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)

บทที่ 1
บทนำ

1.1 ปัญหาและความเป็นมา

ด้วยจำนวนเอกสารที่มีปริมาณเพิ่มมากขึ้นอย่างรวดเร็ว ทำให้งานด้านการวิเคราะห์ความหมายเอกสารจึงมีความยากเพิ่มมากขึ้นตามไปด้วย โดยเอกสารแต่ละประเภทก็มีเนื้อหารายละเอียดที่แตกต่างกันไป เช่น เอกสารการประเภทพยากรณ์อากาศ สิ่งที่ต้องการวิเคราะห์คือ เกิดสภาพอากาศอะไร ที่ไหน เมื่อไหร่ มีผลกระทบต่อใครบ้างและทำให้เกิดผลอะไร และเอกสารประเภทกีฬา สิ่งที่ต้องการวิเคราะห์คือ ใครแข่งกับใคร ผลการแข่งขันเป็นอย่างไร รายละเอียดของการแข่งขันเป็นอย่างไร และมีเหตุการณ์อะไรเกิดขึ้นบ้างในระหว่างการแข่งขัน การกำหนดหัวข้อให้กับเอกสารจึงเป็นงานที่สำคัญในการวิเคราะห์ความหมายเอกสาร โดยเมื่อมีการระบุหัวข้อเอกสารแล้วก็จะทำให้ทราบว่าข้อมูลที่ต้องการวิเคราะห์คืออะไร ทำให้สามารถวิเคราะห์ความหมายได้สอดคล้องกับเนื้อความในเอกสาร

การกำหนดหัวข้อเอกสารนั้นงานวิจัยโดยส่วนใหญ่แล้วจะใช้แนวทางการวิเคราะห์โดยใช้วิธีทางสถิติ และการเรียนรู้จากกลุ่มเอกสารฝึกสอน โดยวิเคราะห์จากเทอมที่ปรากฏในเอกสาร ซึ่งในการวิเคราะห์ความสำคัญของเทอมนั้นจะมีการกำหนดค่าน้ำหนักให้กับแต่ละเทอม โดยเอกสารที่ใช้ในการฝึกสอนจะคัดคำที่เป็นคำฟุ่มเฟือยออกก่อนแล้วจึงทำการหาค่าน้ำหนักของคำที่เหลือที่เรียกว่า content words จากนั้นจึงกำหนดเทอมที่สำคัญที่สามารถใช้เป็นตัวแทนของหัวข้อเอกสารเพื่อระบุหัวข้อให้กับเอกสารใหม่ โดยเรียกเทอมเหล่านั้นว่า topic keywords เช่น เอกสารประเภทข่าวกีฬามักประกอบด้วยกลุ่มคำสำคัญ เช่น tournament, champion, won, defeat, final, player, match, coach, team และเอกสารประเภทสุขภาพมักประกอบด้วยคำสำคัญ เช่น health, disease, medical, patients, medicine, blood, drug ซึ่งเทอมที่มีความน้ำหนักมากก็จะสามารถเป็นตัวแทนที่ดีของหัวข้อนั้น โดยค่าน้ำหนักที่ใช้ในการวิเคราะห์ความสำคัญของเทอมนั้น มีด้วยกันหลายวิธี เช่น Term Frequency Inverse Document Frequency (TFIDF) (Salto and McGill, 1983), Information Gain (Wang et al., 2007, Li et al, 2009), Chi-Square (Caropreso et al., 2001, Li et al, 2009), และ Term Frequency Inverse Class Frequency (TFICF) (Kim et al., 2005)

ดังนั้นในงานวิจัยนี้จึงได้นำเสนอการออกแบบและทำการทดลองเพื่อเปรียบเทียบเทคนิคต่าง ๆ ที่ใช้ในการกำหนดหัวข้อข่าวให้กับเอกสารดังที่กล่าวข้างต้น โดยการประยุกต์ใช้การวิเคราะห์ตาม Zipf's Law () ในการกำหนดกลุ่มตัวแทนของหัวข้อ ซึ่งผลการทดลองนี้สามารถกำหนดกลุ่มของคำสำคัญแต่ละหัวข้อเพื่อใช้ในการระบุหัวข้อข่าวให้กับเอกสารใหม่ และใช้เทคนิคการจัดกลุ่มเอกสารซัพพอร์ทเวกเตอร์แมชชีนในการวัดประสิทธิภาพกลุ่มของเทอมที่สำคัญ นอกจากนี้การกำหนดกลุ่มคำสำคัญของหัวข้อนี้สามารถนำไปเป็นขั้นตอนหนึ่งในการทำงานทางด้านการสกัดความรู้ (Information Extraction) การแบ่งกลุ่มเอกสาร

(Document Clustering) การสรุปเอกสาร (Document Summarization) การวิเคราะห์เนื้อความเอกสาร (Document Content Analysis)

1.2 วัตถุประสงค์ในการวิจัย

เพื่อออกแบบ ทดลอง และเปรียบเทียบเทคนิควิธีที่เหมาะสมในการกำหนดหัวข้อข่าวให้กับเอกสาร ซึ่งเทคนิคที่ใช้ในการเปรียบเทียบได้แก่ TFIDF, TFICF, Chi-Square, และ Information Gain

1.3 ขอบเขตการศึกษา

เพื่อเปรียบเทียบประสิทธิภาพเทคนิคที่ใช้ในการกำหนดหัวข้อให้กับเอกสาร โดยกลุ่มเอกสารที่ใช้ในการทดลองประกอบด้วยเอกสารประเภท กีฬา สภาพอากาศ ธุรกิจ การเมือง สุขภาพ และบันเทิง โดยคำนึงถึงปัจจัยต่าง ๆ ที่มีผลกระทบต่อการกำหนดกลุ่มคำสำคัญที่ใช้เป็นตัวแทนของหัวข้อ ซึ่งได้แก่ จำนวนเอกสาร เทอมในกลุ่มคำสำคัญ และค่า Threshold ที่กำหนดค่าความถี่ต่ำสุดของคำสำคัญ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

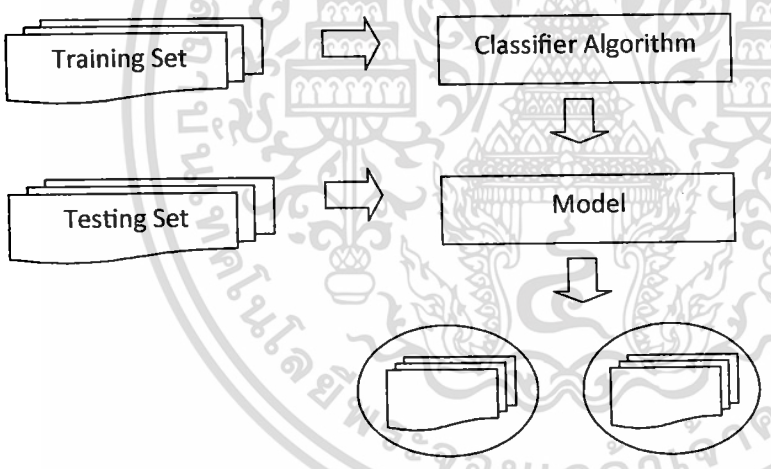
1. ผลสรุปจากการทดลองทำให้ได้เทคนิคที่เหมาะสมในการกำหนดกลุ่มคำที่สำคัญที่ใช้ในการกำหนดหัวข้อข่าว
2. สามารถนำใช้วิธีการกำหนดคำสำคัญนี้ไปใช้ในงานทางด้านการแบ่งกลุ่มเอกสาร การจัดกลุ่มเอกสาร การสกัดข้อมูล การสรุป และการวิเคราะห์ความหมายเอกสารได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้นำเสนอทฤษฎีพื้นฐานต่าง ๆ เกี่ยวกับการจัดกลุ่มเอกสาร ได้แก่ การกำหนดคุณลักษณะของเอกสารเพื่อใช้ในการจัดกลุ่มโดยการพิจารณาจากค่าน้ำหนัก อัลกอริทึมที่ใช้ในการจัดกลุ่มเอกสาร การวัดประสิทธิภาพ และงานวิจัยที่เกี่ยวข้องในการจัดกลุ่มเอกสาร

2.1 การจัดกลุ่มเอกสาร (Document Classification)

การจัดกลุ่มเอกสาร เป็นการจัดเอกสารกลุ่มเอกสารตามลักษณะของเนื้อหาในเอกสาร ซึ่งเอกสารใดมีเนื้อหาที่ใกล้เคียงกันก็จะถูกจัดให้อยู่กลุ่มเดียวกัน ในการจัดกลุ่มเอกสารนั้นเป็นการฝึกสอนโดยใช้ชุดเอกสารตัวอย่างที่เรียกว่า Training Set สำหรับสร้างโมเดลและทดสอบโมเดลนั้นจะใช้เอกสารทดสอบที่เรียกว่า Testing Set ซึ่งเป็นคนละชุดกับเอกสารตัวอย่าง รูปที่ 2.1 แสดงตัวอย่างขั้นตอนการจัดกลุ่มเอกสาร



รูปที่ 2.1 รูปแสดงขั้นตอนการจัดกลุ่มเอกสาร

จากรูปที่ 2.1 Training Set คือเอกสารที่ใช้สำหรับฝึกสอนและเรียนรู้ว่าลักษณะเอกสารที่อยู่ในกลุ่มเดียวกันนั้นควรมีคุณลักษณะอย่างไรบ้าง โดยมีการกำหนดกลุ่มเอกสารให้กับเอกสารทดสอบเหล่านั้นก่อนและฝึกสอนโดยใช้อัลกอริทึมต่าง ๆ เช่น Support Vector Machine, Naïve Bayes เป็นต้น เมื่อเอกสารกลุ่มตัวอย่างถูกทำการฝึกสอนจากข้อมูลทดสอบด้วยอลักอริทึมการจัดกลุ่มข้อมูลใด ๆ แล้วจะทำการสร้างโมเดลเพื่อใช้สำหรับจัดกลุ่มเอกสาร ในการทดสอบโมเดลนั้นจะใช้เอกสารกลุ่มใหม่ที่เรียกว่า Testing Set เพื่อแบ่งกลุ่มเอกสารตามที่ได้เรียนรู้มากจากเอกสารชุดฝึกสอน

2.1.1 วิธีการจัดกลุ่มเอกสาร

เอกสารที่ใช้ในการเรียนรู้จะมีรูปแบบเป็นเอกสารที่ไม่มีโครงสร้าง (Unstructured document) อาจจะเป็นเว็บเอกสาร หรือ เอกสาร Word หรือ เอกสาร PDF ซึ่งเมื่อทำการจัดกลุ่มเอกสารโดยจัดกลุ่มตามเนื้อหาในเอกสาร จะต้องเลือกเอาเฉพาะข้อความที่ปรากฏในเอกสารเท่านั้น แล้วทำการจัดกลุ่มเอกสารเหล่านั้นด้วยมือสำหรับใช้เป็นเอกสารเรียนรู้ จากนั้นเลือกคุณลักษณะที่สำคัญที่จะใช้เป็นตัวระบุในการจัดกลุ่มเอกสาร แล้วทำการฝึกสอนด้วยเทคนิคการเรียนรู้ต่าง ๆ เพื่อให้ได้โมเดลและทดสอบโดเมลนั้นด้วยเอกสารชุดทดสอบ ขั้นตอนในการจัดกลุ่มเอกสารประกอบด้วย

การตัดคำที่ใช้บ่อย (Stop word removal)

เป็นการนำคำไม่มีนัยสำคัญออก หรือคำที่น่าจะไม่มีความสำคัญต่อเอกสาร โดยที่คำเปล่านั้นไม่สามารถกำหนดเป็นตัวแทนของกลุ่มเอกสารได้ และมักจะปรากฏในทุกๆ เอกสาร เช่น คำบุพบท (in, on, at, under, ...) คำสรรพนาม (he, she, they, ...) คำระบุนาม (the, a, ...) คำสันธาน (and, or, but, ...) เป็นต้น ซึ่งมีผลทำให้จำนวนคำในเอกสารลดลง โดยทั่วไปแล้ว ในขั้นตอนนี้จะทำให้จำนวนของคำในเอสการลดลงถึง 40-50% จากคำทั้งหมด (Salton, 1983) นอกจากนั้นคำที่ถูกตัดออกไปนั้นก็มักจะเป็นคำที่มีความถี่ของการปรากฏในเอกสารมากและพบได้ในเอกสารทุกประเภท

การแปลงคำให้กลับไปอยู่ในรูปเดิมหรือ Base form (Stemming)

เนื่องจากคำศัพท์แต่ละคำนั้นสามารถแปลงให้อยู่ได้หลายรูปแบบ เช่น คำกริยาในรูปอดีต และปัจจุบัน คำนาม เป็นต้น ในขั้นตอนนี้จะเป็นการแปลงคำศัพท์เหล่านั้นให้อยู่ในรูปรากศัพท์หรือ Base form เช่นคำศัพท์ prevent, prevents, preventing และ prevention จะมีรากศัพท์ที่เหมือนกันคือ prevent ผลจากการทำเช่นนี้ทำให้จำนวนคำในเอกสารลดลง และมีผลต่อการนับการปรากฏของคำในเอกสารด้วย โดยทั่วไปจะใช้อัลกอริทึมที่ได้รับความนิยมคือ Porter Stemming แต่อย่างไรก็ตามในการแปลงคำศัพท์ให้อยู่ในรูปแบบของ Base form นั้นบางคำศัพท์เมื่อแปลงแล้วจะไม่สามารถแปลความหมายได้จากดิกชันารี

การแทนเอกสารด้วย Vector space Model

อัลกอริทึมในการจัดกลุ่มเอกสารนั้นแต่ละเอกสารจะถูกแทนในรูปแบบของเวกเตอร์โดยแต่ละแอทริบิวต์ของเวกเตอร์แทนด้วยค่าน้ำหนักของเทอม ซึ่งค่าน้ำหนักนั้นจะใช้ความถี่ของการปรากฏของคำในเอกสารที่สามารถคำนวณได้ด้วยวิธีการต่าง ๆ ได้แก่ TFIDF, IG, CHI และค่าน้ำหนักนี้จะเป็นแนวทางในการจัดกลุ่มให้กับเอกสารโดยแยกออกจากกลุ่มเอกสารที่ไม่เกี่ยวข้อง ในการเลือกเทอมและการคำนวณหาค่าน้ำหนักเหล่านี้จะอธิบายในหัวข้อถัดไป โดยรูปแบบการแทนเอกสารแสดงได้ดังตัวอย่าง

	t1	t2	t3	t4	...	tn
D1	[x1	x2	x3	x4	...	xn]

อธิบายได้ว่าเอกสาร D1 ประกอบด้วยคำต่าง ๆ ได้แก่ t1,t2,...,tn โดยที่ n คือจำนวนคำที่ไม่ซ้ำที่ปรากฏในเอกสาร D1 และ x คือค่าน้ำหนักของแต่ละคำ

ตารางที่ 2.1 แสดงตัวอย่างกลุ่มของเอกสารจำนวน 4 เอกสาร (D1, D2, D3, D4)

เอกสาร	ข้อความ
D1	Human machine interface for computer applications
D2	A survey of user opinion of computer system response time
D3	The EPS user interface management system
D4	Systems and human system engineering testing of EPS

เอกสารจากตารางที่ 2.1 เป็นเอกสารที่ต้องการแทนในรูปแบบของเวกเตอร์ โดยแต่ละเอกสารจะผ่านกระบวนการตำคำที่ไม่สำคัญออก (Stop word Removal) จากนั้นหาความถี่ของคำที่ไม่ซ้ำกันทั้งหมดในกล่มุเอกสาร และสร้างเวกเตอร์ของเอกสาร แล้วรวมเวกเตอร์ทั้งหมดให้อยู่ในรูปแบบของเมตริกซ์ โดยแถวของเมตริกซ์คือเอกสารทั้งหมด และคอลัมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในกลุ่มเอกสาร ซึ่งเมตริกซ์นี้จะใช้เป็นเอกสารนำเข้าให้กับขั้นตอนของการจัดกลุ่มเอกสาร ตัวอย่างเมตริกซ์แสดงดังตาราง 2.2

ตารางที่ 2.2 แสดงเมตริซ์ของเอกสารตัวอย่างจากตารางที่ 2.1 ด้วยค่าความถี่ของแต่ละคำที่ปรากฏในเอกสาร

	Human	Machine	Interface	Computer	applications	survey	user	system	...
D1	1	1	1	1	1	0	0	0	
D2	0	0	0	1	0	1	1	1	
D3	0	0	1	0	0	0	0	1	
D4	1	0	0	0	0	0	0	1	

จากตารางที่ 2.2 ค่าน้ำหนักเท่ากับ 0 หมายความว่าคำนั้นไม่ปรากฏในเอกสารเลย และพบว่ายิ่งถ้ามีเอกสารเป็นจำนวนมากก็จะยิ่งมีจำนวนคำที่ไม่ซ้ำที่กำหนดเป็นคอลัมน์มากขึ้นด้วย แม้ว่าจะผ่านกระบวนการตัดคำที่ไม่สำคัญและแปลงคำให้อยู่ในรูปรากศัพท์แล้วก็ตาม เมื่อข้อมูลเหล่านั้นถูกแทนด้วยเมตริกซ์ จะมีผลทำให้เมตริกซ์นั้นมีขนาดใหญ่มาก ดังนั้นจึงควรลดขนาดของเมตริกซ์โดยการเลือกเฉพาะคำที่สำคัญที่สามารถเป็นตัวแทนของแต่ละกลุ่มเอกสารได้

การเลือกคุณลักษณะ (Feature Selection)

ในขั้นตอนนี้จะเป็นการลดขนาดหรือจำนวนของคุณลักษณะให้มีจำนวนลดลง โดยเลือกคุณลักษณะที่มีประสิทธิภาพเพื่อใช้ในการจัดกลุ่มเอกสาร โดยทั่วไปงานวิจัยทางด้านการจัดกลุ่มเอกสารนี้จะใช้คำเป็นคุณลักษณะในการใช้แยกหรือระบุกลุ่มของเอกสาร และใช้ค่าน้ำหนักของคำที่ปรากฏในเอกสารเป็นค่าของคุณลักษณะ

การเลือกคุณลักษณะหรือลดจำนวนคุณลักษณะนี้จะใช้ค่าน้ำหนักที่กำหนดด้วย TFIDF, TFICF, Information gain (IG) และ Chi-square (CHI) อธิบายรายละเอียดดังต่อไปนี้

TFIDF (Term frequency inverse document frequency) เป็นการกำหนดค่าน้ำหนักเพื่อระบุว่าคำนั้นมีความสำคัญต่อกลุ่มเอกสารของเอกสารอย่างไร ถ้าคำนั้นมีความสำคัญมากค่าน้ำหนักก็จะมากตามไปด้วย แสดงดังสูตรที่ (2.1)

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times log \frac{N}{N(t_i)} \tag{2.1}$$

โดยที่ t คือ เทอมหรือคำ

d คือ เอกสาร

tf(t,d) คือ จำนวนครั้งที่ปรากฏหรือความถี่ของเทอม t ในเอกสาร d

N คือ จำนวนเอกสารทั้งหมด

N(t) คือ จำนวนเอกสารที่ปรากฏเทอม t

จากสูตรที่ (2.1) ของแต่ละคำนั้น คำที่ปรากฏในเอกสารหลายๆ เอกสารนั้นจะมีค่าน้ำหนักน้อยกว่าคำที่ปรากฏในบางเอกสาร โดยค่า TFIDF มีค่าเท่ากับ 0 ก็ต่อเมื่อเทอมนั้นปรากฏในทุกเอกสาร นั้นหมายความว่าคำที่ปรากฏบ่อย ๆ ในเอกสารหนึ่งแต่ปรากฏในเอกสารเป็นจำนวนน้อย จะเป็นตัวแทนที่ดีของกลุ่มเอกสารนั้น ๆ

TFICF (Term frequency inverse class frequency) เป็นการคำนวณค่าน้ำหนักของคำโดยคำนึงถึงว่าคำนั้นเกี่ยวข้องกับกลุ่มเอกสารประเภทใดและไม่เกี่ยวข้องกับกลุ่มเอกสารประเภทใด เนื่องจากว่ามีหลาย ๆ คำที่สามารถปรากฏได้หลายกลุ่มเอกสาร ซึ่งคำนั้นอาจจะเป็นคำที่สำคัญสำหรับเอกสารประเภทหนึ่งแต่ไม่ใช่คำที่สำคัญในเอกสารอีกประเภทหนึ่ง แสดงดังสูตร (2.2)

$$ticf(t_i, c_j) = tf(t_i, c_j) \times icf(t) \tag{2.2}$$

$$tf(t_i, c_j) = \frac{\sum_{k=1}^{\#docs_j} freq(t_i, doc_{jk})}{\sum_{k=1}^{\#docs_j} \#token(doc_{jk})} \tag{2.3}$$

$$icf(t) = \log \frac{|c|}{cf(t_i)} \quad (2.4)$$

โดยที่ t คือ เทอมหรือคำ

c คือ ประเภทของเอกสารหรือคลาสของกลุ่มของเอกสาร

doc คือ เอกสาร

freq(t,doc_{jk}) คือ ความถี่ของคำ t ที่ปรากฏในเอกสารที่ k และในประเภทเอกสาร j

#token(doc_{jk}) คือ จำนวนคำที่ปรากฏในเอกสารที่ k และในประเภทเอกสาร j

|C| คือ จำนวนประเภทของเอกสาร

cf(t) คือ จำนวนประเภทเอกสารที่ปรากฏคำ t

จากสูตร (2.2), (2.3) และ (2.4) ถ้าคำนั้นปรากฏในทุกประเภทของเอกสารแล้ว คำนั้นจะมีค่าน้ำหนักเท่ากับ 0 โดยที่ไม่คำนึงว่าคำนั้นจะปรากฏเป็นจำนวนเท่าไหร่ในแต่ละประเภทเอกสาร ความแตกต่างระหว่าง TFIDF และ TFICF คือ TFIDF จะคำนวณโดยไม่สนใจกลุ่มของเอกสาร สนใจเฉพาะคำในเอกสารทั้งหมด เพื่อพิจารณาว่ากลุ่มคำที่สำคัญของแต่ละประเภทเอกสาร ส่วน TFICF นั้นจะคำนึงถึงประเภทของเอกสารด้วย โดยพิจารณาว่าคำใดไปปรากฏในประเภทเอกสารอื่นบ้าง

ไคสแควร์ (Chi-Square หรือ CHI) เป็นการทดสอบทางสถิติเพื่อเปรียบเทียบความสัมพันธ์ระหว่างตัวแปร โดยเปรียบเทียบข้อมูลที่มีอยู่ในรูปแบบของความถี่ที่สามารถจำแนกออกเป็นประเภทหรือหมวดหมู่ได้ โดยในงานการจัดกลุ่มเอกสาร ไคสแควร์ถูกนำมาใช้เพื่อคำนวณหาค่าน้ำหนักของคำในกลุ่มเอกสารซึ่งเป็นการเปรียบเทียบความสัมพันธ์ระหว่างคำกับประเภทหรือกลุ่มของเอกสาร โดยพิจารณาคำที่ไม่ปรากฏในเอกสารร่วมกับคำที่ปรากฏในเอกสารนั้นด้วย แสดงดังสูตร (2.5)

$$\chi^2(t_k, c_i) = \frac{N[P(t_k, c_i)P(\overline{t_k}, \overline{c_i}) - P(t_k, \overline{c_i})P(\overline{t_k}, c_i)]^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})} \quad (2.5)$$

โดยที่ t_k คือ เทอม

c_j คือ หัวข้อหรือกลุ่มเอกสาร

$P(t_k, c_i)$ คือ ความน่าจะเป็นของเอกสารในหัวข้อ c_i เมื่อปรากฏเทอม t_k

$P(t_k, \overline{c_i})$ คือ ความน่าจะเป็นของเอกสารที่ไม่อยู่ในหัวข้อ c_i เมื่อปรากฏเทอม t_k

$P(\overline{t_k}, c_i)$ คือ ความน่าจะเป็นของเอกสารในหัวข้อ c_i เมื่อไม่ปรากฏเทอม t_k

$P(\overline{t_k}, \overline{c_i})$ คือ ความน่าจะเป็นของเอกสารที่ไม่อยู่ในหัวข้อ c_i เมื่อไม่ปรากฏเทอม t_k

หรืออาจเขียนได้ดังนี้

	c	$\bar{c}$
t	A	B
$\bar{t}$	C	D

สามารถเขียนเป็นสูตรไคสแควร์ได้อย่างง่าย แสดงดังสูตรที่ (2.6)

$$\chi^2 = \frac{N(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \tag{2.6}$$

โดยที่ N คือ จำนวนเอกสารทั้งหมด

จากสูตรที่ (2.5) และ (2.6) การคำนวณค่าน้ำหนักด้วยไคสแควร์นั้นได้พิจารณาคำที่เกี่ยวข้องและไม่เกี่ยวข้องกับเอกสารในแต่ละประเภท ในขณะที่ TFIDF ไม่ได้พิจารณาคำแยกตามประเภทเอกสาร

Information Gain (IG) เป็นวิธีในการคำนวณค่าน้ำหนักโดยใช้ในการทำนายประเภทของเอกสาร โดยดูจากการปรากฏและไม่ปรากฏของคำในเอกสารแต่ละประเภท แสดงดังสูตร (2.7)

$$IG(t_k, c_i) = P(t_k, c_i)\log\frac{P(t_k, c_i)}{P(t_k)P(c_i)} + P(\bar{t_k}, c_i)\log\frac{P(\bar{t_k}, c_i)}{P(\bar{t_k})P(c_i)} \tag{2.7}$$

โดยที่ t_k คือ เทอม

c_j คือ หัวข้อหรือกลุ่มเอกสาร

$P(t_k, c_i)$ คือ ความน่าจะเป็นของเอกสารในหัวข้อ c_i เมื่อปรากฏเทอม t_k

$P(\bar{t_k}, c_i)$ คือ ความน่าจะเป็นของเอกสารในหัวข้อ c_i เมื่อไม่ปรากฏเทอม t_k

สามารถเขียนเป็นสูตรได้อย่างง่ายดังนี้

$$IG = -\frac{A+C}{N}log\frac{A+C}{N} + \frac{A}{N}log\left(\frac{A}{A+B}\right) + \frac{C}{N}log\left(\frac{C}{C+D}\right) \tag{2.8}$$

จากสูตรที่ (2.7) และ (2.8) พบว่าการคำนวณค่าน้ำหนักนี้จะพิจารณาเฉพาะเอกสารของแต่ละประเภทว่าในประเภทเอกสารนั้น ๆ มีคำใดที่เกี่ยวข้อง ซึ่งแตกต่างจากไคสแควร์ที่พิจารณาคำนั้นในประเภทเอกสารอื่นด้วย

จากสูตรของ TFICF, CHI และ IG ที่กล่าวข้างต้นนั้น จะเป็นการอ้างถึงเทอมหรือคำที่ปรากฏในแต่ละประเภทเอกสาร เพื่อที่จะกำหนดค่าของเทอมที่เป็นโกลบอลในการระบุค่าน้ำหนักของเทอมนั้นในแต่ละเวกเตอร์เราจะคำนวณได้จาก

