บทที่ 3

การวิเคราะห์และออกแบบแม่แบบ

3.1 บทนำ

ในอดีตปัญหาของเว็บคือผู้ใช้ต้องใช้เวลารอนานสำหรับรับข้อมูลตอบกลับเมื่อทำการติดต่อ และ server ต้องทำงานอย่างหนักเพื่อให้บริการแก่ client เนื่องจากรูปแบบการติดต่อบนเว็บเป็นการติดต่อโดยตรงระหว่าง server และ client เว็บแคชชิงเป็นเทคนิคหนึ่งซึ่งนำมาใช้แก้ปัญหาของเว็บ เพราะเว็บแคชชิงช่วยลดปริมาณงานที่ server ต้องรองรับ และลดระยะเวลาในการรอข้อมูลตอบกลับ ดังนั้นหากสามารถปรับปรุงประสิทธิภาพของเว็บแคชชิงให้ดีขึ้นก็จะช่วยให้การทำงานของเว็บทำงานได้ดีขึ้นด้วยเช่นกัน โดยประสิทธิภาพของเว็บแคชชิงในงานวิทยานิพนธ์นี้วัดจากค่า hit ratio สำหรับการปรับปรุงการทำงานของเว็บแคชชิงในตอนต้นเน้นการพัฒนาวิธีการแทนที่ข้อมูลซึ่งช่วยให้ค่า hit ratio ของเว็บแคชชิงเพิ่มขึ้นประมาณร้อยละ 30 – 50 [Abrams, 1995] และวิธีการข้างต้นอาจทำให้เว็บแคชชิงไม่สามารถรองรับรูปแบบการทำงานของผู้ใช้ได้อย่างครอบคลุมเพราะเว็บแคชชิงใช้ค่าทางสถิติในการตัดสิน หากสามารถทำนายคำร้องขอที่จะเกิดขึ้นแล้วร้องขอข้อมูลนั้นเก็บในแคชก่อนการร้องขอจริงอาจช่วยเพิ่มประสิทธิภาพของเว็บแคชชิงได้ การทำเหมืองข้อมูลเว็บช่วยให้การทำนายคำร้องขอใกล้เคียงกับรูปแบบการทำงานของผู้ใช้เนื่องจากใช้ข้อมูลบันทึกการเข้าใช้งานเว็บในการวิเคราะห์ความสัมพันธ์ ดังนั้นหากนำเทคนิคการทำเหมืองข้อมูลเว็บทำงานร่วมกับเว็บแคชชิงจะช่วยให้เว็บแคชชิงมีประสิทธิภาพการทำงานดีขึ้นและรองรับรูปแบบการทำงานจากผู้ใช้ได้อย่างเหมาะสม หลักการทำงานตามแนวคิดข้างต้นคือการนำข้อมูลจากบันทึกการเข้าใช้งานมาผ่านขั้นตอนการเตรียมข้อมูลและหากฎความเชื่อมโยงจากข้อมูลที่เตรียมไว้ หลังจากนั้นร้องขอข้อมูลจากกฎที่ได้เพื่อบันทึก web object ที่อาจมีการร้องขอในอนาคตลงในแคชก่อนการร้องขอจริง ซึ่งจะส่งผลให้ค่า hit ratio ของเว็บแคชชิงเพิ่มสูงขึ้น

ในรายงานวิทยานิพนธ์นี้นำเสนอแม่แบบการทำงานของเว็บแคชชิงที่นำเทคนิคการทำเหมืองข้อมูลเว็บมาทำงานร่วมกันเพื่อทำนายรูปแบบการร้องขอข้อมูลจากบันทึกการเข้าใช้งานเว็บแคชชิงที่เสมือนเป็นตัวแทนการใช้งานของผู้ใช้เพื่อเพิ่มประสิทธิภาพการทำงานของเว็บแคชชิงให้มีค่า hit ratio สูงขึ้นมากกว่าเดิม เพราะฉะนั้นในบทนี้จะกล่าวถึงการออกแบบแม่แบบเว็บแคชชิงที่ใช้เทคนิคการทำเหมืองข้อมูลเว็บทำนายรูปแบบการร้องขอข้อมูลในอนาคตจากข้อมูลบันทึกการเข้าใช้งาน โดยหัวข้อที่ 3.2 อธิบายถึงรูปแบบของข้อมูลที่ใช้ในงานวิทยานิพนธ์ และใน

หัวข้อที่ 3.3 อธิบายองค์ประกอบต่างๆของแม่แบบซึ่งจะอธิบายหน่วยการทำงานของแม่แบบ ขั้นตอนการทำงาน และหน้าที่ของหน่วยการทำงานแต่ละหน่วย สำหรับหัวข้อสุดท้ายหัวข้อที่ 3.4 นำเสนอบทสรุปของการออกแบบ

3.2 รูปแบบของข้อมูลที่ใช้ในงานวิทยานิพนธ์

สำหรับข้อมูลที่ใช้ในงานวิจัยคือ ข้อมูลบันทึกเข้าใช้งานของเว็บแคชชิง หรือที่เรียกว่า access log ซึ่งเก็บบันทึกข้อมูลเกี่ยวกับการเข้าใช้งานเว็บที่ติดต่อผ่านเว็บแคชชิง รูปแบบของข้อมูลที่บันทึกจะแตกต่างกันไปตามโปรแกรมเว็บแคชชิงที่ใช้ ส่วนใหญ่ข้อมูลที่บันทึกจะเก็บในรูปของ text file ซึ่งบันทึกข้อมูลหนึ่งบรรทัดสำหรับการร้องขอข้อมูลหนึ่งชนิด ในงานวิทยานิพนธ์นี้ใช้ข้อมูลจากเว็บแคชชิงที่ชื่อ cache.psu.ac.th สำหรับการทดสอบการทำงาน โดย cache.psu.ac.th ใช้โปรแกรม squid สำหรับให้บริการเป็นเว็บแคชชิงรูปแบบการบันทึกจัดเป็นประเภท native log format ข้อมูลที่บันทึกประกอบด้วยข้อมูลอย่างน้อยจำนวน 10 รายการและคั่นข้อมูลแต่ละตัวด้วยช่องว่างดังตัวอย่างในภาพประกอบ 3-1

```
1124070796.923 1791 192.168.2.42 TCP_MISS/302 581 GET http://www.telstra.net/bpgstr_u.gif - DIRECT/203.50.5.178 text/html
|--------(1)-------|-(2)-|------(3)------|--------(4)--------|-(5)-|-(6)-|-------------------(7)------------------|-(8)-|-----------(9)------------|-(10)-|
```

ภาพประกอบ 3-1 ตัวอย่างข้อมูลบันทึกการเข้าใช้งานเว็บของโปรแกรม Squid จาก cache.psu.ac.th

โดยรายละเอียดของข้อมูลแต่ละส่วนอธิบายดังนี้

(1) time คือเวลาที่มีการร้องขอข้อมูลโดยบันทึกเวลาในรูปแบบของ UTC มีหน่วยเป็นมิลลิวินาที (Milliseconds)

(2) duration คือเวลาที่ใช้ในการบันทึกข้อมูลลงในแคชมีหน่วยเป็นมิลลิวินาที

(3) client address คือหมายเลข IP ของเครื่องผู้ใช้งานที่ร้องขอข้อมูล

(4) result code คือรหัสผลลัพธ์ของการดำเนินการโดยสามารถแสดงผลลัพธ์ออกเป็น 2 ส่วนคือ รหัสผลลัพธ์ซึ่งประกอบด้วยผลการร้องขอจากแคช เป็นข้อมูลแจ้งถึงลักษณะของการร้องขอข้อมูลจากแคชที่ระบุว่าสำเร็จหรือล้มเหลวอย่างไรเช่น TCP_HIT หรือ TCP_MISS และคั่นด้วยเครื่องหมาย “/” ตามด้วยรหัสสถานะดำเนินการซึ่งใช้รหัสผลลัพธ์ของโพรโทคอล HTTP เช่น 302 โดยรายละเอียดของรหัสผลลัพธ์และรหัสสถานะดำเนินการของโพรโทคอล HTTP นำเสนอในภาคผนวก ก และ ภาคผนวก ข ตามลำดับ

(5) #byte คือขนาดของข้อมูลที่ส่งไปให้ยังผู้ร้องขอ แต่ไม่ใช่ขนาดสุทธิของข้อมูล

(6) request_method คือวิธีการร้องขอเพื่อให้ได้ข้อมูลที่ต้องการเช่น GET หรือ POST

(7) URL คือ URL ของ web object ที่ผู้ใช้ร้องขอ

(8) rfc931 คือข้อกำหนดมาตรฐานสำหรับการอนุญาตให้ใช้บริการบนเครื่องให้บริการ ซึ่งเป็นรูปแบบการระบุตัวตนของผู้ใช้งานที่ติดต่อเครือข่ายบนโพรโทคอล TCP โดยถ้าไม่มีการกำหนดค่าให้บันทึกจะทำการบันทึกเป็นเครื่องหมาย “-”

