บทที่ 2
วรรณกรรมที่เกี่ยวข้อง

ความรู้พื้นฐานเกี่ยวกับเรื่องที่วิจัย

ถึงแม้ว่านิยามอย่างเป็นทางการของกลุ่มเมฆที่เขียนโดย Mell and Grance (2009) มิได้กำหนดโพรโตคอลเฉพาะที่ใช้ในการเข้าถึงทรัพยากรคำนวณไว้ ข้อมูลส่วนใหญ่ที่โหลดจากกลุ่มเมฆซึ่งรวมถึงข้อมูลดิบของผู้ใช้ที่เก็บอยู่ในบริการโครงสร้างพื้นฐานกลุ่มเมฆและข้อมูลผลการประมวลของบริการแพลตฟอร์มกลุ่มเมฆและบริการซอฟต์แวร์กลุ่มเมฆถูกโหลดมายังผู้ใช้ปลายทางผ่านโพรโตคอลการสื่อสารมาตรฐานสากลนั่นคือ เอชทีทีพี (HTTP) ข้อมูลจากกลุ่มเมฆที่ถูกส่งผ่านโพรโตคอลเอชทีทีพีจะถูกนำมาประมวลผลและ(หรือ)แสดงผลผ่านเว็บบราวเซอร์หรือเว็บโอเอส (Web OS) (Alex, 2000, Bennet, et al, 2010) ของผู้ใช้ปลายทาง ดังนั้นเทคนิคพื้นฐานที่สามารถช่วยลดค่าใช้จ่ายของการโหลดข้อมูลจากกลุ่มเมฆได้คือการติดตั้งเว็บแคชที่ฝั่งองค์กรหรือหน่วยงานของผู้ใช้บริการกลุ่มเมฆเพื่อลดปริมาณการโหลดข้อมูลจากแม่ข่ายบนกลุ่มเมฆ

เว็บแคช หรือ เว็บพร็อกซี่แคช (Web proxy cache) (Tanenbaum and Steen, 2002) เป็นระบบการทำสำเนาของข้อมูลต้นฉบับที่ผู้ใช้โหลดจากแม่ข่ายระยะไกลด้วยโพรโตคอลเอชทีทีพีเป็นต้น ผ่านเครือข่ายคอมพิวเตอร์และเก็บไว้ในหน่วยเก็บข้อมูลของแม่ข่ายเว็บแคชที่ฝั่งผู้ใช้ เมื่อผู้ใช้ต้องการใช้ข้อมูลที่เคยโหลดจากแม่ข่ายอีกในภายหลังก็สามารถอ่านสำเนาของข้อมูลดังกล่าวจากเว็บแคชไปใช้ได้อย่างรวดเร็วโดยไม่จำเป็นต้องโหลดข้อมูลต้นฉบับที่อยู่บนแม่ข่ายอีกซึ่งช้ากว่าการอ่านจากเว็บแคช จะเห็นว่านอกจากประสิทธิภาพในการโหลดข้อมูลที่เพิ่มขึ้น การใช้เว็บแคชยังช่วยลดปริมาณข้อมูลที่ถูกโหลดผ่านเครือข่ายคอมพิวเตอร์จากแม่ข่ายมายังคอมพิวเตอร์ด้านผู้ใช้อีกด้วย ทั้งนี้ประสิทธิภาพของเว็บแคชในการลดปริมาณการโหลดข้อมูลขึ้นอยู่กับกลไกสำคัญสองประการ (Shim, Scheuermann, and Vingralek, 1999) ของเว็บแคชที่ถูกกำหนดให้ใช้ในการประมวลผลคำร้องขอ (Request) การโหลดข้อมูล ดังต่อไปนี้

1) เกณฑ์วิธีการรักษาความต้องกันของข้อมูลในเว็บแคช (Web cache consistency protocols) ใช้สำหรับปรับเนื้อหาของข้อมูลในเว็บแคชให้เป็นปัจจุบันเหมือนกับเนื้อหาของข้อมูลต้นฉบับล่าสุดบนแม่ข่ายซึ่งอาจถูกเปลี่ยนแปลงไปจากเนื้อหาเดิมที่เว็บแคชโหลดไปก่อนหน้า ถ้า

เว็บแคชถูกกำหนดให้ใช้เกณฑ์วิธีการรักษาความต้องกันแบบเข้มงวด เนื้อหาของข้อมูลในเว็บแคชจะถูกปรับให้เหมือนกับข้อมูลต้นฉบับบนแม่ข่ายเสมอด้วยวิธีการใดวิธีการหนึ่ง ข้อมูลที่ผู้ร้องขออ่านไปจากเว็บแคชจึงเป็นข้อมูลที่ทันสมัยเสมอ แต่เกณฑ์วิธีแบบนี้จำเป็นต้องทำการโหลดข้อมูลผ่านอินเทอร์เน็ตเป็นปริมาณมาก เว็บแคชส่วนใหญ่ในปัจจุบันใช้เกณฑ์วิธีแบบไม่เข้มงวดโดยข้อมูลในเว็บแคชจะถูกกำหนดวันหมดอายุ หลังจากวันดังกล่าว เว็บแคชจะต้องโหลดข้อมูลใหม่จากแม่ข่าย(หากต้นฉบับมีการเปลี่ยนแปลง)ก่อนส่งต่อให้ผู้ร้องขอข้อมูล ทำให้ปริมาณการโหลดข้อมูลจากแม่ข่ายมายังเว็บแคชน้อยกว่าเกณฑ์วิธีแบบเข้มงวดเพราะไม่จำเป็นต้องโหลดข้อมูลทุกครั้งที่ข้อมูลต้นฉบับถูกเปลี่ยนแปลงแต่โหลดใหม่เฉพาะเมื่อผู้ใช้ร้องขอและข้อมูลหมดอายุแล้ว (Yin, 2002) ซึ่งเป็นผลดีต่อองค์กรที่ต้องการใช้เว็บแคชเป็นเครื่องมือในการประหยัดค่าโหลดข้อมูล

