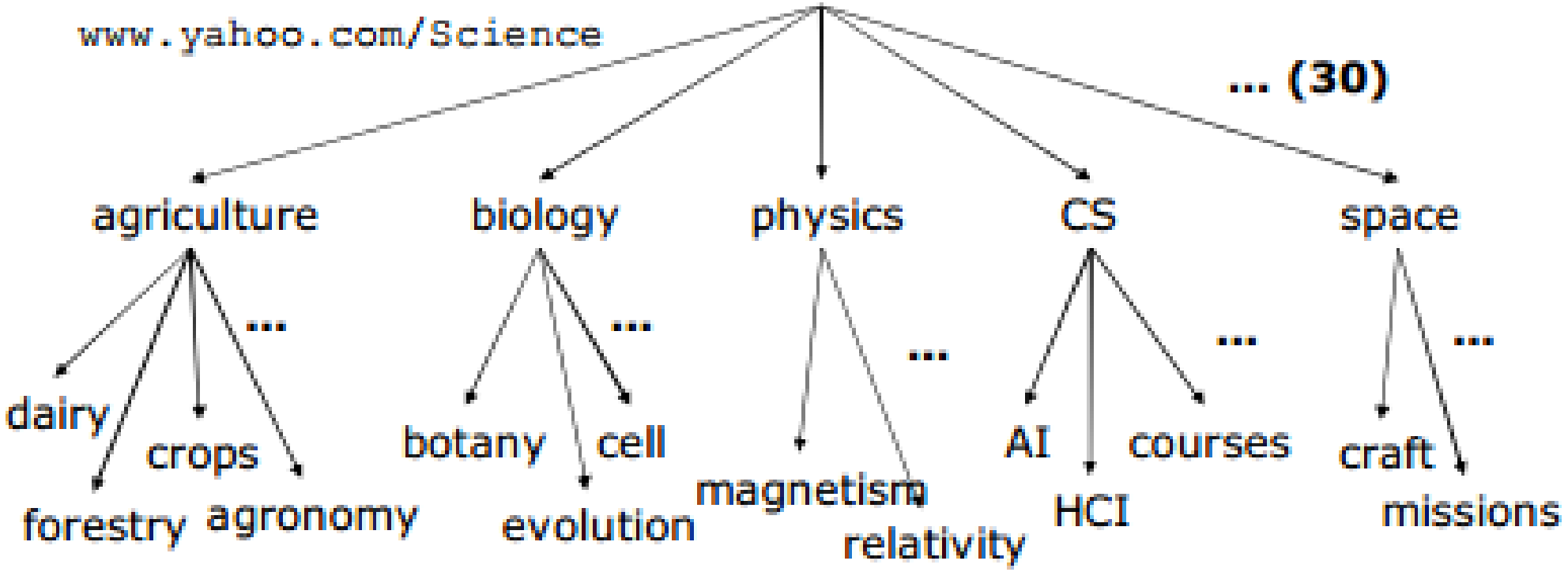
Text Clustering

OLEH : RAHMAT ROBI WALIYANSYAH, M.KOM.

Clustering

- Pengelompokan, penggerombolan
- Proses pengelompokan sekumpulan obyek ke dalam kelas-kelas obyek yang memiliki sifat sama.
- Unsupervised learning

Yahoo! Hierarchy



Menggunakan Cluster

- Hipotesis : dokumen dengan teks yang mirip memiliki keterkaitan.
- Oleh karena itu, untuk meningkatkan recall:
  - Kelompokkan dokumen pada korpus.
  - Jika query ***q*** cocok dengan dokumen ***d***, maka berikan juga dokumen lain yang sekelompok dengan ***d***.
- Contoh: query car juga akan memberikan dokumen tentang automobile (karena satu kelas).

ISU PADA CLUSTERING

- Representasi dokumen:

  - Ruang vektor? Normalisasi?

- Ukuran kesamaan/jarak

- Banyaknya kelas:

  - Tetap

  - Tergantung pada data

  Harus dihindari jumlah kelas yang terlalu sedikit atau terlalu banyak. Mengapa?

  Apabila terlalu banyak, kemungkinan ada frekuensi yang bernilai nol dan apabila terlalu sedikit, konsentrasi pada kelas tertentu

Apa yang membuat dokumen berhubungan?

- Ideal : kesamaan semantik
- Praktis : kesamaan statistik
  - Menggunakan ukuran kesamaan Cosine
  - Dokumen sebagai vektor
  - Untuk beberapa algoritma, lebih mudah memperhatikan jarak antar dokumen, dibanding kesamaannya.