(9) hierarchy code คือข้อมูลเกี่ยวกับการติดต่อกับเว็บแคชชิงอื่นในระดับชั้นของการติดต่อซึ่งประกอบด้วย hierarchy tag เช่น TIMEOUT_ ในกรณีที่ระยะเวลาในการรอการตอบกลับ package ของโพรโทคอล ICP[2] หมดเวลาตามที่กำหนด และ hierarchy code ที่ใช้อธิบายว่ามีการจัดการคำร้องขออย่างไรเช่น DIRECT หมายถึง ร้องขอ web object จากผู้ให้บริการโดยตรง และหมายเลข IP หรือชื่อของเครื่องให้บริการที่จะทำการส่งต่อคำร้องขอในกรณีที่ข้อมูลที่ร้องขอไม่มีในแคช แต่หากร้องขอไปยังผู้ให้บริการโดยตรงหมายเลข IP นี้คือหมายเลข IP ของเครื่องให้บริการ รายละเอียดของ hierarchy code ดูได้จากภาคผนวก ค

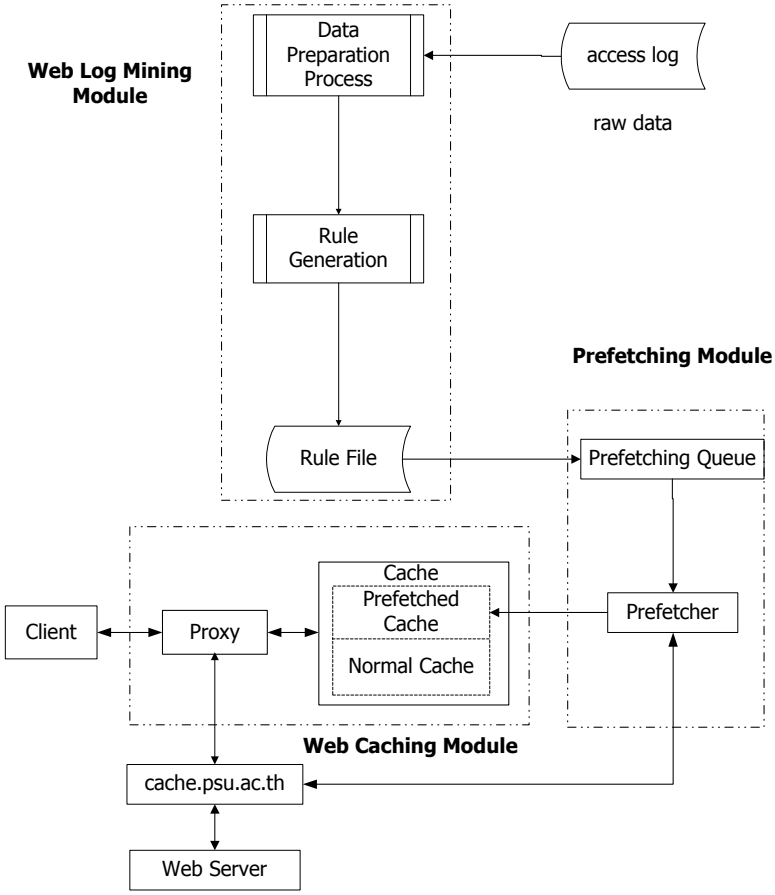
(10) type คือข้อมูลที่บอกชนิดของเนื้อหาที่บรรจุใน web object ที่ทำการร้องขอ ซึ่งสามารถดูได้จาก HTTP header ที่ตอบกลับมาเช่น text/html หมายถึงชนิดของ web object ที่ร้องขอเป็นข้อความแสดงในรูปแบบของ HTML เป็นต้น ในกรณีที่เป็นร้องขอข้อมูลจากเว็บแคชชิงอื่นด้วยโพรโทคอล ICP ข้อมูลนี้จะไม่แสดงชนิดของข้อมูล ดังนั้นจะบันทึกข้อมูลด้วยเครื่องหมาย “-” แทน หรือในบางครั้งข้อมูลที่ตอบกลับมีชนิดของข้อมูลเป็น “:” หรือไม่มีการระบุชนิดของข้อมูล

3.3 องค์ประกอบแม่แบบเว็บแคชชิง

แม่แบบเว็บแคชชิงที่ออกแบบและพัฒนาในงานวิทยานิพนธ์นี้ใช้ข้อมูลบันทึกการเข้าใช้งานเว็บที่มีการจัดรูปแบบตามภาพประกอบ 3-1 สำหรับการทดลองวิเคราะห์แบบอย่างการร้องขอของผู้ใช้ เพื่อทำนายคำร้องขอที่จะร้องขอในอนาคต โดยเรียกแม่แบบที่นำเสนอว่า MineCache ซึ่งสถาปัตยกรรมของแม่แบบนี้ประกอบด้วยหน่วยการทำงาน 3 หน่วยหลักดังภาพประกอบ 3-2 โดยมีรายละเอียดของแต่ละหน่วยการทำงานหลักดังคำอธิบายต่อไปนี้

---

[2] Internet Cache Protocol (ICP) คือโพรโทคอลสำหรับการติดต่อสื่อสารระหว่างเว็บแคชชิงเพื่อแลกเปลี่ยนข้อมูลในการตัดสินใจเลือกเว็บแคชชิงที่เหมาะสมในการดึงข้อมูล[RFC 2186]



ภาพประกอบ 3-2 องค์ประกอบของแม่แบบ MineCache

หน่วยการทำงานหลักของแม่แบบ MineCache ประกอบด้วย

1. *หน่วยการทำเหมืองข้อมูลบันทึกการใช้งานเว็บ (Web Log Mining Module)* ทำหน้าที่ในการหากฎความเชื่อมโยงการใช้งานของผู้ใช้โดยใช้ข้อมูลบันทึกการเข้าใช้งานที่บันทึกโดยเว็บแคชชิงในการทำนาย ซึ่งรายละเอียดการทำงานจะนำเสนอในหัวข้อที่ 3.3.1
2. *หน่วยการทำงานเว็บแคชชิง (Web Caching Module)* เป็นหน่วยการทำงานส่วนกลางที่เชื่อมโยงการทำงานของทั้งระบบ ทำหน้าที่ในการให้บริการต่อคำร้องขอเว็บที่มาจากผู้ใช้ในกรณีที่มีข้อมูลที่ผู้ใช้ต้องการในแคช หรือทำหน้าที่ร้องขอข้อมูลไปยังเครื่องให้บริการจริงเมื่อข้อมูลที่ผู้ใช้ต้องการไม่มีบันทึกในแคชของเว็บแคชชิงซึ่งอธิบายการทำงานในหัวข้อที่ 3.3.2

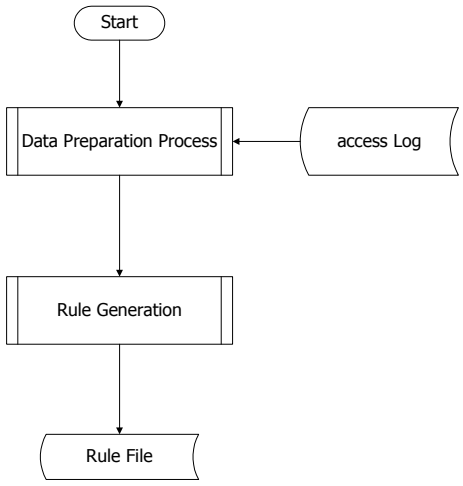
3. *หน่วยการดึงข้อมูลล่วงหน้า (Prefetching Module)* เป็นหน่วยการทำงานที่ทำการร้องขอข้อมูลจากกฎที่ได้จากหน่วยการเหมืองข้อมูลบันทึกการใช้งานเว็บล่วงหน้าเพื่อบันทึกข้อมูลก่อนที่จะมีการร้องขอจริง โดยรายละเอียดของการทำงานอธิบายในหัวข้อที่ 3.3.3

3.3.1 หน่วยการทำงานการทำเหมืองข้อมูลบันทึกการเข้าใช้งานเว็บ (Web Log Mining Module)

หน่วยการทำงานนี้เป็นการวิเคราะห์หาความสัมพันธ์ของการใช้งานเว็บจากข้อมูลบันทึกการเข้าใช้งาน โดยการออกแบบการทำงานของหน่วยการทำงานนี้ประกอบด้วยกระบวนการทำงาน 2 ส่วนย่อยคือ

