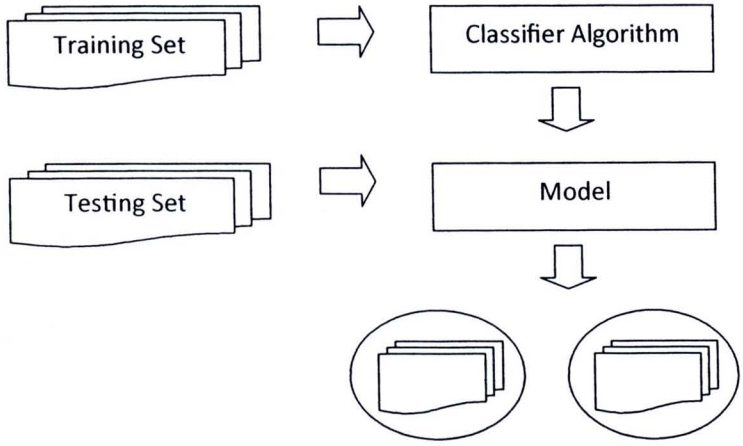
บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

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) นอกจากนนั้นคำที่ถูกตัดออกไปนั้นก็มักจะเป็นคำที่มีความถี่ของการปรากฏในเอกสารมากและพบได้ในเอกสารทุกประเภท

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)

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

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

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

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

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

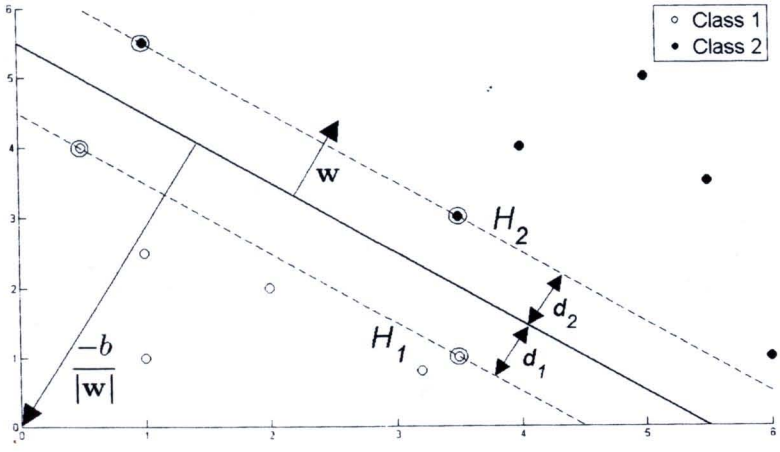
3. ค่าน้ำหนักที่มากที่สุดของเทอมนั้นในจำนวนประเภทเอกสาร จะถูกกำหนดให้เป็นค่าน้ำหนักของเทอมนั้น แสดงดังสูตรที่ 2.11

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

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

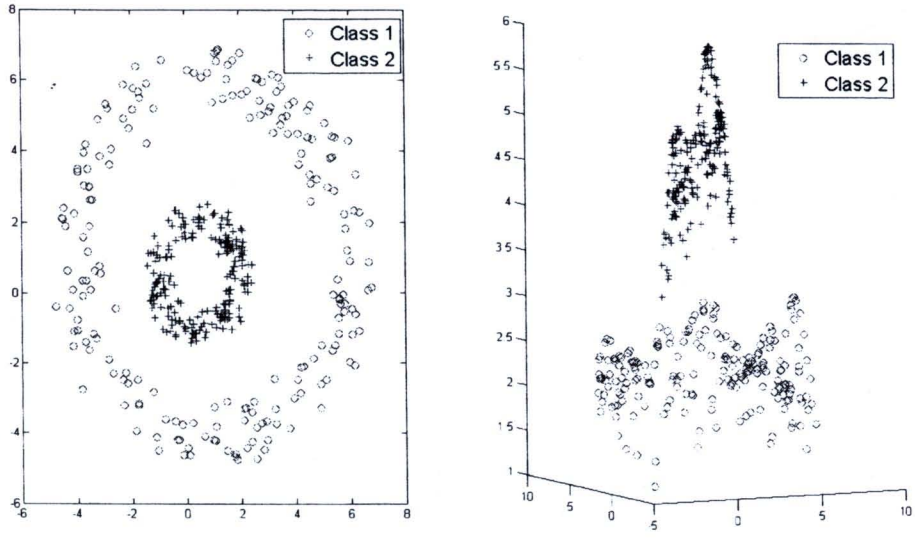
จากที่กล่าวมาข้างต้นเป็นการจัดกลุ่มข้อมูลด้วยไฮเปอร์เพลนในลักษณะเชิงเส้น เพื่อให้อัลกอริทึมนี้สามารถจัดกลุ่มข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นได้นั้น จึงสร้างเมตริกซ์ H จากการ dot product ของข้อมูลนำเข้า (แสดงในรูปแบบเวกเตอร์) ได้ดังสมการ

