ANALISIS CLUSTER PADA DOKUMEN TEKS

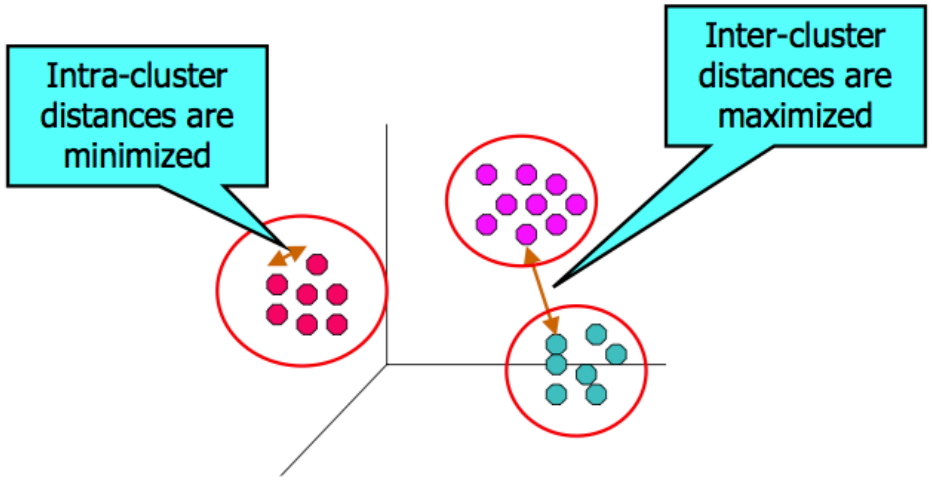
Budi Susanto (versi 1.3)

Tujuan

- Memahami konsep analisis clustering
- Memahami tipe-tipe data dalam clustering
- Memahami beberapa algoritma clustering:
  - K-Means
  - K-Medoids
  - Nearest Neighbor
  - Hierarchical Clustering
- Menjelaskan implementasi algoritma clustering pada text corpus.

Analisis Cluster

- Analisis cluster adalah pengorganisasian kumpulan pola ke dalam cluster (kelompok-kelompok) berdasar atas kesamaannya.
- Pola-pola dalam suatu cluster akan memiliki kesamaan ciri/sifat daripada pola-pola dalam cluster yang lainnya.



Analisis Cluster

- *Clustering* bermanfaat untuk melakukan analisis pola-pola yang ada, mengelompokkan, termasuk *data mining*, *document retrieval*, *segmentasi citra*, dan klasifikasi pola.
- Metodologi clustering lebih cocok digunakan untuk eksplorasi hubungan antar data untuk membuat suatu penilaian terhadap strukturnya.

Problem Statement

- Dinyatakan:
  - himpunan dokumen D = {d_1, d_2, …, d_N}
  - Jumlah cluster target, *K*.
  - Fungsi objektif untuk evaluasi kualitas *clustering*.
    - Fungsi objektif didefinisikan dalam istilah kemiripan atau jarak antar dokumen.
- Ingin dihitung persamaan γ = D➔{1, …, *K*} yang meminimalkan fungsi objektif atau memastikan tidak ada K cluster yang kosong.

Jumlah Cluster, K

- Dalam kebanyakan algoritma clustering, pemilihan jumlah cluster sangat ditentukan oleh proses inisialisasi.
- Beberapa rule of thumb[1]:
  - $k \approx \sqrt{\frac{n}{2}}$
  - *(m x n)/t*
    - m = jumlah dokumen
    - n = jumlah term
    - t = jumlah non-zero entri

D =	t_1	t_2	t_3	t_4	t_5	t_6	
	1	1	0	0	1	0	d_1
	1	1	0	1	1	0	d_2
	0	0	0	0	0	1	d_3
	0	1	1	0	0	1	d_4
	0	1	1	1	0	1	d_5

http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#cite_note-0

Tipe Clustering

- Partitional Clustering
  - Pembagian objek data ke dalam non-overlapping subset (cluster) sehingga setiap objek data adalah tepat satu subset
- Hirerarchical Clustering
  - Sehimpunan cluster bersarang yang diorganisasikan sebagai struktur hirarki pohon.

