บทที่ 2

ทฤษฎี และแนวคิดในงานวิจัย

2.1 บทนำ

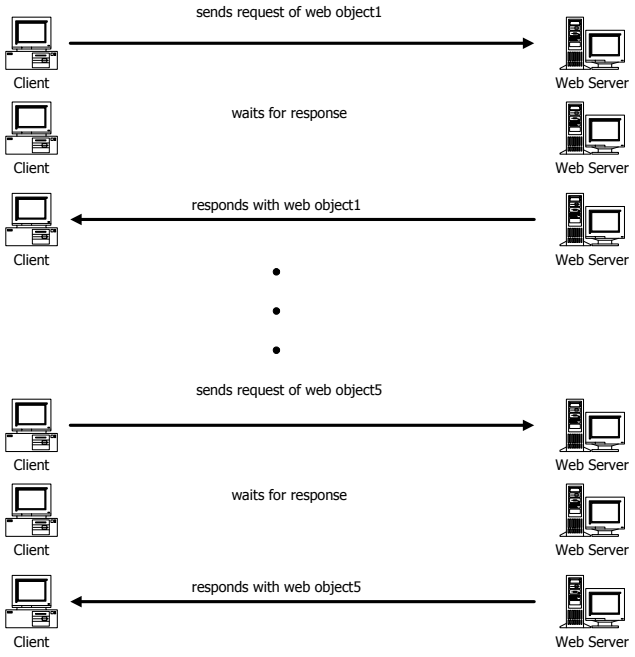
เว็บแคชชิงเป็นเทคโนโลยีหนึ่งที่ทำงานกับเว็บบนเครือข่ายอินเทอร์เน็ต ซึ่งรูปแบบการดำเนินงานของเว็บแคชชิงจะเกี่ยวข้องกับการทำงานของเว็บทั้งรูปแบบการติดต่อและข้อมูลที่ใช้งาน บทนี้จึงนำเสนอทฤษฎีและหลักการพื้นฐานสำหรับงานวิจัย โดยในหัวข้อที่ 2.2 นำเสนอความรู้เกี่ยวกับเทคโนโลยีเว็บ กลไกการทำงานแบบผู้ให้และรับบริการ (client-server mechanism) และโพรโทคอล (Protocol) ในการติดต่อ และหัวข้อที่ 2.3 อธิบายความหมายและหลักการทำงานของเว็บแคชชิง ในหัวข้อที่ 2.4 นำเสนอหลักการและแนวคิดของการทำเหมืองข้อมูลเนื่องจากการทำเหมืองข้อมูลเว็บเป็นเทคนิคการทำงานที่อาศัยหลักการของการทำเหมืองข้อมูล (data mining) เพื่อวิเคราะห์ข้อมูลเว็บ จากนั้นในหัวข้อ 2.5 อธิบายหลักการและวิธีดำเนินการโดยสรุปของการทำเหมืองข้อมูลเว็บ สำหรับหัวข้อที่ 2.6 กล่าวถึงบทสรุป

2.2 เทคโนโลยีเว็บ (Web Technology)

การใช้งานอินเทอร์เน็ตในยุคแรกส่วนใหญ่จำกัดอยู่ในวงการวิจัยและการทหาร เป็นหลักไม่ได้ใช้อย่างกว้างขวางเหมือนในปัจจุบัน จุดเปลี่ยนนั้นเกิดขึ้นเมื่อนักวิทยาศาสตร์แห่งศูนย์ค้นคว้าวิจัยทางฟิสิกส์ชื่อ European Organization for Nuclear Research (CERN) ในประเทศสวิตเซอร์แลนด์ต้องการพัฒนาเทคโนโลยีในการแลกเปลี่ยนข้อมูลข่าวสารระหว่างศูนย์ลูกข่ายซึ่งตั้งในประเทศต่างๆ ทั่วยุโรปให้สะดวกและรวดเร็วขึ้น โดยอาศัยระบบอินเทอร์เน็ตที่มีอยู่เดิมเพียงแต่มีวิธีติดต่อผู้ใช้ (user-interface) ที่ใช้ง่ายขึ้น เทคโนโลยีดังกล่าวมีหลักการทำงานคือเชื่อมโยงเอกสารหลาย ๆ แห่งซึ่งอาจอยู่บนคอมพิวเตอร์ต่างเครื่องเข้าด้วยกันจนคล้ายกับว่ามีเอกสารอยู่ที่เดียว ต่อมาเทคโนโลยีการแลกเปลี่ยนข้อมูลนี้ได้มีการเชื่อมโยงสื่ออื่นที่ไม่ใช่เฉพาะข้อความเท่านั้นเช่น ภาพนิ่ง ภาพเคลื่อนไหว และข้อมูลเสียงเป็นต้น เทคโนโลยีที่พัฒนาขึ้นเรียกว่าเวิลด์ ไวด์ เว็บ (World Wide Web: WWW) หรือเรียกว่าเว็บ หลังจากเว็บได้รับการพัฒนาขึ้นมาใช้ต่อมาสถาบันของมหาวิทยาลัยอิลลินอยส์ชื่อ National Center for Supercomputing Application (NCSA) ได้พัฒนาโปรแกรมเว็บบราวเซอร์ (web browser) ชื่อ Mosaic ซึ่งเป็นโปรแกรมสำหรับ

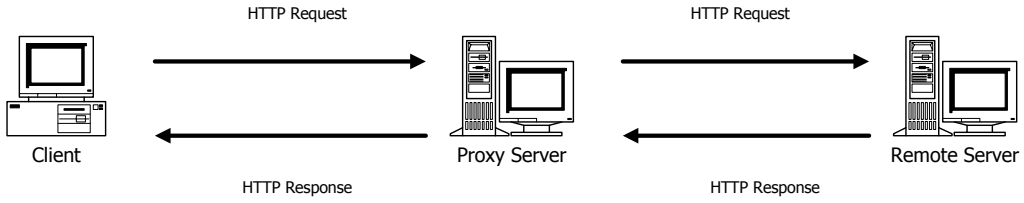
ค้นหาและแสดงข้อมูลเว็บบนหน้าจอเครื่องคอมพิวเตอร์ของผู้ใช้และ NCSA ได้แจกจ่ายโปรแกรม Mosaic ให้แก่สาธารณะทำให้โปรแกรมนี้ได้รับความนิยมอย่างมากจากนั้นบริษัทซอฟต์แวร์ต่าง ๆ ได้พัฒนาโปรแกรม web browser ออกมามากขึ้นและเมื่อมีโปรแกรมให้เลือกใช้งานจำนวนมากก็ช่วยให้ผู้ใช้สามารถใช้บริการอินเทอร์เน็ตง่ายขึ้น เพราะฉะนั้นอาจพิจารณาได้ว่าการใช้อินเทอร์เน็ตที่เกิดขึ้นส่วนใหญ่มาจากการใช้บริการเว็บ

