บทที่ 4

การพัฒนาแม่แบบ

4.1 บทนำ

สำหรับบทนี้จะกล่าวถึงการพัฒนาแม่แบบตามที่ออกแบบในบทที่ 3 โดยจะกล่าวถึงภาษาที่ใช้ในการพัฒนาซึ่งอธิบายรายละเอียดในหัวข้อ 4.2 และเครื่องมือที่ใช้ในการวิจัย เนื่องจากงานวิทยานิพนธ์ที่นำเสนอเป็นการเสนอแนวคิดเพื่อปรับปรุงประสิทธิภาพการทำงานของเว็บแคชชิง โดยนำเสนอแม่แบบการทำงานจึงไม่เน้นการพัฒนาโปรแกรมประยุกต์เพื่องานใดงานหนึ่งโดยเฉพาะ แต่จะใช้เครื่องมือที่มีอยู่หรือมีการพัฒนาขึ้นมาเองบางส่วนในการทำงานเพื่อเสนอให้เห็นภาพการทำงาน และแสดงให้เห็นว่าแม่แบบที่นำเสนอสามารถใช้งานได้จริง เครื่องมือสำหรับงานวิจัยนี้ประกอบด้วย โปรแกรม webmining เป็นโปรแกรมที่ทำงานในหน่วยการทำเหมืองข้อมูลบันทึกการใช้งานเว็บซึ่งได้อธิบายการทำงานและการพัฒนาในหัวข้อ 4.3 ในหัวข้อ 4.4 อธิบายการพัฒนาโปรแกรม Jproxy ทำงานในหน่วยการทำงานเว็บแคชชิง หัวข้อ 4.5 อธิบายการใช้งานโปรแกรม Wget ที่ใช้ทำหน้าที่ร้องขอข้อมูลล่วงหน้าเป็นส่วนหนึ่งของหน่วยการดึงข้อมูลล่วงหน้า และหัวข้อ 4.6 อธิบายการพัฒนาโปรแกรม GetURL ที่ใช้สำหรับทดสอบการทำงานของระบบซึ่งไม่ได้ทำงานอยู่ในหน่วยการทำงานใด แต่เป็นเครื่องมือที่มีส่วนสำคัญเพื่อใช้ในการทดสอบระบบ สำหรับหัวข้อ 4.7 เสนอบทสรุปของการพัฒนาแม่แบบ

4.2 ภาษาที่ใช้ในการพัฒนา

การพัฒนาโปรแกรมเพื่อใช้ในงานวิทยานิพนธ์ในส่วนของการเตรียมข้อมูลพัฒนาโดยใช้ภาษา Perl เนื่องจากเป็นภาษาที่ใช้จัดการข้อมูลประเภทข้อความได้อย่างมีประสิทธิภาพ เนื่องจากข้อมูลหลักที่ใช้ในงานวิจัยเป็นข้อมูลที่บันทึกในรูปของข้อความในแฟ้มข้อมูล สำหรับภาษาที่ใช้พัฒนาเครื่องมืออีกตัวคือ ภาษาจาวา (JAVA) ซึ่งใช้พัฒนาโปรแกรมเว็บแคชชิงที่ใช้ในงานวิทยานิพนธ์ และโปรแกรมร้องขอข้อมูลเว็บสำหรับการทดสอบการทำงาน

4.3 การพัฒนาโปรแกรมส่วนการทำเหมืองข้อมูลบันทึกการเข้าใช้งาน

โปรแกรมการทำงานในส่วนนี้เป็นส่วนการทำนายกฎความเชื่อมโยงเรียกว่า webmining โดยมีการทำงานย่อย 2 ขั้นตอนคือการเตรียมข้อมูล และการสร้างกฎ ซึ่งกระบวนการเตรียมข้อมูลในงานวิทยานิพนธ์นี้จะพัฒนาโปรแกรมเพื่อใช้งานเองชื่อ preparedata สำหรับการทำงานในการสร้างกฎจะใช้โปรแกรมชื่อ apriori จากงานวิจัยของ Borgelt [Borgelt, 2002] สำหรับดำเนินการเรียกใช้โปรแกรม webmining ด้วย คำสั่งดังนี้

```
webmining.pl rawdataname transactionfilename outputfilename
|-----(1)------|-------(2)------|---------(3)----------|-------(4)------|
```

โดยรายละเอียดของคำสั่งประกอบด้วย

(1) webmining.pl คือ ชื่อของโปรแกรมใช้สำหรับเรียกการทำงานเพื่อทำนายคำร้องขอในอนาคต

(2) rawdataname คือชื่อแฟ้มข้อมูลบันทึกการเข้าใช้งานที่บันทึกโดยโปรแกรมเว็บแคชชิง

(3) transactionfilename คือชื่อแฟ้มข้อมูลของข้อมูลที่ผ่านการเตรียมข้อมูลให้เหมาะสำหรับการหากฎความเชื่อมโยง

(4) outputfilename คือชื่อแฟ้มข้อมูลสำหรับบันทึกกฎความเชื่อมโยงที่ได้จากการทำนายด้วยขั้นตอนวิธี apriori

ขั้นตอนการทำงานของคำสั่งข้างต้นประกอบด้วย โปรแกรมทำการอ่านข้อมูลจากบันทึกการเข้าใช้งานของเว็บแคชชิงหลังจากนั้นนำข้อมูลผ่านกระบวนการเตรียมข้อมูลด้วยโปรแกรม preparedata ซึ่งจะอธิบายการพัฒนาโปรแกรมนี้ในหัวข้อที่ 4.3.1 และใช้ข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลนี้หากฎความเชื่อมโยงโดยอธิบายการใช้งานโปรแกรม apriori เพื่อสร้างกฎในหัวข้อที่ 4.3.2

