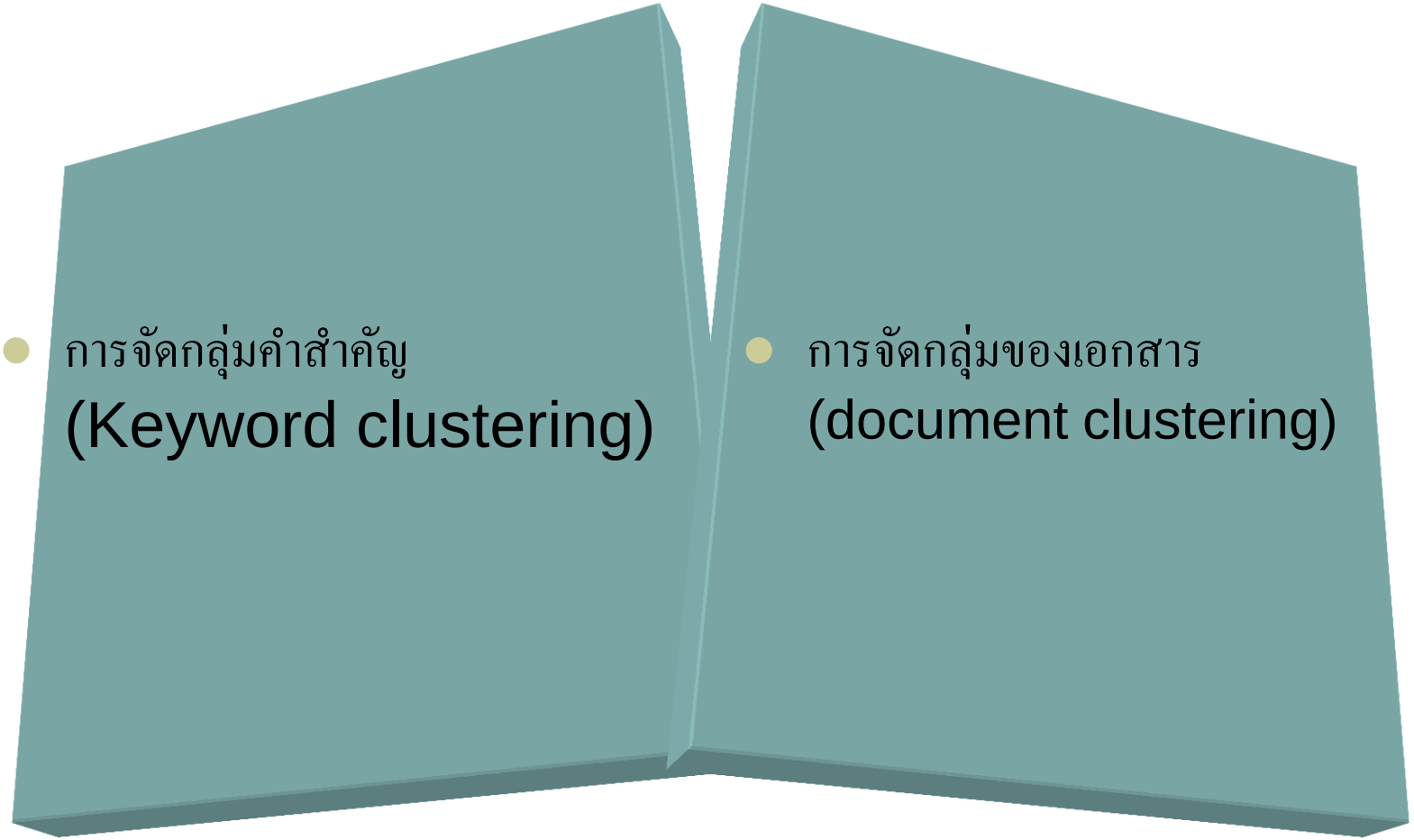
ครั้งที่ 5

การแบ่งกลุ่มข้อมูลอัตโนมัติ

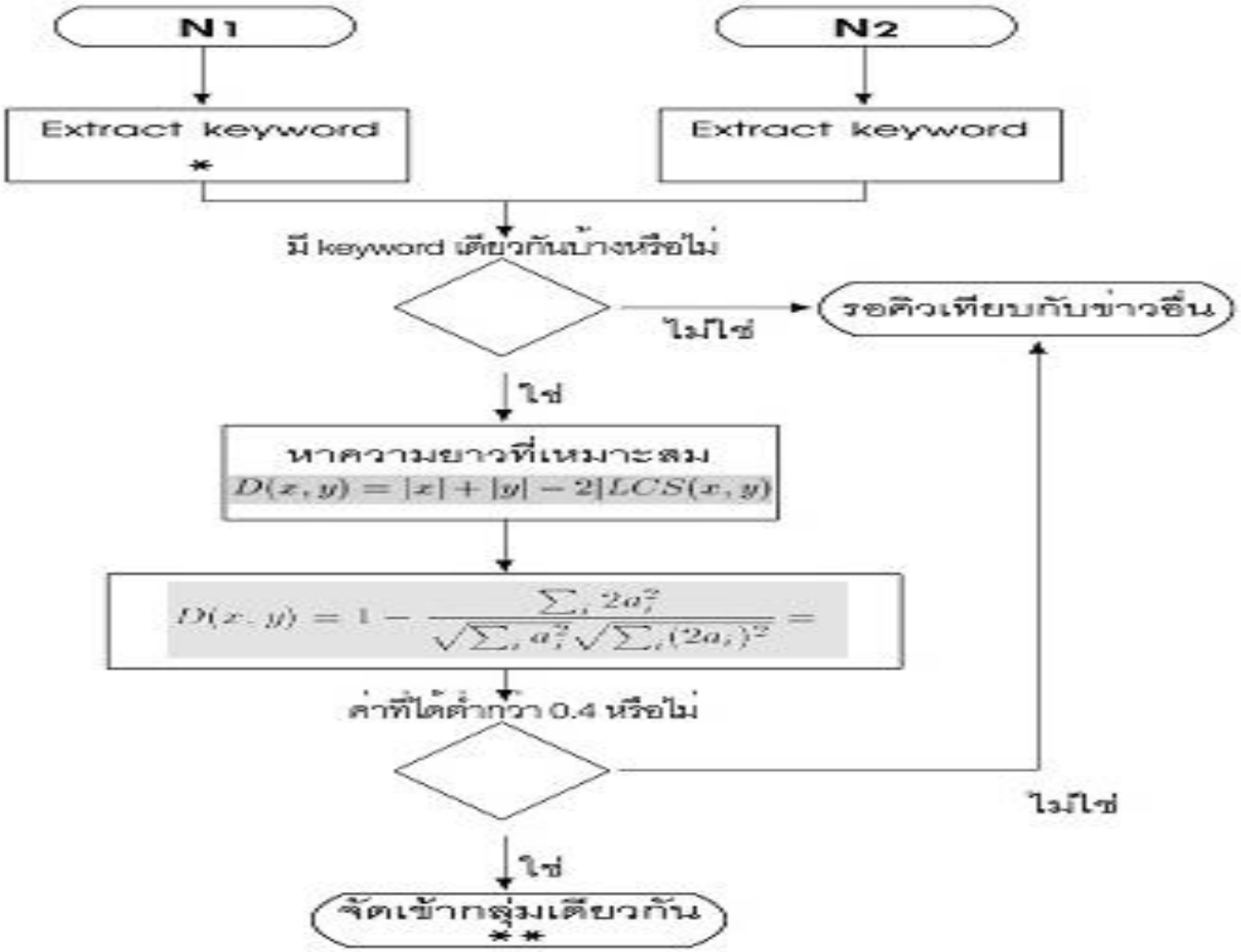
การแบ่งกลุ่มข้อมูลอัตโนมัติในระบบ IR



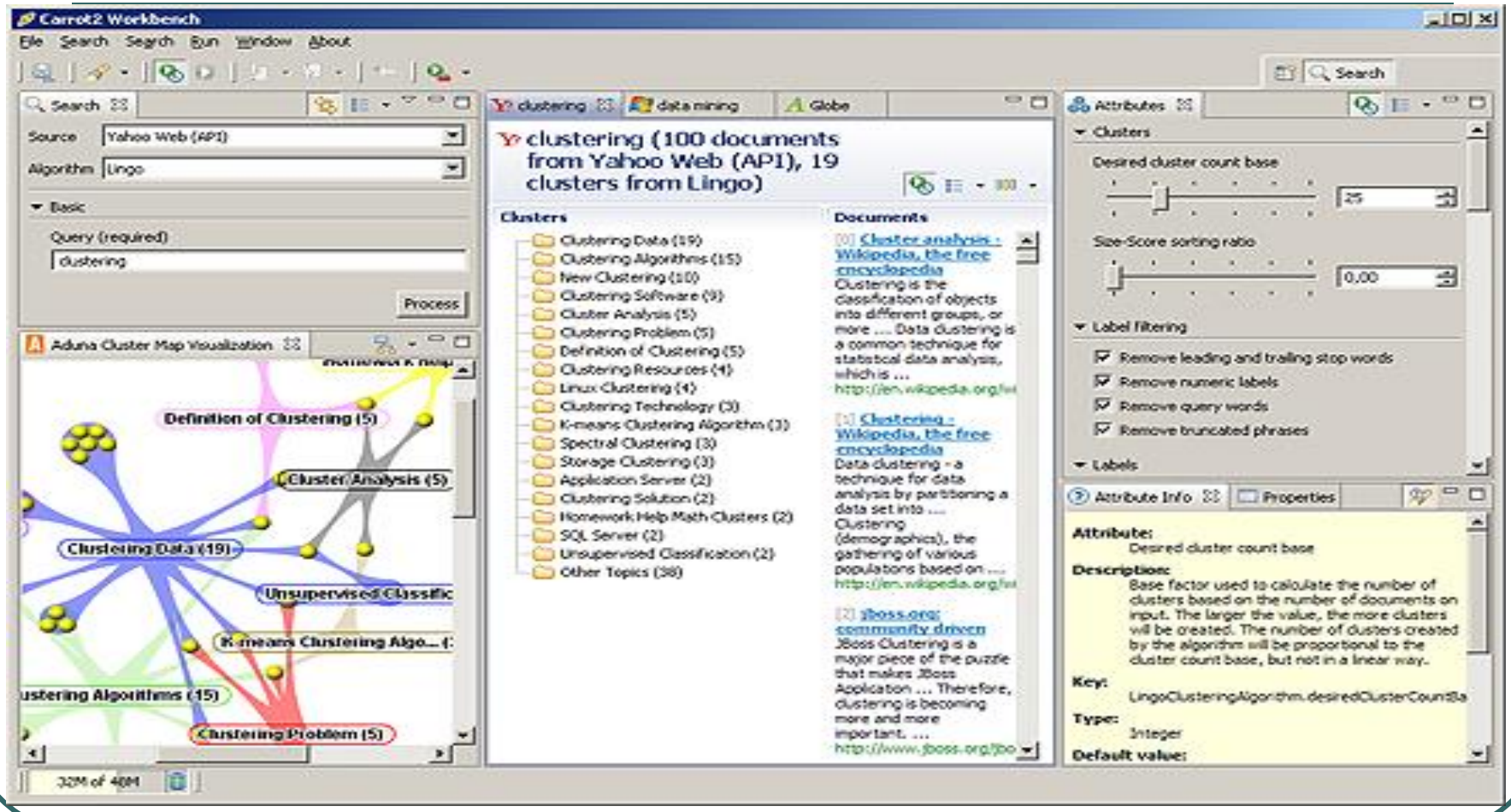
- การจัดกลุ่มคำสำคัญ (Keyword clustering)
- การจัดกลุ่มของเอกสาร (document clustering)


keyword clusterer
5 keywords ranker
1 keyword ranker
1
2
3
Google SERP

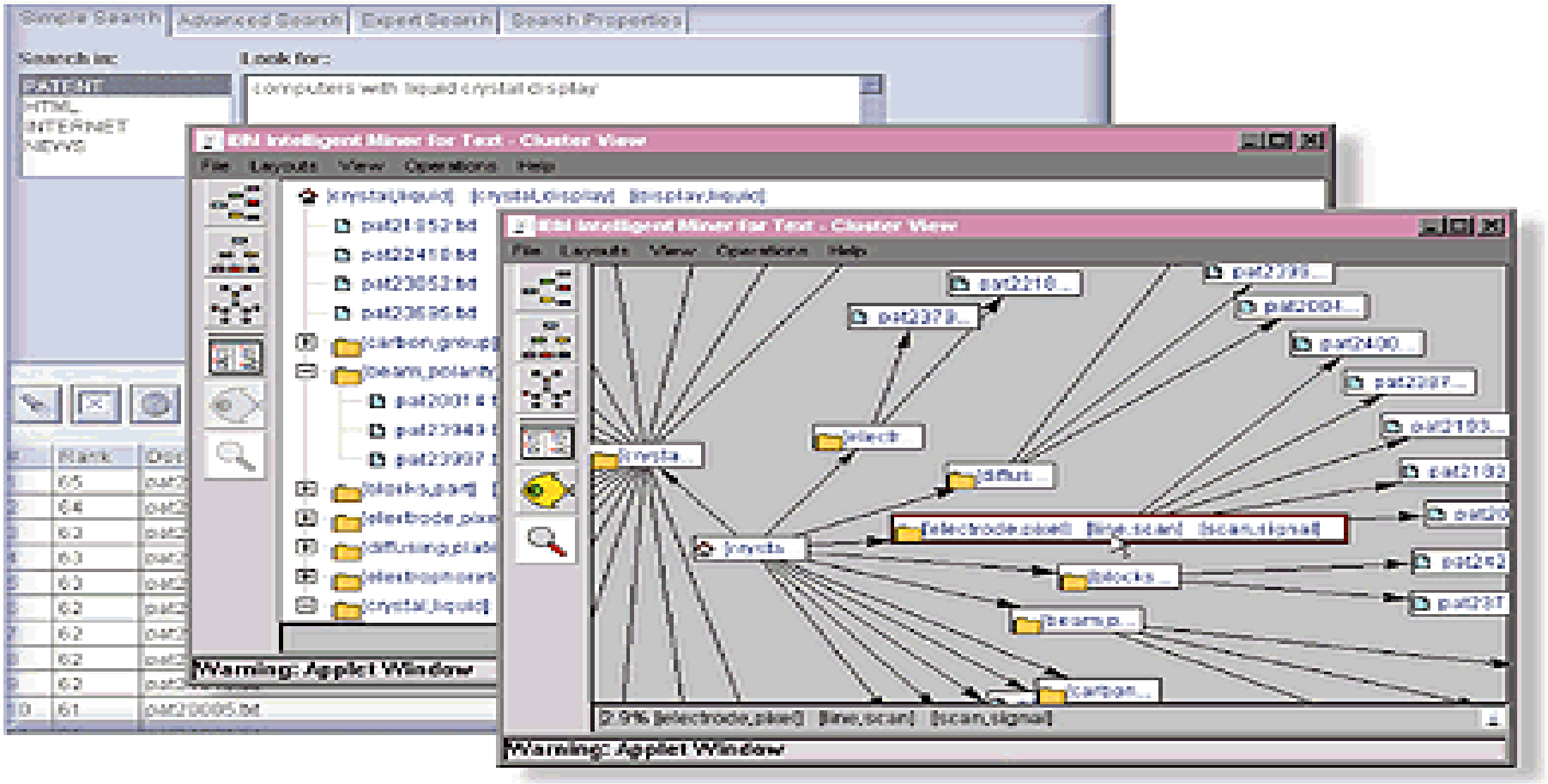
Clustering Architecture



Document Clustering Workbench screenshot



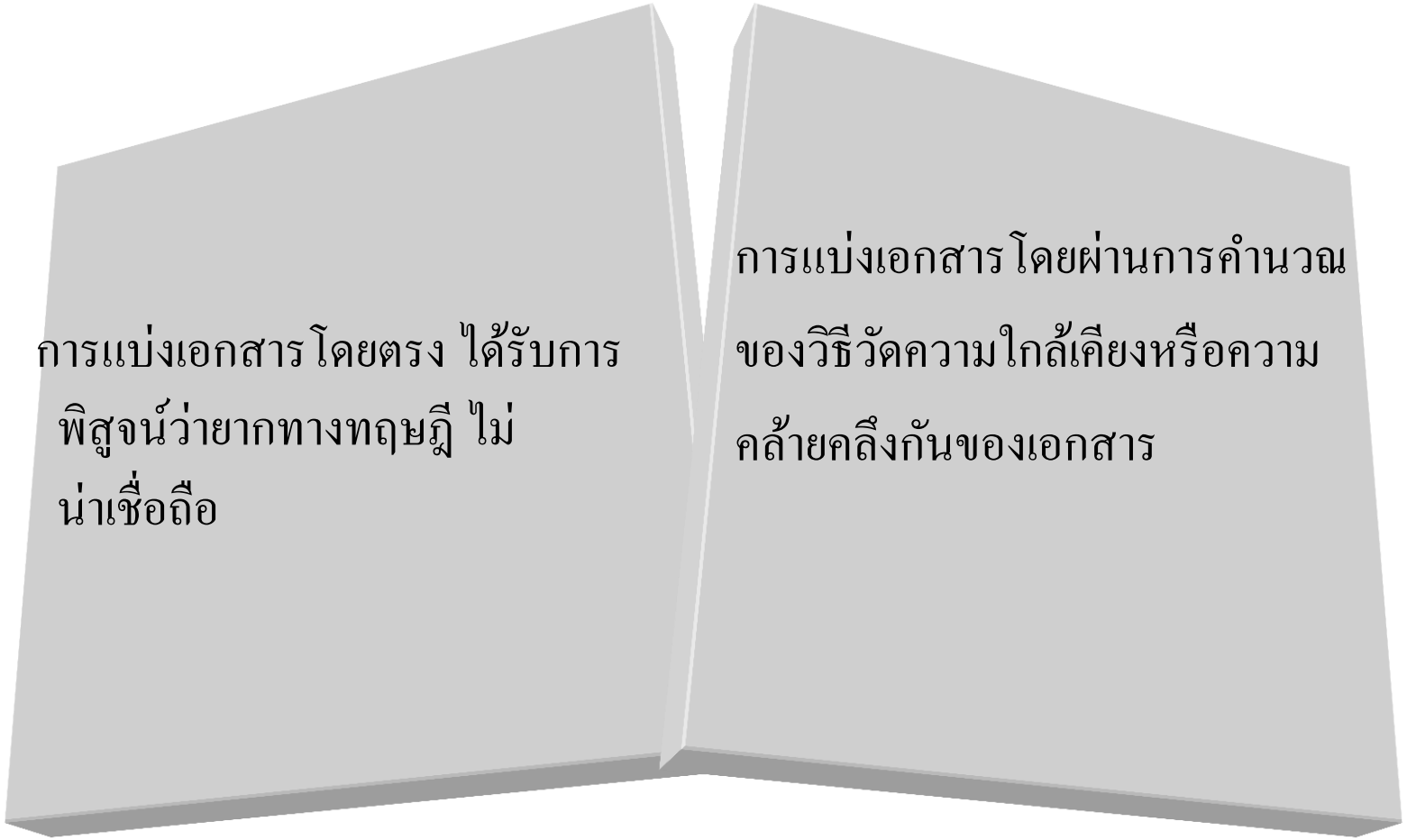
Hierarchical document clustering



R.M.Hayes ได้กล่าวไว้ว่า

- “โครงสร้างของการจัดหมวดหมู่ จะเป็นการรวมกลุ่มของไอเท็ม(item) ซึ่งอาจเป็นเอกสารหรือตัวแทนเอกสารเข้าด้วยกัน ซึ่งไอเท็มเหล่านี้จะถูกจัดให้รวมกลุ่มเปรียบเสมือนให้เป็นหน่วยหนึ่ง ซึ่งไอเท็มเหล่านี้อาจจะสูญเสียคุณสมบัติเฉพาะไปบ้างไม่มากก็น้อย”

โครงสร้างทางตรรกศาสตร์ **(Logical Organization)**



วิธีวัดความคล้ายคลึง
(Measures of Association)

ได้หลายวิธีดังนี้

Simple Coefficient

- เป็นรูปแบบที่ง่ายที่สุด คือ $|X \cap Y|$
- เป็นจำนวนของคำดรรชนีที่อยู่ทั้งในเอกสาร X และ Y วิธีนี้ไม่ได้นำเอาขนาดของ X และ Y มาพิจารณา

สมมุติ $X = \{1, 2, 3\}$

$Y = \{1, 4\}$

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

$|X \cap Y| = 1$

Dice's Coefficient

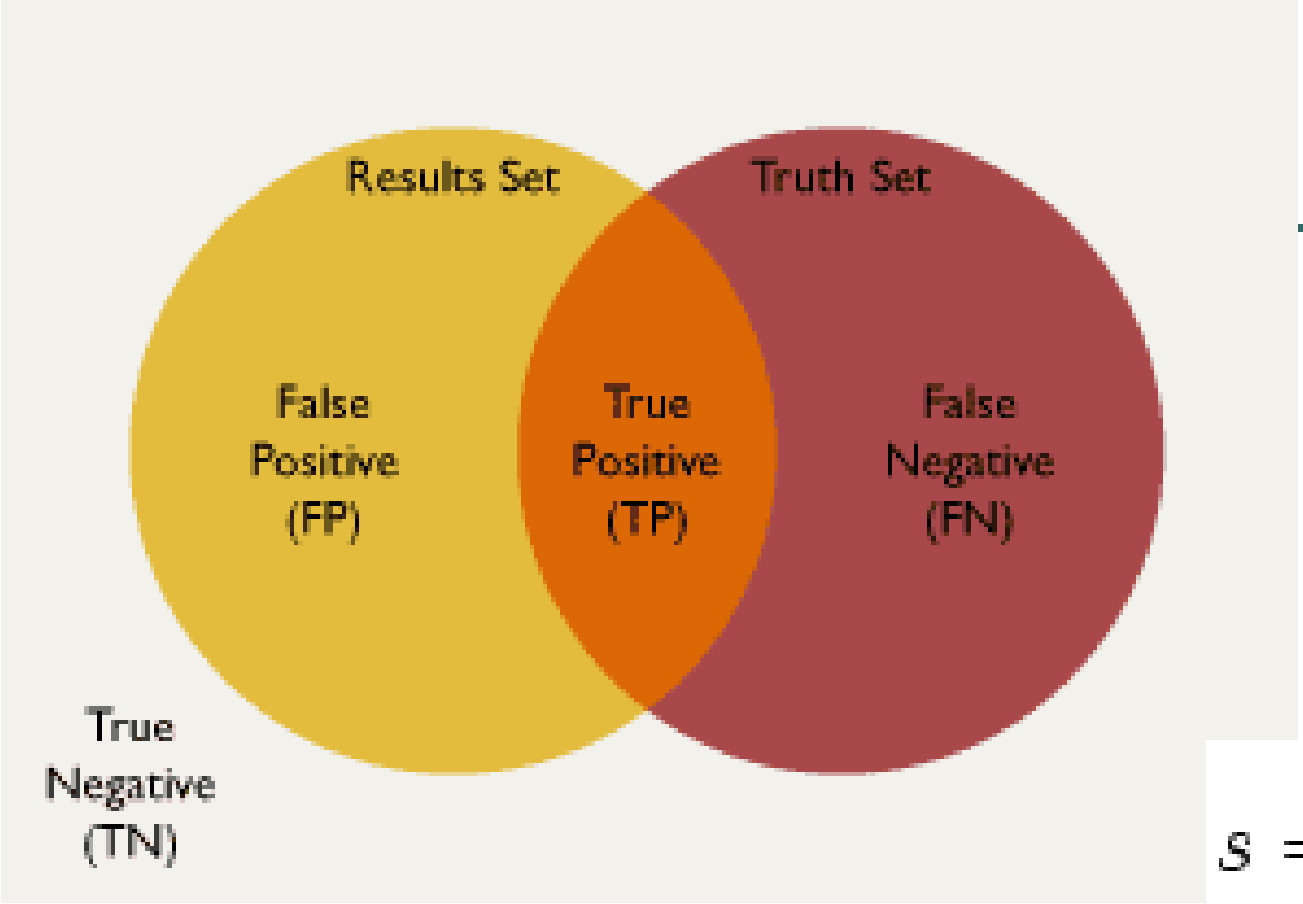
$$2 | X \cap Y | \quad / \quad | X | + | Y |$$

วิธีนี้มีการนำเอาขนาดของ X และ Y มาพิจารณา

สมมุติ $X = \{1, 2, 3\}$ $Y = \{1, 4\}$

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

ดังนั้นค่าสัมประสิทธิ์มีค่าเท่ากับ $2*1 / 3+2 = 2/5$



$$s = \frac{2|X \cap Y|}{|X| + |Y|}$$

Jaccard's Coefficient

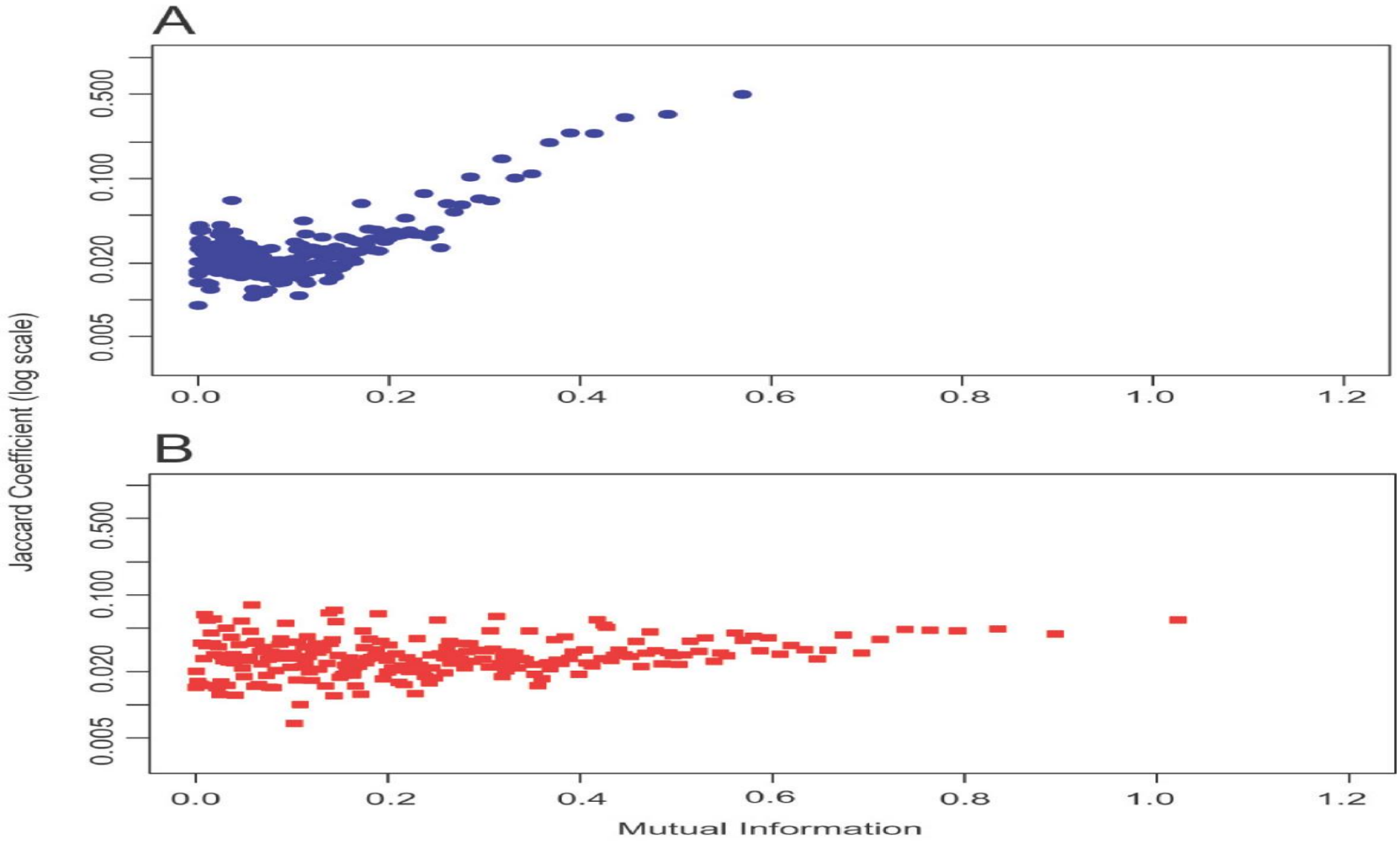
$$|X \cap Y| \;/\; |X \cup Y|$$

สมมุติ $X = \{1, 2, 3\}$ $Y = \{1, 4\}$

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

$X \cup Y = \{1,2,3,4\}$ $|X \cup Y| = 4$

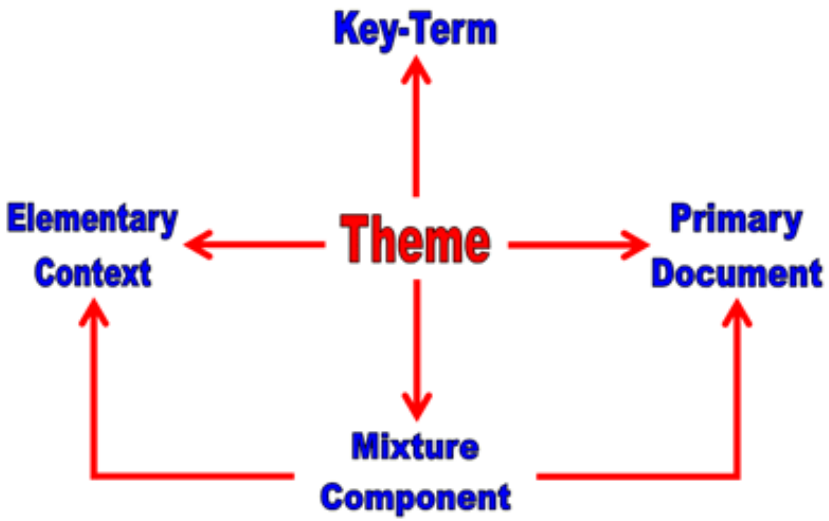
ดังนั้นค่าสัมประสิทธิ์มีค่าเท่ากับ $1 / 4$

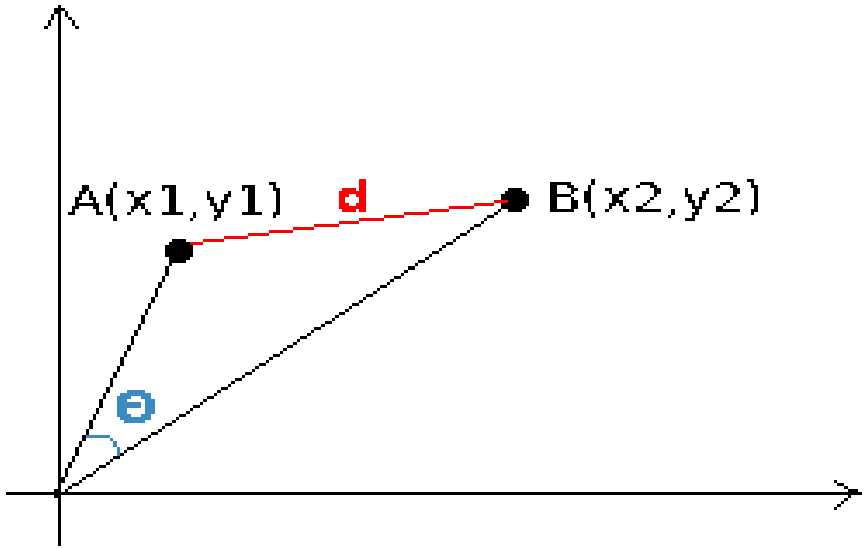

A
B
Jaccard Coefficient (log scale)
0.005
0.020
0.100
0.500
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Mutual Information

Cosine Coefficient

- Salton ได้กำหนดตัวแทนเอกสารเป็นเวกเตอร์ของเลขฐานสองในลักษณะของ n มิติ ซึ่ง n เป็นจำนวนผลรวมของคำดรรชนี ดังนั้น $(X \cap Y) / ||X|| ||Y||$ สามารถอธิบายได้ เป็นมุม cosine ที่แยกเป็น 2 ไบนารีเวกเตอร์ X และ Y ดังนั้น สามารถเขียนตามเวกเตอร์จริงได้ดังนี้

$$| X \cap Y | / |X|^{1/2} * |Y|^{1/2}$$


Key-Term
Elementary
Context
Theme
Primary
Document
Mixture
Component


A(x1,y1)
d
B(x2,y2)
θ

Overlap coefficient

$$| X \cap Y | \; / \; \min(|X| , |Y|)$$

สมมุติ $X = \{ 1 , 2 , 3 \}$ $Y = \{ 1 , 4 \}$

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

ดังนั้นค่าสัมประสิทธิ์มีค่าเท่ากับ $1 / 2$

TABLE 6

Association Norms of German and French Students: Low Overlap Coefficient for "tranquil"

Rank	German Students		French Students	
1	forest	41	country	13
2	sleep	35	forest	11
3	church	20	nature	10
4	night	19	house	9
5	water	17	sea	9
6	read	14	room	8
7	cemetary	13	library	5
8	nature	9	landscape	5
9	bed	8	lonely person	5
10	lake	8	pensioner	5

Percentage of visual associations for the first 10 ranks.
Overlap coefficient = 0,155 (added relative frequencies of common associations, here calculated from all association percentages > 5 %); number of French students 210, number of German students 95.

Source: Institute for Consumer and Behavioral Research, 1992, Saarbrücken.

Similarity Measures

Simple matching (coordination level match)	$\lvert Q \cap D \rvert$
Dice's Coefficient	$2\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert + \lvert D \rvert}$
Jaccard's Coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \cup D \rvert}$
Cosine Coefficient	$\frac{\lvert Q \cap D \rvert}{\lvert Q \rvert^{1/2} \times \lvert D \rvert^{1/2}}$
Overlap Coefficient	$\frac{\lvert Q \cap D \rvert}{\min(\lvert Q \rvert, \lvert D \rvert)}$

การวัดความคล้ายคลึงกันโดยใช้ **Vector Space**

กำหนดให้ D เป็นเอกสาร , Q เป็นข้อสอบถาม ,

W เป็นน้ำหนักของแต่ละเทอม

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

$$Q = w_{q1}, w_{q2}, \ldots, w_{qt}$$

$$w = 0 \text{ if a term is absent}$$

การวัดความคล้ายคลึงกันโดยใช้ **Vector Space**

การวัดความคล้ายคลึงกันกระทำโดยนำน้ำหนักเทอมไป normalize โดยสูตร

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

คำนวณน้ำหนักของแต่ละเอกสาร

$$w_{ik} = tf_{ik} * \log(N / n_k)$$

$T_k = \text{term } k \text{ in document } D_i$

$tf_{ik} = \text{frequency of term } T_k \text{ in document } D_i$

$idf_k = \text{inverse document frequency of term } T_k \text{ in } C$

$N = \text{total number of documents in the collection } C$

$n_k = \text{the number of documents in } C \text{ that contain } T_k$

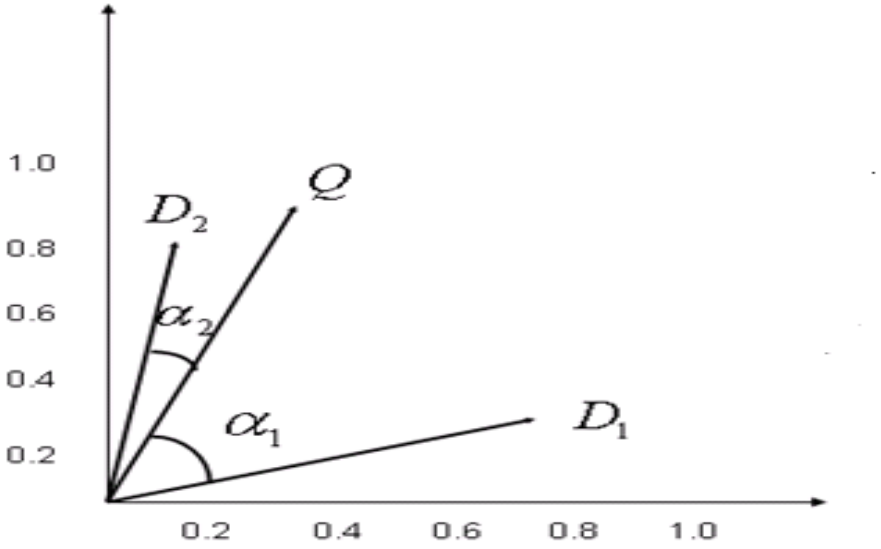
$$idf_k = \log\left(\frac{N}{n_k}\right)$$

ตัวอย่าง

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

เอกสาร D1 มีค่าดังนี้ D1 = (0.8,0.3)

เอกสาร D2 มีค่าดังนี้ D2 =(0.2,0.7)



ตัวอย่าง

- เราสามารถหาความคล้ายคลึงจากเอกสารและข้อสอบถามได้โดยใช้สูตรคือ

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

ตัวอย่าง

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

$Q = (0.4, 0.8)$ D2 =(0.2,0.7)

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

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

ตัวอย่าง

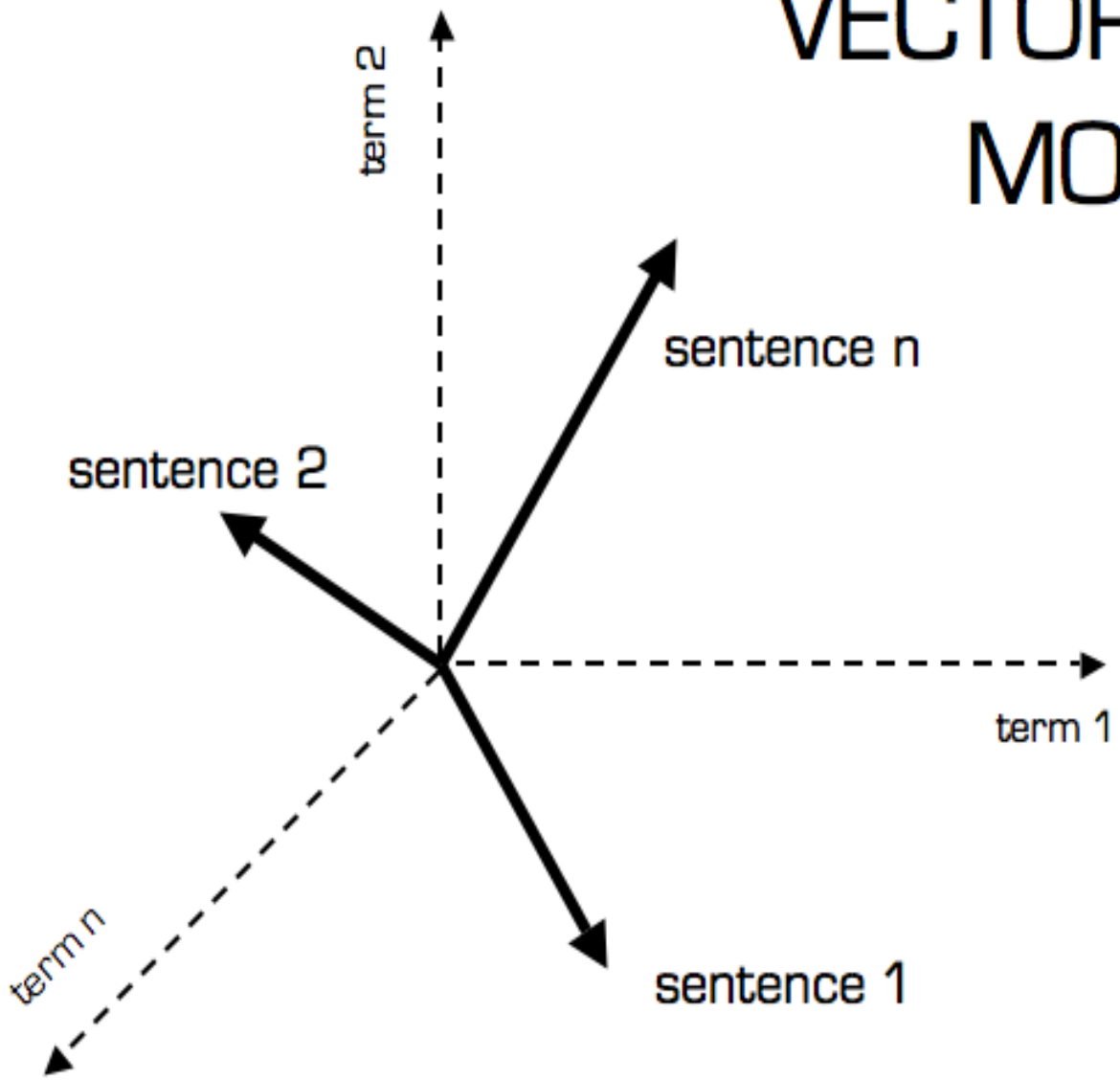
$$sim(Q, D_i) = \frac{\sum_{j=1}^{t} w_{qj} * w_{d_{ij}}}{\sqrt{\sum_{j=1}^{t} (w_{qj})^2 * \sum_{j=1}^{t} (w_{d_{ij}})^2}}$$

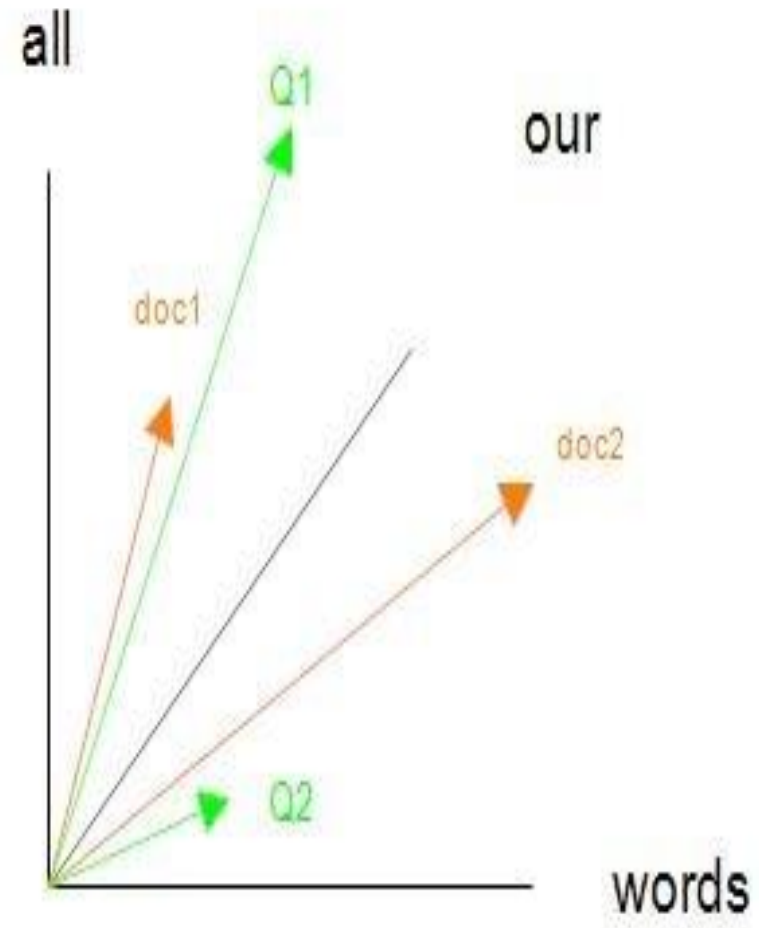
$Q = (0.4,0.8)$ D1 = (0.8,0.3)

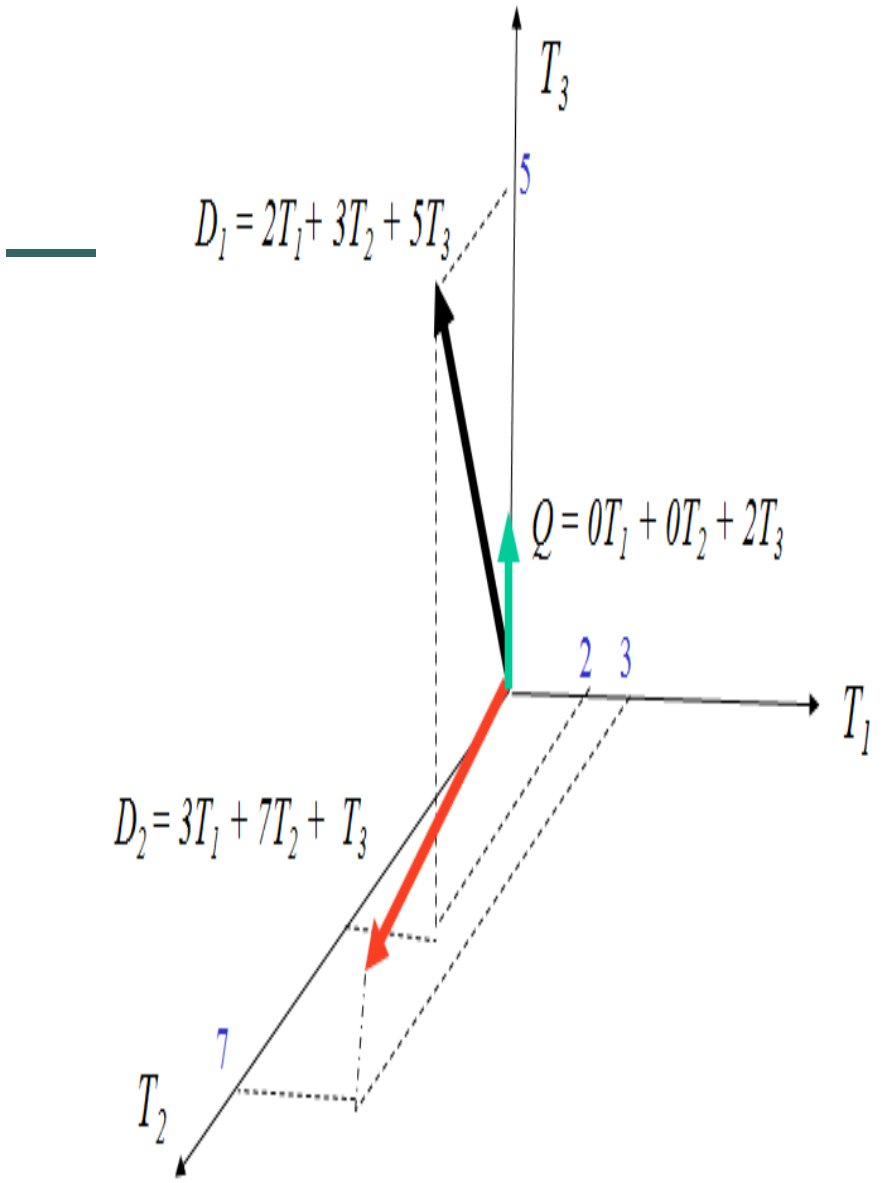
แทนค่าในสูตรระหว่างข้อสอบถาม(Q) และเอกสาร D1 ได้ดังนี้

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

VECTOR SPACE MODEL




all
Q1
our
doc1
doc2
Q2
words


T_3
5
$D_1 = 2T_1 + 3T_2 + 5T_3$
$Q = 0T_1 + 0T_2 + 2T_3$
2
3
T_1
$D_2 = 3T_1 + 7T_2 + T_3$
7
T_2

ความไม่คล้ายคลึงกัน (Dissimilarity)

กำหนด P เป็นเซ็ตของสิ่งของที่ถูกจัดกลุ่ม

D เป็นสัมประสิทธิ์ความไม่คล้ายคลึงกัน เป็นฟังก์ชันหนึ่งจาก

P x P ไปยังเลขจำนวนจริงที่ไม่เป็นค่าลบ

โดยทั่วไป D จะมีคุณสมบัติดังต่อไปนี้

$D(X, Y) > 0$ สำหรับทุกๆ X , Y ของ P

$D(X, X) = 0$ สำหรับทุกๆ X ของ P

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

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

สูตรความไม่คล้ายคลึงกัน

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

ซึ่งสูตรนี้มาจากสูตรความคล้ายคลึงของ Dice's Coefficient

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

สามารถหาความไม่คล้ายคลึงกันได้ดังนี้

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

| X ∆ Y | / |X | + |Y| โดย **| X ∆ Y | = | X ∪ Y | - | X ∩ Y |**

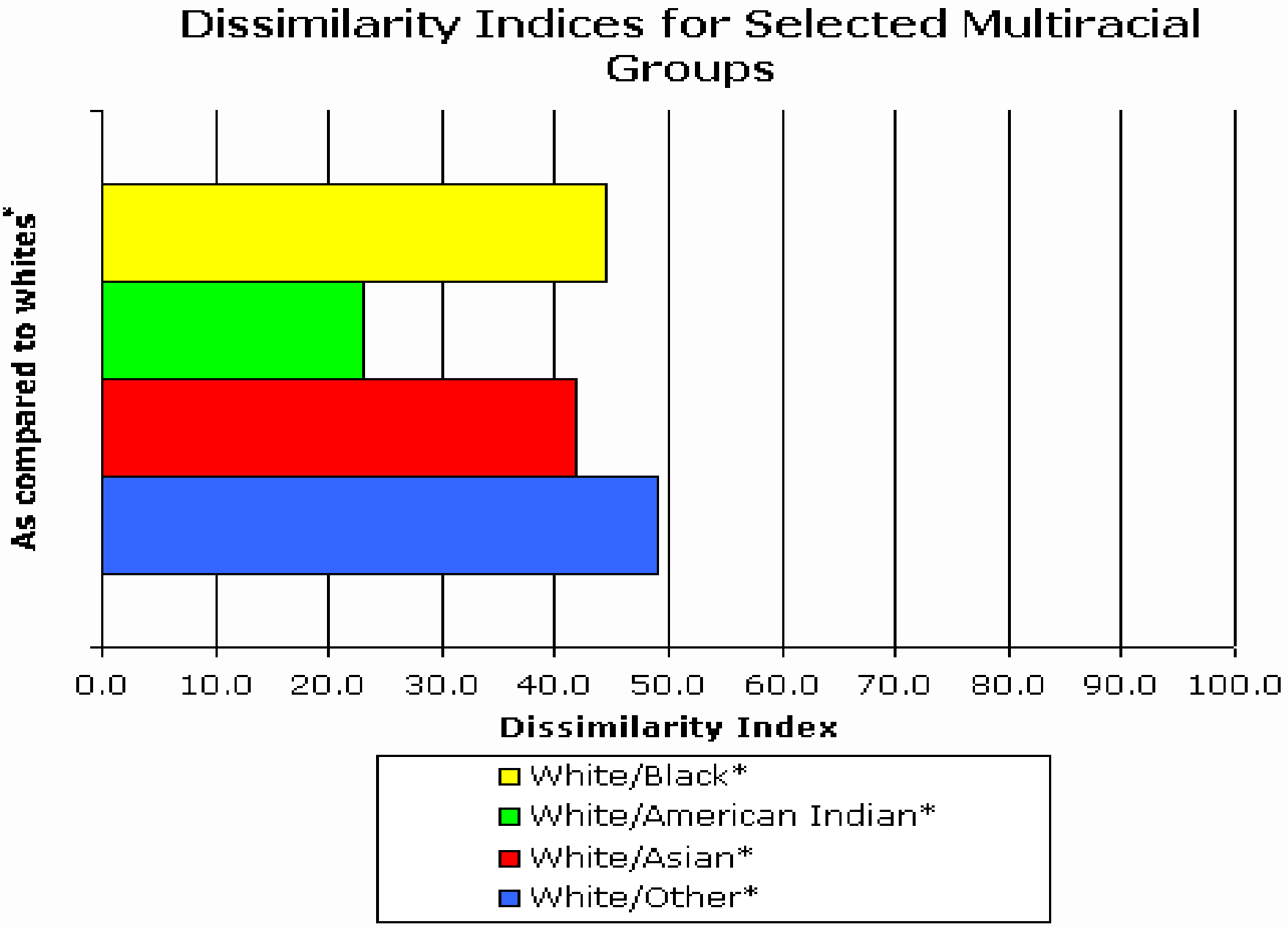
สมมุติ $X = \{1, 2, 3\}$ $Y = \{1, 4\}$

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

$X \cup Y = \{1,2,3,4\}$ $|X \cup Y| = 4$

ดังนั้นค่าสัมประสิทธิ์ความไม่คล้ายคลึงกัน

มีค่าเท่ากับ $(4 - 1) / (3 + 2) = 3/5$


Dissimilarity Indices for Selected Multiracial Groups
As compared to whites*
0.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0
80.0
90.0
100.0
Dissimilarity Index
White/Black*
White/American Indian*
White/Asian*
White/Other*

Jaccard

- ใช้เลขฐานสองแทนค่าดรรชนีของเอกสาร มีผลให้ดรรชนีตัวที่ i จะแทนด้วยเลข 0 หรือ 1 ในตำแหน่งที่ I ตามลำดับ ซึ่งค่า 1 นั้นจะแสดงถึงคำสำคัญอยู่ในเอกสาร สำหรับเลข 0 หมายถึงคำสำคัญไม่ได้อยู่ในเอกสารนั้นในตำแหน่งที่ระบุ

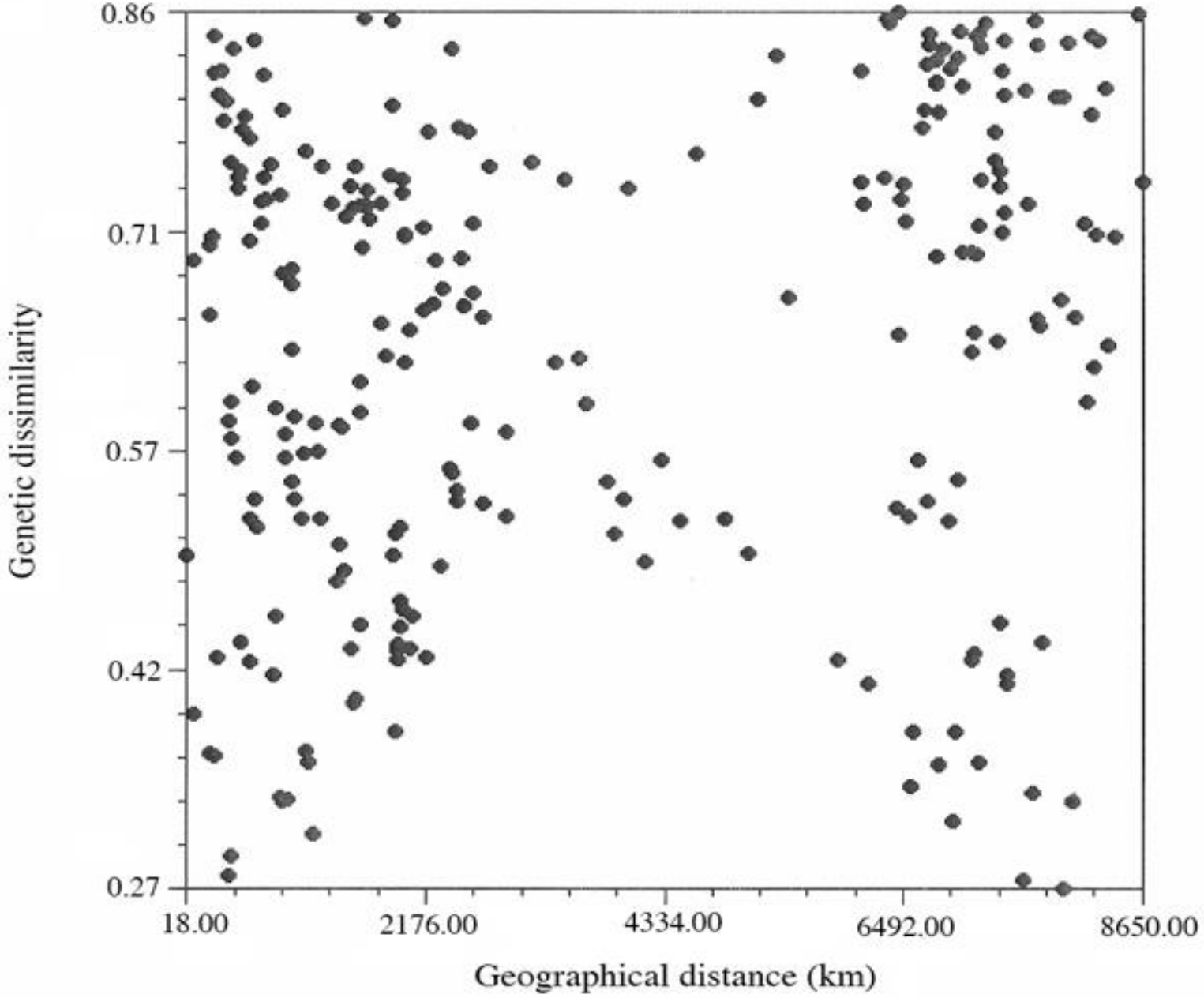
ถ้า $|X| = \Sigma Xi$ โดยที่ $i \in 1..N$

และ N เป็นจำนวนของค่าดรรชนีทั้งหมดในเอกสาร

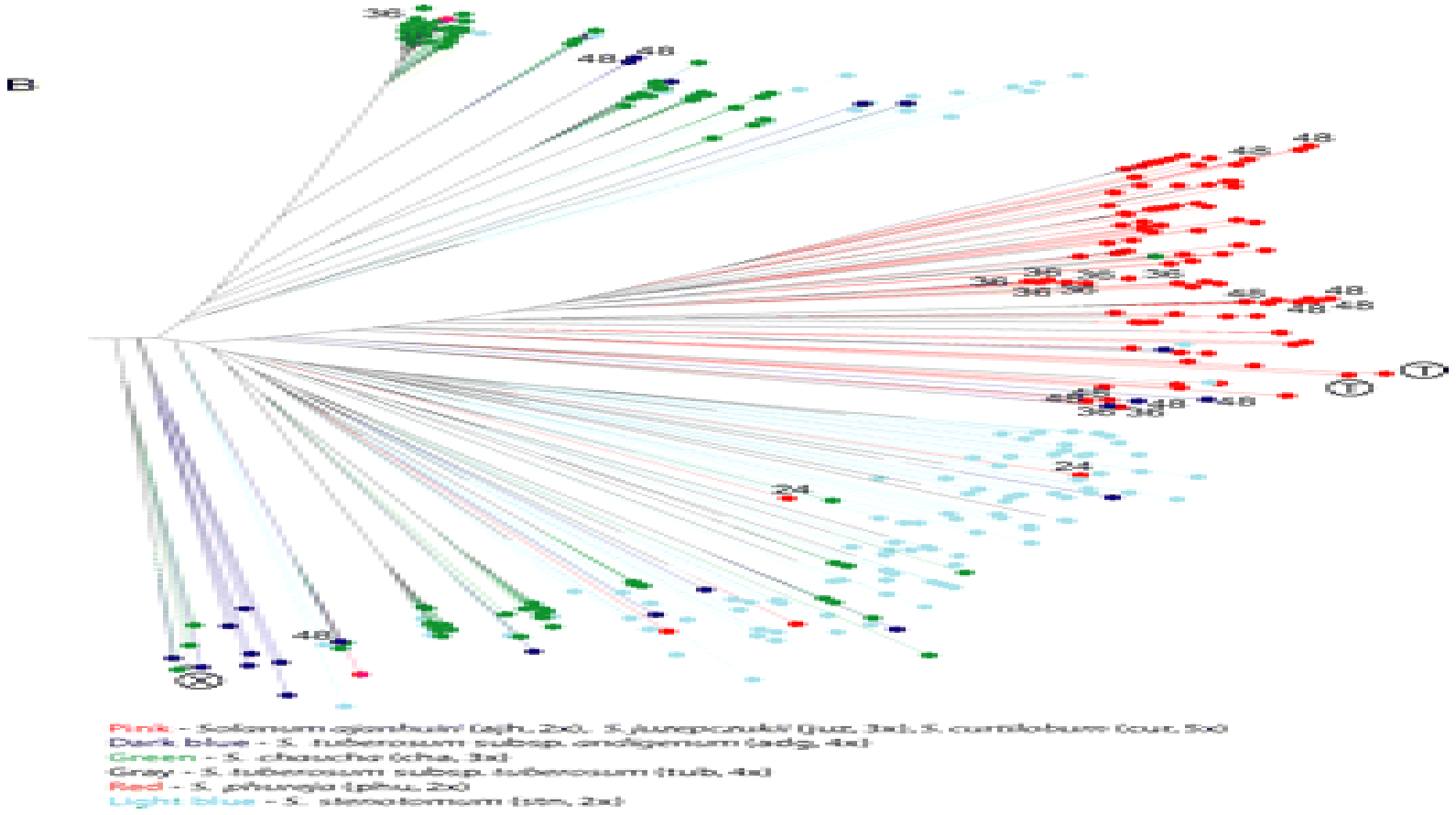
ดังนั้น $|X \cap Y| = \Sigma Xi\ Yi$ โดยที่ $I \in 1..N$

ความไม่คล้ายคลึงกันมีค่าเท่ากับ

$$|X \Delta Y| / |X| + |Y| = (\Sigma Xi(1-Yi) + \Sigma Yi(1-Xi)) / (\Sigma Xi + \Sigma Yi)$$


0.86
0.71
0.57
0.42
0.27
Genetic dissimilarity
18.00
2176.00
4334.00
6492.00
8650.00
Geographical distance (km)

Jaccard's tree based on a dissimilarity matrix



การปรับปรุงประสิทธิภาพระบบค้นคืนสารสนเทศ

ใช้ความคาดหวังเพื่อใช้ในการวัดความเกี่ยวข้องได้

สามารถกระจายความน่าจะเป็น P(Xi) และ P(Xj)

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

Jardine และ Sibson

- วัดความไม่คล้ายคลึงกันระหว่าง สองออปเจ็กต์ โดยแบ่งการการจายความน่าจะเป็นเป็นสองจุดคือ 1 และ 0 ดังนั้น P1(1),p1(0),P2(1),P2(0) เป็นการกระจายความน่าจะเป็นของออปเจ็กต์ทั้งสอง ซึ่งการวัดความไม่คล้ายคลึงกันของ Jardine และ Sibson นี้เรียกว่ารัศมีสารสนเทศ (Information Radius)

มีการคำนวณดังนี้

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

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

วิธีจัดแบ่งกลุ่ม (Classification Methods)

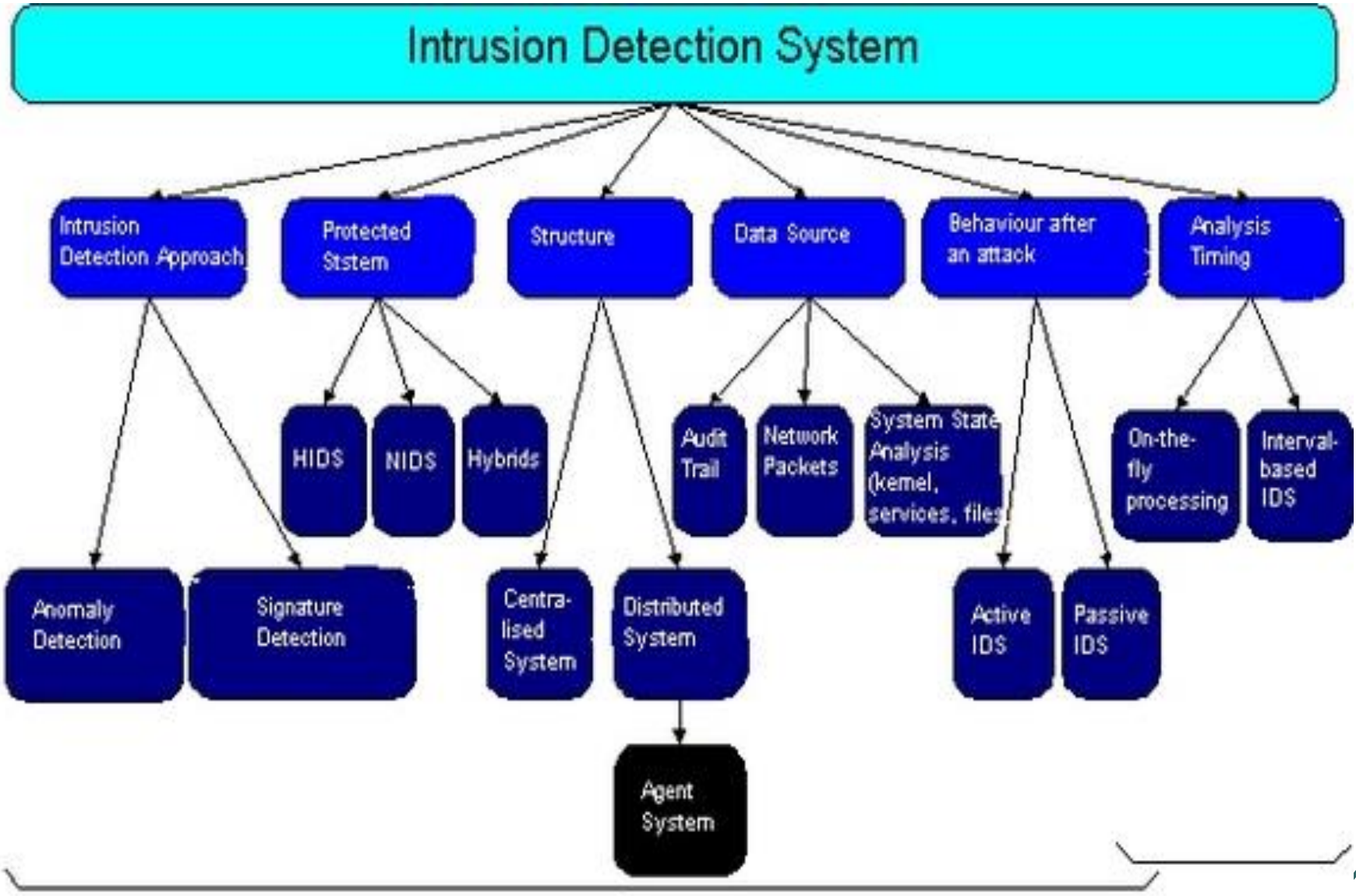
ออปเจ็กต์

อาจเป็นเอกสาร(documents)
คำดรรชนี(keyword)
ตัวอักษรที่เขียนด้วยมือคน
(hand written character)หรือ
ชนิดของพืช(species)

ส่วนคำอธิบาย(Description)

เป็นโครงสร้างของข้อมูลภายใต้ชื่อต่างๆ
อาจเป็นคุณสมบัติในหลายๆสถานะ

- สี
- สถานะของเลขฐานสอง เช่น คำดรรชนี (Keyword)
- จำนวน เช่น มาตราวัดความแข็ง หรือ น้ำหนักของดรรชนี เป็นต้น
- การกระจายความน่าจะเป็น (Probability Distributions)

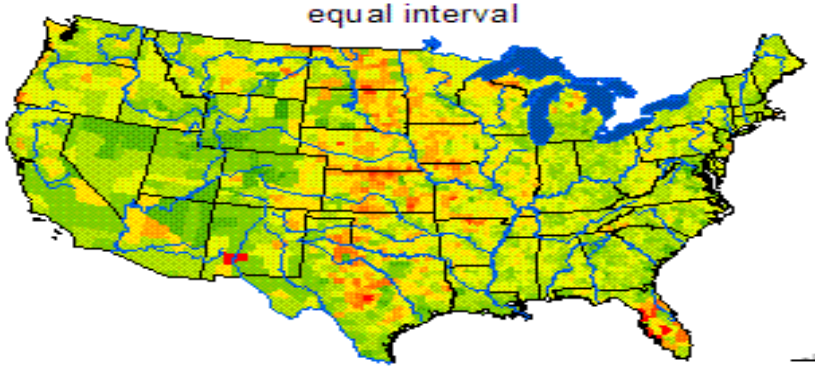

Intrusion Detection System
Intrusion Detection Approach
Protected Ststem
Structure
Data Source
Behaviour after an attack
Analysis Timing
HIDS
NIDS
Hybrids
Audit Trail
Network Packets
System State Analysis (kernel, services, files
On-the-fly processing
Interval-based IDS
Anomaly Detection
Signature Detection
Centra-lised System
Distributed System
Active IDS
Passive IDS
Agent System

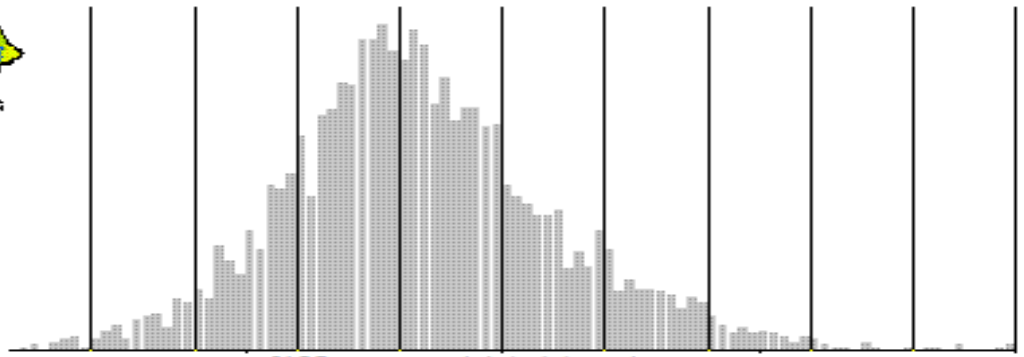
COMPARISON OF CLASSIFICATION METHODS

Percentage of US Population Age 65 or Older, by County

Distributions of Counties

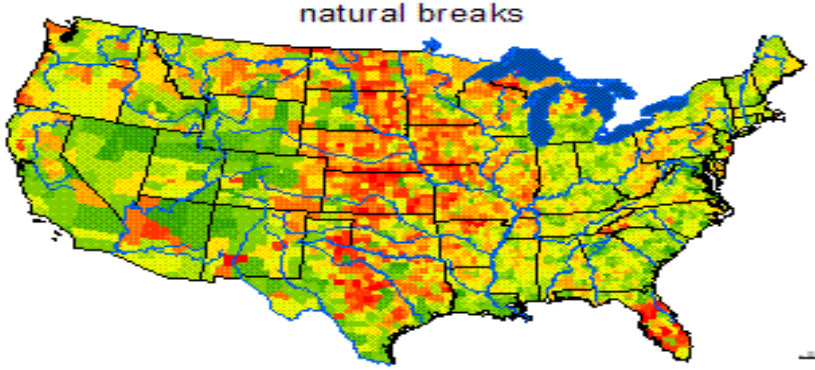
equal interval

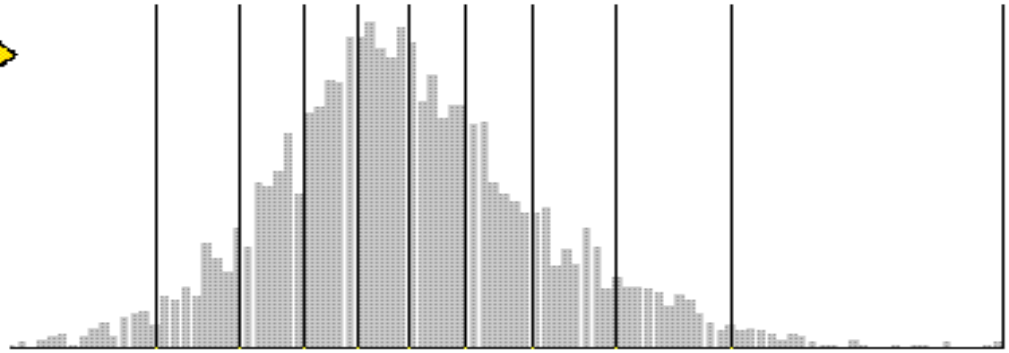




%65up -- equal data intervals

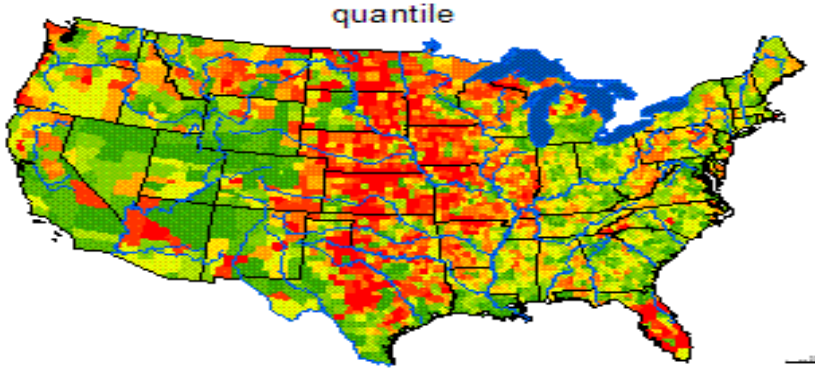
natural breaks

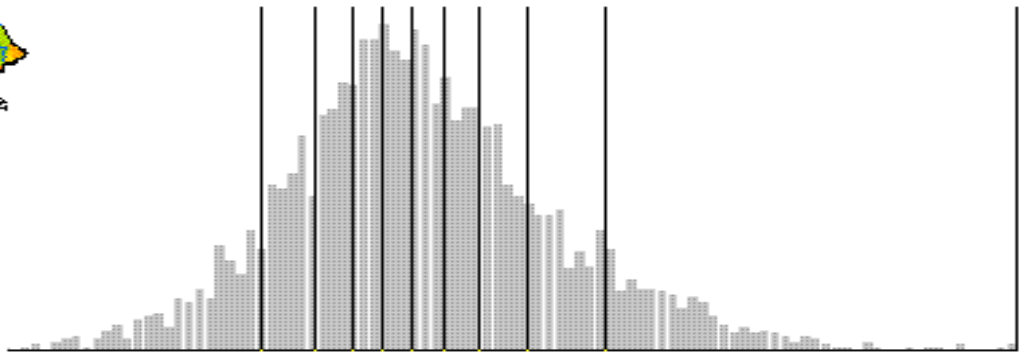




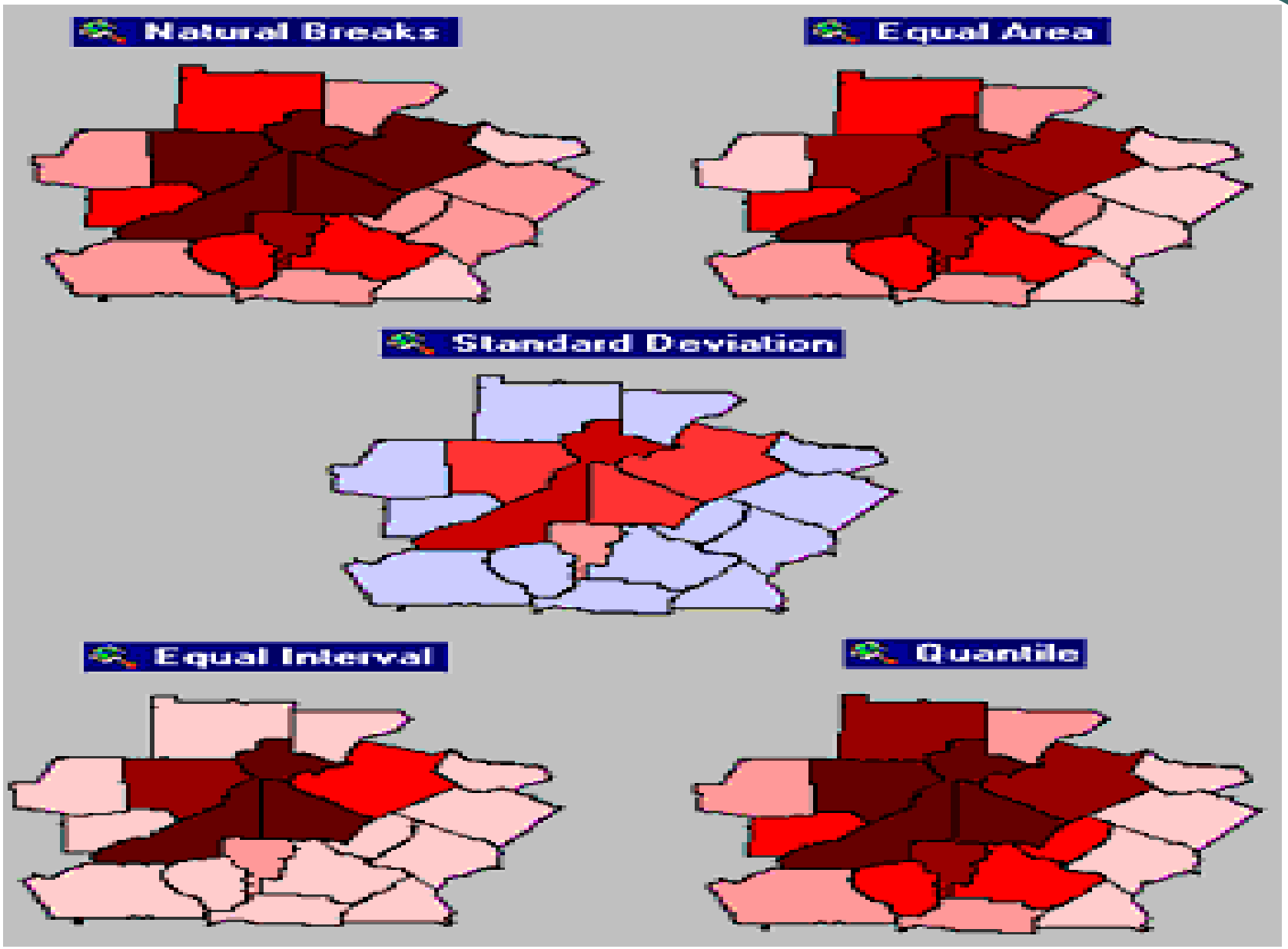
%65up -- Jenks Natural Breaks classification

quantile





%65up -- equal number of counties in each category



are grouped into five classes using different **classification methods**.

Sparck Jones

ได้แบ่งวิธีการจัดกลุ่มตามความสัมพันธ์ได้หลายลักษณะดังนี้

ความสัมพันธ์ระหว่างคุณสมบัติและคลาส

- Monothetic

กลุ่มของ Individual ที่มีคุณสมบัติเหมือนกัน

- Polythetic

คุณสมบัติในสี่ประการจะมีคุณสมบัติที่เหมือนกัน 3 ข้อที่เหมือนกัน เป็นต้น กล่าวคือไม่เหมือนกันหมด

ความสัมพันธ์ระหว่างคุณสมบัติและคลาส

รูป แสดงความแตกต่างระหว่าง **Monothetic** และ **Polythetic**

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

Polythetic

Monothetic

ความสัมพันธ์ระหว่างออปเจ็กต์และคลาส

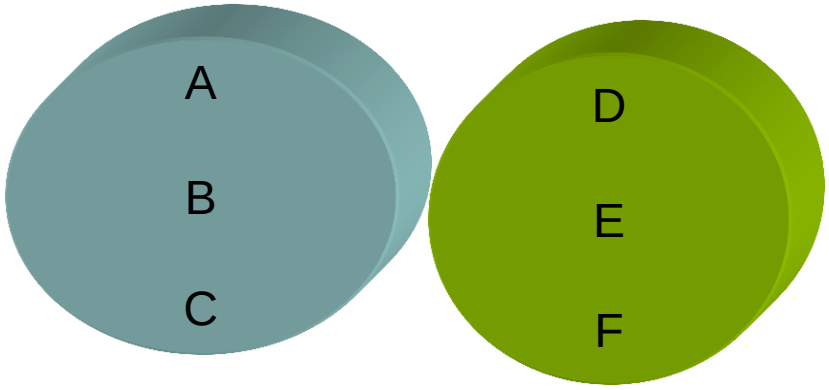
Exclusive Class

กลุ่มของ individual ที่เป็นสมาชิกเฉพาะไม่มี Overlapping Individuals

Overlapping Class

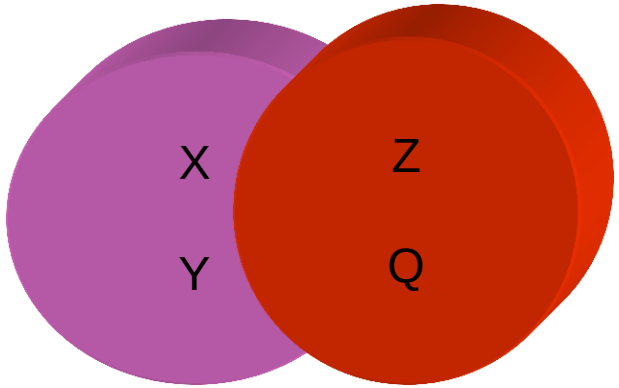
กลุ่มของ Individuals ที่อาจมีสมาชิกหนึ่งของกลุ่มนี้ ไปเป็นสมาชิกของกลุ่มอื่น

ความสัมพันธ์ระหว่างออปเจ็กต์และคลาส



Exclusive Class

Overlapping Class



ความสัมพันธ์ระหว่างคลาสกับคลาส

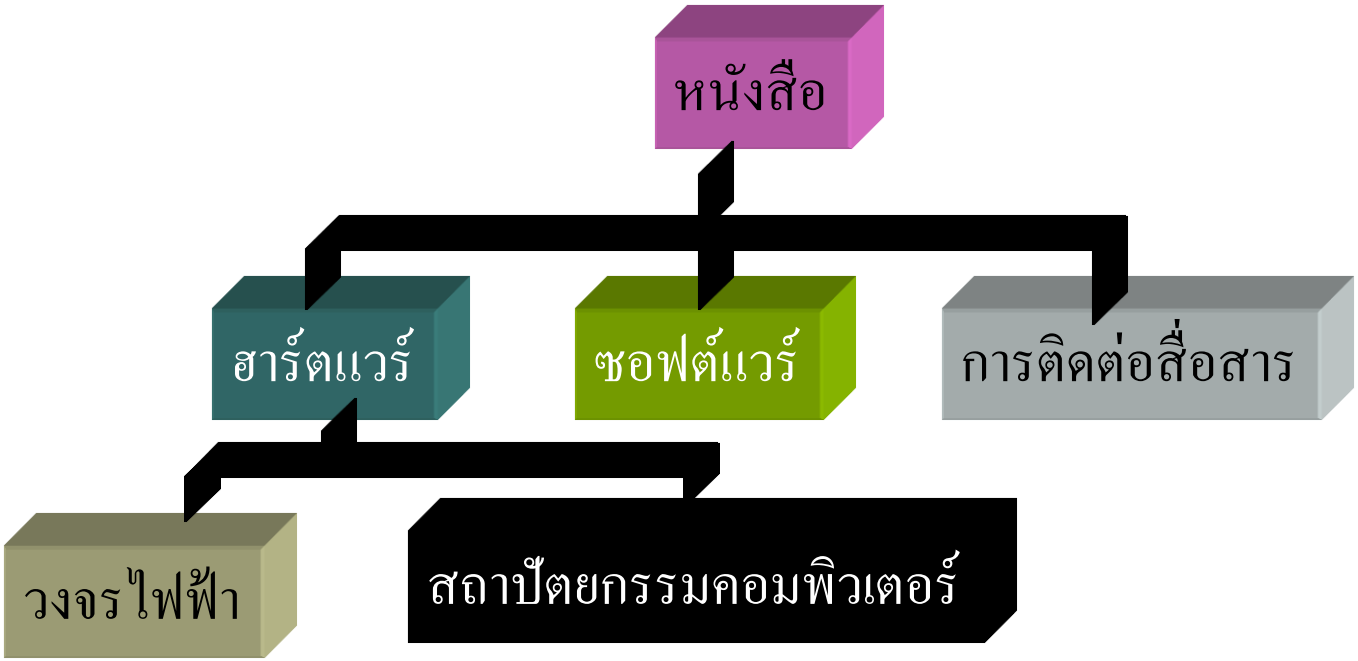
Ordered Classification

เป็นการจัดแบ่งกลุ่มแบบมีการจัดลำดับของคลาส เช่น การจัดแบ่งกลุ่มเป็นลำดับชั้น (Hierarchical) ซึ่งคลาสที่อยู่ในระดับล่างจะอยู่ภายใต้ขอบข่ายของคลาสที่อยู่ระดับบน

Unordered Classification

เป็นการแบ่งกลุ่มแบบไม่มีลำดับ เช่น การสร้าง Thesaurus แบบอัตโนมัติ ซึ่งเป็นพจนานุกรมข้อมูลที่ทำให้วิธีการค้นหาข้อมูลมีประสิทธิภาพ

ความสัมพันธ์ระหว่างคลาสกับคลาส



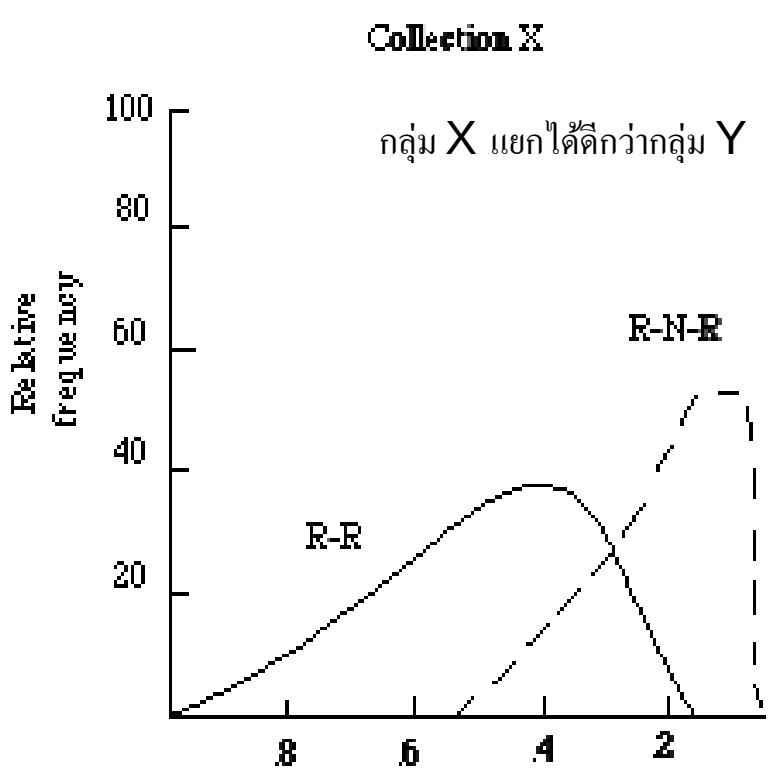
Ordered Classification

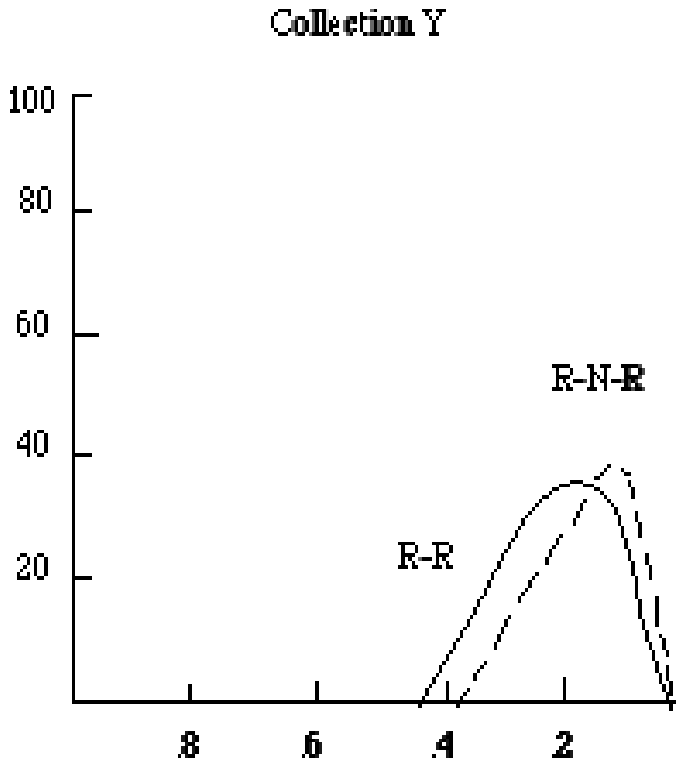
ข้อสมมุติฐานของคลัสเตอร์ (Cluster Hypothesis)

- “ เอกสารที่เกี่ยวข้องกันอย่างใกล้ชิดมีแนวโน้มที่มีการเกี่ยวพันธ์กับข้อคำถามที่ร้องขอเหมือนกัน “
- **“closely associated documents tend to be relevant to the same requests “**

รูป: แสดงถึงผลรวมของกลุ่มข้อเรียกร้อง

relevant-relevant(R-R) และ relevant-non-relevant(R-N-R)





เอกสารที่เกี่ยวข้อง(relevant) จะคล้ายคลึงกันและกัน มากกว่าจะคล้ายกับเอกสารที่ไม่เกี่ยวข้อง(Non-relevant)

ทำไมต้องจัดกลุ่มเอกสาร **(document clustering)**

- ช่วยให้การดึงข้อมูลได้ประสิทธิภาพมากกว่าการค้นหาตามลำดับ เพราะวิธีการค้นหาตามลำดับ จะไม่สนใจความสัมพันธ์ระหว่างเอกสาร แต่คลัสเตอร์สามารถจัดโครงสร้างของกลุ่มโดยให้เอกสารที่คล้ายคลึงกันมากอยู่ในคลาสเดียวกัน ทำให้สามารถเพิ่มความเร็วในการดึงข้อมูล ทำให้มีประสิทธิผลมากขึ้น

clustering

- กระบวนการที่รวมออปเจ็กต์ที่มีลักษณะคล้ายกัน หรือเหมือนกันไว้ในกลุ่มเดียวกัน หรือกำหนดกฎเกณฑ์ขึ้นมาภายในคลัสเตอร์

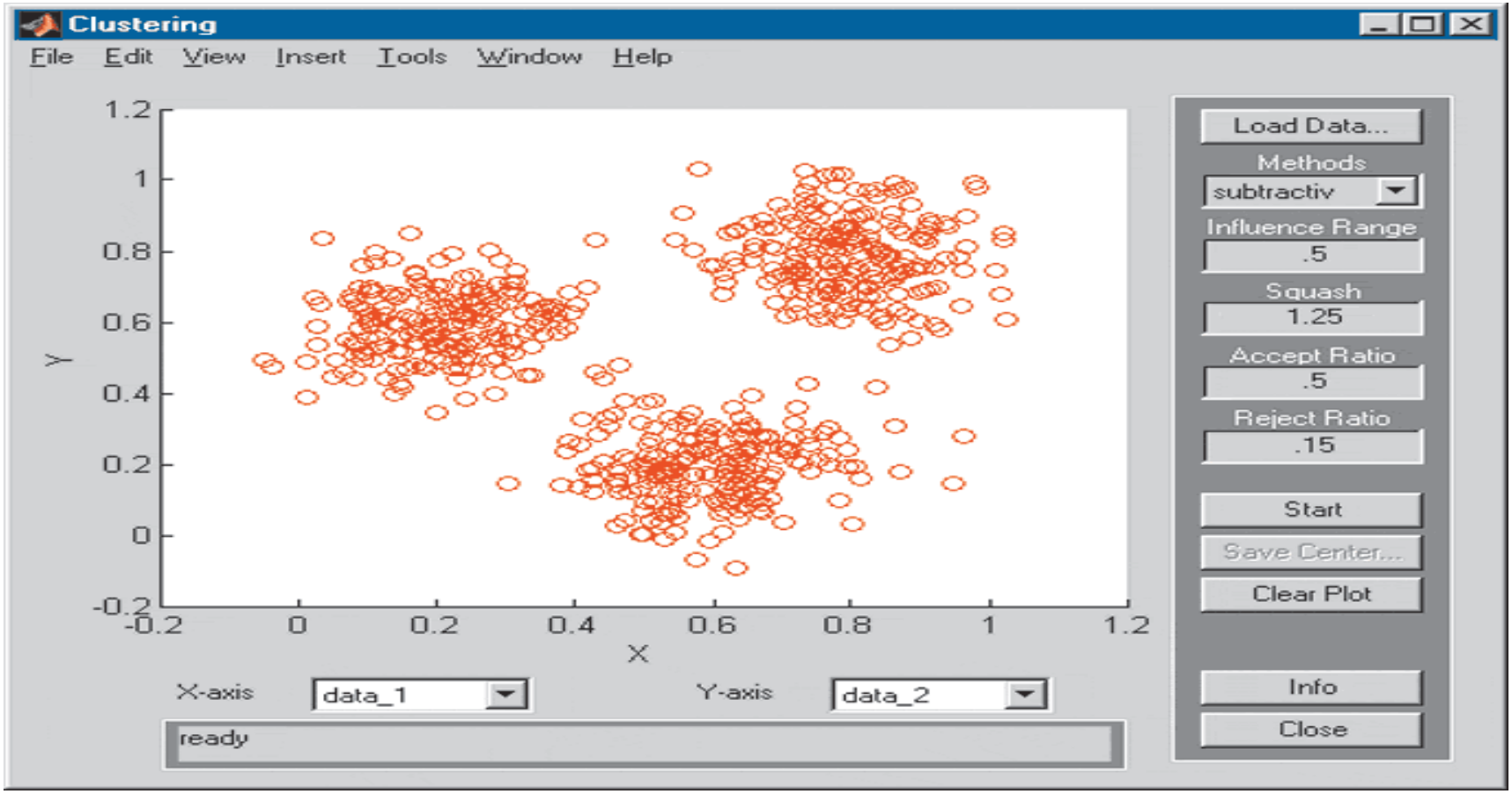
ซึ่งออปเจ็กต์ที่อยู่ภายในกลุ่มจะมีลักษณะตรงตามที่กำหนดเอาไว้นั่นเอง

