บทที่ 3
หลักการของแบบจำลองเวกเตอร์สเปซ

แบบจำลองเวกเตอร์สเปซเป็นการจัดรูปแบบของเอกสารหรือตัวแทนของเอกสาร (Document Representation) เพื่อให้ง่ายต่อการประมวลผลและตีความ ซึ่งรูปแบบจะมีลักษณะของการแสดงความสัมพันธ์ระหว่างคำในเอกสารทั้งหมดด้วยเวกเตอร์ 2 มิติ ซึ่งคำที่ได้นั้นต้องผ่านกระบวนการทำให้เป็นบรรทัดฐาน (Normalization) แล้ว (เช่น การตัดคำหยุดออกไป รวมถึงการให้น้ำหนักกับคำเหล่านั้น) และเอกสารทั้งหมดที่มีอยู่ในฐานข้อมูล รูปแบบเช่นนี้เรียกว่าการจัดรูปแบบแบบจำลองเวกเตอร์สเปซ หรือบางครั้งเรียกตามรูปแบบของเอกสารที่ปรากฏว่าถุงของคำ (Bag of Words)

ตาราง 3.1 การจัดรูปแบบของเอกสารแบบถุงของคำ

	W1	W2	...	Wk	...	Wv
D1	W11	W12	...	W1k	...	W1v
D2	W21	W22	...	W2k	...	W2v
...	...	...	...	...	...	...
Dn	Wn1	Wn2	...	Wnk	...	Wnv

แบบจำลองเวกเตอร์สเปซ เป็นแบบจำลองที่แสดงเนื้อหาของเอกสารในรูปแบบเวกเตอร์ของคำที่ได้รับการปรับน้ำหนักแล้ว โดยเอกสาร (Document) จะถูกแปลงไปเป็นเวกเตอร์ด้วยการแปลงคำแต่ละคำในเอกสารไปเป็นชุดคำที่เหมือน ๆ กัน (Corresponding Word Stem) แล้วทำการปรับน้ำหนักของชุดคำศัพท์ องค์ประกอบที่สอดคล้องกันเหล่านี้จะใช้เป็นตัวแทนของเอกสาร

ความสำเร็จของการใช้แบบจำลองเวกเตอร์สเปซอยู่ที่ความเข้ากันได้ดีในเชิงมิติ (Dimensional Compatibility) คือ การเปรียบเทียบเอกสาร 2 เอกสาร หรือเอกสารกับคำขอจะอยู่บนพื้นฐานของการเปรียบเทียบคำที่เหมือนกันในเอกสารเสมอ

3.1 กระบวนการแบบจำลองเวกเตอร์สเปซ

แบบจำลองเวกเตอร์สเปซ คือ การแทนเอกสาร และข้อสอบถามหรือข้อมูลนำเข้า (Query) ด้วยเวกเตอร์ที่เรียกว่า “เวกเตอร์ของคำ” (Vector of Terms) โดยแกนแต่ละแกนของเวกเตอร์ คือ คำ (Term) ที่ปรากฏอยู่ในเอกสาร คำเหล่านี้ได้จากการคัดเลือกจากคำทั้งหมดที่ปรากฏอยู่ในทุกเอกสาร แล้วบันทึกเก็บไว้ในพจนานุกรมหรือฐานข้อมูล เพื่อใช้อ้างอิงในการสร้างเวกเตอร์ของคำที่ตรงกันทั้งฐานข้อมูล และความยาวของเวกเตอร์ คือ ค่าความถี่ของคำที่ถูกเลือกซึ่งปรากฏอยู่ในเอกสารนั้นประกอบร่วมกัน