เว็บได้รับการออกแบบให้ผู้ใช้สามารถสร้างและอ่านข้อมูลที่สร้างขึ้นได้ ซึ่งผู้สร้างข้อมูลต่างๆจะทำหน้าที่เป็นผู้ให้บริการในขณะที่ผู้อ่านข้อมูลจะเป็นผู้รับบริการ โดยผู้ให้และรับบริการทุกรายจะทำการติดต่อสื่อสารกันด้วยอินเทอร์เน็ตผ่านทางโพรโทคอลของเว็บที่เรียกว่า HyperText Transport Protocol (HTTP) ซึ่งเป็นโพรโทคอลพื้นฐานสำหรับใช้สื่อสารระหว่างเครื่องคอมพิวเตอร์ทำให้ส่งข้อมูลติดต่อกันได้อย่างถูกต้อง โพรโทคอล HTTP ทำงานอยู่บนโพรโทคอล TCP/IP อีกชั้นหนึ่งโดยรูปแบบการทำงานคือผู้รับบริการจะส่งคำร้องขอ (request) ไปยังผู้ให้บริการแล้วรอจนกระทั่งได้รับข้อมูลกลับตามที่ร้องขอ หลังจากนั้นการติดต่อสื่อสารระหว่างกันจึงสิ้นสุดลง ซึ่งในการร้องขอข้อมูลบนเว็บนั้นโพรโทคอล HTTP จะร้องขอข้อมูลหนึ่งชนิดด้วยคำร้องขอหนึ่งครั้ง ด้วยเหตุนี้ถ้าร้องขอข้อมูลจากผู้ให้บริการรายหนึ่งซึ่งมีข้อมูล 5 ชนิดโพรโทคอล HTTP ก็จะส่งคำร้องขอเพื่อร้องขอข้อมูล 5 ครั้งด้วยกัน ตัวอย่างการทำงานแสดงดังภาพประกอบ 2-1



ภาพประกอบ 2-1 ตัวอย่างการติดต่อระหว่าง client และ server ด้วยโพรโทคอล HTTP

วัตถุประสงค์ในตอนต้นของการสร้างเครือข่ายอินเทอร์เน็ตก็เพื่อสร้างเส้นทางการติดต่อที่ผู้ใช้สามารถรับและส่งข้อมูลผ่านโพรโตคอลต่างๆ แต่ปัจจุบันผู้ใช้อินเทอร์เน็ตไม่ได้ต้องการเพียงแค่รับหรือส่งข้อมูลเท่านั้นแต่ยังคำนึงถึงประเด็นเรื่องของความปลอดภัยด้วย เป็นสาเหตุให้มีการนำเทคโนโลยีด้านความปลอดภัยบนเครือข่ายหลากหลายวิธีมาใช้งาน เทคโนโลยีหนึ่งที่นำมาใช้คือการควบคุมการรับส่งข้อมูลเข้าและออกในระบบเครือข่ายซึ่งได้มีการพัฒนาโปรแกรมจากแนวคิดของเทคโนโลยีข้างต้นมาใช้เรียกว่า proxy โดยเครื่องคอมพิวเตอร์หรืออุปกรณ์ที่ติดตั้งโปรแกรมให้ทำหน้าที่เป็น proxy จะตั้งอยู่ตรงกลางระหว่างระบบเครือข่ายภายในองค์กรและอินเทอร์เน็ตภายนอกองค์กร ดังนั้นคำร้องขอจากภายนอกหรือจากภายในองค์กรจะได้รับการตอบสนองจาก proxy แทนที่จะให้เครื่องจากภายนอกติดต่อกับภายในโดยตรง จากลักษณะข้างต้น proxy จึงทำตัวเหมือนเป็น server ตอบสนองต่อคำร้องขอของ client และทำตัวเป็น client ส่งคำร้องขอไปยัง server ที่อยู่ภายนอก รูปแบบการติดต่อของ client และ server ผ่าน proxy แสดงดังภาพประกอบ 2-2



ภาพประกอบ 2-2 การติดต่อระหว่าง Client และ Server ผ่าน Proxy

เว็บเป็นเพียงเทคโนโลยีที่เชื่อมโยงข้อมูลเข้าด้วยกันแต่ข้อมูลที่นำเสนอนั้นจะมีรูปแบบของการแสดงผลที่มีการเรียกต่างกันตามลักษณะของการดำเนินการคือเว็บเพจ (web page) เว็บไซต์ (web site) และโฮมเพจ (home page) โดยเว็บเพจคือ หน้าเว็บที่ประกอบด้วยข้อมูลต่างๆ ซึ่งข้อมูลแต่ละชนิดที่อยู่บนเว็บเพจจะเรียกว่าเป็นวัตถุบนเว็บ (web object) ข้อมูลในเว็บเพจแบ่งเป็นสองส่วนคือ ส่วนที่เป็นตัวข้อมูลและส่วนที่เป็นตัวเชื่อม (link) ทำหน้าที่เชื่อมโยงไปยังข้อมูลอื่นที่เกี่ยวข้อง ดังนั้นในหน้าเว็บเพจหนึ่งจึงประกอบไปด้วย web object ต่างๆจำนวนมากและในเอกสารของงานวิจัยที่นำเสนอในบทต่อไปจะเรียกแทนข้อมูลต่างๆที่อยู่บนเว็บว่า web object เว็บไซต์คือ แหล่งรวบรวมเว็บเพจหลายๆหน้าเข้าด้วยกันเป็นกลุ่มเดียวเปรียบเหมือนหนังสือหนึ่งเล่ม แต่ละเว็บไซต์ต้องมีวิธีการระบุที่อยู่ในการเข้าถึง (address) ไม่ซ้ำกับผู้อื่น และวิธีการระบุที่อยู่ของเว็บไซต์เรียกว่า Uniform Resource Locator (URL) ประกอบด้วยข้อมูลสองส่วนคือ ส่วนแรก

เป็นโพรโทคอล HTTP ตามด้วยเครื่องหมาย “://” และส่วนที่สองใช้บอกตำแหน่งที่เก็บข้อมูล ประกอบด้วยชื่อของผู้ให้บริการเว็บซึ่งเป็นเครื่องคอมพิวเตอร์ในเครือข่ายอินเทอร์เน็ตที่ให้บริการข้อมูล และชื่อไฟล์ที่เก็บเว็บเพจ หรือชื่อโดเมน (domain name) ที่ใช้ในการติดต่อสื่อสารเพื่อการเชื่อมโยงเว็บไซต์บนอินเทอร์เน็ต โดยการใช้ชื่อไฟล์ต้องระบุเส้นทางหรือไดเร็กทอรี (directory) ให้ถูกต้องเช่น http://www.cs.psu.ac.th/index.html เป็นต้น และโฮมเพจคือ เว็บเพจหน้าแรกสุดในเว็บไซต์เปรียบเสมือนหน้าแรก หรือหน้าปกของหนังสือ และเป็นส่วนที่ใช้บอกชื่อเรื่องของเอกสาร ซึ่งจะมี link ไปยังเว็บเพจหน้าต่างๆ ทั้งในเว็บไซต์เดียวกันและต่างเว็บไซต์

2.3 เว็บแคชชิง (Web Caching)