1. รวมค่าน้ำหนักของแต่ละประเภทเอกสารเพื่อกำหนดเป็นค่าน้ำหนักของเทอมนั้น แสดงดังสูตรที่ 2.9

$$f_{sum}(t_k) = \sum_{i=1}^{|c|} f(t_k, c_i) \tag{2.9}$$

2. ค่าเฉลี่ยของเทอมในแต่ละประเภทเอกสาร โดยที่ $P(c_i)$ คือ ความน่าจะเป็นที่เทอมนั้นปรากฏในประเภทเอกสารใดๆ แสดงดังสูตรที่ 2.10

$$f_{avg}(t_k) = \sum_{i=1}^{|c|} P(c_i) f(t_k, c_i) \tag{2.10}$$

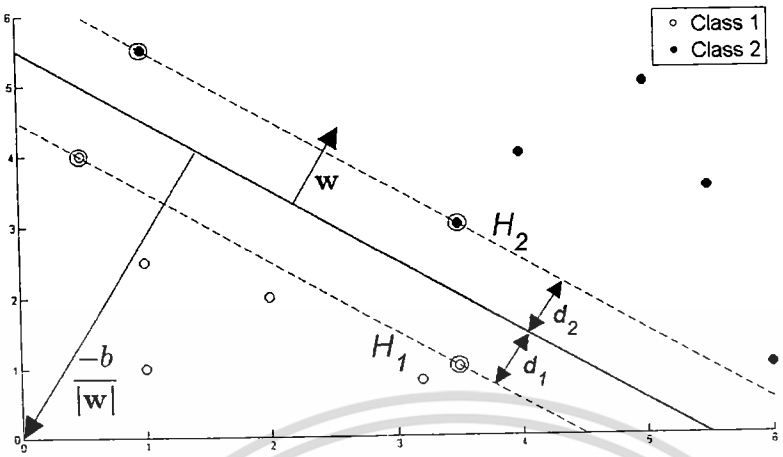
3. ค่าน้ำหนักที่มากที่สุดของเทอมนั้นในจำนวนประเภทเอกสาร จะถูกกำหนดให้เป็นค่าน้ำหนักของเทอมนั้น แสดงดังสูตรที่ 2.11

$$f_{max}(t_k) = \sum_{i=1}^{|c|} f(t_k, c_i) \tag{2.11}$$

อัลกอริทึมในการจัดกลุ่มเอกสาร (Classifier Algorithm)

อัลกอริทึมในการจัดกลุ่มเอกสาร แบ่งออกเป็น 2 ขั้นตอน ได้แก่ เรียนรู้เพื่อสร้างโมเดลสำหรับการจัดกลุ่มเอกสารและการจัดกลุ่มเอกสารตามโมเดลที่ได้เรียนรู้ โดยพิจารณาจากความคล้ายคลึงกันของเนื้อความในเอกสาร ซึ่งข้อมูลนำเข้าอัลกอริทึมการเรียนรู้นี้ คือคุณลักษณะของคำในรูปแบบของเวกเตอร์ที่คำที่ถูกกำหนดเป็นคุณลักษณะนั้นได้ผ่านการคัดเลือกเฉพาะคำสำคัญที่สามารถใช้แยกความแตกต่างระหว่างเอกสารได้และมีการให้ค่าน้ำหนักด้วยวิธีการต่าง ๆ

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) อัลกอริทึมนี้มีหลักการทำงาน คือ การสร้างไฮเปอร์เพลนที่เหมาะสมบนระนาบของกลุ่มตัวอย่างที่ใช้การเรียนรู้ เพื่อแบ่งแยกข้อมูลที่แตกต่างกัน กำหนดให้ระยะห่างระหว่างจุดข้อมูลที่อยู่ใกล้กับไฮเปอร์เพลนมากที่สุดทั้งสองด้าน คือ d1 และ d2 ในการสร้างไฮเปอร์เพลนที่เหมาะสมนั้นคือไฮเปอร์เพลนที่มีค่ามาร์จิน กว้างที่สุด โดยข้อมูลที่อยู่บนขอบของมาร์จิน เรียกว่า support vector และระยะมาร์จินเกิดจากระยะ d1+d2



รูปที่ 2.2 ไฮเปอร์เพลนในการแบ่งข้อมูลสองกลุ่ม (Fletcher, 2009)

จากรูปที่ 2.2 เป็นการแบ่งกลุ่มข้อมูลจำนวน 2 กลุ่ม (Class 1 และ Class 2) และข้อมูลที่ใช้ในการฝึกสอนถูกแสดงในรูปแบบดังสมการที่ (2.12)

$$\{x_i, y_i\} \; where \; i = 1 \ldots L, y_i \in \{-1,1\}, x \in \Re^D \tag{2.12}$$

โดยที่ x_i คือ อินพุตเวกเตอร์ของข้อมูล

y_i คือ กลุ่มหรือคลาสของข้อมูล ประกอบด้วย y=1 และ y=-1

อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนจะสร้างไฮเปอร์เพลนที่เหมาะสมบนระนาบของข้อมูล โดยสมการของไฮเปอร์เพลนแบบเชิงเส้น แสดงดังสมการที่ (2.13)

$$(\mathrm{w} \times \mathrm{x}) + \mathrm{b} = 0 \tag{2.13}$$

$\frac{\mathrm{b}}{\|w\|}$ เป็นระยะตั้งฉากจากไฮเปอร์เพลนถึงจุดออริจิน

โดยที่ w คือ เวกเตอร์ที่ตั้งฉากกับไฮเปอร์เพลน

b คือ ค่าคงที่ซึ่งกำหนดตำแหน่งของเวกเตอร์ที่สัมพันธ์กับตำแหน่งเดิมใน input space

ซัพพอร์ทเวกเตอร์ (Support vector) คือ ข้อมูลที่อยู่ใกล้เส้นไฮเปอร์เพลนที่แยกระหว่าง 2 กลุ่ม แล้วเลือกเวกเตอร์ที่อยู่ใกล้เส้นไฮเปอร์เพลนของทั้งสองกลุ่ม เพื่อหาระยะทางระหว่างเส้นขอบทั้งสองโดยเลือกระยะที่ห่างจากไฮเปอร์เพลนที่น้อยที่สุดเป็นตัวเลือกในการจัดกลุ่มเอกสาร

จากรูป ค่า w และ b ของกลุ่มตัวอย่างข้อมูลอธิบายโดย

$$x_i \times w + b \geq 1 \; for \; y_i = +1 \tag{2.14}$$

$$x_i \times w + b \leq 1 \; for \; y_i = -1 \tag{2.15}$$

จากสองสมการรวมกันได้เป็นสมการ

$$y_i(x_i \times w + b) - 1 \geq 0 \; \forall_i \tag{2.16}$$

พิจารณาจุดข้อมูลที่ใกล้เส้นไฮเปอร์เพลนแล้ว จะได้จุดที่อยู่ใกล้ไฮเปอร์เพลนมากที่สุดของทั้งสองกลุ่มคือ H1 และ H2 ซึ่งเป็นซัพพอร์ท ซึ่งสองจุดนี้อธิบายได้โดย

$$x_i \times w + b = +1 \; for \; H_1 \tag{2.17}$$

$$x_i \times w + b = -1 \; for \; H_2 \tag{2.18}$$

กำหนดให้ d1 เป็นระยะจากจุด H1 ถึงไฮเปอร์เพลน และ d2 เป็นระยะจากจุด H2 ถึงไฮเปอร์เพลนเช่นเดียวกัน ระยะ d1 ถึงไฮเปอร์เพลนมีค่าเท่ากับระยะ d2 ถึงไฮเปอร์เพลน (d1=d2) ซึ่งตามหลักการแล้วเราต้องการระยะมาร์จินที่มากที่สุด

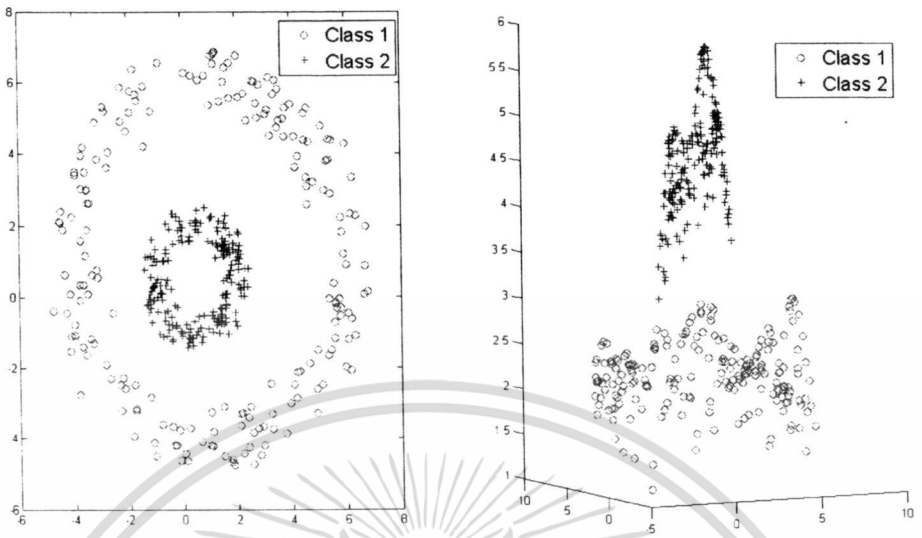
จากที่กล่าวมาข้างต้นเป็นการจัดกลุ่มข้อมูลด้วยไฮเปอร์เพลนในลักษณะเชิงเส้น เพื่อให้อัลกอริทึมนี้สามารถจัดกลุ่มข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นได้นั้น จึงสร้างเมตริกซ์ H จากการ dot product ของข้อมูลนำเข้า (แสดงในรูปแบบเวกเตอร์) ได้ดังสมการ

$$H_{ij} = y_i y_j k(x_i, x_j) = x_i, x_j = x_i^T x_j \tag{2.19}$$

โดยที่ k(xi,xj) เป็นฟังก์ชันที่เรียกว่าฟังก์ชันเคอร์เนล (Kernel function) หรือ Radial Basis Function (RBF) เซตของฟังก์ชันเคอร์เนลประกอบด้วย

$$k(x_i, x_j) = e^{-\left(\frac{\|x_i - x_j\|^2}{2\delta^2}\right)} \tag{2.20}$$

ตัวอย่างข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นซึ่งเป็นการจัดกลุ่มข้อมูลจำนวนสองกลุ่มแสดงดังรูปที่ 2.3



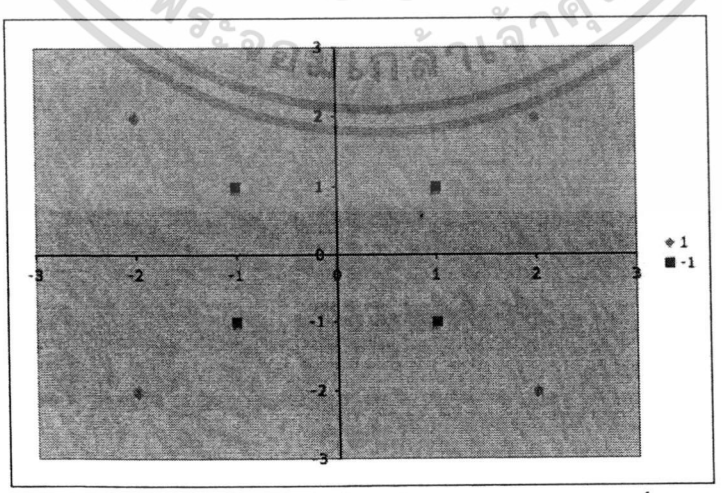
รูปที่ 2.3 การจัดกลุ่มข้อมูลในลักษณะข้อมูลไม่เป็นเชิงเส้น (Fletcher, 2009)

ตัวอย่างการจัดกลุ่มข้อมูลที่มีลักษณะเป็นแบบ Nonlinear จาก (Ventura, 2009) กำหนดให้ข้อมูล 2 กลุ่ม ประกอบด้วย

ข้อมูลกลุ่มที่ 1 $\{\binom{2}{2}, \binom{2}{-2}, \binom{-2}{-2}, \binom{-2}{2}\}$ มีค่าเป็นบวก (กลุ่ม +1)

ข้อมูลกลุ่มที่ 2 $\{\binom{1}{1}, \binom{1}{-1}, \binom{-1}{-1}, \binom{-1}{1}\}$ มีค่าเป็นลบ (กลุ่ม -1)

ข้อมูลที่สองกลุ่มนำมาวาดเป็นกราฟได้ดังรูปที่ 2.4 โดยข้อมูลกลุ่มที่ 1 แทนด้วยจุดสีฟ้า และข้อมูลกลุ่มที่ 2 แทนด้วยจุดสีแดง

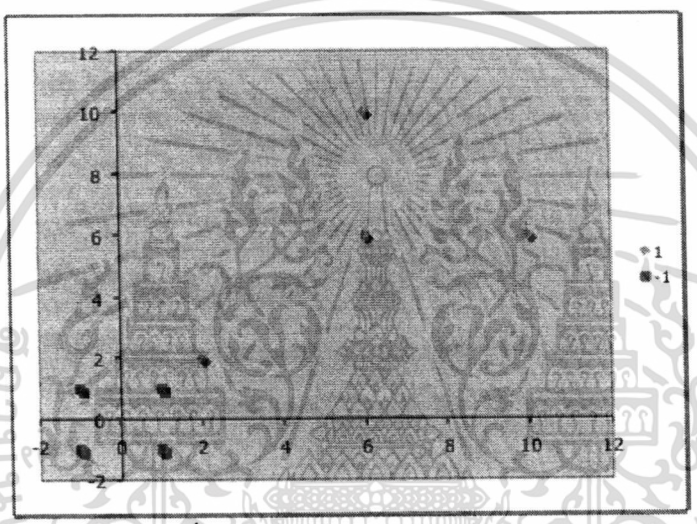


รูปที่ 2.4 แสดงข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นจากตัวอย่างที่กำหนด

จากนั้นทำการหาไฮเปอร์เพลนที่เหมาะสมให้กับข้อมูลทั้งสองกลุ่มนี้ เพื่อแยกข้อมูลทั้งสองกลุ่ม สำหรับข้อมูลในลักษณะที่ไม่เป็นเชิงเส้น นั้นจะต้องปรับข้อมูลจาก input space ให้อยู่ใน feature space ดังนี้

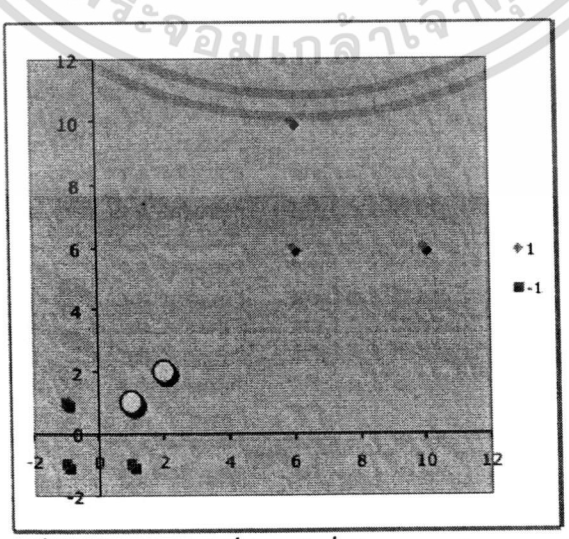
$$\emptyset\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & if\ \sqrt{x_1^2+x_2^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix}\ otherwise\end{cases}$$

จะได้ ดังรูปที่ 2.5 โดยข้อมูลในกลุ่มที่ 1 จะเปลี่ยนเป็น $\{\binom{2}{2}, \binom{6}{2}, \binom{6}{6}, \binom{2}{6}\}$



รูปที่ 2.5 แสดงข้อมูลใน feature space

ดังนั้น จะได้ซัพพอร์ทเวกเตอร์ดังรูปที่ 2.6 โดยที่เวกเตอร์ทั้งสองเป็นเวกเตอร์คนละกลุ่มโดยซัพพอร์ทเวกเตอร์ในกลุ่มที่ 1 คือ $\binom{2}{2}$ และซัพพอร์ทเวกเตอร์ของกลุ่มที่ 2 คือ $\binom{1}{1}$



รูปที่ 2.6 แสดงข้อมูลที่ทำหน้าที่เป็นซัพพอร์ทเวกเตอร์

เมื่อกำหนดซัพพอร์ทเวกเตอร์ของทั้งสองกลุ่มแล้ว จากนั้นคำนวณหาไฮเปอร์เพลนระหว่างซัพพอร์ทเวกเตอร์นี้

$$\alpha_1 \emptyset_1(s_1) \times \emptyset_1(s_1) + \alpha_2 \emptyset_2(s_2) \times \emptyset_2(s_2) = -1$$

$$\alpha_1 \emptyset_1(s_1) \times \emptyset_1(s_2) + \alpha_2 \emptyset_2(s_2) \times \emptyset_2(s_2) = 1$$

สามารถลดรูปได้เป็น

$$\alpha_1 \tilde{s_1} \times \tilde{s_1} + \alpha_2 \tilde{s_2} \times \tilde{s_1} = -1$$

$$\alpha_1 \tilde{s_1} \times \tilde{s_2} + \alpha_2 \tilde{s_2} \times \tilde{s_2} = 1$$

กำหนดให้ $\{s_1 = \binom{1}{1}, s_2 = \binom{2}{2}\}$ และ ใช้เวกเตอร์ 1 เป็น bias input คำนวณกับซัพพอร์ทเวกเตอร์ ทำการ dot product ซัพพอร์ทเวกเตอร์จะได้

$$3\alpha_1 + 5\alpha_2 = -1$$

$$5\alpha_1 + 9\alpha_2 = -1$$

จากสมการจะได้ α_1=-7 และ α_2 = 4 จึงจะทำให้สมการเป็นจริง คำนวณหาไฮเปอร์เพลนโดย

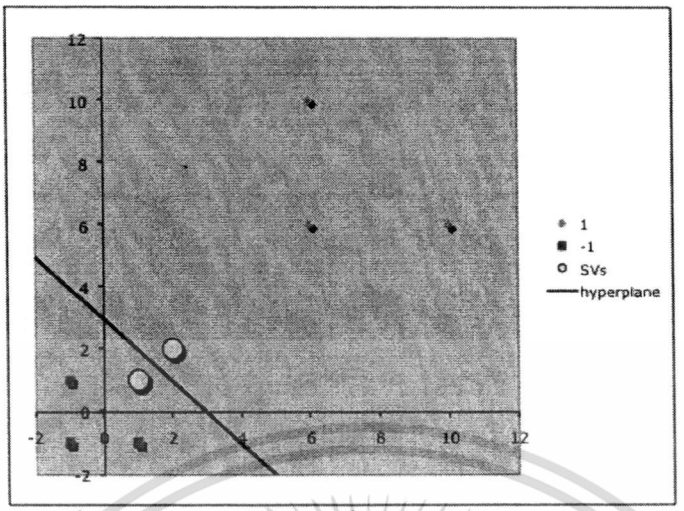
$$\tilde{w} = \sum_i \alpha_i \tilde{s}_i$$

จะได้

$$\tilde{w} = -7\begin{pmatrix}1\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix}$$

$$\tilde{w} = \begin{pmatrix}1\\1\\-3\end{pmatrix}$$

จากสมการ y=wx+b จะได้ $w = \binom{1}{1}$ และ b =-3 แสดงดังรูปที่ 2.7



รูปที่ 2.7 แสดงไฮเปอร์เพลนที่แยกระหว่างข้อมูลสองกลุ่ม

ถ้าต้องการจัดกลุ่มให้กับข้อมูลใหม่คือ x = (4, 5) ว่าควรอยู่ในกลุ่มใด สามารถคำนวณได้จาก

$$f\begin{pmatrix}4\\5\end{pmatrix} = \delta\left(-7\emptyset_1\begin{pmatrix}1\\1\end{pmatrix} \times \emptyset_1\begin{pmatrix}4\\5\end{pmatrix} + 4\emptyset_1\begin{pmatrix}2\\2\end{pmatrix} \times \emptyset_1\begin{pmatrix}4\\5\end{pmatrix}\right)$$

$$f\begin{pmatrix}4\\5\end{pmatrix} = \delta\left(-7\emptyset_1\begin{pmatrix}1\\1\\1\end{pmatrix} \times \begin{pmatrix}0\\1\\1\end{pmatrix} + 4\begin{pmatrix}2\\2\\1\end{pmatrix} \times \begin{pmatrix}0\\1\\1\end{pmatrix}\right)$$

$$f\begin{pmatrix}4\\5\end{pmatrix} = \delta(-2)$$

ดังนั้น ข้อมูล x = (4,5) จัดอยู่ในกลุ่มที่ 2ซึ่งมีค่าเป็นลบ

2.2 การวัดประสิทธิภาพ

การวัดประสิทธิภาพของการจัดกลุ่มเอกสารสามารถคำนวณค่าประสิทธิภาพด้วย ค่าความเที่ยงตรง (Precision), ค่าความระลึก (Recall) และค่าเอฟ (F-measure)

Category		Expert Judgment	
		True	**False**
Classifier Judgment	**True**	TP	FP
	False	FN	TN

Precision เป็นการวัดความสามารถของระบบในการจัดกลุ่มเอกสารที่ถูกต้องจากจำนวนเอกสารทั้งหมดที่ระบบทำการจัดกลุ่ม แสดงดังสมการที่ 2.21

$$\text{ค่าความเที่ยงตรง} = \frac{TP}{TP+FP} \tag{2.21}$$

Recall เป็นการวัดความสามารถของระบบในการจัดกลุ่มเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด แสดงดังสมการที่ 2.22

$$\text{ค่าความระลึก} = \frac{TP}{\text{TP+FN}} \tag{2.22}$$

F-measure เป็นการวัดค่าความสัมพันธ์ระหว่างค่าความแม่นยำและค่าความระลึก แสดงได้ดังสมการที่ 2.23

$$\text{F} - \text{measure} = \frac{2\times\text{ค่าความเที่ยงตรง}\times\text{ค่าความระลึก}}{\text{ค่าความเที่ยงตรง}+\text{ค่าความระลึก}} \tag{2.23}$$

2.3 Zipf's Law

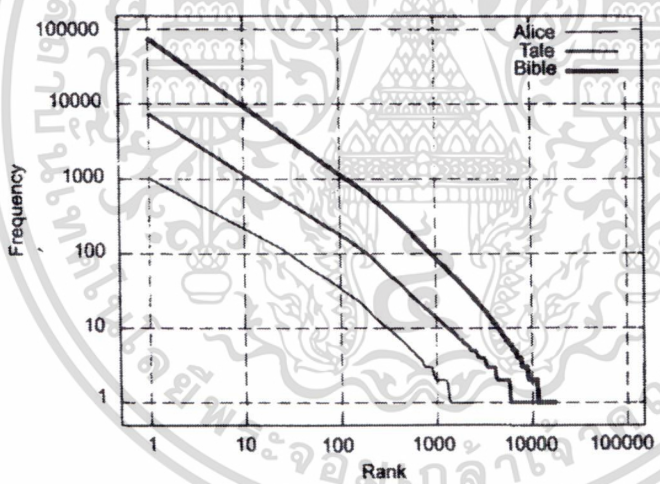
G.K. Zipf (1949) กล่าวว่า เราจะมีการใช้คำที่ไม่มากนักและมักจะเป็นคำที่ซ้ำ ๆ กับที่เคยใช้และมีการใช้คำอื่น ๆ น้อยครั้ง ซึ่งการใช้คำใด ๆ ซ้ำ ๆ นั้นอาจเนื่องจากความคุ้นเคยและการใช้งานคำนั้นบ่อยทำให้ไม่ค่อยได้นำเอาศัพท์ใหม่มาใช้ ในอดีตพบว่ามีคำจำนวนน้อยกว่าร้อยละ 20 ที่สามารถใช้อธิบายข้อความจากข้อความทั้งหมด ดังนั้นจึงเกิดความสัมพันธ์ระหว่างลำดับของคำกับความถี่ของคำที่ปรากฏในเอกสาร

ถ้าเรียงลำดับความถี่ของการปรากฏของคำศัพท์ในเอกสารจากมากไปน้อย คำว่า "the" จะเป็นคำศัพท์ที่พบมากที่สุดในเอกสารเป็นลำดับที่หนึ่ง และ "of" จะปรากฏมากเป็นลำดับที่สอง ซึ่งข้อมูลนี้มากจากการทดสอบจากกลุ่มเอกสารของ Reuters

จาก Zipf's Law กำหนดว่าผลคูณของความถี่ของคำกับลำดับของคำนั้นในเอกสารจะมีค่าใกล้เคียงหรือเหมือนกับผลคูณของความถี่กับลำดับของคำอื่น ถ้าสอดคล้องกับ Zipf's Law แล้ว ผลคูณของลำดับและความถี่นั้นจะเป็นค่าคงที่ที่หยาบ ๆ จากตารางที่ 2.3 พบว่าผลคูณของลำดับและความถี่มีความแปรปรวนตั้งแต่คำแรกจนถึงคำสุดท้ายในตาราง สามารถแสดงเป็นกราฟ logarithm ได้ดังรูปที่ 2.8 โดยกำหนดให้แกน x แทนด้วยลำดับ และแกน y แทนด้วยความถี่ของแต่ละลำดับ เส้นกราฟที่ปรากฏจะไม่เป็นเส้นตรงโดยจะโป่งตรงส่วนกลางของกราฟแต่ละเส้นซึ่งเป็นส่วนที่มีค่าผลคูณมากที่สุด และมีความถี่มากในส่วนต้นของกราฟ และลดลงในส่วนท้ายของกราฟ

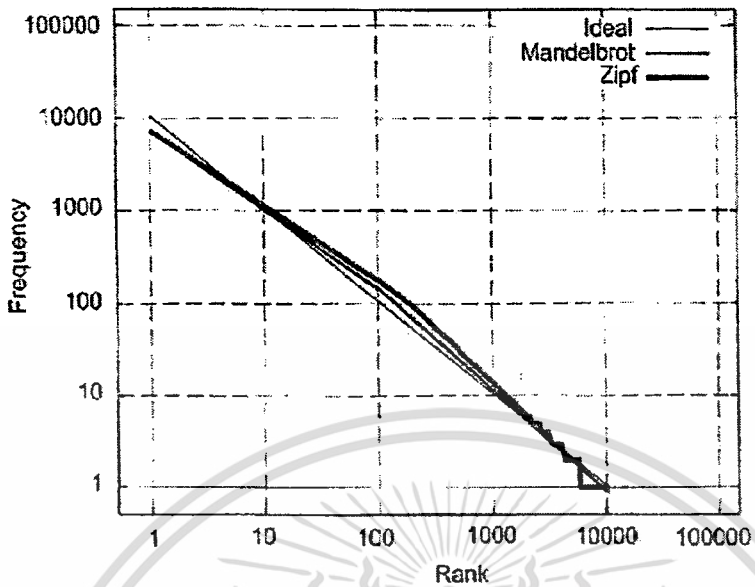
ตารางที่ 2.3 ตารางแสดงค่าลำดับที่ของคำ (Rank) ความถี่ของคำ (Freq) และผลคูณของค่าลำดับกับความถี่