Tipe Cluster

- Well-separated clusters
- Center-based clusters
- Density-based clusters

Well-separated

- Sebuah cluster adalah sehimpunan titik yang memiliki kemiripan dengan titik lain dalam cluster daripada di cluster lain.

Center-based

- Sebuah cluster yang memiliki anggota-anggota yang mirip dengan pusat cluster daripada pusat cluster lain.
- Pusat cluster
  - Centroid: Rata-rata dari semua titik dalam cluster
  - Medoid: memilih titik sebagi titik tengah.

Density-based

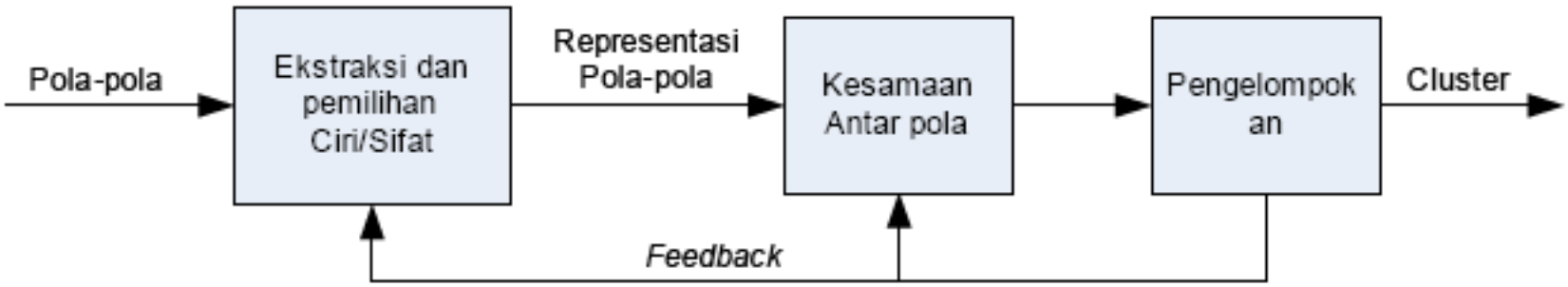
- Sebuah cluster adalah area padat titik, yang dipisahkan dengan area kepadatan rendah, dari area kepadatan tinggi lainnya.
- Digunakan ketika cluster tidak teratur atau saling terkait, dan ketika noise dan outliers hadir.

Komponen

- representasi pola (termasuk ekstraksi sifat/ciri dan atau pemilihan),
- definisi ukuran kedekatan pola sesuai dengan domain data,
- clustering atau pengelompokan,
- jika diperlukan, abstraksi data (proses ekstraksi untuk deksripsi cluster),
- jika diperlukan, penilaian terhadap hasil (menggunakan metode pengukuran dan pengujian terhadap hasil clustering apakah valid atau tidak).

Tahapan Clustering

- Kedekatan pola biasanya diukur dengan fungsi jarak antar dua pasang pola.
  - *cosine similarity*, *manhattan distance*, dan *euclidean distance*.



Tahapan Clustering

- Representasi pola (*pattern representation*) merupakan jumlah kelas, jumlah pola yang ada, jumlah, tipe dan skala ciri/sifat yang tersedia untuk algoritma clustering.
- Pemilihan ciri/sifat (*feature selection*) adalah proses identifikasi ciri/sifat yang lebih efektif untuk digunakan dalam algoritma clustering, sedangkan ekstraksi ciri/sifat adalah pemakaian satu atau lebih transformasi dari ciri/ sifat yang ada sebelumnya untuk mendapatkan ciri/sifat yang lebih menonjol.