Algoritme Clustering

- **Partitional algorithms**
  - Dimulai dengan sebagian secara acak
  - Dilakukan iterasi:
    - K means clustering
    - Model based clustering
- **Hierarchical algorithms**
  - Bottom-up, agglomerative
  - Top-down, divisive

K-means

- Asumsikan tiap dokumen sebagai vektor bernilai bilangan riil.
- Kelompokkan dokumen berdasarkan centroid pada suatu cluster c :

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$

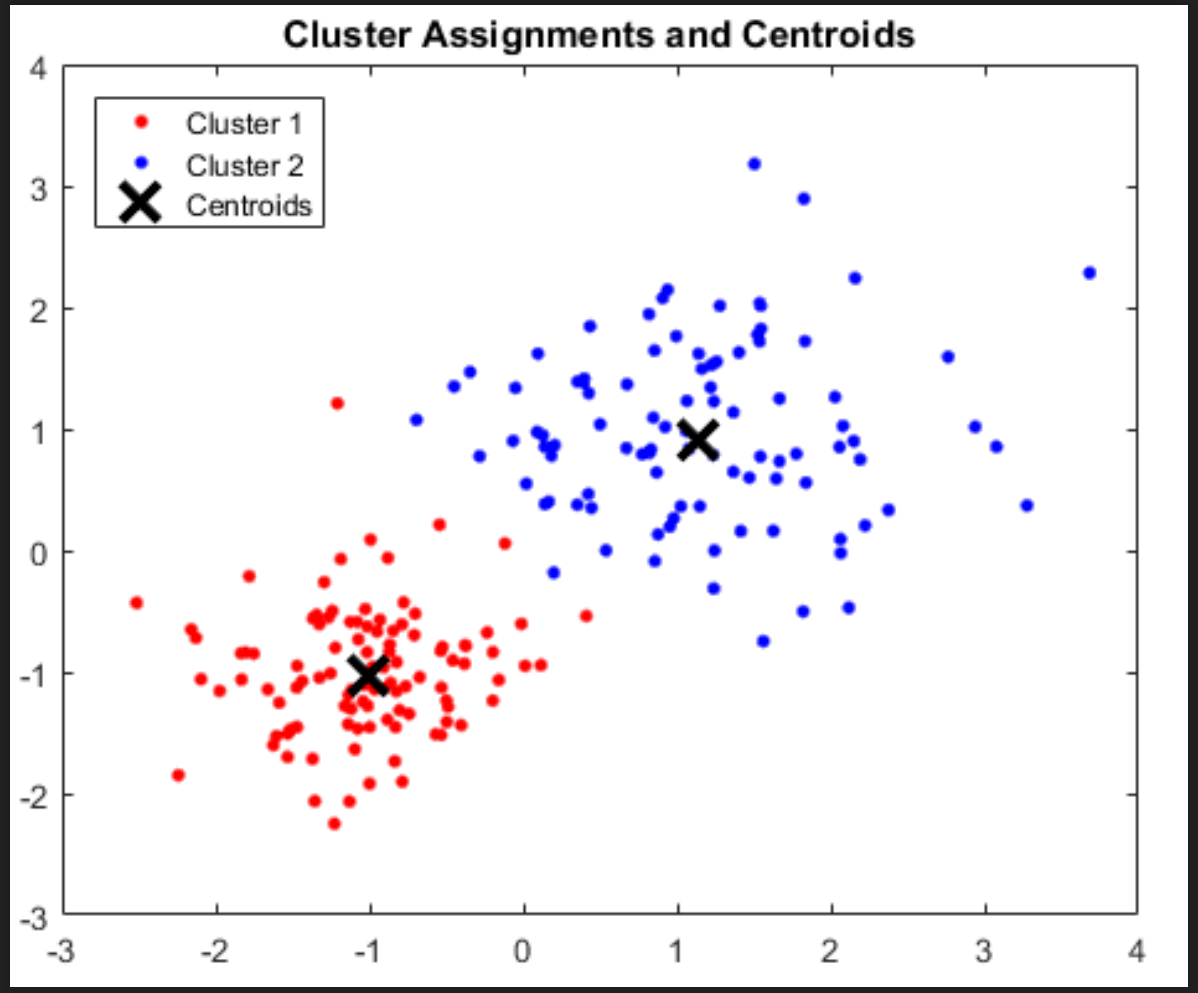
- Penempatan elemen pada clusters berdasarkan jarak terhadap centroid dari cluster yang ada (similarities).