แบบจำลองเวกเตอร์สเปซเป็นรูปแบบหนึ่งในการแสดงเอกสารผ่านทางคำที่เอกสารนั้น ๆ มีคำนั้นอยู่ โดยวิธีนี้จะทำการเปรียบเทียบจากคำที่เป็นข้อสอบถามว่ามีความคล้ายคลึงกับเอกสารที่มีอยู่ส่วนใดบ้าง ส่วนวิธีการกระทำนั้นจะทำโดยการนำคำในเอกสารแต่ละเอกสารมาทำการแยกออกเป็นตารางความถี่ของคำ โดยจะเรียกตารางความถี่ของคำว่าเวกเตอร์ (Vector) และทำการจัดเก็บไว้ในอะเรย์ จากนั้นจะนำเอาเวกเตอร์ซึ่งเป็นตัวแทนของเอกสารนั้น ๆ มาทำการเทียบกับคำศัพท์ โดยคำศัพท์นั้นจะถูกสร้างขึ้นมาจากคำทุกคำในเอกสารทั้งหมดในระบบ

ตัวอย่างเช่น ในเอกสาร A มีคำว่า “การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้หลักการประมวลผลภาษาธรรมชาติ” และในเอกสาร B มีคำว่า “ความคล้ายคลึงของเอกสารภาษาอังกฤษ” สามารถที่จะนำมาแยกออกเป็นตารางความถี่ของคำได้ตามตาราง 3.2 และตาราง 3.3

เอกสาร A “การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้หลักการประมวลผลภาษาธรรมชาติ”

เมื่อทำการตัดคำจะได้เป็น การ|วัด|ความ|คล้ายคลึง|ของ|เอกสาร|ภาษา|ไทย|โดย|ใช้|หลัก|การ|ประมวลผล|ภาษา|ธรรมชาติ

ตาราง 3.2 แสดงความถี่ของคำในเอกสาร A

การ	วัด	ความ	คล้ายคลึง	ของ	เอกสาร	ภาษา	ไทย
2	1	1	1	1	1	1	1
โดย	ใช้	หลัก	ประมวลผล		ธรรมชาติ		
2	1	1	1		1	1	1

เอกสาร B “ความคล้ายคลึงของเอกสารภาษาอังกฤษ”

เมื่อทำการตัดคำจะได้เป็น ความ|คล้ายคลึง|ของ|เอกสาร|ภาษา|อังกฤษ

ตาราง 3.3 แสดงความถี่ของคำในเอกสาร B

ความ	คล้ายคลึง	ของ	เอกสาร	ภาษา	อังกฤษ
1	1	1	1	1	1

โดยมีคำศัพท์ในระบบคือ การ วัด ความ คล้าย คลึง ของ เอก สาร ภาษา ไทย โดย ใช้ หลัก ประมวล ผล ธรรมชาติ และ อังกฤษ ต่อจากนั้นก็จะนำเอกสารทั้ง A และ B มาทำเป็นเวกเตอร์ เพื่อที่จะนำมาทำการเปรียบเทียบกับคำศัพท์ที่มีในระบบ

เอกสาร A เมื่อนำมาทำเป็นเวกเตอร์

เวกเตอร์ของเอกสาร A: (2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 0)

ตาราง 3.4 แสดงค่าของเอกสาร A เมื่อนำมาทำเป็นเวกเตอร์

การ	วัด	ความ	คล้ายคลึง	ของ	เอกสาร	ภาษา	ไทย	โดย
2	1	1	1	1	1	2	1	1
ใช้	หลัก	ประมวลผล		ธรรมชาติ	อังกฤษ			
1	1	1		1	0			

เอกสาร B เมื่อนำมาทำเป็นเวกเตอร์

เวกเตอร์ของเอกสาร B: (0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1)

ตาราง 3.5 แสดงค่าของเอกสาร B เมื่อนำมาทำเป็นเวกเตอร์

การ	วัด	ความ	คล้ายคลึง	ของ	เอกสาร	ภาษา	ไทย	โดย
0	0	1	1	1	1	1	0	0
ใช้	หลัก	ประมวลผล		ธรรมชาติ	อังกฤษ			
0	0	0		0	1			