Word	Rank	Freq	Rank*F	Word	Rank	Freq	Rank*F
the	1	120021	120021	investors	400	828	331200
of	2	72225	144450	Head	800	421	336800
and	4	53462	213848	warrant	1600	184	294400
For	8	25578	204624	Tehran	3200	73	233600
is	16	16739	267824	Guarantee	6400	25	160000
company	32	9340	298880	Pittiston	10000	11	110000
Co.	64	4005	256320	Thinly	20000	3	60000
quarter	100	2677	267700	Morgenthaler	40000	1	40000
unit	200	1489	297800	tabulating	47075	1	47075



รูปที่ 2.8 กราฟแสดงความสัมพันธ์ระหว่างความถี่และลำดับ ของเอกสาร Alice, Tale และ Bible (Konchady, 2006)

การหาค่าความถี่ f=K/r กำหนดให้ r คือลำดับ และ K คือค่าคงที่ ในขณะที่ B.B. Mandelbrot (1953) ได้ปรับปรุงสูตรข้างต้นโดยเพิ่มตัวแปร c และ θ ซึ่งสามารถคำนวณหาค่าความถี่ได้เป็น $f=K/(c+r)^{\theta}$ กำหนดให้ K คือความถี่ที่เพิ่มจนจนกระทั่งถึงจำนวนคำทั้งหมด และค่า c อยู่ระหว่าง 1 ถึง 100 และค่า θ จะขึ้นอยู่กับเอกสาร จากรูป กำหนด θ มีค่าอยู่ระหว่าง 1 และ 2 แสดงได้ดังรูปที่ 2.9



รูปที่ 2.9 กราฟแสดงความสัมพันธ์ระหว่างความถี่และลำดับที่ใช้กฎ Zipf และ Mandelbrot (Konchady, 2006)

จากรูป 2.9 เป็นกราฟที่เปรียบเทียบระหว่างความถี่กับลำดับ โดยเส้นกราฟจะแสดงการกระจายของความถี่ ในการประยุกต์ใช้ Zipf's Law นี้จำนวนของคำที่แตกต่างกันต้องใกล้เคียงกับจำนวนของการเกิดขึ้นความถี่ของคำที่มากที่สุด ถ้ากลุ่มตัวอย่างที่ใช้ในการทดสอบมีไม่มาก เราอาจจะพบจำนวนของคำเป็นจำนวนมากที่เกิดขึ้นเพียงครั้งเดียว ขณะที่มีกลุ่มตัวอย่างที่มากก็อาจแทบจะไม่มีคำที่เกิดขึ้นเพียงครั้งเดียว ขนาดของคำที่เหมาะสมในการประยุกต์ใช้กฎนี้ควรจะมีประมาณ 120,000 คำ

งานวิจัยส่วนใหญ่ได้นำกฎของ Zipf มาใช้ในการวิเคราะห์การกระจายของคำในเอกสาร คำที่ใช้บ่อยในกลุ่มเอกสารและคำที่ใช้ค่อนข้างน้อยในเอกสาร งานวิจัยของ G. Forman (2003) ได้นำกฎของ Zipf มาช่วยในการวิเคราะห์ความถี่ของคำในเอกสาร โดยคำที่ความถี่น้อยซึ่งมีแนวโน้มว่าจะไม่มีความสำคัญต่อเอกสารนั้นจะถูกตัดออก ในการตัดคำที่มีความถี่น้อยนั้นจะถูกกำหนดโดยค่า Threshold = 3 หมายความว่าความถี่ของคำที่มีค่ามากกว่า 3 จะถูกพิจารณา แต่ถ้าความถี่ของคำใดน้อยกว่า 3 แล้วคำนั้นจะถูกตัดออก โดยในงานวิจัยนี้มีคำที่ถูกตัดออกเป็นจำนวน 7333 คำ

นอกจากนั้น งานวิจัย (Dahui et al., 2005) และ (Xiao, 2008) ได้ใช้กฎของ Zipf ในการวิเคราะห์ความถี่หรือการกระจายของคำในรูปแบบต่าง ๆ ของเอกสารภาษาต่าง ๆ ได้แก่ ภาษาจีน ภาษาฝรั่งเศส

2.4 งานวิจัยที่เกี่ยวข้อง

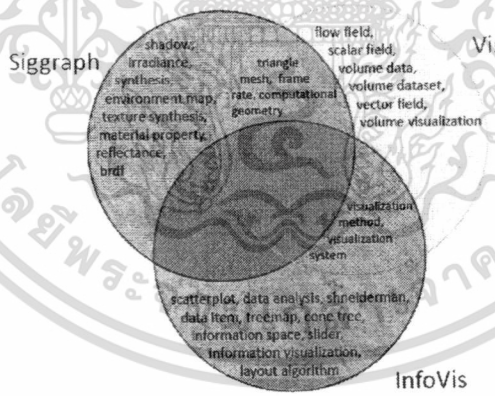
เนื่องด้วยข้อมูลโดยส่วนใหญ่แล้วอยู่ในรูปแบบที่ไม่เป็นโครงสร้าง เช่น เอกสารรายงาน อีเมล์ ข่าว เป็นต้น งานวิจัยทางด้านการจัดกลุ่มเอกสารจึงเป็นงานหนึ่งที่ช่วยให้ผู้ใช้งานสามารถเข้าถึงข้อมูลได้สะดวกขึ้น โดยทำการจัดกลุ่มเอกสารตามลักษณะเนื้อหาหรือข้อความที่คล้ายคลึง ซึ่งระบบจะต้องเรียนรู้การจัด

กลุ่มจากกลุ่มข้อมูลฝึกสอนเพื่อสร้างโมเดล และทดสอบโมเดลที่สร้างขึ้นด้วยข้อมูลทดสอบ ซึ่งงานวิจัยที่ศึกษานี้จะเป็นงานวิจัยที่เกี่ยวข้องกับการเลือกคุณลักษณะที่ใช้ในการจัดกลุ่มข้อมูล

โดยทั่วไปแล้วในการกำหนดค่าน้ำหนักนั้นวิธีการที่เป็นที่นิยมได้แก่ TFIDF (Jing et al, 2002, Liao et al, 2003, Liu et al., 2007), Information Gain (IG) (Bong and Narayanan, 2004, Gabrilovich and Markovitch, 2004, Liu et al., 2007, Brank et al., 2008, Li at el., 2009) และ Chi-Square (CHI) (Bong and Narayanan, 2004, Gabrilovich and Markovitch, 2004, Liu et al., 2007, Li at el., 2009) ซึ่งมักนำไปเป็นมาตรฐานในการเปรียบเทียบวิธีการใหม่ที่นำเสนอ แต่อย่างไรก็ตามวิธีดังกล่าวยังคงให้ผลลัพธ์ที่มีประสิทธิภาพสูงเช่นเดียวกัน

(Daniel et. al, 2009) นำเสนอวิธีการในการกำหนดกลุ่มของคำที่เกี่ยวข้องกับประเภทเอกสาร โดยทั่วไปแล้วคำ ๆ หนึ่งสามารถปรากฏเป็นคำสำคัญได้ในประเภทเอกสารอื่นได้มากกว่าหนึ่งประเภทเอกสาร แต่สำหรับในงานวิจัยนี้จะพิจารณาว่าคำที่กำหนดเป็นคำสำคัญในแต่ละประเภทเอกสารนั้น จะต้องไม่มีความสัมพันธ์กับประเภทเอกสารอื่น โดยแยกคำที่มีความแตกต่างจากกลุ่มเอกสารประเภทอื่น และคำที่ปรากฏร่วมกันในประเภทเอกสารต่าง ๆ ดังนั้นในงานวิจัยนี้จึงทำการระบุกลุ่มคำสำคัญที่ปรากฏเฉพาะประเภทเอกสาร และปรากฏร่วมกันในเอกสารหลายประเภท

ตัวอย่างเช่น ต้องการระบุคำที่สามารถกำหนดเป็นคำสำคัญที่ไม่ปรากฏในเอกสารประเภทอื่น กับคำที่ปรากฏร่วมกันในเอกสารประเภทต่าง ๆ โดยแบ่งประเภทเอกสารออกเป็น 3 ประเภท แสดงดังรูปที่ 2.10



รูปที่ 2.10 แสดงกลุ่มคำศัพท์ที่ปรากฏในแต่ละประเภทเอกสาร (Daniel et. al, 2009)

จากรูปที่ 2.10 ประเภทเอกสารประกอบด้วย Siggraph, Vis และ InfoVis ซึ่งเอกสารเหล่านั้นจะปรากฏคำที่ใช้ร่วมกันระหว่างประเภทเอกสารและคำที่ใช้เฉพาะแต่ละประเภทเอกสาร คำสำคัญเฉพาะประเภทเอกสาร Siggraph ได้แก่ shadow, synthesis, textual synthesis, environment map เป็นต้น คำสำคัญเฉพาะประเภทเอกสาร Vis ได้แก่ flow field, vector field, volume data เป็นต้น และคำสำคัญเฉพาะประเภทเอกสาร InfoVis ได้แก่ scatterplot, data analysis, cone tree เป็นต้น คำสำคัญที่ปรากฏร่วมกันระหว่าง

ประเภทเอกสาร Siggraph และ Vis ได้แก่ triangle, mesh, frame เป็นต้น และคำสำคัญที่ปรากฏร่วมกันระหว่างประเภทเอกสาร Vis และ InfoVis ได้แก่ visualization method, visualization system

ในการแบ่งกลุ่มเอกสารนั้นจะใช้คำที่ปรากฏเฉพาะเอกสารในการคำนวณค่าน้ำหนัก เนื่องจากว่าคำเหล่านี้น่าจะสามารถแบ่งแยกเอกสารได้ดีกว่าที่การใช้คำที่ปรากฏร่วมระหว่างประเภทเอกสารมาพิจารณาร่วมในการแบ่งกลุ่มเอกสาร การคำนวณค่าน้ำหนักนั้นจะใช้วิธี TFICF ซึ่งปรับปรุงมาจากวิธี TFIDF

การพิจารณาคำที่ปรากฏเฉพาะประเภทเอกสารกับคำปรากฏร่วมประเภทเอกสาร จะพิจารณาโดยการคำนวณค่าน้ำหนักของคำในแต่ละประเภทเอกสาร ถ้าคำนั้นมีค่าน้ำหนักสูงในประเภทเอกสารหนึ่งและมีคะแนนน้อยในประเภทเอกสารอื่น และค่าคะแนนนั้นต่ำกว่าค่า Threshold ที่กำหนด สรุปได้ว่าคำนั้นสามารถกำหนดเป็นตัวแทนของเอกสารที่มีคะแนนที่มากเพียงอย่างเดียว ในขณะที่คำที่จะเป็นคำที่ปรากฏร่วมระหว่างประเภทเอกสารนั้น จะเป็นคำที่มีค่าน้ำหนักในประเภทเอกสารต่าง ๆ มากกว่าค่า threshold ที่กำหนด

การหากลุ่มคำสำคัญนั้น ในงานวิจัยนี้ได้ทำการทดลองกับตัวอย่างเอกสารที่ประกอบด้วยเอกสารประเภทงานวิจัยที่เกี่ยวข้องกับหัวข้อ InfoVis, Siggraph และ Vis โดยแต่ละหัวข้อประกอบด้วยเอกสารจำนวนหัวข้อละ 100 เอกสาร ที่ใช้ในการเรียนรู้เพื่อกำหนดกลุ่มคำสำคัญในแต่ละประเภทเอกสาร การทดลองนี้ประกอบด้วยขั้นตอนต่อไปนี้

(1) กำกับหน้าที่ของคำ ระบุว่าเป็นคำนาม คำกริยา คำคุณศัพท์ เป็นต้น

(2) ระบุนามวลีที่ปรากฏในเอกสาร

(3) หาค่าน้ำหนักด้วยวิธี TFICF, TFIDF average, TFIDF max และ differential analysis ให้กับกลุ่มคำในข้อ (2) เพื่อเปรียบเทียบประสิทธิภาพของวิธีการต่าง ๆ ข้างต้น เมื่อคำนวณหาค่าน้ำหนักและเลือกกลุ่มคำสำคัญที่ทำหน้าที่เป็นตัวแทนของประเภทเอกสารเพียงหัวข้อเดียวจำนวนหัวข้อละ 15 กลุ่มคำสำคัญ โดยแต่ละวิธีอธิบายสรุปได้ดังนี้

- TFIDF average คำนวณค่าน้ำหนักให้กับกลุ่มคำในเอกสารจำนวน 300 เอกสารด้วย TFIDF จากนั้นแยกเอกสารออกเป็นแต่ละประเภทตามที่กำหนด หาค่าน้ำหนักเฉลี่ยของแต่ละคำในแต่ละประเภทเอกสาร แล้วทำการเรียงลำดับคำในแต่ละประเภท และเลือกคำอันดับมากสุด 15 อันดับแรก
- TFIDF max วิธีการเหมือนวิธีข้างต้น แต่ใช้การหาค่าน้ำหนักมากที่สุดของแต่ละคำแทนการหาค่าเฉลี่ย
- Different analysis วิธีการนี้จะใช้จะทำการเปรียบเทียบความน่าจะเป็นของการเกิดขึ้นของคำในเอกสารทดสอบกับเอกสารอ้างอิง
- TFICF กำหนดค่า threshold เท่ากับ 2.0 และคำที่ถูกกำหนดจะต้องปรากฏในเอกสารประเภทนั้นมากกว่า 10% ของเอกสารทั้งหมดของแต่ละประเภท

ในการทดสอบประสิทธิภาพการจัดกลุ่มเอกสารจากกลุ่มคำสำคัญที่หาได้จากวิธีการข้างต้นแล้วได้นำเอกสารจำนวน 60 เอกสารและเป็นคนละกับเอกสารข้างต้นมาใช้การทดสอบ แบ่งเป็นหัวข้อละ 20 เอกสาร โดยในการระบุหัวข้อให้กับเอกสารจำนวน 60 เอกสารเหล่านี้ จะพิจารณาว่าเอกสารนี้ประกอบด้วยกลุ่มคำในประเภทใดมากที่สุด ก็จะกำหนดให้เอกสารอยู่ในประเภทนั้น

จากการทดลองสามารถสรุปได้ว่า กลุ่มคำที่คำนวณค่าน้ำหนักด้วยวิธี TFICF นั้นจะให้ค่าความถูกต้องมากที่สุดถึง 0.91 ในขณะที่วิธี TFIDF average, TFIDF max และ Different analysis จะให้ค่าความถูกต้องเท่ากับ 0.71, 0.77 และ 0.78 ตามลำดับ แต่อย่างไรก็ตามยังมีเอกสารจำนวนหนึ่งที่ไม่สามารถระบุประเภทได้

(Bong and Narayanan, 2004) นำเสนอวิธีการคำนวณค่าน้ำหนักของเทอมที่มีชื่อว่า Categorical Descriptor Term (CTD) เพื่อลดขนาดของคุณลักษณะของเวกเตอร์เอกสาร โดยปรับปรุงมาจากเทคนิคการหาค่าน้ำหนัก TFIDF โดยกำหนดคุณลักษณะของแต่ละกลุ่มเอกสารให้มีคุณลักษณะที่แตกต่างกัน คำไหนยิ่งปรากฏน้อยกลุ่มก็ยิ่งมีความสามารถกำหนดเป็นคุณลักษณะของกลุ่มนั้นมาก แสดงได้ดังสมการที่ (2.24)

$$CTD(t_k, c_i) = TF(t_k, c_i) \times IDF(t_k, c_i) \times ICF(t_k) \quad (2.24)$$

$$ICF(t_k) = \log\left[\frac{|C|}{CF(t_k)}\right], IDF(t_k, c_i) = \log\left[\frac{|D(c_i)|}{DF(t_k, c_i)}\right] \quad (2.25)$$

โดยที่ $D(c_i)$ คือ จำนวนเอกสารในกลุ่ม c_i
C คือ จำนวนกลุ่มของเอกสาร
$CF(t_k)$ คือ จำนวนกลุ่มที่เทอม t_k ปรากฏในเอกสารของกลุ่มนั้น
$DF(t_k,c_i)$ คือ จำนวนเอกสารที่ปรากฏเทอม t_k ในกลุ่ม c_i

จากสูตรที่ (2.24) และ (2.25) จะพิจารณาค่าน้ำหนักของคำจากการปรากฏในแต่ละกลุ่มประเภทเอกสาร จำนวนเอกสารที่ปรากฏคำนั้นในแต่ละประเภทเอกสาร และจำนวนประเภทเอกสารที่ปรากฏคำ ๆ นั้น

สำหรับในงานวิจัยนี้ได้ทดสอบวิธีการที่นำเสนอกับกลุ่มเอกสารที่ชื่อ Reuters-21578, 20 newsgroup และ เอกสารงานวิจัยจาก Technology and Teacher Education Annual เอกสารที่ใช้เรียนรู้มีจำนวน 1121 เอกสาร ใน 25 กลุ่มประเภทเอกสาร และเอกสารที่ใช้ในการทดสอบจำนวน 414 เอกสาร โดยทำการเปรียบเทียบกับวิธีการคำนวณค่าน้ำหนักอื่นๆได้แก่ Information gain, Chi-Square, Correlated Coefficient, Odd ratio และ GSS Coefficient จากการทดลองพบว่า CTD สามารถแบ่งกลุ่มได้อย่างมีประสิทธิภาพดีในกลุ่มเอกสารที่มีความใกล้เคียงกัน

(Li at el., 2009) นำเสนอวิธีการของการลดคุณลักษณะในการจัดกลุ่มเอกสารด้วยวิธีการคำนวณค่าน้ำหนักที่เรียกว่า weight frequency and odds (WFO)

การเลือกคุณลักษณะในงานวิจัยนี้ จะเลือกภายใต้เงื่อนไขทั้งสองอย่างคือ

1. คุณลักษณะที่ดีจะมีความถี่ของเอกสารที่ปรากฏคุณลักษณะนั้นมาก
2. คุณลักษณะที่ดีจะมีค่าสัดส่วนของกลุ่มประเภทเอกสารที่มาก (สัดส่วนระหว่างจำนวนเอกสารที่มีเทอม t และอยู่ในคลาส c ต่อ จำนวนเอกสารที่มีเทอม t แต่ไม่อยู่ในคลาส c)

และสูตรการคำนวณค่าน้ำหนักด้วย WFO แสดงได้ดังสมการที่ (2.26)

$$WFO(t,c_i) = P(t|c_i)^{\lambda}\left[log\frac{P(t|c_i)}{P(t|\bar{c}_i)}\right]^{1-\lambda} when\frac{P(t|c_i)}{P(t|\bar{c}_i)} > 1 \quad (2.26)$$

มิฉะนั้น

$$WFO(t,c_i = 0))$$

หรือ

$$WFO = \left(\frac{A_i}{N_i}\right)^{\lambda}\left(log\frac{A_i\times(N_{all}-N_i)}{B_i\times N_i}\right)^{1-\lambda} \quad (2.27)$$

โดยที่ λ คือตัวแปรค่าน้ำหนัก มีค่าอยู่ระหว่าง 0 ถึง 1 โดยค่าที่เหมาะสมนั้นจะได้จากการเรียนรู้จากกลุ่มตัวอย่างฝึกสอน

ในการทดลองงานวิจัยนี้ทดสอบกับกลุ่มข้อมูล Reuters-21578 จำนวน 2,000 เอกสาร 20 Newsgroup จำนวน 20,000 เอกสาร Cornell movie-review และ DVD reviews จำนวนอย่างละ 2,000 เอกสาร และอัลกอริทึมที่ในการจัดกลุ่มข้อมูลคือ ซัพพอร์ทเวกเตอร์แมชชีน (Support vector machine หรือ SVM)

การทดลองเพื่อเปรียบเทียบค่าน้ำหนักและเลือกคุณลักษณะที่ดีเพื่อใช้ในการจัดกลุ่มข้อมูลนี้ ได้ทำการเปรียบเทียบวิธี WFO กับ Document Frequency (DF), Mutual Information (MI), Information gain (IG), Chi-Square (CHI), Bi-Normal Separation (BNS) และ Weighed Log Likelihood Ratio (WLLR)

DF	$DF = \sum_{i=1}^{m} A_i$
MI	$MI = -log\left(1+\frac{1}{\frac{A_i}{B_i}}\right) - log\frac{N_i}{N_{all}}$

IG	$IG = \left(-\sum_{i=1}^{m}\frac{N_i}{N_{all}} log \frac{N_i}{N_{all}}\right) + \left(\sum_{i=1}^{m}\frac{A_i}{N_{all}}\right)\left(\sum_{i=1}^{m}\frac{A_i}{A_i+B_i} log \frac{A_i}{A_i+B_i}\right) + \left(\sum_{i=1}^{m}\frac{C_i}{N_{all}}\right)\left(\sum_{i=1}^{m}\frac{C_i}{C_i+D_i} log \frac{C_i}{C_i+D_i}\right)$
CHI	$CHI = \frac{2N_i(\frac{A_i}{B_i}-1)^2}{\left(\frac{A_i}{B_i}+1\right)\left(\frac{2N_i}{A_i}\times\frac{A_i}{B_i}-\left(\frac{A_i}{B_i}+1\right)\right)}$
BNS	$BNS = \left\lvert F^{-1}\left(\frac{A_i}{N_i}\right) - F^{-1}\left(\frac{B_i}{N_{all}-N_i}\right)\right\rvert$
WLLR	$WLLR = \frac{A_i}{N_i} log \frac{A_{i(N_{all}-N_i)}}{B_i \times N_i}$

โดยที่ A_i คือ จำนวนเอกสารที่ประกอบด้วยเทอม t และเป็นเอกสารในกลุ่ม c_i

B_i คือ จำนวนเอกสารที่ประกอบด้วยเทอม t แต่ไม่ได้อยู่ในกลุ่ม c_i

N_i คือ จำนวนเอกสารทั้งหมดในกลุ่ม c_i

N_{all} คือ จำนวนเอกสารทั้งหมดที่ใช้ในการฝึกสอน

C_i คือ จำนวนเอกสารที่ไม่มีเทอม t แต่อยู่ในกลุ่ม c_i

D_i คือ จำนวนเอกสารที่ไม่มีเทอม t และไม่อยู่ในกลุ่ม c_i

จากการทดลองเพื่อวัดประสิทธิภาพ โดยแบ่งแยกตามจำนวนของคุณลักษณะ ของแต่ละกลุ่มเอกสารฝึกสอน

ถ้าจำนวนของคุณลักษณะมีไม่มาก (น้อยกว่า 1,000) IG, CHI และ WLLR จะให้ค่าผลลัพธ์การจัดกลุ่มที่ดีกว่า ขณะที่ WFO ก็คงให้ประสิทธิภาพดีเช่นเดียวกัน ที่ตัวแปร λ = 0.5 แต่ถ้าเพิ่มขนาดของคุณลักษณะให้มีจำนวนมากขึ้น พบว่า MI และ BNS ให้ประสิทธิภาพดีกว่าในกลุ่มเอกสาร 20 Newsgroup และ Movie ขณะที่ IG และ CHI ให้ประสิทธิภาพดีในกลุ่มเอกสาร DVD ส่วน WFO ก็ยังคงให้ประสิทธิภาพที่ดีทั้งสามกลุ่มเอกสาร

(Liu et al., 2007) นำเสนอวิธีการคำนวณค่าน้ำหนักของเทอมที่ชื่อว่า Category-Based Term Weights (CBTWs) โดยในงานวิจัยนี้จะแทนการคำนวณค่า idf ด้วยค่าที่กำหนดจากขั้นตอนการเลือกเทอมที่สำคัญที่นำเสนอ การคำนวณค่าน้ำหนักของเทอมจะใช้วิธีการดังนี้

กำหนดให้ A คือ จำนวนเอกสารที่อยู่ในกลุ่ม c_i ซึ่งมีเทอม t_k ปรากฏอย่างน้อยหนึ่ง

B คือ จำนวนเอกสารที่ไม่อยู่ในกลุ่ม c_i ซึ่งมีเทอม t_c ปรากฏอย่างน้อยหนึ่ง

C คือ จำนวนเอกสารที่อยู่ในกลุ่ม c_i ซึ่งไม่ปรากฏเทอม t_c เลย

A/B หมายความว่าถ้าเทอม t_k มีความเกี่ยวข้องกับกลุ่ม c_i มากเพียงกลุ่มเดียว จะกล่าวว่า เทอม t_k เป็นคุณลักษณะที่ดีที่จะเป็นตัวแทนของกลุ่ม c_i แล้วค่าของ A/B มีแนวโน้มที่จะสูง

A/C หมายความว่าระหว่างเทอมเทอม t_k กับ t_k เทอมใดมีค่า A/C สูงกว่า แล้วเทอมนั้นจะเป็นคุณลักษณะที่ดีกว่าของกลุ่ม c_i

จากอัตราส่วนที่กล่าว CBTW สามารถแสดงได้ดังสูตร

CBTW1	$\log(1+\frac{A}{B}\frac{A}{C})$
CBTW2	$\log(1+\frac{A}{B}+\frac{A}{C})$
CBTW3	$\log(1+\frac{A}{B})\log(1+\frac{A}{C})$
CBTW4	$\log\left[\left(1+\frac{A}{B}\right)\left(1+\frac{A}{C}\right)\right]$
CBTW5	$\log(1+\frac{A+B}{B}\frac{A+C}{C})$
CBTW6	$\log(1+\frac{A+B}{B}+\frac{A+C}{C})$
CBTW7	$\log(1+\frac{A+B}{B})\log(1+\frac{A+C}{C})$
CBTW8	$\log\left[\left(1+\frac{A+B}{B}\right)\left(1+\frac{A+C}{C}\right)\right]$

A/B สามารถอธิบายได้ว่าเทอมใดยิ่งมีค่าอัตราส่วนนี้สูง แล้วเทอมนั้นยิ่งมีความสำคัญต่อกลุ่ม ในทำนองเดียวกันกับ A/C ว่าเทอมใดที่ถูกลงความเห็นว่ามีความเกี่ยวข้องมากนั้นคือปรากฏเป็นส่วนใหญ่ในกลุ่มใดๆ มากกว่ากลุ่มอื่นแล้วเทอมนั้นจะมีความสำคัญต่อกลุ่มที่ปรากฏเป็นส่วนใหญ่มาก