Algoritma K-means

- Pilih K dokumen secara acak {s1, s2,... sK} sebagai “seed”.
- Lakukan iterasi:
  - Untuk setiap dokumen di, masukkan di ke cluster cj sehingga jarak(xi, sj) adalah minimum.
  - Perbaiki centroid tiap cluster cj

$$sj = \mu(cj)$$

Contoh K-means (K=2)

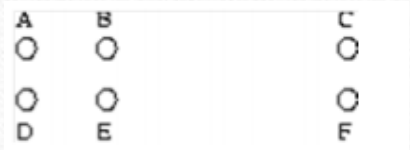


Kapan Iterasi Berhenti?

- Jumlah iterasi ditentukan
- Partisi dokumen tidak berubah
- Posisi centroid tidak berubah

Memilih Seed

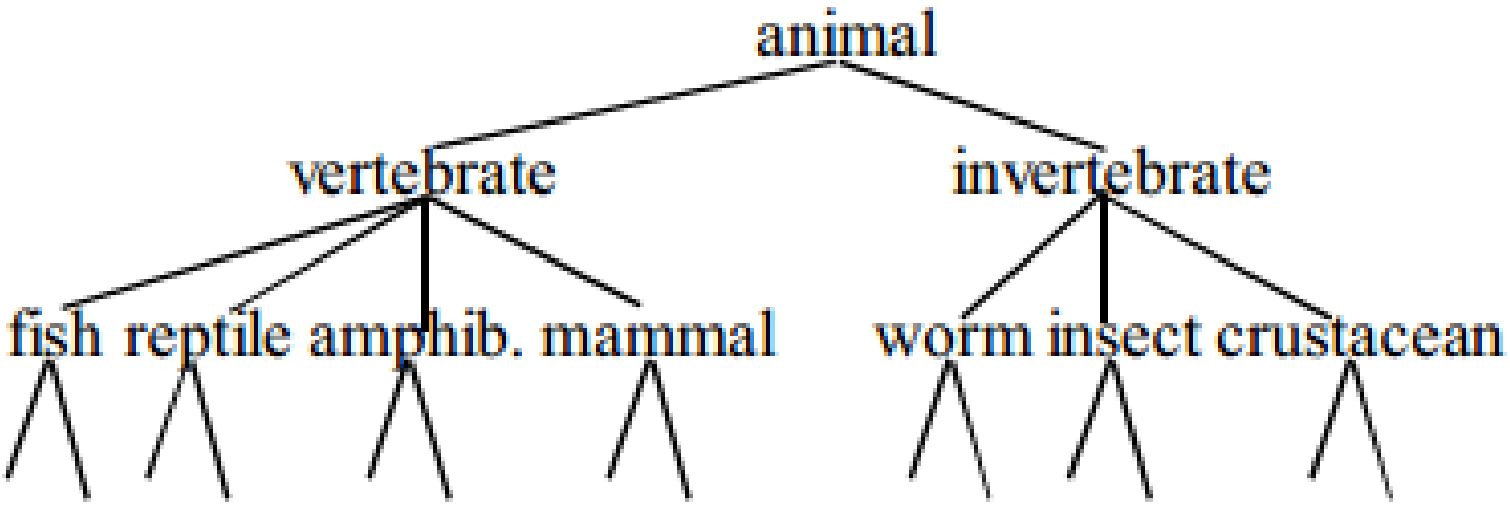
- Cluster yang dihasilkan tergantung pada pemilihan seed di awal (secara acak).
- Contoh



  - Jika mulai dengan B dan E sebagai centroid, maka akan konvergen ke {A,B,C} dan {D,E,F}
  - Jika mulai dengan D dan F sebagai centroid, maka akan konvergen ke {A,B,D,E} dan {C,F}