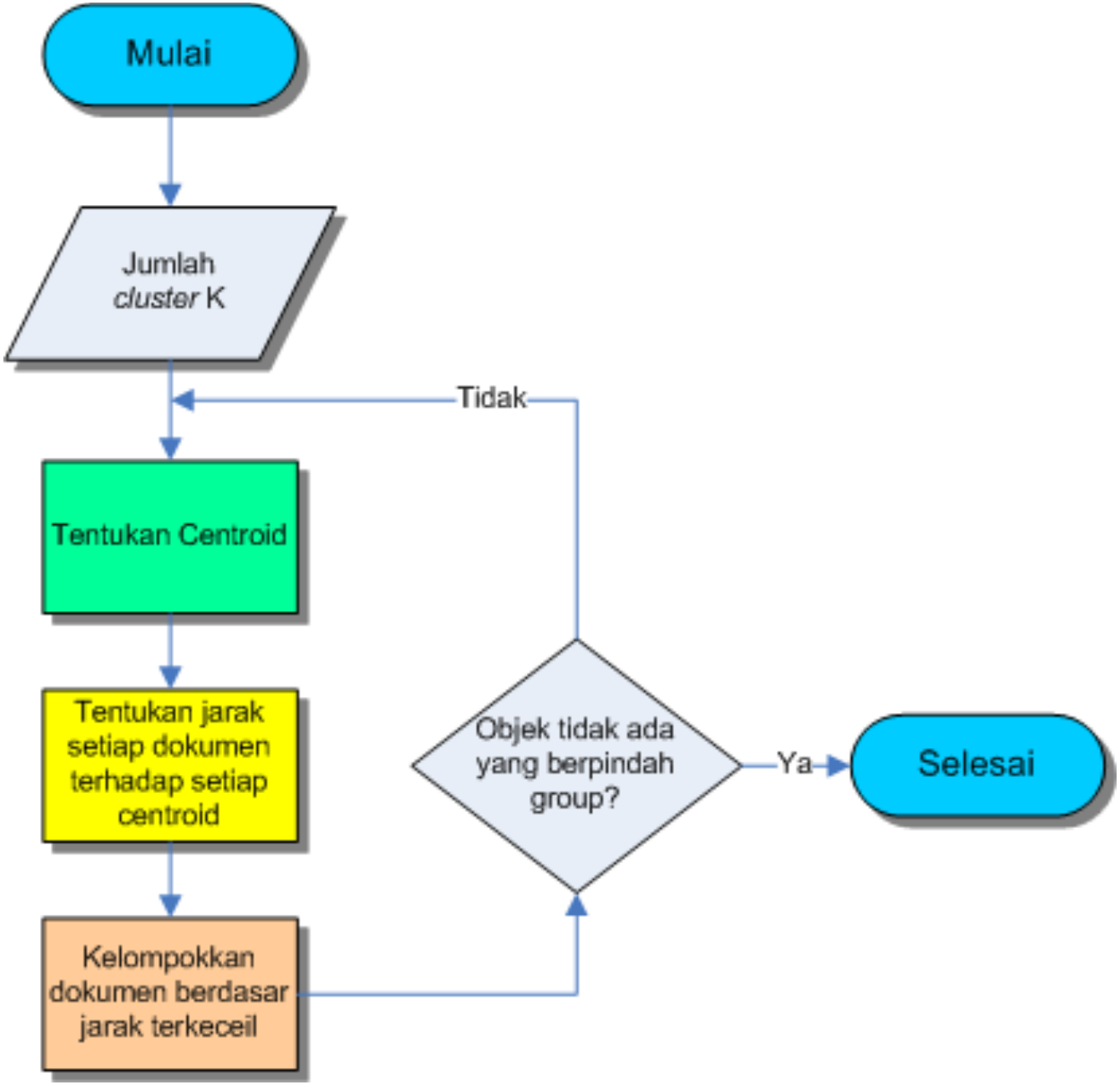
Tahapan Clustering

- Kedekatan pola biasanya diukur dengan fungsi jarak antar dua pasang pola.
- Pengukuran jarak yang sederhana, seperti *Euclidean distance*, *Minkowski*, *Hamming distance*, sering digunakan untuk menyatakan ketidaksamaan antara dua pola
- Sedangkan pengukuran kesamaan lain, seperti *Simple Matching Coefficient*, *Jaccard Coefficient*, *Cosine Similarity*, dapat digunakan untuk menunjukkan kesamaan karakter antar pola-pola.