เอกสารที่ใช้ในการทดสอบได้แก่ MCV1 และ Reuters-21578 โดยที่ MCV1 ประกอบด้วย 18 กลุ่มเอกสารและ Reuters-21578 ประกอบด้วย 13 กลุ่ม และแต่ละกลุ่มนั้นมีจำนวนเอกสารที่ไม่เท่ากัน โดยทำการเปรียบเทียบวิธีการที่นำเสนอกับวิธีคิดค่าน้ำหนักอื่นๆ ในขั้นตอนของการเลือกคุณลักษณะได้แก่ Chi-Square, Correction coefficient, odds ratio, information gain และ relevance frequency ซึ่งสูตรเหล่านี้จะแทนในส่วนของ IDFในสูตร TFIDF และค่า TF สามารถคำนวณได้จากสมการที่ (2.28)

$$tf(t_i,d_j)=\frac{tf(t_i,d_j)}{\max[(tf(d_j))]} \qquad (2.28)$$

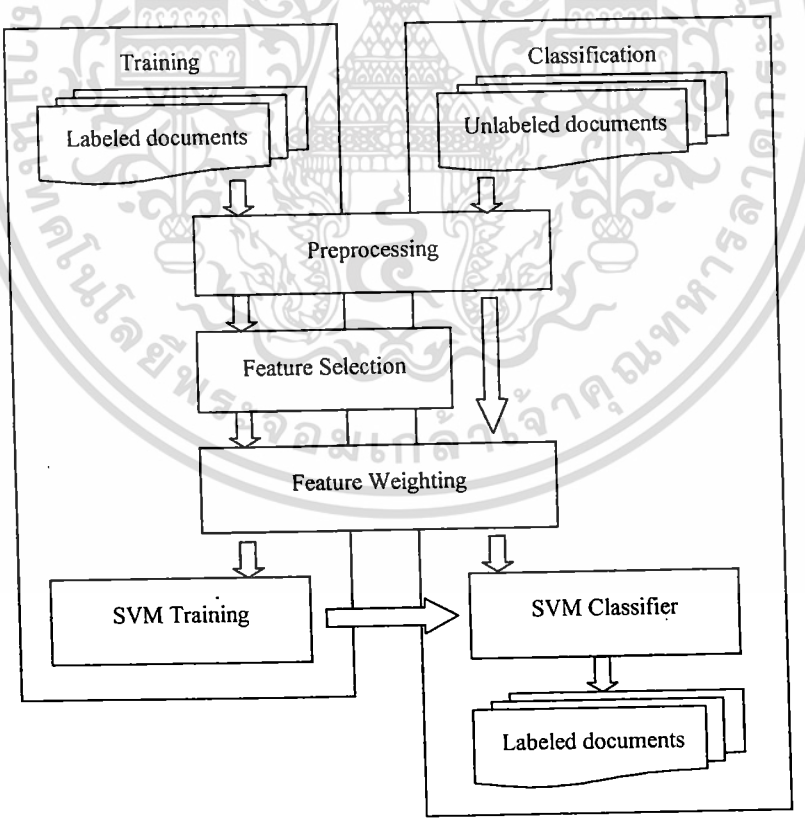
โดยที่ tf(ti,df) คือความถี่ของเทอม t_i ในเอกสาร d_j

max[tf(dj)] คือค่าความถี่ที่มากที่สุดของเทอมในเอกสาร d_j

ในการแบ่งกลุ่มนั้นจะใช้อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนในการจัดกลุ่ม จากการทดลองพบว่า CBTW1 นั้นมีประสิทธิภาพดีที่สุดในเอกสารทั้งสองกลุ่มตัวอย่าง

บทที่ 3
กระบวนการในการกำหนดหัวข้อข่าวให้กับเอกสาร

ในบทนี้กล่าวถึงกระบวนการกำหนดหัวข้อข่าวให้กับเอกสาร โดยใช้เทคนิคการจัดกลุ่มเอกสารเพื่อระบุว่าเอกสารข่าวแต่ละเอกสารนั้นมีหัวข้อข่าวประเภทใด ดังนั้นบทที่จึงกล่าวถึงกระบวนการหรือขั้นตอนในการจัดกลุ่มเอกสาร เพื่อให้ทราบว่าตัวแทนของแต่ละประเภทข่าวประกอบด้วยคำใดบ้าง และคำเหล่านั้นมีประสิทธิภาพพียงใดในการนำมาใช้กำหนดหวัข้อข่าว โดยกระบวนการนั้นประกอบด้วย การตัดคำฟุ่มเฟือย และแปลงคำศัพท์ที่เหลือให้อยู่ในรูปรากศัพท์เดิม ซึ่งเป็นขั้นตอนของการเตรียมข้อมูล จากนั้นเลือกคุณลักษณะที่จะนำมาใช้เป็นข้อมูลในการจัดกลุ่มเอกสาร โดยในการเลือกคุณลักษณะนี้จะอาศัยการคำนวณค่าน้ำหนักของคำโดยในเอกสารนี้คุณลักษณะนั้นหมายถึงคำแต่ละคำที่ทำหน้าที่เป็นตัวแทนของประเภทเอกสารและจะใช้คำนี้แทนคำว่าคุณลักษณะ แล้วจัดรูปแบบเอกสารให้อยู่ในรูปแบบของเวกเตอร์ขนาดตามจำนวนคุณลักษณะที่ถูกเลือก เพื่อนำไปสร้างโมเดลการจัดกลุ่มเอกสารด้วยเทคนิคซัพพอร์ตเวกเตอร์แมทชีน ซึ่งกระบวนการนี้แสดงดังรูปที่ 3.1

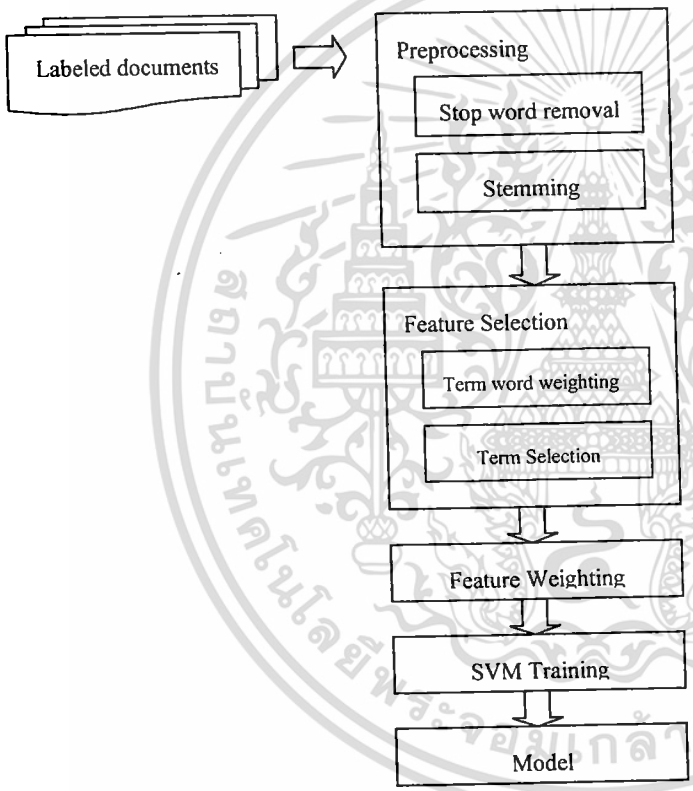


รูปที่ 3.1 กระบวนการในการจัดกลุ่มเอกสารในงานวิจัยนี้

3.1 กระบวนการจัดกลุ่มเอกสาร

จากรูปที่ 3.1 แบ่งขั้นตอนการทำงานออกเป็น 2 ส่วน ได้แก่ ส่วนการเรียนรู้ (Training) โดยการฝึกสอนจากข้อมูลกลุ่มตัวอย่างที่กำหนดให้ โดยผลลัพธ์จากส่วนนี้จะได้โดเมลเพื่อนำไปใช้ในส่วนถัดไป และส่วนการทดสอบการจัดกลุ่ม (Classification) โดยในส่วนนี้จะทำการทดสอบโมเดลที่ได้จากการเรียนรู้กับกลุ่มข้อมูลชุดใหม่ เพื่อทดสอบประสิทธิภาพของโมเดล

3.1.1 ส่วนการเรียนรู้ (Training) เป็นการเรียนรู้จากกลุ่มเอกสารตัวอย่างที่มีกำหนดประเภทเอกสารที่ถูกต้องไว้แล้ว โดยประกอบด้วยขั้นตอนแสดงดังรูปที่ 3.2



รูปที่ 3.2 ขั้นตอนในส่วนของการเรียนรู้

1. **การเตรียมข้อมูลสำหรับการเรียนรู้ (Preprocessing)** เป็นขั้นตอนในการเตรียมเอกสารที่จะนำมาใช้ในการจัดกลุ่มเอกสารสำหรับงานวิจัยนี้ โดยมีกระบวนการดังต่อไปนี้

การรวบรวมเอกสาร

รวบรวมเอกสารจากเว็บข่าวต่าง ๆ ได้แก่ Yahoo (http://news.yahoo.com), Accuweather (http://www.accuweather.com), New York Time (http://www.nytimes.com), CNN (http://www.cnn.com), News Week (http://www.newsweek.com) เป็นต้น ซึ่งเอกสารที่รวบรวมประกอบด้วยหัวข้อข่าวเกี่ยวกับกีฬา การเมือง บันเทิง สุขภาพ สภาพอากาศ และเศรษฐกิจ จำนวนหัวข้อละ 1,000 เอกสาร รวมทั้งหมดเป็น

6,000 เอกสาร โดยเอกสารเหล่านั้นอยู่ในรูปแบบ HTML จากนั้นเลือกเฉพาะข้อความข่าวโดยใช้ API ส่วนที่เป็นรูปภาพ มัลติมีเดีย ข้อคิดเห็นต่าง ๆ เกี่ยวกับข่าว เมนู โฆษณา จะไม่ถูกเลือก

ตัดคำฟุ่มเฟือย (Stop word Removal)

เป็นขั้นตอนอ่านเอกสารเพื่อทำการตัดคำที่ไม่มีความสำคัญ หรือคำที่พบในเอกสารบ่อยแต่ไม่ได้เป็นประเด็นสำคัญในเอกสาร เช่น คำสันธาน คำบุพบท คำระบุนาม คำสรรพนาม เป็นต้น นอกจากนี้คำในกลุ่มที่ใช้บ่อยที่ระบุโดย () ได้ถูกนำมาใช้เป็นคำฟุ่มเฟือยด้วย เช่น country, close, together, seem, stop, best, remember, early, morning, common, miss เป็นต้น

การแปลงคำให้อยู่ในรูปรากศัพท์ (Stemming)

เป็นขั้นตอนการแปลงคำศัพท์ให้กลับไปอยู่ในรูปแบบรากศัพท์ หรือ base form เนื่องจากว่าคำในภาษาอังกฤษสามารถเปลี่ยนรูปแบบไปได้หลายรูปแบบ เช่น ในรูปของอดีต ปัจจุบัน อนาคต ทำให้เป็นคำนาม ซึ่งคำเหล่านั้นอาจจะมาจากรากศัพท์เดียวกัน ในขั้นตอนนี้จึงเป็นการลดจำนวนคำที่ไม่ซ้ำในเอกสาร โดยอัลกอริทึมที่นิยมใช้คือ Porter Stemming ตัวอย่างเช่น

surprise -> surprise

losing, loses -> lose

losses -> loss

intended -> intend

saving, save -> save

recently -> recent

แม้ว่าในการแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์ของแต่ละคำนั้น เมื่อใช้อัลกอริทึม Porter Stemming ในการแปลงแล้ว คำบางคำจะไม่สามารถระบุความหมายได้ ก็ไม่ส่งผลต่อการจัดกลุ่มเอสการ เพราะว่าเราสนใจที่จำนวนของคำที่ปรากฏในเอกสารโดยไม่คำนึงถึงความหมายของคำเหล่านั้น

2. **การเลือกคำที่ทำหน้าที่เป็นคุณลักษณะ (Feature Selection)** ในขั้นตอนนี้จะเลือกคำที่สามารถกำหนดเป็นตัวแทนของกลุ่มเอกสาร โดยคำที่ถูกเลือกนั้นจะพิจารณาจากการคำนวณค่าน้ำหนักให้กับคำนั้น ๆ แล้วเลือกคำที่เหมาะสม นอกจากนั้นแล้วยังทำให้ขนาดของเวกเตอร์ของเอกสารที่ขนาดลดลงแทนที่จะใช้คำทุกคำมาเป็นคุณลักษณะ วิธีการในการเลือกคำนั้นในงานวิจัยนี้ได้ใช้เทคนิค TFIDF, TFICF, IG และ CHI-square โดยเริ่มแรกคำจะถูกคำนวณค่าน้ำหนักด้วยเทคนิคต่าง ๆ จากนั้นเลือกค่าน้ำหนักของคำซึ่งคำที่เลือกนั้นจะทำหน้าที่เป็นคุณลักษณะ แล้วกำหนดค่าน้ำหนักมาให้กับคำต่าง ๆ ในแต่ละเอกสารทดสอบตามคำที่ได้เลือกในข้างต้นในขั้นตอนถัดไป

การคำนวณค่าน้ำหนักของคำในเอกสาร (Term word weighting)

ข้อมูลนำเข้าในขั้นตอนนี้จะเป็นเอกสารที่ขจัดคำที่มีความฟุ่มเฟือย และแปลงคำให้อยู่ในรูปรากศัพท์ โดยในเอกสารนี้จะเรียกแต่ละคำเหล่านั้นว่าเทอม จากนั้นนำมาคำนวณค่าน้ำหนักให้กับแต่ละเทอม เพื่อกำหนดความสำคัญในแต่ละคำ โดยคำที่มีความสำคัญมากจะมีค่าน้ำหนักมาก และจะส่งผลให้ได้กลุ่มคำที่มีประสิทธิภาพที่จะใช้ระบุประเภทของเอกสาร และในทางตรงกันข้ามคำที่มีความสำคัญน้อยก็จะมีค่านำหนักน้อยตามไปด้วย การคำนวณน้ำหนักด้วยเทคนิค 4 เทคนิค ต่อไปนี้ TFIDF, TFICF, IG และ CHI-square ดังจะได้อธิบายแต่ละเทคนิคจากตัวอย่าง

ตัวอย่าง เอกสารกลุ่มที่ 1 (C1) ประกอบด้วย D1 และ D2 เอกสารกลุ่มที่ 2 (C2) ประกอบด้วย D3 และ D4

ตารางที่ 3.1 ตัวอย่างเอกสาร D1, D2, D3 และ D4 ที่ผ่านการตัดคำฟุ่มเฟือยและการแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์

เอกสาร	เนื้อความในเอกสาร
D1	lose invest tough lose fund intend educ height estim percent colleg save account plan suffer loss shed percent skittish burn investor ongo volatil increas competit provid announc product ad featur reduc cost strategi conserv option fidel invest recent announc addit deposit portfolio sold colleg save plan includ manag hampshir california massachusett delawar arizona deposit bear account design preserv princip recent month plan vanguard york reduc expens ratio row price fee colleg educ rise time faster rate inflat save colleg biggest financi challeng famili bob reynold ceo putnam invest introduc design outperform inflat period downturn decad investor todai sensit volatile
D2	america gear holidai american holidai survei american express averag travel month mean typic holidai report base survei adult quarter take longer vacat dine dump fly busi new airlin restaur alik consum remain cautiou spend holidai american plan don similarli consum cite price differ motiv strang wealthi american focus cut vacat spend survei affluent consum annual household incom forego holidai problem friend cost
D3	allergi promis condit tend indoor cope pollen spend surpris health benefit seri studi scientist swap concret confin hour surround forest park place plenti tree increas immun function stress reduct factor scientist chalk phytoncid airborn chemic plant emit rot insect benefit human publish januari includ data healthi japan visit park therapeut popular call shinrin yoku bath instruct wood hour walk trade place scientist plant produc lower concentr cortisol lower puls rate lower pressur thing studi shown visit park forest level

	cell includ walk dai percent spike level killer cell increas cell last expos phytoncid
D4	doctor hunan provinc china report syndrom medicin meet luck club patient medic journal lancet pain swell test normal ultrasound show vein thrombosi danger clot hospit week intraven thinner aspirin wear compress stock recov mainten therapi warfarin thinner health problem thrombosi risk factor birth pill ultim diagnosi doctor mah jongg relat vein thrombosi note spent hour sit motionless previou plai ancient chines tile mah jongg intent consum amount patient economi syndrom underli pathophysiolog mechan mechan mah jongg relat complic stress involv monetari bet depriv

จากเอกสารที่กำหนดทั้ง 4 เอกสารนำมาคำนวณค่าน้ำหนักด้วยวิธีการต่าง ๆ ต่อไปนี้

1. Term Frequency and Inverse Document Frequency (TFIDF) ซึ่งเป็นการคำนวณค่าน้ำหนักของเทอมในแต่ละเอกสาร วิธีคำนวณตามสมการที่ 2.1

$$tfidf(t_i, d_j) = tf(t_i, d_j) \times log \frac{N}{N(t_i)}$$

ตารางที่ 3.2 แสดงตัวอย่างการคำนวณค่าน้ำหนักด้วยวิธี TFIDF จากเอกสาร D1, D2, D3 และ D4

เอกสาร	คำ	ค่าน้ำหนัก
D1	lose	1*log(4/1) = 0.602
D2	consum	3*log(4/2) = 0.903
D4	consum	1*log(4/1) = 0.602
D1	colleg	2*log(4/1)=1.204
D1	financi	1*log(4/1) = 0.602
D2	cost	1*log(4/1) = 0.602
D4	syndrom	2*log(4/1)=1.204
D4	diagnosi	1*log(4/1) = 0.602
D3	cell	3*log(4/1) = 1.806
D1	increas	1*log(4/2) = 0.301

จากตารางที่ 3.2 เป็นตัวอย่างการคำนวณค่าน้ำหนักของคำที่ปรากฏในเอกสารตัวอย่าง (D1, D2, D3, D4) สามารถอธิบายได้ว่า คำว่า "consum" ซึ่งปรากฏในเอกสาร D2 และ D4 โดยที่ในเอกสาร D2 มีค่าน้ำหนักมากกว่า D4 นั้นหมายความว่า "consum" มีความสำคัญต่อเอกสาร D2 มากกว่า เช่นเดียวกับคำว่า

"cell" มีค่าน้ำหนักค่อนข้างมาก เนื่องจากปรากฏในเอกสารเดียวและมีความถี่มาก ทำให้คำนี้มีแนวโน้มจะเป็นคำที่สำคัญที่จะใช้ในการระบุประเภทเอกสาร

2. Term Frequency Inversed Class Frequency (TFICF) เป็นการคำนวณค่าน้ำหนักของคำในแต่ละกลุ่มเอกสาร วิธีคำนวณตามสมการที่ 2.2

$$tficf(t_i, c_j) = tf(t_i, c_j) \times icf(t)$$

ตารางที่ 3.3 แสดงตัวอย่างการคำนวณค่าหนักด้วยวิธี TFICF จากเอกสาร D1, D2, D3 และ D4

กลุ่มที่	คำ	ค่าน้ำหนัก
C1	lose	1*log(2/1) = 0.3010
	consum	3*log(2/2)=0.0
	colleg	2*log(2/1)=0.602
	financi	1*log(2/1) = 0.3010
	cost	1*log(2/1) = 0.3010
C2	consum	1*log(2/2)=0.0
	syndrom	2*log(2/1)=0.602
	diagnosi	1*log(2/1) = 0.3010
	cell	3*log(2/1)=0.903
	increas	1*log(2/2)=0.0

จากตารางที่ 3.3 พบว่าเมื่อคำใดที่ปรากฏในเอกสารทุกประเภท จะมีค่าน้ำหนักเท่ากับศูนย์แล้วคำนั้นจะถือว่าไม่สามารถเป็นตัวแทนของเอกสารได้ เช่น ในประเภทเอกสาร C1 คำว่า "consum" และในประเภทเอกสาร C2 คำว่า "consum" และ "increase" ไม่สามารถกำหนดเป็นตัวของแต่ละประเภทเอกสารได้ ในขณะที่ในประเภทเอกสาร C1 คำว่า "colleg" จะมีค่าน้ำหนักมากกว่าค่าอื่น และในประเภทเอกสาร C2 คำว่า "cell " และ "syndrome" มีค่าน้ำหนักมากกว่าค่าอื่น ๆ ดังนั้นคำเหล่านี้จึงมีแนวโน้มที่จะใช้เป็นคำที่สามารถระบุประเภทเอกสารได้

3. Information Gain (IG) เป็นการคำนวณค่าน้ำหนักของคำในแต่ละกลุ่มเอกสาร โดยเป็นการวัดจำนวนบิตของข้อมูล เพื่อใช้ในการจัดกลุ่มเอกสารโดยใช้การปรากฏของคำที่มีในเอกสาร วิธีคำนวณตามสมการที่ 2.8

$$IG = -\frac{A+C}{N} log \frac{A+C}{N} + \frac{A}{N} log\left(\frac{A}{A+B}\right) + \frac{C}{N} log\left(\frac{C}{C+D}\right)$$

กำหนดให้ A คือจำนวนเอกสารที่ปรากฏเทอม t และอยู่ในกลุ่ม C

B คือจำนวนเอกสารที่ปรากฏเทอม t และไม่อยู่ในกลุ่ม C

C คือจำนวนเอกสารที่ไม่ปรากฏเทอม t และอยู่ในกลุ่ม C

D คือจำนวนเอกสารที่ไม่ปรากฏเทอม t และอยู่ไม่ในกลุ่ม C

ตารางที่ 3.4 แสดงตัวอย่างการคำนวณค่าหนักด้วยวิธี IG จากเอกสาร D1, D2, D3 และ D4

กลุ่มที่	คำ	ค่าน้ำหนัก
C1	lose	0.27
	consum	0.31
	colleg	0.27
	financi	0.27
	cost	0.27
C2	consum	0.31
	syndrom	0.27
	diagnosi	0.27
	cell	0.27
	increas	0.31

จากตารางที่ 3.4 ค่าน้ำหนักของคำในแต่ละประเภทเอกสาร พบว่า "consum" ที่ปรากฏในเอกสารทั้งสองประเภทจะมีค่าสูงกว่าคำที่ปรากฏในเอกสารประเภทใดประเภทหนึ่ง ดังนั้นการคำนวณด้วยวิธีนี้จะให้ค่าน้ำหนักที่แตกต่างจากวิธี TFIDF และ CHI

4. Chi Square (CHI) เป็นการคำนวณค่าน้ำหนักของคำในแต่ละกลุ่มเอกสาร เป็นการคำนวณหาความสัมพันธ์ระหว่างคำกับกลุ่มเอกสาร วิธีคำนวณตามสมการที่ 2.6

$$\chi^2 = \frac{N(AD - BC)^2}{(A + B)(C + D)(A + C)(B + D)}$$

ตารางที่ 3.5 แสดงตัวอย่างการคำนวณค่าหนักด้วยวิธี CHI จากเอกสาร D1, D2, D3 และ D4

กลุ่มที่	คำ	ค่าน้ำหนัก
C1	lose	1.33
	consum	0
	colleg	1.33
	financi	1.33
	cost	1.33
C2	consum	0
	syndrom	1.33
	diagnosi	1.33
	cell	1.33
	increas	0

จากตารางที่ 3.5 คำที่ปรากฏในเอกสารเพียงประเภทเดียวจะมีค่าน้ำหนักมากกว่าคำที่ปรากฏในเอกสารทั้งสองประเภท ซึ่งคำที่พบในเอกสารทั้งสองประเภท ได้แก่ "consum" และ "increase" จะมีค่าน้ำหนักเท่ากับศูนย์ ในขณะที่คำที่ปรากฏในเอกสารประเภทเดียวนั้นจะมีค่าน้ำหนักมากกว่าคำที่ปรากฏทั้งสองประเภทเอกสาร ได้แก่ "lose" "college" "finance" ปรากฏในเอกสารกลุ่ม C1 และ "syndrome" "diagnosi" "cell" ปรากฏในเอกสารกลุ่ม C2

เลือกคำที่ทำหน้าที่เป็นตัวแทนของกลุ่มเอกสารจากค่าน้ำหนักที่คำนวณได้ตามการคำนวณค่าน้ำหนักในการเลือกคำด้วยการคำนวณค่าน้ำหนักด้วยวิธี TFICF, IG และ CHI-square จะกำหนดตาม

(1) จำนวนคำ เลือกคำที่มีน้ำหนักมากที่สุด n คำของแต่ละกลุ่ม แล้วกำหนดให้คำเหล่านั้นเป็นคุณลักษณะ ซึ่งจะได้คำที่เป็นตัวแทนของแต่ละกลุ่มเอกสาร เมื่อจำนวน n มากขึ้นจำนวนคุณลักษณะก็มากขึ้นตามไปด้วย

(2) ค่าเฉลี่ยของน้ำหนักของคำในกลุ่มเอกสาร โดยที่น้ำหนักของเทอมที่ถูกกำหนดให้เป็นคุณลักษณะจะต้องมีค่ามากกว่าค่าเฉลี่ยของคำในกลุ่มเอกสาร ดังนั้นจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะจะถูกกำหนดจากค่าเฉลี่ยของคำในกลุ่มเอกสาร

ในการเลือกคุณลักษณะที่คำนวณค่าน้ำหนักด้วยวิธี TFIDF จะเป็นการคำนวณค่าน้ำหนักของแต่ละเทอมในแต่ละเอกสาร จึงได้นำหลักการของ Zip'f Law เข้ามาช่วยในการเลือกคำ โดยมีหลักการคือ นำความถี่ของคำที่ปรากฏในเอกสารคูณด้วยลำดับของปรากฏของคำ ซึ่งจะได้ค่าคงที่หนึ่ง ซึ่งเป็นตัวแทนของ

เอกสารนั้น ซึ่งแต่ละเอกสารจะมีค่าคงที่เฉพาะตัว ซึ่งการคำนวณวิธีนี้ความถี่ของข้อมูลจะถูกนำมาใช้วัดความสำคัญของคำที่ใช้แทนเอกสาร แสดงได้ดังนี้

ตารางที่ 3.6 แสดงตัวอย่างค่าความคำ ความถี่ ลำดับ และผลคูณของความถี่กับลำดับที่ปรากฏของคำจากเอกสารทดสอบ

Word	Rank	Freq	Rank*Freq
percent	1	4,218	4,218
report	2	3,108	6,218
year	3	2,899	8,697
govern	4	2,340	14,040
storm	5	2,098	18,882
health	6	1,967	19,670
rate	7	1,886	22,632
cancer	8	1,871	24,323
risk	9	1,833	29,328
nation	10	1,726	31,068
...			
adida	6,981	14	97,734
baggag	7,000	14	98,000
servant	9,373	9	84,357
Mmspercept	14,514	4	58,056

การใช้กฎ Zipf's Law จะช่วยลดจำนวนคำที่ไม่สำคัญ โดยคำที่มีความถี่เป็นจำนวนน้อยจะไม่ถูกนำมาจัดในกลุ่มคำที่สำคัญ เนื่องจากคำเหล่านี้ไม่ได้ให้ความสำคัญกับเนื้อความในเอกสาร ดังนั้น เราจึงสร้างจุดที่ใช้สำหรับตัดคำที่ไม่สำคัญออก ซึ่งเป็นคำที่มีความถี่น้อยครั้งในเอกสาร เราจะเรียกจุดนั้นว่าค่า Threshold โดยในการกำหนดค่า Threshold ที่เหมาะสมนั้น จะกำหนดโดยการทำการลองโดยสมมติค่า Threshold เพื่อดูผลลัพธ์แล้วทำการปรับค่า โดยการทำซ้ำ ๆ หลาย ๆ ครั้ง เพื่อหาค่าที่เหมาะสมที่สุด ซึ่งค่า Threshold นั้นจะเป็นการระบุความถี่ขั้นต่ำที่จะนำมาคำนวณหาค่านำหนัก TFIDF