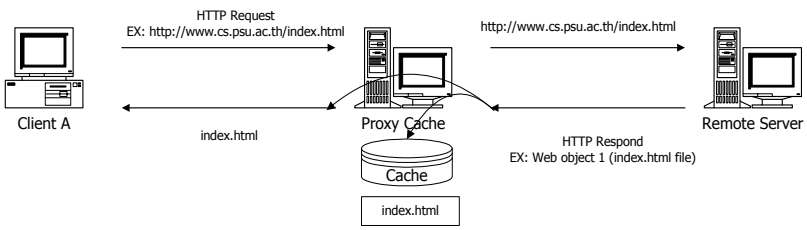
เว็บแคชชิงเป็นเทคโนโลยีที่มีหลักการทำงานลักษณะเดียวกับหน่วยความจำแคชภายในเครื่องคอมพิวเตอร์ซึ่งทำการบันทึกข้อมูลที่คาดว่าหน่วยประมวลผลกลาง (CPU) จะใช้งานเก็บไว้ในแคชก่อนที่จะมีการใช้จริงเพื่อเพิ่มความเร็วในการประมวลผล โดยเว็บแคชชิงจะเป็นแหล่งเก็บ web object ที่ผ่านการร้องขอมาก่อนหน้าและคาดว่าจะได้รับการร้องขออีกในเวลาอันใกล้สำหรับการร้องขอจากผู้ใช้ซึ่งช่วยให้การร้องขอข้อมูลบนเว็บเร็วขึ้นเพราะผู้ใช้ได้ข้อมูลที่ต้องการจากเครื่องให้บริการเว็บแคชชิงที่อยู่ใกล้กับผู้ใช้มากกว่าผู้ให้บริการ เมื่อผู้ใช้บริการร้องขอ web object และได้รับข้อมูลตามที่ร้องขอจากแคชของเว็บแคชชิงก็เรียกว่า cache hit แต่ถ้าร้องขอข้อมูลจากแคชแล้วไม่ได้ตามที่ต้องการเรียกว่า cache miss และเว็บแคชชิงร้องขอข้อมูลจากผู้ให้บริการจริง เมื่อได้รับข้อมูลตอบกลับทำการส่งข้อมูลให้กับผู้ร้องขอ เว็บแคชชิงมีหลายประเภทขึ้นอยู่กับตำแหน่งของแคชที่บันทึกข้อมูลโดยส่วนใหญ่จะแบ่งออกเป็นสองประเภทหลักคือ client cache และ proxy cache โดย

- *client cache* เป็นแคช[1]ที่อยู่ในโปรแกรม web browser โดยมีการจัดแบ่งเนื้อที่ภายในฮาร์ดดิสก์ของเครื่องคอมพิวเตอร์บางส่วนเพื่อบันทึก web object ที่แสดงบนหน้าต่างของโปรแกรมเมื่อผู้ใช้ต้องการย้อนกลับไปดูข้อมูลด้วยการกดปุ่มคำสั่ง back โปรแกรมจะนำข้อมูลจากแคชมาแสดงแทนที่จะต้องรอขอข้อมูลใหม่ client cache สามารถแบ่งออกเป็น 2 รูปแบบตามลักษณะของการบันทึกคือ persistent client cache เป็นการทำงานที่บันทึกข้อมูลที่ร้องขอจาก web browser ในระหว่างที่มีการติดต่อไว้ในแคชตลอด และ non-persistent client cache เป็นลักษณะการ

[1] แคชคือ พื้นที่สำหรับเก็บบันทึกข้อมูลที่มีการร้องขอ

ทำงานที่ทำการคืนพื้นที่ในหน่วยความจำที่ใช้สำหรับบันทึกข้อมูลที่ร้องขอเมื่อผู้ใช้เลิกใช้ web browser

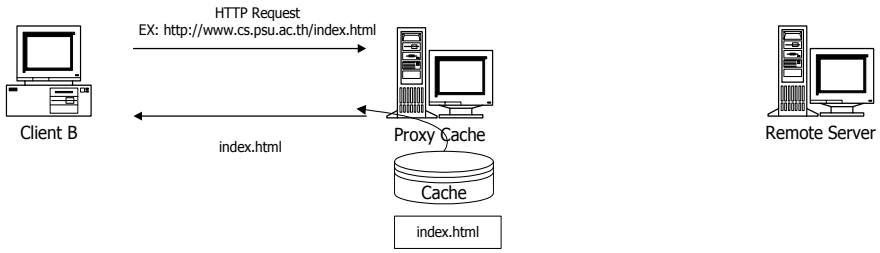
- *proxy cache* เนื่องจากปัจจุบันมีการนำแนวคิดของ proxy มาประยุกต์ใช้ร่วมกับเว็บแคชชิง โดยพัฒนาโปรแกรมที่รวมความสามารถของ proxy ซึ่งเป็นตัวแทนการติดต่อระหว่างเครือข่ายภายในองค์กรกับเครือข่ายภายนอกและเว็บแคชชิงที่บันทึกข้อมูลที่มีการร้องขอเพื่อให้ผู้ใช้ที่ต้องการข้อมูลสามารถใช้ข้อมูลนั้นได้ เมื่อติดตั้งโปรแกรมลักษณะข้างต้นบนเครื่องคอมพิวเตอร์เพื่อให้บริการกับผู้ใช้งานจะเรียกเครื่องให้บริการประเภทนี้ว่า proxy server หรือบางครั้งอาจเรียกว่า cache server โดยไม่มีคำว่า proxy ก็ได้ หรือเรียกสั้นๆว่า proxy ปัจจุบันหากพูดถึง proxy server ก็จะหมายถึงเครื่องให้บริการซึ่งทำหน้าที่ติดต่อกับเครือข่ายภายนอกให้กับ client ที่อยู่ในเครือข่ายภายในและสามารถบันทึก web object ที่ร้องขอได้ แต่ในความจริงแล้วการทำงานของ proxy คือการเป็นตัวแทนการติดต่อเท่านั้นไม่มีการบันทึกข้อมูลร้องขอแต่อย่างใด อย่างไรก็ตาม proxy server กลายเป็นการเรียกบริการที่เป็นตัวแทนติดต่อและบันทึกข้อมูลร้องขออย่างเป็นปกติไปแล้ว ในงานวิจัยที่นำเสนอใช้เว็บแคชชิงประเภท proxy cache นั่นคือเว็บแคชชิงที่ทำการบันทึก web object และเป็นตัวแทนในการติดต่อกับเครือข่ายอื่น รูปแบบการทำงานแสดงดังภาพประกอบ 2-3



ภาพประกอบ 2-3 การทำงานของเว็บแคชชิงประเภท Proxy cache เมื่อมีการร้องขอข้อมูลจากผู้ใช้