2) นโยบายการแทนที่ข้อมูลในเว็บแคช (Web cache replacement policies) ใช้ควบคุมการลบข้อมูลออกจากแคชเมื่อพื้นที่เก็บข้อมูลในแคชไม่เพียงพอสำหรับการบันทึกข้อมูลใหม่ที่แคชโหลดมาตามคำร้องขอข้อมูลของผู้ใช้ นโยบายการแทนที่ข้อมูลในเว็บแคชทวีความสำคัญยิ่งขึ้นในยุคกลุ่มเมฆเนื่องจากข้อมูลที่ส่งผ่านเข้าสู่ระบบเครือข่ายคอมพิวเตอร์ของผู้ใช้ซึ่งจะถูกสำเนาเก็บไว้ในเว็บแคชด้วยมีขนาดเพิ่มขึ้นอย่างต่อเนื่องอันเป็นผลมาจากการโหลดข้อมูลของผู้ใช้ที่ถูกย้ายไปเก็บไว้บนกลุ่มเมฆและข้อมูลตอบกลับจากบริการซอฟต์แวร์กลุ่มเมฆต่างๆ ที่มีจำนวนเพิ่มมากขึ้นในขณะที่หน่วยเก็บข้อมูลของเว็บแคชมีพื้นที่เก็บข้อมูลที่จำกัด ตัวอย่างนโยบายการแทนที่ข้อมูลในเว็บแคชที่ถูกใช้งานอย่างแพร่หลายในผลิตภัณฑ์เว็บแคชส่วนใหญ่รวมทั้ง Squid (Wessels, 2004) ซึ่งมีชื่อเสียงและมีการใช้งานอย่างกว้างขวางทั่วโลก ได้แก่ LRU, LFU-DA และ GDSF (Podlipnig and Böszörmenyi, 2003)

ทฤษฎีที่รองรับหรือกรอบความคิดทางทฤษฎี

ปัจจัยพื้นฐานต่างๆ ที่นโยบายการแทนที่ข้อมูลในเว็บแคชปัจจุบันใช้ในการพิจารณาเลือกข้อมูลที่จะถูกแทนที่ได้แก่ ระยะเวลาที่ข้อมูลถูกร้องขอครั้งล่าสุด (Time since last request หรือนิยมเรียกว่า Recency), จำนวนครั้งที่ข้อมูลถูกร้องขอ (Number of past requests หรือนิยมเรียกว่า Frequency), ขนาดของข้อมูลที่ร้องขอ (Size of requested object), ความถี่ของการเปลี่ยนแปลงเนื้อหาข้อมูลต้นฉบับ (Update frequency), อายุใช้งานคงเหลือ (Freshness lifetime) และต้นทุนทาง

เทคนิคในการโหลดข้อมูล (Cost to fetch object from its origin server) (เช่น ระยะเวลาที่ต้องใช้หรือแบนด์วิดท์ในการโหลดข้อมูลจากแม่ข่าย, อัตราค่าโหลดข้อมูลตามปริมาณ) กลุ่มทฤษฎีที่เป็นเหตุผลสนับสนุนแนวคิดของการพิจารณาแต่ละปัจจัยข้างต้นมีรายละเอียดโดยสังเขปดังนี้

1) ขนาดของข้อมูลที่ร้องขอ, อัตราค่าโหลดข้อมูล, จำนวนครั้งที่ข้อมูลถูกร้องขอ และระยะเวลาที่ข้อมูลถูกร้องขอครั้งล่าสุด

เนื่องจากโดยปกติค่าใช้บริการกลุ่มเมฆคิดจากอัตราค่าโหลดข้อมูลและปริมาณข้อมูลที่รับส่งระหว่างผู้ใช้และผู้ให้บริการกลุ่มเมฆ นโยบายการแทนที่ข้อมูลในเว็บแคชที่มุ่งเน้นการลดค่าโหลดข้อมูลจากกลุ่มเมฆจึงควรนำขนาดของข้อมูลที่ร้องขอมาเป็นปัจจัยพื้นฐานในการพิจารณาเลือกแทนที่ข้อมูลด้วย งานวิจัยของ Podlipnig and Böszörmenyi (2003) และ Williams et al. (1996) พบทฤษฎีว่านโยบายที่เลือกแทนที่ข้อมูลที่มีขนาดใหญ่ (เพื่อให้ได้พื้นที่ในเว็บแคชคืนสำหรับเก็บข้อมูลขนาดเล็กจำนวนมาก (Balamash and Krunz, 2004) มีผลช่วยให้อัตราการพบข้อมูลในแคชเพิ่มขึ้น นั่นคือจำนวนครั้งที่ผู้ใช้ได้รับข้อมูลที่ร้องขออย่างรวดเร็วมีความถี่มากขึ้น แต่ผลกระทบด้านลบคือประสิทธิภาพของเว็บแคชเมื่อใช้อัตราขนาดรวมของข้อมูลที่พบในแคชเป็นหน่วยวัดจะลดลง กล่าวอีกนัยหนึ่งคือปริมาณข้อมูลรวมตลอดระยะเวลาการใช้งานเว็บแคชที่ต้องโหลดใหม่จากแม่ข่ายจะเพิ่มขึ้น สภาวะการได้อย่างเสียอย่างที่ค้นพบนี้สอดคล้องกับการงานวิจัยของ Wong (2006) ที่ว่าไม่มีนโยบายใดสามารถให้ค่าของหน่วยวัดทุกตัวสูงได้พร้อมๆ กัน ดังนั้นนโยบายการแทนที่ข้อมูลในเว็บแคชปัจจุบันเลือกแทนที่ข้อมูลที่มีขนาดใหญ่จึงไม่เหมาะกับองค์กรที่ให้ความสำคัญกับความประหยัดค่าโหลดข้อมูลของการใช้บริการกลุ่มเมฆมากกว่าสมรรถนะด้านอัตราการพบข้อมูลในแคช กอปรกับที่ Gartner (2010) ได้เผยแพร่ข้อมูลว่าภายในปี 2013 ข้อมูลในรูปแบบวีดิทัศน์ (Video), เสียง และภาพ กล่าวอีกนัยหนึ่งคือข้อมูลขนาดใหญ่ จะมีสัดส่วนเพิ่มขึ้นเป็น 25 เปอร์เซ็นต์ โดยเฉพาะวีดิทัศน์จะกลายเป็นแบบชนิดมาตรฐานของสื่อที่มีการใช้งานอย่างแพร่หลายมากขึ้น ดังนั้นนโยบายการแทนที่ข้อมูลในเว็บแคชปัจจุบันที่เลือกแทนที่ข้อมูลขนาดใหญ่ก่อนอาจมีประสิทธิภาพในการลดค่าใช้จ่ายการโหลดข้อมูลจากผู้ให้บริการกลุ่มเมฆน้อยลง

