931	Hieraechy Code	type
1128531608	2113	202.44.135.35	TCP_MISS/200	49268	GET	http://spaces.msn.com/members/artsurapat/	-	DIRECT/65.54.153.254	text/html
1128531609	1955	202.44.135.35	TCP_MISS/200	46858	GET	http://spaces.msn.com/members/mayzai/	-	TIMEOUT_DIRECT/65.54.153.254	text/html

Request Mothod : เก็บข้อมูลรูปแบบของการร้องขอที่เกิดขึ้นกับข้อมูล

URL : เก็บที่อยู่ของข้อมูลที่ถูกร้องขอ

RFC931 : ทำหน้าที่เก็บลักษณะของเครื่องลูกข่ายที่ร้องขอข้อมูล

Hierarchy Code : เก็บข้อมูลที่เกี่ยวข้องกับการทำงานเป็นลำดับชั้นของโปรแกรม

Type :ประเภทของออบเจ็กต์ที่ถูกส่งมาให้ผู้ใช้

3.1.2 ข้อมูลหมวดเว็บ

รวบรวมรายละเอียดของหมวดเว็บต่างๆ จากเว็บที่มีการดำเนินการแบ่งหมวดเว็บไว้แล้ว เช่น google หรือ sanook เป็นต้น หรือจากเนื้อหาการให้บริการของแต่ละเว็บ ซึ่งสามารถแบ่งหมวดเว็บ ได้ดังตารางที่ 4

ตารางที่ 4: แสดงตัวอย่างหมวดเว็บ

	หมวดเว็บ
1	การแพทย์ และสุขภาพ
2	การศึกษา
3	กิจกรรม และเหตุการณ์สำคัญ
4	กีฬา
5	เกม
6	ข่าว และสื่อ
7	ความรู้ และข้อมูลสำคัญ
8	คอมพิวเตอร์
9	ชอปปิ้ง
10	ท่องเที่ยว
11	ธนาคาร และสถาบันการเงิน
12	ธุรกิจ
13	บันเทิงและนันทนาการ
14	บันเทิงและนันทนาการสำหรับผู้ใหญ่
15	บุคคล สังคม และวัฒนธรรม
16	ยานยนต์
17	วิทยาศาสตร์
18	หน่วยงานราชการและองค์กร
19	อสังหาริมทรัพย์ ก่อสร้าง และออกแบบตกแต่ง
20	อินเทอร์เน็ต

จากตารางที่ 3 สามารถคัดเลือกข้อมูลมาแสดงความสัมพันธ์ระหว่างเวลา หมวดของเว็บและเนื้อหาเว็บ

เวลา 17.00.00-18.00.00 น. , คอมพิวเตอร์ → www.pantip.com

ซึ่งจากความสัมพันธ์ข้างต้น สามารถอธิบายได้ดังนี้คือ ที่เวลา 17.00.00-18.00.00 น. หมวดคอมพิวเตอร์ มีแนวโน้มที่ผู้ใช้บริการจะเรียกใช้งานเว็บ www.pantip.com และเมื่อนำข้อมูลทั้งหมดมาหาความสัมพันธ์จะได้ค่าสนับสนุนและค่าความเชื่อมั่น ที่แตกต่างกันออกไป ดังตารางที่ 5 โดยเป็นตัวอย่างแสดงความสัมพันธ์ของเว็บที่ถูกเรียกใช้มากที่สุด 10 อันดับแรก ในช่วงเวลา 1 ชั่วโมง คือ เวลา 17.00.00-18.00.00 น. โดยไม่นำเว็บที่อยู่ในหมวดรวบรวมและค้นหาเว็บไซต์(search engine) และ เหตุการณ์ที่ผิดปกติ เช่น เหตุการณ์ทางการเมือง อุบัติภัย ฯลฯ ที่ทำให้การเรียกใช้เว็บ ผิดปกติ มาพิจารณาการหากฎความสัมพันธ์

ตารางที่ 5: แสดงตัวอย่างกฎความสัมพันธ์

กฎการวิเคราะห์ความสัมพันธ์			ค่าความเชื่อมั่น (%)	ค่าสนับสนุน (%)
เกม	⇒	www.thaibg.com	100	6.68
คอมพิวเตอร์	⇒	au.download.windowsupdate.com	55.27	7.95
คอมพิวเตอร์	⇒	msgr.dlservice.microsoft.com	44.73	6.43
บันเทิงและนันทนาการ	⇒	www.kapook.com	51.75	7.82
บันเทิงและนันทนาการ	⇒	www.yenta4.com	48.25	7.27
บันเทิงและนันทนาการสำหรับผู้ใหญ่	⇒	asclub.net	55.17	17.84
บันเทิงและนันทนาการสำหรับผู้ใหญ่	⇒	www.asclub.net	44.83	14.50
อินเตอร์เน็ต	⇒	rad.msn.com	48.19	15.18
อินเตอร์เน็ต	⇒	webboard.nisitgirl.com	28.19	8.88
อินเตอร์เน็ต	⇒	www.mthai.com	23.62	7.44