k-Means

- Partitional clustering
- Setiap cluster terasosiasi dengan sebuah centroid
- Setiap titip dinyatakan ke suatu cluster yang paling dekat dengan centroidnya.
- Jumlah cluster, K, dinyatakan di awal

K-Means


Mulai
Jumlah cluster K
Tidak
Tentukan Centroid
Tentukan jarak setiap dokumen terhadap setiap centroid
Kelompokkan dokumen berdasar jarak terkeceil
Objek tidak ada yang berpindah group?
Ya
Selesai

Contoh K-Means

- Kelompokkan dataset berikut ke dalam 3 kelompok dengan k-means (2 epoch saja):
  - A1=(2,10)
  - A2=(2,5)
  - A3=(8,4)
  - A4=(5,8)
  - A5=(7,5)
  - A6=(6,4)
  - A7=(1,2)
  - A8=(4,9)

Keterbatasan K-Means

- K-Mean bermasalah ketika cluster-cluster berbeda
  - Ukuran
  - Kepadatan
  - Tidak berbentuk bola
- K-Mean bermasalah ketika data berisi outlier

Partitioning Around Medoids (PAM)

- Seperti metode partisi clustering yang lainnya, metode PAM juga digunakan untuk pengelompokkan dokumen.
- Dalam metode PAM ini setiap cluster dipresentasikan dari sebuah objek di dalam cluster yang disebut dengan *medoid*.
- Tujuannya adalah menemukan kelompok *k-cluster* (jumlah cluster) diantara semua objek data di dalam sebuah kelompok data.
- Clusternya dibangun dari hasil mencocokkan setiap objek data yang paling dekat dengan cluster yang dianggap sebagai *medoid* sementara.

K-Medoids

1. pilih point k sebagai inisial centroid / nilai tengah (medoids) sebanyak k cluster.
2. cari semua point yang paling dekat dengan medoid, dengan cara menghitung jarak vector antar dokumen. (menggunakan Euclidian distance)
3. secara random, pilih point yang bukan medoid.
4. hitung total distance
5. if TD baru < TD awal, tukar posisi medoid dengan medoids baru, jadilah medoid yang baru.
6. ulangi langkah 2 - 5 sampai medoid tidak berubah.

Contoh K-Medoids

x1	**4**	**7**
X2	6	2
X3	7	3
X4	8	5
X5	3	4
X6	7	6

Nearest Neighbor clustering

- Sebuah titik membentuk cluster baru atau bergabung dengan salah satu cluster yang sudah ada tergantung pada seberapa dekat titik tersebut dengan *cluster*.
  - Sebuah treshold, t, untuk menentukan bergabung atau membuat cluster baru.

Nearest Neighbor clustering

Input:

$D=\{t_1, t_2, \ldots, t_n\}$ // Set of elements
A // Adjacency matrix showing distance between elements
θ // threshold

Output:

K //Set of k clusters

Nearest-Neighbor algorithm

$K_1 =\{t_1\}$; add K_1 to K; // t1 initialized the first cluster
k = 1;
for i =2 to n **do** // for t_2 to t_n add to existing cluster or place in new one
find the t_m in some cluster K_m in K such that $d(t_m,t_i)$ is the smallest;
if $d(t_m,t_i) < \theta$ **then**
$K_m = K_m \cup \{t_i\}$ // existing cluster
else
$k = k + 1$; $K_k = \{t_i\}$; add K_k to K // new cluster