อย่างไรก็ตาม การพิจารณาเลือกแทนที่ข้อมูลขนาดใหญ่เพียงอย่างเดียวอาจไม่สามารถลดค่าใช้จ่ายของการโหลดข้อมูลได้อย่างมีประสิทธิภาพเพราะค่าใช้จ่ายที่ลดลงในการโหลดข้อมูลแต่ละชิ้นเกิดจากปริมาณข้อมูลรวมที่ไม่จำเป็นต้องโหลดใหม่จากแม่ข่ายสำหรับข้อมูลชิ้นดังกล่าว

เนื่องจากพบในเว็บแคช โดยสามารถคำนวณค่าใช้จ่ายที่ลดลงในการโหลดข้อมูลแต่ละชิ้นได้จากผลคูณระหว่างขนาดข้อมูลชิ้นนั้นและจำนวนครั้งที่ข้อมูลชิ้นนั้นถูกร้องขอหลังจากถูกบันทึกไว้ในแคชแล้ว งานวิจัยนี้เสนอว่านโยบายแทนที่ข้อมูลที่มุ่งลดค่าใช้จ่ายของการโหลดข้อมูลควรพิจารณาสถิติจำนวนครั้งที่ข้อมูลถูกร้องขอในอดีตตั้งแต่ข้อมูลถูกบันทึกไว้ในแคชด้วย เพราะการเลือกแทนที่ข้อมูลที่ถูกร้องขอไม่บ่อยแม้มีขนาดใหญ่อาจช่วยลดค่าโหลดข้อมูลได้มากกว่าการเลือกแทนที่ข้อมูลขนาดเล็กที่ถูกร้องขอบ่อยกว่าก่อนได้ ข้อเสนอนี้สอดคล้องกับการค้นพบของ Arlitt et al. (2000) ที่ว่านโยบายการแทนที่ข้อมูลที่พิจารณาปัจจัยจำนวนครั้งที่ข้อมูลถูกร้องขอสามารถให้อัตราขนาดรวมของข้อมูลที่พบในแคชได้สูงกว่านโยบายที่ไม่พิจารณาปัจจัยดังกล่าว และยังสอดคล้องกับข้อพิสูจน์ของ Chen, Xiao, and Shen (2006) ที่ว่านโยบายแทนที่ข้อมูลที่ดีควรพยายามเก็บข้อมูลที่ถูกร้องขอบ่อยไว้ในแคชให้นานเพราะมีความเป็นไปได้สูงที่จะถูกร้องขออีก ในขณะที่ข้อมูลที่ถูกร้องขอน้อยครั้งมักไม่ถูกร้องขออีก การค้นพบหรือข้อพิสูจน์นี้แท้จริงคือหลักการที่เรียกว่า Spatial Locality ซึ่งหมายถึงการที่ข้อมูลที่ถูกร้องขอบ่อยครั้งมีโอกาสสูงที่จะถูกร้องขอซ้ำอีก

ข้อเสนอข้างต้นเป็นที่มาของสมมติฐานข้อหนึ่งของงานวิจัยนี้ว่านโยบายการแทนที่ข้อมูลในเว็บแคชที่เลือกแทนข้อมูลที่มีสถิติบ่งชี้ว่าช่วยลดค่าโหลดข้อมูล (ซึ่งเป็นผลคูณของขนาดข้อมูล, อัตราค่าโหลดของแต่ละข้อมูล และจำนวนครั้งที่ข้อมูลดังกล่าวถูกร้องขอตั้งแต่ถูกบันทึกไว้ในแคช) ได้มากก่อน น่าจะสามารถลดค่าใช้จ่ายของการโหลดข้อมูลจากกลุ่มเมฆได้มากกว่านโยบายที่เลือกแทนที่ข้อมูลที่มีสถิติบ่งชี้ว่าช่วยลดค่าโหลดข้อมูลได้น้อยกว่า อย่างไรก็ตาม สมมติฐานนี้อาจมีผลกระทบด้านลบกล่าวคือ ข้อมูลที่เคยช่วยประหยัดค่าโหลดข้อมูลได้มากและยังไม่หมดอายุมีแนวโน้มที่จะไม่ถูกเลือกแทนที่และคงอยู่ในแคชเป็นเวลานานแม้ว่าจะไม่ถูกร้องขออีกต่อไป ก่อให้เกิดมลพิษในแคช (Cache pollution) (Arlitt, et al. 2000) ทำให้การใช้พื้นที่สำหรับจัดเก็บข้อมูลในแคชไม่มีประสิทธิภาพ เพื่อเป็นการป้องกันปัญหาดังกล่าว งานวิจัยนี้จะนำระยะเวลาที่ข้อมูลถูกร้องขอครั้งล่าสุดมาพิจารณาร่วมเพื่อทำให้ข้อมูลในแคชที่เคยช่วยประหยัดค่าโหลดข้อมูลได้มากและยังไม่หมดอายุแต่ค้างอยู่ในแคชนานมีโอกาสถูกเลือกแทนที่โดยข้อมูลใหม่เร็วขึ้น การใช้ระยะเวลาที่ข้อมูลถูกร้องขอครั้งล่าสุดยังอาจช่วยเพิ่มประสิทธิภาพของนโยบายแทนที่ข้อมูลใน

เว็บแคชถ้าคำร้องขอข้อมูลที่ส่งมายังแคชมีแนวโน้มที่จะเกิดขึ้นซ้ำในระยะเวลาอันสั้น (Temporal locality) (Podlipnig and Böszörmenyi, 2003)