กระบวนการคลัสเตอร์ริ่ง(Clustering algorithm) สามารถนำไปใช้ได้หลายทาง

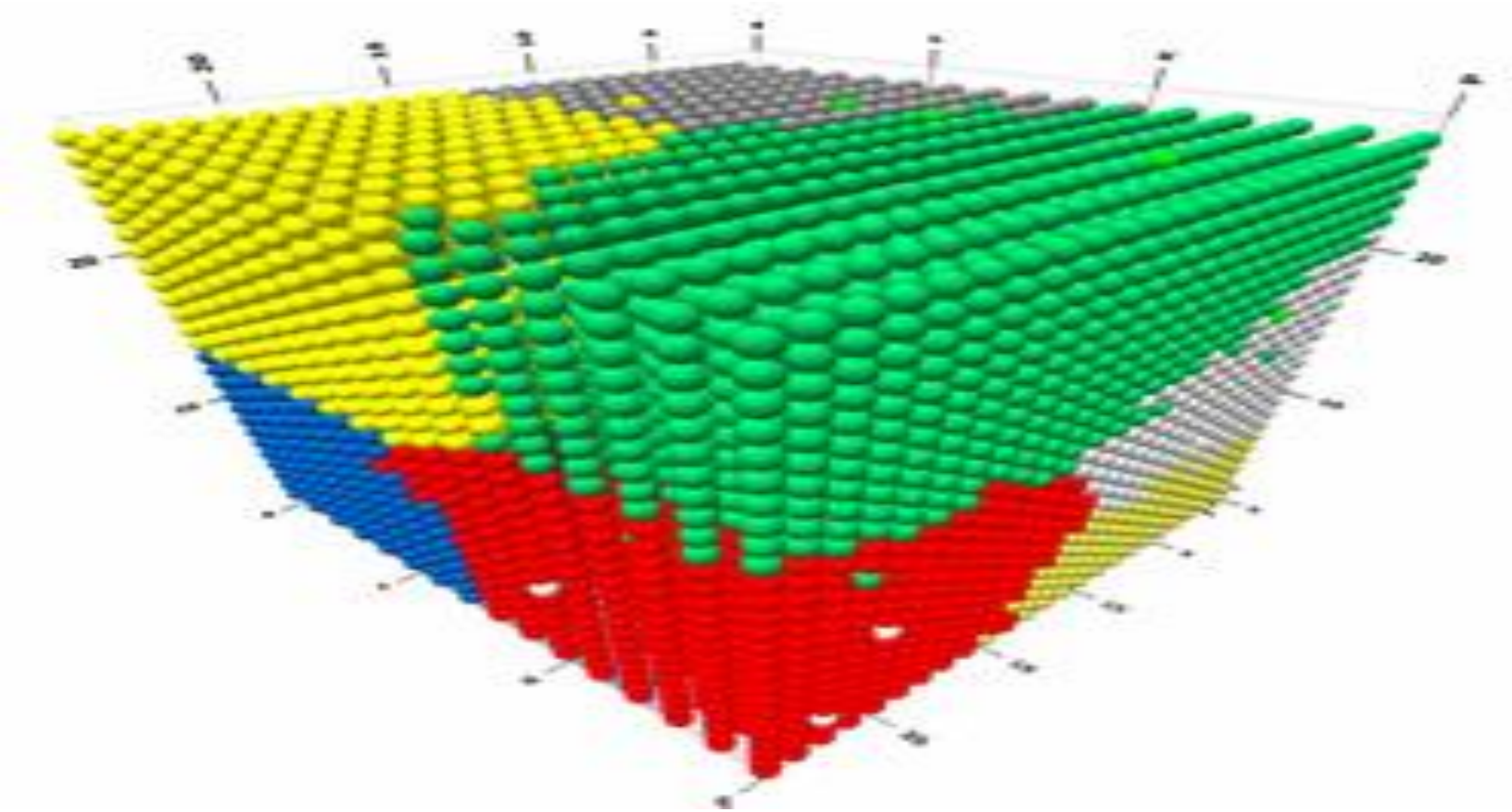
- การจัดกลุ่มลูกค้าที่มีลักษณะของการซื้อขายที่เหมือนกันไว้ภายในฐานข้อมูลเดียวกัน
- ทางชีววิทยาสามารถนำลักษณะเด่นของพืชและสัตว์มาจัดกลุ่มไว้ในกลุ่มเดียวกัน
- ทางด้านบรรณารักษ์ จะนำไปใช้ในการจัดหนังสือและแยกประเภทของหนังสือ
- การประกันภัยสามารถนำไปใช้ในการแบ่งแยกกลุ่มของกรมธรรม์ประกันภัยของลูกค้า
- การจัดทำผังเมือง ใช้แบ่งที่ตั้งที่อาศัยตามสภาพภูมิศาสตร์
- การเฝ้าสังเกตการณ์แผ่นดินไหว ใช้ในการเฝ้าสังเกตว่าส่วนใดเป็นพื้นที่อันตรายที่เสี่ยงต่อการเกิดแผ่นดินไหว

Customer Clustering

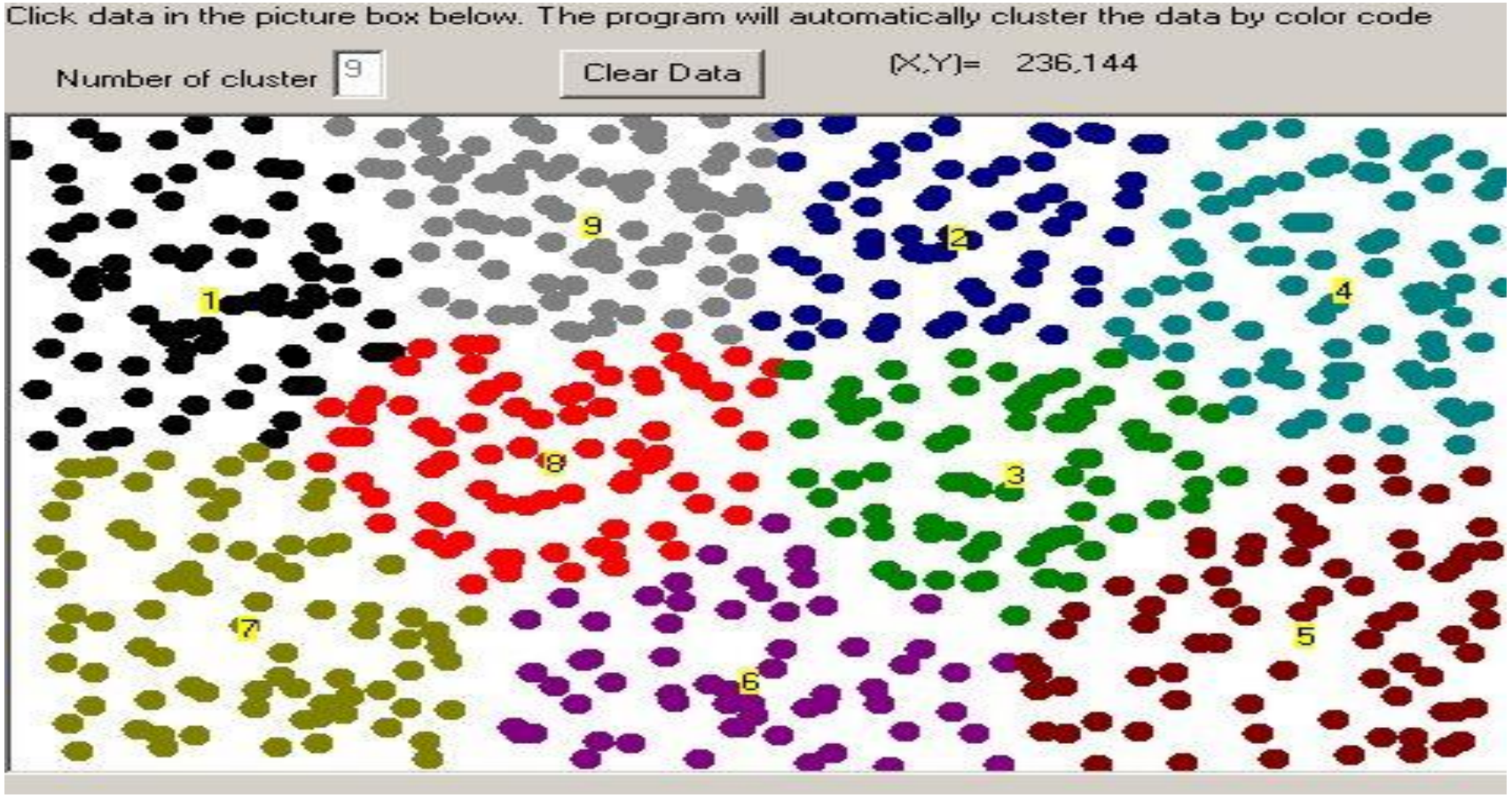
The Fuzzy Clustering GUI uses numerical data to develop classification



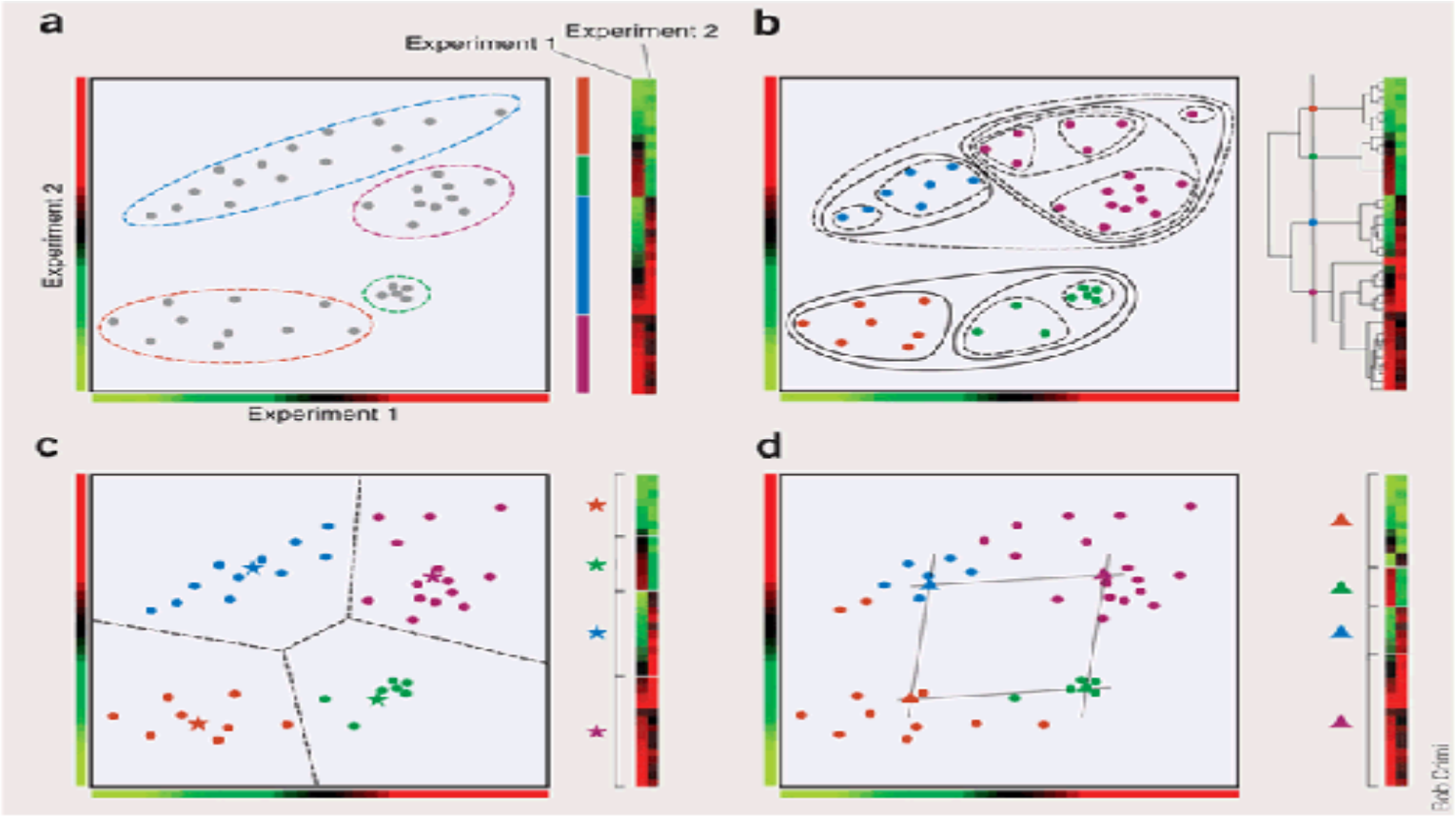
Visual Clustering



K-Mean Clustering



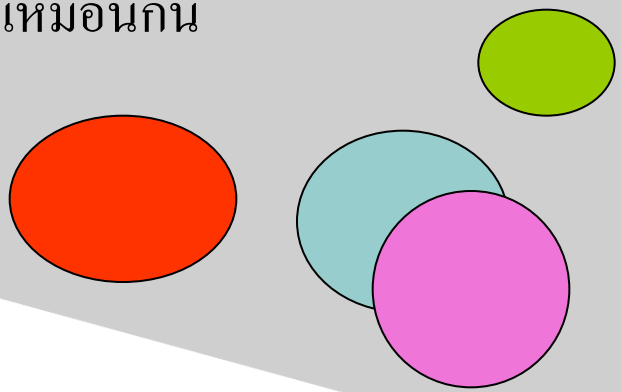
applying cluster analysis to your own gene expression data sets.



การใช้วิธีคลัสเตอร์ในระบบค้นคืนสารสนเทศ

Exclusive Clustering

เป็นการจัดกลุ่มที่ข้อมูลที่อยู่ในกลุ่มเดียวกันจะมีลักษณะเหมือนกัน



Overlapping Clustering

เป็นการจัดกลุ่มที่ข้อมูลภายในกลุ่มแบ่งออกให้อยู่ในรูปของเซตย่อยๆ แต่ละจุดอาจเป็นส่วนหนึ่งของคลัสเตอร์มากกว่าหนึ่งก็ได้และจำนวนสมาชิกภายในเซตต่างๆมีค่าแตกต่างกันได้

การใช้วิธีคลัสเตอร์ในระบบค้นคืนสารสนเทศ

Hierarchical Clustering

เป็นการจัดกลุ่มที่มีการรวมคุณสมบัติพื้นฐานของ

Exclusive Clustering และ Overlapping Clustering ไว้ด้วยกันแต่มีการกำหนดเงื่อนไขหรือกฎที่ใช้ในการกำหนดข้อมูลในแต่ละคลัสเตอร์

Probabilistic Clustering

เป็นการจัดกลุ่มโดยใช้วิธีการทางสถิติ

การเลือกวิธีที่เหมาะสมในการคลัสเตอร์นั้นมีหลักเกณฑ์ 2 ประการ

ความมั่นคงของวิธีการ

(1) เมื่อมี ออปเจ็ก อื่นๆเพิ่มเข้ามา คลัสเตอร์จะไม่ค่อยเปลี่ยนแปลง คือเจริญเติบโตแบบคงที่

(2) วิธีการนั้นจะต้องมั่นคง กล่าวคือความผิดพลาดเล็กๆน้อยๆ ก็จะก่อให้เกิดการเปลี่ยนแปลงอย่างเล็กๆน้อยๆในคลัสเตอร์

(3) วิธีการต้องเป็นอิสระไม่ขึ้นอยู่กับลำดับเริ่มแรกของออกเจ็กต์

ประสิทธิภาพ

ด้านความเร็ว และ ความต้องการเนื้อที่เก็บความจำ ซึ่งเป็นความสามารถในการดึงข้อมูลที่ต้องการ และมีข้อมูลที่ไม่ต้องการมาน้อยที่สุด