1. *การเตรียมข้อมูล (Data Preparation Process)* เป็นขั้นที่ทำการเตรียมข้อมูลที่ได้จากบันทึกการใช้งานของเว็บแคชชิง ให้เป็นข้อมูลที่เหมาะสำหรับการหารูปแบบความสัมพันธ์เพื่อสร้างกฎความเชื่อมโยง
2. *การสร้างกฎความเชื่อมโยง (Rule Generation)* ทำหน้าที่สร้างกฎความเชื่อมโยงจากข้อมูลที่ได้ในขั้นการเตรียมข้อมูล

องค์ประกอบและขั้นตอนการทำงานของหน่วยการทำเหมืองข้อมูลบันทึกการเข้าใช้งานเว็บแสดงดังผังงานการทำงานดังภาพประกอบ 3-3

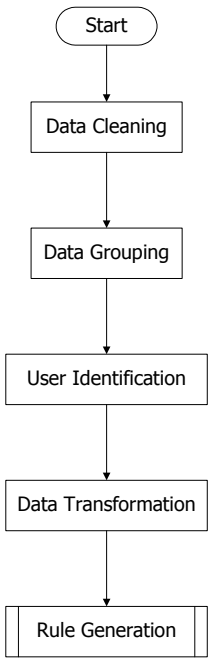


ภาพประกอบ 3-3 ผังการทำงานของหน่วยการทำงานการทำเหมืองข้อมูลเว็บ

ขั้นตอนการทำงานของโปรแกรมส่วนการทำเหมืองข้อมูลบันทึกการเข้าใช้งานเว็บทำงานในลักษณะที่รวบรวมข้อมูลในระยะเวลาหนึ่งมาทำงาน (batch mode) โดยรายละเอียดของการทำงานในขั้นตอนการเตรียมข้อมูล และการสร้างกฎความเชื่อมโยงจะกล่าวถึงในหัวข้อที่ 3.3.1.1 และ 3.3.1.2 ตามลำดับ

3.3.1.1 การเตรียมข้อมูล

เป็นขั้นตอนที่เตรียมข้อมูลที่ได้จากบันทึกการใช้งานของเว็บแคชชิง โดยในงานวิทยานิพนธ์จะเรียกว่า ข้อมูลดิบ (raw data) ตามตัวอย่างในภาพประกอบ 3-1 ให้เป็นข้อมูลที่เหมาะสำหรับการหากฎความสัมพันธ์สำหรับสร้างกฎความเชื่อมโยงเนื่องจากข้อมูลดิบประกอบด้วยข้อมูลต่างๆ จำนวนมากเช่น วันที่ร้องขอข้อมูล เวลาที่ทำการร้องขอ หมายเลข IP ของผู้ร้องขอ วิธีการในการร้องขอเป็นต้น ซึ่งข้อมูลข้างต้นไม่เหมาะสำหรับการนำมาวิเคราะห์ เพราะมีข้อมูลมากเกินความจำเป็นจึงต้องเตรียมข้อมูลตามวิธีการของการทำเหมืองข้อมูลเพื่อให้ได้ข้อมูลที่เหมาะกับการหาความสัมพันธ์ โดยการเตรียมข้อมูลประกอบด้วยการทำงานย่อย 4 กระบวนการคือ การทำความสะอาดข้อมูล การจัดกลุ่มข้อมูล การระบุผู้ใช้งาน และการแปลงรูปข้อมูล ขั้นตอนของกระบวนการทำงานส่วนนี้แสดงดังภาพประกอบ 3-4

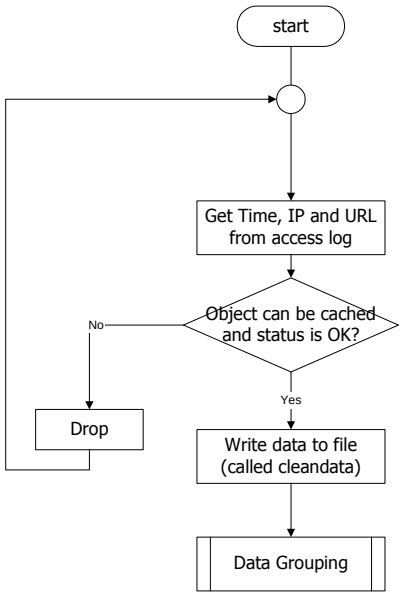


ภาพประกอบ 3-4 กระบวนการเตรียมข้อมูล

โดยวิธีการในการดำเนินการของแต่ละขั้นตอนอธิบายได้ดังนี้

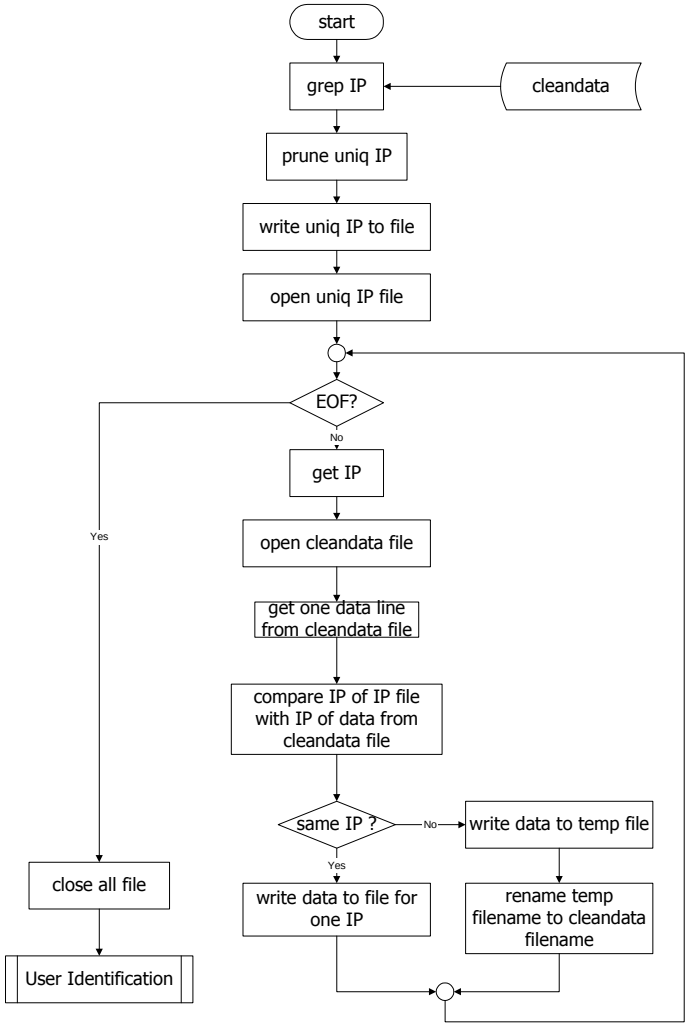
1. *การทำความสะอาดข้อมูล (Data Cleaning)* เป็นกระบวนการที่ทำการคัดข้อมูลที่ไม่จำเป็นออกจากข้อมูลดิบ เพื่อให้เหลือเฉพาะข้อมูลที่จะนำไปวิเคราะห์ แต่การที่จะคัดข้อมูลใดออกจากบันทึกการเข้าใช้งานเว็บหรือเก็บข้อมูลใดไว้ขึ้นอยู่กับวัตถุประสงค์ของการทำเหมืองข้อมูลเว็บ เช่น เว็บไซต์ที่ประกอบด้วยข้อมูลกราฟฟิกเป็นส่วนมากในการวิเคราะห์ไม่ควรทำการคัดข้อมูลประเภทรูปที่มีไฟล์นามสกุลเป็น GIF หรือ JPEG ออกจากบันทึกการใช้งานโดยอัตโนมัติ เพราะในกรณีนี้ข้อมูลกราฟฟิกอาจจะเป็นตัวแทนการดำเนินการของผู้ใช้ที่ชัดเจนควรเหลือข้อมูลนี้ไว้สำหรับการวิเคราะห์ หรือถ้าการทำเหมืองข้อมูลเว็บมีวัตถุประสงค์เพื่อสนับสนุนการทำงานของเว็บแคชชิงหรือการดึงข้อมูลล่วงหน้าก็ไม่ควรคัดข้อมูลคำร้องขอในบันทึกการเข้าใช้งานเว็บที่อ้างถึงรูปภาพหรือข้อมูลอื่นในเว็บออกไป เพราะข้อมูลประเภทนี้ก็เป็นตัวแทนของการดำเนินงานที่เกิดจากการร้องขอในแต่ละครั้ง เช่นเดียวกันกับไฟล์นามสกุล html หรือ htm แต่ข้อมูลบางประเภทที่ไม่สามารถบันทึกได้ควรจะคัดออกเนื่องจากไม่เกี่ยวข้องต่อการทำงานของเว็บแคชชิงตัวอย่างข้อมูลที่บันทึกไม่ได้เช่น ไฟล์ประเภท php หรือ asp เป็นต้น เนื่องจากข้อมูลข้างต้นมีลักษณะของการดำเนินการที่ต้องทำการประมวลผลที่เครื่องให้บริการ โดยข้อมูลและคำสั่งต่างๆ จะมีบนเครื่องผู้ให้บริการเท่านั้นทำให้ไม่สามารถนำข้อมูลออกจากมาเพื่อบันทึกได้ นอกจากนี้ในทางกลับกันถ้าต้องการวิเคราะห์หาโครงสร้างการเชื่อมโยง หรือต้องการวิเคราะห์เพื่อจัดหาการเชื่อมโยงโดยอัตโนมัติสำหรับผู้ใช้แต่ละรายที่เข้ามาเยี่ยมชมเว็บไซต์ ในการนำข้อมูลมาวิเคราะห์ควรนำเฉพาะข้อมูลที่เป็นการร้องขอโดยตรงจากผู้ใช้เท่านั้นเพราะเป็นข้อมูลที่แสดงถึงการดำเนินการของผู้ใช้ เพราะฉะนั้นจึงควรคัดข้อมูลคำร้องขอที่เป็นการร้องขอรูปภาพหรือสื่อหลายมิติออกจากข้อมูลบันทึกการเข้าใช้งาน การกรองข้อมูลที่ไม่มีประโยชน์ออกจากบันทึกการเข้าใช้งานสามารถช่วยลดขนาดของพื้นที่สำหรับเก็บบันทึกข้อมูลนี้ให้ลดลงซึ่งจะช่วยให้การวิเคราะห์ทำงานได้เร็วขึ้น ในงานวิจัยนี้ต้องการเพิ่มประสิทธิภาพเว็บแคชชิงดังนั้นทำการคัดเฉพาะเอาข้อมูลเวลาที่มีการร้องขอ หมายเลขไอพีของผู้ร้องขอ และคำร้องขอที่เป็นการร้องขอ web object ที่สามารถเก็บบันทึกในแคชได้มาใช้ในการวิเคราะห์ โดยทำการพิจารณาจากกลุ่มอักขระที่ปรากฏใน URL ที่ร้องขอ โดย URL ใดที่มีกลุ่มอักขระต่อไปนี้ .cgi, .php, .asp, .jsp, .js, .dll, .cfm, .exe, .pl, .class และสัญลักษณ์ ?, =, & และ comma จะทำการคัด URL นั้นๆออกไปไม่นำมาวิเคราะห์ นอกจากนี้ยังตรวจสอบรหัสสถานะการดำเนินการซึ่งต้องเป็นการดำเนินการที่สำเร็จเช่น

