



การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวม  
กราฟแสดงความสัมพันธ์ของคำศัพท์  
กับการโปรแกรมตรรกะเชิงอุปนัย

โดย

ภริตา บุญปลุก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
วิทยาศาสตรมหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์  
พ.ศ. 2550

การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวม  
กราฟแสดงความสัมพันธ์ของคำศัพท์  
กับการโปรแกรมตรรกะเชิงอุปนัย

โดย

ภริตา บุญปลุก

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
วิทยาศาสตร์มหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์  
พ.ศ. 2550

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

By

PARITA BOONPLOOK

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
วิทยาศาสตร์มหาบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์  
พ.ศ. 2550

มหาวิทยาลัยธรรมศาสตร์  
คณะวิทยาศาสตร์และเทคโนโลยี

วิทยานิพนธ์

ของ

ภริตา บุญปลูก

เรื่อง

การจัดหมวดหมู่เว็บเพจที่คล้ายคลึงกันโดยการรวบรวมกราฟแสดงความสัมพันธ์ของคำศัพท์  
กับการโปรแกรมตรรกะเชิงอุปนัย

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
วิทยาศาสตรมหาบัณฑิต

เมื่อ วันที่ 5 พฤษภาคม 2551

ประธานกรรมการวิทยานิพนธ์

\_\_\_\_\_  
(ดร. รัชฎา คงคะจันทร์)

กรรมการและอาจารย์ที่ปรึกษาวิทยานิพนธ์

\_\_\_\_\_  
(ผู้ช่วยศาสตราจารย์ ดร. สุกวี สิ้นธุภิณฺโญ)

กรรมการวิทยานิพนธ์

\_\_\_\_\_  
(ดร. เด่นดวง ประดับสุวรรณ)

กรรมการวิทยานิพนธ์

\_\_\_\_\_  
(ดร. ชลวิทย์ นันทิ)

คณบดี

\_\_\_\_\_  
(รองศาสตราจารย์ สมชาย วิริยะยุทธกร)

## บทคัดย่อ

การเติบโตอย่างรวดเร็วของเว็บไซต์ทำให้เกิดปัญหาในการค้นหาสารสนเทศในหน้าเว็บเหล่านั้น ซึ่งทางแก้ปัญหานี้ของปัญหานี้คือการจัดหน้าเว็บเหล่านั้นให้เป็นกลุ่มของความสนใจ 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

วิทยานิพนธ์นี้สำเร็จได้ด้วยความช่วยเหลืออย่างยิ่งจากอาจารย์ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ซึ่งให้ข้อคิด แนวทาง และคำปรึกษาตลอดจนเป็นผู้ตรวจทานแก้ไข อีกทั้งสละเวลาอันมีค่า ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง ขอขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณ ดร. รัชฎา คงคะจันทร์ ดร. เด่นดวง ประดับสุวรรณ และ ดร. ชลวิทย์ นันทิ์ ประธานกรรมการและกรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำแนะนำในการแก้ไข วิทยานิพนธ์ให้มีคุณภาพยิ่งขึ้น ขอขอบพระคุณคณาจารย์ในภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ทุกท่านที่ประสิทธิ์ประสาทความรู้ อันมีค่าแก่ผู้วิจัย

ขอขอบพระคุณธนาคารออมสินที่ให้โอกาสและมอบทุนการศึกษาในการศึกษาระดับปริญญาโทในครั้งนี้

ที่สำคัญที่สุดกราบขอบพระคุณ คุณพ่อ คุณแม่ ตลอดจนญาติพี่น้องทุกคน รวมทั้งเพื่อนๆ ที่เป็นแรงผลักดัน คอยให้กำลังใจ ให้ความช่วยเหลือ ให้ความรักและความเข้าใจที่ดี เพื่อให้ข้าพเจ้าได้ทำงานวิจัยนี้จนกระทั่งสำเร็จลุล่วงไปได้ด้วยดี

ข้าพเจ้าขอขอบพระคุณเป็นอย่างสูง แก่ผู้ที่ได้กล่าวถึงข้างต้นและมีได้กล่าวถึงทุกท่าน ขออำนาจของคุณพระศรีรัตนตรัยและสิ่งศักดิ์สิทธิ์ทั้งปวง จงดลบันดาลให้ทุกท่านมีความสุขสวัสดิ์ สุขภาพพลานามัยแข็งแรง ตลอดไปด้วยเทอญ