เมื่อมีข้อสอบถามเข้ามาในระบบ ก็จะนำคำที่ได้จากข้อสอบถามมากระทำเป็นเวกเตอร์ เช่นเดียวกับเอกสาร A และเอกสาร B แล้วจึงนำเวกเตอร์ของข้อสอบถามที่ได้นั้นมาทำการเปรียบเทียบกันว่าเอกสารใดมีความคล้ายคลึงกับคำที่ได้จากข้อสอบถาม

โดยสมมุติให้ข้อสอบถามที่เข้ามาเป็น “การประมวลผลเอกสารภาษาอังกฤษ”

เมื่อทำการตัดคำจะได้เป็น การ|ประมวลผล|เอกสาร|ภาษา|อังกฤษ

ข้อสอบถามเมื่อนำมาทำเป็นเวกเตอร์

เวกเตอร์ของข้อสอบถาม : (1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1)

ตาราง 3.6 แสดงค่าของข้อสอบถามเมื่อนำมาทำเป็นเวกเตอร์

การ	วัด	ความ	คล้ายคลึง	ของ	เอกสาร	ภาษา	ไทย	โดย
1	0	0	0	0	1	1	0	0

ใช้	หลัก	ประมวลผล	ธรรมชาติ	อังกฤษ
0	0	1	0	1

จากนั้นจึงนำเวกเตอร์ที่ได้มาทำการเปรียบเทียบกันในขอบเขตเชิงสถิติเวกเตอร์โดยข้อสอบถามจะถูกเปรียบเทียบกับแต่ละเวกเตอร์เอกสาร โดยการคำนวณผลคูณภายในของทั้ง 2 เวกเตอร์จะได้ว่าจำนวนคำที่ตรงกับในเอกสารเวกเตอร์ของข้อสอบถาม

ข้อสอบถาม (1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1)

เวกเตอร์ A (2, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 0)

ข้อสอบถาม * เวกเตอร์ A = (1*2) + (0*1) + (0*1) + (0*1) + (0*1) + (1*1) + (1*2) + (0*1) + (0*1) + (0*1) + (0*1) + (1*1) + (0*1) + (1*0)

= 6

ข้อสอบถาม (1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1)

เวกเตอร์ B (0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1)

ข้อสอบถาม * เวกเตอร์ B = (1*0) + (0*0) + (0*1) + (0*1) + (0*1) + (1*1) + (1*1) + (0*0) + (0*0) + (0*0) + (0*0) + (1*0) + (0*0) + (1*1)

= 3

จากตัวอย่างจะเห็นได้ว่าเอกสาร A จะตรงกับข้อสอบถามมากกว่าเอกสาร B เพราะว่าผลคูณภายในมีค่าสูงกว่า

รูปแบบการทำงานของแบบจำลองเวกเตอร์สเปซ

1) ให้ความสำคัญกับความถี่ของคำที่ปรากฏอยู่ในเอกสารและความถี่มีผลต่อการให้น้ำหนักของคำในเอกสาร โดยที่

- การนับความถี่ของคำ (Term Frequency : TF) คือ การใช้ความถี่ของคำ เช่น พบ 1 ครั้ง เรียกว่า เทอม (Term) ทั้งนี้ขึ้นอยู่กับจำนวนคำของเอกสาร โดยเทอมจะแทนคำศัพท์ของแต่ละคำ

- ค่าน้ำหนักคำ (Term Weight : W) ความถี่ของคำ ๆ หนึ่งที่พบในทุก ๆ เอกสาร

2) การสร้างตารางความถี่ของคำเป็นขั้นตอนในการหาความถี่ของคำศัพท์ และกำหนดค่าน้ำหนักของคำศัพท์ในแต่ละเอกสารลงในตารางความถี่ของคำ

3) การนำเสนอข้อมูลในเชิงเวกเตอร์ ซึ่งคำศัพท์ในตารางความถี่ของคำจะถูกนำเสนอในเชิงเวกเตอร์ โดยจะถูกมองเป็นอะเรย์ของเวกเตอร์ เช่น (3, 1, 0, 0, 0)

4) การคำนวณค่าความเหมือน โดยคุณสมบัติของเวกเตอร์ทำให้สามารถคำนวณค่าความคล้ายคลึงระหว่างเวกเตอร์ได้จากค่าสัมประสิทธิ์โคไซน์ของมุมระหว่างคู่เวกเตอร์ ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1

