บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบโครงร่างสำหรับสร้างระบบแนะนำข้อมูลบนเว็บโดยใช้เทคนิคการกรองสารสนเทศโดยดูจากเนื้อหา (Content-Based Filtering) เทคนิคการกรองสารสนเทศที่นำเสนอนี้อาศัยการจำแนกหมวดหมู่ของเอกสารโดยการแทนเอกสารด้วยแบบจำลองหัวข้อ (Topic Model) นอกจากนี้ยังได้นำเสนอวิธีใหม่ในการจำแนกหมวดหมู่ของเอกสารโดยประยุกต์ใช้แบบจำลองหัวข้อร่วมกับการใช้โครงสร้างการเชื่อมโยงของหน้าเว็บ

ในโครงร่างที่นำเสนอนี้มีการวิจัยหาวิธีการใหม่สำหรับแนะนำข้อมูลโดยเน้นใน 3 ส่วน ได้แก่ (1) การนำเสนอวิธีการแทนเอกสารด้วยแบบจำลองหัวข้อแทนวิธีการแบบ Bag of Words (BOW) (2) การเพิ่มฟีเจอร์ให้กับหน้าเว็บปัจจุบันโดยใช้ฟีเจอร์ที่มาจากหน้าเว็บข้างเคียง และ (3) การนำเสนอโมเดลในการจำแนกแบบลำดับชั้นโดยประยุกต์ใช้ซัพพอร์ตเวกเตอร์แมชชีน

จากผลการวิจัยโดยใช้สารานุกรมวิกิพีเดียฉบับโรงเรียนพบว่า (1) การแทนเอกสารด้วยแบบจำลองหัวข้อให้ค่าความถูกต้องสูงกว่าการแทนเอกสารด้วยวิธี BOW 17.29 % และให้ค่าความถูกต้องสูงกว่าการแทนเอกสารด้วยวิธีการวิเคราะห์หาความหมายที่แอบแฝง 2.74% (2) การเพิ่มฟีเจอร์ที่มาจากหน้าเว็บข้างเคียงให้ค่าความถูกต้องสูงกว่าการแทนเอกสารด้วยแบบจำลองหัวข้อโดยใช้ฟีเจอร์จากหน้าเว็บปัจจุบัน 6.67% และ (3) โมเดลในการจำแนกแบบลำดับชั้นให้ค่าความถูกต้องในการจำแนกเท่ากับ 95.69% และให้ค่า F1 เท่ากับ 95.66% ซึ่งสูงกว่าโมเดลในการจำแนกแบบไม่เป็นลำดับชั้นโดยคิดเป็น 10.48 % และ 10.65% ตามลำดับ

229537

Abstract

The research objective is to design a framework for building a web-based recommender system using content-based filtering. The proposed content-based filtering is based on text categorization technique by using the topic model for document representation. In addition, we propose a novel text categorization by incorporating the topic model with the link structure of web pages.

The proposed framework consists of three new methodologies: (1) document representation approach using the topic model instead of the bag of words (BOW), (2) integration of additional features derived from neighboring pages for the current page, and (3) a hierarchical classification model based on the support vector machines.

The research study was evaluated by using Wikipedia Selection for Schools. The experimental results can be summarized as follows: (1) the document representation based on the topic model yields higher performance accuracy than the representation using BOW equal to 17.29 % , and higher than the representation using latent semantic analysis equal to 2.74%, (2) integrating additional features from neighboring pages gives higher performance accuracy equal to 6.67% compared to using only current page, and (3) the hierarchical classification model yields the accuracy of 95.69% and the F1 measure of 95.66%, which are equal to 10.48% and 10.65% improvement over the non-hierarchical model.