กำหนดค่าน้ำหนักให้กับเอกสารตามคำที่ทำหน้าที่เป็นคุณลักษณะ (Feature Weighting)

เมื่อได้กลุ่มคำหรือเซตของคำที่เป็นคุณลักษณะแล้ว จากนั้นจึงสร้างเวกเตอร์ของเอกสาร โดยมีขนาดเท่ากับขนาดของคุณลักษณะ โดยน้ำหนักที่กำหนดในแต่ละคุณลักษณะนั้นคำนวณจาก

(1) สำหรับการคำนวณค่าน้ำหนักด้วยวิธี TFICF, IG และ CHI นั้นจะเป็นการกำหนดน้ำหนักของเทอมในแต่ละกลุ่มเอกสาร ดังนั้นจึงต้องทำการหาค่าน้ำหนักของเทอม สำหรับในงานวิจัยนี้ได้ใช้ ตามสูตรที่ 2. ดังนั้นจะได้คะแนนของแต่ละเทอมเพื่อใช้ในการจัดกลุ่มเอกสาร

(2) สำหรับการคำนวณค่าน้ำหนักด้วยวิธี TFIDF นั้นเป็นการกำหนดน้ำหนักให้กับคำในแต่ละเอกสารจึงทำขั้นตอนถัดไปได้เลย

จากนั้นจึงทำการปรับค่าน้ำหนักให้มีค่าอยู่ระหว่าง 0 ถึง 1 โดยคำนวณตามสมการ

$$score_{normalize} = \frac{score - minweight}{maxweight - minweight}$$

โดยที่ *minweight* คือ ค่าน้ำหนักที่น้อยที่สุดในกลุ่มของคุณลักษณะ

maxweight คือ ค่าน้ำหนักที่มากที่สุดในกลุ่มของคุณลักษณะ

score คือ ค่าน้ำหนักของแต่ละคุณลักษณะ

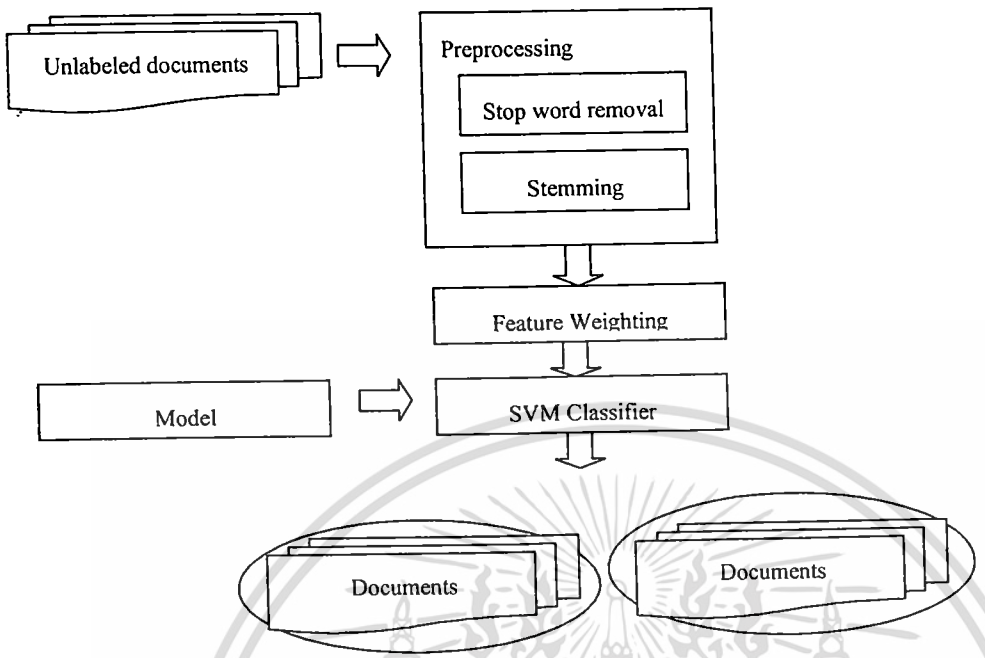
เนื่องจากค่าน้ำหนักของแต่ละคุณลักษณะมีค่าค่อนข้างต่ำ การปรับค่าคะแนนจะช่วยให้ค่าน้ำหนักมีค่าเพิ่มมากขึ้นเพื่อใช้ในการคำนวณได้สะดวกมากขึ้น

เมื่อกำหนดค่าน้ำหนักให้กับแต่ละคุณลักษณะแล้ว แต่ละเอกสารที่ใช้ในการฝึกสอนจะถูกแปลงให้อยู่ในรูปแบบของเวกเตอร์ตามขนาดของคุณลักษณะ โดยเอกสารใดไม่ปรากฏคุณลักษณะนั้นก็จะกำหนดค่าน้ำหนักให้เป็นศูนย์ พร้อมทั้งระบุกลุ่มเอกสารที่ถูกต้องด้วยหรือคำเฉลยของเอกสารว่าอยู่กลุ่มใด

3. การเรียนรู้ (SVM Training) ในขั้นตอนนี้เป็นการเรียนรู้เพื่อสร้างโมเดลในการจัดกลุ่มเอกสารโดยใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน ซึ่งรายละเอียดกล่าวไว้ในบทที่ 2 ข้อมูลนำเข้าไปเรียนรู้ได้แก่ เวกเตอร์ของแต่ละเอกสารพร้อมคำเฉลย และผลลัพธ์ที่ได้จากการเรียนรู้คือโมเดลซึ่งจะนำไปใช้ในการจัดกลุ่มเอกสารกับเอกสารทดสอบต่อไป

ในการเรียนรู้สำหรับงานวิจัยนี้เลือกใช้ Kernel function ที่ชื่อ Radial Basis Function (RBF) เนื่องจากฟังก์ชันนี้เหมาะสมกับข้อมูลที่มีขนาดใหญ่หรือข้อมูลที่มีจำนวนคุณลักษณะเป็นจำนวนมาก และข้อมูลเหล่านั้นมีลักษณะแบบ Nonlinear

3.1.2 ส่วนการทดสอบการจัดกลุ่ม (Classification) เป็นส่วนของการจัดกลุ่มเอกสารโดยใช้โมเดลที่เป็นผลลัพธ์จากเรียนรู้ของเอกสารฝึกสอนข้างต้น การทำงานในขั้นตอนนี้แสดงได้ดังรูปที่ 3.3



รูปที่ 3.3 ขั้นตอนในส่วนของจัดกลุ่ม

จากรูปที่ 3.3 เป็นการแสดงขั้นตอนของการจัดกลุ่มเอกสารที่ไม่มีการระบุประเภทเอกสาร โดยในขั้นแรกเอกสารจะถูกตัดคำฟุ่มเฟือยและแปลงคำศัพท์เหล่านั้นให้อยู่ในรากศัพท์เดิม จากนั้นกำหนดค่าน้ำหนักตามคุณลักษณะที่ได้การจากเรียนรู้ โดยคำที่ไม่ใช่คุณลักษณะก็จะไม่ถูกกำหนดค่าน้ำหนัก และในการกำหนดค่าน้ำหนักนั้นจะทำการสร้างเวกเตอร์ของแต่ละเอกสารตามขนาดของคำที่ทำหน้าที่เป็นคุณลักษณะ แล้วจึงจึงทำการจัดกลุ่มเอกสารด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน และใช้โมเดลที่ได้จากการเรียนรู้ข้างต้นในการจัดกลุ่มเอกสาร

บทที่ 4

การทดลองและผลการทดลอง

ในบทนี้จะกล่าวถึงการออกแบบการทดลองการกำหนดหัวข้อข่าวให้กับเอกสาร โดยการใช้เทคนิคการจัดกลุ่มเอกสารเพื่อเปรียบเทียบประสิทธิภาพ โดยคำนึงถึงปัจจัยต่าง ๆ ที่เกี่ยวข้องกับการจัดกลุ่มเอกสาร ได้แก่ จำนวนเอกสาร จำนวนคุณลักษณะ และค่า Threshold ที่ใช้ในการเลือกคำที่ทำหน้าที่เป็นคุณลักษณะ และวัดประสิทธิภาพการจัดกลุ่มโดยใช้ค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) และ ค่าเอฟ (F-measure)

4.1 ข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่ใช้ในงานวิจัยนี้รวบรวมมาจากเว็บไซต์ข่าวภาษาอังกฤษ ได้แก่ ได้แก่ Yahoo (http://news.yahoo.com), Accuweather (http://www.accuweather.com), New York Time (http://www.nytimes.com), CNN (http://www.cnn.com), News Week (http://www.newsweek.com) เป็นต้น ในช่วงเดือน ธันวาคม 2553-มีนาคม 2554 แบ่งออกเป็น 6 ประเภทข่าว ได้แก่ การเมือง พยากรณ์อากาศ สุขภาพ กีฬา บันเทิง และข่าวธุรกิจ ประเภทข่าวละ 1,000 เอกสาร รวมทั้งหมดเป็น 6,000 เอกสาร ใช้เป็นข้อมูลที่ใช้ในฝึกสอนและทดสอบ โดยจะเลือกเฉพาะข้อความข่าวเท่านั้น

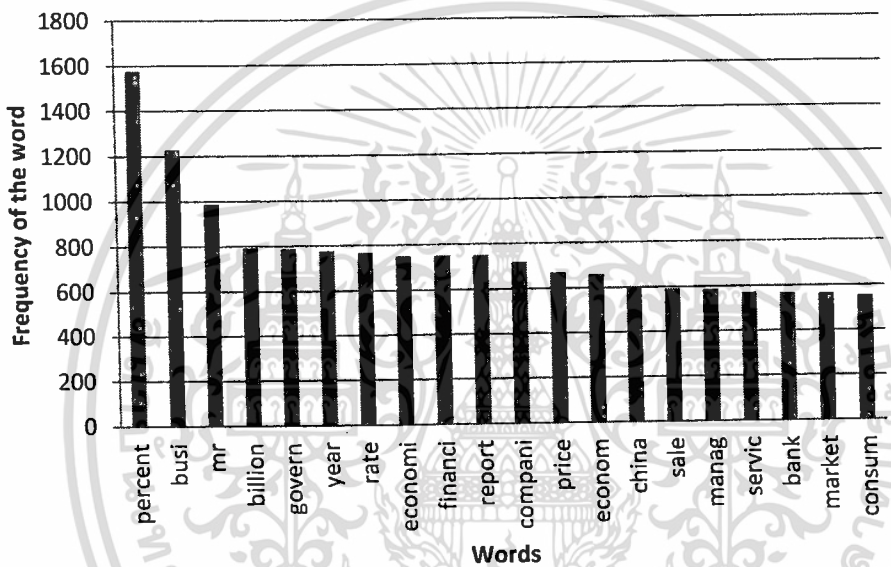
ตารางที่ 4.1 แสดงจำนวนคำในเอกสารแต่ละประเภทที่ตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ และจำนวนคำไม่ซ้ำ พบว่าเอกสารข่าวที่รวบรวมนั้น ข่าวธุรกิจมีจำนวนคำทั้งหมดมากที่สุด ดังนั้นข่าวประเภทนี้ผู้เขียนข่าวจะมีการอธิบายรายละเอียดและนำเสนอข้อมูลที่มากกว่าข่าวประเภทอื่น ข่าวการเมือง สุขภาพ กีฬา พยากรณ์อากาศ และบันเทิงตามลำดับ และข่าวประเภทการเมืองมีการใช้คำซ้ำ ๆ เป็นจำนวนมากกว่าข่าวประเภทอื่น ๆ และข่าวบันเทิงมีการใช้คำที่หลากหลายมากกว่าข่าวประเภทอื่น

ตารางที่ 4.1 แสดงจำนวนคำทั้งหมดและจำนวนคำที่ไม่ซ้ำในแต่ละประเภทเอกสาร

ประเภทเอกสาร	จำนวนคำที่ไม่ซ้ำ	จำนวนคำทั้งหมด	ลดลง (เท่า)
การเมือง	11,218	179,478	16
พยากรณ์อากาศ	9,259	118,882	13
สุขภาพ	12,068	160,426	13
กีฬา	11,494	119,357	10
บันเทิง	12,574	104,479	8
ธุรกิจ	15,178	206,823	14
รวม	71,791	889,445	

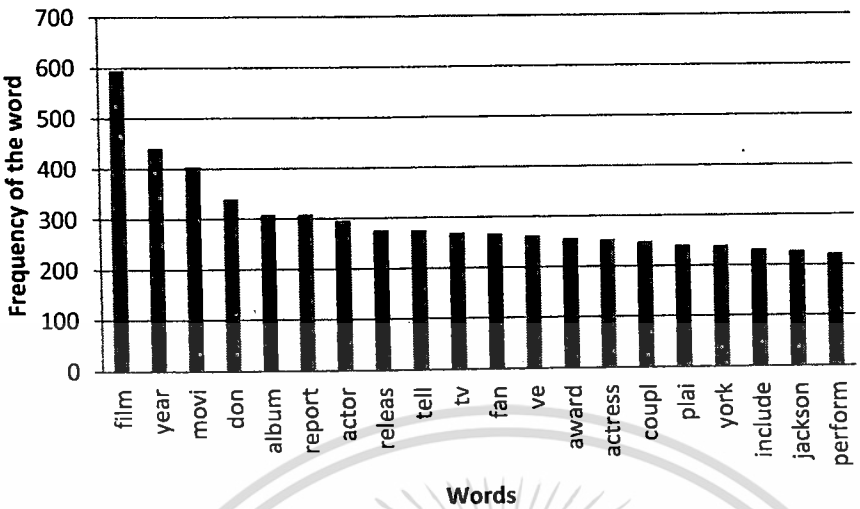
การกระจายของข้อมูลในประเภทข่าวต่าง ๆ หลังจากตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ คำทั้งหมดในเอกสารที่ไม่ซ้ำกัน โดยคำเหล่านั้นเป็นคำที่ไม่ซ้ำกันในแต่ละประเภทเอกสาร มีทั้งหมด 71,791 คำ แต่คำเหล่านั้นอาจเป็นคำที่ซ้ำกันกับคำในเอกสารประเภทอื่น ดังนั้นจึงตัดคำซ้ำออกไปจึงเหลือคำทั้งหมดที่ไม่ซ้ำในเอกสารทั้งหมด 6 ประเภท ซึ่งมีทั้งหมด 33,700 คำ

แต่ละประเภทเอกสารสามารถแสดงการกระจายของคำได้ดังรูปที่ 4.1, 4.2, 4.3, 4.4, 4.5 และ 4.6 โดยแกน y แสดงความถี่ของคำ และแกน x แสดงคำที่ปรากฏในแต่ละประเภทเอกสาร 20 คำแรกที่มีความถี่สูงสุด



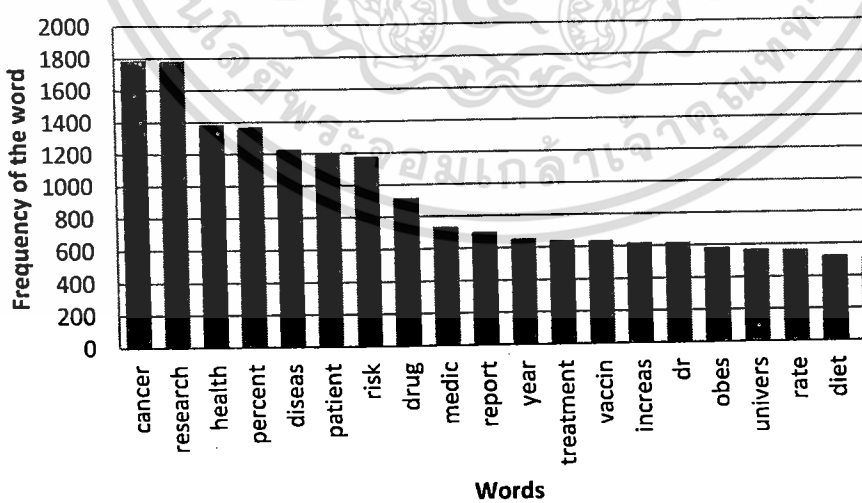
รูปที่ 4.1 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวธุรกิจ

จากรูปที่ 4.1 ในข่าวธุรกิจ 10 คำแรกที่พบมากที่สุดได้แก่ "percent" มีความถี่เท่ากับ 1,577 ครั้ง "busi" มีความถี่เท่ากับ 1,229 ครั้ง "mr" มีความถี่เท่ากับ 987 ครั้ง "billion" มีความถี่เท่ากับ 793 ครั้ง "govern" มีความถี่เท่ากับ 788 ครั้ง "year" มีความถี่เท่ากับ 777 ครั้ง "rate" มีความถี่เท่ากับ 769 ครั้ง "economi" มีความถี่เท่ากับ 752 ครั้ง "financi" มีความถี่เท่ากับ 752 ครั้ง และ "report" มีความถี่เท่ากับ 752 ครั้ง ซึ่งจำนวนคำที่มีความถี่เท่ากับ 1 คือคำนี้ปรากฏเพียงหนึ่งครั้งในเอกสารประเภทนี้ มีจำนวนคำเท่ากับ 5,481 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 คือ คำที่ปรากฏสองครั้งในเอกสารประเภทนี้ โดยมีจำนวนคำเท่ากับ 2,359 คำ คิดเป็นร้อยละ 36 และ 15 ตามลำดับ



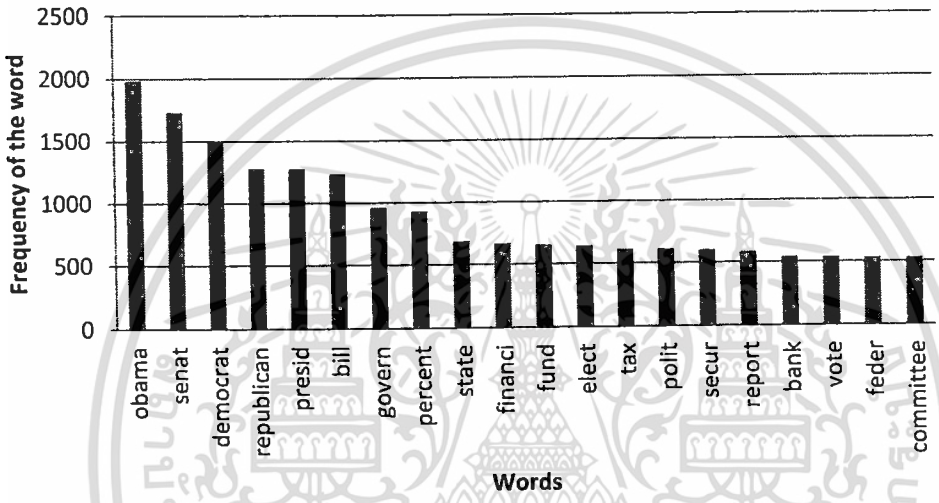
รูปที่ 4.2 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวบันเทิง

จากรูปที่ 4.2 ในข่าวประเภทบันเทิง 10 คำแรกที่พบมากที่สุดได้แก่ "film" มีความถี่เท่ากับ 595 ครั้ง "year" มีความถี่เท่ากับ 441 ครั้ง "movi" มีความถี่เท่ากับ 405 ครั้ง "don" มีความถี่เท่ากับ 340 ครั้ง "album" มีความถี่เท่ากับ 309 ครั้ง "report" มีความถี่เท่ากับ 309 ครั้ง "actor" มีความถี่เท่ากับ 296 ครั้ง "releas" มีความถี่เท่ากับ 277 ครั้ง "tell" มีความถี่เท่ากับ 276 ครั้ง และ "tv" มีความถี่เท่ากับ 270 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,742 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,966 คำ คิดเป็นร้อยละ 37 และ 15 ตามลำดับ



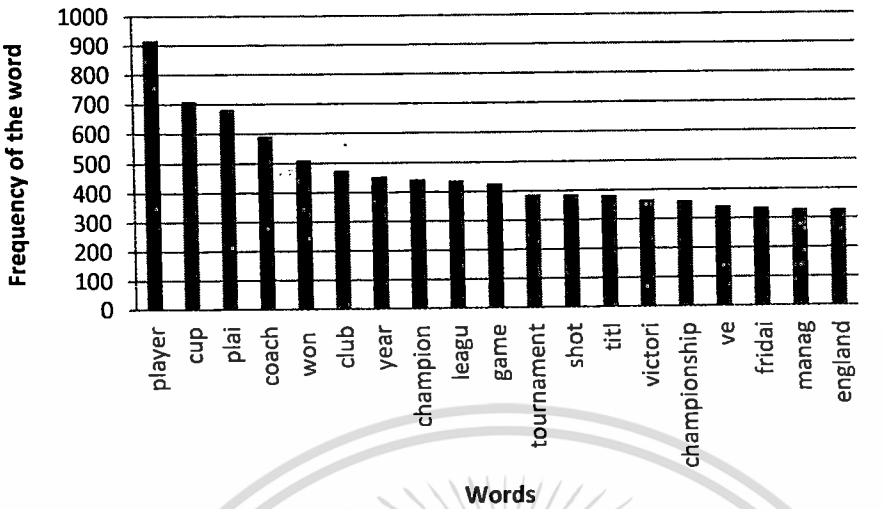
รูปที่ 4.3 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวสุขภาพ

จากรูปที่ 4.3 ในข่าวประเภทข่าวสุขภาพ 10 คำแรกที่พบมากที่สุดได้แก่ "cancer" มีความถี่เท่ากับ 1,785 ครั้ง "research" มีความถี่เท่ากับ 1,784 ครั้ง "health" มีความถี่เท่ากับ 1,388 ครั้ง "percent" มีความถี่เท่ากับ 1,369 ครั้ง "diseas" มีความถี่เท่ากับ 1,226 ครั้ง "patient" มีความถี่เท่ากับ 1,192 ครั้ง "risk" มีความถี่เท่ากับ 1,177 ครั้ง "drug" มีความถี่เท่ากับ 917 ครั้ง "medic" มีความถี่เท่ากับ 733 ครั้ง และ "report" มีความถี่เท่ากับ 698 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,160 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,893 คำ คิดเป็นร้อยละ 34 และ 15 ตามลำดับ



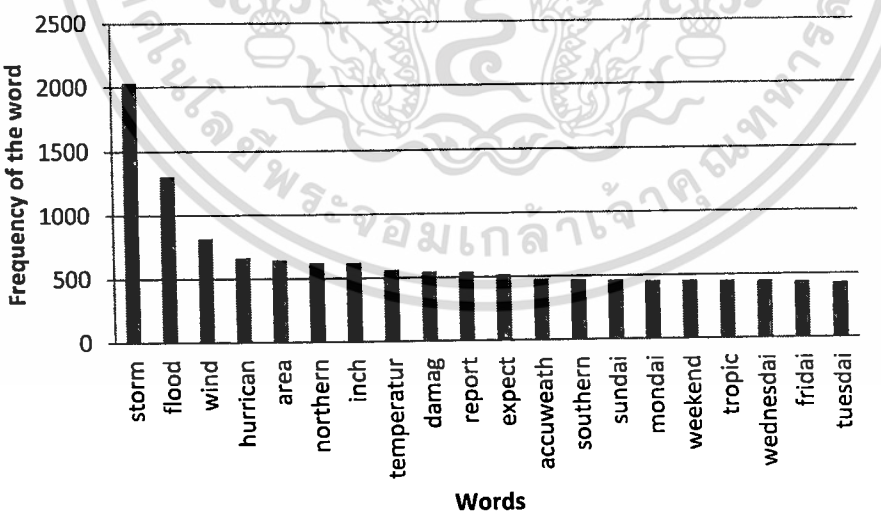
รูปที่ 4.4 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวการเมือง

จากรูปที่ 4.4 ในข่าวประเภทการเมือง 10 คำแรกที่พบมากที่สุดได้แก่ "obama" มีความถี่เท่ากับ 1,987 ครั้ง "senat" มีความถี่เท่ากับ 1,731 ครั้ง "democrat" มีความถี่เท่ากับ 1,489 ครั้ง "republican" มีความถี่เท่ากับ 1,281 ครั้ง "presid" มีความถี่เท่ากับ 1,278 ครั้ง "bill" มีความถี่เท่ากับ 1,235 ครั้ง "govern" มีความถี่เท่ากับ 961 ครั้ง "percent" มีความถี่เท่ากับ 931 ครั้ง "state" มีความถี่เท่ากับ 686 ครั้ง และ "financi" มีความถี่เท่ากับ 663 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 3,635 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,623 คำ คิดเป็นร้อยละ 32 และ 14 ตามลำดับ



รูปที่ 4.5 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทข่าวกีฬา

จากรูปที่ 4.5 ในข่าวประเภทกีฬา 10 คำแรกที่พบมากที่สุดได้แก่ "cup" มีความถี่เท่ากับ 709 ครั้ง "plai" มีความถี่เท่ากับ 682 ครั้ง "coach" มีความถี่เท่ากับ 592 ครั้ง "won" มีความถี่เท่ากับ 510 ครั้ง "club" มีความถี่เท่ากับ 475 ครั้ง "year" มีความถี่เท่ากับ 453 ครั้ง "champion" มีความถี่เท่ากับ 444 ครั้ง "leagu" มีความถี่เท่ากับ 438 ครั้ง และ "game" มีความถี่เท่ากับ 428 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 4,018 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,675 คำ คิดเป็นร้อยละ 35 และ 14 ตามลำดับ



รูปที่ 4.6 การกระจายของคำ 20 คำแรกที่มีความถี่สูงสุดในเอกสารประเภทพยากรณ์อากาศ

จากรูปที่ 4.6 ในข่าวประเภทพยากรณ์อากาศ 10 คำแรกที่พบมากที่สุดได้แก่ "storm" มีความถี่เท่ากับ 2,034 ครั้ง "flood" มีความถี่เท่ากับ 1,306 ครั้ง "wind" มีความถี่เท่ากับ 817 ครั้ง "hurricane" มีความถี่

เท่ากับ 664 ครั้ง “area” มีความถี่เท่ากับ 648 ครั้ง “northern” มีความถี่เท่ากับ 622 ครั้ง “inch” มีความถี่เท่ากับ 620 ครั้ง “temperatur” มีความถี่เท่ากับ 562 ครั้ง “damag” มีความถี่เท่ากับ 546 ครั้ง และ “report” มีความถี่เท่ากับ 537 ครั้ง โดยจำนวนคำที่มีความถี่เท่ากับ 1 มี 3,260 คำ และจำนวนคำที่มีความถี่เท่ากับ 2 มี 1,496 คำ คิดเป็นร้อยละ 35 และ 16 ตามลำดับ