4.3.1 การพัฒนาโปรแกรมส่วนการเตรียมข้อมูล

การพัฒนาโปรแกรม prepapredata ใช้สำหรับเตรียมข้อมูลมีการทำงานในโปรแกรมดังนี้

- Clean เป็นส่วนการทำงานที่ทำหน้าที่อ่านข้อมูลดิบจากบันทึกการเข้าใช้งานเว็บแล้วคัดเฉพาะข้อมูลที่ใช้วิเคราะห์บันทึกลงในแฟ้มข้อมูล
- Group ทำหน้าที่รวมข้อมูลบันทึกการใช้งานเว็บที่มีหมายเลขไอพีเดียวกันเป็นกลุ่มเดียวกัน
- Cluster ทำหน้าที่ระบุผู้ใช้งานโดยพิจารณาเปรียบเทียบช่วงเวลาของการร้องขอข้อมูลของผู้ใช้
- Extract ทำหน้าที่สกัดข้อมูลเพื่อสร้างเป็นรายการการดำเนินการสำหรับนำไปสร้างกฎความเชื่อมโยง

รายละเอียดของการพัฒนาโปรแกรมการทำงานทั้ง 4 ขั้นตอนอธิบายในหัวข้อที่ 4.3.1.1, 4.3.1.2, 4.3.1.3 และ 4.3.1.4 ตามลำดับดังนี้

4.3.1.1 *ฟังก์ชันการทำความสะอาด (Clean)* ทำหน้าที่อ่านข้อมูลจากบันทึกการเข้าใช้งานเว็บแล้วคัดเอาเฉพาะข้อมูลที่ต้องการโดยอาศัยคำสั่งของโปรแกรม awk ซึ่งเป็นโปรแกรมที่มีใช้งานบนระบบปฏิบัติการยูนิกซ์ บันทึกผลลัพธ์จากการคัดข้อมูลลงในแฟ้มข้อมูล ซึ่งข้อมูลที่ผ่านการทำความสะอาดจะนำไปใช้ต่อในส่วนการทำงานการจัดกลุ่ม ตัวอย่างการทำงานเช่นจากข้อมูลดิบตามภาพประกอบ 4-1 เมื่อผ่านกระบวนการทำความสะอาดข้อมูลผลที่ได้แสดงดังภาพประกอบ 4-2

```
1124070108.630 6 172.30.3.54 TCP_IMS_HIT/304 218 GET http://www.startnow.com/ieb/res/navhlp-config.xml - NONE/- application/xml
1124070415.377 0 172.30.3.54 TCP_IMS_HIT/304 218 GET http://www.startnow.com/ieb/res/navhlp-config.xml - NONE/- application/xml
1124070795.026 1760 192.168.2.42 TCP_MISS/200 1123 GET http://tcruskit.telstra.net/cgi-bin/trace - DIRECT/203.50.1.77 text/html
1124070796.923  1791 192.168.2.42 TCP_MISS/200 581 GET http://www.telstra.net/bpgstr_u.gif - DIRECT/203.50.5.178 text/html
1124070796.923  1760 192.168.2.42 TCP_MISS/200 577 GET http://www.telstra.net/bpgstr.gif - DIRECT/203.50.5.178 text/html
1124070798.180 1230 192.168.2.42 TCP_MISS/200 482 GET http://tcruskit.telstra.net/bpgstr_u.gif - DIRECT/203.50.1.77 text/html
1124070798.180   1244 192.168.2.42 TCP_MISS/200 480 GET http://tcruskit.telstra.net/bpgstr.gif - DIRECT/203.50.1.77 text/html
1124070799.506   1260 192.168.2.42 TCP_MISS/404 481 GET http://tcruskit.telstra.net/favicon.ico - DIRECT/203.50.1.77 text/html
```

ภาพประกอบ 4-1 ตัวอย่างข้อมูลดิบจากบันทึกการเข้าใช้งานเว็บของ cache.psu.ac.th

```
1124070796.923  192.168.2.42  http://www.telstra.net/bpgstr_u.gif
1124070796.923  192.168.2.42  http://www.telstra.net/bpgstr.gif
1124070798.180  192.168.2.42  http://tcruskit.telstra.net/bpgstr_u.gif
1124070798.180  192.168.2.42  http://tcruskit.telstra.net/bpgstr.gif
```

ภาพประกอบ 4-2 ตัวอย่างข้อมูลที่ผ่านขั้นตอนการทำความสะอาดข้อมูล

4.3.1.2 *ฟังก์ชันการจัดกลุ่ม (Group)* โปรแกรมในส่วนนี้ทำหน้าที่รวมกลุ่มข้อมูลให้เป็นชุดเดียวกัน โดยจัดกลุ่มตามไอพี ซึ่งทำการเปรียบเทียบข้อมูลคำร้องขอที่มีไอพีเดียวกันเก็บรวบรวมและบันทึกในแฟ้มข้อมูลเดียวกัน

4.3.1.3 *ฟังก์ชันการระบุผู้ใช้ (Cluster)* ทำหน้าที่ระบุผู้ใช้จากกลุ่มผู้ใช้ที่มีหมายเลขไอพีเดียวกัน ซึ่งแบ่งแยกการทำงานของผู้ใช้ด้วยช่วงเวลาของการใช้งาน (Time gap: Δt) ซึ่งเท่ากับ 73 นาที โดยค่าตัวเลขที่ได้มาจากการวิเคราะห์ในหัวข้อที่ 4.3.2.1 กล่าวคือถ้าคำร้องขอของผู้ใช้ที่ลำดับติดกันมีระยะเวลาความห่างเกินกว่าค่าที่กำหนดจะพิจารณาว่า คำร้องขอ ณ. ตำแหน่งนั้นเป็นต้นไปเป็นการร้องขอของผู้ใช้รายอื่น ทำการแบ่งและบันทึกข้อมูลลงแฟ้มข้อมูล