2) ระยะเวลาของการโหลดข้อมูล

เป้าหมายสำคัญอีกประการหนึ่งของเว็บแคชนอกจากการลดปริมาณข้อมูลที่โหลดผ่านระบบเครือข่ายคือการเพิ่มความเร็วในการโหลดข้อมูลของผู้ใช้ โดยเฉพาะในยุคกลุ่มเมฆที่เดสก์ท็อปแอพพลิเคชั่น (Desktop applications) มีวิวัฒนาการด้านรูปแบบไปสู่บริการซอฟต์แวร์กลุ่มเมฆมากขึ้น ทำให้การโต้ตอบระหว่างผู้ใช้และเดสก์ท็อปแอพพลิเคชั่นที่เดิมเป็นไปอย่างรวดเร็วเพราะไม่ผ่านเครือข่ายคอมพิวเตอร์เปลี่ยนรูปแบบเป็นการโต้ตอบผ่านทางเครือข่ายซึ่งใช้เวลานานขึ้น นโยบายการแทนที่ข้อมูลในเว็บแคชที่เหมาะกับบริการกลุ่มเมฆจึงควรพิจารณาต้นทุนทางเทคนิคระยะเวลาของการโหลดข้อมูลซึ่งเกี่ยวข้องโดยตรงกับการลดระยะเวลาการตอบสนองผู้ใช้ (User response time) อันเป็นความต้องการพื้นฐานของผู้ใช้บริการซอฟต์แวร์กลุ่มเมฆเชิงโต้ตอบ (Interactive SaaS) ด้วยการเลือกแทนที่ข้อมูลที่มีระยะเวลาของการโหลดจากแม่ข่ายน้อยก่อนข้อมูลที่ใช้เวลาในการโหลดมาก ทฤษฎีการลดระยะเวลาการตอบสนองผู้ใช้ด้วยการเลือกแทนที่ข้อมูลที่ใช้เวลาในการโหลดน้อยก่อนถูกเสนอและพิสูจน์ในงานวิจัยของ Wooster and Abrams (1997) และ Jin and Bestavros (2000: 254-261) อย่างไรก็ตาม ระยะเวลาของการโหลดข้อมูลต้นฉบับเดิมจากแม่ข่ายในครั้งต่อๆ ไปเมื่อเกิด Cache miss อาจไม่เท่าเดิมโดยแปรเปลี่ยนตามสภาพความคั่งของเครือข่าย (Network congestion) และภาระของแม่ข่าย (Server load) ในแต่ละช่วงเวลา งานวิจัยของ Jin and Bestavros (2000: 254-261) และงานวิจัยของ Shin et al. (2003) เสนอว่าระยะเวลาของการโหลดข้อมูลแต่ละชิ้นควรเป็นผลรวมของเวลาแฝง (Latency) เฉลี่ยที่ใช้ในการสถาปนาการเชื่อมต่อกันทางเครือข่ายระหว่างเว็บแคชกับแม่ข่ายและอัตราส่วนระหว่างขนาดของข้อมูลที่โหลดกับความเร็วแท้จริงในการส่งข้อมูลผ่านเครือข่าย (Network throughput) ระหว่างเว็บแคชกับแม่ข่ายซึ่งขึ้นอยู่กับแบนด์วิดท์และความคั่งเฉลี่ยของเส้นทางเครือข่ายที่ถูกใช้ในการโหลดข้อมูลรวมถึงภาระเฉลี่ยของแม่ข่ายเป็นสำคัญ ส่วนเวลาแฝงของการปิดการเชื่อมต่อกันทางเครือข่ายระหว่างแคชกับแม่ข่ายไม่ถูกพิจารณาเพราะเกิดขึ้นภายหลังจากแคชได้รับข้อมูลที่ร้องขอสมบูรณ์แล้ว การวัดความเร็วเฉลี่ยแท้จริงของการส่งข้อมูลผ่านเครือข่ายระหว่างเว็บแคชกับแต่ละแม่ข่ายของผู้ให้บริการกลุ่มเมฆทุกรายที่องค์กรใช้บริการอยู่สามารถดำเนินการโดยใช้ซอฟต์แวร์วิเคราะห์

โพรโตคอลเครือข่าย (Haugdahl, 2000) เช่น ping, Iperf (NLANR/DAST, 2010) หรือฟังก์ชั่น Throughput Graph ของ Wireshark (Wireshark Foundation, 2011)

3) อายุใช้งานคงเหลือของข้อมูล

เนื่องจากข้อมูลมีอายุการใช้งานที่บ่งบอกถึงระยะเวลาก่อนเนื้อหาของข้อมูลจะถูกเปลี่ยนแปลงแก้ไขหรือยุติการใช้งาน กอปรกับผู้ผลิตข้อมูลต้นฉบับสามารถใช้คำสั่งเอชทีเอ็มแอลทำการกำหนดวันหมดอายุหรือระยะเวลาที่อนุญาตให้ใช้งานได้ของข้อมูล นโยบายการแทนที่ข้อมูลในเว็บแคชจึงควรพิจารณาปัจจัยวันหมดอายุของข้อมูล ซึ่งถูกระบุอยู่ในฟิลด์ส่วนหัว Expires ของโพรโตคอลเอชทีทีพีที่ใช้ในการส่งสำเนาของข้อมูลต้นฉบับมายังแคช หรืออายุใช้งานได้ของข้อมูล ซึ่งถูกระบุอยู่ในฟิลด์ส่วนหัว Max-age ของโพรโตคอลเอชทีทีพี (Fielding, et al. 1999) ทั้งนี้เพื่อทำการเลือกแทนที่ข้อมูลที่หมดอายุแล้วหรืออายุใช้งานคงเหลือสั้น (Short freshness lifetime first) เป็นการป้องกันมิให้มีข้อมูลหมดอายุค้างอยู่ในแคชเป็นการสิ้นเปลืองพื้นที่จัดเก็บของแคช ทฤษฎีการเลือกแทนที่ข้อมูลดังกล่าวเป็นส่วนหนึ่งของงานวิจัยของ Wessels (1995) งานวิจัยของ Fielding et al. (1999) เสนอแนวทางการคำนวณอายุใช้งานคงเหลือไว้ดังมีรายละเอียดคือ ถ้าคำสนองเอชทีทีพี (HTTP Response) ที่ตอบกลับจากแม่ข่ายมายังแคชมีฟิลด์ Max-age ให้กำหนดอายุใช้งานคงเหลือ (Freshness lifetime) เท่ากับค่า Max-age แต่ถ้าคำสนองเอชทีทีพีไม่มีฟิลด์ Max-age แต่มีฟิลด์ Expires ให้กำหนดอายุใช้งานคงเหลือเท่ากับค่า Expires ลบด้วยวันเวลาที่คำสนองถูกสร้างขึ้นที่แม่ข่ายซึ่งถูกระบุอยู่ในฟิลด์ส่วนหัว Date ของโพรโตคอลเอชทีทีพี อย่างไรก็ตาม ข้อมูลบางประเภท ได้แก่ เว็บเพจแบบพลวัต (Dynamic web pages) ซึ่งจะไม่ถูกอ่านใช้ซ้ำ (จึงไม่ควรถูกเก็บไว้ในแคช) และข้อมูลสถิต (Static data) ที่ไม่ถูกกำหนดวันหมดอายุหรืออายุใช้งาน (เนื่องจากอายุการใช้งานนานมาก) จะไม่มีทั้งฟิลด์ Max-age และ Expires ปรากฏในคำสนองเอชทีทีพี งานวิจัยนี้กำหนดอายุใช้งานคงเหลือของข้อมูลประเภทเหล่านี้ให้เท่ากับค่าโดยปริยาย 0 วินาที