ภาพประกอบ 2-3 แสดงการทำงานของเว็บแคชชิงโดย client A ร้องขอข้อมูลเช่น http://www.cs.psu.ac.th/index.html คำร้องขอจาก client A ก็จะส่งไปยัง proxy server เมื่อได้รับคำร้องขอ proxy ทำการตรวจสอบว่ามีข้อมูลที่ client ต้องการในแคชหรือไม่ ถ้าไม่มีจะทำการร้องขอข้อมูลจาก server เมื่อได้รับข้อมูลตอบกลับก็จะบันทึกข้อมูลนั้นในแคชและส่งข้อมูลไปให้ยัง client A หากมีผู้ใช้รายอื่นเช่น client B ร้องขอข้อมูล เมื่อคำร้องขอส่งไปยัง proxy เครื่องให้บริการ proxy ก็จะทำการตรวจสอบข้อมูลที่ร้องขอซึ่งมีการทำงานเหมือนกับที่ให้บริการ client A หากพบว่าข้อมูลที่ร้องขอมีอยู่ในแคช proxy จะส่งข้อมูลนั้นไปให้ client B ทันทีโดยไม่ต้องร้องขอไปยัง

server อีกครั้งดังแสดงการทำงานในภาพประกอบ 2-4 แต่ถ้าไม่พบข้อมูลก็จะส่งคำร้องไปขอข้อมูลจาก server

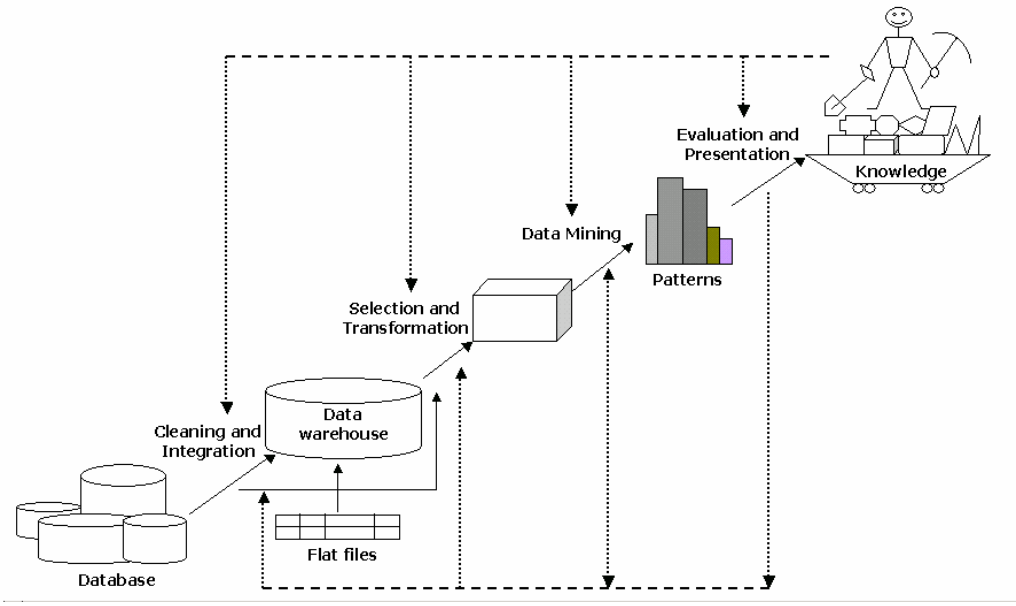


ภาพประกอบ 2-4 การทำงานของเว็บแคชชิงเมื่อผู้ใช้ร้องขอและข้อมูลที่ร้องขอมีในแคช

หลักการทำงานของเว็บแคชชิงคือการบันทึก web object ในพื้นที่ของแคชซึ่งมีขนาดจำกัดตามอุปกรณ์ที่ทำหน้าที่เป็นแคช จากข้อจำกัดข้างต้นพื้นที่บันทึกจึงเป็นข้อด้อยประการหนึ่งของเว็บแคชชิงโดยเมื่อเว็บแคชชิงทำงานไประยะเวลาหนึ่ง web object จะบันทึกในแคชจนกระทั่งไม่มีพื้นที่เหลือสำหรับบันทึก web object ใหม่ที่มีการร้องขอทั้งๆที่ควรจะได้รับการบันทึก ทำให้ต้องร้องขอข้อมูลข้างต้นใหม่ทุกครั้งที่ต้องการ ดังนั้นจึงมีการนำเทคนิคการทำงานเพื่อจัดการพื้นที่ของแคชขนาดจำกัดให้สามารถรองรับการทำงานและให้ผลการทำงานที่มีประสิทธิภาพมากที่สุด เทคนิคหนึ่งที่นำมาใช้คือขั้นตอนวิธีการแทนที่ (replacement algorithm) โดยหากพิจารณาในมุมมองของการทำงานในหน่วยความจำแคชขั้นตอนวิธีการแทนที่เป็นวิธีการที่ใช้เพื่อตัดสินว่าข้อมูลในแคชตัวใดที่ต้องทำการคัดออกจากพื้นที่ของแคชกลับไปยังหน่วยความจำหลักเมื่อแคชต้องการพื้นที่สำหรับบันทึกข้อมูลใหม่ สำหรับในมุมมองการทำงานของเว็บแคชชิงขั้นตอนวิธีการแทนที่ก็จะเป็นวิธีการเพื่อตัดสินคัด web object ออกจากแคชเพื่อให้สามารถบันทึก web object ใหม่ที่ทำการร้องขอได้ ขั้นตอนวิธีแทนที่ซึ่งใช้โดยทั่วไปมีอยู่หลายวิธีเช่นวิธี Least Recently Used (LRU) มีหลักการทำงานคือนำข้อมูลใหม่มาแทนที่ข้อมูลที่มีการใช้งานน้อยที่สุด หรือวิธี First-In-First-Out (FIFO) ที่พิจารณาคัดข้อมูลออกเมื่อข้อมูลนั้นบันทึกในแคชนานที่สุดเป็นต้น แต่จากการศึกษาพบว่าไม่ว่าจะใช้ขั้นตอนวิธีแทนที่วิธีการใดค่า hit ratio ที่เป็นไปได้ที่มากสุดจะมีค่าประมาณ 30 ถึง 50 เปอร์เซ็นต์เท่านั้น [Abrams, 1995] เนื่องจากเทคโนโลยีเว็บแคชชิงเป็นเทคนิคที่ช่วยเพิ่มประสิทธิภาพของเว็บ ดังนั้นหากมีวิธีการที่ช่วยเพิ่มประสิทธิภาพการทำงานของเว็บแคชชิงก็จะช่วยปรับปรุงการทำงานของเว็บให้ดีขึ้นเช่นกัน

2.4 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (data mining) หรือที่รู้จักในอีกชื่อหนึ่งคือ Knowledge Discovery in Databases (KDD) ซึ่งหมายถึงการสกัดหรือค้นหาความรู้จากข้อมูลขนาดใหญ่ โดยทั่วไปแล้วคำว่าการทำเหมืองข้อมูลและ KDD มักมีการนำมาใช้แทนกันแต่ในความเป็นจริงแล้ว การทำเหมืองข้อมูลเป็นเพียงส่วนหนึ่งของกระบวนการในการค้นหาความรู้ของ KDD ดังภาพประกอบ 2-5