5) สามารถจัดลำดับ (Ranking) ของเอกสาร โดยใช้เกณฑ์ความสำคัญของคำและการเข้ากันได้ของคำ

ข้อดีของแบบจำลองเวกเตอร์สเปซ คือ ใช้คณิตศาสตร์เรียบง่ายในการคิด มีการพิจารณาจากความถี่ของคำ และสามารถจัดลำดับของเอกสารได้ และสามารถใช้กับเอกสารที่มีข้อมูลมาก ๆ ได้ดี แบบจำลองเวกเตอร์เป็นการจับคู่แบบไม่เที่ยงตรง (Inexact Match) เช่นเดียวกับ แบบจำลองความน่าจะเป็น (Probabilistic Model) ทำให้สามารถค้นหาแบบเป็นบางส่วน (Partial Search) ได้

ข้อเสียของแบบจำลองเวกเตอร์สเปซ คือ แบบจำลองเวกเตอร์สเปซจะไม่สนใจในความหมายของคำ วลี โครงสร้างของคำ หรือคำที่มีความหมายเหมือนกัน เพราะแบบจำลองเวกเตอร์สเปซอยู่บนข้อสมมุติที่ว่าคำทุกคำไม่มีความสัมพันธ์กัน ซึ่งในความเป็นจริงคำจะมีความสัมพันธ์หรืออ้างอิงกันในเชิงความหมาย

3.2 การให้ค่าน้ำหนักคำ (Term Weighting)

คำที่มีความสำคัญหรือใช้เป็นตัวแทนของเอกสารที่ดี ควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยมากในชุดเอกสารที่เหลือทั้งหมด แต่ถ้าคำนั้นปรากฏเป็นจำนวนมากในทุก ๆ เอกสาร แสดงว่าคำดังกล่าวไม่สามารถเป็นตัวแทนของเอกสารใดๆ ได้ ซึ่งคำเหล่านี้เรียกว่า คำหยุด (Stop Word) เช่น คำว่า “ที่” “และ” “ซึ่ง” เป็นต้น ดังนั้นการให้ค่า

น้ำหนักคำ ๆ หนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากการนับความถี่ของคำที่ปรากฏในเอกสารนั้น และจำนวนของเอกสารทั้งหมดที่มีคำ ๆ นั้นปรากฏอยู่

ธีรศักดิ์ สังข์ศรี (2548) กล่าวไว้ว่า การให้ค่าน้ำหนักคำ ๆ หนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากการนับความถี่ของคำที่ปรากฏในเอกสารนั้น ๆ และจำนวนของเอกสารทั้งหมดที่มีคำ ๆ นั้นปรากฏอยู่

วิธีการให้ค่าน้ำหนักคำที่นิยมใช้กันมากคือ การกำหนดค่าน้ำหนักคำในเอกสารตามแนวคิดของซอลตั่น (1989) ซึ่งจะประกอบด้วยกระบวนการ 2 กระบวนการ คือ

1) การนับความถี่ของคำ ถ้าคำที่ปรากฏยิ่งบ่อยในเอกสารจะยิ่งมีความสำคัญต่อเอกสารนั้นมากยกเว้นคำหยุด การนับความถี่ของคำเป็นการนับจำนวนครั้งของคำที่คำนั้นปรากฏในเอกสาร เช่น ถ้าพบคำว่า “เวกเตอร์” ในเอกสาร 10 ครั้ง ค่าการนับความถี่ของคำจะมีค่าเท่ากับ 10 การนับความถี่ของคำนั้นยังเป็นการปรับปรุงการวัดค่าความถูกต้องของข้อมูลอีกด้วย