ผลการวิจัยที่เกี่ยวข้อง

ในปัจจุบันมีงานวิจัยนโยบายการแทนที่ข้อมูลในเว็บแคชจำนวนมาก ซึ่งหากมิได้ระบุแหล่งข้อมูลอ้างอิงเป็นอย่างอื่นนโยบายที่สังเคราะห์สรุปไว้ในหัวข้อนี้อ้างอิงมาจากงานวิจัยเชิง

สำรวจของ Podlipnig and Böszörmenyi (2003) การทบทวนวรรณกรรมต่อไปนี้ถูกจัดเรียงตามลำดับหัวข้อย่อยในหัวข้อที่ 2.2

1) ขนาดของข้อมูลที่ร้องขอ, อัตราค่าโหลดข้อมูล และจำนวนครั้งที่ข้อมูลถูกร้องขอ

นโยบายการแทนที่ข้อมูลในเว็บแคชสามารถถูกแยกประเมินตามตัวแปรต้นของนโยบาย คือ ขนาดของข้อมูล ได้เป็นสองกลุ่มดังนี้

กลุ่มนโยบายการแทนที่ข้อมูลในเว็บแคชที่ไม่มีการพิจารณาขนาดของข้อมูลที่จะทำการแทนที่ ได้แก่ LRU, LFU-DA, EXP1, Value-Aging, HLRU, LFU, LFU-Aging, α-Aging, swLFU, SLRU, Generational Replacement, LRU*, LRU-Hot, Server-assisted cache replacement, LR, RAND, LRU-C, Randomized replacement with general value functions, ARC (Megiddo and Modha, 2004), CSOPT (Jeong and Dubois, 2006), LA2U (Chen, et al., 2006), LAUD (Chen, et al., 2006), SEMALRU (Geetha, et al., 2009), LRU-SLFR (Shin, Kim, and Jang, 2003) และ BHRO (Shi and Zhang, 2008) นโยบายเหล่านี้มีเป้าหมายเพื่อลดระยะเวลาตอบสนองต่อผู้ใช้ให้ได้บ่อยครั้งที่สุด (ยกเว้น BHRO ที่มุ่งลดปริมาณข้อมูลที่ต้องส่งผ่านระบบเครือข่าย) โดยพิจารณาหลายปัจจัยยกเว้นขนาดของข้อมูลในการเลือกข้อมูลที่จะถูกแทนที่ ดังนั้นถ้าเกิดสถานการณ์ที่ข้อมูลขนาดใหญ่ในแคชที่กำลังจะถูกร้องขอถูกแทนที่ด้วยข้อมูลอื่นบ่อยครั้ง ก็จะทำให้เว็บแคชต้องโหลดข้อมูลขนาดใหญ่นี้ใหม่จากแม่ข่ายซึ่งถ้าตั้งอยู่ด้านผู้ให้บริการกลุ่มเมฆบ่อยครั้งด้วย ทำให้อัตราประหยัดค่าโหลดข้อมูลต่ำ ดังนั้นนโยบายในกลุ่มนี้จึงไม่มีประสิทธิภาพในการลดค่าโหลดข้อมูลในการใช้บริการกลุ่มเมฆ

กลุ่มที่สองเป็นนโยบายการแทนที่ข้อมูลในเว็บแคชที่พิจารณาขนาดของข้อมูลในการแทนที่ ได้แก่ GDSF, LRU-Threshold, SIZE, LOG2-SIZE, LRU-Min, PSS, LRU-LSC, Partitioned Caching, HYPER-G, CSS, LRU-SP, GD-Size, GD*, TSP, MIX, M-Metric, HYBRID, LNC-R-W3, LNC-R-W3-U (Shim, et al., 1999), LRV, LUV, HARMONIC, LAT, GDSP, LRU-S, SE (Sarma and Govindarajan, 2003), R-LPV (Chand, et al., 2006), Min-SAUD (Xu, et al., 2004), OPT (Yin, et al., 2005), LPPB-R (Kim and Park, 2001), OA (Li, et al., 2007), NNPCR-2 (ElAarag and Romano, 2009), CSP (Triantafillou, et al., 2003), GA-Based Cache Replacement Policy (Chen, Li, and Wang, 2004), นโยบายที่เสนอโดย Bolot and Hoschka (1996), HRO (Shi