ค่าความเชื่อมั่นและค่าสนับสนุนสามารถหาจากสมการดังนี้

ค่าสนับสนุน คือ ร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนข้อมูลทั้งหมด สามารถเขียนเป็นสมการดังนี้

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

ค่าความเชื่อมั่น คือร้อยละของข้อมูลที่มีเงื่อนไขและผลลัพธ์สอดคล้องตามกฎต่อจำนวนรายการข้อมูลที่เป็นเงื่อนไข สามารถเขียนเป็นสมการดังนี้

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

กฎความสัมพันธ์ที่สร้างได้มีจำนวนมากดังนั้นจะต้องตัดความสัมพันธ์บางส่วนออกโดยการกำหนดค่าความเชื่อมั่นต่ำสุดและค่าสนับสนุนต่ำสุด

3.3 การนำตัวแบบที่ได้มาทดสอบและประยุกต์ใช้

3.3.1 ข้อมูล

งานวิจัยนี้ได้นำข้อมูลการเรียกใช้เว็บจากพร็อกซีเซิร์ฟเวอร์ มหาวิทยาลัยศิลปากร มาเพื่อทำนายการเรียกใช้งานเว็บโดยใช้ตัวแบบที่สร้างได้จากตารางที่ 5

ตารางที่ 6: แสดงตัวอย่างข้อมูล

เวลา	เนื้อหาเว็บ	หมวดเว็บ
17.00.00-18.00.00 น.	asclub.net	บันเทิงและนันทนาการสำหรับผู้ใหญ่
17.00.00-18.00.00 น.	rad.msn.com	อินเตอร์เน็ต
17.00.00-18.00.00 น.	www.asclub.net	บันเทิงและนันทนาการสำหรับผู้ใหญ่
17.00.00-18.00.00 น.	webboard.nisitgirl.com	อินเตอร์เน็ต
17.00.00-18.00.00 น.	au.download.windowsupdate.com	คอมพิวเตอร์
17.00.00-18.00.00 น.	www.kapook.com	บันเทิงและนันทนาการ
17.00.00-18.00.00 น.	www.mthai.com	อินเตอร์เน็ต
17.00.00-18.00.00 น.	www.yenta4.com	บันเทิงและนันทนาการ
17.00.00-18.00.00 น.	www.thaibg.com	เกม
17.00.00-18.00.00 น.	msgr.dlservice.microsoft.com	คอมพิวเตอร์
17.00.00-18.00.00 น.	www.pantip.com	คอมพิวเตอร์
17.10.00-18.00.00 น.	www.sanook.com	บันเทิงและนันทนาการ
17.10.00-18.00.00 น.	gateway.messenger.hotmail.com	อินเตอร์เน็ต
17.10.00-18.00.00 น.	www.googig.com	บุคคล สังคม และวัฒนธรรม
17.10.00-18.00.00 น.	…….	…….

3.3.2 การจำแนกข้อมูลจากกฎความสัมพันธ์

จากตารางที่ 6 ถ้าต้องการทำนายการเรียกใช้งานเว็บในช่วงเวลา 17.00.00 – 18.00.00 น. จะได้ความสัมพันธ์ตามตารางที่ 4 ซึ่งในกรณีที่พบว่ากฎที่เกิดขึ้น 2 กฎ ดังนั้นจะต้องกำหนดเกณฑ์ในการเลือกว่าจะใช้กฎความสัมพันธ์ใด โดยดูลำดับความสำคัญของเกณฑ์ดังนี้

1. พิจารณาจากค่าความเชื่อมั่นที่สูงสุดของแต่ละเงื่อนไข
2. ถ้าค่าความเชื่อมั่นเท่ากัน ให้พิจารณาค่าสนับสนุนที่สูงสุดของแต่ละเงื่อนไข
3. ถ้าค่าความเชื่อมั่นและค่าสนับสนุนมีค่าเท่ากันให้พิจารณากฎที่มาก่อนให้มีค่าความสำคัญมากกว่า