รหัสสถานะ 200 เป็นต้น ถ้าพบว่ารหัสสถานะการดำเนินการไม่สำเร็จก็จะไม่นำ URL นั้นมาวิเคราะห์เช่นกัน ขั้นตอนการทำงานของกระบวนการนี้แสดงดังภาพประกอบ 3-5



ภาพประกอบ 3-5 ผังงานแสดงขั้นตอนการทำความสะอาดข้อมูล

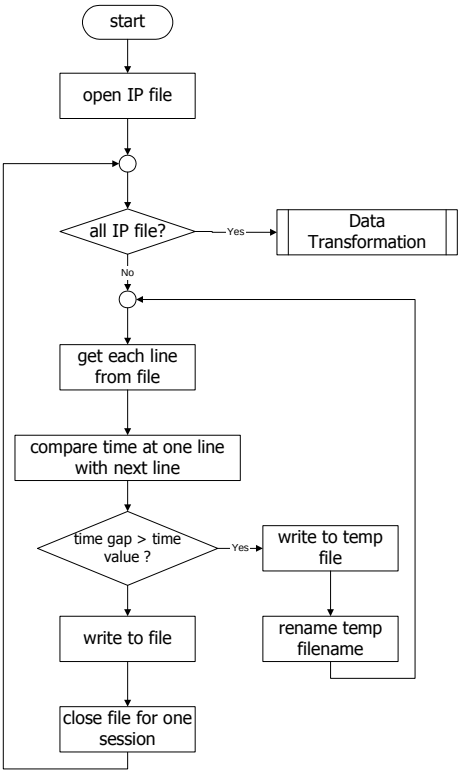
2. *การจัดกลุ่มข้อมูล (Data Grouping)* แฟ้มข้อมูลบันทึกการใช้บางประเภทจะมีการบันทึกข้อมูลเฉพาะที่สามารถใช้ระบุเครื่องคอมพิวเตอร์เช่นชื่อของเครื่องคอมพิวเตอร์ หรือหมายเลขไอพี และโปรแกรมตัวแทนการทำงานของผู้ใช้ (user agent) เป็นต้น หรือบางเว็บไซต์ที่ต้องมีการสมัครใช้งานหรือลงชื่อเพื่อใช้งาน บันทึกการเข้าใช้งานจะมีการบันทึกชื่อการเขาใช้งาน (user login) ซึ่งในกรณีข้างต้นสามารถใช้ข้อมูลชื่อผู้เข้าใช้สำหรับระบุผู้ใช้เพื่อจัดกลุ่มข้อมูล แต่ถ้าไม่มีการให้ผู้ใช้ลงชื่อเข้าใช้งานอาจจะใช้ IP ในการพิจารณาได้ ในงานวิทยานิพนธ์นี้ขั้นตอนจัดกลุ่มข้อมูลจะใช้หมายเลข IP ในการจัดกลุ่ม โดยข้อมูลที่มีหมายเลข IP เดียวกันจะถูกบันทึกลงในแฟ้มข้อมูลเดียวกัน ขั้นตอนการทำงานแสดงดังภาพประกอบ 3-6



ภาพประกอบ 3-6 ผังงานแสดงขั้นตอนการจัดกลุ่ม

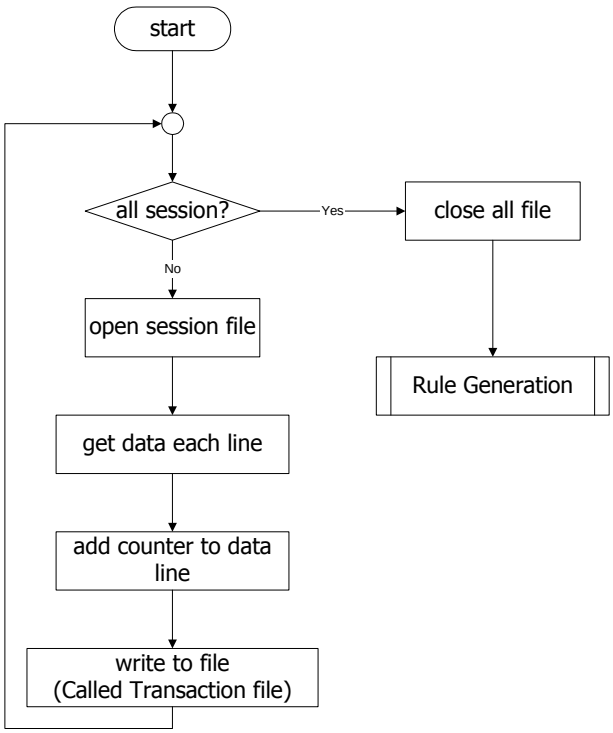
3. *การระบุผู้ใช้ (User Identification)* เป็นกระบวนการที่ทำการระบุผู้ใช้ของข้อมูลในกลุ่มเดียวกัน เนื่องจากถึงแม้ว่าการจัดกลุ่มข้อมูลในขั้นตอนข้างต้นจะได้ข้อมูลที่มีหมายเลขไอพีเดียวกันอยู่รวมเป็นกลุ่มเดียวกันแล้วก็ตาม แต่ก็ยังไม่สามารถชี้ชัดได้ว่าข้อมูลที่เกิดขึ้นนั้นเป็นการเข้าใช้งานของผู้ใช้คนหนึ่งๆ เพราะเครื่องคอมพิวเตอร์เครื่องหนึ่งจะมีผู้ใช้ที่สามารถใช้งานได้หลายคน ดังนั้นจึงจำเป็นต้องทำการระบุให้ได้ว่าคำร้องที่เกิดขึ้นจากเครื่องคอมพิวเตอร์เครื่องหนึ่งๆนั้นเป็นการร้องขอของผู้ใช้รายบุคคล ซึ่งจะใช้ความห่างของระยะเวลา (time gap: Δt) เป็นวิธีการแบ่งข้อมูลเพื่อระบุความแตกต่าง โดยระยะเวลาห่างของเวลาที่ใช้ในงานวิทยานิพนธ์นี้เท่ากับ 73 นาที รายละเอียดของการหาค่า time gap นำเสนอในบทที่ 4 การที่ใช้ time gap เป็นตัวแบ่งรายการดำเนินการที่เกิดขึ้นเพื่อระบุตัว