ภริตา บุญปลุก

นักศึกษาระดับบัณฑิตศึกษา

ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์

พ.ศ. 2551

## สารบัญ

|   | หน้า |
|---|------|
| บทคัดย่อ .....                                  | (2)  |
| กิตติกรรมประกาศ.....                            | (4)  |
| สารบัญตาราง.....                                | (8)  |
| สารบัญภาพประกอบ .....                           | (9)  |
| บทที่   |      |
| 1. บทนำ .....                                   | 1    |
| 1.1 ความเป็นมาและความสำคัญของงานวิจัย .....     | 1    |
| 1.2 วัตถุประสงค์.....                           | 4    |
| 1.3 ขอบเขตของงานวิจัย.....                      | 5    |
| 1.4 ผลที่คาดว่าจะได้รับ .....                   | 5    |
| 1.5 รายละเอียดของวิทยานิพนธ์.....               | 6    |
| 2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง .....          | 7    |
| 2.1 ทฤษฎีที่เกี่ยวข้อง.....                     | 7    |
| 2.1.1 การสืบค้นข้อมูลบนเว็บ.....                | 7    |
| 2.1.1.1 เสิร์ชเอนจินแบบโต้ตอบ.....              | 8    |
| 2.1.1.2 ไดร็กทอรี.....                          | 8    |
| 2.1.1.3 เมตาเสิร์ชเอนจิน.....                   | 9    |
| 2.1.2 สถาปัตยกรรมของการสืบค้นข้อมูลบนเว็บ ..... | 9    |
| 2.1.2.1 เครื่องมือรวบรวมหน้าเว็บ.....           | 9    |
| 2.1.2.2 ตัวทำดัชนี .....                        | 9    |
| 2.1.2.3 ตัวสืบค้น.....                          | 10   |
| 2.1.2.4 การจัดอันดับ .....                      | 10   |
| 2.1.3 เว็บเชิงความหมาย .....                    | 11   |

|         |   |    |
|---------|---|----|
| 2.1.3.1 | เอ็กซ์เอ็มแอล.....                                  | 11 |
| 2.1.3.2 | Uniform Resource Identifier .....                   | 13 |
| 2.1.3.3 | Resource Description Framework .....                | 13 |
| 2.1.3.4 | Web Ontology Language.....                          | 16 |
| 2.1.4   | การทำเหมืองข้อมูลเอกสาร .....                       | 18 |
| 2.1.4.1 | แบบจำลองการแปลงข้อความ .....                        | 18 |
| 2.1.4.2 | เครื่องมือวัดความเหมือนกัน .....                    | 20 |
| 2.1.4.3 | ขั้นตอนการทำเหมืองข้อมูล .....                      | 21 |
| 2.1.4.4 | การจัดกลุ่มเอกสาร .....                             | 22 |
| 2.1.4.5 | วิธีการประเมินผล .....                              | 23 |
| 2.1.5   | โปรแกรมตรรกะเชิงอุปนัย .....                        | 23 |
| 2.1.5.1 | วิธีการโปรแกรมตรรกะเชิงอุปนัย.....                  | 23 |
| 2.1.5.2 | เทคนิคไอน์แอลพีพื้นฐาน.....                         | 27 |
| 2.1.6   | ทฤษฎีกราฟ และสมมติฐานของกราฟย่อย .....              | 32 |
| 2.1.7   | ฐานความรู้ภิกษานศัพท์เวิร์ดเน็ต.....                | 33 |
| 2.2     | งานวิจัยที่เกี่ยวข้อง .....                         | 35 |
| 3.      | วิธีการดำเนินงานวิจัย .....                         | 40 |
| 3.1     | กราฟความสัมพันธ์ของคำศัพท์แนวคิด .....              | 4  |
| 3.2     | การตัดคำหลักและสกัดคำหลัก.....                      | 41 |
| 3.3     | การค้นหาคำหลัก.....                                 | 42 |
| 3.4     | การจำแนกข้อมูลด้วยกฎจากโปรแกรมตรรกะเชิงอุปนัย ..... | 43 |
| 3.4.1   | การสร้างความรู้ภูมิหลัง.....                        | 43 |
| 3.4.1.1 | กลุ่มโครงสร้างกราฟ.....                             | 43 |
| 3.4.1.2 | กลุ่มความหมายของคำ .....                            | 44 |
| 3.4.1.3 | กลุ่มแปลงหน้าเว็บเป็นกราฟ.....                      | 44 |
| 3.4.2   | การสร้างตัวอย่างบวก และตัวอย่างลบ.....              | 45 |
| 3.4.3   | การสอน และการหากฎ.....                              | 45 |
| 3.4.4   | ค่าความแม่นยำ.....                                  | 45 |

