

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการการวิจัยแห่งชาติ



249865

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 เนติโสภาคกุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินรายได้ประจำปีงบประมาณ 2554

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## กิตติกรรมประกาศ

งานวิจัยเรื่องการกำหนดหัวข้อข่าว โดยใช้ค่า TFIDF และ Zipf's Law ในการเรียนรู้แบบมี การสอนนี้ ได้รับทุนสนับสนุนการวิจัยจากเงินรายได้คณะเทคโนโลยีสารสนเทศ สถาบัน เทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ประจำปี พ.ศ. 2554 ผู้วิจัยจึงขอกราบขอบพระคุณ เป็นอย่างสูงมา ณ ที่นี้

พรฤดี เนติโสภาคกุล

## 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 เนติโสภาคย์ 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

249865

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 ในการวิเคราะห์สัมพันธระหว่างค่าน้ำหนักของเทอมกับลำดับของค่าน้ำหนักนั้น เพื่อกำหนดกลุ่มตัวแทนให้กับหัวข้อข่าว นอกจากนั้นยังได้ศึกษาถึงผลกระทบต่าง ๆ ที่มีผลต่อประสิทธิภาพของการกำหนดหัวข้อข่าว ได้แก่ จำนวนของเอกสารที่ใช้ฝึกสอน จำนวนของเทอมที่ใช้เป็นตัวแทนของเอกสาร และค่า threshold ที่เหมาะสมที่ใช้กำหนดในการกำหนดจำนวนเทอม

## **Abstract**

**249865**

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

	หน้า
กิตติกรรมประกาศ	I
บทคัดย่อ	II
สารบัญ	IV
สารบัญตาราง	IV
สารบัญรูป	VII
บทที่ 1 บทนำ	1
1.1 ปัญหาและความเป็นมา	1
1.2 วัตถุประสงค์ในการวิจัย	2
1.3 ขอบเขตการศึกษา	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
บทที่ 2 แนวคิดและเทคนิคที่เกี่ยวข้อง	3
2.1 การจัดกลุ่มเอกสาร	3
2.1.1 วิธีการจัดกลุ่มเอกสาร	4
2.2 การวัดประสิทธิภาพ	15
2.3 Zipf's Law	16
2.4 งานวิจัยที่เกี่ยวข้อง	18
บทที่ 3 กระบวนการในการกำหนดหัวข้อข่าวให้กับเอกสาร	26
3.1 กระบวนการจัดกลุ่มเอกสาร	27
3.1.1 ส่วนการเรียนรู้	27
3.1.2 ส่วนการทดสอบการจัดกลุ่ม	35
บทที่ 4 การทดลองและผลการทดลอง	37
4.1 ข้อมูลที่ใช้ในการทดลอง	37
4.2 การออกแบบการทดลองและผลการทดลอง	46
4.3 อธิบายผลการทดลอง	59
บทที่ 5 สรุปการวิจัย	64
5.1 สรุปงานวิจัย	64
บรรณานุกรม	66

## สารบัญตาราง (ต่อ)

ตารางที่	หน้า
4.10	52
แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 3,000 เอกสาร	
4.11	54
ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร	
4.12	54
แสดงค่าความเที่ยงตรงและค่าความระลึกลักษณะตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร	
4.13	55
แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 4,200 เอกสาร	
4.14	56
ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร	
4.15	57
แสดงค่าความเที่ยงตรงและค่าความระลึกลักษณะตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร	
4.16	58
แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 6,000 เอกสาร	
4.17	62
แสดงค่าเฉลี่ยค่าเอฟแบ่งตามจำนวนคำที่เลือกในเอกสารแต่ละประเภท และจำนวนเอกสารตามชุดเอกสารทดสอบที่ 1, 2 และ 3	

# สารบัญรูป

รูปที่		หน้า
2.1	รูปแสดงขั้นตอนการจัดกลุ่มเอกสาร	3
2.2	ไฮเปอร์เพลนในการแบ่งข้อมูลสองกลุ่ม (Fletcher, 2009)	10
2.3	การจัดกลุ่มข้อมูลในลักษณะข้อมูลไม่เป็นเชิงเส้น (Fletcher, 2009)	12
2.4	แสดงข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นจากตัวอย่างที่กำหนด	12
2.5	แสดงข้อมูลใน feature space	13
2.6	แสดงข้อมูลที่ทำหน้าที่เป็นซัพพอร์ตเวกเตอร์	13
2.7	แสดงไฮเปอร์เพลนที่แยกระหว่างข้อมูลสองกลุ่ม	15
2.8	กราฟแสดงความสัมพันธ์ระหว่างความถี่และลำดับ ของเอกสาร Alice, Tale และ Bible (Konchady, 2006)	17
2.9	กราฟแสดงความสัมพันธ์ระหว่างความถี่และลำดับที่ใช้กฎ Zipf และ Mandelbrot (Konchady, 2006)	18
2.10	แสดงกลุ่มคำศัพท์ที่ปรากฏในแต่ละประเภทเอกสาร (Daniel et. al, 2009)	19
3.1	กระบวนการในการจัดกลุ่มเอกสารในงานวิจัยนี้	26
3.2	ขั้นตอนในส่วนของการเรียนรู้	27
3.3	ขั้นตอนในส่วนของการจัดกลุ่ม	36
4.1	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวธุรกิจ	38
4.2	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวบันเทิง	39
4.3	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวสุขภาพ	39
4.4	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวการเมือง	40
4.5	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวกีฬา	41
4.6	การกระจายของค่า 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทพยากรณ์อากาศ	41
4.7	แสดงจำนวนคำที่ไม่ซ้ำในแต่ละประเภทเอกสาร	44
4.8	แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสาร	45
4.9	แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารแยกตามประเภทเอกสาร	45
4.10	แสดงค่าเอฟตามค่า Threshold และกลุ่มเอกสารชุดทดสอบ	48
4.11	แสดงค่าเอฟตามกลุ่มเอกสารชุดทดสอบ และวิธีการคำนวณค่าน้ำหนัก	50

## สารบัญญรูป (ต่อ)

รูปที่		หน้า
4.12	แสดงค่าเอฟแยกตามจำนวนคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 1	53
4.13	แสดงค่าเอฟแยกตามจำนวนคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 2	55
4.14	แสดงค่าเอฟแยกตามจำนวนคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 3	58
4.15	แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนเอกสารที่ใช้ในการทดสอบ (a) ไม่มีการเลือกคุณลักษณะ (b) เลือกคุณลักษณะ	60
4.16	แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนคุณลักษณะ ในแต่ละชุดเอกสารทดสอบ	61