บุคคลพิจารณาจากการทำงานของผู้ใช้เครื่องที่ผู้ใช้แต่ละคนจะมีการใช้งานเครื่องคอมพิวเตอร์ในระยะเวลาหนึ่งข้อมูลที่เกิดจากการทำงานจะเกิดขึ้นในเวลาใกล้เคียงกัน เพราะฉะนั้นหากเวลาที่บันทึกของข้อมูลที่อยู่ติดกันมีความห่างของระยะเวลามากกว่าค่าที่กำหนดแสดงว่าอาจมีการเปลี่ยนการใช้งานของผู้ใช้ ขั้นตอนการทำงานแสดงดังภาพประกอบ 3-7



ภาพประกอบ 3-7 ผังงานแสดงขั้นตอนการระบุผู้ใช้

4. *การแปลงข้อมูล (Data Transformation)* เป็นกระบวนการสำหรับการแปลงข้อมูลจากกระบวนการข้างต้นให้อยู่ในรูปแบบที่เหมาะสำหรับการนำไปหาความสัมพันธ์โดยใช้เทคนิคการทำเหมืองข้อมูล ตัวอย่างเช่นข้อมูลที่เกี่ยวข้องกับเวลานั้นไม่จำเป็นสำหรับการค้นหาความสัมพันธ์ด้วยเทคนิค association rule เพราะฉะนั้นการเตรียมข้อมูลเพื่อหาความสัมพันธ์จึงควรคัดข้อมูลเวลาในแต่ละรายการการดำเนินการออกไปและจัดเตรียมข้อมูลอื่นที่จำเป็นโดยให้หมายเลขกับ URL แล้วบันทึกลงในแฟ้มข้อมูลโดยขั้นตอนการทำงานแสดงดังภาพประกอบ 3-8



ภาพประกอบ 3-8 ผังงานแสดงขั้นตอนการแปลงข้อมูล

3.3.1.2 การสร้างกฎความเชื่อมโยง

เป็นขั้นตอนการหาความสัมพันธ์ของข้อมูลที่ผ่านการเตรียมข้อมูลแล้วโดยนำเทคนิคการหากฎความเชื่อมโยงค้นหารูปแบบและความเชื่อมโยงของการใช้งาน ซึ่งในงานวิจัยนี้ใช้ขั้นตอนวิธี Apriori ที่เสนอโดย Agrawal ในปีค.ศ.1993 [Agrawal, 1993] เนื่องจากขั้นตอนวิธี Apriori เป็นขั้นตอนวิธีที่เหมาะกับการหาความสัมพันธ์ของข้อมูลขนาดใหญ่ กระบวนการในการสร้างกฎความเชื่อมโยงแบ่งออกเป็น 2 กระบวนการย่อยคือ กระบวนการหา Frequent Itemset และกระบวนการสร้างกฎ รายละเอียดการทำงานนำเสนอในหัวข้อที่ 3.3.1.2.1 และ 3.3.1.2.2 ตามลำดับ

ตารางที่ 3-1 คำนิยามของสัญลักษณ์ที่ใช้สำหรับขั้นตอนวิธี Apriori

k-itemset	ชุดของไอเท็มที่มีสมาชิก k ไอเท็ม โดยที่ k =1, 2, 3, …, n
L_k	ชุดของ Frequent k-itemset นั่นคือชุดของ Frequent Itemset ที่ประกอบด้วยสมาชิก k-Item
C_k	ชุดของ Candidate k-itemset นั่นคือชุดของ Candidate Itemset ที่ประกอบด้วยสมาชิก k-Item

3.3.1.2.1 การหา Frequent Itemset

Frequent Itemset คือชุดของข้อมูลที่คาดว่าจะมีการใช้งานบ่อยๆ ซึ่ง Frequent Itemset ในงานวิจัยนี้ก็คือ URL ที่คาดว่าจะมีการร้องขอ โดยวิธีการการหา Frequent Item ด้วยขั้นตอนวิธี Apriori มีกระบวนการทำงานดังนี้คือ รอบแรกของการทำงานจะทำการนับจำนวนความถี่ของ Item ที่มีในรายการการดำเนินการในที่นี้คือรายการของ URL ที่มีการร้องขอจากผู้ใช้เพื่อหา Frequent 1-Itemset นั่นคือชุดของ Itemset ที่ประกอบด้วยสมาชิกหนึ่งตัวที่เป็น Frequent Itemset ซึ่งความถี่ที่นับได้นำมาคำนวณหาค่าสนับสนุนของ Item นั้นๆ โดยสามารถคำนวณค่าสนับสนุนได้จากสมการ (1) ดังนี้

$$\text{Support (S)} = (|U| / |T|) * 100\% \quad (1)$$

โดยที่ Support (S) คือค่าสนับสนุนของ Itemsset

|U| คือ จำนวนของรายการที่มี Item S ปรากฏอยู่ใน Itemset

|T | คือ จำนวนของรายการทั้งหมดที่เกิดขึ้น

ตัวอย่างเช่น สมมติให้รายการการดำเนินงานที่เกิดขึ้นมีทั้งหมด 4 รายการคือ {ABC, BD, ABCD, CDE} ค่าสนับสนุนของ Item A คือ Support (A) = (2/4)*100% เท่ากับ 50% เป็นต้น

จากนั้นพิจารณาเปรียบเทียบค่าสนับสนุนที่คำนวณได้กับค่าสนับสนุนต่ำสุดที่กำหนดเพื่อระบุว่า Itemset ใดบ้างที่เป็น Frequent 1- Itemset ในรอบถัดมาของการทำงาน นำ Frequent Itemset ที่ได้จากการทำงานในรอบที่แล้วมาสร้างชุด Itemset ที่อาจเป็น Frequent Itemset เรียกว่า Candidate Itemset นับความถี่ของ Candidate Itemset ที่สร้างขึ้นจากข้อมูลในรายการการดำเนินงาน คำนวณหาค่าสนับสนุนแล้วเปรียบเทียบกับค่าสนับสนุนที่กำหนด ถ้า Candidate Itmeset ใดที่มีค่าความสนับสนุนมากกว่าค่าสนับสนุนต่ำสุดที่กำหนดพิจารณา Candidate Itemset ดังกล่าวเป็น Frequent k-Itemset และมีการทำงานเป็นเช่นนี้ซ้ำๆจนกระทั่งไม่สามารถสร้าง Candidates Itemset ได้อีกจึงจบการทำงานในการหา Frequent Itemset ขั้นตอนวิธีการหา Frequent Itemset ของ Apriori แสดงดังภาพประกอบ 3-9

```
L1 = {frequent 1-itemsets};
for (k = 2; Lk-1 ≠ ∅; k++) do begin
     Ck = apriori-gen (Lk-1);   // New candidates
     forall transactions t in the database do begin
          Ct = subset (Ck, t);    // Candidates contained in t
              forall candidates c ∈ Ct do
                   c.count++;
      end
     L1 = { c ∈ Ck  | c.count ≥ minsup}
End
Answer = Uk Lk;
```

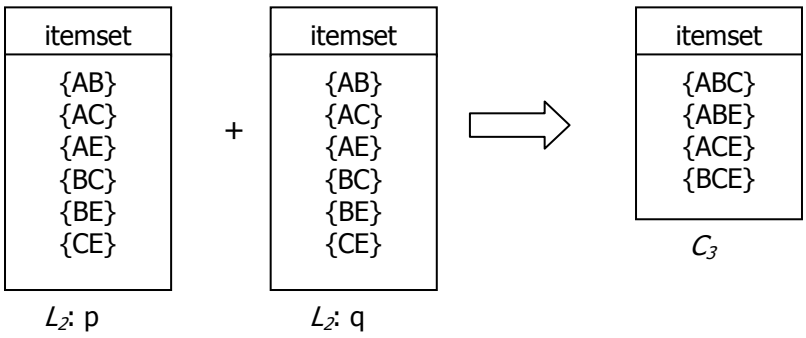
ภาพประกอบ 3-9 การสร้าง Frequent Itemset ของขั้นตอนวิธี Apriori