จากการพิจารณาตามเกณฑ์ที่กำหนดจะได้ว่าในช่วงเวลา 17.00.00 – 18.00.00 น. ผู้ใช้เว็บที่จะเรียกใช้เว็บในหมวด เกม มีแนวโน้มที่จะเรียกใช้เว็บ www.thaibg.com และเมื่อพิจารณาจากข้อมูลการเข้าเว็บทั้งหมดจะสามารถทำนายการเรียกใช้เว็บในอนาคตได้ และนำข้อมูลที่ได้มาใช้ประโยชน์ต่อไป

3.3.3 การทดสอบตัวแบบ

ตารางที่ 7: แสดงการทดสอบความถูกต้องของตัวแบบ

หมวดเว็บ	เนื้อหาเว็บ	ค่าความเชื่อมั่น (%)		ความถูกต้อง (True / False)
		โมเดลข้อมูลเรียนรู้	โมเดลข้อมูลตรวจสอบ	
เกม	www.thaibg.com	100	100	T
คอมพิวเตอร์	au.download.windowsupdate.com	55.27	54.53	F
คอมพิวเตอร์	msgr.dlservice.microsoft.com	44.73	45.47	T
บันเทิงและนันทนาการ	www.kapook.com	51.75	52.37	T
บันเทิงและนันทนาการ	www.yenta4.com	48.25	47.65	F
บันเทิงและนันทนาการสำหรับผู้ใหญ่	asclub.net	55.17	55.72	T
บันเทิงและนันทนาการสำหรับผู้ใหญ่	www.asclub.net	44.83	44.28	F
อินเตอร์เน็ต	rad.msn.com	48.19	48.84	T
อินเตอร์เน็ต	webboard.nisitgirl.com	28.19	27.51	F
อินเตอร์เน็ต	www.mthai.com	23.62	23.65	T

การตรวจสอบความถูกต้องของตัวแบบ ผู้วิจัยได้ทำการแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลการเรียนรู้ และข้อมูลตรวจสอบ ซึ่งการแบ่งข้อมูลจะใช้การเลือกตัวอย่างแบบสุ่มมาทำการค้นหากฎความสัมพันธ์ดังตารางที่ 7 ซึ่งผลลัพธ์ของการทดสอบตัวแบบจะแสดงเป็นค่า T และ F (โดยที่ T หมายถึง การทดสอบตัวแบบมีความถูกต้อง , F คือ การทดสอบตัวแบบมีความผิดพลาด)

จากลำดับที่ 1 สามารถอธิบายได้ว่าในหมวดเว็บ ซึ่งเป็นหมวดเกม ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการ 100 % ที่เข้าเว็บ www.thaibg.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 100 % ของผู้ใช้เว็บ มีผู้ใช้บริการในหมวดเกมที่เข้าใช้ www.thaibg.com ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากลำดับที่ 2 ในหมวดเว็บ ซึ่งเป็นหมวดคอมพิวเตอร์ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 55.27 % ที่เข้าเว็บ au.download.windowsupdate.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 54.53 % ของผู้ใช้เว็บ มีผู้ใช้บริการในหมวดคอมพิวเตอร์ที่เข้าใช้ au.download.windowsupdate.com ดังนั้นจึงมีความถูกต้องเท่ากับ F

จากลำดับที่ 3 ในหมวดเว็บ ซึ่งเป็นหมวดคอมพิวเตอร์ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 44.73 % ที่เข้าเว็บ msgr.dlservice.microsoft.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 45.47 % ของผู้ใช้เว็บ มีผู้ใช้บริการในหมวดคอมพิวเตอร์ที่เข้าใช้ msgr.dlservice.microsoft.com ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากลำดับที่ 4 ในหมวดเว็บ ซึ่งเป็นหมวดบันเทิงและนันทนาการ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการ 51.75 % ที่เข้าเว็บ www.kapook.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 52.37 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ www.kapook.com ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากลำดับที่ 5 ในหมวดเว็บ ซึ่งเป็นหมวดบันเทิงและนันทนาการ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 48.25 % ที่เข้าเว็บ www.yenta4.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 47.65 % ของผู้ใช้เว็บ มีผู้ใช้บริการในหมวดคอมพิวเตอร์ที่เข้าใช้ www.yenta4.com ดังนั้นจึงมีความถูกต้องเท่ากับ F

