บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันมีเว็บไซต์มากมายอยู่บนอินเตอร์เน็ตซึ่งเว็บไซต์เหล่านี้ถูกสร้างขึ้นเพื่อวัตถุประสงค์ที่แตกต่างกัน เช่น ใช้เป็นช่องทางในการติดต่อสื่อสาร ใช้เป็นช่องทางในการติดต่อธุรกิจ ใช้เป็นช่องทางในการเผยแพร่ข้อมูลข่าวสาร เป็นต้น และเนื่องจากในปัจจุบันการสร้างเว็บไซต์เป็นเรื่องง่ายเพราะมีเครื่องมือสำหรับใช้สร้างเว็บไซต์ให้เลือกอย่างมากมายซึ่งเป็นสาเหตุหนึ่งที่ทำให้อัตราการเกิดของเว็บไซต์ใหม่เพิ่มขึ้นอย่างรวดเร็ว แต่ข้อมูลที่สำคัญซึ่งเจ้าของเว็บไซต์ส่วนมากต้องการทราบคือ มีใครเข้ามาเยี่ยมชมเว็บไซต์บ้าง มีจำนวนคนเข้ามาเยี่ยมชมมากน้อยเท่าไร เข้ามาเยี่ยมชมเมื่อไรและเข้ามาทำอะไรบนเว็บไซต์บ้าง โดยเราสามารถทราบข้อมูลดังกล่าวได้โดยทำการวิเคราะห์ล็อกไฟล์ซึ่งบันทึกกิจกรรมต่างๆที่เกิดขึ้นบนเว็บเซิร์ฟเวอร์ที่เป็นที่ตั้งของเว็บไซต์

เครื่องมือวิเคราะห์ล็อกไฟล์และแทร็กเกอร์ (tracker) ที่มีอยู่ส่วนมากจะรายงานกิจกรรมของผู้ใช้บริการในเว็บเซิร์ฟเวอร์ซึ่งการใช้เครื่องมือดังกล่าวจะทำให้สามารถทราบถึงจำนวนครั้งในการเยี่ยมชมไฟล์ต่างๆในเว็บเซิร์ฟเวอร์ ช่วงเวลาของการเยี่ยมชม โดเมนเนมและ URL ของผู้ชม โดยที่แทร็กเกอร์ที่มีอยู่ในปัจจุบันแสดงรายงานโดยใช้ตัววัดประสิทธิภาพของเว็บไซต์ดังนี้ ฮิต (hit) คือจำนวนครั้งของการกระทำที่เกิดขึ้นกับไฟล์แต่ละไฟล์ที่อยู่บนเว็บเซิร์ฟเวอร์ในช่วงระยะเวลาหนึ่ง (วัน เดือน ปี) เพจวิว (page view) คือจำนวนครั้งที่เว็บเพจถูกเยี่ยมชมโดยไม่นับจำนวนครั้งของการร้องขอจากการรีโหลดเว็บเพจ ยูสเซอร์เซสชัน (user session) คือจำนวนผู้ชมจริงๆที่เข้าเยี่ยมชมเว็บไซต์ โดยทั่วไปแล้วแทร็กเกอร์เหล่านี้ถูกออกแบบมาเพื่อใช้กับเว็บเซิร์ฟเวอร์ที่มีอัตราการเข้าใช้บริการที่ไม่สูงนักและแทบไม่มีการวิเคราะห์หาความสัมพันธ์ระหว่างเว็บเพจที่ถูกเยี่ยมชมและลำดับการเยี่ยมชมเว็บเพจเลย

ความสัมพันธ์ระหว่างเว็บเพจและลำดับการเยี่ยมชมจะสามารถวิเคราะห์ได้จากล็อกไฟล์ ดังนั้นแทร็กเกอร์ที่ดีต้องสามารถค้นหาความสัมพันธ์ในหมู่เว็บเพจ รูปแบบลำดับของการเยี่ยมชม และสามารถจำแนกผู้ชมโดยพิจารณาจากข้อมูลการเยี่ยมชมได้โดยอัตโนมัติ นอกจากนี้แทร็กเกอร์ที่ดียังต้องสามารถค้นหาความสัมพันธ์ในเว็บเซิร์ฟเวอร์ที่มีอัตราการเข้าใช้บริการข้อมูลสูงและมีล็อกไฟล์ขนาดใหญ่ (ซึ่งอาจเป็นไปได้ที่จะอยู่กระจายกัน)

การทำเหมืองข้อมูลเป็นกระบวนการในการหาความสัมพันธ์ที่น่าสนใจจากกลุ่มของข้อมูลจำนวนมากโดยใช้เทคนิคทางปัญญาประดิษฐ์หรือเทคนิคทางสถิติมาประมวลผลและเมื่อไม่นานมานี้ได้มีนักวิจัยหลายคนเสนอการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลเพื่อทำการค้นหาข้อมูลบนอินเตอร์เน็ต จุดสำคัญของข้อเสนอเหล่านี้คือการค้นหาความรู้จากข้อมูลที่อยู่บนอินเตอร์เน็ต