and Zhang, 2008) และ LRO (Shi and Zhang, 2008) นโยบายเกือบทั้งหมดนี้ (ยกเว้น LRU-Min, M-Metric, NNPCR-2 และนโยบายของ Bolot and Hoschka (1996) ดังจะอธิบายในลำดับถัดไป) พิจารณาเลือกแทนที่ข้อมูลที่มีขนาดใหญ่ก่อนเสมอซึ่งถึงแม้จะทำให้อัตราการพบข้อมูลในแคชสูงขึ้นแต่อัตราขนาดรวมของข้อมูลที่พบในแคชจะลดลง นโยบายในกลุ่มนี้จึงไม่มีประสิทธิภาพในการประหยัดค่าใช้บริการกลุ่มเมฆ นโยบาย LRU-Min พยายามลบข้อมูลที่มีขนาดใหญ่กว่าข้อมูลที่แคชโหลดมาใหม่ก่อนแต่ถ้าไม่มีข้อมูลที่ใหญ่กว่าข้อมูลที่กำลังต้องการบันทึกลงในแคชจึงทำการลบกลุ่มของข้อมูลที่มีขนาดเล็กลงเพื่อให้ได้พื้นที่พอสำหรับการบันทึกข้อมูล อย่างไรก็ตามนโยบายนี้ไม่พิจารณาจำนวนครั้งการร้องขอข้อมูลและไม่รองรับอัตราค่าโหลดข้อมูล สำหรับนโยบาย M-Metric เป็นนโยบายที่อนุญาตให้เว็บแคชเลือกแทนที่ข้อมูลขนาดเล็กก่อนใหญ่ได้แต่ไม่รองรับ (ไม่มีตัวแปรต้นที่สามารถนำมาประยุกต์ใช้เป็น) ปัจจัยอัตราค่าโหลดข้อมูลที่แตกต่างกันสำหรับแต่ละผู้ให้บริการกลุ่มเมฆ และนโยบาย NNPCR-2 เลือกข้อมูลที่จะถูกแทนที่โดยอาศัยหลักการ Neural Network เพื่อเรียนรู้จากชุดตัวอย่างข้อมูลนำเข้าซึ่งรวมถึงขนาดของข้อมูลและชุดตัวอย่างผลลัพธ์ที่สอดคล้อง นโยบายนี้ไม่รองรับปัจจัยจำนวนครั้งการร้องขอข้อมูลและอัตราค่าโหลดข้อมูล นโยบายของ Bolot and Hoschka (1996) เลือกแทนที่ข้อมูลที่มีขนาดเล็กก่อนแต่ไม่มีการพิจารณาจำนวนครั้งของการร้องขอข้อมูล

จากการศึกษางานวิจัยข้างต้นสรุปได้ว่า นโยบายการแทนที่ข้อมูลในเว็บแคชปัจจุบันเกือบทั้งหมดมุ่งเน้นการลดจำนวนครั้งของการโหลดข้อมูลจากแม่ข่ายหรือระยะเวลาตอบสนองต่อผู้ใช้ที่ร้องขอข้อมูล ยังไม่มีนโยบายใดถูกออกแบบให้มีเป้าหมายของการลดค่าใช้จ่ายการโหลดข้อมูลของเว็บแคช

2) ระยะเวลาของการโหลดข้อมูล

นโยบายการแทนที่ข้อมูลในเว็บแคชหลายนโยบายทำการลดระยะเวลาการตอบสนองผู้ใช้โดยเลือกแทนที่ข้อมูลที่มีต้นทุนทางเทคนิคเกี่ยวกับระยะเวลาของการโหลดข้อมูลต่ำก่อน เช่น นโยบาย GD-Size, GDSF, GD* (Jin and Bestavros, 2000) และ GDSP (Jin and Bestavros, 2000: 254-261) ใช้จำนวนกลุ่มข้อมูล (Packets) ที่ใช้บรรทุกข้อมูลที่โหลดจากแม่ข่ายเพื่อสะท้อนถึงต้นทุนทางเทคนิคด้านระยะเวลาการโหลด กล่าวคือ ถ้าข้อมูลมีขนาดใหญ่จะทำให้มีจำนวนกลุ่มข้อมูลที่ต้องส่งผ่านเครือข่ายจำนวนมาก บอกเป็นนัยถึงระยะเวลาการโหลดข้อมูลที่มากด้วย