สำหรับขั้นตอนการสร้าง Candidate Itemset จะเป็นขั้นตอนที่นำ Frequent Itemset ในระดับก่อนหน้ามาเชื่อมต่อกันเป็น Itemset ชุดเดียวกันโดยพิจารณาเปรียบเทียบจาก Item ตัวแรกต้องเหมือนกันจึงจะทำการเชื่อม Item ให้กลายเป็น Itemset ชุดใหม่หลังจากนั้นทำการคัด Itemset ที่ประกอบด้วย Item ที่ไม่ได้เป็นสมาชิกใน Frequent Itemset ในระดับก่อนหน้า โดยขั้นตอนวิธีการสร้าง และตัวอย่างการสร้างแสดงดังภาพประกอบ 3-10 และ 3-11 ตามลำดับ

```
Step 1: join Lk-1 with Lk-1
          insert into Ck
          select p.item1, p.item2, …, p.itemk-1, q.itemk-1
          from Lk-1 p, Lk-1 q
           where p.item1 = q.item1, …, p.itemk-2 = q.itemk-2, p.itemk-1 < q.itemk-1;
Step 2: prune
          forall itemsets c ∈ Ck do
              forall (k-1)-sunsets s of c do
                   if (s ∉ Lk-1) then
                       delete c from Ck
```

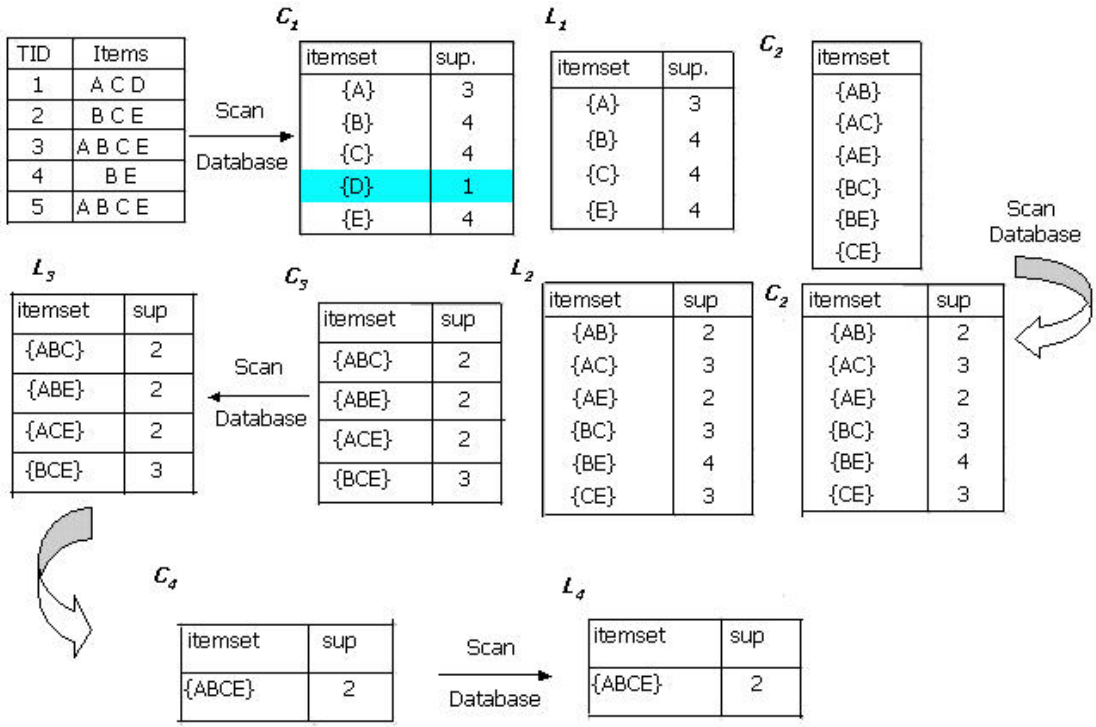
ภาพประกอบ 3-10 ขั้นตอนการสร้าง Candidate Itemset

insert into C_3
select c.item$_1$, p.item$_2$, q.item$_2$
from L_2p, L_2q
where p.item$_1$ = q.item$_1$, p.item$_2$ < q.item$_2$;



ภาพประกอบ 3-11 ตัวอย่างการสร้าง Candidate Itemset

ตัวอย่างการสร้าง Frequent Itemset ของขั้นตอน Apriori แสดงดังภาพประกอบ 3-12



สมมติให้ค่าสนับสนุนต่ำสุด (min_sup) เท่ากับ 40%

ภาพประกอบ 3-12 ตัวอย่างการสร้าง Frequent Itemset ของขั้นตอนวิธี Apriori

จากภาพประกอบ 3-12 มีการทำงานดังนี้คือ เริ่มจากพิจารณาหา 1-frequent itemset ของรายการดำเนินการซึ่งมีรายการทั้งหมด 5 รายการ พิจาณา item ทั้งหมดในที่นี้ประกอบด้วย A, B, C,D และ E และนับความถี่ของ item ที่ปรากฏในรายการ จะได้ว่า A มีปรากฏในรายการทั้งหมด 4 ครั้งหรือ 4 รายการนั่นเอง จากนั้นทำการนับความถี่ของ item ทุกตัวผลที่ได้แสดงดังตาราง C_1 ซึ่งเรียก itemset ที่เกิดว่า 1-Candidate itemset หลังจากนั้นเปรียบเทียบความถี่ที่ได้กับค่าสนับสนุนต่ำสุดที่กำหนดซึ่งตัวอย่างกำหนดให้เท่ากับ 40% ดังนั้น 40% ของการดำเนินการ 5 รายการเท่ากับ 1 เพราะฉะนั้น itemset ที่ในการดำเนินการต้องปรากฏมากกว่า 1 ครั้งจึงจะพิจารณาว่าเป็น 1-frequent itemset ดังแสดงในตาราง L_1 ในการสร้าง 2-Candidate itemset เป็นการเชื่อม 1-frequent itemset กับ 1-frequent itemset โดยเทียบสมาชิกตัวหน้าของแต่ละ itemset หากมีสมาชิกเหมือนกันนำ itemset มาเชื่อมต่อกันผลการทำงานดังตาราง C_2 หลังจากนั้นนับค่าความถี่ของการเกิดขึ้นต้องนับรายการที่มี Candidate itemset ปรากฏพร้อมกันในรายการเช่น itemset {AB} ปรากฏในรายการทั้งหมด 2 รายการเป็นต้น แล้วทำการกรอง 2- Candidate itemset ที่มีค่าความถี่ต่ำกว่าสนับสนุนต่ำสุดผลลัพธ์ที่ได้แสดงดังตาราง L_2 การทำงานรอบถัดไปสร้าง 3-Candidate itemset นับค่าความถี่และคัด Candidate itemset ที่ไม่ผ่านเกณฑ์ผลที่ได้ดังตาราง C_2และ L_3 ตามลำดับ ทำการสร้าง Candidate itemset ในลำดับถัดไปคือ 4-Candidate itemset นับความถี่และกรองCandidate itemset ที่ไม่ผ่านค่าที่กำหนดเพื่อสร้าง 4-frequent itemset ซึ่งจะพบว่าในขั้นนี้ 4-frequent itemset ไม่สามารถสร้าง 5-Candidate itemset ได้อีกแล้วเพราะฉะนั้นหยุดการทำงาน ผลลัพธ์สุดท้ายที่ได้คือ frequent itemset ที่มีสมาชิกคือ {A}, {B}, {C}, {E}, {AB}, {AC}, {AE}, {BC}, {BE}, {CE}, {ABC}, {ABE}, {ACE}, {BCE} และ {ABCE}

3.3.1.2.2 การสร้างกฎ

เมื่อได้ชุดของ Frequent Itemset จากขั้นตอนข้างต้นนำ Frequent Itemset ที่ได้ทั้งหมดมาสร้างกฎความเชื่อมโยง โดยสร้างกฎ (R) ให้อยู่ในรูปของ $X \rightarrow Y$ และทำการคำนวณหาค่าความเชื่อมั่นของแต่ละกฎที่สร้างขึ้นมาเพื่อคัดเลือกกฎมาใช้ โดยพิจารณาเปรียบเทียบค่าความเชื่อมั่นของกฎที่ได้กับค่าความเชื่อมั่นต่ำสุดที่กำหนด ถ้ากฎใดมีค่าความเชื่อมั่นมากกว่าค่าความเชื่อมั่นต่ำสุดพิจารณาว่ากฎที่ได้นั้นสามารถนำมาใช้ได้ ซึ่งการคำนวณหาค่าความเชื่อมั่นสามารถคำนวณได้จากสมการ (2) ดังนี้