4.3.1.4 *ฟังก์ชันการสกัดข้อมูล (Extract)* ทำหน้าที่แปลงข้อมูลให้เหมาะกับการวิเคราะห์ โดยนำข้อมูลของแต่ละกลุ่มผู้ใช้จากกระบวนการการระบุผู้ใช้รวบรวมเป็นแฟ้มข้อมูลเดียวกัน ซึ่งทำการคัดข้อมูลเวลาออกไปเนื่องจากไม่ใช้ในการวิเคราะห์ และให้หมายเลขคำร้องขอเพื่อสร้างรายการการดำเนินของผู้ใช้รายหนึ่ง ตัวอย่างการทำงานเช่น จากตัวอย่างข้อมูลในภาพประกอบ 4-2 เมื่อผ่านกระบวนการจัดกลุ่มและระบุผู้ใช้จะได้ข้อมูลบันทึกในแฟ้มข้อมูลเดียวกันเนื่องจากเป็นข้อมูลที่มีหมายเลข IP เดียวกันและเวลาในการร้องขอต่อเนื่องจึงเป็นการดำเนินการของผู้ใช้รายหนึ่งตามสมมติฐานที่ได้อธิบายในบทที่ 3 ดังนั้นเมื่อนำข้อมูลข้างต้นมาสกัดเพื่อให้เหลือเฉพาะข้อมูลที่จะนำไปวิเคราะห์ด้วยขั้นตอนวิธี Apriori จะได้ตัวอย่างรายการการดำเนินการของผู้ใช้รายหนึ่ง แสดงดังภาพประกอบ 4-3

```
15  http://www.telstra.net/bpgstr_u.gif
15  http://www.telstra.net/bpgstr.gif
15  http://tcruskit.telstra.net/bpgstr_u.gif
15  http://tcruskit.telstra.net/bpgstr.gif
```

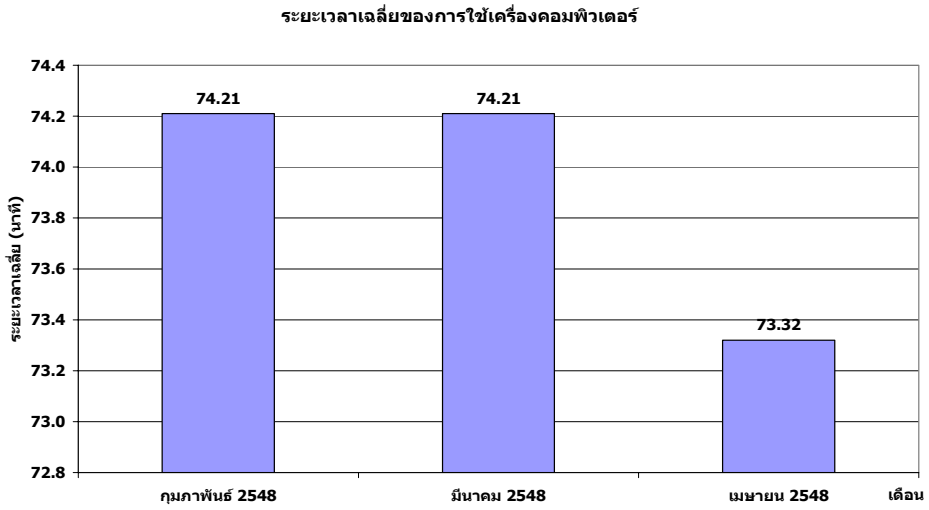
ภาพประกอบ 4-3 ตัวอย่างข้อมูลที่ผ่านการแปลงเพื่อนำไปสร้างกฎความเชื่อมโยง

4.3.2 การพัฒนาโปรแกรมส่วนการสร้างกฎความเชื่อมโยง

การสร้างกฎความเชื่อมโยงใช้ข้อมูลสองชนิดในการทำงานคือ ข้อมูลบันทึกการเข้าใช้เครื่องคอมพิวเตอร์ของศูนย์คอมพิวเตอร์มหาวิทยาลัยสงขลานครินทร์ ใช้เพื่อวิเคราะห์ค่าเฉลี่ยช่วงเวลาการใช้งานของผู้ใช้และข้อมูล web access log ของ cache.psu.ac.th สำหรับการสร้างกฎความเชื่อมโยง ซึ่งอธิบายการดำเนินการในแต่ละส่วนดังนี้

4.3.2.1 การวิเคราะห์หาค่าเฉลี่ยช่วงเวลาการใช้งาน

การทดสอบใช้ข้อมูลบันทึกการเข้าใช้งานเครื่องคอมพิวเตอร์ของศูนย์คอมพิวเตอร์มหาวิทยาลัยสงขลานครินทร์ในระยะเวลา 3 เดือนคือ เดือนกุมภาพันธ์ เดือนมีนาคม และเดือนเมษายน พ.ศ. 2548 การที่เลือกรวบรวมข้อมูลในช่วงเวลาข้างต้นมาใช้เนื่องจากเป็นช่วงเวลาที่สามารถนำมาเป็นตัวแทนการใช้งานเครื่องคอมพิวเตอร์ของผู้ใช้โดยภาพรวมได้ เพราะมีลักษณะของการทำงานครอบคลุมทั้งช่วงที่มีการใช้งานมากคือ เดือนกุมภาพันธ์ซึ่งเป็นระยะเวลาที่มีการเรียนการสอนปกติ และเดือนมีนาคมที่มีการใช้งานลดลงเพราะเป็นช่วงของการสอบปลายภาคและเริ่มปิดภาคเรียน สำหรับเดือนเมษายนก็จะเป็นช่วงเวลาที่ปิดภาคเรียนของภาคเรียนปกติ แต่เริ่มมีการเรียนการสอนของภาคการเรียนฤดูร้อนซึ่งจะมีจำนวนผู้ใช้งานเครื่องคอมพิวเตอร์เพิ่มขึ้นแต่อาจไม่สูงเท่าช่วงเวลาเปิดเรียนตามปกติ นำข้อมูลทั้ง 3 เดือนมาวิเคราะห์หาระยะเวลาเฉลี่ยการใช้งานเครื่องคอมพิวเตอร์ของผู้ใช้รายหนึ่งๆ ในแต่ละเดือน ผลที่ได้แสดงดังภาพประกอบ 4-4