Latihan NN

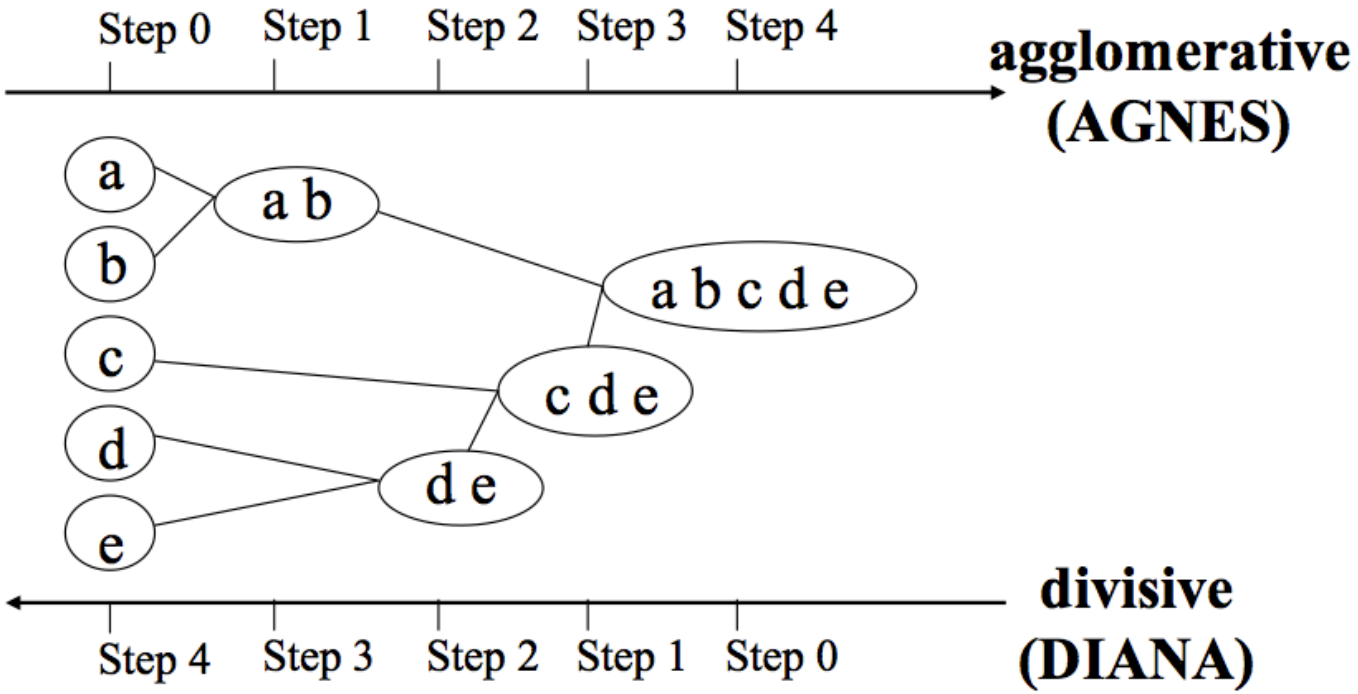
- Kelompokkan dataset berikut ke dalam 3 kelompok dengan NN clustering (2 epoch saja):
  - A1=(2,10)
  - A2=(2,5)
  - A3=(8,4)
  - A4=(5,8)
  - A5=(7,5)
  - A6=(6,4)
  - A7=(1,2)
  - A8=(4,9)

Hierarchical Clustering

- Membentuk beberapa himpunan cluster
  - Jumlah cluster tidak dimasukkan di awal
- Struktur hirarki cluster dapat dipresentasikan sebagai dendrogram.
  - Daun berisi 1 item.
  - Setiap item masuk dalam satu cluster
  - Root mewakili semua item
  - Internal node menyatakan cluster yang dibentuk oleh penggabungan cluster anak.
  - Setiap level diasosiasikan dengan suatu treshold jarak yang digunakan untuk menggabungkan cluster
    - Jika jarak antar 2 cluster lebih kecil dari treshold, maka digabungkan.
    - Jarak akan bertambah sesuai dengan level.

Hierarchical Clustering

- Menggunakan matrik jarak sebagai kriteria clustering. Metode ini tidak memerlukan jumlah cluster, K, sebagai inputan, namun butuh kondisi terminasi.



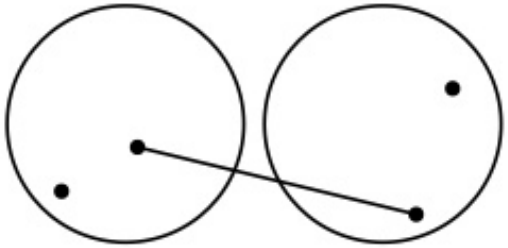
Hierarchical Clustering

- Agglomerative
  - dimulai dari asumsi bahwa setiap objek dalam kumpulan data sebagai cluster individu (*singleton cluster*),
  - langkah selanjutnya menggabungkan antar singleton cluster berdasar jarak terdekatnya.
- Divisive
  - dimulai dengan asumsi bahwa seluruh objek dalam kumpulan data sebagai satu cluster,
  - cluster tersebut akan dipecah sampai semua objek merupakan *singleton cluster*.