จากรูปทั้ง 6 รูป (รูปที่ 4.1-4.6) ข้างต้นพบว่าคำที่มีความถี่น้อยหรือคำที่ปรากฏในเอกสารน้อยครั้งนั้นมีเป็นจำนวนมากหรือมีประมาณเกือบร้อยละ 40 ของคำที่ไม่ซ้ำในเอกสารที่ปรากฏหนึ่งครั้งและประมาณร้อยละ 15 ของคำที่ไม่ซ้ำในเอกสารที่ปรากฏสองครั้ง ในทางกลับกันคำที่มีความถี่มากมักจะเป็นมีเพียงคำเดียวดังอธิบายข้างต้น ทำให้เราสามารถกล่าวอ้างได้ว่าคำที่มีความถี่ต่ำมาก ๆ จะเป็นคำที่น่าจะมีความสำคัญน้อย และในทำนองเดียวกันคำที่มีความถี่มาก ๆ ในประเภทเอกสารหลาย ๆ ประเภทนั้นน่าจะเป็นคำที่มีความสำคัญน้อยด้วยเช่นกัน จากกราฟพบว่า “report” ปรากฏในประเภทเอกสารสภาพอากาศ สุขภาพ ธุรกิจและบันเทิง ซึ่งมีความถี่สูงในประเภทเอกสารเหล่านี้ นั้นแสดงว่า “report” ไม่ควรจะถูกกำหนดให้เป็นคำสำคัญสำหรับใช้ระบุประเภทเอกสาร

จากแต่ละประเภทเอกสาร ได้ทำการเลือกตัวอย่างคำที่ปรากฏมากที่สุดในแต่ละประเภทเอกสาร 40 คำแรก ซึ่งจะพบว่าแต่ละคำนั้นมีความหมายที่เกี่ยวข้องกับแต่ละประเภทเอกสาร แสดงดังตารางที่ 4.2

ตารางที่ 4.2 แสดงตัวอย่างคำที่ปรากฏมากที่สุดในแต่ละประเภทเอกสาร 40 คำแรก

ประเภทเอกสาร	คำที่ปรากฏมากที่สุด 40 คำแรก
การเมือง	obama senat democrat republican presid bill govern percent state financi fund elect tax polit secur report bank vote feder committe unit offici issu american administr congress leader afghanistan includ call propos court regul lawmak legisl reform washington year nation rate
พยากรณ์อากาศ	storm flood wind hurrican area northern inch temperatur damag report expect accuweath southern sundai mondai weekend tropic wednesdai fridai tuesdai offici condit forecast saturdai includ atlant central thursdai mile dai part nation todai eastern thunderstorm meteorologist home western airport road
สุขภาพ	cancer research health percent diseas patient risk drug medic report year treatment vaccin increas dr obes univers rate diet find american studi develop brain includ prevent effect result test state breast clinic nation hospit cell medicin gene call higher journal
กีฬา	player cup plai coach won club year champion leagu game tournament shot titl goal victori championship ve fridai manag england jame score in win saturdai start final germani don didn footbal return ad sign point contract team run

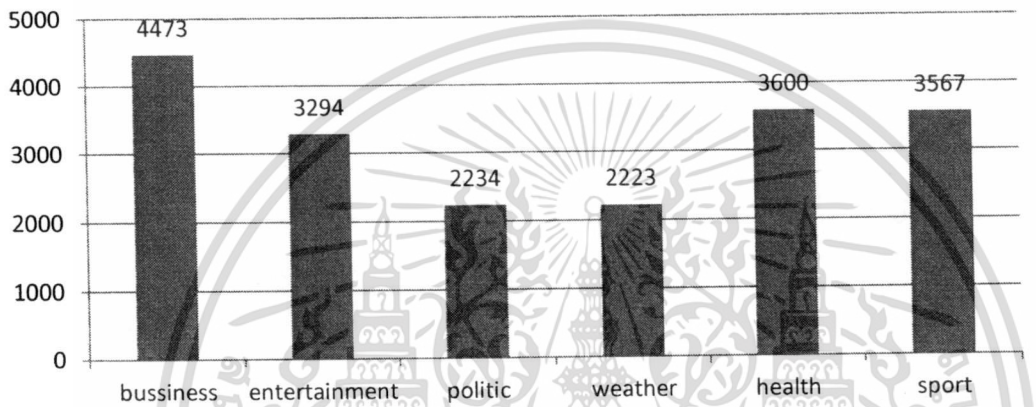
	wimbledon nada
บันเทิง	film year movi don album report actor releas tell tv fan ve award actress coupl plai york includ jackson perform star work recent singer michael show hollywood june kid video call celebr make seri didn angel lo dai thing tour
ธุรกิจ	percent busi mr billion govern year rate economi financi report compani price econom china sale manag servic bank market consum invest quarter growth tax increas american execut expect includ credit presid stock don firm make recent investor product site global

แต่อย่างไรก็ตามจากกลุ่มคำในตารางที่ 4.2 จะพบว่า 'includ', 'report' และ 'year' ปรากฏทั้ง 5 กลุ่มเอกสาร ขณะที่ 'call', 'don', 'percent' และ 'rate' ปรากฏใน 3 กลุ่มเอกสาร และ 'dai', 'didn', 'expect', 'financi', 'fridai', 'govern', 'increas', 'make', 'mange', 'offici', 'plai' เป็นต้น ซึ่งปรากฏใน 2 กลุ่มเอกสาร ดังนั้นจึงพบว่าการใช้คำเพียงอย่างเดียวในการระบุกลุ่มเอกสารนั้นไม่เพียงพอ จึงต้องใช้อัลกอริทึมการจัดกลุ่มเอกสารมาช่วยการทำนายว่าแต่ละเอกสารที่ประกอบด้วยกลุ่มคำเหล่านี้ควรจะจัดอยู่ในประเภทเอกสารใดจึงจะเหมาะสม

สำหรับคำที่ปรากฏในเอกสารทั้ง 6 ประเภทเอกสาร มีทั้งหมด 2,964 คำ แต่ละคำก็จะมีจำนวนความถี่ในแต่ละประเภทเอกสารที่แตกต่างกันโดยจะยกตัวอย่างบางคำที่ปรากฏต่อไปนี้

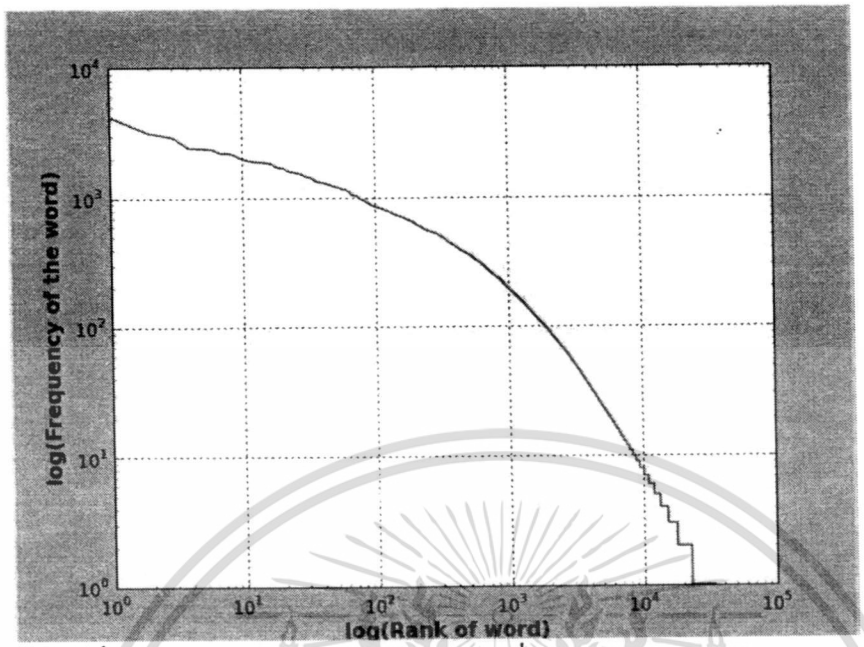
'bear', 'foul', 'protest', 'sleep', 'upsid', 'climb', 'hate', 'request', 'accus', 'accur', 'swai', 'edward', 'pride', 'worth', 'digit', 'risk', 'rise', 'voic', 'tenni', 'loom', 'jack', 'govern', 'affect', 'vast', 'disturb', 'wooden', 'ignit', 'huddl', 'correct', 'wednesdai', 'miller', 'direct', 'histor', 'enjoi', 'consequ', 'second', 'street', 'supervis', 'hide', 'wreck', 'neg', 'calcul', 'asia', 'spokesman', 'toll', 'new', 'net', 'succumb', 'liberti', 'specialist', 'elimin', 'hero', 'avert', 'carv', 'lodg', 'met', 'voter', 'china', 'aftermath', 'enrol', 'interpret', 'incom', 'deterior', 'forum', 'militari', 'anymor', 'loos', 'precis', 'jame', 'smoke', 'permit', 'studi', 'controversi', 'counti', 'golden', 'volunt', 'carl', 'campaign', 'newspap', 'julia', 'mitchel', 'thrust', 'attitud', 'moral', 'total', 'unit', 'highli', 'plot', 'describ', 'prescript', 'overshadow', 'insult', 'concret', 'call', 'telegraph', 'recommend', 'strike', 'indiana', 'type', 'tell', 'relax', 'relat', 'award', 'hurt', 'warn', 'phone', 'connecticut', 'exce', 'adult', 'wari', 'midst', 'hold', 'shoot', 'accid', 'join', 'room', 'henri', 'work', 'wors', 'era', 'ms', 'mr', 'root', 'advocaci', 'shook', 'climat', 'give', 'household', 'dolphin', 'india', 'indic', 'caution', 'refus', 'want', 'basebal', 'david', 'attract', 'vanish', 'end', 'quot', 'polic', 'travel', 'faulti', 'ceremoni', 'recoveri', 'answer', 'gate', 'negoti', 'perspect', 'confid', 'grown', 'recogn', 'lai', 'mess', 'chines', 'lag', 'lab', 'badli', 'modest', 'beauti', 'law', 'demonstr', 'domin', 'third', 'amid', 'grant', 'greet', 'think', 'perform', 'dispar', 'environ', 'reloc', 'enter', 'exclus', 'worst', 'order', 'wind', 'oper', 'offici', ...

ในขณะเดียวกันแต่ละประเภทเอกสารก็จะมีคำที่ไม่ซ้ำกันเลย แสดงดังรูปที่ 4.7 โดยแกน x คือ ประเภทเอกสาร และแกน y คือจำนวนคำที่ไม่ซ้ำ พบว่ากลุ่มเอกสารประเภทพยากรณ์อากาศ (weather) จะมีคำที่ไม่ซ้ำน้อยกว่าข่าวประเภทอื่น และข่าวประเภทธุรกิจ (business) จะมีคำที่ไม่ซ้ำมากกว่าข่าวประเภทอื่น ดังนั้นจากกราฟจึงอธิบายได้ว่าข่าวประเภทธุรกิจจะมีคำที่มีลักษณะเฉพาะสำหรับข่าวประเภทนั้นค่อนข้างมากกว่าข่าวประเภทอื่น ๆ รองลงมาได้แก่ ข่าวสุขภาพ (health) กีฬา (sport) บันเทิง (entertainment) การเมือง (politic) และ พยากรณ์อากาศ (weather) ตามลำดับ

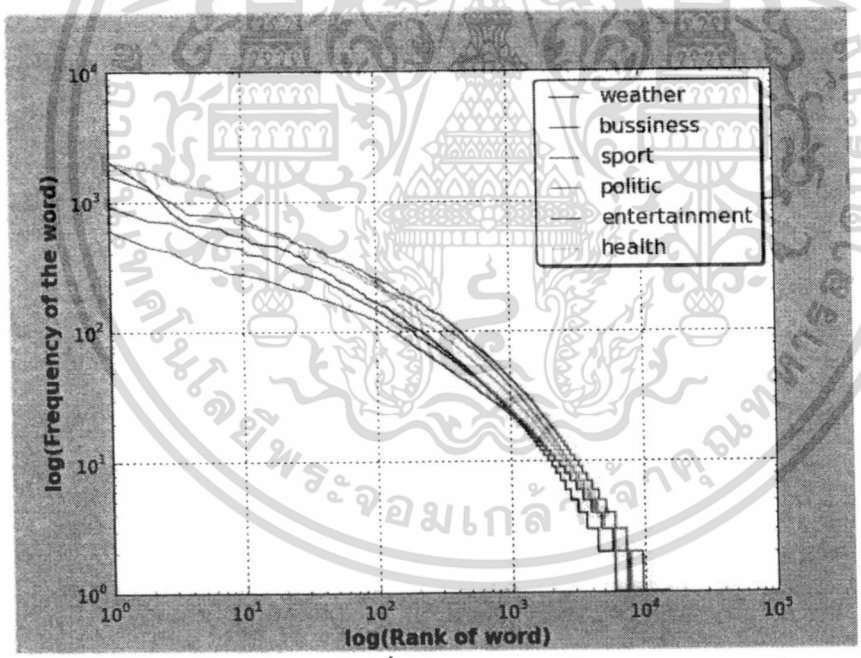


รูปที่ 4.7 แสดงจำนวนคำที่ไม่ซ้ำในแต่ละประเภทเอกสาร

การกระจายของความถี่ของคำในกลุ่มเอกสารตามกฎของ Zipf นั้นกล่าวว่าความถี่ของคำจะแปรผกผันกับลำดับของคำ ซึ่งความถี่ของคำสามารถถูกใช้เพื่อนำมาวัดความสำคัญของคำที่ใช้แทนเอกสารหนึ่งได้ โดยให้ f เป็นความถี่ของคำที่ปรากฏในกลุ่มเอกสาร และ r เป็นลำดับความสำคัญของคำนั้น ๆ ความสัมพันธ์ของ f และ r จะกล่าวได้ว่า เมื่อค่า f สูงจะส่งผลให้ r มีค่าต่ำ ดังที่กล่าวข้างต้น และนำความสัมพันธ์ระหว่าง f และ r มาสามารถแสดงได้ดังกราฟในรูปที่ 4.8 และ 4.9 เมื่อใช้ข้อมูลจากเอกสารที่รวบรวมมาได้



รูปที่ 4.8 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสาร



รูปที่ 4.9 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารแยกตามประเภทเอกสาร

รูปที่ 4.8 แสดงแสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารทั้งหมด และรูปที่ 4.9 แสดงความสัมพันธ์ระหว่างความถี่และลำดับของกลุ่มเอกสารแยกตามประเภทเอกสาร จากรูปทั้งสองนี้คำที่มีความถี่มากจะอยู่ในลำดับที่ต่ำ และกราฟค่อย ๆ ลาดลงโดยในช่วงกลางของกราฟนั้นจะมีลักษณะโค้งค่อนข้างมากและส่วนปลายเส้นกราฟจะมีลักษณะการซ้ำกันของความถี่ค่อนข้างชัดเจน นั้นแสดงว่าคำศัพท์ในส่วนลำดับท้าย ๆ จะมีความถี่ต่ำมาก ๆ และมีจำนวนคำศัพท์เป็นในระดับความถี่นี้เป็นจำนวนมาก

4.2 การออกแบบการทดลองและผลการทดลอง

งานวิจัยนี้ใช้วิธีการทดสอบแบบการตรวจสอบแบบไขว้ (Cross-Validation) โดยแบ่งข้อมูลออกเป็น 5 กลุ่ม แล้วทำการใช้ 1 กลุ่มมาเป็นข้อมูลทดสอบ (Testing set) ส่วนที่เหลือจำนวน 4 กลุ่มจะเป็นข้อมูลฝึกสอน (Training set) แล้วทำการวน 5 ครั้ง ซึ่งจะเปลี่ยนกลุ่มทดสอบไปเรื่อย ๆ ตามลำดับจนครบข้อมูลทั้งหมด

ในการทดสอบนี้แบ่งกลุ่มข้อมูลออกเป็น 3 ชุดข้อมูลทดสอบ ประกอบด้วยข้อมูลจำนวน 3,000 เอกสาร 4,200 เอกสาร และ 6,000 เอกสาร ซึ่งในแต่ละกลุ่มก็จะแบ่งเอกสารออกเป็นเอกสารที่ใช้ในการฝึกสอนและเอกสารที่ใช้ในการทดสอบ ดังรายละเอียดข้างต้น โดยในการอธิบายนี้จะแทนแต่ละกลุ่มเอกสารดังนี้

ชุดเอกสารทดสอบที่ 1 ได้แก่ เอกสารจำนวน 3,000 เอกสาร

ชุดเอกสารทดสอบที่ 2 ได้แก่ เอกสารจำนวน 4,200 เอกสาร

ชุดเอกสารทดสอบที่ 3 ได้แก่ เอกสารจำนวน 6,000 เอกสาร

การออกแบบการทดลองในงานวิจัยนี้ ได้ตั้งสมมติฐานเกี่ยวกับการกำหนดหัวข้อข่าวกับปัจจัยต่าง ๆ ที่มีผลกระทบต่อการจัดกลุ่มได้ดังนี้

1. ค่า Threshold ที่ใช้ในการเลือกคำที่กำหนดที่มีความถี่ต่ำสุดมาเป็นคุณลักษณะ ด้วยการคำนวณวิธี TFIDF มีผลอย่างไร
2. จำนวนคำที่กำหนดเป็นคุณลักษณะมีผลต่อการกำหนดหัวข้อข่าวหรือไม่ (Effect of feature size on performance)
3. จำนวนเอกสารที่ใช้ในการฝึกสอนมีผลต่อการกำหนดหัวข้อข่าวหรือไม่ (Effect of training set size on Topic identification)
4. เปรียบเทียบวิธีการคำนวณค่าน้ำหนักระหว่าง TFIDF, TFICF, IG และ CHI (Performance comparison between four approaches) ให้กับกลุ่มคำสำคัญ

ในการวัดประสิทธิภาพของแต่ละสมมติฐานนั้นจะใช้ค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ (รายละเอียดอธิบายในบทที่ 2) ซึ่งค่าเหล่านี้จะคำนวณโดย นำเอกสารในกลุ่มฝึกสอนมาสร้างโมเดล และทดสอบโมเดลนี้โดยใช้เอกสารในกลุ่มทดสอบเพื่อวัดประสิทธิภาพในค่าต่าง ๆ

ดังนั้นในการออกแบบการทดลองในงานวิจัยนี้จึงต้องสามารถตอบคำถามข้อมูลข้างต้นได้ ซึ่งสามารถกำหนดได้เป็น

การทดลองที่ 1: ค่า Threshold ที่ใช้กำหนดความถี่ต่ำสุดในการเลือกคำที่กำหนดเป็นคุณลักษณะด้วยการคำนวณวิธี TFIDF มีผลอย่างไร

ในการทดลองนี้เป็นการเลือก Threshold ที่เหมาะสม ได้ใช้กฎของ Zipf สร้างค่า Threshold เพื่อกำหนดจุดที่เหมาะสมในการเลือกกลุ่มคำที่ไม่สำคัญหรือคำที่ปรากฏค่อนข้างน้อยครั้งในเอกสาร โดยถ้าค่า

น้ำหนักน้อยกว่าค่า Threshold จะถือว่าเป็นกลุ่มคำที่มีความสำคัญ ซึ่งในการเลือกค่า Threshold จะอ้างอิงที่ค่าความถี่ของการปรากฏของคำในกลุ่มเอกสาร

ตารางที่ 4.3 แสดงจำนวนคำที่กำหนดคุณลักษณะของกลุ่มเอกสารทดสอบเมื่อกำหนดค่า Threshold = 3, 4, 5 และ 6

ชุดเอกสารทดสอบ	จำนวนคำที่กำหนดคุณลักษณะ			
	Threshold=3	Threshold=4	Threshold=5	Threshold=6
1	10,997	9,565	8,521	7,752
2	11,828	10,337	9,168	8,294
3	14,349	12,615	11,284	10,310

จากตารางที่ 4.3 เป็นการแสดงจำนวนคำที่กำหนดเป็นคุณลักษณะของเอกสารกลุ่มทดสอบแยกตามค่า Threshold ที่กำหนด เมื่อ กำหนดให้ค่า Threshold มีค่าเพิ่มขึ้นจำนวนคำก็จะมีจำนวนลดลงอยู่ระหว่าง 9-13 %

เมื่อกำหนดจำนวนคำที่เป็นคุณลักษณะแต่ละ Threshold จึงทำการเรียนรู้จากเอกสารฝึกสอนเพื่อสร้างโมเดลและทดสอบโมเดลจากเอกสารทดสอบในแต่ละชุดเอกสารทดสอบ โดยใช้อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน ตามวิธีการที่กล่าวข้างต้น และทำในทุก ๆ ชุดเอกสารทดสอบ เพื่อวัดประสิทธิภาพการระบุประเภทเอกสาร ด้วยตัววัดความเที่ยงตรง ความระลึกและค่าเอฟ ผลการทดลองแสดงดังตารางที่ 4.4

ในการคำนวณเพื่อวัดประสิทธิภาพนั้นคำนวณได้จาก

Category		Expert Judgment	
		True	False
Classifier Judgment	True	TP	FP
	False	FN	TN

ค่าความเที่ยงตรงนั้นคำนวณได้จาก (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง) / (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง + จำนวนเอกสารที่โมเดลทำนายว่าอยู่ในประเภทนี้ แต่ในความเป็นจริงแล้วไม่ใช่)

ค่าความระลึกคำนวณได้จาก (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง) / (จำนวนเอกสารที่โมเดลสามารถทำนายได้ถูกต้อง + จำนวนเอกสารที่โมเดลทำนายว่าเอกสารไม่อยู่ในประเภทนี้ แต่ในความเป็นจริงแล้วใช่)

ค่าเอฟคำนวณได้จาก 2*(ค่าความเที่ยงตรง*ค่าความระลึก)/ (ค่าความเที่ยงตรง+ค่าความระลึก)

จากตัววัดทั้ง 3 ตัววัดนั้น ในการระบุว่าโมเดลที่ใช้ในการระบุประเภทเอกสารดีสามารถระบุเอกสารได้ถูกต้องมากที่สุดจะต้องมีค่าความเที่ยงตรง ค่าความระลึกและค่าเอฟที่สูง

ตารางที่ 4.4 แสดงค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ (เปอร์เซ็นต์) ในแต่ละค่า Threshold และ แต่ละชุดเอกสารทดสอบ

ชุดเอกสารทดสอบ	Threshold=3			Threshold=4			Threshold=5			Threshold=6		
	P	R	F	P	R	F	P	R	F	P	R	F
1	92.99	92.81	92.90	92.94	92.98	92.86	93.07	92.92	93.00	92.90	92.75	92.82
2	92.73	92.57	92.65	92.82	92.67	92.74	92.82	92.69	92.76	92.77	92.64	92.70
3	93.09	92.96	93.03	93.13	93.02	93.08	93.11	93.00	93.05	93.11	93.00	93.06

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก F=ค่าเอฟ

รูปที่ 4.10 แสดงค่าเอฟตามค่า Threshold และกลุ่มเอกสารชุดทดสอบ

จากตารางที่ 4.4 สามารถแสดงในรูปแบบกราฟดังในรูปที่ 4.10 โดยที่ กราฟสีเขียวแสดงชุดเอกสารทดสอบที่ 1 สีน้ำเงินแสดงชุดเอกสารทดสอบที่ 2 และสีแดงแสดงชุดเอกสารทดสอบที่ 3 พบว่าค่า Threshold ที่มีค่าเท่ากับ 5 มีผลให้ประสิทธิภาพดีกว่าค่า Threshold อื่น ๆ ในกลุ่มเอกสารทดสอบชุดที่ 1 และ 2 แต่ค่า Threshold เท่ากับ 4 จะให้ผลดีกว่าค่าอื่น ๆ ในกลุ่มเอกสารทดสอบชุดที่ 3 ดังนั้นในการเปรียบเทียบประสิทธิภาพในการทดลองถัดไปจึงเลือกค่า Threshold เท่ากับ 5 ในวิธีการคำนวณค่าน้ำหนักด้วย TFIDF

ดังนั้นจึงกล่าวได้ว่า เมื่อกำหนดค่า Threshold เท่ากับ 5 แล้วจะทำให้การระบุประเภทเอกสารดีกว่ากำหนดด้วยค่า Threshold อื่นๆ ที่กำหนดในการทดลอง

การทดลองที่ 2: จำนวนคำที่กำหนดเป็นคุณลักษณะและจำนวนเอกสารมีผลต่อการจัดกลุ่มหรือไม่

ในการทดลองนี้เป็นการเปรียบเทียบประสิทธิภาพวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยมีเงื่อนไขคือจำนวนเอกสารที่แตกต่างกัน และจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันในแต่ละวิธีการคำนวณค่าน้ำหนัก โดยคุณลักษณะที่ได้นั้นจะเป็นคำที่ปรากฏในเอกสาร ในการเลือกจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะนั้นจะเป็นไปตามเงื่อนไขที่กำหนด

การทดลองที่ 2.1 เปรียบเทียบประสิทธิภาพของการวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยปราศจากการเลือกคำที่ทำหน้าที่เป็นคุณลักษณะ ซึ่งในการทดลองนี้จะมีจำนวนคำเพื่อใช้ในการจัดกลุ่มเป็นจำนวนมาก แสดงดังตารางที่ 4.5 และตารางที่ 4.6 แสดงผลการทดลองการระบุประเภทเอกสารโดยแสดงค่าความเที่ยงตรงและค่าความระลึก ตามลำดับ และตารางที่ 4.7 แสดงค่าเอฟ

ตารางที่ 4.5 ตารางแสดงผลรวมของจำนวนคำที่ไม่ซ้ำที่กำหนดเป็นคุณลักษณะของแต่ละชุดเอกสารทดสอบ

ชุดเอกสารทดสอบ	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
1	8,521	23,217	25,032	25,098
2	9,168	24,492	26,528	26,615
3	11,284	28,747	31,016	31,478

จากตารางที่ 4.5 แสดงจำนวนคุณลักษณะหรือคำที่ถูกกำหนดให้ทำหน้าที่เป็นตัวแทนของเอกสารในการระบุประเภทเอกสาร โดยจะพบว่าแต่ละชุดเอกสารทดสอบจะมีจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะอยู่ระหว่าง 8,000 – 32,000 คำ ซึ่งในการคำนวณค่าน้ำหนักด้วยวิธี CHI จะมีคำที่ถูกกำหนดเป็นคุณลักษณะมากที่สุด แต่อย่างไรก็ตามจะมีจำนวนคำที่ไม่แตกต่างจากวิธี IG มากนัก ส่วนการคำนวณค่าน้ำหนักด้วยวิธี TFICF นั้นก็จะมีจำนวนน้อยกว่ากลุ่มแรกที่กล่าว เพราะว่าถ้าคำใดปรากฏในทุก ๆ ประเภทเอกสารคำนั้นจะไม่ถูกเลือกมากำหนดเป็นคุณลักษณะ ในขณะที่วิธี TFIDF จะมีจำนวนคุณลักษณะน้อยที่สุด เนื่องจากว่าถูกกำหนดด้วยว่าความถี่ของคำจะต้องมีค่ามากกว่าค่า Threshold จากการทดลองที่ 1

ตารางที่ 4.6 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนเอกสาร

ชุดเอกสารทดสอบ	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
1	93.07	92.92	87.80	86.72	89.62	89.44	90.20	90.11
2	92.82	92.69	87.76	87.17	89.45	89.29	89.83	89.73
3	93.11	93.00	86.40	83.04	90.09	89.96	90.57	90.50