ภาพประกอบ 4-4 กราฟแสดงค่าเฉลี่ยระยะเวลาการใช้งานคอมพิวเตอร์ของผู้ใช้ในแต่ละเดือน

เพราะฉะนั้นระยะเวลาเฉลี่ยของการใช้งานจะได้ว่า (74.21+74.21+73.32)/3 เท่ากับ 73 นาที ซึ่งค่าเฉลี่ยที่ได้นำไปใช้ในการวิเคราะห์เพื่อแบ่งช่วงการดำเนินการของสำหรับระบุผู้ใช้ในขั้นตอนการเตรียมข้อมูล

4.3.2.2 การสร้างกฎความเชื่อมโยง

สร้างกฎความเชื่อมโยงจากข้อมูลที่ได้ในขั้นตอนการเตรียมข้อมูล แต่เนื่องจากการหากฎความเชื่อมโยงจากข้อมูล URL ที่เป็นตัวอักษรทำให้เสียเวลาและพื้นที่บันทึกกฎสำหรับการทำงาน เพราะฉะนั้นเพื่อช่วยให้สามารถทำการสร้างกฎได้ดีจะทำการเข้ารหัสข้อมูล URL เป็นตัวเลข ตัวอย่างเช่นจากข้อมูลในภาพประกอบ 4-4 ทำการเข้ารหัสข้อมูล URL ส่วนหมายเลขการดำเนินการยังคงเดิมบันทึกข้อมูลลงแฟ้มข้อมูล โดยข้อมูลในเขตข้อมูลแรกหมายถึงลำดับเลขของการดำเนินการเพื่อใช้สำหรับแปลงเป็นรายการการดำเนินการ และเขตข้อมูลที่สองคือURL ที่มีการเข้ารหัสเป็นตัวเลข โดยทำการกำหนดค่าตัวเลขให้ URL แต่ละตัวด้วยโปรแกรมการเข้ารหัสแสดงผลที่ได้จากการทำงานดังภาพประกอบ 4-5

```
15  731
15  732
15  5827
15  5828
```

ภาพประกอบ 4-5 ตัวอย่างข้อมูลที่เข้ารหัส URL เพื่อใช้ในการหากฎความเชื่อมโยง

นำข้อมูลที่เข้ารหัสทั้งหมดแปลงให้อยู่ในรูปของรายการการดำเนินการ โดยการคัดข้อมูลที่มีหมายเลขการดำเนินการเดียวกันรวมข้อมูลเป็นกลุ่มเดียวกันและบันทึกข้อมูลแต่ละการดำเนินการ รายละเอียดของรหัสโปรแกรมดูได้จากภาคผนวก ค ตัวอย่างรายการการดำเนินการจากข้อมูลที่เข้ารหัสแสดงดังภาพประกอบ4-6 ซึ่งแต่ละบรรทัดจะหมายถึงรายการการดำเนินการร้องขอข้อมูลของผู้ใช้รายหนึ่ง

```
140 146 141 141 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
170 171
200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225
24 25 40 43 44 45 46 47 48 49 50 51 54 55 57
1112 1113 1114 1115 1116 1117 1118 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1130 1131
731 732 5827 5828
```

ภาพประกอบ 4-6 ตัวอย่างข้อมูลเข้ารหัส URL ที่แปลงให้เป็นรายการการดำเนินการ

สร้างกฎความเชื่อมโยงจากข้อมูลเข้ารหัสที่ค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุดที่แตกต่างกันเพื่อหาค่าสนับสนุนและค่าความเชื่อมั่นต่ำสุดที่เหมาะสำหรับการทำงานของแม่แบบ MineCache โดยเรียกใช้โปรแกรม apriori ซึ่งมีรูปแบบของคำสั่งในการเรียกใช้งานดังนี้

```
apriori -s5 -c60 convert_filename coderule_filename
|-(1)-|-(2)-|-(3)-|----------(4)-------|---------(5)---------|
```