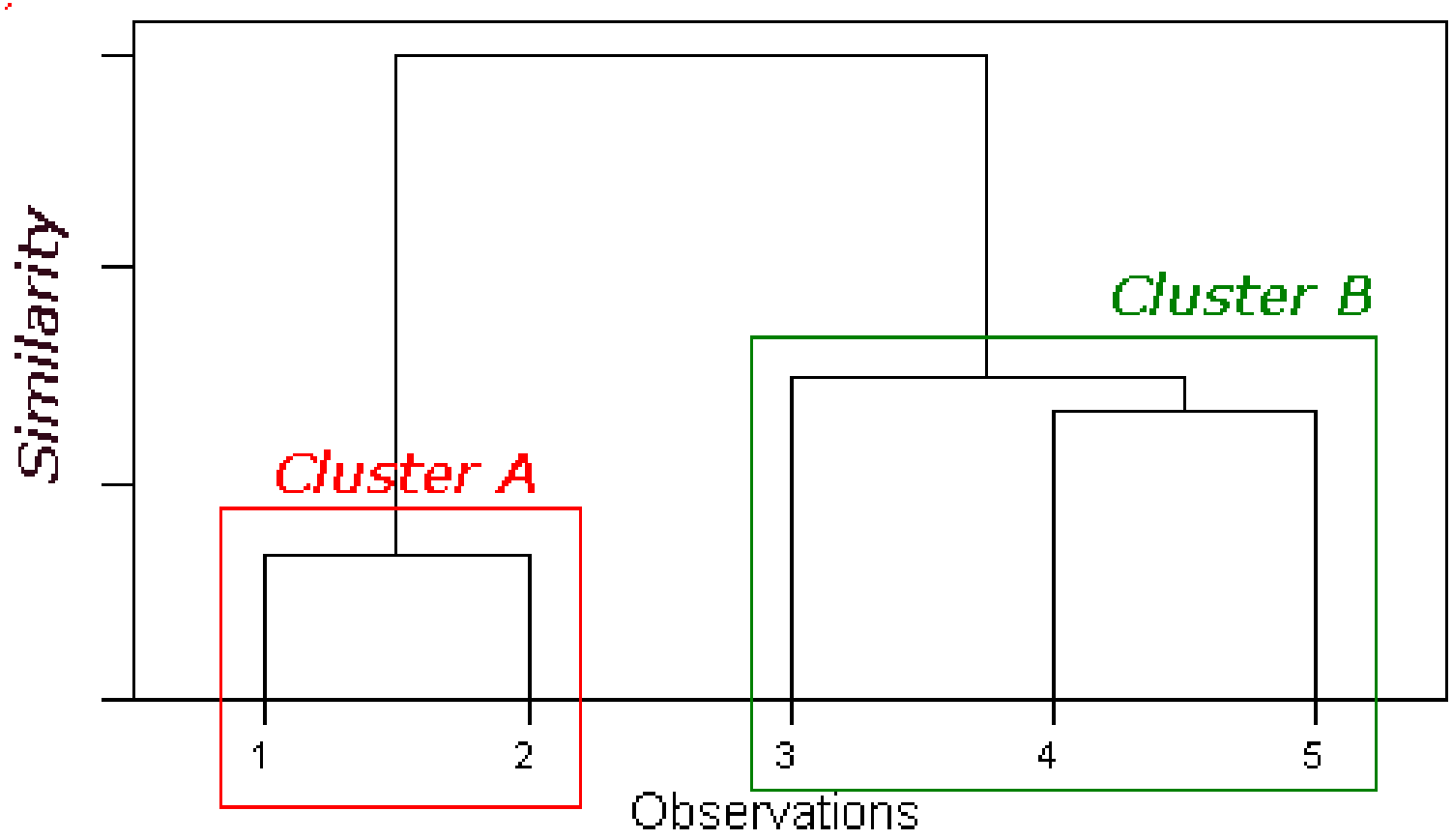
Hierarchical Clustering

- Membangun hirarki taksonomi berbasis denah pohon (dendogram) dari sekumpulan dokumen.



Dendogram

"Cluster diperoleh dengan memotong dendogram pada level Tertentu"



Hierarchical Agglomerative Clustering (HAC)

- Mulai dengan setiap dokumen sebagai suatu obyek tersendiri
- Gabungkan setiap obyek yang memiliki sifat sama (ukuran kesamaan paling tinggi, atau ukuran jarak paling kecil)
- Lakukan langkah kedua di atas seterusnya, dan berhenti jika semua obyek berada pada satu kelompok.

MENGGABUNGKAN CLUSTER

- **Single-link**
  - Menggunakan obyek yang paling dekat atau paling sama
- **Complete-link**
  - Menggunakan obyek yang paling jauh atau paling tidak sama
- **Average-link**
  - Menggunakan nilai rata-rata dari setiap anggota tiap cluster

Single Link

- Menggunakan ukuran kesamaan yang terbesar dari tiap pasangan.