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

โดยที่ k(xi,xj) เป็นฟังก์ชันที่เรียกว่าฟังก์ชันเคอน์เนล (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



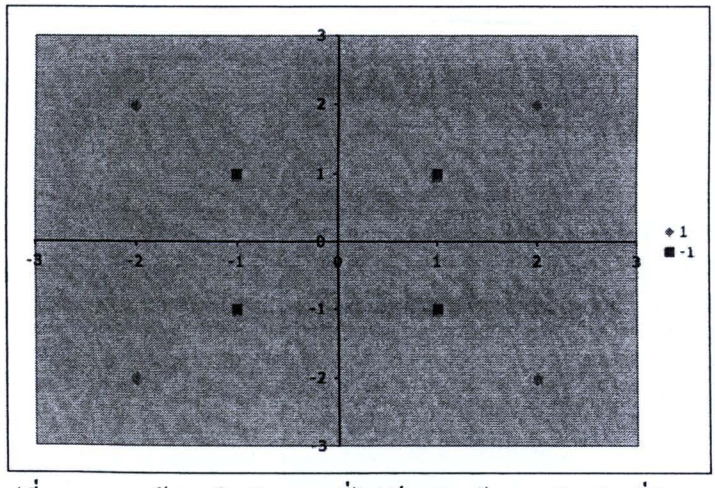
รูปที่ 2.3 การจัดกลุ่มข้อมูลในลักษณะข้อมูลไม่เป็นเชิงเส้น (Fletcher, 2009)

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

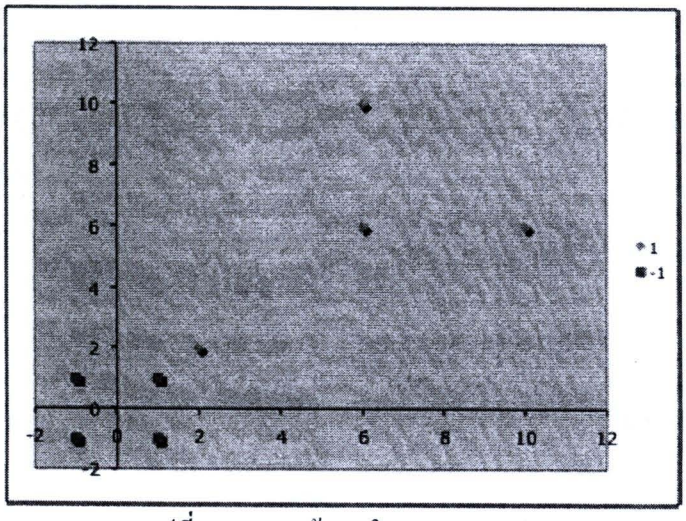


รูปที่ 2.4 แสดงข้อมูลในลักษณะที่ไม่เป็นเชิงเส้นจากตัวอย่างที่กำหนด

จากนั้นทำการหาไฮเปอร์เพลนที่เหมาะสมให้กับข้อมูลทั้งสองกลุ่มนี้ เพื่อแยกข้อมูลทั้งสองกลุ่ม สำหรับข้อมูลในลักษณะที่ไม่เป็นเชิงเส้น นั้นจะต้องปรับข้อมูลจาก input space ให้อยู่ใน feature space ดังนี้