2) การหาค่าส่วนกลับความถี่ของเอกสาร (Inverse Document Frequency: IDF) เป็นการช่วยปรับปรุงการวัดค่าความแม่นยำของข้อมูล สูตรคำนวณการหาค่าส่วนกลับความถี่ของเอกสารตามสมการ (3.1)

$$idf = \log\left(\frac{N}{df}\right) \qquad \text{..... (3.1)}$$

โดยที่ idf คือ ค่าส่วนกลับความถี่ของเอกสาร

N คือ จำนวนเอกสารในชุดเอกสารทั้งหมด

df คือ ค่าความถี่ของเอกสาร (Document Frequency) ของคำแต่ละคำ ถ้าคำใดปรากฏในเอกสารทุกฉบับจะมีค่าเท่ากับ N ซึ่งทำให้ค่าส่วนกลับความถี่ของเอกสารมีค่าเท่ากับศูนย์

เมื่อรวมปัจจัยหรือกระบวนการทั้ง 2 อย่างนี้เข้าด้วยกัน สามารถนำไปคำนวณหาค่าน้ำหนักคำได้ดังสมการ (3.2)

$$w = tf * idf \quad \text{หรือ} \quad w = tf * \log\left(\frac{N}{df}\right) \qquad \text{..... (3.2)}$$

ตัวอย่าง เช่น มีเอกสารอยู่ 3 เอกสาร ได้แก่

D1: “การวัดความคล้ายคลึงของเอกสารภาษาไทยโดยใช้หลักการประมวลผลภาษาธรรมชาติ”

เมื่อทำการตัดคำจะได้เป็น การ|วัด|ความ|คล้ายคลึง|ของ|เอกสาร|ภาษา|ไทย|โดย|ใช้|หลัก|การ|ประมวลผล|ภาษา|ธรรมชาติ

D2: “ความคล้ายคลึงของเอกสารภาษาอังกฤษ”

เมื่อทำการตัดคำจะได้เป็น ความ|คล้ายคลึง|ของ|เอกสาร|ภาษา|อังกฤษ

D3: “การประมวลผลเอกสารภาษาอังกฤษ”

เมื่อทำการตัดคำจะได้เป็น การ|ประมวลผล|เอกสาร|ภาษา|อังกฤษ

และมีข้อสอบถามที่เข้ามาคือ

Q: “เอกสารภาษาธรรมชาติ”

เมื่อทำการตัดคำจะได้เป็น เอกสาร|ภาษา|ธรรมชาติ

สามารถที่จะนำเอกสารที่มีทั้งหมดมาทำเป็นดัชนีคำศัพท์ (Index Term) เพื่อที่จะนำไปใช้ในการนับความถี่ของคำและค่าส่วนกลับความถี่ของเอกสารได้

ตาราง 3.7 แสดงคำในเอกสารทั้งหมด เมื่อนำมาทำเป็นดัชนีคำศัพท์

เทอม	คำ
T1	การ
T2	ของ
T3	คล้ายคลึง
T4	ความ
T5	ใช้
T6	โดย
T7	ไทย
T8	ธรรมชาติ
T9	ประมวลผล
T10	ภาษา
T11	วัด
T12	หลัก
T13	อังกฤษ
T14	เอกสาร

จะได้การนับความถี่ของคำได้ตามตาราง 3.8

ตาราง 3.8 แสดงความถี่ของคำแต่ละคำที่พบในเอกสารแต่ละเอกสาร

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
D1	2	1	1	1	1	1	1	1	1	2	1	1	0	1
D2	0	1	1	1	0	0	0	0	0	1	0	0	1	1
D3	1	0	0	0	0	0	0	0	1	1	0	0	1	1

เมื่อนำการนับความถี่ของคำมาเขียนในรูปแบบเมตริกซ์จะได้ตามรูป 3.1

$$TF_{ij} = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

รูป 3.1 การนับความถี่ของคำเมื่อนำมาเขียนในรูปเมตริกซ์

จะได้ค่าความถี่ของเอกสาร หรือความถี่ของเอกสารที่พบของดัชนีคำศัพท์แต่ละคำในเอกสาร เพื่อที่จะนำไปใช้ในการหาค่าส่วนกลับความถี่ของเอกสารต่อไป โดยแสดงในรูปของเมตริกซ์ตามรูป 3.2

$$DF_1^1 = [2 \quad 2 \quad 2 \quad 2 \quad 1 \quad 1 \quad 1 \quad 1 \quad 2 \quad 3 \quad 1 \quad 1 \quad 2 \quad 3]$$