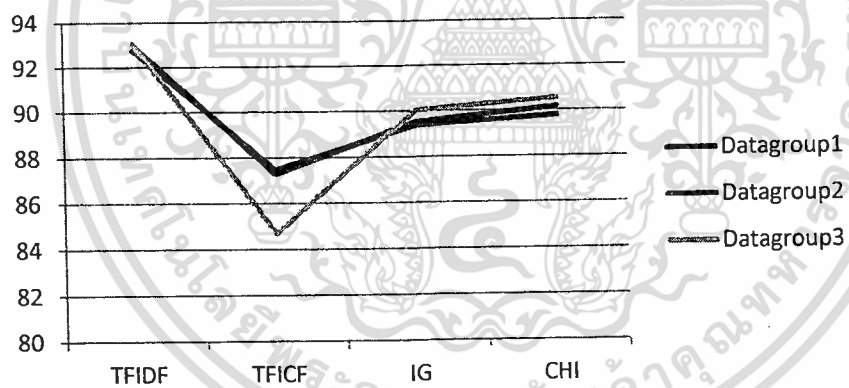
หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

จากตารางที่ 4.6 ค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก แต่อย่างไรก็ตามค่าทั้งสองก็มีค่าที่สูง โดยค่าความเที่ยงตรงมีค่าอยู่ระหว่าง 86-93% และค่าความระลึกมีค่าอยู่ระหว่าง 86-93% เช่นเดียวกัน ซึ่งถือว่ามีประสิทธิภาพในการระบุประเภทเอกสารที่ค่อนข้างดี จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.7

ตารางที่ 4.7 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนเอกสาร

ชุดเอกสารทดสอบ	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก			
	TFIDF	TFICF	IG	CHI
1	93.00	87.26	89.53	90.16
2	92.76	87.46	89.37	89.78
3	93.05	84.68	90.03	90.54

จากตารางที่ 4.7 นำมาสร้างกราฟได้ดังรูปที่ 4.11



รูปที่ 4.11 แสดงค่าเอฟตามกลุ่มเอกสารชุดทดสอบ และวิธีการคำนวณค่าน้ำหนัก

จากรูปที่ 4.11 กำหนดให้ Datagroup1 คือ ชุดเอกสารทดสอบที่ 1 Datagroup2 คือ ชุดเอกสารทดสอบที่ 2 และ Datagroup3 คือ ชุดเอกสารทดสอบที่ 3 อธิบายได้ว่า ในการคำนวณค่าน้ำหนักด้วยวิธี TFIDF พบว่าเอกสารที่มีจำนวนทดสอบมากที่สุด (ชุดเอกสารทดสอบที่ 3) จะให้ประสิทธิภาพในการระบุประเภทเอกสารดีที่สุด ซึ่งมีค่าเอฟเท่ากับ 93.05% ซึ่งหมายความว่าสามารถระบุประเภทเอกสารได้ถูกเมื่อเทียบเป็นเปอร์เซ็นต์ได้ 93.05% การคำนวณค่าน้ำหนักด้วยวิธี TFICF นั้นเอกสารทดสอบที่ 2 จะให้ค่าเอฟมากที่สุด เท่ากับ 87.46% การคำนวณค่านำหนักด้วยวิธี IG และ CHI นั้น เอกสารทดสอบที่ 3 จะให้ค่าเอฟมากที่สุด เท่ากับ 90.03% และ 90.54% ตามลำดับ ดังนั้นจึงอาจกล่าวได้ว่า ยิ่งมีจำนวนเอกสารที่ใช้ในการเรียนรู้มากเท่าไหร่แล้วก็ยิ่งทำให้ประสิทธิภาพระบุประเภทเอกสารมีเพิ่มมากขึ้นไปด้วย นอกจากนั้นแล้ว

จากค่าเอฟที่คำนวณได้จากการใช้วิธีการคำนวณค่าน้ำหนักที่แตกต่างกันนั้น การใช้วิธี TFIDF (Threshold=5) นั้นจะให้ประสิทธิภาพการระบุประเภทเอกสารได้ดีกว่าวิธี CHI IG และ TFICF ตามลำดับ หรือสามารถบอกได้ว่าเมื่อใช้วิธีการคำนวณค่าน้ำหนักด้วย TFIDF โดยกำหนดค่า Threshold เท่ากับ 5 นั้น สามารถระบุประเภทเอกสารได้ถูกต้องมากกว่าคำนวณค่านำหนักด้วยวิธีอื่นที่เปรียบเทียบแสดงดังข้างต้น

การทดลองที่ 2.2 เปรียบเทียบประสิทธิภาพวิธีการคำนวณค่าน้ำหนักต่าง ๆ โดยมีเงื่อนไขคือจำนวนเอกสารที่แตกต่างกันและจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกัน ในการเลือกจำนวนคำนั้นก็คือคำที่จะใช้เป็นตัวแทนของแต่ละประเภทเอกสารนั้น จะกำหนดโดยเลือกคำที่มีค่าน้ำหนักมากที่สุดของแต่ละประเภทเอกสารจำนวน n คำ แล้วนำมารวมกันเป็นตัวแทนของกลุ่มเอกสาร เช่น กำหนดให้เลือกคำจากแต่ละประเภทเอกสารประเภทละ 500 คำ (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นตัวแทนประเภทเอกสารจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท จึงตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำที่เป็นตัวแทนประเภทเอกสารเท่ากับ 1,507 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,932 คำ จากวิธี IG และ 2,161 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นตัวแทนของกลุ่มเอกสาร แสดงในตารางที่ 4.8

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นตัวแทนจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คำเท่ากับ 2,937 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,819 คำ จากวิธี IG และ 4,158 คำ จากวิธี CHI

ในการทดลองนี้แบ่งออกเป็นการทดลองย่อยแยกตามชุดเอกสารทดสอบ จำนวน 3 ชุดเอกสาร ซึ่งจะประกอบด้วยรายละเอียดต่อไปนี้คือ จำนวนคำที่ทำหน้าที่เป็นคุณลักษณะ แยกตามวิธีการคำนวณค่าน้ำหนักต่าง ๆ ประสิทธิภาพของแต่ละวิธี ด้วยตัววัดค่าความเที่ยงตรง ค่าความระลึก และค่าเอฟ

ชุดเอกสารทดสอบ กลุ่มที่ 1 ซึ่งมีเอกสารจำนวน 3,000 เอกสาร แล้วทดลองตามจำนวนที่กำหนดเป็นตัวแทนที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.8

ตารางที่ 4.8 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,507	1,932	2,161

1,000	6,000	2,937	3,819	4,158
1,500	8,251	4,168	5,584	5,504
2,000	-	5,821	6,975	6,836
4,000	-	13,652	12,328	10,394

ตารางที่ 4.9 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
500	91.55	91.22	80.14	75.86	68.85	67.89	69.99	67.81
1,000	92.92	92.78	80.90	77.00	74.58	74.25	76.62	75.61
1,500	93.07	92.92	83.01	79.97	78.91	78.42	79.82	79.39
2,000	-	-	84.90	82.22	79.47	78.97	80.34	79.89
4,000	-	-	86.32	84.25	81.30	80.58	86.53	86.19

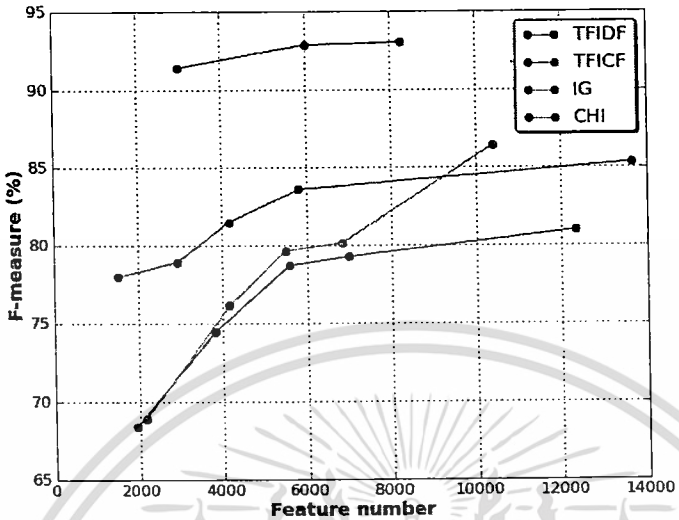
หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

จากตารางที่ 4.9 พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้น นั้นหมายความว่า การระบุประเภทเอกสารมีความถูกต้องเพิ่มมากขึ้น จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.10

ตารางที่ 4.10 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 3,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	91.38	77.94	68.37	68.88	76.64
1,000	92.85	78.90	74.42	76.11	80.57
1,500	93.00	81.46	78.66	79.60	83.18
2,000	-	83.54	79.22	80.11	80.97
4,000	-	85.27	80.94	86.36	84.19

จากตารางที่ 4.10 นำค่าเอฟดังกล่าวมาสร้างกราฟได้ดังรูปที่ 4.12



รูปที่ 4.12 แสดงค่าเอฟแยกตามจำนวนคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 1

จากตารางที่ 4.10 และ รูปที่ 4.12 พบว่าเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้น ค่าเอฟจะเพิ่มมากขึ้นด้วยในทุก ๆ วิธีการคำนวณค่าน้ำหนัก ได้แก่ ในการวิธีคำนวณค่าน้ำหนักด้วย TFIDF ค่าเอฟจะเพิ่มขึ้นจาก 91.38%, 92.85% และ 93.00% เมื่อกำหนดคุณลักษณะให้ n = 500, 1000 และ 1,500 ตามลำดับ เมื่อใช้วิธี TFICF ค่าเอฟจะเพิ่มขึ้นจาก 77.94% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 85.27% เมื่อกำหนดคุณลักษณะให้ n=4,000 เมื่อใช้วิธี IG ค่าเอฟจะเพิ่มขึ้นจาก 68.37% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 80.94% เมื่อกำหนดคุณลักษณะให้ n=4,000 เมื่อใช้วิธี CHI ค่าเอฟจะเพิ่มขึ้นจาก 68.88% เมื่อกำหนดคุณลักษณะให้ n=500 เป็น 86.36% เมื่อกำหนดคุณลักษณะให้ n=4,000

ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 76.64%, 80.57%, 83.18%, 80.97% และ 84.19% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบ กลุ่มที่1 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น โดยการคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ค่าเอฟที่มากที่สุด

ชุดเอกสารทดสอบ กลุ่มที่ 2 ซึ่งมีเอกสารจำนวน 4,200 เอกสาร แล้วทดลองตามจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.11

ตารางที่ 4.11 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,440	1,883	2,188
1,000	6,000	2,853	3,745	4,187
1,500	9,000	4,340	5,563	5,655
2,000	-	5,549	6,976	7,005
4,000	-	13,506	12,328	26,615

จากตารางที่ 4.11 เมื่อกำหนดให้เลือกคำที่ทำหน้าที่เป็นคุณลักษณะจากแต่ละประเภทเอกสาร ประเภทละ 500 คำ (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นตัวแทนจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท ทำให้ต้องตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำเท่ากับ 1,440 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,883 คำ จากวิธี IG และ 2,188 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นคุณลักษณะ

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นคุณลักษณะจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คุณลักษณะเท่ากับ 2,853 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,745 คำ จากวิธี IG และ 4,187 คำ จากวิธี CHI

ตารางที่ 4.12 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
	P	R	P	R	P	R	P	R
500	90.93	90.52	76.04	73.55	64.91	64.29	68.33	67.81
1,000	92.64	92.48	79.33	76.07	74.43	74.19	74.94	74.52
1,500	92.87	92.74	81.75	79.67	77.61	77.33	79.32	79.21
2,000	-	-	83.38	81.33	79.54	79.33	81.06	80.98
4,000	-	-	85.01	83.81	83.76	83.50	84.83	84.69

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

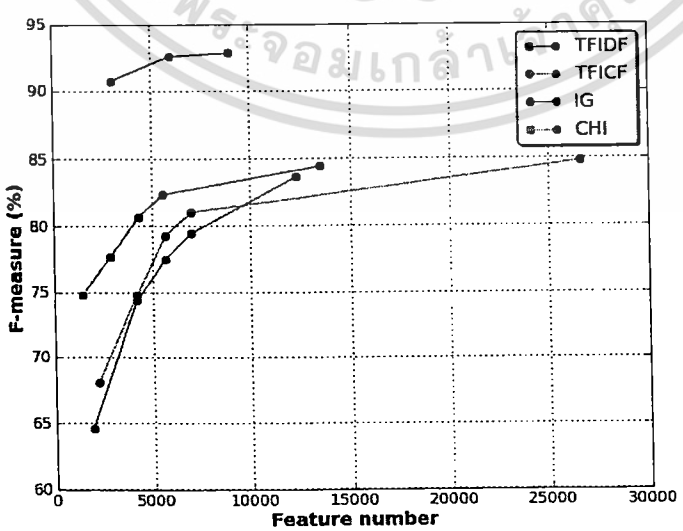
จากตารางที่ 4.12 พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้น ในชุดเอกสารทดสอบที่ 2 โดยการคำนวณด้วยวิธี TFIDF จะให้ค่าความเที่ยงตรงและค่าความระลึกสูงที่สุด ในขณะเดียวกันเมื่อคำนวณค่าน้ำหนักด้วยวิธี IG จะให้ค่าความเที่ยงตรงและค่าความระลึกน้อยที่สุด

โดยค่าความเที่ยงตรงและค่าความระลึกมีค่าสูง หมายความว่า มีประสิทธิภาพการระบุประเภทเอกสารได้มาก โดยสามารถระบุประเภทเอกสารได้ถูกต้องเป็นจำนวนมาก จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก แสดงดังตารางที่ 4.13

ตารางที่ 4.13 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 4,200 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	90.73	74.77	64.60	68.07	74.54
1,000	92.56	77.67	74.31	74.73	79.82
1,500	92.80	80.69	77.47	79.26	82.53
2,000	-	82.34	79.43	81.02	80.93
4,000	-	84.41	83.63	84.76	84.27

จากตารางที่ 4.13 นำมาสร้างกราฟได้ดังรูปที่ 4.13



รูปที่ 4.13 แสดงค่าเอฟแยกตามจำนวนคำและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 2

จากตารางที่ 4.13 และ รูปที่ 4.13 พบว่าเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น ค่าเอฟจะเพิ่มมากขึ้นด้วยในทุก ๆ วิธีการคำนวณค่าน้ำหนักของชุดเอกสารทดสอบที่ 2 ดังรายละเอียดต่อไปนี้ ในวิธีคำนวณค่าน้ำหนักด้วยวิธีต่าง ๆ เมื่อกำหนดคุณลักษณะให้ n = 500, 1,000, 1,500, 2,000 และ 4,000 ตามลำดับ เมื่อค่า n มากขึ้นจำนวนคุณลักษณะก็จะมากขึ้นตามไปด้วย โดยเมื่อใช้วิธี TFICF ค่าเอฟจะมีค่าเท่ากับ 74.77%, 77.67%, 80.69%, 82.34% และ 84.41% ตามลำดับ สำหรับวิธี IG ค่าเอฟจะมีค่าเท่ากับ 64.60%, 74.31%, 77.47%, 79.43 และ 83.63% ตามลำดับ และสำหรับวิธี CHI ค่าเอฟจะมีค่าเท่ากับ 68.07%, 74.73%, 79.26%, 81.02% และ 84.76% ตามลำดับ จากค่าเอฟเหล่านี้อธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 2 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น

ในทำนองเดียวกันการคำนวณค่าน้ำหนักด้วยวิธี TFIDF นั้น ค่าเอฟจะมีค่าเพิ่มสูงขึ้นเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น โดยที่เมื่อกำหนดให้ n=500, 1,000 และ 1,500 ค่าเอฟมีค่าเท่ากับ 90.73%, 92.56% และ 92.80%

ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 74.54%, 79.82%, 82.53%, 80.93% และ 84.27% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 2 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น และการคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ค่าเอฟที่มากกว่าการคำนวณค่าน้ำหนักด้วยวิธีอื่นที่ทำการเปรียบเทียบ ดังนั้นเมื่อใช้วิธี TFIDF เมื่อกำหนดค่า Threshold เท่ากับ 5 ในการกำหนดคำที่เป็นคุณลักษณะจะให้การระบุประเภทเอกสารมีความถูกต้องมากกว่าวิธีอื่น

ชุดเอกสารทดสอบ กลุ่มที่ 3 ซึ่งมีเอกสารจำนวน 6,000 เอกสาร แล้วทดลองตามจำนวนคำที่กำหนดเป็นคุณลักษณะที่แตกต่างกันไป โดยกำหนดจากค่า n ซึ่งแสดงในตารางที่ 4.14

ตารางที่ 4.14 ตารางแสดงจำนวนคำไม่ซ้ำในเอกสารแต่ละประเภทตามจำนวนคำที่เลือกของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	รวมจำนวนคำที่ไม่ซ้ำในเอกสาร 6 ประเภท			
	TFIDF	TFICF	IG	CHI
500	3,000	1,453	1,870	2,235

1,000	6,000	2,851	3,729	4,260
1,500	9,000	4,523	5,565	6,014
2,000	-	5,626	7,360	7,410
4,000	-	12,453	12,601	11,685

จากตารางที่ 4.14 เมื่อกำหนดให้เลือกคำที่ทำหน้าที่เป็นคุณลักษณะจากแต่ละประเภทเอกสาร ประเภทละ 500 8e (n=500) จะได้เซตของคำทั้งหมดที่ทำหน้าที่เป็นคุณลักษณะจำนวน 3,000 คำ จาก 6 ประเภทเอกสาร ซึ่งในเซตของคำนี้จะมีคำซ้ำกัน เนื่องจากคำ ๆ หนึ่งสามารถเป็นตัวแทนของประเภทเอกสารได้หลายประเภท ทำให้ต้องตัดคำที่ซ้ำกันออก ดังนั้น จะได้จำนวนคำเท่ากับ 1,453 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 1,870 คำ จากวิธี IG และ 2,235 คำ จากวิธี CHI และในวิธี TFIDF จะเป็นการคำนวณค่านำหนักของคำในแต่ละเอกสาร ไม่ได้แยกตามประเภทเอกสาร ทำให้ไม่มีคำที่ซ้ำกันจึงเลือกคำทั้งหมดนั้นกำหนดเป็นคุณลักษณะ

เมื่อกำหนดค่า n=1,000 จะได้เซตของคำที่ทำหน้าที่เป็นคุณลักษณะจำนวน 6,000 คำ จาก 6 ประเภทเอกสาร เมื่อตัดคำที่ซ้ำกันจะได้คุณลักษณะเท่ากับ 2,851 คำ จากวิธีการคำนวณค่าน้ำหนักด้วย TFICF 3,729 คำ จากวิธี IG และ 4,260 คำจากวิธี CHI

ตารางที่ 4.15 แสดงค่าความเที่ยงตรงและค่าความระลึกตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท ของเอกสารจำนวน 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท	ค่าความเที่ยงตรงและค่าความระลึกแบ่งตามวิธีการคำนวณค่าน้ำหนัก							
	TFIDF		TFICF		IG		CHI	
(n)	P	R	P	R	P	R	P	R
500	91.34	90.95	72.41	67.86	62.66	61.71	66.03	65.77
1,000	92.74	92.61	77.60	75.43	71.92	70.88	73.24	72.88
1,500	93.04	92.93	80.60	75.60	77.33	76.71	77.15	76.40
2,000	-	-	82.07	78.43	79.74	79.45	79.22	78.80
4,000	-	-	84.70	83.30	83.06	82.95	85.14	85.06

หมายเหตุ P=ค่าความเที่ยงตรง R=ค่าความระลึก

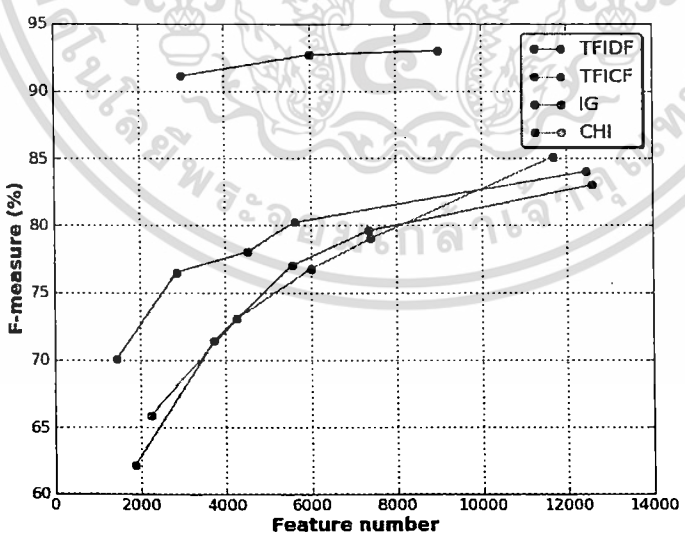
จากตารางที่ 4.15พบว่าค่าความเที่ยงตรงในแต่ละการทดลองจะมีค่ามากกว่าค่าความระลึก โดยเมื่อจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มมากขึ้นค่าความเที่ยงตรงและค่าความระลึกก็จะมีค่าสูงขึ้นในชุดเอกสารทดสอบที่ 3 โดยการคำนวณด้วยวิธี TFIDF ยังคงให้ค่าความเที่ยงตรงและค่าความระลึกสูงที่สุด ซึ่งหมายความว่า การคำนวณค่าน้ำหนักด้วยวิธี TFIDF จะให้ประสิทธิภาพในการระบุประเภทเอกสารดีที่สุด

ในขณะเดียวกันเมื่อคำนวณค่าน้ำหนักด้วยวิธี IG ก็ยังคงให้ค่าความเที่ยงตรงและค่าความระลึกน้อยที่สุด หมายความว่า การคำนวณค่าน้ำหนักด้วยวิธี IG จะให้ประสิทธิภาพในการระบุประเภทเอกสารน้อยที่สุด หรือจะกล่าวว่าระบุประเภทเอกสารให้กับเอกสารได้ถูกต้องน้อยที่สุดเมื่อเปรียบเทียบกับการคำนวณค่าน้ำหนักด้วยวิธีอื่น ๆ จากนั้นคำนวณค่าเอฟจากค่าความเที่ยงตรงและค่าความระลึก ได้ดังตารางที่ 4.16

ตารางที่ 4.16 แสดงค่าเอฟตามวิธีการคำนวณค่าน้ำหนักและจำนวนคำที่เลือกเป็นคุณลักษณะของเอกสารแต่ละประเภท โดยที่จำนวนเอกสารเท่ากับ 6,000 เอกสาร

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเอฟแบ่งตามวิธีการคำนวณค่าน้ำหนัก				ค่าเอฟเฉลี่ยทุกวิธีการคำนวณค่าน้ำหนัก
	TFIDF	TFICF	IG	CHI	
500	91.14	70.06	62.19	65.90	72.32
1,000	92.68	76.50	71.39	73.06	78.41
1,500	92.98	78.02	77.02	76.77	81.20
2,000	-	80.21	79.59	79.00	79.60
4,000	-	84.00	83.00	85.10	84.03

จากตารางที่ 4.16 นำมาสร้างกราฟได้ดังรูปที่ 4.14



รูปที่ 4.14 แสดงค่าเอฟแยกตามจำนวนคำที่กำหนดเป็นคุณลักษณะและวิธีการคิดค่าน้ำหนัก ของเอกสารชุดทดสอบที่ 3

จากตารางที่ 4.16 และรูปที่ 4.14 พบว่าเมื่อมีจำนวนคำที่ทำหน้าที่เป็นคุณลักษณะเพิ่มขึ้น ค่าเอฟก็จะมากขึ้นด้วยในทุก ๆ วิธีการคำนวณค่าน้ำหนักของชุดเอกสารทดสอบที่ 3 ดังรายละเอียดต่อไปนี้ ในวิธีคำนวณค่าน้ำหนักด้วยวิธีต่าง ๆ เมื่อค่า n มากขึ้นจำนวนคำที่กำหนดเป็นคุณลักษณะก็จะมากขึ้นตามไปด้วย โดยกำหนดคุณลักษณะให้ n = 500, 1000, 1500, 2000 และ 4000 ตามลำดับ เมื่อใช้วิธี TFICF ค่าเอฟจะมีค่าเท่ากับ 70.06%, 76.50%, 78.02%, 80.21% และ 84.00% ตามลำดับ สำหรับวิธี IG ค่าเอฟจะมีค่าเท่ากับ 64.60%, 71.39%, 76.77%, 79.59% และ 83.00% ตามลำดับ และสำหรับวิธี CHI ค่าเอฟจะมีค่าเท่ากับ 65.90%, 73.06%, 79.26%, 79.00% และ 85.10% ตามลำดับ จากค่าเอฟเหล่านี้อธิบายได้ว่าเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 3 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น

ในทำนองเดียวกันการคำนวณค่าน้ำหนักด้วยวิธี TFIDF นั้น ค่าเอฟจะมีค่าเพิ่มสูงขึ้นเมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น โดยที่เมื่อกำหนดให้ n=500, 1000 และ 1500 ค่าเอฟมีค่าเท่ากับ 91.14%, 92.68% และ 92.98%

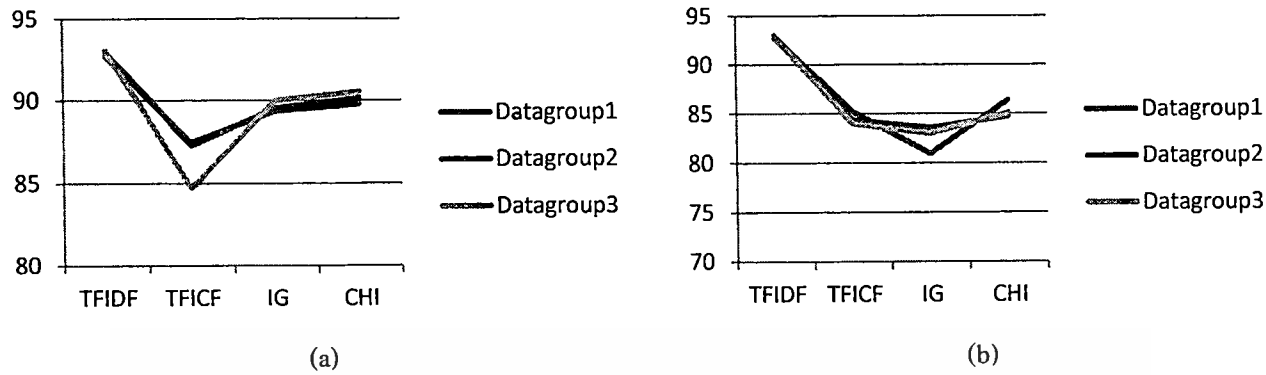
ค่าเอฟเฉลี่ยของทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่กำหนดให้เป็นคุณลักษณะจะมีแตกต่างกัน เมื่อกำหนดให้ n มีค่า เท่ากับ 500, 1000, 1500, 2000 และ 4000 จะมีค่าเอฟเฉลี่ยเท่ากับ 72.32%, 78.41%, 81.20%, 79.60% และ 84.03% ตามลำดับ โดยเมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะมีค่าเพิ่มขึ้นแล้วค่าเอฟเฉลี่ยจะมีค่าเพิ่มขึ้นด้วย ยกเว้นเมื่อเพิ่ม n จาก 1,500 เป็น 2,000