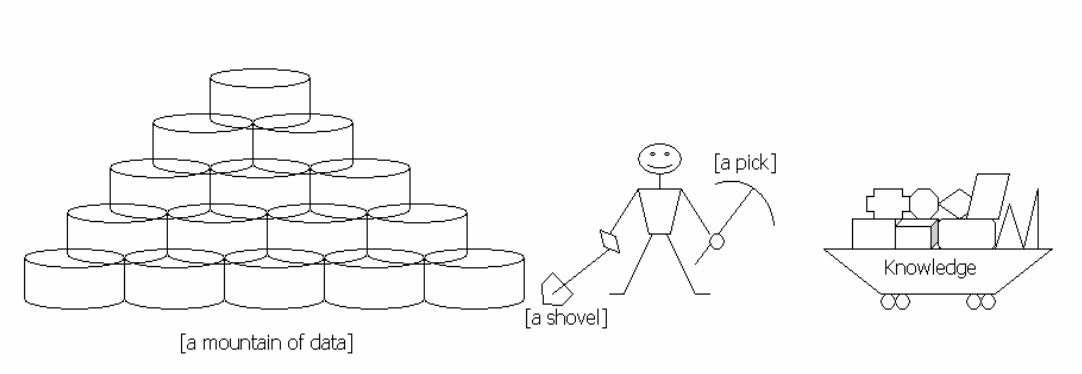
ภาพประกอบ 2-5 กระบวนการการค้นหาความรู้ (Knowledge Discovery Process) [Han, 1999]

กระบวนการ KDD ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลดิบให้กลายเป็นความรู้ใหม่ ซึ่งประกอบด้วยลำดับขั้นตอนดังนี้ data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation และ knowledge representation โดยขั้นตอนข้างต้นสามารถวนกลับไปทำงานยังขั้นตอนที่ผ่านมาได้ สำหรับรายละเอียดของการทำงานอธิบายได้ดังนี้

- *data cleaning* หรือบางครั้งเรียกว่า *data cleansing* เป็นขั้นตอนสำหรับคัดข้อมูลที่เป็นส่วนรบกวน หรือข้อมูลที่ไม่เกี่ยวข้องออกไป
- *data integration* เป็นขั้นตอนรวมแหล่งข้อมูลซึ่งมีข้อมูลคล้ายกันหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน

- *data selection* เป็นขั้นตอนในการดึงข้อมูลสำหรับการวิเคราะห์จากข้อมูลที่บันทึก
- *data transformation* หรือ *data consolidation* เป็นขั้นตอนแปลงข้อมูลที่ผ่านการคัดเลือกในขั้นตอน data selection ให้อยู่เหมาะสำหรับขั้นตอนการทำเหมืองข้อมูล
- *data mining* เป็นขั้นตอนสำหรับสกัดรูปแบบที่มีประโยชน์จากข้อมูลที่เตรียมไว้
- *pattern evaluation* เป็นขั้นตอนในการประเมินรูปแบบที่ได้จากขั้นตอน data mining เพื่อให้ได้รูปแบบที่เป็นตัวแทนของความรู้ที่ตรงความต้องการ
- *knowledge representation* เป็นขั้นตอนสุดท้ายที่นำเสนอความรู้ที่ค้นหาได้ต่อผู้ใช้ โดยจำเป็นต้องมีเทคนิคในการแสดงความรู้ที่ได้เพื่อช่วยให้ผู้ใช้เข้าใจและแปลความผลลัพธ์ได้ถูกต้อง

ชื่อเรียกการทำเหมืองข้อมูลได้มาจากลักษณะการทำงานที่คล้ายกับการทำเหมืองแร่ซึ่งทำการขุดหินเพื่อค้นหาเส้นทางหลักของสายแร่ที่มีค่า เพราะฉะนั้นการทำเหมืองข้อมูลคือการค้นหาข้อมูลสารสนเทศที่มีประโยชน์ในฐานข้อมูลขนาดใหญ่ โดยลักษณะที่คล้ายกันระหว่างการทำงานทั้งสองข้างต้นคือ การเลือกเฟ้นสิ่งที่ต้องการจากวัตถุหรือข้อมูลจำนวนมาก หรือการค้นหาวัตถุหรือข้อมูลอย่างฉลาดเพื่อหาตำแหน่งที่แน่นอนของสิ่งที่ต้องการ อย่างไรก็ตามการเรียกการทำเหมืองข้อมูลเป็นการเรียกชื่อที่ไม่ถูกต้องเนื่องจากในกรณีที่ทำเหมืองเพื่อค้นหาทองคำที่อยู่ในหิน โดยทั่วไปจะเรียกว่าการทำเหมืองทองคำ (gold mining) ไม่ใช่การทำเหมืองหิน (rock mining) ดังนั้นเช่นเดียวกันกับการทำเหมืองข้อมูลจึงควรเรียกว่าการทำเหมืองความรู้ (knowledge mining) ดังภาพประกอบ 2-6 แต่คำว่าการทำเหมืองข้อมูลกลายเป็นคำที่ยอมรับและมีการใช้อย่างเป็นปกติ ทั้งยังมีแนวโน้มในการใช้เพิ่มขึ้นอย่างรวดเร็วจนลดความสำคัญของความหมายพื้นฐานของการค้นหาความรู้ในฐานข้อมูล (KDD) ที่อธิบายได้สมบูรณ์กว่า นอกจากการทำเหมืองข้อมูลแล้วยังมีคำศัพท์อื่นที่มีความหมายคล้ายคลึงกันคือ data dredging, knowledge extraction และ pattern discovery [Han, 2000] เป็นต้น



ภาพประกอบ 2-6 การทำเหมืองข้อมูลโดยการค้นหาความรู้จากข้อมูลจำนวนมาก [Han, 2000]

ผลที่ได้จากทำเหมืองข้อมูลจะมีได้หลายรูปแบบด้วยกันขึ้นอยู่กับลักษณะงานของการทำเหมืองข้อมูลที่นำมาใช้เพื่อค้นหา ซึ่งโดยส่วนใหญ่แล้วจะแบ่งออกเป็น 2 วิธีการหลักคือ descriptive data mining และ predictive data mining โดย descriptive data mining เป็นงานที่อธิบายคุณสมบัติทั่วไปหรืออธิบายรูปแบบของข้อมูลที่มีอยู่แล้วที่อาจนำมาใช้เป็นแนวทางในการตัดสินใจ และ predictive data mining เป็นงานที่ทำนายพฤติกรรมจากข้อมูลที่เกิดขึ้นในอดีตโดยใช้ข้อมูลที่รู้ถึงผลลัพธ์หรือรูปแบบที่เกิดขึ้นสร้างต้นแบบที่สามารถใช้ทำนายกับข้อมูลใหม่หรือข้อมูลที่ต่างไปที่เกิดขึ้นภายหลัง นักวิจัยด้านการทำเหมืองข้อมูล [Osmar, 1999 และ Han, 2000] ได้อธิบายเทคนิคการทำเหมืองข้อมูลแบบต่างๆจากรูปแบบของการทำเหมืองข้อมูลข้างต้นดังต่อไปนี้