Confidence (R) = (Support (X⟶ Y) / Support (X)) *100% (2)

โดยที่ Confidence (R) คือค่าความเชื่อมั่นของกฎ

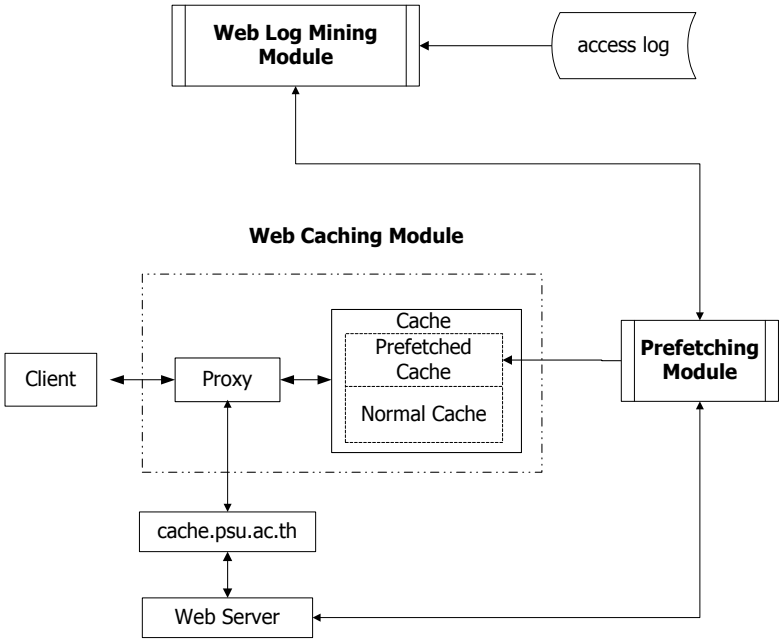
Support (X⟶ Y) คือค่าสนับสนุนของกฎซึ่งก็คือจำนวนรายการที่ประกอบด้วย Item X และ Item Y อยู่ร่วมกัน

Support (X) คือค่าสนับสนุนของ Item X

ตัวอย่างเช่น สมมติให้ Frequent Itemset {A, B, C} มีค่าสนับสนุนเท่ากับ 60% {A, B} มีค่าสนับสนุนเท่ากับ 80% และ {C} มีค่าสนับสนุนเท่ากับ 80% เพราะฉะนั้นเมื่อพิจารณากฎความเชื่อมโยง A, B ⟶ C ค่าความเชื่อมั่นของกฎนี้จะเท่ากับ (Support ({A, B, C}) / Support ({A, B})) * 100% = (60 % /80%)*100% = 75% เป็นต้น

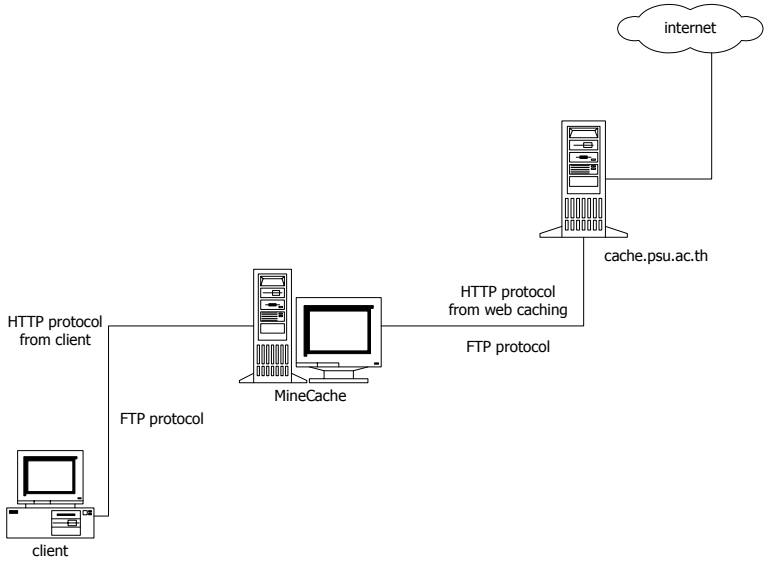
ข้อมูลความสัมพันธ์ที่ได้จากการทำเหมืองข้อมูลเว็บในงานวิจัยนี้จะเรียกว่า กฎความเชื่อมโยง ซึ่งในงานวิทยานิพนธ์นี้จะใช้แทน URL ที่ผู้ใช้จะทำการร้องขอในอนาคตนั่นเอง

3.3.2 หน่วยการทำงานเว็บแคชชิง (Web Caching Module)



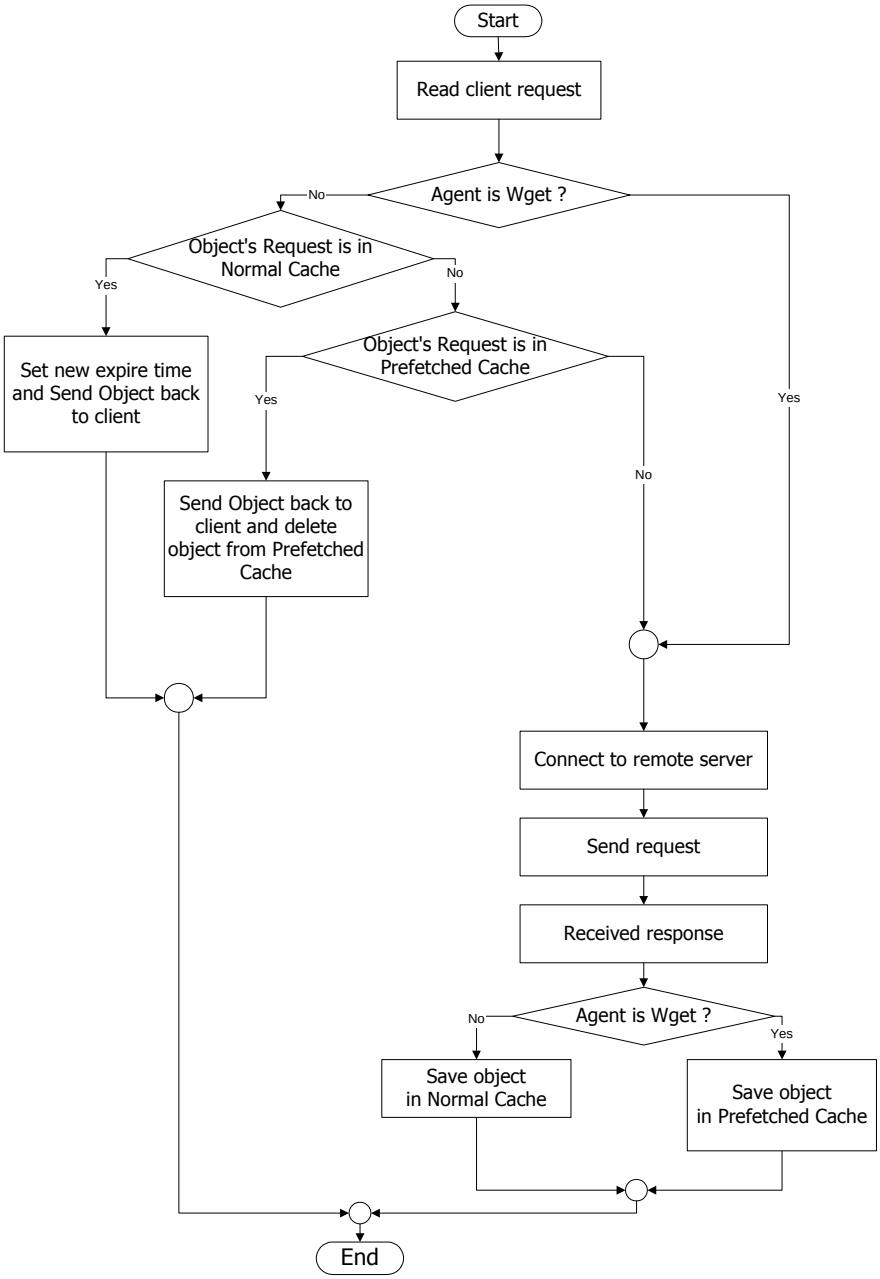
ภาพประกอบ 3-13 หน่วยการทำงานเว็บแคชชิง

แม่แบบเว็บแคชชิงที่นำเสนอประกอบด้วยหน่วยการทำงานหลัก 2 หน่วยคือ proxy และ cache โดย proxy ทำหน้าที่เป็นตัวกลางในการติดต่อระหว่างเครื่องผู้ใช้บริการกับเครือข่ายอื่น ซึ่งในงานวิจัยใช้โปรแกรม Shell script ชื่อ forward ทำการส่งต่อคำร้องจากเครื่องคอมพิวเตอร์ที่อยู่ภายในเครือข่ายติดต่อสู่ภายนอก ซึ่งจะทำการกรองข้อมูลพิจารณาจากพอร์ต (port) โดยคำร้องขอที่มีการร้องขอผ่านโพรโทคอล HTTP จะทำการเปลี่ยนทิศทาง (redirect) การส่งข้อมูลให้ส่งผ่านโปรแกรมเว็บแคชชิงที่ใช้ในงานวิทยานิพนธ์ และสำหรับข้อมูลอื่นๆให้ส่งต่อไปยัง เครือข่ายภายนอกทันที รูปแบบของการติดตั้งโปรแกรมเว็บแคชชิงของหน่วยการทำงานนี้แสดงดังภาพประกอบ 3-14