โดยวิเคราะห์รูปแบบการเยี่ยมชมของผู้ใช้บริการบนเครื่องเว็บเซิร์ฟเวอร์เพียงเครื่องเดียวและใช้ล็อกไฟล์ของเว็บเซิร์ฟเวอร์เพื่อทดสอบการประยุกต์ใช้งานของการทำเหมืองข้อมูลนี้ ดังนั้นการทำเหมืองเว็บจึงเป็นการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูลกับข้อมูลเว็บ ถึงอย่างไรก็ตามก็ไม่ใช่เรื่องง่ายที่จะปรับอัลกอริทึมที่มีอยู่ให้สามารถใช้ได้กับข้อมูลเว็บเพราะว่ามีข้อจำกัดของแบบจำลองทางไคลเอนท์-เซิร์ฟเวอร์ซึ่งมีลักษณะเฉพาะหลายอย่างในอินเตอร์เน็ตและความแตกต่างในการจัดการข้อมูลเว็บทางกายภาพกับทางจินตภาพ ดังนั้นจึงทำให้มีความจำเป็นที่จะต้องพัฒนากรอบงานใหม่ซึ่งรวมไปถึงสถาปัตยกรรม แบบจำลองข้อมูลและอัลกอริทึมเพื่อทำการสร้างเครื่องมือการทำเหมืองเว็บที่มีประสิทธิภาพ

งานวิจัยนี้จะมุ่งเน้นวิธีการสร้างเครื่องมือเพื่อหาความสัมพันธ์ระหว่างเว็บเพจและรูปแบบลำดับการเยี่ยมชมเว็บเพจโดยทำการวิเคราะห์จากล็อกไฟล์ของเว็บเซิร์ฟเวอร์

1.2 วัตถุประสงค์ของการวิจัย

เพื่อค้นหารูปแบบการเยี่ยมชมเว็บไซต์โดยทำการประมวลผลจากล็อกไฟล์ของการเข้าถึงเว็บเซิร์ฟเวอร์

1.3 ขอบเขตของการวิจัย

1. ทำการวิเคราะห์รูปแบบการเยี่ยมชมเว็บไซต์บนเครื่องเว็บเซิร์ฟเวอร์เพียงเครื่องเดียวเท่านั้น
2. นำล็อกไฟล์ของการเข้าถึงเว็บเซิร์ฟเวอร์แบบเอ็กซ์เทนด์มาประมวลผลเท่านั้น
3. ทำการค้นหากฎความสัมพันธ์(association rule)ในการเข้าถึงเว็บเพจจากล็อกไฟล์
4. ทำการค้นหารูปแบบลำดับ(sequential pattem)ในการเข้าถึงเว็บเพจจากล็อกไฟล์
5. พัฒนาโปรแกรมให้สามารถประมวลผลได้บนเว็บเซิร์ฟเวอร์ที่เป็นเครื่องคอมพิวเตอร์ส่วนบุคคล

1.4 ขั้นตอนและวิธีการดำเนินการวิจัย

1. ศึกษางานวิจัยที่เกี่ยวข้อง
2. ศึกษารูปแบบของล็อกไฟล์และเว็บเซิร์ฟเวอร์ของเว็บไซต์
3. ศึกษาความต้องการของผู้ใช้เกี่ยวกับรูปแบบและข้อมูลของรายงานที่ต้องการ
4. ศึกษาทฤษฎีและแนวทางที่ใช้ในการประมวลผล
5. ออกแบบและพัฒนาวิธีการหารูปแแบบการท่องเว็บไซต์
6. ออกแบบวิธีการทดสอบขั้นตอนวิธี

7. ทดสอบและปรับปรุงคุณภาพของขั้นตอนวิธี
8. สรุปผลการวิจัย และจัดทำรายงานวิทยานิพนธ์

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. ช่วยให้สามารถวิเคราะห์ความนิยมของเว็บไซต์ได้
2. ช่วยให้รู้ข้อมูลเกี่ยวกับผู้ชมที่เข้ามายังเว็บไซต์
3. ช่วยแสดงลำดับการเยี่ยมชมว่าผู้ชมไปที่เว็บเพจใดในเว็บไซต์

1.6 โครงสร้างของวิทยานิพนธ์

เนื้อหาของวิทยานิพนธ์ฉบับนี้ถูกแบ่งออกเป็น 6 บทดังนี้ คือ บทที่ 1 เป็นบทนำ บทที่ 2 จะกล่าวถึงทฤษฎีและงานวิจัยต่าง ๆ ที่เกี่ยวข้อง เช่น รูปแบบของล็อกไฟล์ การเรียนรู้กฎความสัมพันธ์ เป็นต้น บทที่ 3 กล่าวถึงขั้นตอนวิธีการประมวลผลข้อมูลเบื้องต้น ส่วนบทที่ 4 จะกล่าวถึงขั้นตอนวิธีการเรียนรู้กฎความสัมพันธ์และรูปแบบลำดับการท่องเว็บไซต์ บทที่ 5 กล่าวถึงการทดลองและผลการทดลองของขั้นตอนวิธีที่นำเสนอ และบทที่ 6 ซึ่งเป็นบทสุดท้ายจะเป็นบทสรุปของการวิจัย รวมทั้งข้อเสนอแนะต่างๆในการพัฒนาขั้นตอนวิธีการหารูปแบบการท่องเว็บไซต์ให้ดียิ่งขึ้น