จากลำดับที่ 6 ในหมวดเว็บ ซึ่งเป็นหมวดบันเทิงและนันทนาการสำหรับผู้ใหญ่ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 55.17 % ที่เข้าเว็บ asclub.net เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 55.72 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ asclub.net ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากลำดับที่ 7 ในหมวดเว็บ ซึ่งเป็นหมวดบันเทิงและนันทนาการสำหรับผู้ใหญ่ ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 44.83 % ที่เข้าเว็บ www.asclub.net เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 44.28 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ www.asclub.net ดังนั้นจึงมีความถูกต้องเท่ากับ F

จากลำดับที่ 8 ในหมวดเว็บ ซึ่งเป็นหมวดอินเตอร์เน็ต ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 48.19 % ที่เข้าเว็บ rad.msn.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 48.84 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ rad.msn.com ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากลำดับที่ 9 ในหมวดเว็บ ซึ่งเป็นหมวดอินเตอร์เน็ต ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 28.19 % ที่เข้าเว็บ webboard.nisitgirl.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 27.51 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ webboard.nisitgirl.com ดังนั้นจึงมีความถูกต้องเท่ากับ F

จากลำดับที่ 10 ในหมวดเว็บ ซึ่งเป็นหมวดอินเตอร์เน็ต ของโมเดลข้อมูลเรียนรู้มีผู้ใช้บริการเว็บ 23.65 % ที่เข้าเว็บ www.mthai.com เมื่อนำข้อมูลตรวจสอบมาทดสอบจะพบว่า 23.65 % ของผู้ใช้เว็บ มีผู้ใช้บริการเว็บในหมวดคอมพิวเตอร์ที่เข้าใช้ www.mthai.com ดังนั้นจึงมีความถูกต้องเท่ากับ T

จากผลการดสอบตัวแบบพบว่าโมเดลการเรียนรู้มีความถูกต้องคิดเป็นร้อยละ 60 ของจำนวนรายการทั้งหมด

3.4 การเพิ่มอัตราการพบ (Hit rate)

หลังจากที่ได้มีการคาดการณ์ว่าจะมีเว็บใดบ้างที่จะถูกเรียกใช้ในช่วงเวลาต่างๆกันแล้ว ได้มีการประมาณค่า Hit rate ใหม่ ซึ่งโดยเฉลี่ยค่า Hit rate ของ พร็อกซี่ ที่มหาวิทยาลัยศิลปากรมีค่า 38.15 % เมื่อได้นำวิธีการเหมืองข้อมูลมาใช้ในการทำนายเว็บที่จะมีการเรียกใช้ผลปรากฏว่าค่า Hit rate ของระบบพร็อกซี่จะเพิ่มขึ้นเป็น 54.54 % (ค่า Hit rate ที่ได้จากการนำเทคนิคเหมืองข้อมูล มาจากการเปรียบเทียบเว็บที่ถูกจัดเก็บในระบบพร็อกซี่ด้วยวิธีการเหมืองข้อมูล กับ เว็บที่ถูกเรียกใช้จริงในช่วงเวลาเดียวกันของวันต่อมา โดยไม่นำเว็บที่เป็นsearch engine มาพิจารณา)และผลการวิจัยเบื้องต้นสามารถเพิ่มประสิทธิภาพการพบ (hit rate) ได้ถึง 42.96% ของร้อยละของการทำงานของระบบพร็อกซี่เดิม ซึ่งเป็นอัตราการพบดังกล่าวจะเป็นการเพิ่มประสิทธิภาพของพร็อกซี่ได้เป็นอย่างดี

4 บทสรุปและข้อเสนอแนะ

จากการใช้เทคนิคเหมืองข้อมูล โดยการวิเคราะห์ความสัมพันธ์เพื่อนำมาใช้ในการทำนายการใช้เว็บเพจพบว่าสามารถนำมาทำนายการเรียกใช้เว็บเพจได้ การทำนายการเรียกใช้งานเว็บในอนาคตมีประโยชน์อย่างยิ่งในการเพิ่ม Hit Rate ของระบบพร็อกซี่ได้เป็นอย่างมาก เมื่อ Hit rate สูงขึ้นขึ้นอย่างจะช่วยให้การใช้งานเว็บรวดเร็วขึ้นเพราะเรียกใช้จาก พร็อกซี่ โดยตรง อันจะเป็นการเพิ่มประสิทธิภาพของระบบ พร็อกซี่ และลดปริมาณการใช้งานเครือข่าย ทำให้ภาพรวมของการใช้งานเครือข่ายอินเตอร์เน็ตดีขึ้น การทำงานโดยใช้เทคนิคเหมืองข้อมูลในงานวิจัยนี้ความถูกต้องจะขึ้นอยู่กับการสร้าง ตัวแบบและข้อมูลของพร็อกซี่ ที่นำมาสร้างความสัมพันธ์