อย่างไรก็ตาม การเลือกแทนที่ข้อมูลโดยพิจารณาเพียงจำนวนกลุ่มข้อมูลเพื่อประมาณระยะเวลาของการโหลดข้อมูลอาจขาดความแม่นยำเป็นอย่างมากเพราะข้อมูลที่มีจำนวนกลุ่มข้อมูลเท่ากันอาจใช้ระยะเวลาของการโหลดไม่เท่ากันถ้าถูกส่งผ่านเส้นทางในระบบเครือข่าย (Network path) คนละเส้นทางที่มีความเร็วของการส่งข้อมูลไม่เท่ากัน, นโยบาย HYBRID (Wooster and Abrams, 1997) และนโยบาย LAT (Wooster and Abrams, 1997) พิจารณาระยะเวลาของการโหลดข้อมูลขนาดต่างๆ จากแม่ข่ายมายังเว็บแคชโดยคำนวณจากระยะเวลาเฉลี่ยในการสถาปนาการเชื่อมต่อและแบนด์วิดท์ของการเชื่อมต่อเครือข่ายระหว่างแม่ข่ายและเว็บแคช อย่างไรก็ตาม ระยะเวลาที่ได้จากการประมาณนี้อาจมีความคลาดเคลื่อนเพราะแม้ว่าข้อมูลที่มีขนาดเท่ากันที่ถูกโหลดผ่านเส้นทางเครือข่ายที่มีแบนด์วิดท์เท่ากันก็อาจใช้ระยะเวลาการโหลดต่างกันหากเส้นทางหนึ่งมีความคั่งของข้อมูลเครือข่ายรวมกับภาระของแม่ข่ายมากกว่าอีกเส้นทางหนึ่ง, นโยบาย LUV (Bahn, et al., 2002) และ MIX (Niclausse, et al. 1998) พิจารณาระยะเวลาในการส่งข้อมูลจากแม่ข่ายมายังแคชซึ่งสามารถนำมาจากชุดข้อมูลคำร้องขอการโหลดข้อมูลเว็บในอดีต (Cache proxy traces) แต่ทั้งสองนโยบายไม่พิจารณาระยะเวลาที่ใช้ในการสถาปนาการเชื่อมต่อ, นโยบาย LNC-R-W3 (Scheuermann, et al., 1997) ใช้ระยะเวลาของการโหลดข้อมูลซึ่งเป็นผลต่างระหว่างเวลาที่ข้อมูลถูกโหลดมาบันทึกไว้ในแคชและเวลาที่คำร้องขอข้อมูลไปถึงแม่ข่ายโดยค่าเวลาเหล่านี้นำมาจากชุดข้อมูลคำร้องขอการโหลดข้อมูลเว็บในอดีต จึงเป็นนโยบายที่ไม่พิจารณาระยะเวลาในการสถาปนาการเชื่อมต่อ, นโยบาย LNC-R-W3-U พิจารณาระยะเวลาของการโหลดข้อมูลซึ่งเป็นค่าผลรวมถ่วงน้ำหนักระหว่างระยะเวลาที่วัดครั้งล่าสุดในการโหลดข้อมูลตัวอย่างและระยะเวลาการโหลดข้อมูลจริงที่วัดและบันทึกไว้ในอดีต อย่างไรก็ตาม เป็นการยากที่จะหาค่าคงที่ที่เหมาะสมสำหรับใช้ในการถ่วงน้ำหนัก, นโยบาย LRU-SLFR คำนวณระยะเวลาของการโหลดข้อมูลจากแม่ข่ายมายังเว็บแคชจากระยะเวลาเฉลี่ยในการสถาปนาการเชื่อมต่อ (ซึ่งคำนวณจากการหาผลรวมถ่วงน้ำหนักระยะเวลาในการสถาปนาการเชื่อมต่อที่วัดค่าสะสมไว้ในอดีต) และความเร็วแท้จริงเฉลี่ยของการส่งข้อมูลผ่านเครือข่ายจากแม่ข่ายมายังเว็บแคช (ซึ่งคำนวณจากการหาผลรวมถ่วงน้ำหนักความเร็วแท้จริงของการส่งข้อมูลผ่านเครือข่ายที่วัดค่าสะสมไว้ในอดีต) ค่าในอดีตนี้ถูกวัดเก็บไว้สำหรับแต่ละแม่ข่ายที่ถูกร้องขอข้อมูล (Per-server basis) ทำให้มีระดับการปรับขนาดได้ (Scalability) สูงกว่าการวัดเก็บไว้สำหรับข้อมูลทุกชิ้นที่โหลดมาแม้จากแม่ข่ายเดียวกัน (Per-document basis) อย่างไรก็ตาม เป็น

การยากที่จะหาค่าคงที่ที่เหมาะสมสำหรับใช้ในการถ่วงน้ำหนัก, นโยบาย GDSP นอกเหนือจากรุ่นที่พิจารณาจำนวนกลุ่มข้อมูลดังกล่าวแล้วข้างต้น ยังมีรุ่นที่คำนวณระยะเวลาของการโหลดข้อมูลจากแม่ข่ายมายังเว็บแคชจากระยะเวลาเฉลี่ยในการสถาปนาการเชื่อมต่อและความเร็วเฉลี่ยแท้จริงของการส่งข้อมูลผ่านเครือข่ายจากแม่ข่ายมายังเว็บแคชซึ่งวิเคราะห์จากชุดข้อมูลคำร้องขอการโหลดข้อมูลเว็บในอดีต แนวทางนี้จะถูกนำมาใช้ในงานวิจัยนี้

3) อายุใช้งานคงเหลือของข้อมูล

นโยบายการแทนที่ข้อมูลในเว็บแคชที่พิจารณาปัจจัยที่เกี่ยวข้องกับวันหมดอายุหรือระยะเวลาที่อนุญาตให้ใช้งานได้ของข้อมูลก่อนเกิดการเปลี่ยนแปลงเนื้อหาได้แก่ นโยบายการแทนที่ข้อมูลในเว็บแคช LA2U (Chen, et al., 2006), LAUD (Chen, et al., 2006) และ LNC-R-W3-U นโยบายเหล่านี้พิจารณาความถี่ของการเปลี่ยนแปลงข้อมูลต้นฉบับโดยเลือกแทนที่ข้อมูลที่ต้นฉบับของมันถูกเปลี่ยนแปลงบ่อยก่อนข้อมูลที่ต้นฉบับไม่ค่อยถูกเปลี่ยนแปลง อย่างไรก็ตาม Chen, et al. (2006) ไม่ได้อธิบายวิธีการหาความถี่ของการเปลี่ยนแปลงข้อมูลต้นฉบับของนโยบาย LA2U และ LAUD ไว้ ส่วนนโยบาย LNC-R-W3-U คำนวณความถี่ของการเปลี่ยนแปลงข้อมูลต้นฉบับจากค่าเฉลี่ยของระยะเวลาระหว่างการตรวจพบแต่ละครั้งว่าข้อมูลต้นฉบับมีการเปลี่ยนแปลงซึ่งรู้ได้จากการทดสอบค่าในฟิลด์ Last-Modified ของโพรโตคอลเอชทีทีพีที่เป็นคำสนองกลับมายังแคช ข้อเสียของวิธีนี้คือถึงแม้ว่าข้อมูลต้นฉบับจะมีการเปลี่ยนแปลงถี่ ถ้าผู้ใช้มีการเรียกใช้ข้อมูลนี้ไม่บ่อย เว็บแคชจะส่งคำร้องขอไปยังแม่ข่ายไม่บ่อยด้วย ทำให้จำนวนครั้งที่ตรวจพบการเปลี่ยนแปลงค่าในฟิลด์ Last-Modified ในคำสนองไม่บ่อยเช่นกัน ความถี่ของการเปลี่ยนแปลงข้อมูลต้นฉบับที่วัดได้จึงต่ำกว่าความเป็นจริง นอกจากนี้การพิจารณาความถี่ของการเปลี่ยนแปลงข้อมูลต้นฉบับยังไม่เหมาะกับเว็บเพจแบบพลวัตซึ่งหมดอายุทันทีที่ส่งไปยังผู้ใช้ เว็บแคชจึงไม่ควรเสียเวลาคำนวณความถี่ของข้อมูลประเภทนี้ ปัญหาเหล่านี้ไม่พบเมื่อใช้อายุใช้งานคงเหลือของข้อมูลแทนความถี่ของการเปลี่ยนแปลงข้อมูลดังในนโยบายที่เสนอโดย Bolot and Hoschka (1996)