โดยรายละเอียดของคำสั่งประกอบด้วย

(1) apriori คือชื่อโปรแกรมสำหรับการเรียกใช้งานโปรแกรม

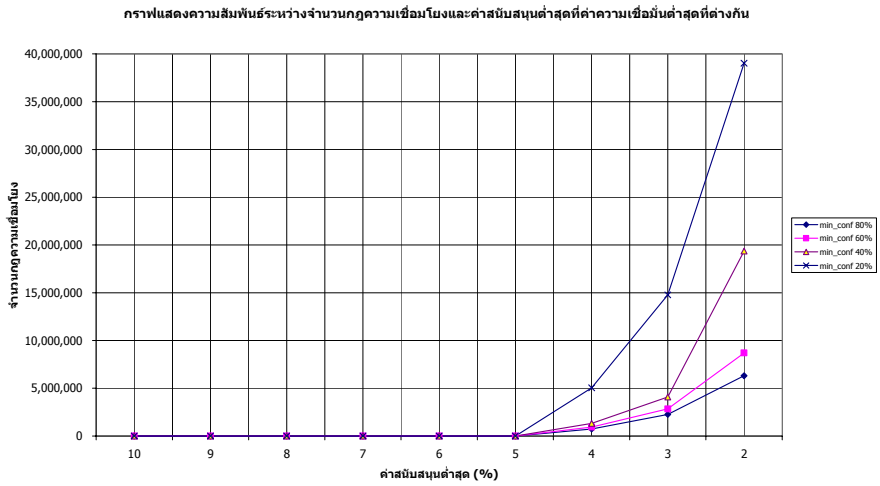
(2) –s5 คือตัวเลือกการทำงานเพื่อกำหนดค่าสนับสนุนต่ำสุดสำหรับหากฎความเชื่อมโยง โดยพิมพ์สัญลักษณ์ “-” ตามด้วยตัวอักษร s (support) และระบุค่าสนับสนุนที่ต้องการซึ่งมีหน่วยเป็นเปอร์เซ็นต์ จากตัวอย่าง –s5 หมายถึง กำหนดให้ค่าสนับสนุนต่ำสุดเท่ากับ 5 เปอร์เซ็นต์

(3) –c60 คือตัวเลือกการทำงานเพื่อกำหนดค่าความเชื่อมั่นต่ำสุดสำหรับหากฎความเชื่อมโยง โดยพิมพ์สัญลักษณ์ “-” ตามด้วยตัวอักษร c (confident) และระบุค่าความเชื่อมั่นที่ต้องการซึ่งมีหน่วยเป็นเปอร์เซ็นต์ จากตัวอย่าง –c60 หมายถึง กำหนดให้ค่าความเชื่อมั่นต่ำสุดเท่ากับ 60 เปอร์เซ็นต์

(4) convert_filename คือชื่อแฟ้มข้อมูลที่บันทึกข้อมูลรายการการดำเนินการสำหรับหากฎความเชื่อมโยง

(5) coderule_filename คือชื่อแฟ้มข้อมูลสำหรับบันทึกกฎความเชื่อมโยงที่ทำนายได้จากโปรแกรม apriori

จำนวนกฎความเชื่อมโยงที่ได้จากการทำงานของโปรแกรมสำหรับค่าสนับสนุนต่ำสุดและค่าความเชื่อมั่นต่ำสุดที่แตกต่างกันแสดงดังภาพประกอบ 4-7



ภาพประกอบ 4-7 กราฟความสัมพันธ์ระหว่างจำนวนกฎที่ได้กับค่าสนับสนุนและค่าความเชื่อมั่นต่ำสุดที่แตกต่างกัน

จากกราฟพบว่าจำนวนกฎความเชื่อมโยงที่ได้เมื่อค่าสนับสนุนต่ำสุดเท่ากับ 5 เปอร์เซ็นต์ และค่าความเชื่อมั่นเท่ากับ 60 เปอร์เซ็นต์ จะไม่แตกต่างกัน เพราะฉะนั้นในงานวิทยานิพนธ์ที่นำเสนอจะใช้ค่าสนับสนุนต่ำสุดเท่ากับ 5 เปอร์เซ็นต์ และค่าความเชื่อมั่นต่ำสุด 60 เปอร์เซ็นต์ สำหรับการสร้างกฎความเชื่อมโยง โดยกฎความเชื่อมโยงจะอยู่ในรูปแบบของ ถ้า...และ... และ ... แล้ว... (*if this and this and this then this*) ซึ่งกฎที่ได้ยังเป็นข้อมูลเข้ารหัสที่เป็นตัวเลข ดังนั้นต้องทำการแปลงกลับมาเป็นข้อมูลปกติที่เป็นข้อความอ่านเข้าใจได้ซึ่งรายละเอียดของรหัสโปรแกรมดูได้จากภาคผนวก ค ตัวอย่างของกฎความเชื่อมโยงที่ได้แสดงดังภาพประกอบ 4-8

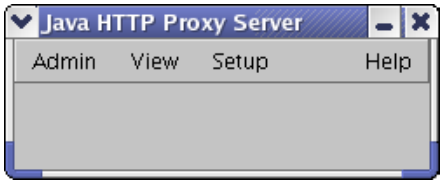
ค่าความเชื่อมั่นต่ำสุด (%)	กฎความเชื่อมโยง
78.8%	http://us.i1.yimg.com/us.yimg.com/i/tv/mischabarton104d.jpg → http://www.yahoo.com/
67.2%	http://view.atdmt.com/MSN/iview/msnnkhac001728x90Xhotj4300140msn/direct/01→ http://global.msads.net/ads/8013/0000008013_00000000000000000195503.jpg

ภาพประกอบ 4-8 ตัวอย่างกฎความเชื่อมโยง

จากตัวอย่างที่เสนอในภาพประกอบ4-8 กฎที่ได้จะมีความหมายว่าถ้าผู้ใช้เว็บมีการร้องขอข้อมูล http://us.i1.yimg.com/us.yimg.com/i/tv/avrillavigne104d.jpg แล้วมีความเป็นได้ 78.8 เปอร์เซ็นต์ที่จะมีการร้องขอ http://www.yahoo.com/ ด้วย สำหรับการแปลความในแถวที่สองก็จะเป็นเช่นเดียวกับแถวที่หนึ่ง