รูป 3.2 ความถี่ของเอกสารที่พบของดัชนีคำศัพท์แต่ละคำในเอกสาร

หลังจากนั้นจึงนำค่าที่ได้จากการหาค่าความถี่ของเอกสารมาทำการหาค่าส่วนกลับความถี่ของเอกสาร โดยมีขั้นตอนดังนี้

(ก) ถ้าดัชนีคำศัพท์หนึ่ง ๆ ปรากฏเพียง 1 ครั้งในเอกสารทั้ง 3 จะได้

$$idf = \log\left(\frac{N}{df}\right)$$
$$= \log\left(\frac{3}{1}\right)$$
$$= 0.477$$

(ข) ถ้าดัชนีคำศัพท์หนึ่ง ๆ ปรากฏเพียง 2 ครั้งในเอกสารทั้ง 3 จะได้

$$idf = \log\left(\frac{N}{df}\right)$$
$$= \log\left(\frac{3}{2}\right)$$
$$= 0.176$$

(ค) ถ้าดัชนีคำศัพท์หนึ่งๆ ปรากฏเพียง 3 ครั้งในเอกสารทั้ง 3 จะได้

$$idf = \log\left(\frac{N}{df}\right)$$
$$= \log\left(\frac{3}{3}\right)$$
$$= 0$$

และหลังจากคำนวณค่าทั้งหมดเสร็จสามารถนำมาเขียนในรูปแบบของเมตริกซ์ ได้ตามรูป 3.3

IDF_i = [0.176 0.176 0.176 0.176 0.477 0.477 0.477 0.477 0.176 0.0 0.477 0.477 0.176 0.0]

รูป 3.3 ค่าที่ได้หลังจากการคำนวณหาค่าส่วนกลับความถี่ของเอกสารแล้ว

หลังจากนั้นจึงนำค่าการนับความถี่ของคำและค่าส่วนกลับความถี่ของเอกสารที่ได้มาทำการหาค่าน้ำหนักคำแต่ละคำในเอกสารตามรูป 3.4

$$W_{ij} = [0.176 \quad 0.176 \quad 0.176 \quad 0.176 \quad 0.477 \quad 0.477 \quad 0.477 \quad 0.477 \quad 0.176 \quad 0.0 \quad 0.477 \quad 0.477 \quad 0.176 \quad 0.0] * \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

รูป 3.4 การหาค่าน้ำหนักคำแต่ละคำในเอกสาร

หลังจากการคำนวณหาค่าน้ำหนักคำแต่ละคำในเอกสารแล้ว จะได้ค่าน้ำหนักคำแต่ละคำตามตาราง 3.9

ตาราง 3.9 ค่าที่ได้หลังจากการคำนวณค่าน้ำหนักคำในเอกสารแล้ว

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
D1	0.352	0.176	0.176	0.176	0.477	0.477	0.477
D2	0	0.176	0.176	0.176	0	0	0
D3	0.176	0	0	0	0	0	0
	T8	**T9**	**T10**	**T11**	**T12**	**T13**	**T14**
D1	0.477	0.176	0	0.477	0.477	0	0
D2	0	0	0	0	0	0.176	0
D3	0	0.176	0	0	0	0.176	0

จากนั้นจึงนำข้อสอบถามที่ทำการรับเข้ามาในตอนแรกมาทำการเข้ากระบวนการเช่นเดียวกับคำในเอกสารในการหาค่าน้ำหนักคำ ตั้งแต่ต้นจนกระทั่งได้เป็นค่าน้ำหนักคำออกมาตามตาราง 3.10