- *Characterization Technique* เป็นการทำงานที่สรุปคุณสมบัติหรือลักษณะของกลุ่มข้อมูลเป้าหมายจากข้อมูลที่เหมือนกันกับลักษณะเฉพาะที่ระบุ ตัวอย่างเช่น ถ้าผู้จัดการฝ่ายขายต้องการศึกษาคุณลักษณะของผลิตภัณฑ์ซอฟต์แวร์ที่มียอดขายเพิ่มขึ้น 10 เปอร์เซ็นต์ในปีล่าสุด ดังนั้นต้องมีการรวบรวมข้อมูลที่เกี่ยวข้องของแต่ละผลิตภัณฑ์ซึ่งมีวิธีการหลายวิธี เป็นต้น ผลลัพธ์ที่ได้สามารถแสดงในหลายรูปแบบเช่น แผนภาพวงกลม แผนภาพแท่ง เป็นต้น นอกจากนี้การอธิบายผลลัพธ์ยังสามารถแสดงในรูปของความสัมพันธ์ทั่วไปหรืออยู่ในรูปของกฎที่เรียกว่า characteristic rules
- *Discrimination Technique* เป็นการทำงานที่เปรียบเทียบลักษณะทั่วไปของกลุ่มวัตถุเป้าหมายกับลักษณะทั่วไปของกลุ่มวัตถุเปรียบเทียบ โดยทำการกำหนดกลุ่มเป้าหมายและกลุ่มเปรียบเทียบได้ เทคนิคนี้คล้ายกับ Characterization Technique ผลลัพธ์ที่นำเสนอก็คล้ายกันเพียงแต่คำอธิบายผลลัพธ์จะมีการรวมค่าความแตกต่างจากการเปรียบเทียบลงไปด้วยเพื่อช่วยให้

สามารถแยกระหว่างกลุ่มเป้าหมายและกลุ่มเปรียบเทียบ ซึ่งคำอธิบายผลลัพธ์จะแสดงในรูปของกฎเรียกว่า discriminant rules ตัวอย่างเช่น ต้องการแบ่งกลุ่มลูกค้าของบริษัท AllElectronics ออกเป็น2 กลุ่มคือ กลุ่มที่มักซื้อผลิตภัณฑ์คอมพิวเตอร์โดยมีการซื้อสินค้ามากกว่า 4 ครั้งต่อเดือน กับกลุ่มลูกค้าที่ซื้อสินค้าประเภทนี้นานๆครั้งโดยซื้อสินค้าน้อยกว่า 3 ครั้งต่อปี เพราะฉะนั้นผลลัพธ์ที่ได้สามารถเป็นประวัติเปรียบเทียบของลูกค้าเช่น 80 เปอร์เซ็นต์ของลูกค้าที่ซื้อผลิตภัณฑ์คอมพิวเตอร์บ่อยๆเป็นลูกค้าที่มีอายุระหว่าง 20 ถึง 40 ปี และเป็นนักศึกษา ในขณะที่ 60 เปอร์เซ็นต์ของลูกค้าที่มีการซื้อสินค้าประเภทนี้น้อยมากเป็นลูกค้าทั้งที่มีอายุมากและอายุน้อยมาก ซึ่งไม่ได้เป็นนักศึกษาเป็นต้น

- *Association Analysis Technique* คือการค้นหากฎความเชื่อมโยงหรือที่เรียกว่า association rule ซึ่งเป็นการแสดงคุณลักษณะของข้อมูลที่มีเงื่อนไขว่าเกิดขึ้นด้วยกันบ่อยๆ โดยการสร้างกฎความเชื่อมโยงใช้ค่าแบ่งวัดสองตัวคือ ค่าสนับสนุน (support) เป็นค่าแสดงความถี่ของรายการที่เกิดขึ้นและค่าความเชื่อมั่น (confidence) เป็นค่าแสดงเงื่อนไขความน่าจะเป็นของรายการที่ปรากฏในการดำเนินการเมื่อรายการอีกรายหนึ่งเกิดขึ้น เทคนิคนี้นิยมใช้สำหรับการวิเคราะห์การซื้อ-ขาย (Market-basket analysis) เช่น ถ้าผู้จัดการร้านเช่าวิดีโอสามารถรู้ว่าภาพยนต์เรื่องอะไรบ้างที่มักได้รับการเช่าพร้อมกัน หรือมีความสัมพันธ์กันอย่างไรระหว่างการเช่าภาพยนต์ประเภทหนึ่งกับการซื้อข้าวโพดคั่ว ข้อมูลข้างต้นอาจนำมาใช้ปรับปรุงการจัดร้านให้ลูกค้าเลือกหนังได้ง่ายขึ้นทำให้ยอดการเช่าเพิ่มขึ้นตามไปด้วย กฎความเชื่อมโยงจะอยู่ในรูปแบบ $X \rightarrow Y$ [s, c] เช่น $A_1 \wedge A_2 \wedge A_3 \wedge \ldots \wedge A_m \rightarrow B_1 \wedge \ldots \wedge B_n$ ซึ่ง A_i ($i \in \{1, \ldots, m\}$) และ B_j ($j \in \{1, \ldots, n\}$) โดย X และ Y คือคู่ความสัมพันธ์ทางลักษณะของ “ถ้า ... แล้ว” และ s คือความน่าจะเป็นที่ X และ Y เกิดขึ้นพร้อมกันในการดำเนินการ ส่วน c เป็นความน่าจะเป็นแบบมีเงื่อนไขที่ Y ปรากฏในการดำเนินการเมื่อ X เกิดขึ้น โดยกฎที่ได้มีความหมายว่าเมื่อ X เกิดขึ้น Y ก็จะเกิดขึ้นตามมาด้วย ตัวอย่างเช่น

RentType (X, “game”) $\wedge$ Age (X, “13-19”) $\rightarrow$ Buys (X, “pop”) [s=2%, c= 55%]

จากกฎความเชื่อมโยงข้างต้นสามารถแปลความได้ดังนี้คือ 2 เปอร์เซ็นต์ของรายการดำเนินการที่พิจารณาจะมีลูกค้าอายุระหว่าง 13 ถึง 19 ปีเช่าเกมและซื้อ

ข้าวโพดคั่ว และมีความเป็นไปได้ 55 เปอร์เซ็นต์ที่ลูกค้าวัยรุ่นเช่าเกมและยังซื้อข้าวโพดคั่วด้วย สำหรับในงานวิจัยนี้จะใช้เทคนิคการทำเหมืองข้อมูลประเภทนี้เพื่อค้นหาความสัมพันธ์ของการร้องขอข้อมูลเว็บว่าเป็นอย่างไรเนื่องจากเป็นวิธีการที่ใช้ทำงานกับข้อมูลที่มีจำนวนมาก ซึ่งเหมาะกับข้อมูลเว็บ โดยเทคนิค association rule จะมีขั้นตอนวิธีสำหรับหาความสัมพันธ์หลายวิธีแต่ขั้นตอนวิธีที่เลือกใช้ในงานวิจัยจะใช้ apriori algorithm ซึ่งอธิบายการทำงานในบทที่ 3