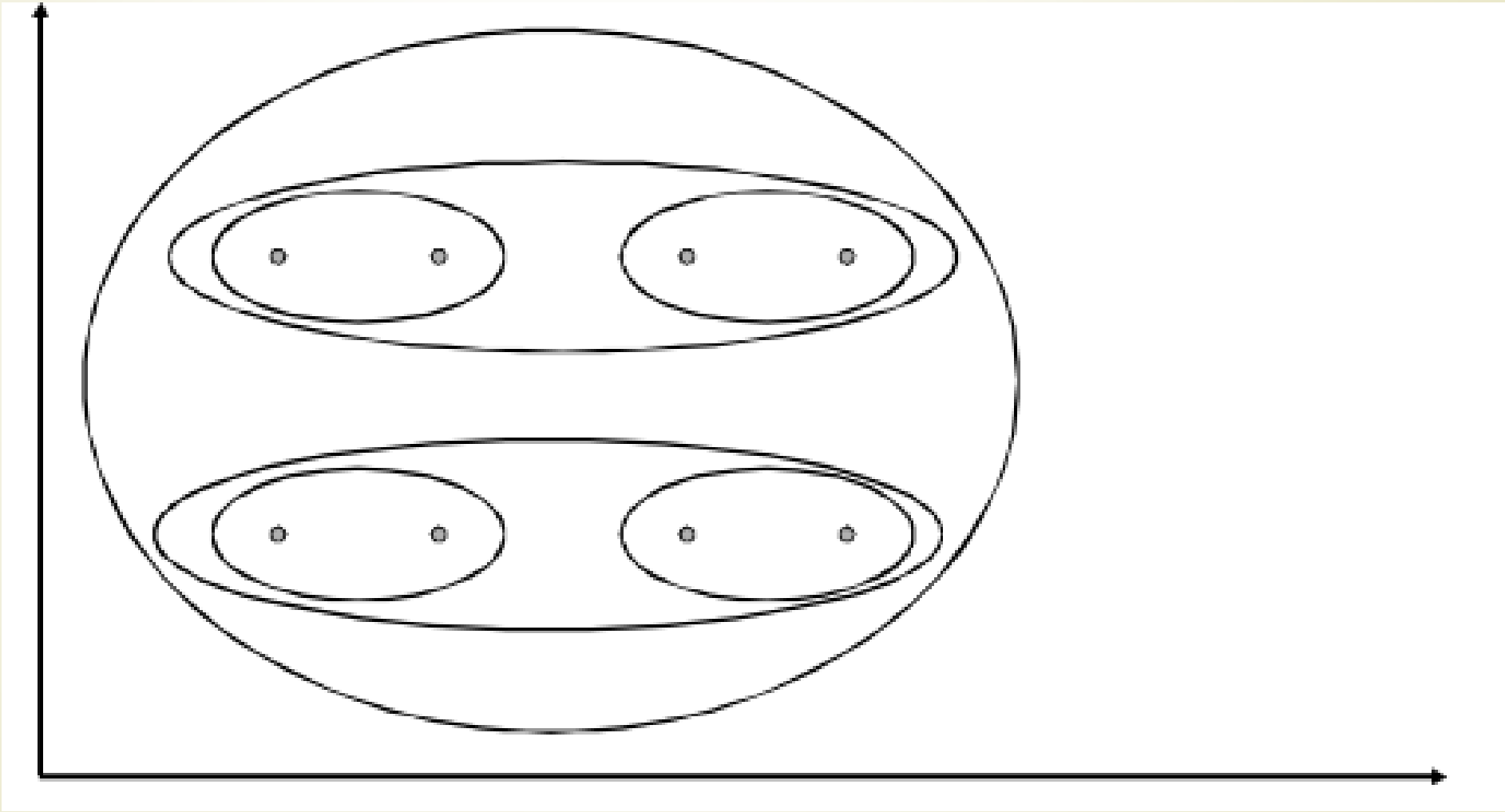
$$sim(C_i, C_j) = \max sim(x, y)$$

$$x \in C_i, y \in C_j$$

- Setelah C_i dan C_j digabung, maka ukuran kesamaan dari cluster yang dihasilkan dengan cluster lainnya, C_k adalah:

$$sim\left(\left(C_i \cup C_j\right), C_k\right) = max\left(sim(C_i, C_k), sim(C_j, C_k)\right)$$

Single Link



COMPLETE LINK

- Menggunakan ukuran kesamaan yang terkecil dari tiap pasangan.