ตาราง 3.10 ค่าที่ได้หลังจากการคำนวณค่าน้ำหนักคำของข้อสอบถามที่รับเข้ามา

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Q	0	0	0	0	0	0	0
	T8	**T9**	**T10**	**T11**	**T12**	**T13**	**T14**
Q	0.477	0	0	0	0	0	0

3.3 การวัดความคล้ายคลึง (Similarity Measure)

การวัดความคล้ายคลึงของเอกสารนั้นสามารถจำแนกได้เป็น 2 ประเภท คือ

(1) การวัดความคล้ายคลึงจากการมีหรือไม่มีคำสำคัญในเอกสาร

(2) การวัดความคล้ายคลึงจากความถี่ของคำสำคัญที่พบในเอกสาร

สำหรับการค้นคว้าอิสระนี้จะสนใจเฉพาะในส่วนของการวัดความคล้ายคลึงของเอกสารจากความถี่ของคำสำคัญที่เกิดขึ้น โดยที่เอกสารใด ๆ ก็ตามสามารถนำเสนอในรูปแบบของเวกเตอร์หรือรายการคำสำคัญซึ่งเกิดขึ้นในเอกสาร โดยกำหนดให้เวกเตอร์แสดงถึงทุก ๆ คำสำคัญที่มีอยู่

ขั้นตอนวิธีการพื้นฐานก่อนที่จะทำการจัดกลุ่มเอกสารหรือการค้นคืนเอกสาร จำเป็นต้องศึกษาถึงการเลือกขั้นตอนวิธี ทั้งนี้เพราะคุณลักษณะของข้อมูลมีผลต่อการจัดกลุ่มเอกสารและการค้นคืนเอกสาร และที่สำคัญที่สุดคือวิธีการที่ใช้ในการวัดค่าความคล้ายคลึง ซึ่งสามารถจำแนกค่าความคล้ายคลึงออกเป็น 3 กลุ่มใหญ่ ๆ คือ สัมประสิทธิ์ระยะทาง (Distance Coefficients) สัมประสิทธิ์ความสัมพันธ์ (Association Coefficients) และสัมประสิทธิ์ความน่าจะเป็น (Probabilistic Coefficients)

ตาราง 3.11 ตัวอย่างสัมประสิทธิ์ความสัมพันธ์แบบต่าง ๆ

วิธีวัดค่าความเหมือน	การวัดค่าแบบไบนารี่เวกเตอร์	การวัดค่าแบบให้น้ำหนักเวกเตอร์
Inner Product	$X \cap Y$	$\sum_{i=1}^{t} X_i \bullet Y_i$
Dice Coefficient	$2\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$	$\frac{2\sum_{i=1}^{t} X_i Y_i}{\sum_{i=1}^{t} X_i^2 + \sum_{i=1}^{t} Y_i^2}$
Cosine Coefficient	$\left\lvert \frac{\lvert X \cap Y \rvert}{\lvert X \rvert^{\frac{1}{2}} + \lvert Y \rvert^{\frac{1}{2}}} \right\rvert$	$\frac{\sum_{i=1}^{t} X_i Y_i}{\sqrt{\sum_{i=1}^{t} X_i^2 + \sum_{i=1}^{t} Y_i^2}}$
Jacquard Coefficient	$\frac{\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert - \lvert X \cap Y \rvert}$	$\frac{\sum_{i=1}^{t} X_i Y_i}{\sum_{i=1}^{t} X_i^2 + \sum_{i=1}^{t} Y_i^2 - \sum_{i=1}^{t} X_i Y_i}$

การวัดความคล้ายคลึงของเอกสารจะอาศัยค่าน้ำหนักคำหลักที่มีอยู่ในแต่ละเอกสารและค่าน้ำหนักคำของข้อสอบถาม โดยค่าน้ำหนักคำหลักในเอกสารคำนวณได้จากขั้นตอนการเตรียมข้อมูล ส่วนค่าน้ำหนักคำของข้อสอบถามจะคำนวณได้จากขั้นตอนการเปรียบเทียบเอกสารในแต่ละครั้ง