4.4 การพัฒนาโปรแกรมส่วนการทำงานเว็บแคชชิง

ในส่วนการทำงานเว็บแคชชิงจะใช้โปรแกรม Jproxy [Hsueh, 1997] พัฒนาด้วยภาษาจาวา ซึ่งเป็นโปรแกรมแม่แบบที่สามารถรองรับการทำงานกับผู้ใช้บริการจำนวนหนึ่ง โดยทำการปรับปรุงโปรแกรม Jproxy เพิ่มเติมเพื่อให้ทำงานรองรับการทำการทำเหมืองข้อมูลเว็บ โปรแกรม Jproxy ทำหน้าที่เป็นทั้งเว็บแคชชิงเพื่อบันทึก web object ที่ร้องขอและ proxy สำหรับการเป็นตัวกลางการติดต่อระหว่างเครือข่ายภายในการทำงานและเครือข่ายของภาควิชาวิทยาการคอมพิวเตอร์โปรแกรมประกอบด้วยเมนูหลักสำหรับการสั่งงานทั้งหมด 4 เมนูดังภาพประกอบ 4-9



ภาพประกอบ 4-9 หน้าจอหลักของโปรแกรม Jproxy

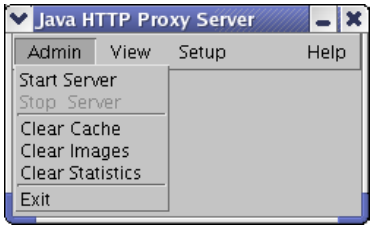
โปรแกรม Jproxy มีการทำงานแบบ Graphic User Interface (GUI)[3] โดยเมนูหลักของโปรแกรมประกอบด้วยเมนู Admin, เมนู View, เมนู Setup และเมนู Help ซึ่งแต่ละเมนูมีการทำงานดังต่อไปนี้

เมนู Admin เป็นเมนูคำสั่งในการทำงานสำหรับผู้ดูแล (Administrator) โดยทำหน้าที่จัดการเกี่ยวกับ

- การเริ่มต้นการทำงานของโปรแกรม
- การหยุดการทำงานของโปรแกรม
- การลบ web object ที่บันทึกในแคช
- การลบรูปภาพที่มาพร้อมกับ web object ที่บันทึกในแคช
- การเตรียมค่าทางสถิติใหม่ให้กับโปรแกรม กล่าวคือเป็นการตั้งค่าเริ่มต้นใหม่ให้กับโปรแกรม
- การออกจากโปรแกรม

[3] Graphic User Interface คือการติดต่อใช้งานโปรแกรมประยุกต์บนหน้าจอคอมพิวเตอร์ผ่านทางรูปภาพ หรือเมนู แทนการใช้คำสั่งที่เป็นตัวอักษร โดยใช้เมาส์เป็นอุปกรณ์ในการนำเข้าข้อมูล

หน้าจอแสดงรายการดำเนินของเมนู Admin แสดงดังภาพประกอบ 4-10

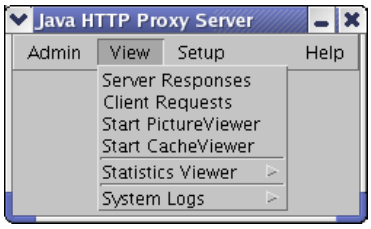


ภาพประกอบ 4-10 เมนู Adminของโปรแกรม Jproxy

เมนู View เป็นเมนูคำสั่งในการทำงานสำหรับแสดงสถานะการทำงานของโปรแกรม และ web object ที่บันทึกในโปรแกรม ซึ่งสามารถแสดงสถานะดังนี้

- การตอบสนองจากเครื่องผู้ให้บริการ
- การร้องขอข้อมูลจากผู้ใช้บริการ
- แสดงรูปภาพที่บันทึกในแคช
- แสดงข้อมูลของ web object ที่บันทึกในแคช
- แสดงค่าทางสถิติขอการทำงาน ซึ่งประกอบด้วย อัตราการพบข้อมูลตามที่ต้องการในแคช (hit ratio) และอัตราจำนวนข้อมูลเป็นจำนวนไบต์ที่พบในแคช (byte hit ratio)
- แสดงบันทึกสถานะของระบบ โดยประกอบด้วยสถานะการเข้าใช้งานเว็บ (access log) และสถานะการทำงานที่ผิดพลาด (error log)

หน้าจอแสดงรายการดำเนินของเมนู View แสดงดังภาพประกอบ 4-11

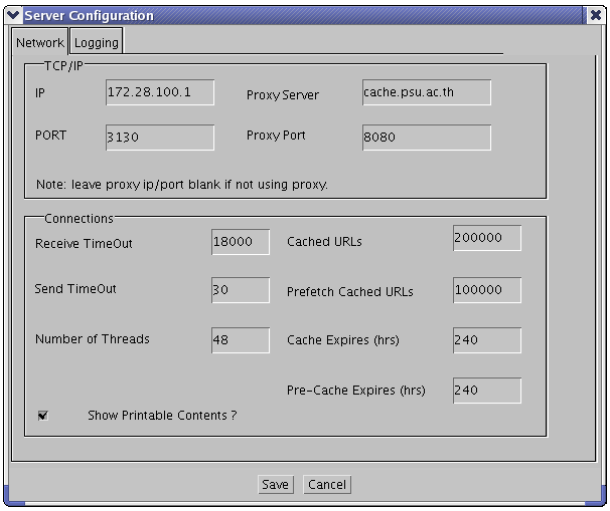


ภาพประกอบ 4-11 เมนู View ของโปรแกรม Jproxy

เมนู Setup เป็นเมนูคำสั่งสำหรับการตั้งค่าการทำงานต่างๆของโปรแกรม ประกอบด้วยการทำงานย่อยๆอีก 2 การทำงานคือ