- *Classification Technique* เป็นการจัดแบ่งประเภทของข้อมูล โดยหาชุดต้นแบบ หรือชุดของการทำงานที่อธิบายและแบ่งประเภทข้อมูล วัตถุประสงค์เพื่อให้สามารถใช้เป็นต้นแบบทำนายประเภทของวัตถุหรือข้อมูลที่ไม่มีการระบุประเภทหรือชนิดของข้อมูล ซึ่งต้นแบบสร้างจากการวิเคราะห์ชุดของข้อมูลฝึกสอน (Training data) โดยอาจจะเป็นกลุ่มข้อมูลที่มีการระบุประเภทหรือกลุ่มเรียบร้อยแล้ว รูปแบบของต้นแบบแสดงได้หลายแบบเช่น classification (IF-THEN) rules, decision trees หรือ neural networks เป็นต้น ตัวอย่างเช่น ผู้จัดการฝ่ายขายของบริษัทแห่งหนึ่งควรจะสามารถแบ่งประเภทของสินค้าจำนวนมากในคลังสินค้า โดยอาจจะแบ่งประเภทตามการตอบรับในการจัดรายการส่งเสริมการขายคือสินค้าที่ได้รับการตอบรับดี สินค้าที่ได้รับปานกลาง และสินค้าที่ไม่ได้รับการตอบรับ ซึ่งสร้างต้นแบบสำหรับแต่ละประเภทจากคุณสมบัติของสินค้าเช่น ราคา ตราสินค้า สถานที่ผลิต รูปแบบสินค้าและชนิดสินค้า เป็นต้น ผลลัพธ์ที่ได้ควรแยกความแตกต่างของแต่ละประเภทอย่างชัดเจนอาจใช้โครงสร้างต้นไม้แสดงผลการทำงาน โดยให้ราคาเป็นปัจจัยที่ดีที่สุดสำหรับแยกความแตกต่าง และใช้ตราสินค้าและสถานที่ผลิตเป็นปัจจัยถัดมาตามลำดับในการแบ่งประเภท ซึ่งโครงสร้างต้นไม้ที่ได้ช่วยให้เข้าใจถึงผลกระทบที่เกิดขึ้นเมื่อใช้รายการส่งเสริมการขายทำให้สามารถออกแบบและสร้างรายการส่งเสริมการขายที่ดีขึ้นในอนาคต
- *Clustering Analysis Technique* คือการจัดกลุ่มข้อมูลซึ่งมีลักษณะคล้ายกับการแบ่งประเภทแต่จะไม่เหมือนกันโดยการแบ่งประเภทจะวิเคราะห์ข้อมูลตามต้นแบบ แต่สำหรับการแบ่งกลุ่มเป็นการวิเคราะห์โดยไม่พิจารณาจัดกลุ่มตามประเภทที่มีหรือที่รู้จักแต่จะใช้ขั้นตอนวิธีการจัดกลุ่มเพื่อค้นหากลุ่มที่สามารถยอมรับได้เพื่อจัดเข้ากลุ่ม กล่าวคือกลุ่มของวัตถุมีการสร้างขึ้นโดยเปรียบเทียบวัตถุที่มีความเหมือนกันจัดเข้ากลุ่มเดียวกัน

- *Evolution Analysis Technique* เป็นการวิเคราะห์ที่เกี่ยวข้องกับการศึกษาความสัมพันธ์ของเวลากับข้อมูลที่มีการเปลี่ยนแปลงตลอดเวลา สำหรับการวิเคราะห์นี้จะมีการสร้างต้นแบบแนวโน้มการเปลี่ยนแปลงในข้อมูลที่สามารถแสดงคุณสมบัติ เปรียบเทียบ จัดประเภทและแบ่งกลุ่มตามเวลาที่เกี่ยวข้อง
- *Deviation Analysis Technique* คือการค้นหาการเปลี่ยนแปลงในข้อมูลที่มีนัยสำคัญมากที่สุดจากค่าที่วัดก่อนหน้าหรือค่าที่ได้ตั้งเป็นเกณฑ์

จากที่กล่าวมาข้างต้นการทำเหมืองข้อมูลเป็นการค้นหาความรู้จากข้อมูลจำนวนมาก ซึ่งสามารถนำมาปรับใช้กับงานได้หลากหลายประเภท แต่ต้องมีการเตรียมข้อมูลที่จะวิเคราะห์ให้เหมาะกับงาน เพราะฉะนั้นการทำเหมืองข้อมูลสามารถนำมาใช้กับข้อมูลที่อยู่บนเว็บได้เช่นกัน เรียกว่าการทำเหมืองข้อมูลเว็บ

2.5 การทำเหมืองข้อมูลเว็บ (Web Mining)

การทำเหมืองข้อมูลเว็บคือการใช้เทคนิคการทำเหมืองข้อมูลเพื่อค้นหาและสกัดข้อมูลสารสนเทศจากเอกสารเว็บและบริการบนเว็บโดยอัตโนมัติ เพื่อนำความรู้ที่ได้มาแก้ปัญหาที่ต้องการทั้งทางตรงและทางอ้อม จากงานวิจัยของการทำเหมืองข้อมูลเว็บ [Kosala, 2000] นักวิจัยได้แบ่งประเภทของการทำเหมืองข้อมูลเว็บโดยพิจารณาจากข้อมูลที่นำมาวิเคราะห์ออกเป็น 3 ประเภท ดังนี้คือ web content mining, web structure mining และ web usage mining โดยรายละเอียดของแต่ละประเภทนำเสนอในหัวข้อที่ 2.5.1 - 2.5.3 ตามลำดับ

2.5.1 Web Content Mining

เป็นการค้นหาข้อมูลที่มีประโยชน์จากข้อมูลที่อยู่ภายในเว็บเช่น ข้อความ รูปภาพ เป็นต้น โดยงานวิจัยใน web content mining สามารถแบ่งออกเป็น 2 ประเภทตามมุมมองคือ มุมมองทางด้าน Information Retrieval (IR) และมุมมองทางด้านฐานข้อมูล (database) สำหรับเป้าหมายของ web content mining จากมุมมองของ IR คือการทำเหมืองข้อมูลเว็บเพื่อปรับปรุงการหาข้อมูล หรือกรองข้อมูลให้ผู้ใช้โดยพิจารณาจากข้อมูลที่ผู้ใช้อ้างอิงหรือร้องขอ ในขณะที่เป้าหมายของ web content mining ในมุมมองของฐานข้อมูลส่วนใหญ่พยายามจำลองข้อมูลบนเว็บและรวมข้อมูลนั้นเพื่อให้การสอบถามทำงานดีขึ้นมากกว่าการใช้คำหลักเป็นตัวค้นหาเพียงอย่างเดียว