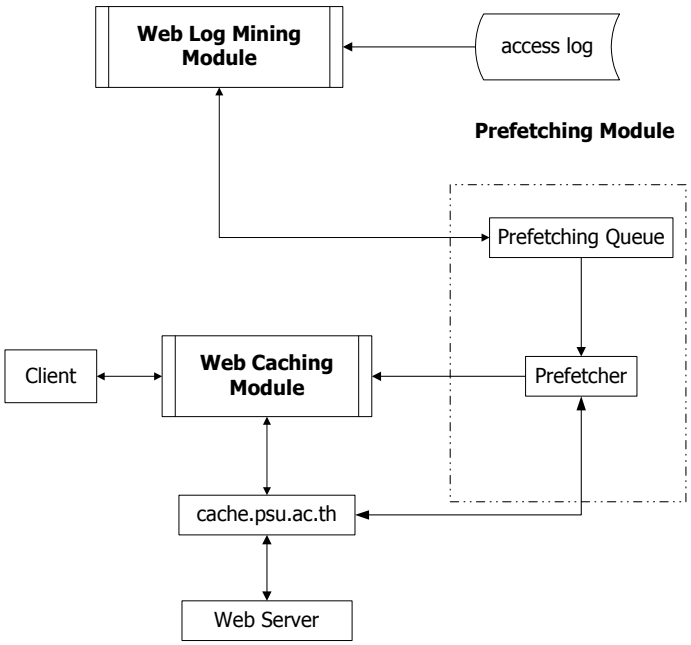
ภาพประกอบ 3-14 รูปแบบการติดตั้งเครื่องให้บริการเว็บแคชชิงในงานวิทยานิพนธ์

ส่วนการทำงาน cache ทำหน้าที่เป็นแหล่งที่เก็บ web object ต่างๆที่ได้จากการตอบกลับจากเครื่องผู้ให้บริการมายังผู้ใช้บริการ โดยในงานวิจัยจะทำการแบ่งพื้นที่ในแคชออกเป็น 2 ส่วนคือ Prefetched cache และ Normal cache โดย Prefetched cache เป็นพื้นที่ที่ใช้เก็บ web object ที่ได้จากการดึงข้อมูลล่วงหน้าจากตัวดำเนินการดึงข้อมูลล่วงหน้า และ Normal cache ใช้สำหรับเก็บ web object จากการใช้งานเว็บแคชชิงตามปกติ ซึ่งแคชทั้งสองนี้ใช้วิธีการแทนที่ข้อมูลสำหรับการบันทึกข้อมูลในแคชต่างกัน โดย Normal cache จะใช้ขั้นตอนวิธี LRU (Least Recently Used) เป็นวิธีสำหรับการจัดการแทนที่ข้อมูลและ Prefetched cache จะใช้ขั้นตอนวิธี FIFO (First In First Out) โดยขั้นตอนการทำงานของเว็บแคชชิงแสดงดังภาพประกอบ 3-15



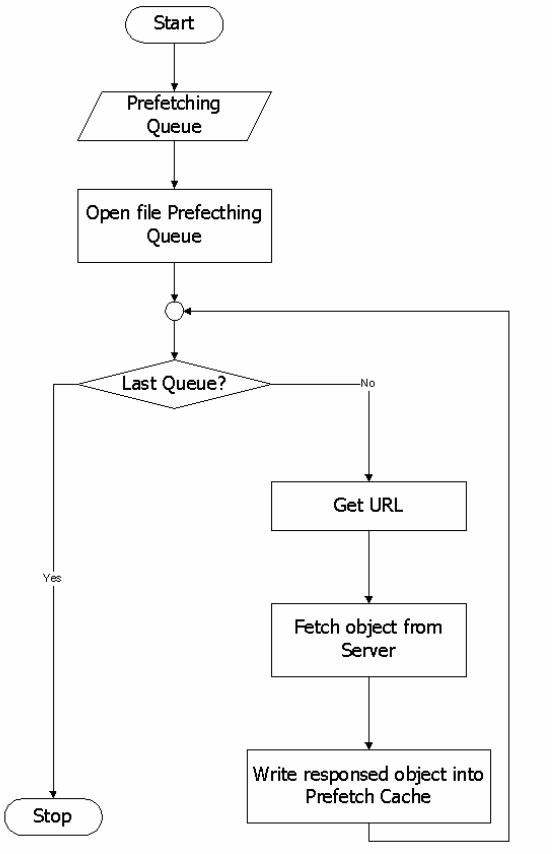
ภาพประกอบ 3-15 ผังงานแสดงขั้นตอนการดำเนินการของหน่วยการทำงานเว็บแคชชิง

3.3.3 หน่วยการทำงานดึงข้อมูลล่วงหน้า (Prefetching Module)



ภาพประกอบ 3-16 หน่วยการทำงานดึงข้อมูลล่วงหน้า

หน่วยการทำงานนี้ประกอบด้วยคิวลำดับการดึงข้อมูล (Prefetching queue) และตัวดำเนินการดึงข้อมูลล่วงหน้า (Prefetcher) โดยคิวลำดับการดึงข้อมูลล่วงหน้ามีหน้าที่เป็นที่เก็บรายการของ URL ที่ต้องทำการดึงมาล่วงหน้าซึ่งก็คือกฎความเชื่อมโยงที่ได้จากขั้นตอนการสร้างกฎ และทำการดึง URL จากแฟ้มข้อมูลบันทึกของคิวลำดับการดึงข้อมูลทีละ URL เพื่อทำการร้องขอข้อมูลนั้นมาเก็บในแคช โดยตัวดำเนินการดึงข้อมูลจะทำหน้าที่ในการร้องขอข้อมูลจาก URL ที่กำหนดจากเครื่องผู้บริการเพื่อร้องขอ web object ที่ผู้ใช้ต้องการผ่านทางเว็บแคชชิง เพราะฉะนั้น web object ที่ตอบกลับมาจะมีการบันทึกในแคชของเว็บแคชชิงโดยอัตโนมัติ แต่จะบันทึกในพื้นที่ของ Prefetched cache ขั้นตอนการทำงานในส่วนการทำงานนี้แสดงดังภาพประกอบ 3-17



ภาพประกอบ 3-17 ผังงานการดึงข้อมูลล่วงหน้าจากคิวลำดับการดึงข้อมูล

3.4 สรุป

บทนี้กล่าวถึงรายละเอียดรูปแบบบันทึกการเข้าใช้งานเว็บซึ่งเป็นข้อมูลสำคัญในการทำงานของแม่แบบ และการออกแบบแม่แบบเว็บแคชชิงที่มีประสิทธิภาพ โดยแม่แบบที่นำเสนอประกอบด้วยหน่วยการทำงานหลัก 3 หน่วยคือ หน่วยการทำเหมืองข้อมูลบันทึกการเข้าใช้งานเว็บ ทำหน้าที่ในการวิเคราะห์ความสัมพันธ์และทำนายคำร้องขอในอนาคตโดยสร้างออกมาในรูปของกฎความเชื่อมโยง หน่วยการทำงานเว็บแคชชิงเป็นส่วนการทำงานให้บริการร้องขอข้อมูลแทนผู้ใช้หรือให้ข้อมูลที่ผู้ใช้ร้องขอ และหน่วยการดึงข้อมูลล่วงหน้า ทำหน้าที่ดึงข้อมูลจากกฎความเชื่อมโยงที่ได้ล่วงหน้าก่อนที่จะมีการร้องขอจริง และเพื่อแสดงให้เห็นการทำงานตามแนวคิดที่นำเสนอจึงทำการพัฒนาแม่แบบโดยนำเสนอเครื่องมือต่างๆที่เกี่ยวข้องในการทำงาน ซึ่งจะกล่าวถึงรายละเอียดของการสร้างตามที่ออกแบบในบทที่ 4