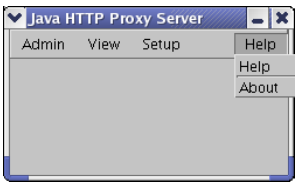
- การตั้งค่าสำหรับเครือข่ายเช่น หมายเลข IP พอร์ตของเครื่องที่ทำหน้าที่เป็นเว็บแคชชิงซึ่งคือเครื่องที่โปรแกรม Jproxy ทำงานอยู่นั่นเอง และขีดจำกัดเวลาของการส่งคำร้องขอและการรอรับการตอบสนองเป็นต้น
- การตั้งค่าสำหรับการบันทึกข้อมูลการดำเนินการของโปรแกรม ซึ่งจะบันทึกข้อมูลการเข้าใช้งานเว็บ ความผิดพลาดที่เกิดขึ้น และจำนวนของบันทึกที่จะบันทึก

หน้าจอแสดงรายการดำเนินของเมนู Setup แสดงดังภาพประกอบ 4-12



ภาพประกอบ 4-12 หน้าต่างสำหรับการตั้งค่าของเครือข่ายในเมนู Setup ของโปรแกรม Jproxy

เมนู Help เป็นเมนูสำหรับแสดงข้อความช่วยเหลือ โดยเมนูย่อย Help เป็นเมนูที่จะแสดงหน้าจอเพื่ออธิบายการใช้งานโปรแกรม Jproxy โดยสรุปและเมนูย่อย About แสดงข้อมูลระบบที่โปรแกรม Jproxy ทำงานอยู่ หน้าจอแสดงรายการดำเนินของเมนู Help แสดงดังภาพประกอบ 4-13



ภาพประกอบ 4-13 เมนู Help ของโปรแกรม Jproxy

การพัฒนาปรับปรุงโปรแกรม Jproxy จะทำการเพิ่มเติมการทำงานในส่วนของการแบ่งการบันทึกข้อมูลการร้องขอระหว่างการร้องขอด้วยโปรแกรม browser อื่นๆที่ไม่ใช้โปรแกรม browser สำหรับการร้องขอข้อมูลล่วงหน้าในที่นี้คือโปรแกรม Wget โดยพิจารณาแยกการบันทึกด้วย Agent และเพิ่มส่วนการทำงานในการวัดประสิทธิภาพการทำงานของเว็บแคชชิงซึ่งวัดด้วยค่า hit ratio และ byte hit ratio ให้กับโปรแกรม Jproxy

4.5 โปรแกรม Wget

โปรแกรม Wget เป็นโปรแกรมประเภท Open source พัฒนาโดยกลุ่ม GNU project [Tortonesi, 2003] ซึ่งใช้ภาษาซีในการพัฒนา ในงานวิทยานิพนธ์ที่นำเสนอใช้โปรแกรม Wget เป็นโปรแกรมสำหรับร้องขอข้อมูลจากเว็บผ่านทางโพรโทคอล HTTP โดยโปรแกรม Wget มีการทำงานแบบ command line ซึ่งคือการเรียกใช้งานโปรแกรมด้วยการพิมพ์คำสั่ง และสามารถระบุตัวเลือกการทำงานเพื่อกำหนดรูปแบบการทำงานเช่น ร้องขอข้อมูลโดยอ่านข้อมูลคำร้องขอจากแฟ้มข้อมูล เป็นต้น ซึ่งข้อมูลที่ตอบกลับบันทึกใน พื้นที่ที่ทำการเรียกโปรแกรม แต่สามารถกำหนดให้บันทึกในพื้นที่อื่นได้ตามตัวเลือกที่ระบุ การทำงานของโปรแกรม Wget ในงานวิทยานิพนธ์นี้จะร้องขอข้อมูลโดยนำคำร้องที่จะร้องขอจากแฟ้มข้อมูล และข้อมูลที่ส่งกลับโปรแกรม Jproxy ทำการบันทึกข้อมูลเมื่อมีการร้องขอข้อมูลนั้นๆผ่านโปรแกรมเว็บแคชชิงและบันทึกข้อมูลในพื้นที่ที่กำหนดซึ่งบันทึก web object ตามประเภทของ Agent ที่ร้องขอ เช่นถ้า Agent ที่ร้องขอข้อมูลคือ Wget บันทึก web object ในพื้นที่ของ Prefetched cache เป็นต้น ตัวอย่างรูปแบบของคำสั่งดังนี้

```
Wget –delete-after –nd –i  c:\request\rule.txt
|-(1)-|----------(2)----------|----------(3)--------|
```

รายละเอียดของคำสั่งประกอบด้วย

(1) Wget คือคำสั่งเพื่อเรียกใช้โปรแกรม

(2) option คือตัวเลือกการทำงานซึ่งสามารถระบุค่าเพื่อกำหนดการทำงานได้หลายรูปแบบ แต่ในงานวิจัยนี้จะใช้ตัวเลือกการทำงาน 3 ชนิดคือ –delete-after, –nd และ -i โดย –delete-after หมายถึงให้โปรแกรม Wget ลบข้อมูลที่ร้องขอมาได้ทันทีหลักจากที่ได้รับข้อมูลเนื่องจากโปรแกรม Jproxy จะบันทึกข้อมูลแล้วจึงไม่จำเป็นต้องบันทึกเพิ่ม ตัวเลือก –nd หมายถึงบังคับให้โปรแกรมไม่สร้างไดเร็กทอรี่เพื่อบันทึกข้อมูล และตัวเลือก –i หมายถึงระบุให้โปรแกรมร้องขอข้อมูลตามคำร้องขอที่อยู่ในแฟ้มข้อมูล

(3) ชื่อแฟ้มข้อมูลซึ่งบันทึกคำร้องขอหรือ URL ที่ต้องการร้องขอข้อมูล

ตัวอย่างการทำงานแสดงดังภาพประกอบ 4-14

```
--21:25:33--      http://bg2.gator.com/gbsf/gbax12.dat
                    ==>  'gbax12.dat'
Connecting to 172.28.100.1:3130 … connected.
Proxy request sent, awaiting response … 200 OK
Length: 11,164 [text/plain]

100% [======================================== >] 11,164        72.20 K/s

21:25:35 (71.92 KB/s) - 'gbax12.dat' saved [11164/11164]

Removing gbax12.dat.
```

