การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวมกราฟแสดงความสัมพันธ์ของคำศัพท์กับการโปรแกรมตรรกะเชิงอุปนัย

โดย

ภริตา บุญปลูก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวมกราฟแสดงความสัมพันธ์ของคำศัพท์กับการโปรแกรมตรรกะเชิงอุปนัย

โดย

ภริตา บุญปลูก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

Similarity Web Pages Classification by Combining Relation Graphs to Inductive Logic Programming

By

PARITA BOONPLOOK

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2550

มหาวิทยาลัยธรรมศาสตร์
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

ภริตา บุญปลูก

เรื่อง

การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวมกราฟแสดงความสัมพันธ์ของคำศัพท์กับการโปรแกรมตรรกะเชิงอุปนัย

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต

เมื่อ วันที่ 5 พฤษภาคม 2551

ประธานกรรมการวิทยานิพนธ์ _____
(ดร. รัชฎา คงคะจันทร์)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์ _____
(ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ)

กรรมการวิทยานิพนธ์ _____
(ดร. เด่นดวง ประดับสุวรรณ)

กรรมการวิทยานิพนธ์ _____
(ดร. ชลวิท นัทธี)

คณบดี _____
(รองศาสตราจารย์ สมชาย วิริยะยุทธกร)

บทคัดย่อ

การเติบโตอย่างรวดเร็วของหน้าเว็บทำให้เกิดปัญหาในการค้นหาสารสนเทศในหน้าเว็บเหล่านั้น ซึ่งทางแก้ปัญหาหนึ่งของปัญหานี้ คือการจัดหน้าเว็บเหล่านี้ให้เป็นกลุ่มของความสนใจ วิธีการทั่วไปที่ใช้จัดกลุ่มหน้าเว็บคือการใช้โครงสร้างของหน้าเว็บเหล่านั้น เช่น ใช้ไฮเปอร์ลิงค์ คำสำคัญ แท็กเมตา เป็นต้น ซึ่งเป็นลักษณะสำคัญที่ใช้ในการจัดกลุ่ม ในบทความนี้ผู้วิจัยได้นำเสนอวิธีการที่ใช้การโปรแกรมตรรกะเชิงอุปนัย (Inductive Logic Programming) หรือ ILP เพื่อจำแนกคู่ของหน้าเว็บที่เหมือนกันโดยใช้ข้อความจากหน้าเว็บประกอบกับพจนานุกรมเชิงความหมายจากเวิร์ดเน็ต (WordNet) โดยนำมาเขียนใหม่เป็นกราฟความสัมพันธ์ของคำศัพท์ที่กำหนดขึ้นโดยผู้วิจัย ผลการทดลองพบว่าวิธีการนำเสนอให้อัตราความถูกต้องสูงกว่าวิธีการอื่นที่ทดสอบในการทดลองครั้งนี้

Abstract

The rapid growth of web pages makes difficulties for searching for information stored in them. One solution of this problem is to organize those web pages into several groups of interest. The common method in organizing the pages is a use of structure of the pages, for example, hyperlinks, keywords, meta tag, etc., as important features to cluster them. In this paper, we propose a novel method which uses Inductive Logic Programming (ILP) to classify tuples of similar pages using the text in the pages combining to semantic dictionary from WordNet which are rewritten in our relation graph format. The results obtained from the experiments show that the proposed method yields the better accuracy rate than other methods run in our experiments.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จได้ด้วยความอนุเคราะห์ และความช่วยเหลืออย่างยิ่งจากอาจารย์ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ซึ่งให้ข้อคิด แนวทาง และคำปรึกษาตลอดจนเป็นผู้ตรวจทานแก้ไข อีกทั้งสละเวลาอันมีค่า ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ดร. รัชฎา คงคะจันทร์ ดร. เด่นดวง ประดับสุวรรณ และ ดร. ชลวิท นัทธี ประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำในการแก้ไขวิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น ขอขอบพระคุณคณาจารย์ในภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ทุกท่านที่ประสิทธิ์ประสาทความรู้อันมีค่ายิ่งแก่ผู้วิจัย

ขอขอบพระคุณธนาคารออมสินที่ให้โอกาสและมอบทุนการศึกษาในการศึกษาระดับปริญญาโทในครั้งนี้

ที่สำคัญที่สุดกราบขอบพระคุณ คุณพ่อ คุณแม่ ตลอดจนญาติพี่น้องทุกคน รวมทั้งเพื่อนๆ ที่เป็นแรงผลักดัน คอยให้กำลังใจ ให้ความช่วยเหลือ ให้ความรักและความเข้าใจที่ดี เพื่อให้ข้าพเจ้าได้ทำงานวิจัยนี้จนกระทั่งสำเร็จลุล่วงไปได้ด้วยดี

ข้าพเจ้าขอขอบพระคุณเป็นอย่างสูง แก่ผู้ที่ได้กล่าวถึงข้างต้นและมิได้กล่าวถึงทุกท่าน ขออำนาจของคุณพระศรีรัตนตรัยและสิ่งศักดิ์สิทธิ์ทั้งปวง จงดลบันดาลให้ทุกท่านมีความสุขสวัสดิ์ สุขภาพพลานามัยแข็งแรง ตลอดไปด้วยเทอญ

ภริตา บุญปลูก
นักศึกษาระดับบัณฑิตศึกษา
ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์
พ.ศ. 2551

สารบัญ

สารบัญตาราง

สารบัญภาพประกอบ