เมื่อคำนวณได้ค่าน้ำหนักคำหลักในเอกสารและค่าน้ำหนักคำของข้อสอบถามได้แล้วจะนำไปแทนที่ในสมการ (3.3) เพื่อคำนวณหาค่าความคล้ายคลึงของเอกสาร โดยใช้วิธีแบบเวกเตอร์ (Vector Based Similarity Measurement)

$$sim(Q, D_j) = \frac{\sum_{\forall i}(w_{Q,j} \bullet w_{i,j})}{\sqrt{\sum_{\forall j} w_{Q,j}^{\ 2}}\sqrt{\sum_{\forall i} w_{i,j}^{\ 2}}} \quad \text{..... (3.3)}$$

เมื่อ

$sim(Q, D_j)$ คือ ค่าความคล้ายคลึงระหว่างข้อสอบถามกับเอกสาร j

$\sum_{\forall i}(w_{Q,j} \bullet w_{i,j})$ คือ ค่าผลรวมของ Cross Product ระหว่างค่าน้ำหนักคำของข้อสอบถาม Q กับค่าน้ำหนักคำหลัก i ในเอกสาร j

$\sqrt{\sum_{\forall j} w_{Q,j}^{\ 2}}$ คือ ผลรวมของค่าน้ำหนักทั้งหมดของข้อสอบถาม Q

$\sqrt{\sum_{\forall i} w_{i,j}^{\ 2}}$ คือ ผลรวมของค่าน้ำหนักทั้งหมดของคำหลัก i ในเอกสาร j

3.3.1 การวัดความคล้ายคลึงเชิงมุม (Cosine Similarity Measurement)

การวัดความคล้ายคลึงของเอกสารด้วยวิธีการวัดความคล้ายคลึงเชิงมุมนั้นเป็นวิธีการเปรียบเทียบความคล้ายคลึงของเอกสารสองเอกสาร โดยแต่ละเอกสารจะถูกแทนด้วยเวกเตอร์ขนาดเอ็น (N-Dimensional Vector) ซึ่งเก็บค่าน้ำหนักคำแต่ละคำในเอกสารนั้น (N-Dimensional Vector in Term Space) การเปรียบเทียบความคล้ายคลึงของเอกสารจะเปรียบเทียบโดยดูจากมุมโคไซน์ของมุมระหว่าง 2 เวกเตอร์ของเอกสาร หากเอกสารทั้งสองเอกสารคล้ายคลึงกันมาก เวกเตอร์ของเอกสารทั้ง 2 จะทับกันเกือบสนิท มุมจึงมีค่าน้อย ค่าโคไซน์ที่ได้จะมีค่ามาก ตัวอย่างในการวัดค่าความคล้ายคลึงระหว่างเวกเตอร์ d และเวกเตอร์ d' จะมีค่าดังสมการ (3.4)

$$Similarity = \frac{d * d'}{|d||d'|} \quad \text{..... (3.4)}$$

โดย $d * d'$ คือ เวกเตอร์ของโปรดักซ์ (Product)

ตัวอย่างเช่น

$$d = (2, 1, 1, 1, 0)$$

$$d' = (0, 0, 0, 1, 0)$$

$$d * d' = (2 * 0) + (1 * 0) + (1 * 0) + (1 * 1) + (0 * 0)$$

$$= 1$$

$$|d| = \sqrt{2^2 + 1^2 + 1^2 + 1^2 + 0^2}$$
$$= \sqrt{7}$$
$$= 2.646$$
$$|d'| = \sqrt{0^2 + 0^2 + 0^2 + 1^2 + 0^2}$$
$$= \sqrt{1}$$
$$= 1$$

ค่าความคล้ายคลึง $= \frac{1}{1 * 2.646}$

$= 0.378$

ทั้งนี้ค่าความคล้ายคลึงสูงสุดที่วัดด้วยวิธีนี้จะมีค่าเท่ากับ 1 โดยมีความหมายคือ เวกเตอร์ทั้งสองทำมุมระหว่างกัน 0 องศา นั้นคือเวกเตอร์ทั้งสองมีทิศทางเดียวกัน