นโยบายที่ใกล้เคียงกับกรอบความคิดของงานวิจัยที่เสนอนี้มากที่สุดเป็นผลวิจัยของ Bolot and Hoschka (1996) ซึ่งมีการพิจารณาปัจจัยต่างๆ ได้แก่ ระยะเวลาตั้งแต่ข้อมูลถูกเข้าถึงครั้งล่าสุดถึงปัจจุบัน, ขนาดของข้อมูล, ระยะเวลาของการโหลดข้อมูล และอายุใช้งานคงเหลือของข้อมูล

อย่างไรก็ตาม นโยบายดังกล่าวถูกสร้างและวัดผลโดยปราศจากการนำอายุใช้งานคงเหลือของข้อมูลมาใช้จริง อีกทั้งไม่สามารถรองรับปัจจัยอัตราค่าโหลดข้อมูลที่แตกต่างกันสำหรับแต่ละผู้ให้บริการกลุ่มเมฆได้

สรุป

กรอบความคิดทางทฤษฎีของงานวิจัยฉบับนี้มีสาระสำคัญว่า ปัจจัยพื้นฐานต่างๆ ที่ถูกพิจารณาในนโยบายการแทนที่ข้อมูลในเว็บแคชต่างๆ ที่มุ่งเน้นการลดค่าใช้จ่ายการโหลดข้อมูลจากกลุ่มเมฆควรรวมถึง

1) ขนาดของข้อมูลที่ร้องขอ ซึ่งจำเป็นต่อการคำนวณค่าใช้บริการกลุ่มเมฆ
2) อัตราค่าโหลดข้อมูลจากผู้ให้บริการกลุ่มเมฆที่คิดตามขนาดข้อมูลที่โหลด ซึ่งจะต้องไม่ใช่ค่าคงที่สำหรับข้อมูลทุกชิ้นเนื่องจากแต่ละองค์กรสามารถใช้บริการกลุ่มเมฆจากหลายผู้ให้บริการพร้อมๆ กัน ทำให้ข้อมูลขนาดเดียวกันที่เว็บแคชโหลดมาจากแม่ข่ายของผู้ให้บริการกลุ่มเมฆต่างกันอาจมีค่าใช้จ่ายของการโหลดแตกต่างกัน งานวิจัยนี้เลือกแทนที่ข้อมูลที่มีค่าโหลด(ซึ่งขึ้นอยู่กับขนาดของข้อมูลที่ร้องขอและอัตราค่าโหลด)ต่ำก่อน
3) ระยะเวลาของการโหลดข้อมูล ซึ่งเป็นองค์ประกอบหนึ่งของระยะเวลาการตอบสนองผู้ใช้ในการใช้บริการซอฟต์แวร์กลุ่มเมฆที่กำลังมีจำนวนเพิ่มขึ้นอย่างต่อเนื่อง งานวิจัยนี้เลือกแทนที่ข้อมูลที่มีระยะเวลาการโหลดจากแม่ข่ายสั้นก่อน
4) อายุใช้งานคงเหลือของข้อมูลในแคช ซึ่งเกี่ยวข้องโดยตรงกับพื้นที่จัดเก็บข้อมูลของเว็บแคช งานวิจัยนี้เลือกแทนที่ข้อมูลที่มีอายุใช้งานคงเหลือน้อยก่อน

จากการศึกษาวรรณกรรมที่เกี่ยวข้องกับงานวิจัยนี้ พบว่า

1) มีนโยบายที่พิจารณาขนาดข้อมูลในการแทนที่ซึ่งส่วนใหญ่รวมทั้งนโยบายที่ใช้แพร่หลายในปัจจุบันเลือกแทนที่ข้อมูลขนาดใหญ่ก่อนซึ่งอาจช่วยลดระดับค่าโหลดข้อมูลจากกลุ่มเมฆได้น้อยกว่าการเลือกแทนที่ข้อมูลขนาดเล็กก่อน ส่วนนโยบายที่แทนที่ข้อมูลขนาดเล็กก่อนก็มิได้พิจารณาจำนวนครั้งที่ข้อมูลถูกร้องขอ

2) มีนโยบายที่รองรับอัตราค่าโหลดข้อมูลในรูปแบบของค่าคงที่ที่คูณกับขนาดข้อมูล (นโยบายอื่นเป็นค่าคงที่หารกับขนาดข้อมูลจึงไม่สามารถใช้เป็นอัตราค่าโหลดข้อมูลได้) อย่างไรก็ตาม ค่าคงที่ในนโยบายดังกล่าวเป็นค่าคงที่เดียวเดียวสำหรับข้อมูลที่โหลดทั้งหมดซึ่งจะเกิดปัญหาถ้าข้อมูลถูกโหลดมาจากผู้ให้บริการกลุ่มเมฆหลายรายที่คิดอัตราค่าโหลดต่างกัน
3) มีนโยบายที่พิจารณาระยะเวลาของการโหลดข้อมูลในหลายแนวทาง แนวทางที่คำนวณระยะเวลาของการโหลดข้อมูลจากแม่ข่ายมายังเว็บแคชโดยการหาผลรวมของระยะเวลาในการสถาปนาการเชื่อมต่อและอัตราส่วนระหว่างขนาดของข้อมูลที่โหลดกับความเร็วแท้จริงของการส่งข้อมูลผ่านเครือข่ายจากแม่ข่ายยังเว็บแคช ถูกนำมาใช้ในงานวิจัยนี้
4) มีนโยบายที่พิจารณาอายุใช้งานคงเหลือของข้อมูล แต่อยู่ในระดับแนวคิดที่ยังไม่ถูกสร้างให้ใช้งานได้จริงและยังไม่ผ่านการวัดผล

นอกจากนี้ นโยบายการแทนที่ข้อมูลในเว็บแคชที่มีอยู่มิได้มีการพิจารณาปัจจัยพื้นฐานตามกรอบความคิดของงานวิจัยนี้อย่างบูรณาการ