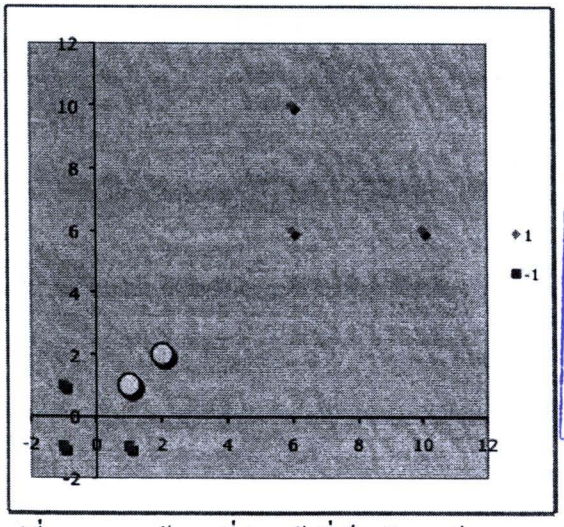
$$\emptyset\begin{pmatrix}x_1\\x_2\end{pmatrix} = \begin{cases}\begin{pmatrix}4-x_2+|x_1-x_2|\\4-x_1+|x_1-x_2|\end{pmatrix} & if\ \sqrt{x_1^2+x_2^2} > 2\\ \begin{pmatrix}x_1\\x_2\end{pmatrix} & otherwise\end{cases}$$

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 คำนวณกับซัพพอร์ทเวกเตอร์ทำการ 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

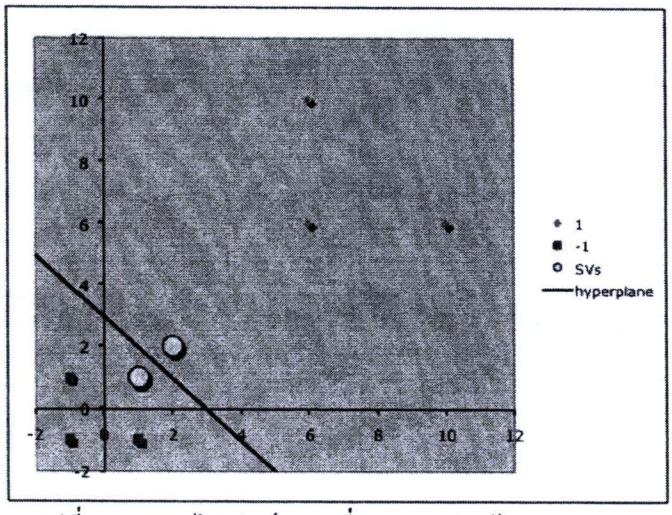
$$\tilde{w} = \sum_i \alpha_i \tilde{s}_i$$

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

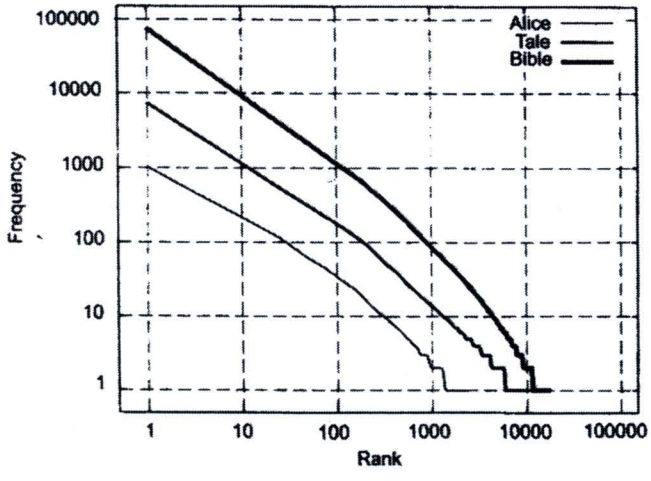
G.K. Zipf (1949) กล่าวว่า เราจะมีการใช้คำที่ไม่มากนักและมักจะเป็นคำที่ซ้ำ ๆ กับที่เคยใช้และมีการใช้คำอื่น ๆ น้อยครั้ง ซึ่งการใช้คำใด ๆ ซ้ำ ๆ นั้นอาจเนื่องจากความคุ้นเคยและการใช้งานคำนั้นบ่อยทำให้ไม่ค่อยได้นำเอาศัพท์ใหม่มาใช้ ในอดีตพบว่ามีคำจำนวนน้อยกว่าร้อยละ 20 ที่สามารถใช้อธิบายข้อความจากข้อความทั้งหมด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