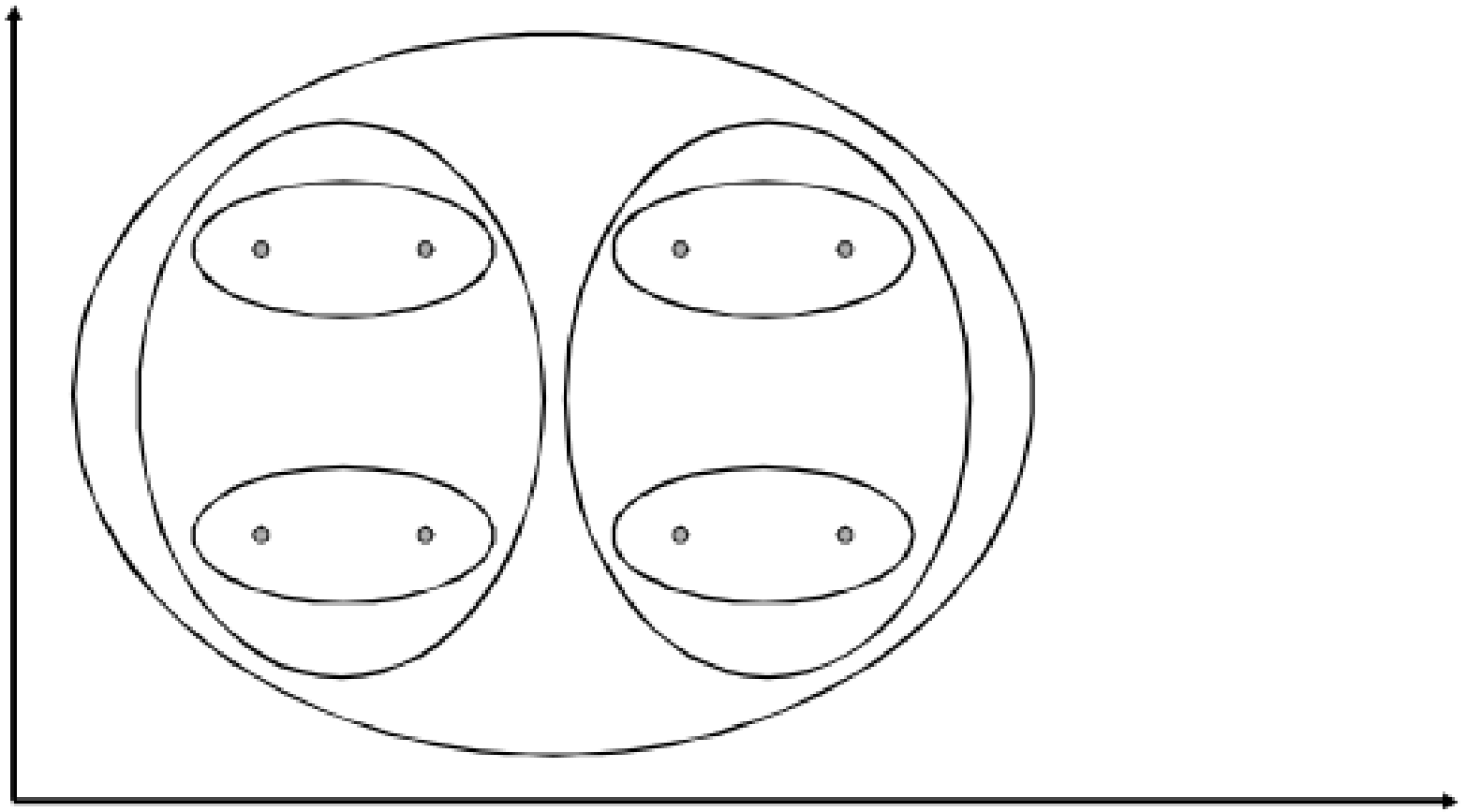
$$sim(C_i, C_j) = \min sim(x, y)$$

$$x \in C_i \text{ , y } \in C_j$$

- Setelah C_i dan C_j digabung, maka ukuran kesamaan dari cluster yang dihasilkan dengan cluster lainnya, C_k adalah:

$$sim\left(\left(C_i \cup C_j\right), C_k\right) = min\left(sim(C_i, C_k), sim(C_j, C_k)\right)$$

COMPLETE LINK



Average Link

- Menggunakan rata-rata dari pasangan ukuran kesamaan.

$$sim(c_i, c_j) = \frac{1}{|c_i \cup c_j|(|c_i \cup c_j| - 1)} \sum_{\vec{x} \in (c_i \cup c_j)} \sum_{\vec{y} \in (c_i \cup c_j): \vec{y} \neq \vec{x}} sim(\vec{x}, \vec{y})$$

- Merupakan kompromi dari pendekatan single link dan complete link.

BAGAIMANA CLUSTER YANG BAIK ?

