ปุณณวัฒน์ ธาดาภาคย์ 2554: การระบุเว็บโฮสต์ไทยโดยใช้เครื่องจักรเรียนรู้ ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์ อานนท์ รุ่งสว่าง, Ph.D. 79 หน้า

เว็บเพจบนอินเทอร์เน็ตถือว่าเป็นแหล่งความรู้ขนาดใหญ่ และมีการเพิ่มจำนวนขึ้นอย่างต่อเนื่อง จนทำให้ไม่สามารถเก็บรวบรวมเว็บเพจทั้งหมดได้ภายใต้ทรัพยากรระบบที่มีอยู่อย่างจำกัด เช่น ระยะเวลา พื้นที่เก็บข้อมูล หรือแบนด์วิดท์ เป็นต้น ด้วยเหตุนี้การใช้เว็บคราวเลอร์แบบเจาะจงภาษาจึงเป็นทางเลือกหนึ่งที่จะมาช่วยให้สามารถเลือกเก็บเฉพาะเว็บเพจที่เขียนด้วยภาษาตามที่ต้องการได้ ตัวอย่างการนำไปใช้งานเช่น ผู้ให้บริการระบบสืบค้นในแต่ละประเทศมักจะแสดงเว็บเพจผลลัพธ์ตามภาษาประจำชาติของผู้ใช้ เพื่อให้ผู้ใช้เกิดความพึงพอใจต่อระบบ หรือการสร้างคลังเว็บเพจประจำชาติ เพื่อเก็บรักษาเว็บเพจที่มีข้อมูลสำคัญด้านสังคม วัฒนธรรม วิถีการดำเนินชีวิตให้กับคนรุ่นต่อไป

วิทยานิพนธ์ฉบับนี้นำเสนอระบบต้นแบบเว็บคราวเลอร์เจาะจงเว็บเพจภาษาไทยแบบแยกเก็บตามไซต์ ซึ่งประกอบด้วยส่วนของการทำนายหรือระบุภาษาของโฮสต์ และส่วนของเว็บคราวเลอร์แบบแยกเก็บตามไซต์ โดยที่ส่วนของการทำนายภาษาใช้เทคนิคทางเครื่องจักรเรียนรู้แบบหลายตัว ซึ่งเครื่องจักรเรียนรู้แต่ละตัวจะเรียนรู้จากคุณลักษณะทั้งตัวอย่างของเว็บโฮสต์ที่เกี่ยวข้อง และเว็บโฮสต์ที่ไม่เกี่ยวข้อง ด้วยเทคนิคทางเครื่องจักรเรียนรู้ที่แตกต่างกัน และส่วนทำนายภาษายังมีหน้าที่ช่วยระบุภาษาของเว็บโฮสต์เป้าหมายพร้อมทั้งให้คะแนนความมั่นใจอีกด้วย แต่สำหรับเว็บคราวเลอร์แบบแยกเก็บตามไซต์มีหน้าที่เก็บเว็บเพจภายใต้เว็บไซต์ที่กำหนดเรียงลำดับตามคะแนนความมั่นใจที่ได้จากส่วนการทำนายภาษา จากผลการทดลองแสดงให้เห็นว่าประสิทธิภาพการเก็บเว็บเพจจากอินเทอร์เน็ตของวิธีการที่นำเสนอดีกว่าวิธีการของเว็บคราวเลอร์เจาะจงภาษาในระดับเว็บเพจที่ผ่านมา ซึ่งวัดผลจากอัตราการเก็บเกี่ยวเว็บเพจภาษาไทย

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก