

บทคัดย่อ

รหัสโครงการ: RSA/08/2544

ชื่อโครงการ: การทำเหมืองเว็บไทยโดยเทคนิคการเรียนรู้ของเครื่องและการโปรแกรมตรรกะเชิงอุปนัย

ชื่อนักวิจัย: นายบุญเสริม กิจศิริกุล

E-mail Address: boonserm.k@chula.ac.th

ระยะเวลาโครงการ: 1 ธค. 2543 – 30 พย. 2546

ปัจจุบันการเติบโตของอินเทอร์เน็ตเป็นไปอย่างรวดเร็วมาก มีเว็บเพจจำนวนมากหลายพันล้านเพจที่เข้าถึงได้บนอินเทอร์เน็ตและมีเว็บเพจหลายล้านเพจเกิดขึ้นใหม่ทุกวัน ผู้ใช้ข้อมูลบนอินเทอร์เน็ตต้องใช้เวลาและความพยายามอย่างมากในการค้นหาเอกสารที่ต้องการ ระบบค้นหาเว็บในปัจจุบันครอบคลุมเอกสารเพียงบางส่วนของเอกสารทั้งหมดบนอินเทอร์เน็ต ระบบค้นหาเว็บเหล่านี้มักค้นคืนได้เอกสารที่ไม่ตรงกับความต้องการของผู้ใช้เพราะใช้การค้นหาตามคำสำคัญ ส่วนระบบไดเรกทอรีโครงข่าย เช่น Yahoo! จัดโครงสร้างของเว็บเพจแยกตามหมวดหมู่ของเว็บเพจ ทำให้สามารถค้นคืนเอกสารได้ตรงกับความต้องการของผู้ใช้มากกว่า อย่างไรก็ตามวิธีการไดเรกทอรีโครงข่ายก็มีข้อจำกัดที่ปริมาณเว็บเพจที่ครอบคลุมจะน้อยมาก เนื่องจากต้องใช้แรงงานคนจำนวนมากในการแบ่งหมวดหมู่ของเอกสาร

ในงานวิจัย เรานำเสนอวิธีการเพื่อแก้ปัญหานี้โดยการจำแนกประเภทของเว็บเพจออกเป็นหมวดหมู่โดยอัตโนมัติ วิธีการที่นำเสนอนี้ใช้เทคนิคของการเรียนรู้ของเครื่องและการโปรแกรมตรรกะเชิงอุปนัย หัวข้อสำคัญที่เน้นทำวิจัยคือ (1) การวิจัยพื้นฐานเพื่อเพิ่มประสิทธิภาพของการโปรแกรมตรรกะเชิงอุปนัย และ (2) การวิจัยเทคนิคการเรียนรู้ของเครื่องที่สามารถใช้ประโยชน์จากข้อมูลแบบไม่มีฉลาก

การโปรแกรมตรรกะเชิงอุปนัย (ไอแอลพี) สามารถนำมาประยุกต์ใช้กับการจำแนกเว็บเพจเป็นหมวดหมู่โดยอัตโนมัติได้ และมีจุดเด่นอยู่ที่ผู้สอนสามารถป้อนความรู้เบื้องต้นในรูปแบบของโปรแกรมตรรกะอันดับที่หนึ่งได้ ซึ่งจะช่วยให้การจำแนกข้อมูลทำได้อย่างมีประสิทธิภาพยิ่งขึ้น ไอแอลพีจะให้เอนโทรปีเป็นเซตของกฎที่สอดคล้องกับตัวอย่างสอน อย่างไรก็ตามไอแอลพีมีข้อด้อยที่กฎที่สร้างได้อาจไม่ตรงพอดีกับตัวอย่างทดสอบโดยเฉพาะข้อมูลที่มีสัญญาณรบกวน และทำให้ข้อมูลเหล่านี้ไม่สามารถจำแนกหมวดหมู่ได้อย่างถูกต้อง ในกรณีเช่นนี้ เราจำเป็นต้องใช้วิธีการที่สามารถหากกฎที่ตรงกับข้อมูลมากที่สุด ในงานวิจัยนี้ เราใช้กระบวนการดึงลักษณะสำคัญและวิธีการแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์ก เพื่อหากกฎที่ตรงกับข้อมูลมากที่สุด ผลการทดลองที่ได้แสดงให้เห็นว่า วิธีการดึงลักษณะสำคัญและแบ็กพรอพาเกชันนิวรอลเน็ตเวิร์กทำให้การโปรแกรมตรรกะเชิงอุปนัยมีประสิทธิภาพสูงขึ้น