|   |    |
|---|----|
| 3.4.5 การทดสอบ .....  | 45 |
| 3.5 วิธีการทดสอบความถูกต้องแบบไขว้ข้าม .....  | 46 |
| 4. ผลการทดลอง .....   | 48 |
| 4.1 วิธีขั้นตอนในการทดลอง.....  | 48 |
| 4.1.1 กลุ่มข้อมูลตัวอย่างที่ใช้ในการทดลอง.....  | 48 |
| 4.1.2 ชุดข้อมูลที่ใช้ในการทดลอง .....   | 49 |
| 4.1.3 ค่าที่ใช้วัดผลในการทดลอง.....   | 49 |
| 4.1.3.1 ค่าความถูกต้องของการจำแนกข้อมูล.....  | 49 |
| 4.1.3.2 จำนวนกฎที่เกิดขึ้นในขั้นตอนการเรียนรู้.....   | 49 |
| 4.2 ข้อมูลคำค้นที่ใช้ในการทดลอง.....  | 50 |
| 4.3 ผลการทดลอง .....  | 51 |
| 5. สรุปผลการศึกษาและข้อเสนอแนะ.....   | 54 |
| 5.1 สรุปการศึกษา .....  | 54 |
| 5.2 ข้อเสนอแนะ.....   | 55 |
| ภาคผนวก   |    |
| ก. ตัวอย่างชุดข้อมูลเรียนรู้และกราฟแสดงความสัมพันธ์คำศัพท์แนวคิด..                          | 57 |
| ข. การใช้เครื่องมือในการสร้างกฎด้วยวิธีการโปรแกรมตรรกะเชิงอุปนัย<br>ด้วยโปรแกรม Aleph ..... | 62 |
| บรรณานุกรม .....  | 64 |

## สารบัญตาราง

| ตารางที่ |   | หน้า |
|----------|---|------|
| 2.1      | แสดงโครงสร้างครอปรวในการเรียนรู้.....   | 26   |
| 2.2      | แสดงจำนวนคำ, จำนวนกลุ่มคำศัพท์, และจำนวนความหมายที่ถูกจัดเก็บ<br>และแสดงผลในเวิร์ดเน็ต (เวอร์ชัน 2.1).....        | 34   |
| 4.1      | แสดงค่าความถูกต้องที่ได้จากการเรียนรู้ของโปรแกรมตรรกะเชิงอุปนัย .....   | 50   |
| 4.2      | แสดงผลเปรียบเทียบค่าความแม่นยำ.....   | 51   |
| 4.3      | แสดงกฎที่ได้หลังจากการเรียนรู้ คำค้น "jaguar"<br>ในการเรียนรู้ชุดที่ 3 ด้วย minpos=2.....                         | 52   |
| 4.4      | แสดงกฎที่ได้หลังจากการเรียนรู้ คำค้น "jaguar"<br>โดยเพิ่มคุณลักษณะพิเศษ ในการเรียนรู้ชุดที่ 3 ด้วย minpos=2 ..... | 52   |
| ก.1      | แสดงคำค้นจำแนกโดเมนที่ใช้ในการทดลอง .....   | 57   |

## สารบัญภาพประกอบ

| ภาพที่ |   | หน้า |
|--------|---|------|
| 2.1    | แสดงสถาปัตยกรรมของการสืบค้นข้อมูล (บรินและเพจ, 1998) .....        | 10   |
| 2.2    | แสดงข้อความในรูปแบบเอกสารเอ็กซ์เอ็มแอล .....                      | 12   |
| 2.3    | แสดง XML Declaration .....  | 12   |
| 2.4    | แสดงเอกสาร XML .....  | 13   |
| 2.5    | แสดงโครงสร้างของ Triple .....                                     | 14   |
| 2.6    | แสดงการแทนข้อมูลในโครงสร้าง RDF (Triple) .....                    | 14   |
| 2.7    | แสดงการแทนข้อมูลในโครงสร้าง RDF (N-Triple) .....                  | 15   |
| 2.8    | แสดงการอธิบายโครงสร้างของเมตาดาตา.....                            | 16   |
| 2.9    | แสดง OWL Document ของ Music Ontology.....                         | 17   |
| 2.10   | แสดงความสมบูรณ์และความสอดคล้องกันของสมมติฐาน.....                 | 25   |
| 2.11   | แสดงส่วนหนึ่งของกราฟรีไฟน์เมนต์ของปัญหาความสัมพันธ์ของครอบครัว .. | 31   |
| 2.12   | แสดงกราฟย่อย G1 เป็นสมมติฐานของกราฟ G2 .....                      | 33   |
| 2.13   | กราฟแสดงความหมายและความสัมพันธ์ของข้อมูล ของ Eric Miller.....     | 37   |
| 3.1    | แสดงลำดับขั้นฐานความรู้ของ Apple .....                            | 40   |
| 3.2    | แสดงเอกสารบนเว็บ .....  | 41   |
| 3.3    | แสดงการเรียนรู้แบบสมบูรณ์ของโปรแกรมตรรกะเชิงอุปนัย.....           | 43   |
| 3.4    | แสดงกระบวนการทำงานจำแนกเว็บ .....                                 | 47   |
| 4.1    | แสดงการจัดแบ่งข้อมูลในการทดลอง .....                              | 48   |
| ก.1    | แสดงความสัมพันธ์ของคำค้น Apple .....                              | 58   |
| ก.2    | แสดงความสัมพันธ์ของคำค้น Jaguar.....                              | 59   |
| ก.3    | แสดงความสัมพันธ์ของคำค้น Mouse .....                              | 60   |
| ก.4    | แสดงความสัมพันธ์ของคำค้น Virus .....                              | 61   |