จากข้อมูลข้างต้นอธิบายได้ว่าเมื่อจำนวนคำที่เป็นคุณลักษณะเพิ่มมากขึ้นในชุดเอกสารทดสอบกลุ่มที่ 3 ประสิทธิภาพในการระบุประเภทเอกสารจะเพิ่มมากขึ้น หมายความว่ามีการระบุประเภทเอกสารได้ถูกต้องมากขึ้น โดยพิจารณาจากค่าเอฟที่สูงขึ้น และการคำนวณค่าน้ำหนักด้วยวิธี TFIDF ก็ยังคงให้ค่าเอฟที่มากกว่าการคำนวณค่าน้ำหนักด้วยวิธีอื่นที่ทำการเปรียบเทียบเหมือนกับการทดลองกับชุดเอกสารทดสอบกลุ่มที่ 1 และ 2 ดังนั้นเมื่อใช้วิธี TFIDF ในการกำหนดคุณลักษณะจะให้การระบุประเภทเอกสารมีความถูกต้องมากกว่าวิธีอื่น

4.3 อธิบายผลการทดลอง

จากการทดลองข้างต้นสามารถสรุปตามสมมติฐานดังนี้

1. จำนวนเอกสารที่ใช้ในการฝึกสอนมีผลต่อการระบุประเภทเอกสารหรือไม่ (Effect of training set size on performance)



(a) (b)

รูปที่ 4.15 แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนเอกสารที่ใช้ในการทดสอบ
(a) ไม่มีการเลือกคุณลักษณะ (b) เลือกคุณลักษณะ

จากรูปที่ 4.15 Datagroup1 (ชุดเอกสารทดสอบที่ 1) ใช้เอกสารทดสอบจำนวน 3,000 เอกสาร Datagroup2 (ชุดเอกสารทดสอบที่ 2) ใช้เอกสารทดสอบจำนวน 4,200 เอกสาร Datagroup3 (ชุดเอกสารทดสอบที่ 3) ใช้เอกสารทดสอบจำนวน 6,000 เอกสาร อธิบายได้ว่า ในกรณีที่ไม่มีการเลือกจำนวนคำที่กำหนดเป็นคุณลักษณะ (รูปที่ 4.15(a)) แสดงว่าทุกคำจะถูกกำหนดให้เป็นคุณลักษณะ เมื่อจำนวนเอกสารทดสอบเพิ่มมากขึ้น ค่าเอฟก็จะสูงขึ้นด้วย ซึ่งค่าเอฟสูงนั้นหมายความว่าการระบุประเภทเอกสารมีความถูกต้องมาก เมื่อกำหนดค่าน้ำหนักด้วยวิธี TFIDF แต่ในวิธี IG และ CHI ค่าเอฟจะลดลงใน Datagroup2 และเพิ่มขึ้นใน Datagroup3 ขณะที่ TFICF ค่าเอฟจะเพิ่มขึ้นใน Datagroup2 และลดลงใน Datagroup3

ในกรณีที่มีการเลือกคำที่กำหนดเป็นคุณลักษณะ (รูปที่ 4.15(b)) อธิบายได้ว่า ในวิธี TFIDF และ CHI ค่าเอฟจะลดลงใน Datagroup2 และเพิ่มขึ้นใน Datagroup3 ขณะที่ TFICF ค่าเอฟจะลดลงเมื่อจำนวนเอกสารทดสอบเพิ่มมากขึ้น และ IG ค่าเอฟจะเพิ่มขึ้นใน Datagroup2 และลดลงใน Datagroup3

ดังนั้นจึงสามารถสรุปได้ดังนี้คือ ความแตกต่างของจำนวนเอกสารที่ใช้ในการเรียนรู้หรือฝึกสอนนั้นส่งผลต่อประสิทธิภาพการระบุประเภทเอกสาร โดยถ้าเราไม่เลือกจำนวนคุณลักษณะทุกวิธีการคำนวณค่าน้ำหนักยกเว้น TFICF จะมีประสิทธิภาพดีในการระบุประเภทเอกสาร นั้นหมายความว่าจำนวนเอกสารมากก็ยิ่งทำให้การระบุประเภทเอกสารดีขึ้น

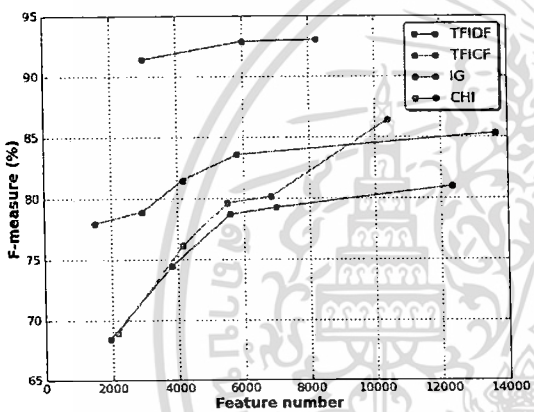
แต่เมื่อมีการลดจำนวนคำที่กำหนดเป็นคุณลักษณะเพื่อลดขนาดของเวกเตอร์แต่ละเอกสารจะมีความไม่แน่นอนในประสิทธิภาพของการระบุประเภทของเอกสารขึ้นอยู่กับแต่ละวิธีการคำนวณค่าน้ำหนัก อย่างไรก็ตามการคำนวณค่าน้ำหนักด้วย TFIDF และ CHI มีแนวโน้มไปทางเดียวกัน คือจำนวนเอกสารที่ใช้ในการทดสอบมีจำนวนมากจะมีผลทำให้ประสิทธิภาพของการระบุประเภทเอกสารเพิ่มมากขึ้นด้วย

2. จำนวนคุณลักษณะมีผลต่อการจัดกลุ่มหรือไม่ (Effect of feature size on performance)

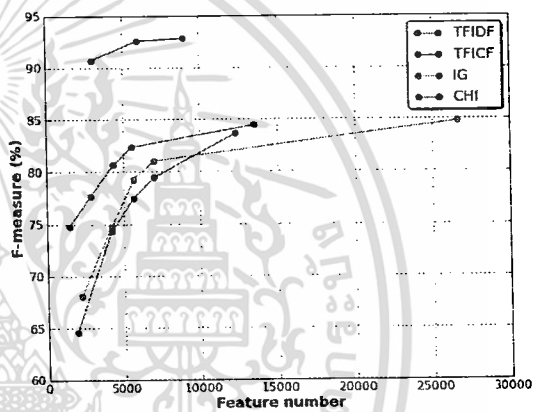
จากรูปที่ 4.16 อธิบายได้ว่าในทุก ๆ วิธีการคำนวณค่าน้ำหนักและทุก ๆ กลุ่มเอกสารทดสอบ เมื่อจำนวนคำที่กำหนดเป็นคุณลักษณะเพิ่มมากขึ้นมีผลทำให้ค่าเอฟมีค่าเพิ่มมากขึ้นตามไปด้วย ดังนั้นจึง

สามารถสรุปได้ว่าจำนวนคุณลักษณะมีผลต่อประสิทธิภาพการจัดกลุ่มเอกสาร โดยเมื่อค่าเอฟสูงหมายความว่าการระบุประเภทเอกสารมีความถูกต้องมาก ในทางตรงกันข้ามถ้าค่าเอฟมีค่าต่ำแสดงว่าการระบุเอกสารนั้นมีความผิดพลาดค่อนข้างมาก โดยระบุประเภทเอกสารไม่ถูกต้อง

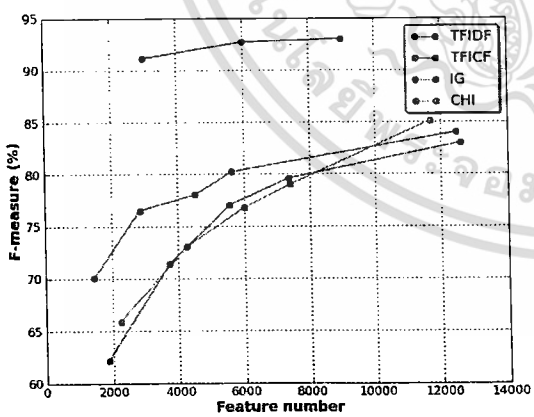
จากกราฟสามารถอธิบายเพิ่มได้ว่า ในทุกชุดเอกสารทดสอบการคำนวณค่านำหนักด้วยวิธี IG และ CHI จะให้ค่าเอฟ (F-measure) ที่ใกล้เคียงกัน คือเป็นค่าที่ระบุความถูกต้องของการระบุประเภทเอกสารที่ไม่ค่อยดีนัก ซึ่งมีการกำหนดจำนวนคำที่กำหนดปันคุณลักษณะที่ไม่มาก คืออยู่ระหว่าง 2,000-7,000 คำ แต่เมื่อจำนวนเพิ่มจำนวนคุณลักษณะ (มากกว่า 8,000 คำ) แล้วจะมีความแตกต่างของค่าเอฟอย่างชัดเจน แสดงดังกราฟ แต่อย่างไรก็ตามวิธี TFIDF (โดยกำหนด Threshold=5) นั้นให้ค่าเอฟที่สูงในทุก ๆ ชุดเอกสารทดสอบ แสดงว่ามีความถูกต้องในการระบุประเภทเอกสารที่มีประสิทธิภาพ



(a) ชุดเอกสารทดสอบที่ 1



(b) ชุดเอกสารทดสอบที่ 2



(c) ชุดเอกสารทดสอบที่ 3

รูปที่ 4.16 แสดงความสัมพันธ์ระหว่างค่าเอฟกับจำนวนคำที่กำหนดเป็นคุณลักษณะ ในแต่ละชุดเอกสารทดสอบ

เมื่อคำนวณค่าเฉลี่ยของค่าเอฟทุกวิธีการคำนวณค่าน้ำหนักแยกตามจำนวนคำที่เลือกในเอกสารแต่ละประเภท ตามตารางที่ 4.17 พบว่าเมื่อเพิ่มจำนวนคำที่เลือกเพื่อกำหนดเป็นคำสำคัญมากขึ้นจาก 500, 1000, 1500, 2000 และ 4000 ในเอกสารชุดทดสอบที่ 1 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 76.64%, 80.57%, 83.18%, 80.97% และ 84.19% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 3.93%, 2.61%,-2.21% และ 3.22% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 1.88% ในเอกสารชุดทดสอบที่ 2 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 74.54%, 79.82%, 82.53%, 80.93% และ 84.27% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 5.28%, 2.71%, -1.60% และ 3.34% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.43% ในเอกสารชุดทดสอบที่ 3 ค่าเฉลี่ยค่าเอฟมีค่าเท่ากับ 72.32%, 78.41%, 81.20%, 79.60% และ 84.03% ตามลำดับ และค่าเฉลี่ยค่าเอฟมีการเพิ่มขึ้นเป็น 6.09%, 2.79%, -1.60% และ 4.43% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.93% จากการทดลองในทุกชุดเอกสารทดสอบแล้วค่าเฉลี่ยค่าเอฟจะมีค่าเพิ่มขึ้นยกเว้น เมื่อกำหนดจำนวนคำจากเดิมเท่ากับ 1500 เป็น 2000 ค่าเฉลี่ยค่าเอฟจะมีค่าลดลง

ค่าเฉลี่ยของผลรวมค่าเฉลี่ยเอฟของทุกชุดเอกสารทดสอบ เมื่อกำหนดให้จำนวนคำที่เลือกเป็นคำสำคัญเท่ากับ 500, 1000, 1500, 2000 และ 4000 นั้น ผลรวมของค่าเฉลี่ยของค่าเฉลี่ยเอฟเท่ากับ 74.50%, 79.60%, 82.30%, 80.50% และ 84.16% ตามลำดับ และค่าเฉลี่ยมีการเพิ่มขึ้นเป็น 5.10%, 2.70%, -1.80% และ 3.66% ตามลำดับ โดยเฉลี่ยมีความถูกต้องเพิ่มขึ้น 2.41%

ตารางที่ 4.17 แสดงค่าเฉลี่ยค่าเอฟแบ่งตามจำนวนคำที่เลือกในเอกสารแต่ละประเภท และจำนวนเอกสารตามชุดเอกสารทดสอบที่ 1, 2 และ 3

จำนวนคำที่เลือกในเอกสารแต่ละประเภท (n)	ค่าเฉลี่ยค่าเอฟแบ่งตามชุดเอกสารทดสอบ						ค่าเฉลี่ยของผลรวมค่าเฉลี่ยเอฟของทุกชุดเอกสารทดสอบ	
	ชุดเอกสารทดสอบที่ 1 (3,000 เอกสาร)		ชุดเอกสารทดสอบที่ 2 (4,200 เอกสาร)		ชุดเอกสารทดสอบที่ 3 (6,000 เอกสาร)			
	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง	ค่าเฉลี่ย	ผลต่าง
500	76.64	-	74.54	-	72.32	-	74.50	-
1,000	80.57	3.93	79.82	5.28	78.41	6.09	79.60	5.10
1,500	83.18	2.61	82.53	2.71	81.20	2.79	82.30	2.70
2,000	80.97	-2.21	80.93	-1.60	79.60	-1.60	80.50	-1.80
4,000	84.19	3.22	84.27	3.34	84.03	4.43	84.16	3.66
เพิ่มขึ้นโดยเฉลี่ย (%)	1.88		2.43		2.93		2.41	

3. ค่า Threshold ที่ใช้ในการเลือกคำที่กำหนดเป็นคุณลักษณะด้วยการคำนวณวิธี TFIDF มีผลอย่างไร

ในการกำหนดค่า Threshold ในงานวิจัยนี้หมายถึง การกำหนดค่าต่ำสุดของความถี่ของคำที่จะนำมาพิจารณาเพื่อกำหนดเป็นคุณลักษณะโดยเลือกค่า Threshold ในการทดลองจำนวน 4 ค่า ได้แก่ 3, 4, 5 และ 6 สามารถอธิบายได้ว่า กำหนดให้ Threshold เท่ากับ 3 หมายความว่า เลือกคำ (ที่ผ่านการตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์) ที่มีความถี่มากกว่า 3 นำมาคำนวณค่าน้ำหนักด้วยวิธี TFIDF

จากการทดลองที่ 1 ค่า Threshold ที่เหมาะสมที่ทำให้การจัดกลุ่มเอกสารมีประสิทธิภาพดีที่สุด ได้แก่ ค่า Threshold เท่ากับ 5 นั้นหมายความว่า คำที่มีความถี่มากกว่า 5 ครั้งนั้นคาดว่าน่าจะคำที่มีความสำคัญที่สามารถนำมาใช้เป็นตัวแทนของเอกสาร โดยสามารถระบุประเภทเอกสารได้ดีกว่าการกำหนดด้วยค่า Threshold อื่น ๆ

4. เปรียบเทียบวิธีการคำนวณค่าน้ำหนักระหว่าง TFIDF, TFICF, IG และ CHI (Performance comparison between four approaches)

จากรูปที่ 1.14 และ 4.15 อธิบายได้ว่า การจัดกลุ่มโดยการใช้วิธีการคำนวณค่าน้ำหนักด้วย TFIDF ในงานวิจัยนี้ ให้ประสิทธิภาพการระบุประเภทเอกสารดีที่สุด นั้นคือมีการระบุประเภทเอกสารที่ถูกต้องค่อนข้างมากเมื่อเทียบกับวิธีอื่น ๆ โดยพิจารณาจากค่าเอฟที่มากกว่าวิธีอื่น โดยค่าเอฟที่มากที่สุดที่คำนวณด้วยวิธีการคำนวณค่าน้ำหนัก TFIDF ที่ Threshold เท่ากับ 5 นั้นมีค่าเท่ากับ 93.05% โดยที่การคำนวณค่าน้ำหนักด้วยวิธีอื่น ๆ นั้นมีค่าอยู่ระหว่าง 80-90% นั้นหมายความว่าการใช้วิธีการคำนวณค่าน้ำหนักด้วยวิธี TFIDF ที่ Threshold เท่ากับ 5 มีความถูกต้องสูงกว่าการคำนวณด้วยวิธีอื่น

บทที่ 5
สรุปผลการวิจัย

5.1 สรุปผลการวิจัย

งานวิจัยฉบับนี้เป็นงานวิจัยที่ทำการทดลองเพื่อเปรียบเทียบประสิทธิภาพการกำหนดหัวข้อข่าวให้กับเอกสารโดยคำนึงถึงปัจจัยต่าง ๆ ที่มีผลกระทบต่อประสิทธิภาพการกำหนดหัวข้อข่าว ได้แก่ จำนวนเอกสารที่ใช้ในการเรียนรู้ จำนวนคุณลักษณะที่เหมาะสม และค่า Threshold ที่ใช้ในการกำหนดค่าน้ำหนักด้วยวิธี TFIDF ในการกำหนดหัวข้อข่าวนั้น เราจะใช้เทคนิคการกำหนดคุณลักษณะเพื่อให้ได้กลุ่มคำที่สามารถใช้ระบุเพื่อเป็นตัวแทนของเอกสารในแต่ละประเภทกลุ่มเอกสาร แล้วประยุกต์ใช้อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนในการจัดกลุ่มเอกสารเพื่อระบุหัวข้อให้กับแต่ละเอกสารว่าเป็นเอกสารประเภทใด

โดยกระบวนการทำงานของงานวิจัยฉบับนี้ ได้รวบรวมเอกสารข่าวจากเว็บไซต์ข่าวต่าง ๆ ที่เป็นภาษาอังกฤษ แบ่งข้อมูลข่าวออกเป็น 6 ประเภทข่าว ได้แก่ การเมือง กีฬา สุขภาพ ธุรกิจ บันเทิง และสภาพอากาศ ในขั้นตอนการทำงานนั้นแบ่งออกเป็นสองส่วนคือ ขั้นตอนของการเรียนรู้และการทดสอบหรือการจัดกลุ่มเอกสาร โดยทั้งสองขั้นตอนมีการทำงานคือ การเลือกเนื้อความข่าวจากเอกสาร HTML ตัดคำฟุ่มเฟือย และแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์ หลังจากได้คำทั้งหมดแล้วจึงนำมาหาค่าความถี่และค่าน้ำหนักด้วยวิธีการต่าง ๆ ได้แก่ TFIDF TFICF IG และ CHI แล้วทำการกำหนดคุณลักษณะตามเงื่อนไข ๆ เพื่อหาคุณลักษณะที่เหมาะสมเพื่อกำหนดให้เป็นตัวแทนของกลุ่มเอกสารแต่ละประเภท จากนั้นสร้างเวกเตอร์ของแต่ละเอกสารตามคุณลักษณะข้างต้นในรูปแบบ Vector Space Model แล้วนำข้อมูลเหล่านี้เข้าสู่ขั้นตอนการเรียนรู้เพื่อสร้างโมเดลสำหรับการจัดกลุ่มเอกสารโดยใช้อัลกอริทึมเวกเตอร์ซัพพอร์ตแมชชีน จากการเรียนรู้นี้ผลลัพธ์ที่ได้คือโมเดลการเรียนรู้การจัดกลุ่มเอกสาร

โมเดลการเรียนรู้การจัดกลุ่มเอกสาร จะถูกนำมาทดสอบกับเอกสารกลุ่มใหม่ที่อยู่ในรูปแบบของเวกเตอร์เรียบร้อยแล้ว โดยมีการวัดประสิทธิภาพด้วยค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) และค่าเอฟ (F-Measure)

จากผลการทดสอบในงานวิจัยฉบับนี้สามารถสรุปได้ว่า

(1) ค่า Threshold ที่ใช้ในการกำหนดค่าต่ำสุดของความถี่ของคำที่ปรากฏในกลุ่มเอกสาร เพื่อนำมาใช้กำหนดในการเลือกคุณลักษณะนั้น พบว่าจากค่า Threshold ที่ใช้ในการทดสอบ คือ 3, 4, 5 และ 6 ค่า Threshold ที่มีค่าเท่ากับ 5 ให้ค่าความถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าวดีที่สุดในกลุ่ม Threshold ที่กำหนด โดยมีค่าเอฟเท่ากับเฉลี่ยเท่ากับ 92.94 %

(2) จำนวนเอกสารที่ใช้ในการเรียนรู้มีผลต่อความค่าถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าว เมื่อเพิ่มจำนวนเอกสารที่ใช้ในการเรียนรู้มากขึ้น เมื่อคำนวณค่าน้ำหนักด้วยวิธี TFIDF (กำหนด Threshold=5) ค่าเอฟมีค่าเท่ากับ 93.05 % เมื่อใช้เอกสารในเอกสารทดสอบชุดที่ 3 ซึ่งมีจำนวนเอกสาร

เท่ากับ 6,000 เอกสาร โดยค่าเอฟเมื่อเทียบกับเอสการทดสอบชุดที่ 1 ซึ่งมีค่าเท่ากับ 93% ซึ่งมีจำนวนเอกสาร 3,000 เอกสาร เมื่อไม่มีการกำหนดจำนวนคุณลักษณะ และในทำนองเดียวกันเมื่อมีการระบุจำนวนคุณลักษณะเมื่อจำนวนเอกสารที่ใช้ในการเรียนรู้เพิ่มขึ้นค่าเอฟก็มีค่าที่สูงขึ้น ในการคำนวณค่าน้ำหนักด้วยวิธี TFIDF โดยเฉลี่ยเพิ่มขึ้นประมาณ 2% เมื่อกำหนดเงื่อนไขการเลือกคุณลักษณะ (n=500 และ n=1,500) โดยเมื่อค่าเอฟมีค่าที่สูงนั้นหมายความว่ามีกำหนดหัวข้อข่าวให้กับเอกสารที่ถูกต้องมาก

(3) จำนวนคุณลักษณะมีผลต่อค่าความถูกต้องหรือประสิทธิภาพของกำหนดหัวข้อข่าว เมื่อค่าเอฟสูงขึ้นหมายความว่า มีการระบุประเภทเอกสารที่ถูกเพิ่มมากขึ้นด้วย โดยจากการทดลองเพิ่มจำนวนคำสำคัญในแต่ละประเภทเอกสารครั้งละ 500 คำ พบว่า เมื่อจำนวนคุณลักษณะเพิ่มมากขึ้น ค่าความถูกต้องของการกำหนดหัวข้อข่าวจะเพิ่มมากขึ้นด้วย โดยในกลุ่มเอกสารทดสอบชุดที่ 1 จำนวน 3000 เอกสาร เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นเฉลี่ยประมาณ 0.8% TFICF เพิ่มขึ้นเฉลี่ยประมาณ 1.8% IG เพิ่มขึ้นเฉลี่ยประมาณ 3.1% และ CHI เพิ่มขึ้นประมาณ 4.3% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 1.88% เมื่อทดลองในกลุ่มเอกสารทดสอบชุดที่ 2 จำนวน 4,200 เอกสาร เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นประมาณ 1% TFICF เพิ่มขึ้นประมาณ 2.4% IG เพิ่มขึ้นประมาณ 4.7% และ CHI เพิ่มขึ้นประมาณ 4.1% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 2.43% และในกลุ่มเอกสารทดสอบชุดที่ 3 เมื่อใช้วิธี TFIDF ค่าเอฟเพิ่มขึ้นประมาณ 0.9% TFICF เพิ่มขึ้นประมาณ 3.4% IG เพิ่มขึ้นประมาณ 5.2% และ CHI เพิ่มขึ้นประมาณ 4.8% โดยเฉลี่ยรวมมีค่าความถูกต้องเพิ่มขึ้น 2.93% ดังนั้น ค่าความถูกต้องของทุกกลุ่มเอกสารทดสอบเพิ่มขึ้นโดยเฉลี่ย 2.41%

(4) จากการเปรียบเทียบค่าความถูกต้องหรือประสิทธิภาพของการกำหนดหัวข้อข่าว โดยการกำหนดคุณลักษณะด้วยการคำนวณค่าน้ำหนักด้วยวิธี TFIDF TFICF IG และ CHI นั้น จากการทดลองพบว่าวิธี TFIDF ให้ค่าความถูกต้องมากกว่าวิธีอื่น ๆ โดยค่าเอฟมีค่าประมาณ 93.05% เมื่อใช้ทุกคำที่ผ่านการตัดคำฟุ่มเฟือยและแปลงให้อยู่ในรูปรากศัพท์ กำหนดเป็นคุณลักษณะ และเมื่อทำการระบุจำนวนคุณลักษณะในเอกสารทดสอบทุกชุดจำนวน 3 ชุดนั้นค่าเอฟก็มีค่าเฉลี่ยเท่ากับ 92.93% ซึ่งสูงกว่าการคำนวณด้วยวิธี TFICF IG และ CHI ที่มีค่าเอฟเฉลี่ยเท่ากับ 84.56% 82.52% และ 85.41% ตามลำดับ

บรรณานุกรม

Brank J., Mladenic D., Grobelnik M. and Milic-Frayling N., "Feature Selection for the Classification of Large Document Collections", Journal of Universal Computer Science vol. 14, no. 10(2008), pp. 1562-1596, 2008.

Caopresp M. F., Matwin S., and Sebastiani F., "A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization," Text Databases and Document Management: Theory and Practice, Chin AG (ed.). Idea Group Publish: Hershey, PA, 2001.

Dahui W., Menfhui L. and Zengru D., "True reason for Zipf's law in language", Physica A: Statistical Mechanics and its Applications, Volume 358, Issues 2-4. 15 December 2005, pp. 545-550.

Fletcher T., "Support vector machine explained", available from http://www.cs.ucl.ac.uk/staff/T.Fletcher , 2009.

Forman G., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research, Volume 3, pp. 1289-1305, 2003.

How B. and Narayanan K., "An Empirical Study of Feature Selection for Text Categorization based on Term Weightage", Proceedings of Web Intelligence International Conference, 20-24 September, 2004.

Jing L., Huang H. and Shi H., "Improved Feature Selection Approach TFIDF in Text Mining", Proceedings of the First International Conference on Machine Learning Cybernetics, Beijing, 4-5 November, 2002.

Keim D., Oelke D. and Rohrdantz C., "Analyzing document collections via context-aware term extraction", In Proceedings of NLDB, -, Saarbrucken, Germany, June 24-26, 2009. pp. 154-168.

Kim J., Huang J., Jung H., and Choi K., "Patent Document Retrieval and Classification at KAIST," Proceeding of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan.

Konchady M., "Text Mining Application Programming", Charles River Media, Inc. Rockland, MA, USA. 2006.

Li S., Xia R., Zong X. and Hunag C., "A Framework of Feature Selection Methods for Text Categorization", Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 692-700, Suntec, Singapore, 2-7 Augst 2009.

Liao C., Alpha S. and Dixon P., "Feature Preparation in Text Categorization", Oracle Corporation, available from http://www.oracle.com/technology/products/text/pdf/feature_preparation.pdf, 2007

Salton G. and McGill M. J., "Introduction to Modern Information Retrieval," McGraw-Hill, New York, USA, 1983.

VenTura D., "SVM Examplae", available from http://axon.cs.byu.edu/Dan/678/miscellaneous/SVM.example.pdf, 2009 .

Wang P., Morgan A. A., Zhanf Q., Sette A., and Peters B., "Automating document classification for the Immune Epitope Database," Journal of BMC Bioinformatics, Volume 8, 2007.

Xiao H., "On the Applicability of Zipf's Law in Chinese Word Frequency Distribution", Journal of Chinese Language and Computing, Volume 18, no 1. pp. 33-46, 2008.