- Kriteria internal: menghasilkan cluster yang baik dimana:
  - Kesamaan antar anggota dalam suatu cluster (intracluster) adalah tinggi
  - Kesamaan antar anggota dari cluster yang berbeda (inter-cluster) adalah rendah
  - Kualitas ukuran tergantung pada representasi dokumen dan ukuran kesamaan yang digunakan

Bagaimana Cluster yang Baik ?

- Kriteria eksternal: diukur dengan menggunakan data kelas yang baik yang sudah diketahui (gold standard).
- Asumsikan ada C kelas-kelas yang baik (gold standard), sedangkan algoritma cluster kita menghasilkan *k* clusters, Π_1, Π_2, …, Π_k dengan n_j anggota.
- Purity, rasio antara kelas yang dominan pada cluster Π_i dan ukuran cluster Π_i

$$Purity(\omega_i) = \frac{1}{n_i} \max_j (n_{ij}) \quad j \in C$$

Studi Kasus Clustering dengan algoritma K-Means

Tabel Data nasabah

Nasabah	Jumlah Rumah	Jumlah Mobil
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Clustering yang diharapkan mampu menghasilkan kelompok nasabah yang memenuhi sifat berikut:

1. Nasabah yang jumlah rumah dan mobilnya hampir sama akan berada pada kelompok nasabah yang sama.
2. Nasabah yang jumlah rumah dan mobilnya cukup berbeda akan berada pada kelompok nasabah yang berbeda.

Berikut langkah-langkah clustering menggunakan algoritma K-Means.

1. **Langkah 1**: Tentukan jumlah cluster yang diinginkan (misl:k=3)
2. **Langkah 2**: Pilih centroid awal secara acak : Pada langkah ini secara acak akan dipilih 3 buah data sebagai centroid, misalnya: data {B,E,F}

   M1=(3,3) ,M2=(1,2),M3=(4,2)
3. **Langkah 3:** Hitung jarak dengan centroid ..... (iterasi 1)

   Pada langkah ini setiap data akan ditentukan centroid terdekatnya, dan data tersebut akan ditetapkan sebagai anggota kelompok yang terdekat dengan centroid.

   Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

   Data: (1,3) , centroid M1: (3,3), centroid M2: (1,2), centroid M3: (4,2)

   DM1 = $\sqrt{(1-3)^2 + (3-3)^2}$ = 2

   DM2 = $\sqrt{(1-1)^2 + (3-2)^2}$ = 1

   DM3 = $\sqrt{(1-4)^2 + (3-2)^2}$ = 3.162

Tabel hasil perhitungan jarak selengkapnya antara masing-masing data dengan centroid:

Nasabah	Jarak ke centroid cluster1	Jarak ke centroid cluster2	Jarak ke centroid cluster3	Jarak terdekat
A	2	1	3.162	C2
B	0	2.236	1.414	C1
C	1	3.162	1	C3
D	2	4.123	1.414	C3
E	2.236	0	3	C2
F	1.414	3	0	C3
G	2.828	1	3.162	C2
H	2.236	1.414	2.236	C2

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

Cluster 1 = {B},cluster 2 ={A,E,G,H},cluster 3= {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

Karena centroid M1=(3,3) ,M2=(1,2),M3=(4,2)

d(m1,m2) = $\sqrt{(3-1)^2+(3-2)^2}$ = 2.236

d(m1,m3) = $\sqrt{(3-4)^2+(3-2)^2}$ = 1.414

d(m2,m3) = $\sqrt{(1-4)^2+(2-2)^2}$ = 3

BCV=d(m1,m2)+d(m1,m3)+d(m2,m3) = 2.236+1.414+3 = 6,650

Dalam hal ini d(mi,mj) menyatakan jarak Euclidean dari m ke mj

Menghitung WCV

Yaitu dengan memilih jarak terkecil antara data dengan centroid pada masing-masing cluster:

nasabah	Jarak ke centroid terkecil
A	1
B	0
C	1
D	1.414
E	0
F	0
G	1
H	1.414

$WCV=1^2+0^2+1^2+1.414^2+0^2+0^2+1^2+1.414^2=7$

Sehingga Besar Rasio = BCV/WCV = 6.650 / 7 = 0.950

Karena langkah ini merupakan iterasi 1 maka lanjutkan ke langkah berikutnya

PEMBARUAN CENTROID DENGAN MENGHITUNG RATA-RATA NILAI PADA MASING-MASING CLUSTER.