ภาพประกอบ 4-14 ตัวอย่างการทำงานของโปรแกรมWget

ภาพประกอบ 4-14 เป็นผลที่แสดงออกทางหน้าจอเมื่อโปรแกรม Wget ทำงาน ซึ่งแสดงรายละเอียดของการทำงานคือ ระบุเวลาที่ทำการร้องขอข้อมูล URL ที่ร้องขอ และชื่อไฟล์ข้อมูลที่ต้องการ ในบรรทัดถัดมาแสดงหมายเลข IP ของการเครื่องให้บริการ proxy ที่ติดต่อเพื่อร้องขอข้อมูล และสถานะผลลัพธ์ของการร้องขอเช่น 200 เป็นต้น พร้อมทั้งแสดงขนาดของข้อมูลที่ได้รับกลับมา โดยในขณะที่ถ่ายโอนข้อมูลจะแสดงสัญลักษณ์ของสถานะการถ่ายโอนเมื่อถ่ายโอนข้อมูลเรียบร้อยสรุปความเร็วของการร้องขอข้อมูล ระบุเวลาที่ทำการบันทึก บรรทัดสุดท้ายแจ้งสถานะของข้อมูลเนื่องจากการทำงานในงานวิทยานิพจน์นี้จะกำหนดตัวเลือกให้ลบข้อมูลที่ได้รับมา เพราะฉะนั้นจึงมีการแจ้งสถานะว่าทำการลบข้อมูลออกไป

4.6 โปรแกรม GetURL

โปรแกรม GetURL พัฒนาโดยผู้วิจัยเป็นโปรแกรมสำหรับร้องขอข้อมูลอีกตัวหนึ่ง พัฒนาด้วยภาษาจาวา ซึ่งโปรแกรม GetURL มีลักษณะการทำงานคล้ายกับการทำงานของโปรแกรม Wget กล่าวคือเป็นโปรแกรมที่ทำงานแบบ command line และร้องขอข้อมูลจาก URL ที่บันทึกในแฟ้มข้อมูลได้ แต่เนื่องจากโปรแกรม GetURL เป็นโปรแกรมสำหรับใช้ทดสอบการทำงานของระบบ จึงทำหน้าที่เฉพาะส่งคำร้องขอเท่านั้น จะไม่มีการรับข้อมูลที่ตอบกลับมาบันทึกหรือใช้งานต่อ เพื่อความสะดวกในการทดสอบการทำงานของแม่แบบ โดยตัวอย่างรูปแบบของคำสั่งในการเรียกใช้งานเป็นดังนี้

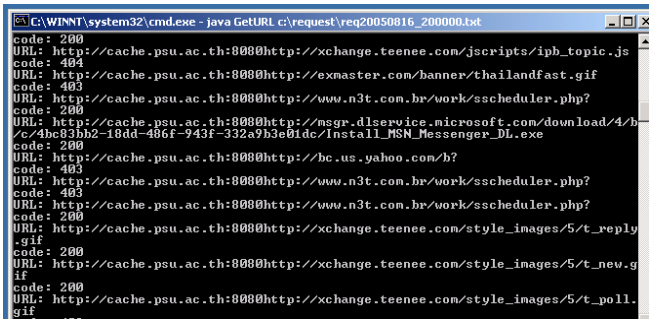
```
GetURL c:\request\rule.txt
|--(1)--|---------(2)---------|
```

รายละเอียดของคำสั่งข้างต้นประกอบด้วย

(1) GetURL หมายถึง ชื่อโปรแกรมสำหรับการเรียกใช้งาน

(2) [path/directory] filename หมายถึง ชื่อแฟ้มข้อมูลที่บันทึก URL ที่ต้องการร้องขอข้อมูล โดยสามารถทำการระบุเส้นทาง (path) และไดเร็กทอรี่ที่จัดเก็บ URL ได้

ตัวอย่างผลจากการทำงานของโปรแกรมแสดงดังภาพประกอบ 4-15



ภาพประกอบ 4-15 ตัวอย่างการทำงานของโปรแกรม GetURL

4.7 บทสรุป

สำหรับบทนี้ได้อธิบายการพัฒนาโปรแกรมสำหรับการเตรียมข้อมูล ซึ่งใช้ภาษา Perl ในการพัฒนาโปรแกรม และโปรแกรมที่เป็นเครื่องมือในการวิจัย โดยจุดเด่นของแม่แบบที่นำเสนอคือการนำเทคนิคการทำเหมืองข้อมูลบันทึกการใช้งานเว็บมาประยุกต์ใช้ร่วมกับเว็บแคชชิง เพื่อทำนายคำร้องขอในอนาคต ซึ่งส่วนใหญ่มักจะใช้เทคนิคการแทนที่ข้อมูลในแคชโดยใช้การคำนวณค่าทางสถิติในการปรับปรุงประสิทธิภาพการทำงานของเว็บแคชชิง เนื่องจากการพิจารณาค่าทางสถิติเพียงอย่างเดียวอาจไม่ครอบคลุมรูปแบบการทำงานของผู้ใช้ตามจริง แต่การทำนายการร้องขอด้วยการทำเหมืองข้อมูลบันทึกการใช้งานเว็บจะช่วยให้วิเคราะห์รูปแบบการใช้งานใกล้เคียงกับความเป็นจริงทำให้ประสิทธิภาพการทำงานของเว็บแคชชิงดีขึ้น ในบทต่อไปจะกล่าวถึงการทดสอบการทำงานของแม่แบบที่พัฒนาขึ้นและผลการทดสอบ