รายงานการวิจัย

การกำหนดหัวข้อข่าว โดยใช้ค่า TFIDF และ Zipf's Law ในการเรียนรู้แบบมีการสอน

News Topic Identification using TFIDF and Zipf's Law in Supervised Learning

ผู้วิจัย

ผศ.ดร. พรฤดี เนติโสภากุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ประจำปีงบประมาณ 2554

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงานการวิจัย

การกำหนดหัวข้อข่าว โดยใช้ค่า TFIDF และ Zipf's Law ในการเรียนรู้แบบมีการสอน

News Topic Identification using TFIDF and Zipf's Law in Supervised Learning

ผู้วิจัย

ผศ.ดร. พรฤดี เนติโสภากุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ประจำปีงบประมาณ 2554

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

กิตติกรรมประกาศ

งานวิจัยเรื่องการกำหนดหัวข้อข่าว โดยใช้ค่า TFIDF และ Zipf's Law ในการเรียนรู้แบบมีการสอนนี้ ได้รับทุนสนับสนุนการวิจัยจากเงินรายได้คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ประจำปี พ.ศ. 2554 ผู้วิจัยจึงขอกราบขอบพระคุณเป็นอย่างสูงมา ณ ที่นี้

พรฤดี เนติโสภากุล

บทคัดย่อ

ชื่อโครงการ (ภาษาไทย) การกำหนดหัวข้อข่าว โดยใช้ค่า TFIDF และ Zipf's Law ในการเรียนรู้แบบมีการสอน

ชื่อโครงการ (ภาษาอังกฤษ) News Topic Identification using TFIDF and Zipf's Law in Supervised Learning

แหล่งเงิน คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ประจำปีงบประมาณ 2554 จำนวนเงินที่ได้รับการสนับสนุน 48,000 บาท

ระยะเวลาการทำวิจัย ตั้งแต่ 1 ตุลาคม พ.ศ. 2553 ถึง 31 กรกฎาคม พ.ศ. 2554

ชื่อ-สกุล หัวหน้าโครงการ

ผศ.ดร. พรฤดี เนติโสภากุล คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง เบอร์โทรศัพท์ 02-723-4957 E-mail ponrudee@it.kmitl.ac.th

คำสำคัญ (Keywords) topic identification, Term Frequency Inverse Document Frequency, Zipf's Law

บทคัดย่อ

งานวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพการกำหนดหัวข้อให้กับเอกสารข่าวออนไลน์ โดยการวิเคราะห์จากค่าน้ำหนักของเทอมในเอกสาร ในการเปรียบเทียบประสิทธิภาพนั้น เป็นการเปรียบเทียบประสิทธิภาพผลลัพธ์จากวิธีการคำนวณหาค่าน้ำหนักของเทอมด้วย Term Frequency Inverse Document Frequency (TFIDF) กับวิธีอื่น ๆ ได้แก่ Chi-Square, Information Gain และ Term Frequency Inverse Document Frequency (TFICF) และประยุกต์ใช้ Zip's Law ในการวิเคราะห์สัมพันธ์ระหว่างค่าน้ำหนักของเทอมกับลำดับของค่าน้ำหนักนั้น เพื่อกำหนดกลุ่มตัวแทนให้กับหัวข้อข่าว นอกจากนั้นยังได้ศึกษาถึงผลกระทบต่าง ๆ ที่มีผลต่อประสิทธิภาพของการกำหนดหัวข้อข่าว ได้แก่ จำนวนของเอกสารที่ใช้ในฝึกสอน จำนวนของเทอมที่ใช้เป็นตัวแทนของเอกสาร และค่า threshold ที่เหมาะสมที่ใช้กำหนดในการกำหนดจำนวนเทอม

Abstract

This research compares performance of several term weighting methods on a topic identification task using web news data. Those methods are term Frequency Inverse Document Frequency (TFIDF) and three methods: Chi-square, Information Gain and Term Frequency Inverse Document Frequency (TFICF). Besides, we combine Zipf's Law for analyzing the relationship between term weighting and its rank. We also observe the impacts of the size of the training corpus, the size of the terms that represent the topic, and the appropriate threshold value.

สารบัญ

สารบัญตาราง (ต่อ)

สารบัญรูป

สารบัญรูป (ต่อ)