Cluster 1		
Nasabah	Jml Rumah	Jml Mobil
B	3	3
Mean	3	3
Cluster 2		
Nasabah	Jml Rumah	Jml Mobil
A	1	3
E	1	2
G	1	1
H	2	1
Mean	1.25	1.75
Cluster 3		
Nasabah	Jml Rumah	Jml Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

Sehingga didapatkan centroid baru yaitu :

m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

Langkah 3: (Iterasi-2) Kembali kelangkah 3, jika masih ada data yang berpindah cluster atau jika nilai centroid diatas nilai ambang, atau jika nilai pada fungsi obyektif yang digunakan masih diatas ambang. Selanjutnya pada langkah ini dilakukan penempatan lagi data dalam centroid terdekat sama seperti yang dilakukan dilangkah-3. Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

Data : (1,3) , m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

DM1 = $\sqrt{(1-3)^2 + (3-3)^2}$ = 2

D M2 = $\sqrt{(1-1.25)^2 + (3-1.75)^2}$ = 1.275

DM3 = $\sqrt{(1-4.33)^2 + (3-2.67)^2}$ = 3.350

Nasabah	Jarak ke centroid custer1	Jarak ke centroid custer2	Jarak ke centroid custer3	Jarak terdekat
A	2	1.275	3.350	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

Cluster 1 = {B},cluster 2 ={A,E,G,H},cluster 3= {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

BCV=d(m1,m2)+d(m1,m3)+d(m2,m3) = 6,741

$WCV=1.275^2+0^2+0.471^2+0.745^2+0.354^2+0.745^2+0.791^2+1.061^2=4.833$

Sehingga Besar Rasio = BCV/WCV = 6.741 /4.833 = 1.394

Bila dibandingkan maka rasio sekarang (1.394) lebih besar dari rasio sebelumnya (0.950) oleh karena itu algoritma dilanjutkan kelangkah berikutnya

Langkah ke 4 – iterasi 3
Pada langkah ini dilakukan pembaruan centroid lagi:

Cluster 1		
Nasabah	Jml Rumah	Jml Mobil
B	3	3
Mean	3	3
Cluster 2		
Nasabah	Jml Rumah	Jml Mobil
A	1	3
E	1	2
G	1	1
H	2	1
Mean	1.25	1.75
Cluster 3		
Nasabah	Jml Rumah	Jml Mobil
C	4	3
D	5	3
F	4	2
Mean	4.33	2.67

Langkah ketiga iterasi-3

Untuk menghitung jarak ke centroid masing-masing cluster pada nasabah A sbb:

Data nasabah A : (1,3) , m1=(3,3),m2=(1.25,1.75),m3=(4.33,2.67)

DM1 $=\sqrt{(1-3)^2+(3-3)^2}=2$

D M2 $=\sqrt{(1-1.25)^2+(3-1.75)^2}=1.275$

DM3 $=\sqrt{(1-4.33)^2+(3-2.67)^2}=3.350$

Nasabah	Jarak ke centroid custer1	Jarak ke centroid custer2	Jarak ke centroid custer3	Jarak terdekat
A	2	1.275	3.350	C2
B	0	1.768	1.374	C1
C	1	3.021	0.471	C3
D	2	3.953	0.745	C3
E	2.236	0.354	3.399	C2
F	1.414	2.813	0.745	C3
G	2.828	0.791	3.727	C2
H	2.236	1.061	2.867	C2

Dari tabel diatas didapatkan keanggotaan nasabah sbb:

Cluster 1 = {B},cluster 2 ={A,E,G,H},cluster 3= {C,D,F}

Pada langkah ini dihitung pula rasio antara besaran BCV (*Between Cluster Variation*) dengan WCV (*Within Cluster Variation*) :

BCV=d(m1,m2)+d(m1,m3)+d(m2,m3) = 6,741

$WCV=1.275^2+0^2+0.471^2+0.745^2+0.354^2+0.745^2+0.791^2+1.061^2=4.833$

Sehingga Besar Rasio = BCV/WCV = 6.741 /4.833 = 1.394

Bila dibandingkan maka rasio sekarang (1.394) sudah tidak lagi lebih besar dari rasio sebelumnya (1.394) oleh karena itu algoritma akan dihentikan.

1. Tentukan jumlah cluster (k=3)
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat cluster (centroid) menggunakan mean untuk masing-masing kelompok
4. Alokasikan masing-masing data ke centroid terdekat
5. Kembali ke langkah 3, jika masih ada data yang berpindah cluster atau jika nilai centroid diatas nilai ambang, atau jika nilai pada fungsi obyektif yang digunakan masih diatas ambang

	d1	d2	d3	d4	d5
car	3	4	4	4	3
auto	7	3	0	0	1
insurance	0	3	9	0	2
best	4	0	7	5	0