งานวิจัยนี้เป็นการวิจัยเชิงวิเคราะห์และการจำลองสถานการณ์ (Simulation) หากจะนำวิธีการนี้ไปใช้จริงจำเป็นจะต้องมีการพัฒนาซอฟต์แวร์เพิ่มเติมและเสริมเข้าไปในระบบพร็อกซี่ ซึ่งถ้าระบบ พร็อกซี่ เป็นระบบเปิดเผยรหัส (Open Source System) ก็สามารถจะแก้ไขได้ทันที

เอกสารอ้างอิง

[1] กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ, และธนาวินท์ รักธรรมานนท์. (2001) “การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษานิสิตคณะวิศวกรรมศาสตร์”. *The Necptec Technical Journal* v.3, no.11 : 134-142.

[2] กฤษณะ ไวยมัย และธีระวัฒน์ พงษ์ศิริปรีดา.(2001) “การใช้เทคนิค Association Rule Discovery เพื่อการจัดสรรกฎหมายการพิจารณาคดีความ”. *The Nectec Technical Journal* v.3, no.11 : 143-152.

[3] ศูนย์คอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีสุรนารี.(2005) “เลือกใช้ พร็อกซี่ อย่างไรให้มีประสิทธิภาพ” [www document] http://www.sut.ac.th/ccs/news/tip_tech/tip002.asp (Accessed 5 Novembe 2005).

[4] บุญเสริม กิจศิริกุล.(2546) “รายงานฉบับสมบูรณ์ โครงการวิจัยร่วมภาครัฐและเอกชน ปีงบประมาณ 2545 โครงการย่อยที่ 7 อัลกอริทึมการทำเหมืองข้อมูล”.

[5] ธีรภัทร มนตรีศาสตร์.“Squid Proxy Caching Server”. [www document]. http://micro.se-ed.com/content/mc205/MC205_181.asp (Accessed 5 November 2005).

[6] Berson, Alex and Smith, Stephen J. (1997) “ Data Warehousing, Data Mining and OLAP”, Singapore :McGraw Hill.

[7] Dick Ng’ambi. (2002) “Pre_empting User Questions through Anticipation – Data Mining FAQ Lists”. *Preceedings of SAICSIT 2002*, p.101-109.

[8] Han, Jiawei and Kamber, Micheline.(2001) “Data Mining Concepts and Techniques” ,USA :Morgan Kaufman.

[9] Viveros, Marisa S., Nearhos, John P. and Rothman, Michale J. (1996)“Applying Data Mining Techniques to a Health Insurance Information System”. *Proceedings of the 22nd VLDB Conference Mumbai(Bombay)*, India.

[10] Mohan, Sujaa Rani, Park E.K. ,and Yijie Han. (2005) “ Association Rule Based Data Mining Agents for Personalized Web Caching” , *roceeding of the 29th Annual International Computer Software and Applications Conference (COMPSAC’05)* , Kansas City.

[11] Wu, Yi-Hung and ArbeeL., P.Chen . (2002) “ Prediction of Web Page Accesses by Proxy Server Log”, World Wide Web:Internet and Web Information System ,5,67-68.

[12] Yang,Qiang, Hui Wang, and Wei Zhang.(2002) “Web-log Mining for Quantitative Temporal-Event Prediction”, *IEEE Computational Intelligence Bulletin*,Vol.1 No.1, Decenber.

ประวัติผู้วิจัย

ชื่อ-สกุล นางสาวพิจิตรา จอมศรี
ที่อยู่ 283/16 ศรีเผื่อน ถนนริมคลองประปาฝั่งซ้าย บางซื่อ กทม. 10800
ที่ทำงาน สำนักงานคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา เขตดุสิต กรุงเทพ 10300

ประวัติการศึกษา

พ.ศ. 2545 สำเร็จการศึกษาปริญญาบัณฑิต สาขาวิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยธรรมศาสตร์

พ.ศ. 2546 ศึกษาต่อระดับปริญญามหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยศิลปากร

ประวัติการทำงาน

พ.ศ. 2545-ปัจจุบัน เจ้าหน้าที่ฝ่ายสารสนเทศและฝึกประสบการณ์วิชาชีพ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏสวนสุนันทา