บทที่ 3
การแบ่งกลุ่มข้อมูลอัตโนมัติ

วัตถุประสงค์

1. เพื่อให้นักศึกษาทราบถึงวิธีการวัดความคล้ายคลึงกันของเอกสาร
2. เพื่อให้นักศึกษาทราบถึงวิธีการวัดความไม่คล้ายคลึงกันของเอกสาร
3. เพื่อให้นักศึกษาทราบถึงวิธีการจัดแบ่งกลุ่ม
4. เพื่อให้นักศึกษาทราบถึงการใช้วิธีคลัสเตอร์ในระบบค้นคืนสารสนเทศ
5. เพื่อให้นักศึกษาทราบถึง Clustering algorithm

สารบัญ

3.1 บทนำ

ในบทนี้จะกล่าวถึง การแบ่งกลุ่มข้อมูลอัตโนมัติ (Automatic Classification) ในระบบค้นคืนสารสนเทศเน้นการใช้โปรแกรมประยุกต์ที่พัฒนาขึ้นเพื่อใช้ในการแบ่งกลุ่มเอกสาร (Document Clustering) ซึ่งความคิดของการแบ่งกลุ่มนี้สามารถนำไปประยุกต์ใช้กับเรื่องของการจดจำรูปแบบ(pattern recognition) การวินิจฉัยทางการแพทย์ (automatic medical diagnosis) การจัดกลุ่มของคำสำคัญ(Keyword Clustering) ได้ สำหรับการแบ่งกลุ่มข้อมูลอัตโนมัติในระบบ IR นั้นเรามีการนำมาประยุกต์ใช้ใน 2 ส่วนหลักคือ การจัดกลุ่มคำสำคัญ (Keyword clustering)และการจัดกลุ่มของเอกสาร (document clustering) เป็นหลัก R.M.Hayes ได้กล่าวไว้ว่า

“โครงสร้างของการจัดหมวดหมู่ จะเป็นการรวมกลุ่มของไอเท็ม(item) ซึ่งอาจเป็นเอกสารหรือตัวแทนเอกสารเข้าด้วยกัน ซึ่งไอเท็มเหล่านี้จะถูกจัดให้รวมกลุ่มเปรียบเสมือนให้เป็นหน่วยหนึ่ง ซึ่งไอเท็มเหล่านี้อาจจะสูญเสียคุณสมบัติเฉพาะไปบ้างไม่มากก็น้อย”

อาจกล่าวได้ว่า ถ้ามีการนำเอกสารใดๆมาจัดให้อยู่ในกลุ่มเดียวกัน ถือว่าเอกสารต่างๆเหล่านั้นมีความคล้ายคลึงกันมาก ถ้าจะแยกแยะเอกสารในกลุ่มต้องมีการตรวจสอบเป็นไอเท็มๆโดยใช้ความสัมพันธ์ทางตรรกศาสตร์ (logical Relationship) ถึงจะแยกเอกสารออกจากกลุ่มได้

โครงสร้างทางตรรกศาสตร์ (Logical Organization) แบ่งเป็น 2 วิธีคือ

1. การแบ่งเอกสารโดยตรง ได้รับการพิสูจน์ว่ายากทางทฤษฎี ไม่น่าเชื่อถือ
2. การแบ่งเอกสารโดยผ่านการคำนวณของวิธีวัดความใกล้เคียงหรือความคล้ายคลึงกันของเอกสาร นิยมใช้มากในปัจจุบัน ในการสร้างกลุ่มของเอกสารนั้นจะทำการวัดความสัมพันธ์กันโดยใช้วิธีการเปรียบเทียบเอกสารโดยอัตโนมัติซึ่งเอกสารในกลุ่มเดียวกันจะมีความใกล้เคียงกันหรือความคล้ายคลึงกันมาก กลุ่มของเอกสารแต่ละกลุ่มจะถูกแทนด้วยเวกเตอร์กลุ่ม (Group Vector) ซึ่งแต่ละกลุ่มจะมีเวกเตอร์กลุ่มแตกต่างกัน และการค้นหาเอกสารต่างๆเหล่านี้จะกระทำโดยเปรียบเทียบกับข้อคำถาม(query) ซึ่งจะกระทำได้โดยตรวจสอบกับทุกๆเวคเตอร์กลุ่ม เอกสารกลุ่มใดที่ได้คะแนนสูงสุดจากการเปรียบเทียบกับข้อคำถาม(query) จะถูกดึงออกมาให้แก่ผู้ร้องขอ

ส่วนมากแล้ว ระบบค้นคืนสารสนเทศขึ้นอยู่กับแนวคิดที่ว่า คำหรือเทอมที่คล้ายคลึงกันจะมีความเกี่ยวพันธ์ในการค้นหาได้จากการสอบถามเดียวกัน(query) โดยปกติแล้วการค้นหาเอกสารที่ต้องการนั้นจะมีการจับคู่กับคำของข้อสอบถาม(query) ความคล้ายคลึงกันนั้นสามารถวัดได้หลายวิธีเช่น String Matching / comparison , Same vocabulary used , Probability that documents arise from same model , Same meaning of text เป็นต้น

3.2 วิธีวัดความคล้ายคลึง (Measures of Association)

ในระบบค้นคืนสารสนเทศมีวิธีการวัดความคล้ายคลึง หรือความสัมพันธ์ของเอกสารเด่นๆอยู่หลายวิธี

สมมุติว่า เอกสาร และข้อคำถาม นั้นถูกแสดงแทนโดย รายการของดรรชนีซึ่งเป็นขนาดของเซ็ตของดรรชนีที่ใช้แทนสิ่งหนึ่ง (object) วิธีการวัดความคล้ายคลึงกันสามารถหาได้หลายวิธีดังนี้

1. Simple Coefficient

เป็นรูปแบบที่ง่ายที่สุด คือ $|X \cap Y|$

เป็นจำนวนของคำดรรชนีที่อยู่ทั้งในเอกสาร X และ Y วิธีนี้ไม่ได้นำเอาขนาดของ X และ Y มาพิจารณา

สมมุติ $X = \{1, 2, 3\}$

$Y = \{1, 4\}$

$X \cap Y = \{1\}$

$|X \cap Y| = 1$

2. Dice's Coefficient

คือ $2|X \cap Y| \quad / \quad |X| + |Y|$

วิธีนี้มีการนำเอาขนาดของ X และ Y มาพิจารณา

3. Jaccard's Coefficient

คือ $|X \cap Y| \quad / \quad |X \cup Y|$

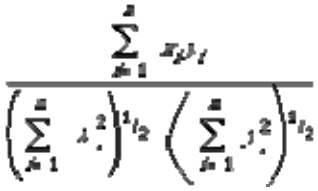
4. Cosine Coefficient

บางครั้งเรียกว่า Cosine Correlation เป็นวิธีการวัดความสัมพันธ์ที่ Salton ได้นำไปใช้กับระบบ SMART เป็นระบบที่ใช้ดึงข้อมูลโดยอัตโนมัติ Salton ได้กำหนดตัวแทนเอกสารเป็นเวกเตอร์ของเลขฐานสองในลักษณะของ n มิติ ซึ่ง n เป็นจำนวนผลรวมของคำดรรชนี ดังนั้น (X,Y) / ||X|| ||Y|| สามารถอธิบายได้ เป็นมุม cosine ที่แยกเป็น 2 ไบนารีเวกเตอร์ X และ Y ดังนั้นสามารถเขียนตามเวกเตอร์จริงได้ดังนี้

| X ∩ Y | / |X|1/2 * |Y| ½

ซึ่ง (X,Y) คือเทอมภายใน และ ||.|| เป็นความยาวของเวกเตอร์

ถ้า X = (X1,…,Xn) และ Y = (Y1,..Yn)



Vector Space Similarity เป็นการวัดความคล้ายคลึงกันของเอกสารโดยใช้เวกเตอร์ สำหรับการวัดความคล้ายคลึงกันของเอกสารนั้นสามารถคำนวณได้จากสูตรดังนี้

$$sim(D_i, D_j) = \sum_{k=1}^{t} w_{ik} * w_{jk}$$

กำหนดให้ Di , Dj เป็นเอกสารใดๆ

$$w_{ik} = \frac{tf_{ik} \log(N / n_k)}{\sqrt{\sum_{k=1}^{t} (tf_{ik})^2 [\log(N / n_k)]^2}}$$

เป็นการ Normalize น้ำหนักของเทอม(the term weights) ซึ่งมีผลให้ช่วงของน้ำหนักของแต่ละเทอมมีค่าอยู่ระหว่าง 0 และ 1 บางครั้งเราเรียกวิธีการนี้ว่า cosine normalized inner product

5. Overlap coefficient

คือ | X ∩ Y | / min(|X| ,|Y|)

ตัวอย่างที่ 3.1 การวัดความคล้ายคลึงกันโดยใช้ Vector Space

กำหนดให้ D เป็นเอกสาร , Q เป็นข้อสอบถาม ,w เป็นน้ำหนักของแต่ละเทอม

$$D_i = w_{d_{i1}}, w_{d_{i2}}, \ldots, w_{d_{it}}$$

$$Q = w_{q1}, w_{q2}, \ldots, w_{qt} \qquad w = 0 \text{ if a term is absent}$$

การวัดความคล้ายคลึงกันกระทำโดยนำน้ำหนักเทอมไป normalize โดยสูตร

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

ตัวอย่างที่ 3.2 การวัดความคล้ายคลึงกันโดยใช้ Vector Space

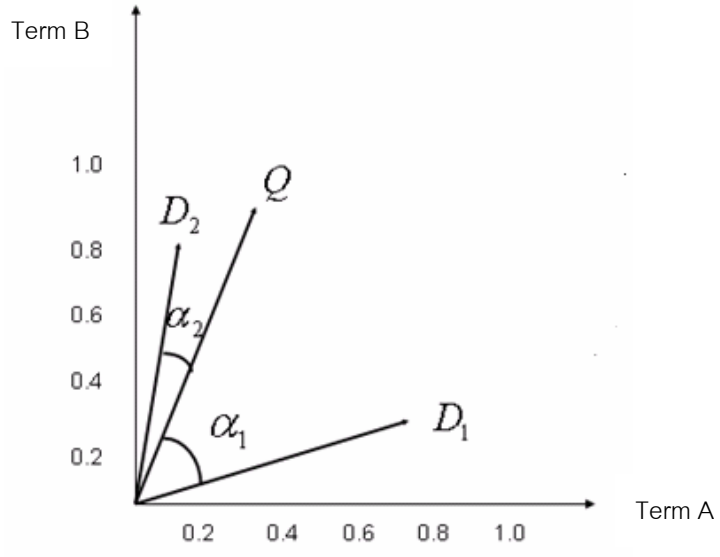
กำหนดให้เวกเตอร์ Q มีค่าดังนี้ $Q = (0.4, 0.8)$

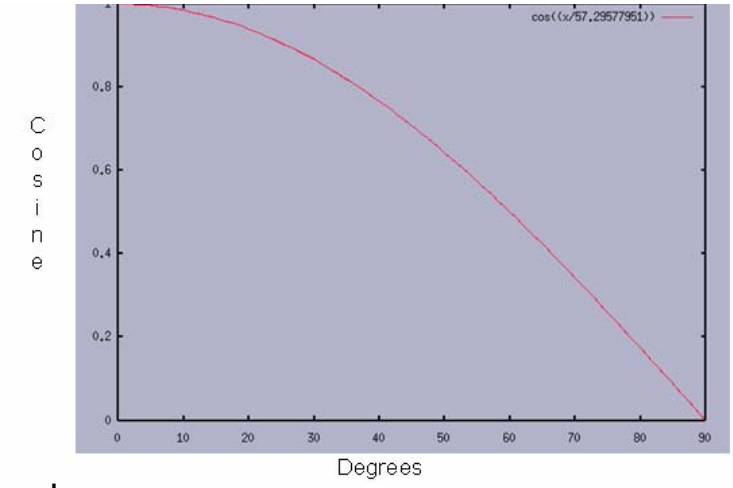
เอกสาร D1 มีค่าดังนี้ D1 = (0.8,0.3)

เอกสาร D2 มีค่าดังนี้ D2 =(0.2,0.3)

$\cos\alpha 1 = 0.74$

$\cos\ \alpha\ 2 = 0.98$





รูปที่ 3.1 : แสดงความสัมพันธ์ระหว่างค่า cosine กับ Degrees

เราสามารถหาความคล้ายคลึงจากเอกสารและข้อสอบถามได้โดยใช้สูตรคือ

$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

แทนค่าในสูตรระหว่างข้อสอบถาม(Q) และเอกสาร D2 ได้ดังนี้

$$sim(Q, D_2) = \frac{(0.4*0.2)+(0.8*0.7)}{\sqrt{[(0.4)^2+(0.8)^2]*[(0.2)^2+(0.7)^2]}}$$

$$= \frac{0.64}{\sqrt{0.42}} = 0.98$$

แทนค่าในสูตรระหว่างข้อสอบถาม(Q) และเอกสาร D1 ได้ดังนี้

$$\text{SIM(Q,D1)} = \frac{0.56}{\sqrt{0.58}} = 0.74$$

3.3 ความไม่คล้ายคลึงกัน (Dissimilarity)

ในการแบ่งกลุ่มนั้นบางกรณีมีการนำความไม่คล้ายคลึงกันของเอกสารมาใช้ในการวัดด้วยเช่นกัน

กำหนด P เป็นเซ็ตของสิ่งของที่ถูกจัดกลุ่ม

D เป็นสัมประสิทธิ์ความไม่คล้ายคลึงกัน เป็นฟังก์ชันหนึ่งจาก P x P ไปยังเลขจำนวนจริงที่ไม่เป็นค่าลบ

โดยทั่วไป D จะมีคุณสมบัติดังต่อไปนี้

(1) D(X , Y) > 0 สำหรับทุกๆ X , Y ของ P

(2) D(X , X) = 0 สำหรับทุกๆ X ของ P

(3) D(X, Y) = D(Y ,X) สำหรับทุกๆ X , Y ของ P

(4) D(X, Y) < D(X , Z) + D(Y, Z)

ซึ่งจากกฎของทั้ง 4 ข้อสามารถสรุปเป็นสูตรได้ดังนี้

1. $|X \Delta Y| / |X| + |Y|$ โดย $|X \Delta Y| = |X \cup Y| - |X \cap Y|$

ซึ่งสูตรนี้มาจากสูตรความคล้ายคลึงของ Dice's Coefficient

$$2|X \cap Y| / |X| + |Y|$$

สามารถหาความไม่คล้ายคลึงกันได้ดังนี้

$$\text{Dissimilarity} = 1 - (2|X \cap Y| / |X| + |Y|)$$
$$= (|X| + |Y| - 2|X \cap Y|) / (|X| + |Y|)$$
$$= (|X \cup Y| - |X \cap Y|) / (|X| + |Y|)$$
$$= |X \Delta Y| / |X| + |Y|$$

2. อีกกรณีหนึ่งถ้าเราใช้สัมประสิทธิ์ของ Jaccard โดยใช้เลขฐานสองแทนค่าดรรชนีของเอกสาร มีผลให้ดรรชนีตัวที่ i จะแทนด้วยเลข 0 หรือ 1 ในตำแหน่งที่ I ตามลำดับ ซึ่งค่า 1 นั้นจะแสดงถึงคำสำคัญอยู่ในเอกสาร สำหรับเลข 0 หมายถึงคำสำคัญไม่ได้อยู่ในเอกสารนั้นในตำแหน่งที่ระบุ เราสามารถรวมจำนวนทั้งหมดของคำสำคัญที่อยู่ในเอกสารได้

ถ้า $|X| = \sum X_i$ โดยที่ $i \in 1..N$

และ N เป็นจำนวนของค่าดรรชนีทั้งหมดในเอกสาร

ดังนั้น $|X \cap Y| = \sum X_i Y_i$ โดยที่ $I \in 1..N$

ความไม่คล้ายคลึงกันมีค่าเท่ากับ

$$| X \Delta Y | \ / \ |X| + |Y| \ = (\sum X_i (1-Y_i) \ + \sum Y_i(1-X_i)) \ / \ (\sum X_i + \sum Y_i)$$

การวัดความเกี่ยวข้องนี้อยู่บนโมเดลทางสถิติซึ่งวัดความเกี่ยวพันธ์กันระหว่าง 2 ออปเจ็กต์ การปรับปรุงประสิทธิภาพระบบค้นคืนสารสนเทศนั้น สามารถใช้ความคาดหวังเพื่อใช้ในการวัดความเกี่ยวข้องได้ การ สามารถกระจายความน่าจะเป็น P(Xi) และ P(Xj)

$$P(x_i, x_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i) P(x_j)}$$

3. สำหรับการคาดหวังสารสนเทศที่คล้ายๆกันนั้น โดย Jardine และ Sibson ในการวัดความไม่คล้ายคลึงกันระหว่าง สองออปเจ็กต์ โดยแบ่งการการจายความน่าจะเป็นเป็นสองจุดคือ 1 และ 0 ดังนั้น P1(1),p1(0),P2(1),P2(0) เป็นการกระจายความน่าจะเป็นของออปเจ็กต์ทั้งสอง ซึ่งการวัดความไม่คล้ายคลึงกันของ Jardine และ Sibson นี้เรียกว่ารัศมีสารสนเทศ (Information Radius) มีการคำนวณดังนี้

$$u P_1(1) \log \frac{P_1(1)}{u P_1(1) + v P_2(1)} + v P_2(1) \log \frac{P_2(1)}{u P_1(1) + v P_2(1)} +$$

$$u P_1(0) \log \frac{P_1(0)}{u P_1(0) + v P_2(0)} + v P_2(0) \log \frac{P_2(0)}{u P_1(0) + v P_2(0)}$$

โดย u และ v คือการเพิ่มน้ำหนักที่เป็นเลขจำนวนบวกไปยังหน่วย ฟังก์ชันนี้จะทำการสร้างในหลายๆสถานะ หรือการกระจายที่มีความต่อเนื่อง

3.4 วิธีจัดแบ่งกลุ่ม (Classification Methods)

โดยปกติข้อมูลประกอบด้วย ออปเจ็กต์ และคำอธิบายที่สอดคล้อง ซึ่งออปเจ็กต์ อาจเป็นเอกสาร(documents) คำดรรชนี(keyword) ตัวอักษรที่เขียนด้วยมือคน (hand written character)หรือเป็นชนิดของพืช(species) ก็ได้ ส่วนคำอธิบาย(Description) เป็นโครงสร้างของข้อมูลภายใต้ชื่อต่างๆ อาจเป็น

- คุณสมบัติในหลายๆสถานะเช่น สี
- สถานะของเลขฐานสอง เช่น คำดรรชนี9Keyword)
- จำนวน เช่น มาตราวัดความแข็ง หรือ น้ำหนักของดรรชนี เป็นต้น
- การกระจายความน่าจะเป็น(Probability Distributions)

Sparck Jones ได้แบ่งวิธีการจัดกลุ่มตามความสัมพันธ์ได้หลายลักษณะดังนี้

1. ความสัมพันธ์ระหว่างคุณสมบัติและคลาส ซึ่งแบ่งออกเป็น 2 แบบคือ Monothetic และ Polythetic ความแตกต่างระหว่าง Monothetic และ Polythetic นั้น

สมมุติว่า $G = \{ f1 , f2 , f3 , \ldots , fn \}$

โดย fi เป็นคุณสมบัติต่างๆ มีกลุ่มของ Individuals ที่อาจเป็นคีย์เวิร์ดหรือเอกสาร แต่ละ Individuals มีคุณสมบัติของ G เป็นจำนวนมาก (row) แต่ละ f ใน G มี Individuals จำนวนมากเป็นเจ้าของ (Column) ไม่มี f ใน G ที่ทุกๆ Individuals ในกลุ่มเป็นเจ้าของ ซึ่งกลุ่มของ Individual ที่มีคุณสมบัติเหมือนกัน เราจัดให้เป็น Monothetic

จากรูปที่ 3.2 จะเห็นว่า แถวที่ 5 และ 6 เป็น Monothetic รวมทั้งแถว 7 และแถวเป็น Monothetic เช่นกันเพราะมีคุณสมบุติเหมือนกัน สำหรับ แถวที่ 1 , 2 , 3 , 4 , 5 เป็น Polythetic เพราะคุณสมบัติในสี่ประการจะมีคุณสมบัติที่เหมือนกัน 3 ข้อที่เหมือนกัน

	A	B	C	D	E	F	G	H
1	+	+	+					
2	+	+		+				
3	+		+	+				
4		+	+	+				
5					+	+	+	
6					+	+	+	
7					+	+		+
8					+	+		+

รูปที่ 3.2 :แสดงความแตกต่างระหว่าง Monothetic และ Polythetic

2. ความสัมพันธ์ระหว่างออปเจ็กต์และคลาส แบ่งออกเป็น 2 แบบคือ Exclusive และ Overlapping ข้อแตกต่างระหว่าง Overlapping และ Exclusive คือ

Overlapping Class หมายถึง กลุ่มของ Individuals ที่อาจมีสมาชิกหนึ่งของกลุ่มนี้ ไปเป็นสมาชิกของกลุ่มอื่น

Exclusive Class หมายถึง กลุ่มของ individual ที่เป็นสมาชิกเฉพาะไม่มี Overlapping Individuals

เนื่องจากขั้นตอนในการแบ่งหมวดหมู่นั้น ข่าวสารบางอย่างจะถูกตัดทิ้งไป เพื่อให้สมาชิกของคลาสเดียวกันมีความคล้ายคลึงกันหรือใกล้เคียงกับข้อมูลดั้งเดิมมากที่สุด

3. ความสัมพันธ์ระหว่างคลาสกับคลาส แบ่งออกเป็น 2 แบบคือ Ordered และ Unordered ข้อแตกต่างระหว่าง Ordered และ Unordered Class คือ

Ordered Classification เป็นการจัดแบ่งกลุ่มแบบมีการจัดลำดับของคลาส เช่น การจัดแบ่งกลุ่มเป็นลำดับชั้น(Hierarchical) ซึ่งคลาสที่อยู่ในระดับล่างจะอยู่ภายใต้ขอบข่ายของคลาสที่อยู่ระดับบน

Unordered Classification เป็นการแบ่งกลุ่มแบบไม่มีลำดับ เช่นการสร้าง Thesaurus แบบอัตโนมัติ ซึ่งเป็นพจนานุกรมข้อมูลที่ทำให้วิธีการค้นหาข้อมูลมีประสิทธิภาพ

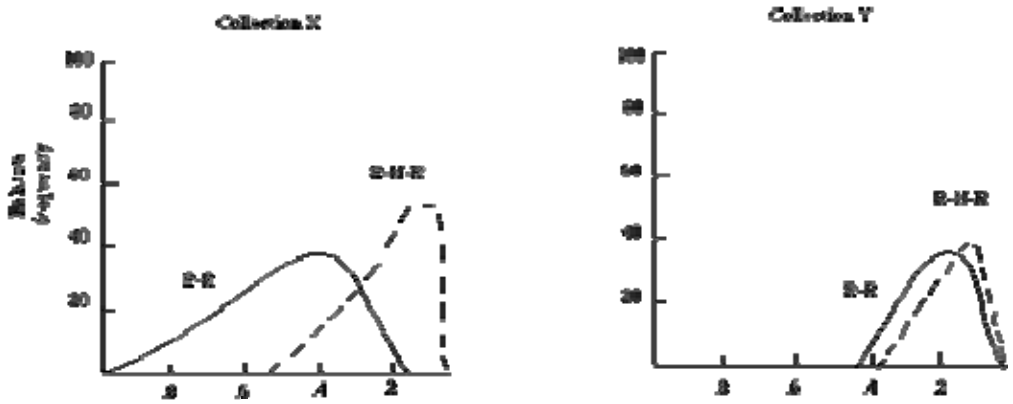
3.5 ข้อสมมุติฐานของคลัสเตอร์ (Cluster Hypothesis)

ข้อสมมุติฐานของคลัสเตอร์นั้นกล่าวไว้ว่า

“ เอกสารที่เกี่ยวข้องกันอย่างใกล้ชิดมีแนวโน้มที่มีการเกี่ยวพันธ์กับข้อคำถามที่ร้องขอเหมือนกัน “

“closely associated documents tend to be relevant to the same requests “

สำหรับข้อสมมุติฐานพื้นฐานในระบบดึงข้อมูล คือเอกสารที่เกี่ยวข้องกับข้อคำถามข้อหนึ่งจะถูกแยกออกจากเอกสารที่ไม่เกี่ยวข้องกับข้อเรียกร้องนั้น หรือเอกสารที่เกี่ยวข้อง (relevant) จะคล้ายคลึงกันและกัน มากกว่าจะคล้ายกับเอกสารที่ไม่เกี่ยวข้อง(Non-relevant)



รูปที่ 3.3 : แสดงถึงผลรวมของกลุ่มข้อเรียกร้อง

จากรูป 3.3 นี้แสดงผลรวมของกลุ่มข้อเรียกร้อง(request) โดยกำหนดการกระจายสัมพันธ์ของ relevant-relevant(R-R) และ relevant-non-relevant(R-N-R) โดยมีการพล็อตจุดของความถี่ที่สัมพันธ์กัน จากผลที่ปรากฏเห็นว่า การแยกของกลุ่ม X จะดีกว่ากลุ่ม B และจุดแข็งของการเกี่ยวข้องระหว่างเอกสารที่เกี่ยวข้องของกลุ่ม X จะดีกว่ากลุ่ม Y ซึ่งจากผลที่เกิดขึ้นนี่เองทำให้เป็นเหตุผลที่ว่าทำไมต้องมีการจัดกลุ่มของเอกสาร(document clustering) เพราะ การแบ่งกลุ่มเอกสาร จะช่วยให้การดึงข้อมูลได้ประสิทธิภาพมากกว่าการค้นหาตามลำดับ เพราะวิธีการค้นหาตามลำดับ จะไม่สนใจความสัมพันธ์ระหว่างเอกสาร แต่คลัสเตอร์สามารถจัดโครงสร้างของกลุ่มโดยให้เอกสารที่คล้ายคลึงกันมากอยู่ในคลาสเดียวกัน ทำให้สามารถเพิ่มความเร็วในการดึงข้อมูล ทำให้มีประสิทธิผลมากขึ้น

ดังนั้น **clustering คือ กระบวนการที่รวมออปเจ็กต์ที่มีลักษณะคล้ายกัน หรือเหมือนกันไว้ในกลุ่มเดียวกัน หรือกำหนดกฎเกณฑ์ขึ้นมาภายในคลัสเตอร์ ซึ่งออปเจ็กต์ที่อยู่ภายในกลุ่มจะมีลักษณะตรงตามที่กำหนดเอาไว้นั่นเอง**

กระบวนการคลัสเตอร์ริ่ง(Clustering algorithm) สามารถนำไปใช้ได้หลายทาง เช่น

- การจัดกลุ่มลูกค้าที่มีลักษณะของการซื้อขายที่เหมือนกันไว้ภายในฐานข้อมูลเดียวกัน
- ทางชีววิทยาสามารถนำลักษณะเด่นของพืชและสัตว์มาจัดกลุ่มไว้ในกลุ่มเดียวกัน

- ทางด้านบรรณารักษ์ จะนำไปใช้ในการจัดหนังสือและแยกประเภทของหนังสือ
- การประกันภัยสามารถนำไปใช้ในการแบ่งแยกกลุ่มของกรมธรรม์ประกันภัยของลูกค้า
- การจัดทำผังเมือง ใช้แบ่งที่ตั้งที่อาศัยตามสภาพภูมิศาสตร์
- การเฝ้าสังเกตการณ์แผ่นดินไหว ใช้ในการเฝ้าสังเกตว่าส่วนใดเป็นพื้นที่อันตรายที่เสี่ยงต่อการเกิดแผ่นดินไหว

ซึ่งการนำไปใช้ในเรื่องใดนั้นต้องมีข้อกำหนดหรือข้อบังคับที่แน่นอน รวมทั้งกำหนดข้อแตกต่างของชนิดของคุณลักษณะต่าง ๆให้ชัดเจน กำหนดขอบเขตความต้องการต่ำสุดที่จะใช้ในการค้นหา และสามารถที่จะจัดการควบคุมกับสิ่งที่ไม่เกี่ยวข้องออกไปให้ได้ ปัญหาของวิธีการนี้นั้นกล่าวคือเทคนิคของ Clustering นี้จะไม่สามารถจัดสรรตำแหน่งที่อยู่ให้เพียงพอกับจำนวนของข้อมูลได้เมื่อมีกรณีที่ซ้ำกัน และสำหรับเลขที่มีจำนวนมากหรือหลายมิติและขนาดข้อมูลใหญ่มาก ๆอาจจะไม่สามารถแก้ไขปัญหาที่ซับซ้อนได้ ซึ่งประสิทธิภาพและวิธีการยังขึ้นอยู่กับความชัดเจนของระยะทาง(Distance-based Clustering) และบางครั้งผลลัพธ์ที่ได้สามารถอธิบายได้หลายทาง

3.6 การใช้วิธีคลัสเตอร์ในระบบค้นคืนสารสนเทศ

วิธีการคลัสเตอร์(Clustering algorithm) มีอยู่ด้วยกันหลายวิธี แบ่งออกเป็น 4 ลักษณะคือ

(1) Exclusive Clustering เป็นการจัดกลุ่มที่ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีลักษณะเหมือนกัน
(2) Overlapping Clustering เป็นการจัดกลุ่มที่ข้อมูลภายในกลุ่มจะแบ่งออกให้อยู่ในรูปของเซตย่อย ๆแต่ละจุดอาจเป็นส่วนหนึ่งของคลัสเตอร์มากกว่าหนึ่งก็ได้และจำนวนสมาชิกภายในเซตต่าง ๆมีค่าแตกต่างกันได้
(3) Hierarchical Clustering เป็นการจัดกลุ่มที่มีการรวมคุณสมบัติพื้นฐานของ Exclusive Clustering และ Overlapping Clustering ไว้ด้วยกันแต่มีการกำหนดเงื่อนไขหรือกฎที่ใช้ในการกำหนดข้อมูลในแต่ละคลัสเตอร์
(4) Probabilistic Clustering เป็นการจัดกลุ่มโดยใช้วิธีการทางสถิติ

การเลือกวิธีที่เหมาะสมในการคลัสเตอร์นั้นมีหลักเกณฑ์ 2 ประการคือ

1. ความมั่นคงของวิธีการ โดยต้องเป็นไปตามเงื่อนไขดังนี้

(1) เมื่อมี ออปเจ็ก อื่นๆเพิ่มเข้ามา คลัสเตอร์จะไม่ค่อยเปลี่ยนแปลง คือ เจริญเติบโตแบบคงที่

(2) วิธีการนั้นจะต้องมั่นคง กล่าวคือความผิดพลาดเล็กๆน้อยๆ ก็จะก่อให้เกิดการเปลี่ยนแปลงอย่างเล็กๆน้อยๆในคลัสเตอร์เช่นกัน

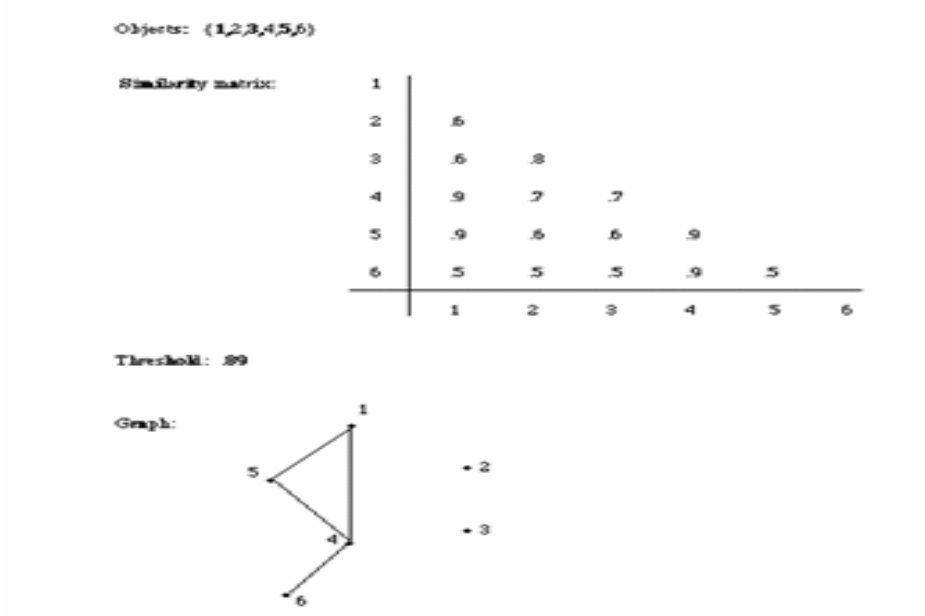
(3) วิธีการต้องเป็นอิสระไม่ขึ้นอยู่กับลำดับเริ่มแรกของออกเจ็กต์

2. ประสิทธิภาพ ในด้านความเร็ว และ ความต้องการเนื้อที่เก็บความจำ ซึ่งเป็นความสามารถในการดึงข้อมูลที่ต้องการ และมีข้อมูลที่ไม่ต้องการมาน้อยที่สุด

ซึ่งในที่นี้เราสามารถสร้างอัลกอริทึมวิธีคลัสเตอร์ริ่ง โดยกำหนดวิธีคลัสเตอร์ริ่ง ดังนี้

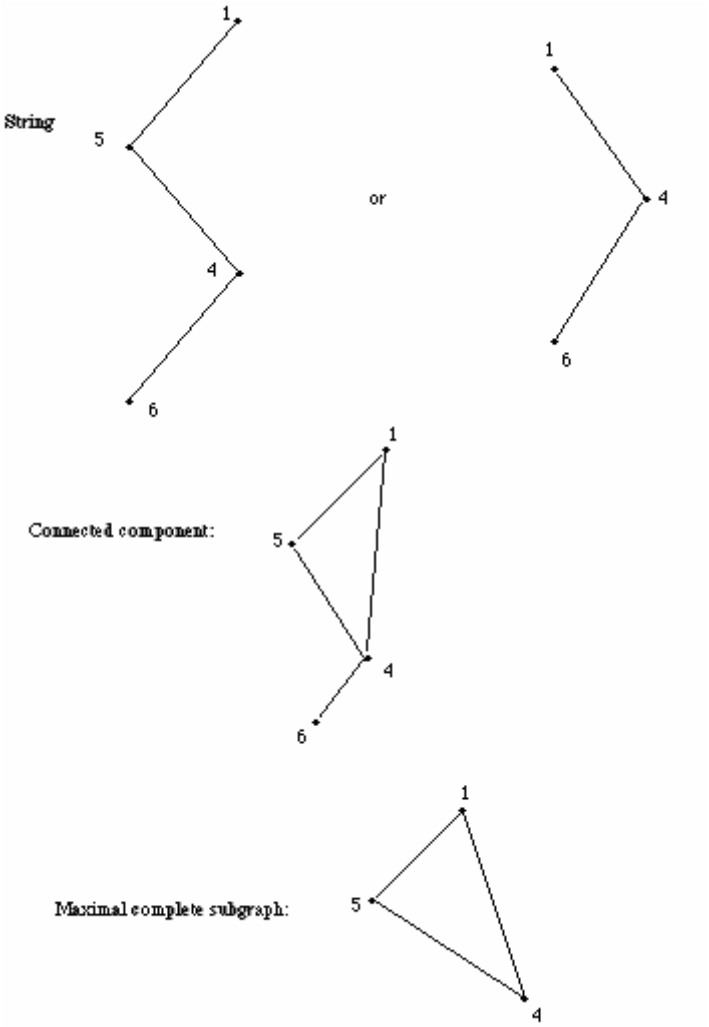
1. วิธีการคลัสเตอร์ริ่งที่อยู่บนรากฐานของวิธีวัดความคล้ายคลึงกันระหว่าง Object ที่จะถูกแบ่งกลุ่ม ตัวอย่างเช่น

1.1 Graph Theoretic Method ที่กำหนดนิยามของคลัสเตอร์ในรูปแบบของกราฟที่ได้จากการวัดความคล้ายคลึงกัน



รูปที่ 3.3 :แสดงสัมประสิทธิ์ของความคล้ายคลึงกันของ 6 ออปเจ็กต์ที่สามารถใช้ดึงออปเจ็กต์ที่มีความคล้ายคลึงกันได้

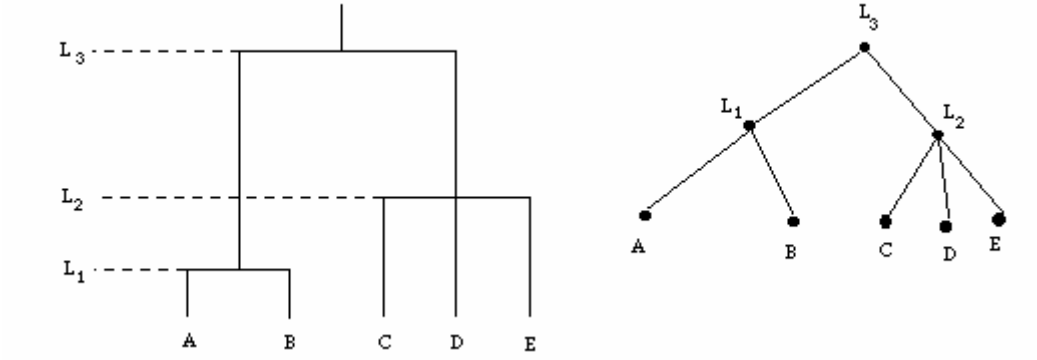
จากรูปที่ 3.3 เป็นกลุ่มของออปเจ็กต์ที่นำมาจัดเป็นคลัสเตอร์ เราคำนวณค่าความคล้ายคลึงกันของแต่ละคู่ของออปเจ็กต์ กราฟที่ได้นั้นจะสอดคล้องกับค่าความคล้ายคลึงที่กำหนด เราเรียกค่าที่กำหนดขึ้นนี้ว่า threshold ค่า threshold ที่กำหนดนี้จะทำให้ออปเจ็กต์ 2 ออปเจ็กต์ที่มีค่าความคล้ายคลึงกันมีค่าเท่ากันหรือมากกว่าถูกเชื่อมโยงเป็นโหนด 2 โหนดในกราฟที่มีการเชื่อมโยงกัน



รูปที่ 3.5 : แสดงการนิยามคลัสเตอร์ที่เป็นไปได้ในเทอมของกราฟย่อย

จากรูปที่ 3.5 นั้น เป็นวิธีการ Keyword clustering ของ Sparck Jones and Jackon ,Auguestson and Minker และ Vaswani and Cameron โดย String เป็นลำดับการเชื่อมโยงของออปเจ็กต์จากจุดเริ่มต้น สำหรับ connected component นั้นเป็นเซตของออปเจกต์ ที่แต่ละออปเจกต์ถูกเชื่อมโยงไปยังสมาชิกอื่นในกลุ่มอย่างน้อย 1 เส้น สำหรับ maximal compleate subgraph เป็นกราฟย่อยที่แต่ละโหนดถูกเชื่อมโยงไปยังโหนดอื่นๆในกราฟย่อยนั้น

1.2 Single Link เป็นวิธีการคลัสเตอร์ริ่งที่ทำให้เกิด Hierarchic Cluster ซึ่งใช้การคำนวณค่าความสัมพันธ์ของ Objects อัลกอริทึมนี้อยู่บนพื้นฐานของสัมประสิทธิ์ความไม่คล้ายคลึงกันของข้อมูลนำเข้า (dissimilarity coefficient) ผลลัพธ์ที่ได้เป็นลำดับขั้น ของระดับตัวเลขที่มีส่วนร่วมกัน เรียกว่า dendrogram ในลักษณะโครงสร้างต้นไม้ (tree structure) ซึ่งแต่ละโหนดแทนหนึ่งคลัสเตอร์

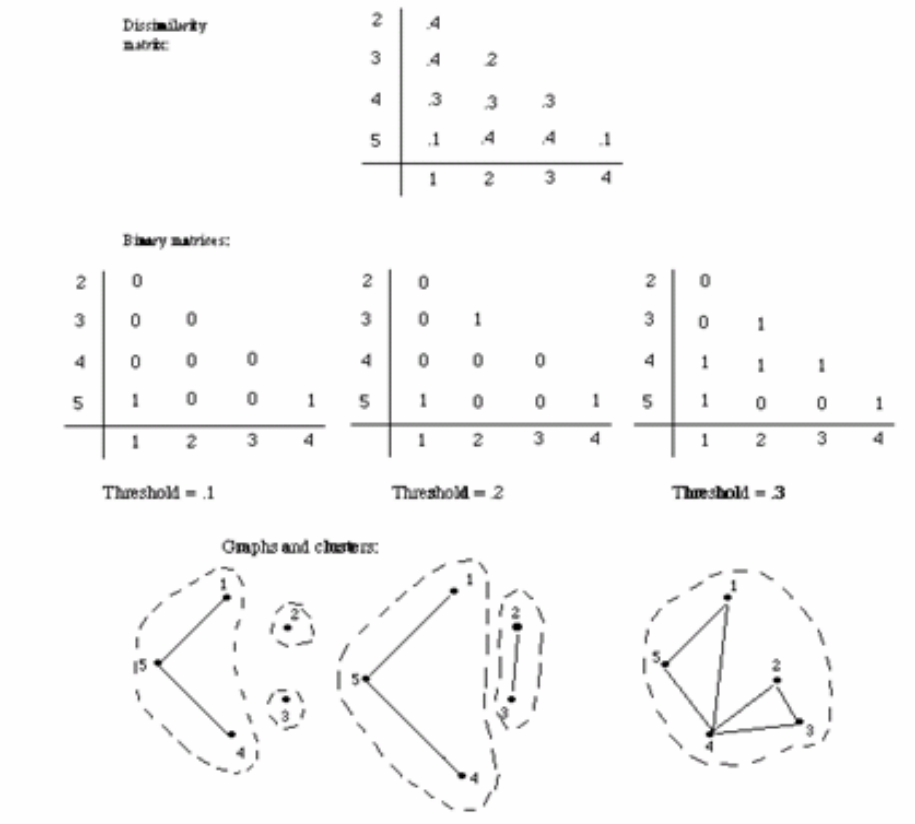


รูปที่ 3.6 :แสดง Dendrogram

ตามรูปที่ 3.6 นั้น เป็นกลุ่มของออปเจ็กต์ {A,B,C,D,E} คลัสเตอร์ในระดับ L1 ประกอบด้วย {A,B}, {C} ,{D} ,{E} และที่ระดับ L2 ประกอบด้วย {A,B} และ {C,D,E} สำหรับที่ระดับ L3 นั้นคือ (A,B,C,D,E} ซึ่งในแต่ละระดับสามารถกำหนดเซตของคลาส และสามารถเคลื่อนย้ายลำดับชั้นที่ต่ำกว่าให้สูงขึ้นได้ แต่การกำหนดคณิตศาสตร์ของ dendrogram ยังมีน้อยมาก ส่วนมากจะสนใจวิธีของ Jardine and Sibson ที่มีการจัดกลุ่ม Single-link โดยวัดจากสัมประสิทธิความไม่คล้ายคลึงกัน(dissimilarity:DC) สมมุติว่าจากตัวอย่างนี้เราสามารถวัด

คล้ายคลึงกันจากการกำหนด thresholding

มีวิธีการของ hierarchic cluster อีกหลายวิธีเช่น complete-link , average-link ที่พัฒนาขึ้นมาเพื่อนำมาใช้เป็นกลยุทธ์ในการค้นคืนสารสนเทศ โดยการวิเคราะห์ระดับตำแหน่งของพารามิเตอร์ หรือ matching function threshold ในการค้นหาแบบลำดับ ซึ่งการค้นคืนคลัสเตอร์ที่ดีสำหรับการสอบถาม(request) นั้นจะมีการจับคู่ในระดับตำแหน่งที่ต่ำ (low level) จะมีค่า high precision และ row recall เป็น cut-off ที่ตำแหน่งระดับต่ำ(low rank position) สำหรับการค้นคืนคลัสเตอร์ที่ดีสำหรับการสอบถามที่ระดับสูงในลำดับชั้นนั้นมีแนวโน้มที่จะให้ high recall แต่ low precision

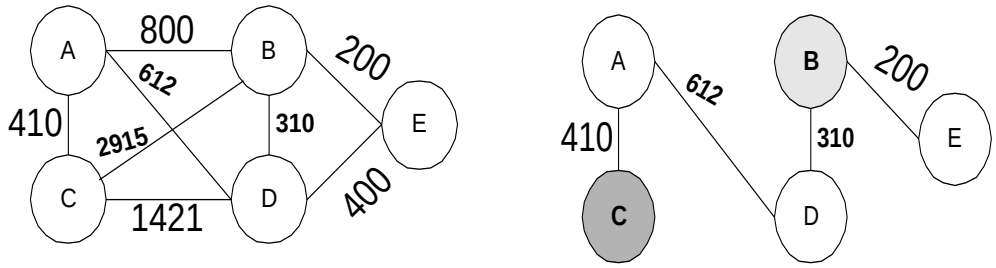


รูปที่ 3.7 :แสดงถึง single-link clusters ที่ดึงจากสัมประสิทธิ์ความไม่คล้ายคลึงกันจากการกำหนด thresholding

ระบบ Hierarchic ของคลัสเตอร์ที่เหมาะสมนั้นมีเหตุผลสามประการที่สำคัญ คือ

- ต้องใช้กลยุทธ์ที่มีประสิทธิภาพที่สามารถค้นหา hierarchic ได้
- การสร้างระบบ hierarchic นั้นต้องมีความเร็วมากกว่าการสร้างที่ไม่เป็นลำดับชั้น
- เนื้อที่ในการจัดเก็บโครงสร้างของ hierarchic นั้นต้องใช้เนื้อที่น้อยกว่าวิธีอื่นๆ

วิธีการของ hierarchic นี้เหมาะสมหรับการจัดกลุ่มของเอกสารที่ขึ้นอยู่กับคำถาม สำหรับต้นไม้ชนิดอื่นๆ เช่น Minimum Spanning Tree หรือ MST นั้นก็เป็นต้นไม้ที่ได้จากสัมประสิทธิ์ความไม่คล้ายคลึงกันเหมือนกับ Single-link tree ความแตกต่างที่เห็นได้ชัดเจน คือ โหนดของ Single-linked tree นั้น จะแทนคลัสเตอร์ แต่ Minimum Spanning Tree นั้นจะแทนออปเจ็กต์แต่ละตัวที่ถูกรวมกลุ่ม ซึ่ง MST คือต้นไม้ที่มีความยาวในการเชื่อมโยงแต่ละออปเจ็กต์ที่มีค่าน้อยที่สุด ซึ่งความยาวนี้อาจเป็นน้ำหนักของการเชื่อมโยงในต้นไม้นั้น ซึ่งความคล้ายคลึงกันนี้ ซึ่ง MST จะมีการบรรจุสารสนเทศมากกว่า Single-link hierarchy และ Single link Cluster แม้ว่าเราจะสามารถดึง single-link hierarchy จากการประมวลผลของ thresholding ก็ตามแต่เราไม่สามารถย้อนกลับได้ ซึ่ง MST จะมีประโยชน์มากกว่าในการเชื่อมโยงเพราะเป็นอิสระในการทำงาน ลดหน่วยความจำในการจัดเก็บรายละเอียดของออปเจ็กต์ อัลกอลิทึมของ minimum spanning tree ให้ผลลัพธ์เฉพาะเส้นที่เชื่อม (edge) ระหว่างโหนด เมื่อนำค่าของ edge มารวมกันแล้วจะได้ความยาวที่สั้นที่สุดบนต้นไม้ดังรูป



minimum spanning tree

2. วิธีการคลัสเตอร์ริ่งที่กระทำได้จากคำอธิบาย(Descriptions) ของ Object โดยตรงไม่มีการวัดความคล้ายคลึงกัน แต่เก็บโครงสร้างที่เหมาะสมโดยจำกัดจำนวนของคลัสเตอร์และขนาดของแต่ละคลัสเตอร์ นั่นคือแนวคิดของการสร้างตัวแทนคลัสเตอร์ (cluster representative) ซึ่งเราเรียกว่า cluster profile หรือ classification vector หรือ centroid จากการสรุปจากหลักเกณฑ์บางอย่างเพื่อเป็นตัวแทนของออปเจ็กต์ในกลุ่มซึ่งตัวแทนนี้จะมีความ

- จำนวนของคลัสเตอร์ที่ต้องการ
- ขนาดต่ำสุดและขนาดสูงสุดของแต่ละคลัสเตอร์
- ค่าของ threshold บน matching function ซึ่งออปเจ็กต์ที่มีค่าต่ำกว่า threshold ที่กำหนดจะไม่ถูกรวมในคลัสเตอร์นี้
- การควบคุมการซ้อนทับ(Overlap)ระหว่างคลัสเตอร์ .
- พารามิเตอร์ที่เลือกอย่างไม่มีกฎเกณฑ์จะถูกปรับให้อยู่ในระดับที่ให้ผลที่ดีที่สุด

การทำงานของอัลกอริทึมนี้จะกระทำซ้ำๆ เพื่อให้ได้ผลลัพธ์ที่เหมาะสมที่สุด

วิธีการคลัสเตอร์ริ่งที่กระทำได้จากคำอธิบาย(Descriptions) ของ Object โดยตรง อย่างเช่น

1. อัลกอริทึมของ Rocchio's clustering นั้นแบ่งการปฏิบัติงานออกเป็น 3 ระยะ

ระยะแรก จะเป็นระยะที่ทำการเลือก ออปเจ็กต์จำนวนหนึ่งเป็นศูนย์กลางของคลัสเตอร์ โดยใช้เกณฑ์ทีกำหนด สำหรับออปเจ็กต์ที่เหลือกำหนดให้เป็นกลุ่ม rag-bag พื้นฐานของการกำหนดค่าเริ่มต้นนั้นตัวแทนคลัสเตอร์จะถูกคำนวณจากออปเจ็กต์ทั้งหมด ซึ่งกฎที่กำหนดขึ้นนั้นกำหนดในเทอมของ thresholds บน matching function คลัสเตอร์สุดท้ายอาจจะมีการซ้อนทับกัน(overlap) ซึ่งออปเจ็กต์หนึ่งอาจมีการกำหนดมากกว่าหนึ่งคลัสเตอร์ได้

ระยะที่สองนั้น เป็นการกระทำขั้นตอนซ้ำๆเพื่อที่จะหาพารามิเตอร์นำเข้าที่สามารถกระทำได้บรรลุวัตถุประสงค์ที่ต้องการหรือตรงตามเงื่อนไขที่กำหนด พารามิเตอร์ต่างๆเหล่านี้อาจเป็น ขนาดของคลัสเตอร์ หรืออื่นๆ

ระยะที่สามนั้นเป็นเป็นการพิจารณาออปเจ็กที่เหลือที่ไม่ได้ถูกกำหนด และมีการขจัดออปเจ็กต์ที่มีการซ้อนทับกันระหว่างคลัสเตอร์ ให้น้อยลง

ซึ่งส่วนมากของอัลกอริทึมเหล่านี้จะมีวัตถุประสงค์เพื่อลดจำนวนการส่งผ่านที่ต้องมีการกระทำของแฟ้มของรายละเอียดออปเจ็กต์ มีอัลกอริทึมในการแบ่งกลุ่มซึ่งมีการทำงานเป็น Single-Pass algorithm ซึ่งมีการปฏิบัติการพื้นฐานดังนี้

- คำอธิบายออปเจ็กต์ถูกดำเนินการเป็นลำดับ
- ออปเจ็กต์ตัวแรกจะถูกกำหนดให้เป็นตัวแทนคลัสเตอร์ของคลัสเตอร์แรก
- ออปเจ็กต์หลังจากนั้นจะถูกจับคู่กับตัวแทนคลัสเตอร์ทั้งหมด

- ถ้าออปเจ็กต์มีการซ้อนทับกัน จะมีการกำหนดให้ออปเจ็กต์มีค่าตามเงื่อนไงบน matching function
- เมื่อออปเจ็กต์ใดถูกกำหนดเป็นตัวแทนของคลัสเตอร์ จะมีการคำนวณใหม่
- ถ้าออปเจ็กต์ใดพลาดจากการทดสอบ(test) จะนำกลับไปเป็นตัวแทนคลัสเตอร์สำหรับคลัสเตอร์ใหม่

ซึ่งการกระทำซ้ำๆนี้ การจัดกลุ่มสุดท้ายจะขึ้นกับพารามิเตอร์ข้อมูลเข้า(input parameter) ที่เกิดจากการกำหนดจากการสังเกตของความแตกต่างของเซตออปเจ็กต์

2. อัลกอริทึมของ Dattola นั้นเริ่มจากการกำหนดส่วนเริ่มต้นและเซตของตัวแทนคลัสเตอร์ก่อน ต่อจากนั้นจะทำการจัดสรรตัวแทนคลัสเตอร์ใหม่ ซึ่งตัวแทนคลัสเตอร์ใหม่จะแทนที่ตัวแทนเก่าถ้าตัวแทนใหม่นั้นมีค่าใกล้เคียงกับวัตถุในคลัสเตอร์ใหม่มากกว่าตัวแทนเก่า วัดจากความรู้สึก ซึ่งอัลกอริทึมนี้จะสร้างกลุ่มที่เป็นลำดับชั้น(hierarchic)

ต่อจากนั้นมีการวิจัยอัลกอริทึมที่เหมาะสมอีกหลายท่าน มีการผสมผสานระหว่างวิธีของ graph-theoretic และ heuristic approaches (วิธีสำนึก) ซึ่งคลัสเตอร์ระยะแรกจะถูกกำหนดโดยวิธีการของ graph-theoretic บนพื้นฐานของการวัดการมีส่วนร่วม(association measure) ต่อจากนั้นจะถูกจัดสรรใหม่ตามเงื่อนไขบน matching function ความเร็วของอัลกอริทึมนี้จะอยู่ในรูปของ n log n โดย n คือจำนวนของออปเจ็กต์ที่ถูกนำมารวมกลุ่ม การเปรียบเทียบอยู่บนพื้นฐานของการวัดการมีส่วนร่วม

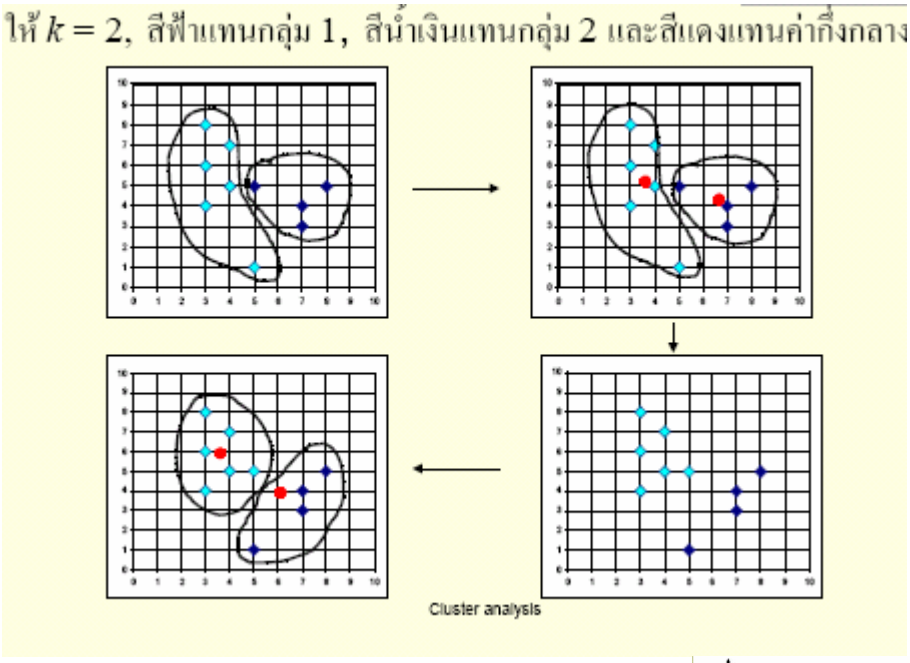
3.7 Clustering algorithm

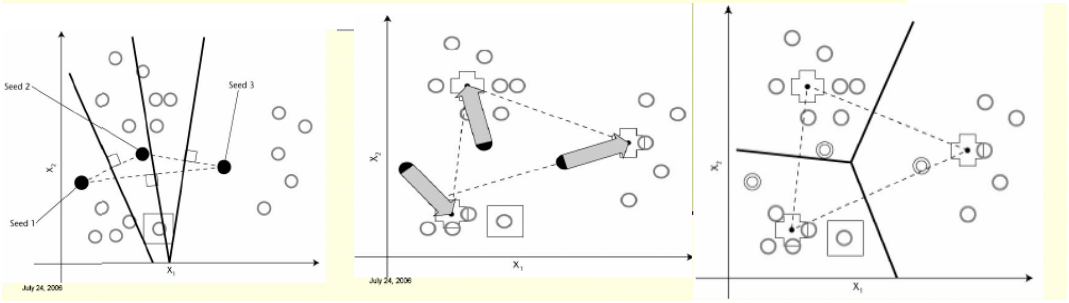
วิธีการที่จะใช้สำหรับ Clustering algorithm มีอยู่ด้วยกันหลายวิธี ดังตัวอย่างต่อไปนี้

(1) K-means เป็นขั้นตอนวิธีการจัดกลุ่มโดยวิธีแบ่งกั้น(Partitioning) ระเบียนข้อมูลถูกแบ่งกั้นเป็นกลุ่มที่ไม่มีสมาชิกร่วมกันเลย โดยใช้การกั้นระหว่างกลุ่มด้วยระยะทาง วิธีการแบ่งกั้นนั้น กำหนดให้ข้อมูล n ระเบียน แบ่งเป็น K กลุ่มที่ไม่มีสมาชิกร่วมกัน วิธีการเกาะกลุ่มโดยใช้วิธีแบ่งกั้นมีขั้นตอนดังนี้

a. แบ่งกลุ่มข้อมูลเป็น k กลุ่มที่ไม่ใช่เซตว่าง
b. คำนวณจุดกึ่งกลาง(centroid) ของกลุ่ม โดยใช้ค่าเฉลี่ยเลขคณิต(mean)
c. สำหรับแต่ละระเบียน นำระเบียนเทียบกับจุดกึ่งกลาง เพื่อกำหนดกลุ่มให้กับระเบียน โดยเลือกระยะจากระเบียนไปจุดกึ่งกลางที่ใกล้ที่สุด
d. วนซ้ำจนกระทั่งไม่มีการเปลี่ยนกลุ่มของระเบียน หรือครบจำนวนรอบสูงสุดที่กำหนดไว้

ตัวอย่าง การทำงานของขั้นตอนวิธี k-mean





K-means clustering1 K-means clustering2 K-means clustering3

ผลลัพธ์ที่ได้จากการเกาะกลุ่มมาจากโครงสร้างในข้อมูล ดังนั้นการเกาะกลุ่มไม่มีลักษณะการบอกว่าการเกาะกลุ่มผิดหรือถูก

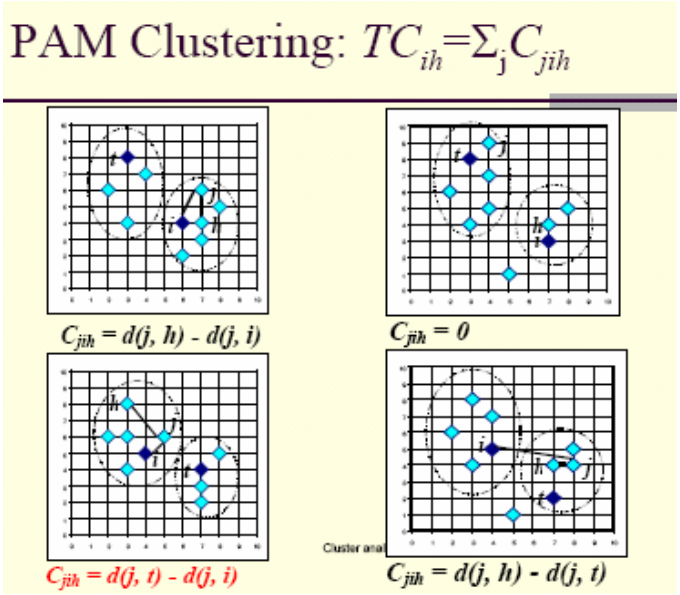
ข้อดีของวิธีการนี้คือ ประสิทธิภาพมีค่าเท่ากับ O(tkn) เมื่อ n คือจำนวนข้อมูล k คือจำนวนกลุ่ม และ t คือจำนวนรอบที่ต้องวนซ้ำโดยปกติแล้วค่า k และ t จะน้อยกว่า n มาก ผลลัพธ์ที่ได้มักเป็นผลเฉลยเฉพาะที่(local optimal) ถ้าต้องการผลเฉลยที่ให้ค่าที่ดีที่สุด(global optimal) จำเป็นต้องใช้วิธีอื่นช่วย

สำหรับข้อเสียของวิธีนี้คือ ใช้ได้เฉพาะลักษณะประจำที่เป็นจำนวนเพราะต้องใช้ค่าเฉลี่ย นอกจากนี้ต้องมีการกำหนดค่า k ก่อนเริ่มขั้นตอนวิธี อีกทั้งข้อมูลที่ผิดปกติจะทำให้

ขั้นตอนวิธีที่ใช้หลักการคล้ายกับขั้นตอนวิธี K-means เริ่มจากกำหนดตัวกึ่งกลางก่อน ใช้ตัววัดความคล้ายคลึงที่ไม่ใช่ค่าเฉลี่ยเลขคณิต ใช้ยุทธวิธีในการคำนวณค่าเฉลี่ย อาจเป็นฐานนิยม(mode) หรือใช้เมตริกซ์ที่วัดความคล้ายคลึงของลักษณะประจำประเภท categorical หรือใช้ frequency-bases ในการกำหนดค่ากึ่งกลางของกลุ่ม หรือใช้วิธี k-prototype ที่มีตัววัดผสมระหว่าง categorical และจำนวน ในกรณีที่ตัวแทนที่ใช้ต้องเป็นระเบียนหนึ่งในชุดข้อมูล เรานิยามตัวแทนของกลุ่มว่า medoid เช่นวิธีที่เรียกว่า PAM(Partitioning Around Medoids,1987) ซึ่งวิธีการนี้มีการสุ่มเลือกระเบียนเป็น medoid มีการคำนวณระยะรวมทั้งหมดของระเบียนกับ medoid ของกลุ่ม ทำการวนซ้ำโดยคำนวณการสลับ medoid กับระเบียนอื่นไม่ทำให้ระยะทั้งหมดดีขึ้นจึงหยุด ซึ่งวิธีการนี้ไม่เหมาะกับข้อมูลปริมาณมาก

ขั้นตอนวิธี PAM

- เลือกข้อมูล k ตัวใช้เป็นตัวแทนข้อมูลอย่างสุ่ม
- สำหรับแต่ละคู่ของข้อมูล h ที่ไม่ใช้ medoid กับ iที่เป็น medoid
- คำนวณค่าการสลับ TCih
    - ถ้า TCih<0 ให้แทน iด้วย h กล่าวคือมีการเปลี่ยนตัวแทน
- ทำซ้ำจนกระทั่งไม่มีการเปลี่ยนกลุ่มของ medoid อีก



อัลกอริทึมที่ข้อมูลที่อยู่ในคลัสเตอร์เดียวกันมีลักษณะเดียวกัน เป็นวิธีการอีกหนึ่งวิธีที่อัลกอริทึมจะทำการแยกประเภทของข้อมูลให้ตรงกับสิ่งที่แต่ละคลัสเตอร์กำหนดไว้ หลักการที่สำคัญคือกำหนดค่าของตัวแปร K ในแต่ละคลัสเตอร์ นำแต่ละจุดที่มีความสัมพันธ์กันของข้อมูลมาเชื่อมโยงกัน จนกว่าที่ไม่สามารถเชื่อมโยงได้ ต่อจากนั้นนำจุดเหล่านั้นมาคำนวณเป็นค่า K ใหม่กระทำซ้ำไปจนค่าของ K ไม่เปลี่ยนค่า สามารถสรุปขั้นตอนการทำงานของอัลกอริทึมได้ดังนี้

a. กำหนดตำแหน่งของจุด K ไปยังคลัสเตอร์ที่ยังไม่มีข้อมูล จุดนี้จะแทนค่าของกลุ่มแรก

b. กำหนดออปเจ็กต์ให้แต่ละคลัสเตอร์ โดยให้มีความใกล้เคียงกับค่าที่กำหนด

c. คำนวณตำแหน่งของค่า K ใหม่

d. กระทำซ้ำในขั้นตอนที่ b และ c จนกระทั่งค่าของ K ไม่เปลี่ยนแปลง ซึ่งขั้นตอนนี้เป็นการแบ่งออปเจ็กต์ที่อยู่ในกลุ่มออกเป็นเมตริกซ์ เพื่อที่จะสามารถนำมาคำนวณหาค่าให้น้อยลงให้มากที่สุด

วิธีการนี้ไม่จำเป็นต้องมีโครงสร้างที่ดีเยี่ยม เพียงแต่ให้สอดคล้องกับวัตถุประสงค์ที่ต้องการมากที่สุด และสามารถประมวลผลแบบ multiple time ได้ และสามารถนำไปแก้ปัญหากับ fuzzy feature vector

สมมุติว่ามีเวกเตอร์ n ตัวคือ X1,X2,...,Xn ซึ่งอยู่ในคลาสเดียวกัน มีข้อตกลงว่า K< n ให้ mi เป็นค่าเฉลี่ยของเวกเตอร์ภายใน cluster i ถ้าเราแยกคลัสเตอร์ออกโดยใช้การหาระยะทางที่ต่ำที่สุด ดังนั้นเราสามารถกล่าวได้ว่า X อยู่ในคลัสเตอร์ i ถ้า ||X-mi|| มีค่าต่ำสุดของค่า K ทุกค่า การทำงานเริ่มจาก

- การคาดคะเนค่าเฉลี่ย m1,m2,…mk
- Until เปลี่ยนค่าเฉลี่ยไปเรื่อยๆ
    - ใช้ค่าเฉลี่ยที่ประมาณไว้ แยกออกจากคลัสเตอร์ได้
    - For I = 1 to K
        - แทนค่า mi ด้วยค่าเฉลี่ยข้างต้น
- end_until

อัลกอริทึมนี้จะแยก m ออกจากคลัสเตอร์ของ K เพื่อหาค่าของผลรวมที่น้อยที่สุดของระยะทางยกกำลังสองภายในคลัสเตอร์

ถึงแม้ว่าอัลกอริทึม K-mean จะง่ายและทำงานในเวลาเชิงเส้น แต่เมื่อใช้ในการแก้ปัญหาเวกเตอร์แบบหลายมิติขนาดใหญ่ก็จะใช้การทำงานที่มากและซับซ้อน มีงานวิจัยที่มีการปรับปรุงเช่น CLARA(Cluster Large Applications,1990) , CLARANS(Ng&Han,1994) มีการพัฒนาให้ใช้กับข้อมูลปริมาณมากได้ รวมทั้ง งานวิจัยของ Sanpawat (http://www.cs.tufts.edu/~{sanpawat,couch}) ที่ได้มีการนำเสนอวิธีการปรับปรุงอัลกอริทึมเคมีนโดยการนำการคำนวณแบบขนานเข้ามาร่วมด้วย ซึ่งทำให้ประสิทธิภาพดีขึ้นด้วยปัจจัย O(K/2) โดยที่ K คือจำนวนกลุ่มที่ต้องการ ซึ่งสามารถทำงานในเครื่องหลายๆเครื่องที่มีหน่วยความจำแบบสะสมได้ขนาดใหญ่

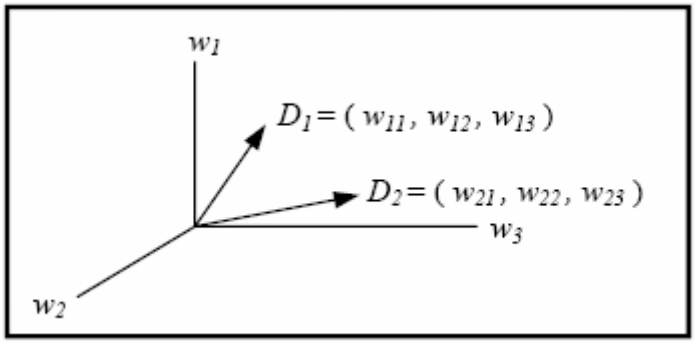
งานวิจัยของ ฆนาศัย กรึงไกร และ ชุลีรัตน์ จรัสกุลชัย ได้ทำการวิจัยเรื่อง การจัดกลุ่มเอกสารข้อความภาษาไทยด้วยขั้นตอนวิธี Spherical K-Means แบบขนานบนพิรุณลีนุกซ์คลัสเตอร์ โดยพยายามคิดค้นวิธีที่จะจัดกลุ่มเอกสารปริมาณมากๆแบบอัตโนมัติ ซึ่งจัดกลุ่มแบบโกลบอล(global) ซึ่งเอกสารถูกจัดกลุ่มตามที่ปรากฏอยู่ในชุดเอกสารทั้งหมด จุดประสงค์เพื่อลดเวลาการประมวลผลกับชุดเอกสารข้อความเต็ม(full text) จำนวนมากๆ แนวคิดคือให้เอกสารทั้งหมดถูกแบ่งออกไปประมวลผลเป็นชุดย่อยๆ และสามารถถูกจัดกลุ่มในเวลาเดียวกันได้อย่างอิสระ หลังจากนั้นแยกรายงานออกเป็นกลุ่มตามเนื้อหาที่คล้ายคลึงกัน ในงานวิจัยนี้มีการแทนเอกสารที่ไม่มีโครงสร้าง(unstructured text document) ด้วย Vector Space Model(VSM) ใน VSM เอกสารแต่ละฉบับ เปรียบเสมือนกับเวกเตอร์คำ โดยที่ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้น

กำหนดให้ w_{ik} คือค่าน้ำหนักของคำ k ที่ปรากฏในเอกสาร i

เวกเตอร์สำหรับเอกสาร D_i สามารถเขียนแทนด้วย

$D_i = (w_{i1}, w_{i2}, \ldots, w_{it})$ ซึ่ง คือจำนวนของคำที่ไม่ซ้ำกันในชุดเอกสารทั้งหมด

ดังนั้น ใน space ของชุดเอกสารหนึ่งจะมีมิติเท่ากับ t-มิติ ดังรูปแสดงเวกเตอร์ใน 3-มิติ



โดยคุณสมบัติของเวคเตอร์ทำให้สามารถคำนวณค่าความคล้ายคลึง(similarity) ของเอกสารคู่หนึ่งๆได้จากค่าสัมประสิทธิ์ cosine ของมุมระหว่างคู่ของเวคเตอร์ซึ่งจะมีค่าอยู่ระหว่าง 0 ถึง 1 มีสูตรดังนี้

$$\mathrm{sim}(D_i, D_j) = \frac{D_i \bullet D_j}{\|D_i\| \times \|D_j\|}$$

ต่อจากนั้นนำมาให้น้ำหนักของคำ คำหนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำ(term frequency) ที่ปรากฏในเอกสารนั้นและจำนวนของเอกสารทั้งหมดที่มีคำนั้นปรากฏอยู่ วิธีการให้น้ำหนักคำที่ใช้กันอย่างมากในการสืบค้นข้อมูลคือ tf*idf(term frequency * inverse document frequency) โดยที่ค่า idf คำนวณจากค่า log(N/df) ซึ่ง N คือจำนวนเอกสารในชุดเอกสารทั้งหมด และ df คือจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ วิธีการให้น้ำหนักของคำมีการ normalization ทำให้เวคเตอร์เอกสามรมีขนาด 1 หน่วยมีสูตรดังนี้

$$w_{ik} = \frac{tf_{ik} \times \log(N/df_k)}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 \times (\log(N/df_j))^2}}$$

โดยที่ tf_{ik} คือความถี่ของคำ k ในเอกสาร i

N คือจำนวนของเอกสารในชุดเอกสาร

df_k คือจำนวนของเอกสารในชุดเอกสารซึ่งบรรจุคำ i

สามารถหาค่าความคล้ายคลึงกันของคู่เอกสารได้จาก inner product เนื่องจาก $||D_i|| = ||D_j|| = 1$ ดังนั้น

$$\mathrm{sim}(D_i, D_j) = D_i \bullet D_j = \sum_{k=1}^{t} (w_{i,k} \times w_{j,k})$$

ตัวอย่างที่ 3.3 ในชุดเอกสารประกอบด้วยเอกสาร D1, D2, D3 เอกสารแต่ละฉบับผ่านการตัดคำ (word segmentation) และดึงคำหยุดออกไป ดังนั้นเหลือคำสำคัญดังนี้

D_1: คอมพิวเตอร์ สารสนเทศ คอมพิวเตอร์ คอมพิวเตอร์
D_2: อินเตอร์เน็ต คอมพิวเตอร์ อินเตอร์เน็ต ข้อมูล
D_3: ระบบ อินเตอร์เน็ต

จากนั้นหาความถี่ของคำที่ไม่ซ้ำกันในเอกสารแต่ละฉบับ แล้วหาคำที่ไม่ซ้ากันทั้งหมดในชุดเอกสาร และกำหนดค่า df และ idf ให้แต่ละคำ ดังนี้

ตารางที่ 1. ความถี่ของคำในชุดเอกสาร

เอกสาร	คำ	*tf*
D_1	คอมพิวเตอร์	3
D_1	สารสนเทศ	1
D_2	อินเตอร์เน็ต	2
D_2	คอมพิวเตอร์	1
D_2	ข้อมูล	1
D_3	ระบบ	1
D_3	อินเตอร์เน็ต	1

ตารางที่ 2. ค่า *idf* ของคำในชุดเอกสาร

คำ	*df*	*idf*
T_1: คอมพิวเตอร์	2	0.18
T_2: สารสนเทศ	1	0.48
T_3: อินเตอร์เน็ต	2	0.18
T_4: ระบบ	1	0.48
T_5: ข้อมูล	1	0.48

เวกเตอร์เอกสารทั้งหมดแทนด้วยเมตริกซ์ โดยแถวของเมตริกซ์คือเอกสารทั้งหมด และสดมภ์คือคำที่ไม่ซ้ำกันทั้งหมดในชุดเอกสาร ถ้าคำในสดมภ์ปรากฏอยู่ในเวคเตอร์เอกสารฉบับหนึ่งๆให้ค่าน้ำหนัก แต่ถ้าไม่ปรากฏคำนั้นในเอกสารก็ให้คำนั้นมีค่าน้ำหนักเป็น 0 ซึ่งเอกสารที่ไม่มีโครงสร้างสามารถถูกแทนได้อย่างเป็นระบบด้วย VSM ซึ่งอยู่ในรูปของเมตริกซ์เอกสารคำ ซึ่งจะใช้เป็น input ในขั้นตอนการจำกลุ่มเอกสารต่อไป และนำไปหาค่าความคล้ายคลึงของเอกสารคู่หนึ่งๆได้

	T_1	T_2	T_3	T_4	T_5
D_1	0.74	0.67	0	0	0
D_2	0.57	0	0.29	0	0.77
D_3	0	0	0.35	0.94	0

รูปที่ 3.8 : เมตริกซ์ เอกสาร-คำ

ในงานวิจัยนี้ได้ใช้วิธีการจัดกลุ่มเอกสารที่เรียกว่า spherical K-mean โดยประยุกต์จากขั้นตอนวิธี k-mean ซึ่งปกติวัดระยะทางระหว่างคู่ของเวคเตอร์ใดๆ โดยใช้ระยะทาง Euclidean แต่ใช้ค่าสัมประสิทธิ์ cosine แทน แนวคิดในการจัดกลุ่มเอกสารคือ หาเวคเตอร์ที่เป็นตัวแทนของกลุ่มเวคเตอร์เอกสารแต่ละกลุ่ม ให้เวคเตอร์เอกสารทั้งหมดเปรียบเทียบกับเวคเตอร์เหล่านี้เพื่อหากลุ่มที่อยู่ใกล้ที่สุดจนกว่าจะพบเงื่อนไขการหยุด

กำหนดให้ X_1 แทนเวคเตอร์เอกสาร

π_j แทนส่วนของเวคเตอร์เอกสารโดยที่

$$\bigcup_{j=1}^{k} \pi_j = \{ x_1, x_2, \ldots, x_n \} \text{ และ } \pi_j \cap \pi_l = \phi \text{ ถ้า } j \neq l$$

ทำการวัดคุณภาพของ π_j , $1 \leq j \leq k$ โดยใช้ฟังก์ชัน objective ตามสมการข้างล่างนี้

$$Q(\pi_j) = \sum_{j=1} \sum_{x \in \pi_j} x^T c_j$$

โดยที่ $x^T c_j$ แทน inner product ระหว่างเวคเตอร์เอกสารและเวคเตอร์ concept $c_j = \mathrm{m}_j / \|\mathrm{m}_j\|$, m_j คือเวคเตอร์เฉลี่ย (centroid) ของเวคเตอร์เอกสารที่บรรจุอยู่ใน π_j ขั้นตอนวิธี spherical k-means

ขั้นตอนวิธี Spherical k-means แบบลำดับ

1. เลือกจำนวนกลุ่ม k แล้วแบ่งเวคเตอร์เอกสารอย่างสุ่มออกเป็น $\pi_j^{(0)}$, $1 \le j \le k$ ให้ $c_j^{(0)}$ แทนเวคเตอร์ concept ที่สอดคล้องกับส่วนของเวคเตอร์ที่กำหนด และกำหนดค่าของ iteration $t = 0$
2. สำหรับแต่ละเวคเตอร์เอกสาร x_i หาเวคเตอร์ concept ที่อยู่ใกล้ x_i มากที่สุด และคำนวณ $\pi_j^{(t+1)}$ ใหม่ซึ่งถูกกำหนดโดย $c_j^{(t)}$ โดยที่

$$\pi_j^{(t+1)} = \{x \in x_i, 1 \le i \le n : x^T c_j^{(t)} \ge x^T c_l^{(t)}, \quad 1 \le l \le n\}, 1 \le j \le k \qquad (6)$$

3. คำนวณเวคเตอร์ concept $c_j^{(t+1)}$ ซึ่งสอดคล้องกับ $\pi_j^{(t+1)}$ ที่ได้จากขั้นตอนที่ 2
4. ถ้าฟังก์ชัน objective เปลี่ยนแปลงน้อยกว่าค่ากำหนด (threshold) แล้วจบการทำงาน ไม่เช่นนั้นเพิ่มค่า t ด้วย 1 และย้อนกลับไปขั้นตอนที่ 2

ในงานวิจัยนี้เอกสารที่ใช้ในการทดลองนำมาจากหนังสือพิมพ์เดลินิวส์ จำนวน 4800 ข่าวโดยมีข่าว 5 ชนิดคือเศรษฐกิจจำนวน1146 ข่าว ต่างประเทศจำนวน 1653 ข่าว การเมืองจำนวน 828 ข่าว พระราชสำนักจำนวน 47 ข่าว สังคมจำนวน 1126 ข่าว ซึ่งเรากำหนดให้ 1 ข่าวแทน 1 เอกสาร ในขั้นตอนแรกของการเตรียมข้อมูลต้องแยกเอกสารออกเป็นคำๆ ใช้วิธีกฎทางภาษาศาสตร์ และวิธี Longest Matching ซึ่งใช้รายการคำไทยจำนวน 32675 คำ มีการดึงเอาคำหยุดออก ผลการทดลองมีการวัดผลของการจัดกลุ่มเอกสารด้วย F-measure ซึ่งเป็น

(2) Fuzzy C-means(FCM) เป็นอัลกอริทึมที่ยอมให้ข้อมูลในแต่ละคลัสเตอร์มีการซ้อนทับกันหรือซ้ำกันได้ วิธีการนี้เป็นการจัดกลุ่มที่มีใช้อย่างแพร่หลายในงานด้านต่างๆเช่น การแพทย์ วิทยาศาสตร์ วิศวกรรมศาสตร์ โดยอาศัยการให้ค่าการเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ การได้มาซึ่งค่าการเป็นสมาชิกส่วนหนึ่งมาจากการวัดระยะทางระหว่างข้อมูลและจุดศูนย์กลางของกลุ่มเหล่านั้น การวัดระยะทางจึงมีความสำคัญต่อการจัดกลุ่ม ซึ่งวิธีการวัดระยะทางนั้นมีหลายวิธีการอาจเป็นการวัดระยะทางแบบยูคลิเดียน (Eulidence distance) หรือการวัดระยะทางแบบมหาลาโนบิส (Mahalanobis distance) สำหรับการวัดระยะทางแบบยูคลีเดียนนั้นไม่เหมาะกับข้อมูลที่เกี่ยวเนื่องกัน สำหรับการวัดระยะทางแบบมหาลาโนบิสนั้นเหมาะสำหรับกลุ่มข้อมูลที่มีข้อมูลโดดออกจากกลุ่ม(outlier) และกลุ่มข้อมูลที่มีข้อมูลหนาแน่นต่างๆ

การจัดกลุ่มแบบฟัซซี่นั้นเป็นเทคนิคในการจัดกลุ่มที่แก้ไขข้อเสียของ K-mean เนื่องจาก k-mean ไม่เหมาะกับข้อมูลที่มีความสัมพันธ์กัน(correlation) เนื่องจากข้อมูลมีโอกาสเป็นสมาชิกเพียงกลุ่มใดกลุ่มหนึ่งเท่านั้น การจัดกลุ่มแบบฟัซซี่นั้น สมาชิกของกลุ่มมีโอกาสหรือค่าการเป็นสมาชิกของข้อมูลระดับต่างๆในทุกๆกลุ่ม สำหรับการแบ่งกลุ่มแบบฟัซซี่(fuzzy clustering) Dunn ได้มีการปรับปรุงโดย Bezdek

ขั้นตอนการทำงานของฟัซซี่ซีมีน(fuzzy C-Means) ประกอบด้วย

- กำหนดกลุ่มข้อมูลที่ต้องการจัดกลุ่ม เพื่อกำหนดค่าเพื่อเป็นเงื่อนไขในการให้ข้อมูลหยุดการจัดกลุ่ม(ε) กำหนดค่าฟัซซี่พารามิเตอร์ (m) ซึ่งต้องมากกว่าหนึ่ง และกำหนดจุดศูนย์กลางเริ่มต้นของข้อมูล
- คำนวณค่าการเป็นสมาชิกของข้อมูลต่อกลุ่มข้อมูลต่างๆ
- คำนวณจุดศูนย์กลางกลุ่มข้อมูลใหม่และตรวจสอบเงื่อนไขโดยตรวจสอบค่าการเป็นสมาชิกใหม่ลบค่าการเป็นสมาชิกก่อนหน้า
- ถ้าเงื่อนไขเป็นจริงคำนวณค่าการเป็นสมาชิกและ objective function ถ้าเงื่อนไขเป็นเท็จ คำนวณค่าการเป็นสมาชิกจากจุดศูนย์กลางล่าสุด(วนรอบ)

การคำนวณ Objective Function สามารถคำนวณจาก

$$J = \sum_{i=1}^{c}\sum_{j=1}^{n}(\mu_{ij})^{m} d^{2}(X_j, Z_i)$$

โดย J แทน Objective Function ของขั้นตอนวิธีฟัซซี่ซีมีน

กำหนดให้เซตของข้อมูล $X = \{X_1, X_2, \ldots X_n\}$

n แทนจำนวนข้อมูล

c แทนจำนวนกลุ่มข้อมูล

m แทนฟัซซี่พารามิเตอร์ที่ต้องมีค่ามากกว่า 1

μ_{ij} คือค่าการเป็นสมาชิก (membership) ของข้อมูลที่ J ในกลุ่มที่ i

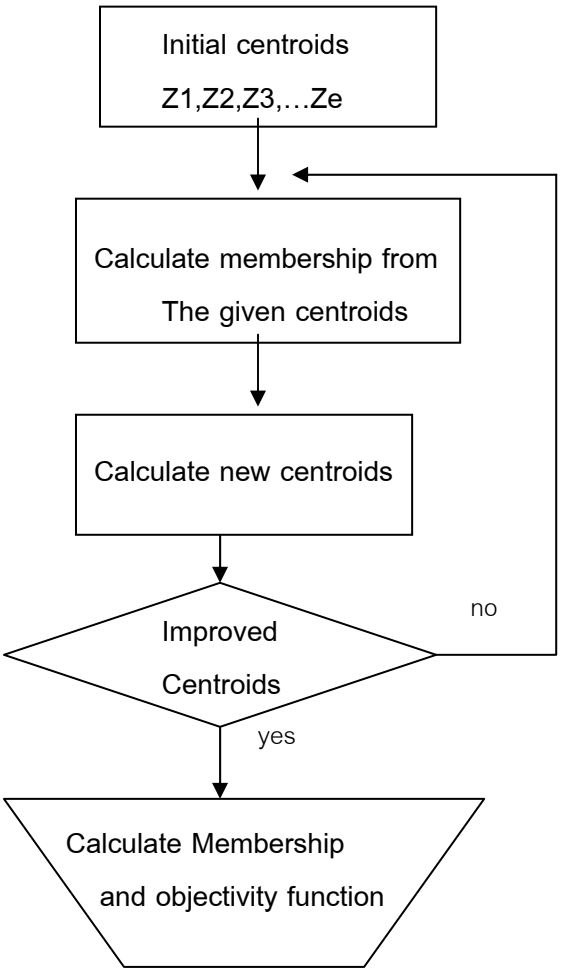
$d^{2}(X_j, Z_i)$ แทนระยะทางยกกำลังสองระหว่างข้อมูล x ที่ j และจุดศูนย์กลางของข้อมูล z กลุ่มที่ i โดย

$$Z_i = \frac{\sum_{j=1}^{n}(\mu_{ij})^{m} X_j}{\sum_{j=1}^{n}(\mu_{ij})^{m}}$$

การหาค่าการเป็นสมาชิก μ_{ij} แสดงได้จากสมการดังนี้

$$\mu_{ij} = \frac{[1/d^{2}(X_j - Z_i)]^{1/(m-1)}}{\sum_{i=1}^{c}[1/d^{2}(X_j - Z_i)]^{1/(m-1)}}$$

รายละเอียดการทำงานของฟัซซี่ซีมีนมีการทำงานดังนี้



สำหรับการวัดระยะทางระหว่างข้อมูลและจุดศูนย์กลางของข้อมูล นั้น แบบยูคลิเดียน (Euclidean distance) สามารถหาได้จากสูตร

$$ED_{ji} = \sqrt{(X_j - Z_i)(X_j - Z_i)^T}$$

โดย ED_{ji} แทนระยะทางแบบยูคลิเดียนระหว่างข้อมูล
X ที่ j และจุดศูนย์กลางข้อมูล Z กลุ่มที่ i และ T แทน Transpose matrix

สำหรับการวัดระยะทางแบบมหาลาโนบิส(Mahalanobis distance) นั้นเหมาะกับข้อมูลที่มีความสัมพันธ์ต่อกัน สามารถหาค่าได้จากสูตร

$$MD_{ji} = \sqrt{(X_j - Z_i)A^{-1}(X_j - Z_i)^T}$$

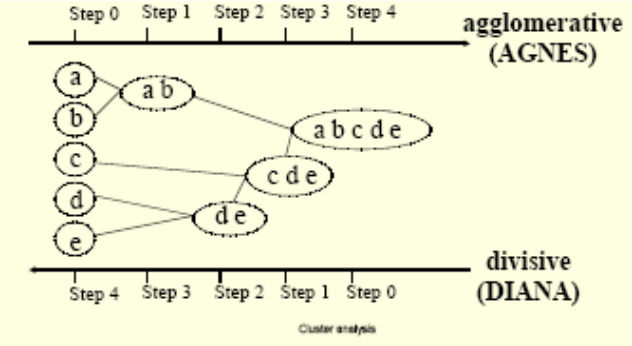
โดย MDji แทนระยะทางแบบมหาลาโนบิสระหว่างข้อมูล X ตัวที่ j และจุดศูนย์กลางข้อมูล Z กลุ่มที่ i

A คือ variance-covariance matrix คำนวณจากสมการดังนี้

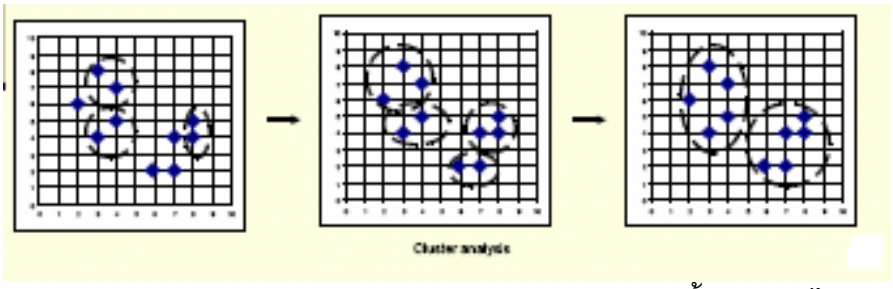
$$A = \frac{\sum_{j=1}^{n}(X_j - Z_i)^T(X_j - Z_i)}{n-1}$$

สมชาย จำปาทอง, ศาสตรา วงศ์ธนวสุ, คำรณ สุนัติ, สิรภัทร เชี่ยวชาญวัฒนา ได้ทำการศึกษาในเรื่อง "อัลกอริทึมการแบ่งกลุ่มข้อมูลโดยใช้ฟัซซี่ซีมีนกับการวัดระยะทาง" (www.cs.buu.ac.th/~deptdoc/proceedings/JCSSE2005/pdf/a-315.pdf) ผลงานวิจัยแสดงให้ทราบว่าการวัดระยะทางมีผลต่อการจัดกลุ่มแบบฟัซซี่ซีมีนและพฤติกรรมของข้อมูบมีส่วนต่อการจัดกลุ่ม ในงานวิจัยทำให้ทราบว่าในกรณีที่ข้อมูลมีข้อมูลโดด การวัดระยะทางแบบยูคลิเดียนไม่เหมาะกับข้อมูลในลักษณะนี้ แต่จะเหมาะกับข้อมูลที่การกระจายของข้อมูลอยู่ในช่วงแคบ ๆและมีความหนาแน่นไม่ต่างกันมากนัก

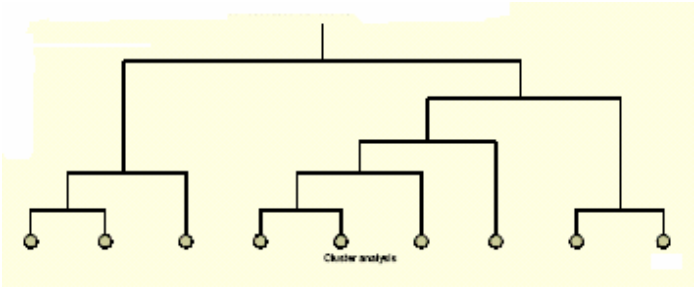
(3) Hierarchical Clustering เป็นการเกาะกลุ่มโดยระดับชั้น ซึ่งระเบียนข้อมูลถูกรวมกลุ่มโดยการใช้ระดับชั้น (Hierarchical decomposition) มีการใช้ distance matrix เป็นเงื่อนไขในการเกาะกลุ่ม วิธีการนี้ไม่ต้องการให้กำหนดค่า k แต่ต้องกำหนดเงื่อนไขในการหยุด



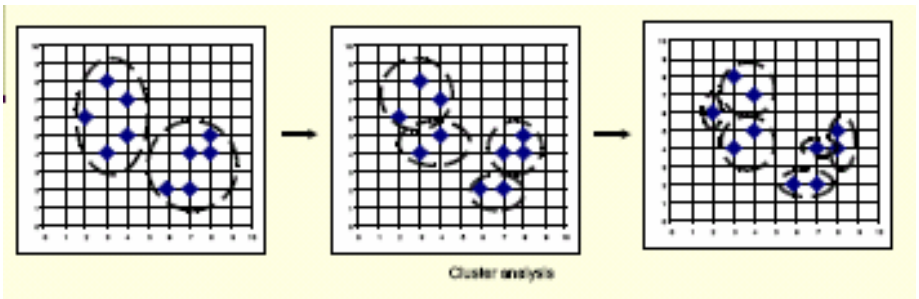
อทิเช่น AGNES(Agglomerative) นำเสนอโดย Kaufmann และ Rousseeuw(1990) ใช้วิธีการ Single-link และเมตริกซ์แสดงความต่างของข้อมูล มีการรวมจุดที่มีความต่างน้อยที่สุดเข้าด้วยกัน มีการทำซ้ำโดยทำการรวมซึ่งไม่มีการแยกออกจนกระทั่งสอดคล้องเงื่อนไขการหยุด ในกรณีที่เราทำไปไม่หยุดจะได้ว่าทุกข้อมูลจะรวมอยู่กลุ่มเดียวกัน



ผัง dendrogram เป็นการแยกกลุ่มข้อมูลเป็นระดับชั้นแบบต้นไม้ (tree of cluster) วิธีการเกาะกลุ่มของข้อมูลคือการตัดผัง dendrogram ตามระดับที่ต้องการ ส่วนประกอบที่เชื่อมกันแสดงถึงการเกาะกลุ่มที่เกิดขึ้น



วิธีการ DIANA(Divisive Analysis) นำเสนอโดย Kaufmann และ Rousseeuw(1990) ใช้วิธี Single-Link และเมตริกซ์แสดงความต่างของข้อมูล เริ่มจากจัดข้อมูลทุกตัวให้อยู่ในกลุ่มเดียวกัน ทำซ้ำ โดยแบ่งกลุ่มใหญ่ออกเป็นกลุ่มย่อย หยุดถ้าเงื่อนไขสอดคล้องการหยุด ถ้าไม่มีการตรวจสอบการหยุด ข้อมูลแต่ละตัวจะถูกจัดในกลุ่มที่ต่างกันหมด



สำหรับวิธีการ agglomerative นี้นั้นไม่เหมาะสำหรับข้อมูลปริมาณมาก ความซับซ้อนของเวลาคือ $O(n^2)$ เมื่อ n แทนจำนวนข้อมูลทั้งหมด นอกจากนี้วิธีการนี้ไม่มีการแยกกลุ่มข้อมูลออกจากกัน เมื่อรวมเล้วข้อมูลจะอยู่ในกลุ่มนั้นตลอด ซึ่งวิธีการแก้ไขคือใช้วิธีการของ BIRCH(1996) ซึ่งมีการใช้ CF-tree และตรวจสอบประสิทธิภาพของกลุ่มย่อย หรืออาจเป็นวิธีการ CURE(1998) ที่มีการเลือกจุดที่กระจายทั่วเป็นตัวแทนของกลุ่ม แล้วหดเข้าสู่จุดกลางตามอัตราส่วนที่กำหนด หรืออาจใช้วิธีการของ CHAMELEON(1990) ซึ่งผนวกทั้งวิธีการ BIRCH และ CURE เข้าด้วยกัน

ขั้นตอนพื้นฐานของ Hierarchical algorithm นั้น กำหนดภายใน 1 เซตประกอบด้วย N คลัสเตอร์ และเมตริกซ์ของความแตกต่างมีขนาด N*N มีขั้นตอนดังนี้

ขั้นตอนที่ 1 เริ่มต้นโดยการกำหนดแต่ละ item ให้ cluster ดังนั้นถ้าเรามี N ไอเท็ม ต้องมี cluster จำนวน N คลัสเตอร์เหมือนกัน เพื่อที่จะบรรจุ item ได้พอดี อนุญาตให้ความแตกต่างระหว่างคลัสเตอร์เหมือนกับความแตกต่างระหว่างคลัสเตอร์ที่บรรจุ item

ขั้นตอนที่ 2 เมื่อพบ คลัสเตอร์ 2 คลัสเตอร์ที่มีความใกล้เคียงกัน ให้ทำการรวมทั้งสองคลัสเตอร์เข้าด้วยกัน ดังนั้นเราจะได้คลัสเตอร์เพียงคลัสเตอร์เดียว

ขั้นตอนที่ 3 คำนวณความแตกต่างระหว่างคลัสเตอร์ใหม่และคลัสเตอร์เก่าของแต่ละคลัสเตอร์

ขั้นตอนที่ 4 ทำซ้ำขั้นตอนที่ 2 และ 3 จนกว่าคลัสเตอร์ทั้งหมด จะกลายเป็นเพียงคลัสเตอร์เดียว

ซึ่ง Hierarchical Algorithm เป็นลักษณะที่เป็นการจับกันเป็นกลุ่ม วิธีการนี้จะให้ทำการรวมคลัสเตอร์ไปเรื่อยๆจนให้เป็นคลัสเตอร์เดียวกัน โดยเฉพาะวิธีนี้เป็นการแบ่งแยกที่ใช้ด้านตรงข้ามของทุกๆออปเจ็กต์ภายในคลัสเตอร์มาเริ่มต้นแทน และจะแบ่งย่อยลงไปอีกจนกลายเป็นออกเจ็กต์ที่เล็กที่สุด วิธีการแบ่งแยกนั้นไม่ได้ใช้กันทั่วไปและใช้ไม่บ่อยนัก แน่นอนว่าจะต้องมีจุดไม่ครบ N ไอเท็มของกลุ่มใน single cluster แต่อาจมีสักครั้งที่ทำให้ ต้นไม้สมบูรณ์ได้ ถ้าต้องการให้มี K คลัสเตอร์ก็ต้องมีจำนวน link ที่มากที่สุดคือ K-1 ลิงค์

อัลกอริทึมของ Single-Linkage Clustering อัลกอริทึมนี้ นำเมตริกที่มีความใกล้เคียงกันมาแยกแถวและคอลัมภ์ออกจะเห็นเป็นคลัสเตอร์เดิมก่อนที่จะนำมารวมกันเป็นคลัสเตอร์ใหม่ เมตริกซ์ N*N ที่มีลักษณะที่ใกล้เคียงกันคือ D =[d(i,j)] ซึ่งภายในคลัสเตอร์นี้จะกำหนดลำดับของตัวเลขเป็น 0,1,2,…,(n-1) และ L(k) เป็น level ของ k clustering m จะแสดงถึงตัวเลขที่ต่อเนื่องกันและความใกล้เคียงกันระหว่าง cluster (r) และ (s) สามารถแสดงได้เป็น d[(r)(s)] ขั้นตอนการทำงานเป็นดังนี้

ขั้นตอนที่ 1 เริ่มต้นด้วยการแยกคลัสเตอร์ที่มี level L(0) =0 และเลขที่อยู่ในลำดับที่ m=0

ขั้นตอนที่ 2 หาคู่ลำดับของคลัสเตอร์ที่มีความแตกต่างกันน้อยที่สุด โดยที่คู่ลำดับจะขึ้นอยู่กับ d[(r)(s)] = min d[(i)(j)] โดยที่จะต้องเป็นค่าที่น้อยที่สุดของคู่ลำดับทั้งหมดของทุกๆคลัสเตอร์ที่อยู่ในนั้น

ขั้นตอนที่ 3 เพิ่มเลขลำดับจาก m = m+1 และทำการรวมคลัสเตอร์ (r) และ (s) ให้เป็นคลัสเตอร์เดียวกัน เป็นคลัสเตอร์ m ใหม่ กำหนด level ของคลัสเตอร์เป็น

L(m) = d[(r)(s)]

ขั้นตอนที่ 4 แก้ไข proximity เมตริกซ์ D โดยตัดแถวและคอลัมภ์ที่ตรงกันของคลัสเตอร์ (r) และ (s) ออกและทำการเพิ่มแถวและคอลัมภ์ที่ตรงกันนี้ไปยังคลัสเตอร์ใหม่ ความใกล้เคียงกันระหว่างคลัสเตอร์ใหม่ denoted(r,s) และคลัสเตอร์ (k) เก่าสามารถนิยามได้ใหม่ดังนี้ d[(k)(r,s)] = min d[(k)(r), d[(k)(s)]

ขั้นตอนที่ 5 ถ้ามีเพียงคลัสเตอร์เดียว ให้จบการคำนวณแต่ถ้าไม่ ก็ให้ทำซ้ำตั้งแต่ขั้นตอนที่ 2-5

ตัวอย่างที่ 3.4 การใช้ Hierarchical clustering ในการหาระยะทางระหว่างเมืองต่างๆ โดยระยะทางมีหน่วยเป็นกิโลเมตร ภายในประเทศอิตาลี มีการใช้อัลกอริทึม single-linkage

กำหนดให้ BA ,FI, MI ,NA , RM ,TO แทนเมืองในประเทศอิตาลี

Input distance matrix โดยทุกๆคลัสเตอร์มีค่า L=0

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

เมืองสองเมืองที่มีระยะทางใกล้กันที่สุดคือเมือง MI และ TO ซึ่งมีระยะทางห่างกัน 138 กิโลเมตร ทำการรวมเมืองทั้งสองเข้าด้วยกันเป็นคลัสเตอร์เดียวกันคือ MI/TO เพราะฉะนั้น Level ของคลัสเตอร์ใหม่น้มีค่าเท่ากับ L(MI/TO) = 138 และค่า m= 1

ต่อจากนั้นคำนวณหาระยะทางของเมืองอื่นๆโดยรอบ ซึ่ง single-linkage clustering มีกฎอยู่ว่า ถ้าบริเวณโดยรอบออปเจ็กซ์นั้นมีค่าเท่ากันหรือน้อยกว่าแต่ต้องน้อยที่สุดของ

หลังจากที่เรานำ MI มารวมกับ TO เราจะได้เมตริกซ์ใหม่ดังนี้

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

ซึ่ง min d(i,j) = d(NA,RM) = 219 จะนำ NA และ RM มารวมเข้าด้วยกัน เป็นคลัสเตอร์ใหม่ เรียกว่า NA/RM โดยมีค่า L(NA/RM) = 219 และ m =2

	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0

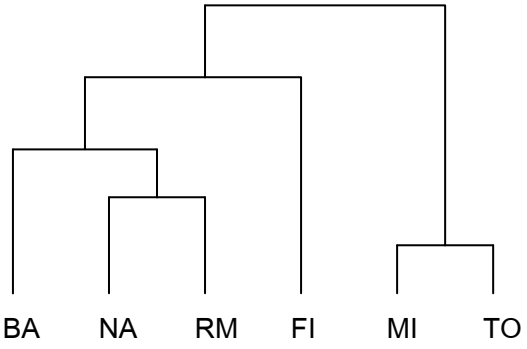
ซึ่ง min d(i,j) = d(BA,NA/RM) = 255 จะนำ BA และ NA/RM มารวมกันจะได้เป็นคลัสเตอร์ใหม่ เรียกว่า BA/NA/RM โดยมีค่า L(BA/NA/RM) = 255 และ m = 3

	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

ซึ่ง min d(i,j) = d(BA/NA/RM.FI) = 268 จะนำ BA/NA/RM และ FI มารวมกันจะได้เป็นคลัสเตอร์ใหม่ เรียกว่า BA/NA/RM/FI โดยมีค่า L(BA/NA/RM/FI) = 268 และ m = 4

	BA/NA/RM/FI	MI/TO
BA/NA/RM/FI	0	295
MI/TO	295	0

สุดท้ายเราจะทำการรวม 2 คลัสเตอร์สุดท้ายได้ level = 295 สามารถเขียนเป็น hierarchical tree ได้ดังนี้



จุดอ่อนที่สำคัญที่สุดของวิธีการนี้คือ เราไม่สามารถรู้ระยะทางที่แน่นอนได้ ซึ่งจะต้องนำมาใช้ในการคำนวณเลขเชิงซ้อนที่น้อยที่สุดคือ $O(n^2)$ โดยที่ n คือจำนวนทั้งหมดของออปเจ็กต์ และเราไม่สามารถที่จะเปลี่ยนแปลงแก้ไขในส่วนที่ได้กระทำไปแล้ว

(4) Mixture of Gaussians เป็นวิธีการอีกวิธีหนึ่งที่ใช้ในการจัดการปัญหาของ clustering โดยใช้วิธีการของ model-based ซึ่งประกอบด้วย การใช้โครงสร้างชนิดใดชนิดหนึ่งของคลัสเตอร์ และพยายามทำให้ดีขึ้นและทำให้เหมาะสมระหว่างข้อมูลและโครงสร้าง ในส่วนของการปฏิบัติแต่ละคลัสเตอร์จะใช้หลักการทางคณิตศาสตร์ในการอธิบายปัจจัยกำหนดในการจำแนกอาจเป็น a Gaussian เป็นแบบต่อเนื่องหรือ a Poisson แบบไม่ต่อเนื่อง เพราะฉะนั้นเซตของข้อมูลทั้งหมดจะสร้างแบบการผสมผสานของวิธการจำแนก แต่ละแบบของการจำแนกจะใช้โครงสร้างพิเศษเฉพาะ ของคลัสเตอร์ ในการอ้างอิงถึงส่วนประกอบของการจำแนกเสมอ

ลักษณะเฉพาะของ Mixture Model นั้น component distribution มีจุดสูงที่สุด และ mixture model ต้องปิดบังข้อมูลได้เป็นอย่างดี จุดเด่นของวิธีการนี้นั้นเหมาะที่ใช้ในการวิเคราะห์ข้อมูลทางด้านสถิติ สามารถที่จะเปลี่ยนแปลงหรือเลือกวิธีการจำแนกได้ อีกทั้งประเมินความหนาแน่นของข้อมูลภายในของแต่ละคลัสเตอร์ได้ รวมทั้งง่ายต่อการที่จะใช้ในการจัดแบ่งหมวดหมู่

อัลกอริทึมของวิธีการนี้มีขั้นตอนดังนี้

ขั้นตอนที่ 1 เลือกส่วนประกอบ(the Gaussian) โดยการสุ่มจากความน่าจะเป็น $P(\infty)$ เป็นจุดที่ใช้ทดลอง $N[\mu_i, \delta^2 I]$

ความน่าจะเป็น $P(X/\mu_i) = \sum_i P(\infty_i)P(X/\infty_i, \mu_1, \mu_2, ..., \mu_k)$

ต่อมามีการประยุกต์ใช้อัลกอริทึมที่ง่ายกว่าอทิเช่น EM (Expectation-Maxixization) ที่ใช้ในการหา the mixture ของ Gaussian สามารถใช้โครงสร้างของเซต

สมมุติให้ X_k คือคะแนนของนักเรียนภายในห้อง โดยมีความน่าจะเป็นดังนี้

X1 = 30 P(X1) = 0.5

X2 = 18 P(X2) = μ

X3 = 0 P(X3) = 2 μ

X4 = 23 P(X4) = 0.5-3 μ

กรณีที่ 1 : คะแนนมีการกระจายอยู่ในหมู่ของนักเรียน

โดย X1 : a students

X2 : b students

X3 : c students

X4 : d students

ดังนั้นความน่าจะเป็นดังนี้

$$P(a,b,c,d \mid \mu) \infty (0.5)^a * \mu^b * (2\mu)^c * (0.5-3\mu)^d$$

จะเห็นได้ว่าค่าที่มากที่สุดในฟังก์ชันนี้ โดยการคำนวณด้วยสูตร $\frac{\partial P}{\partial \mu} = 0$

นำค่าที่ได้จากการคำนวณไปแทนในอัลกอริทึม

$$P(L) = \log(0.5)^a + \log(\mu)^b + \log(2\mu)^c + \log(0.5-3\mu)^d$$

$$\frac{\partial P_L}{\sigma\mu} = \frac{b}{\mu} + \frac{2c}{2\mu} - \frac{3d}{\frac{1}{2} - 3\mu} = 0$$

$$\mu = \frac{b+c}{6(b+c+d)}$$

ในกรณีที่เราให้ a=14 , b=6 , c=9 และ d=10 เราสามารถคำนวณได้ $\mu = \frac{1}{10}$

กรณีที่ 2 : สังเกตได้ว่าเกิดการกระจายของคะแนนของนักเรียน

ให้ x1 + x2 : h students
x3 : c students
x4 : d students

ดังนั้นเราจะทำการแพร่กระจายโดยแบ่งเป็น 2 ขั้นตอน

ขั้นตอนที่ 1 สิ่งที่เราคาดคะเน

$$\mu \to a = \frac{\frac{1}{2}}{\frac{1}{2}+\mu}h, b = \frac{\mu}{\frac{1}{2}+\mu}h$$

ขั้นตอนที่ 2 ทำให้มีจำนวนมากที่สุด

$$a,b \to \mu = \frac{b+c}{6(b+c+d)}$$

การทำให้แพร่กระจายไปทั่วสามารถช่วยแก้ปัญหาการทำซ้ำได้

สรุปได้ว่า EM algorithm ที่ใช้กับ mixture of Gaussian ประกอบด้วยขั้นตอนดังต่อไปนี้
ขั้นตอนที่ 1 : Initialize parameters:

$$\lambda_0 = \{\mu_1, \mu_2, ..., \mu_k, p_1, p_2, ...p_k\}$$

ขั้นตอนที่ 2 : E-step:

$$p(\infty_j \mid X_k, \lambda_t) = \frac{p(X_k, \lambda_t)\, p(\infty_j \mid \lambda_t)}{p(X_k, \lambda_t)} = \frac{p(X_k \mid \infty_i, \mu_i^{\,t}.\sigma^2)\, p_i^{\,t}}{\sum_k p(X_k \mid \infty_i, \mu_i^{\,t}.\sigma^2)\, p_i^{\,t}}$$

ขั้นตอนที่ 3 : M-step:

$$\mu_i^{(t+1)} = \frac{\sum_k p(\infty_i \mid X_k, \lambda_t) x_k}{\sum_k p(\infty_i \mid X_k, \lambda_t)}$$

$$p_i^{(t+1)} = \frac{\sum_k p(\infty_i \mid X_k, \lambda_t)}{R}$$ โดย R เป็นจำนวนของเรคอร์ด

(5) Genetic Algorithm ถูกคิดค้นขึ้นโดย John Holland ในปีค.ศ. 1975 เป็นการนำขบวนการวิวัฒนาการของสิ่งมีชีวิตมาประยุกต์ใช้ในงานปัญญาประดิษฐ์ เพื่อใช้สำหรับหาคำตอบที่ดีที่สุด(Optimization) ของปัญหาต่างๆจากจำนวนคำตอบที่เป็นไปได้ทั้งหมดของการแก้ปัญหานั้น ปัจจุบันได้มีการประยุกต์ใช้จีเนติกอัลกอริทึมในงานต่างๆอย่างแพร่หลายเช่นการใช้จีเนติกอัลกอริทึมในการแก้ปัญหาทางคณิตศาสตร์ , การแก้ปัญหาการหาเส้นทางที่มีระยะทางที่สั้นที่สุด , การจัดตารางสอน , การควบคุมหุ่นยนต์ รวมทั้งการค้นคืนสารสนเทศในส่วนของการปรับปรุงข้อคำถาม(query) ให้สามารถค้นคืนข้อมูลได้ตรงตามความต้องการของผู้ใช้มากขึ้น รวมทั้งในเรื่องอื่นๆ

จีเนติกอัลกอริทึมประกอบด้วยองค์ประกอบที่สำคัญ 5 ส่วนคือ

a. รูปแบบโครโมโซมที่ใช้ในการนำเสนอทางเลือกที่สามารถจะเป็นได้ของแต่ละปัญหา
b. วิธีสร้างประชากรต้นกำเหนิดของทางเลือกที่สามารถจะเป็นไปได้
c. ฟังก์ชันสำหรับประเมินค่าความเหมาะสมเพื่อให้คะแนนแต่ละทางเลือก
d. จีเนติกโอเปอเรเตอร์ซึ่งใช้ในการปรับเปลี่ยนองค์ประกอบของข้อมูลตลอดกระบวนการ ได้แก่ การคัดเลือก การครอสโอเวอร์ และการมิวเตชั่น
e. ค่าพารามิเตอร์ต่างๆที่ต้องใช้สำหรับจีเนติกอัลกอริทึม เช่น ขนาดของประชากร , ความน่าจะเป็นของการใช้จีเนติกโอเปอเรอเตอร์ และจำนวนรุ่นเป็นต้น

การทำงานของจีเนติกอัลกอริทึม

เริ่มจากการกำหนดฟังก์ชันความหมาะสมรวมทั้งรูปแบบโครโมโซมเสียก่อน จากนั้นจึงเริ่มสร้างประชากรต้นกำเนิดตามรูปแบบโครโมโซมที่ได้กำหนดไว้แล้ว เมื่อได้ประชากรต้นกำเนิดแล้วก็ทำการวัดค่าความเหมาะสมของแต่ละโครโมโซม เพื่อคัดเลือกเข้าสู่กระบวนการจีเนติกโอเปอร์เรเตอร์โดยทำการคัดเลือกเอาเฉพาะโครโมโซมที่มีค่าความเหมาะสมเป็นที่น่าพอใจชุดหนึ่งเก็บไว้ โครโมโซมที่คัดเลือกไว้นั้นจะถูกนำมาทำการคลอสโอเวอร์และมิวเตชันได้เป็นโครโมโซมชุดใหม่ ต่อจากนั้นจะนำโครโมโซมชุดใหม่นี้มาวัดค่าความเหมาะสมเพื่อทำการคัดเลือกและดำเนินการต่อไปจนสิ้นสุดตามเงื่อนไขที่กำหนดไว้ ผลของการทำงานจะได้โครโมโซมที่มีค่าความเหมาะสมเป็นที่น่าพอใจ

ตัวอย่างที่ 3.5 การค้นคืนสารสนเทศโดยใช้จีเนติกอัลกอริทึม

สมมุติมีเอกสาร 5 ฉบับประกอบด้วยคำสำคัญดังนี้

DOC1 ={Database, Query, Data Retrieval , Computer,Network, DBMS}

DOC2={Artificial Intelligence, Internet, Indexing, Natural Language Processing}

DOC3={Database , Expert System, Information Retrieval System, Multimedia}

DOC4={Fuzzy Logic, Neural Network, Computer Networks}

DOC5-{Object-Oriented, DBMS , Query ,Indexing}

นำคำสำคัญทั้งหมดมาจัดเรียงลำดับจากน้อยไปมากรวม 16 คำดังนี้

Artificial Intelligence , Computer Network, Data Retrival, Database
DBMS, Expert System , Fuzzy Logic, Indexing
Information Retrieval System, Internet,Multimedia,Natural Language Processing,
Neural Network, Object Oriented, Query, Relational Database

นำเสนอรูปแบบโครโมโซมดังนี้

DOC1=0110100000000011

DOC2=1000000101010000

DOC3=0001010010100000

DOC4=0100001000001000

DOC5=0000100100000110

โครโมโซมชุดแรกที่ได้มานี้จะเรียกว่าประชากรต้นกำเนิด ซึ่งจะนำไปผ่านกระบวนการจีแนติกต่อไป ความยาวของโครโมโซมเหล่านี้จะขึ้นอยู่กับจำนวนคำสำคัญของชุดเอกสารทั้งหมดที่ตรงตามข้อเรียกร้อง(query) จากตัวอย่างนี้มีความยาวเท่ากับ 16 บิต

ต่อจากนั้นนำมาวัดค่าความเหมาะสม เพื่อประเมินแต่ละทางเลือกมีความเหมาะสมสามารถใช้แก้ปัญหาได้ดีเพียงใด โดยเลือกใช้เวคเตอร์สเปซโมเดลที่จะกำหนดว่าเอกสารใดตรงหรือคล้ายคลึงกันกับข้อเรียกร้องนั้น อาจเลือกใช้ Dice coefficient หรือ Cosine coefficient หรือ Jaccard coefficient ก็ได้ ซึ่งผลที่ได้จะอยู่ระหว่างค่า 0.0 ถึง 1.0 ถ้าเอกสารใดเข้าใกล้ 1.0 หมายถึงเอกสารและข้อคำถามนั้นเหมือนกันมีความสัมพันธ์กันมาก ค่าที่ได้นี้เรียกว่า ค่าความเหมาะสม(fitness)

หลังจากได้ค่าความเหมาะสมของแต่ละโครโมโซมแล้ว ขั้นตอนต่อไปคือการคัดเลือกสายพันธ์ ซึ่งเป็นไปตามหลักการอยู่รอดของสิ่งที่เหมาะสมที่สุด(Survival of the fittest) โดยโครโมโซมที่มีค่าความเหมาะสมเป็นที่น่าพอใจจะได้รับการคัดเลือกไว้

ต่อจากนั้นนำมาครอสโอเวอร์(Crossover) คือการนำโครโมโซม2โครโมโซมมาทำการผสมกันเพื่อให้ได้ค่าโครโมโซมใหม่เพื่อพยายามสร้างทางเลือกใหม่และปรับปรุงทางเลือกให้ดีขึ้น เป็นการรวมลักษณะที่ดีของแต่ละโครโมโซมเข้าด้วยกัน โครโมโซมที่มีค่าความเหมาะสมสูงกว่าจะถูกเลือกมาครอสโอเวอรร์มากกว่า

การครอสโอเวอร์แบบหนึ่งตำแหน่ง โดยครอสโอเวอร์ ณ ตำแหน่งที่ 8
ก่อนทำการครอสโอเวอร์

101111110011101

100110011110000

หลังการทำการครอสโอเวอร์

101111111110000

100110010011101

ในบางครั้งอาจกระทำการผ่าเหล่าหรือการมิวเตชัน(Mutation) คือการนำโครโมโซมเก่ามาสุ่มแก้ไขบางส่วนของโครโมโซม เช่นกระทำให้บิตบางบิตเปลี่ยนไป ทำให้ได้โครโมโซมใหม่ที่มีสายพันธ์ต่างไปจากเดิม ซึ่งการกระทำเช่นนี้อาจทำให้โครโมโซมดีขึ้นหรือเลวลงก็ได้

ก่อนการทำการมิวเตชั่น

101111110011101

สุ่มเลือกเปลี่ยนโครโมโซมตำแหน่งที่10

101111110111101

จีเนติกอัลกอริทึม จะทำเป็นวัฏจักรหมุนเวียนอยู่จนกระทั่งถึงจุดหนึ่งที่ตรงตามเงื่อนไขตามที่กำหนด หรือสิ้นสุดเมื่อพบคำตอบที่ดีที่สุดแล้ว หรือถึง threshold ตาที่ได้กำหนดไว้ล่วงหน้า

แบบฝึกหัด

1. จงอธิบายถึงวิธีวัดความคล้ายคลึงกันของวิธี Simple Coefficient , Dice's Coefficient , Jaccard's Coefficient , Cosine Coefficient และ Overlap coefficient
2. จงอธิบายถึงวิธีการวัดความไม่คล้ายคลึงกันโดยใช้สัมประสิทธิ์ของ Jaccard มาพอเข้าใจ
3. จงอธิบายถึงวิธีการวัดความไม่คล้ายคลึงกันโดยใช้สัมประสิทธิ์ของ Dice's Coefficient มาพอเข้าใจ
4. จงอธิบายถึงวิธีการแบ่งกลุ่มแบบ Monothetic และ polythetic ว่าเหมือนหรือต่างกันอย่างไร
5. จงอธิบายถึงความแตกต่างระบ่าง Exclusive Class และ Overlapping Class
6. การจัดแบ่งกลุ่มแบบ Ordered Classification เป็นอย่างไร จงเขียนรูปภาพประกอบการอธิบาย
7. Clustering คืออะไร จงอธิบายพอเข้าใจ
8. ประโยชน์ของ Clustering algorithm สามารถนำไปประยุกต์กับงานในเรื่องใดบ้าง
9. หลักเกณฑ์ที่ใช้เลือกวิธีที่เหมาะสมในการคลัสเตอร์มีเงื่อนไขอย่างไรบ้าง
10. จงอธิบายถึงวิธีการคลัสเตอร์ริ่งที่เรียกว่า Graph Theoretic Method มาพอเข้าใจ
11. จงอธิบายถึงวิธีการคลัสเตอร์ริ่งที่เรียกว่า Single Link Method มาพอเข้าใจ
12. จงอธิบายถึงอัลกอริทึมของ Rocchio มาพอเข้าใจ
13. จงอธิบายถึงอัลกอริทึม K-means มาพอเข้าใจ
14. จงอธิบายถึงอัลกอริทึมของวิธีการ PAM มาพอเข้าใจ
15. จงอธิบายถึงอัลกอริทึมของวิธีการ Fuzzy C-means(FCM) มาพอเข้าใจ
16. จงอธิบายถึงอัลกอริทึมของวิธีการ Single-Linkage Clustering มาพอเข้าใจ
17. จงอธิบายถึงอัลกอริทึม Genetic มาพอเข้าใจ

บรรณานุกรม

อุไร ทองหัวไผ่ ,"โครงสร้างโปรแกรม" , มหาวิทยาลัยรามคำแหง

สุมาลี เมืองไพศาล ,"การจัดการข้อมูล และการเรียกใช้ข้อมูล" ,สำนักพิมพ์มหาวิทยาลัยรามคำแหง ,CS337 ,2535

บังอร กลับบ้านเกาะ, " การค้นคืนสารสนเทศออนไลน์โดยใช้จีเนติกอัลกอริทึม" ,Technical Journal ,Vol11.No7,March-June, 2000

Sanpawat Kantabutra and Alva L.Couch ,"Parallel K-means Clustering Algorithm on NOW's" , Department of Computer Science Tufts University, Medford, Massachusetts, www.nectec.or.tn/NTJ/n06/papers/No6_short_1.pdf

ดร.กรุง สินอภิรมย์สราญ, "การวิเคราะห์การเกาะกลุ่มโดยวิธีแบ่งกั้นและระดับชั้น" ,ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

สมชาย จำปาทอง, ศาสตรา วงศ์ธนวสุ, คำรณ สุนัติ, สิรภัทร เชี่ยวชาญวัฒนา, "อัลกอริทึมการแบ่งกลุ่มข้อมูลโดยใช้ฟัซซี่ซีมีนกับการวัดระยะทาง" , The Joint Conference on Computer Science and Software Engineering. November 17-18, 2005 , ,(www.cs.buu.ac.th/~deptdoc/proceedings/JCSSE2005/pdf/a-315.pdf)

ฆนาศัย กรึงไกร และ ชุลีรัตน์ จรัสกุลชัย , "การจัดกลุ่มเอกสารข้อความภาษาไทยด้วยขั้นตอนวิธี Spherical K-Means แบบขนานบนพิรุณลีนุกซ์คลัสเตอร์ " ,Intelligent Information Retrieval and Database Laboratory, Department of Computer Science,Faculty of Science Kasetsart University,Bangkok,10900,Thailand