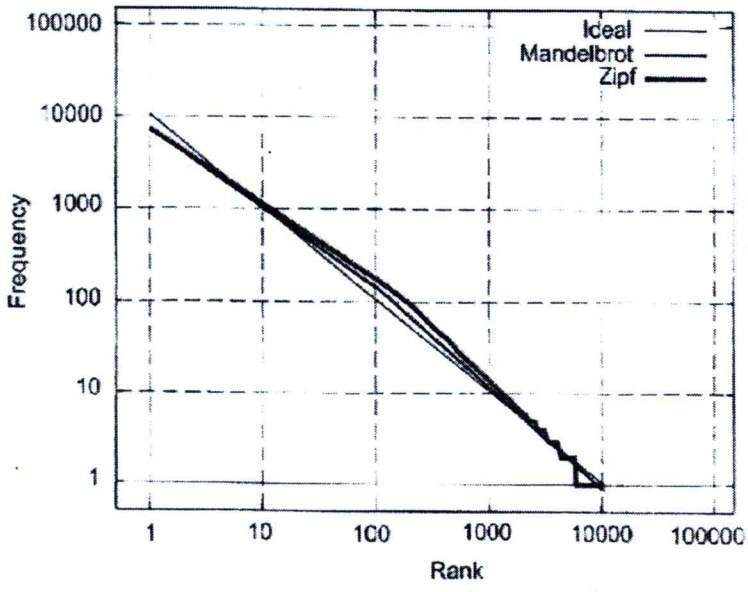
ตารางที่ 2.3 ตารางแสดงค่าลำดับที่ของคำ (Rank) ความถี่ของคำ (Freq) และผลคูณของค่าลำดับกับความถี่

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 มาใช้ในการวิเคราห์การกระจายของคำในเอกสาร 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

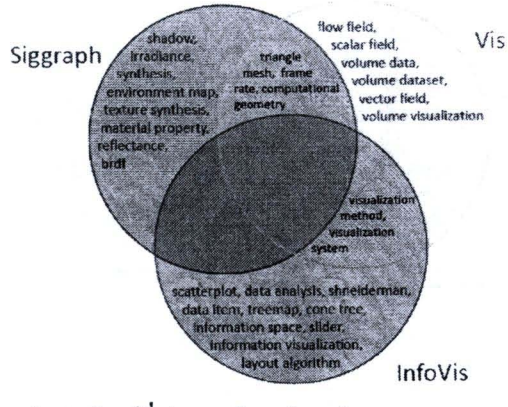
เนื่องด้วยข้อมูลโดยส่วนใหญ่แล้วอยู่ในรูปแบบที่ไม่เป็นโครงสร้าง เช่น เอกสารรายงาน อีเมล์ ข่าว เป็นต้น งานวิจัยทางด้านการจัดกลุ่มเอกสารจึงเป็นงานหนึ่งที่ช่วยให้ผู้ใช้งานสามารถเข้าถึงข้อมูลได้สะดวกขึ้น โดยทำการจัดกลุ่มเอกสารตามลักษณะเนื้อหาหรือข้อความที่คล้ายคลึง ซึ่งระบบจะต้องเรียนรู้การจัด

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 เมื่อคำนวณหาค่าน้ำหนักและเลือกกลุ่มคำสำคัญที่ทำหน้าที่เป็นตัวแทนของประเภทเอกสารเพียงหัวข้อเดียว จำนวนหัวข้อละ 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)

$$CTD(t_k, c_i) = TF(t_k, c_i) \times IDF(t_k, c_i) \times ICF(t_k) \tag{2.24}$$

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

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