

เอกสิทธิ์ ศรีสุขะ 2552: การเก็บเว็บเพจแบบเฉพาะเจาะจงภาษาไทยโดยใช้ฐานความรู้
ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรม
คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก:
ผู้ช่วยศาสตราจารย์อานนท์ รุ่งสว่าง, Ph.D. 90 หน้า

เว็บคราเวลอร์แบบเฉพาะเจาะจงภาษาคือเครื่องมือที่ใช้สำหรับเก็บรวบรวมเว็บเพจที่
เขียนขึ้นจากภาษาที่ผู้ใช้กำหนด สำหรับการเก็บรวบรวมเว็บเพจภาษาไทยนั้น วิธีที่ง่ายที่สุดคือ
กำหนดให้เว็บคราเวลอร์เก็บรวบรวมเว็บเพจตามอินเทอร์เน็ตโดเมนท้องถิ่นในประเทศไทย
เช่น .th เป็นต้น อย่างไรก็ตามปัญหาหลักคือเว็บเพจภาษาไทยส่วนใหญ่จำนวนมากว่าครึ่งหนึ่ง
อยู่ภายนอกขอบเขตของโดเมนดังกล่าว ดังนั้นวิธีการนี้จึงยังไม่ครอบคลุมถึงปริมาณเว็บเพจ
ภาษาไทยส่วนใหญ่ที่มีอยู่ทั้งหมดได้ วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการค้นหาเว็บเพจภาษาไทยซึ่ง
จะคำนวณเพื่อหาค่าความน่าจะเป็นที่ลิงค์ถัดไปจะเป็นเว็บเพจภาษาไทยโดยใช้หลักทฤษฎีความ
น่าจะเป็นแบบเบย์อย่างง่ายมาคำนวณกับฐานความรู้ที่ได้จากการเก็บรวบรวมเว็บเพจมาในครั้ง
แรก คะแนนความน่าจะเป็นดังกล่าวจะถูกนำมาใช้ในการจัดลำดับของยูอาร์แอลในยูอาร์แอลคิว
เพื่อให้เว็บคราเวลอร์เลือกเก็บเว็บเพจจากยูอาร์แอลที่มีคะแนนสูงสุดก่อนนั่นเอง จากผลการ
ทดลองพบว่าวิธีการที่นำเสนอดังกล่าวให้ผลในด้านการเก็บเกี่ยวและความครอบคลุมได้ดีกว่า
วิธีการอื่น