นอกจากนั้นในงานวิจัยนี้เรายังได้นำเสนอวิธีการเรียนรู้แบบใหม่ที่เรียกว่า การสอนไขว้แบบวนซ้ำ ซึ่งสามารถใช้ประโยชน์จากข้อมูลไม่มีฉลากได้ วิธีการนี้มีข้อดีว่าการเรียนรู้แบบสอนทั่วไปที่ต้องใช้ข้อมูลมีฉลากทั้งหมดและต้องอาศัยแรงงานคนจำนวนมากเพื่อติดฉลากให้กับข้อมูล แนวคิดของการสอนไขว้แบบวนซ้ำคือการรวมตัวแยกแยะย่อยสองตัว ซึ่งจะสอนโต้ตอบกันเองไปมาเพื่อปรับปรุงประสิทธิภาพของระบบโดยรวม เมื่อให้ข้อมูลไม่มีฉลากสองเซต แต่ละเซตสำหรับตัวแยกแยะย่อยแต่ละตัว ตัวแยกแยะจะติดฉลากให้กับตัวแยกแยะอีกตัว ด้วยการโต้ตอบที่ถี่ระหว่างตัวแยกแยะทั้งสอง ทำให้ประสิทธิภาพของระบบโดยรวมค่อยๆ ดีขึ้น ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอสามารถใช้ประโยชน์จากข้อมูลไม่มีฉลากได้อย่างมีประสิทธิภาพ เรายังได้ปรับปรุงประสิทธิภาพของการสอนไขว้แบบวนซ้ำโดยใช้การโปรแกรมตรรกะเชิงอุปนัยมาเป็นตัวแยกแยะย่อย ผลที่ได้พบว่า อัลกอริทึมการสอนไขว้แบบวนซ้ำชนิดการโปรแกรมตรรกะเชิงอุปนัยสามารถเพิ่มประสิทธิภาพของการสอนไขว้แบบวนซ้ำแบบดั้งเดิมได้อย่างมาก และมีประสิทธิภาพสูงกว่าวิธีการอื่นๆ ทุกวิธีที่นำทดสอบรวมทั้งการเรียนรู้แบบสอนซึ่งใช้ข้อมูลแบบมีฉลากทั้งหมดอีกด้วย

คำหลัก: การทำเหมืองเว็บ, การโปรแกรมตรรกะเชิงอุปนัย, การเรียนรู้ของเครื่อง

Abstract

Project Code: RSA/08/2544

Project Title: Thai Web Mining using Machine Learning and Inductive Logic Programming

Investigator: Boonserm Kijisirikul

E-mail Address: boonserm.k@chula.ac.th

Project Period: December 1, 2000 – November 30, 2003

With the explosive growth of the Internet, today there are billions Web page accessible on the Internet with several million pages being added daily. The user must spend a great deal of time and effort looking for document he needs. Web search engines available today cover only some fraction of all documents in the Internet, and because of the use of keyword search, they also return documents not related to the real user-interest. Net directory systems, such as Yahoo!, organize their Web resources in category-specific style and thus can provide documents better matching the user needs. However, these systems have the limitations that the number of Web pages covered by the systems is even small as they need a lot of human effort to categorize the documents.

In this research, we propose an approach to solving this problem by automatically classifying Web pages into categories. The proposed approach employs the techniques of machine learning and inductive logic programming. Two main issues are studied in this research; (1) a basic research for improving techniques of inductive logic programming, and (2) machine learning techniques that can make use of unlabeled data.

Inductive Logic Programming (ILP) can be applied to automatic classification of Web pages. It has advantage that the user can provide background knowledge in the form of first-order logic programs. This makes more efficient classification of data. Given training examples and background knowledge as the input, ILP outputs a set of rules that is consistent with training examples. However, ILP has disadvantage that the obtained rules may not be exactly match with test data, especially noisy data, and thus the data cannot be correctly classified. In such a case, we need a method that finds the best matching rule. In this work, we employ first-order feature extraction and backpropagation neural networks to find the best matching rule. The experimental results show that feature extraction and the neural network improve the performance of ILP alone.

In the research, we also propose a new learning method, called Iterative Cross-Training (ICT) which can make use of unlabeled data. The method has advantage over traditional supervised learning which needs all labeled data and requires a lot of human effort to label the data. The idea of ICT is to combine two sub-classifiers which iteratively train each other for improving the performance of the whole system. Given two sets of unlabeled data, each of which is for each classifier, the classifier labels data for the other. With good interaction between two classifiers, the performance of the whole system is increasingly improved. The experimental results show that the proposed method can effectively use unlabeled data. We then further improve the performance of ICT by employing an ILP system as a sub-classifier. The results show that ICT with ILP gives significant improvement over the original ICT, and performs better than other methods tested in our experiment including the supervised learning method which uses all labeled data.

Keywords: Web Mining, Inductive Logic Programming, Machine Learning