2.5.2 Web Structure Mining

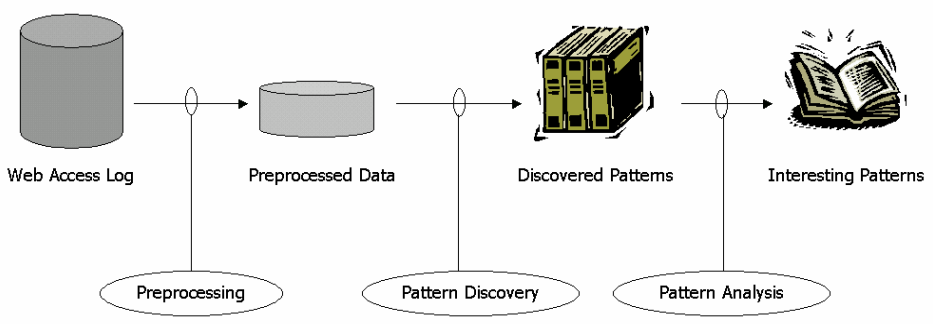
เป็นวิธีการที่พยายามค้นหารูปแบบโครงสร้างการเชื่อมโยงที่สำคัญและซ่อนอยู่ในเว็บ ซึ่งรูปแบบนี้จะขึ้นอยู่กับรูปแบบการเชื่อมโยงเอกสารภายในเว็บ โดยนำรูปแบบที่ได้มาใช้เพื่อจัดกลุ่มเว็บเพจและใช้สร้างข้อมูลสารสนเทศที่เป็นประโยชน์เช่น นำมาใช้ในการปรับโครงสร้างของเว็บให้สามารถให้บริการผู้ใช้ได้อย่างรวดเร็วเป็นต้น

2.5.3 Web Usage Mining

เป็นวิธีการที่พยายามค้นหาความหมายของข้อมูลที่สร้างจากช่วงการทำงานหนึ่งของผู้ใช้หรือสร้างจากพฤติกรรมของผู้ใช้เรียกอีกชื่อหนึ่งว่า web log mining โดยในขณะที่ web content mining และ web structure mining ใช้ประโยชน์จากข้อมูลจริง หรือข้อมูลพื้นฐานบนเว็บแต่ web usage mining ทำการค้นหาความรู้จากข้อมูลการติดต่อสื่อสารระหว่างกันของผู้ใช้ในขณะที่ติดต่อกับเว็บ โดย web usage mining ทำการรวบรวมข้อมูลจากบันทึกในการดำเนินการต่างๆ เช่น บันทึกการใช้งานของ proxy (proxy server log) ข้อมูลการลงทะเบียน (registration data) ข้อมูลการคลิกและเลื่อนเมาส์ (mouse click and scroll) เป็นต้น หรือข้อมูลอื่นอันเป็นผลจากการทำงานร่วมกันมาใช้วิเคราะห์ เพราะฉะนั้น web usage mining จึงเป็นวิธีการทำงานที่เน้นใช้เทคนิคที่สามารถทำนายพฤติกรรมของผู้ใช้ในขณะที่ผู้ใช้ทำงานกับเว็บ กระบวนการทำงานของ web usage mining สามารถแบ่งออกเป็น2 วิธีคือ

- *วิธีการแรก*ทำการจับคู่ข้อมูลการใช้งานของเครื่องให้บริการเว็บให้อยู่ในรูปของตารางความสัมพันธ์ก่อนที่นำข้อมูลนี้มาปรับใช้กับเทคนิคการทำเหมืองข้อมูล
- *วิธีการที่สอง*ใช้ประโยชน์จากข้อมูลในบันทึกการใช้งานโดยตรงซึ่งจะใช้เทคนิคการเตรียมข้อมูล (preprocessing) เพื่อเตรียมข้อมูลก่อนหาความสัมพันธ์ (pattern discovery) และวิเคราะห์รูปแบบ (pattern analysis) สำหรับการเตรียมข้อมูลในงานวิจัยด้านการทำเหมืองข้อมูลเว็บนักวิจัย [Cooley, 1999] แบ่งกระบวนการทำงานเป็น 4 ขั้นตอนคือ data cleaning เป็นกระบวนการกำจัดข้อมูลที่ไม่เกี่ยวข้องออกไป user identification เป็นกระบวนการเพื่อระบุผู้ใช้ session identification เป็นกระบวนการสำหรับระบุช่วงการดำเนินการของผู้ใช้ และ data transformation เป็นกระบวนการเพื่อแปลงข้อมูลที่ได้จากกระบวนการข้างต้นให้อยู่ในรูปแบบที่เหมาะสำหรับการหาความสัมพันธ์ หลังจากนั้นใช้เทคนิคการทำเหมืองข้อมูลหาความสัมพันธ์

ซึ่งเป็นขั้นตอนที่เรียกว่า pattern discovery แล้ววิเคราะห์รูปแบบเพื่อระบุรูปแบบที่น่าสนใจจากรูปแบบความสัมพันธ์ทั้งหมดที่หาได้และนำเสนอให้เข้าใจได้ง่ายเรียกขั้นตอนนี้ว่า pattern analysis กระบวนการทำงานของ web usage mining วิธีนี้แสดงดังภาพประกอบ 2-7



ภาพประกอบ 2-7 กระบวนการทำงานของ Web Usage Mining [Srivastava, 2000]

สำหรับงานวิทยานิพนธ์ที่นำเสนอซึ่งอธิบายรายละเอียดการทำงานในบทที่ 3 จะใช้เทคนิค web usage mining วิธีที่สองเพื่อหาความสัมพันธ์โดยใช้ข้อมูลบันทึกการใช้งานเว็บของเว็บแคชชิงในการดำเนินการวิจัยเนื่องจากต้องการวิเคราะห์ความสัมพันธ์ของการร้องขอของผู้ใช้จากพฤติกรรมการร้องขอและเข้าใช้งานเว็บจริงที่เกิดขึ้น

2.6 บทสรุป

สำหรับบทนี้ได้กล่าวถึงเทคโนโลยีเว็บซึ่งอธิบายลักษณะการติดต่อพื้นฐานบนเว็บแนวคิดและจุดด้อยของเทคนิคเว็บแคชชิง รวมถึงเทคนิคการทำเหมืองข้อมูลเว็บซึ่งเป็นการทำนายรูปแบบการใช้งานจากข้อมูลบันทึกการใช้งานเว็บเพื่อแก้จุดด้อยของเว็บแคชชิงทำให้เว็บแคชชิงมีประสิทธิภาพการทำงานดีขึ้น จากข้อดีและประโยชน์ของการทำเหมืองข้อมูลเว็บที่ได้กล่าวมาแล้วข้างต้น การดำเนินการในงานวิทยานิพนธ์นี้จึงเลือกการทำเหมืองข้อมูลเว็บมาใช้ในการวิเคราะห์และทำนายรูปแบบการใช้งานโดยใช้ข้อมูลบันทึกการใช้งานเว็บของเว็บแคชชิงเพื่อเพิ่มค่า hit ratio ของเว็บแคชชิงให้สูงขึ้น ซึ่งรายละเอียดของรูปแบบของบันทึกข้อมูลการใช้งานเว็บของเว็บแคชชิงและการออกแบบระบบเพื่อนำเสนอแม่แบบจากแนวคิดที่ได้ศึกษาในงานวิจัยจะกล่าวถึงในบทที่ 3