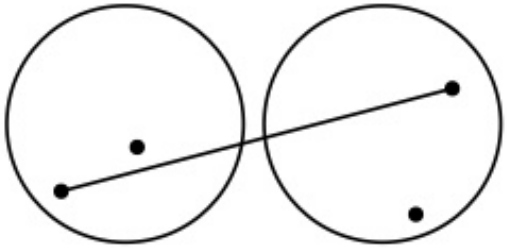
Penentuan Nilai Proximity Cluster

- Single Link
  - Nilai proximity cluster diperoleh dari nilai proximity terdekat (paling mirip) antara dua objek yang berada di cluster yang berbeda.
- Complete Link
  - Nilai proximity cluster diperoleh dari nilai proximity terjauh (paling tidak mirip) antara dua objek yang berada di cluster yang berbeda.
- Centroid
  - Nilai proximity cluster merupakan nilai rata-rata jarak pasangan objek antar cluster.
- Group Average
  - Nilai proximity cluster merupakan nilai rata-rata dari seluruh pasangan objek di cluster yang berbeda.

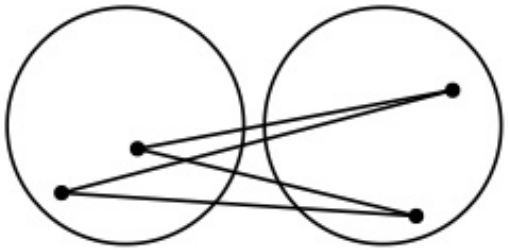
Ilustrasi Cluster Simmilarity



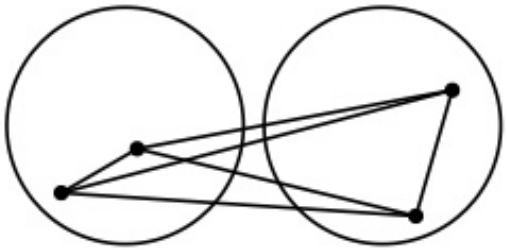
(a) single-link: maximum similarity



(b) complete-link: minimum similarity

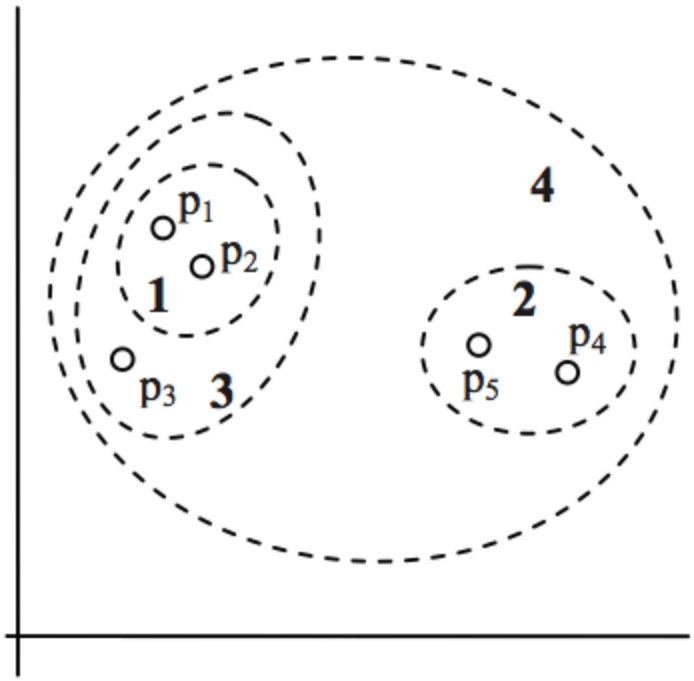


(c) centroid: average inter-similarity

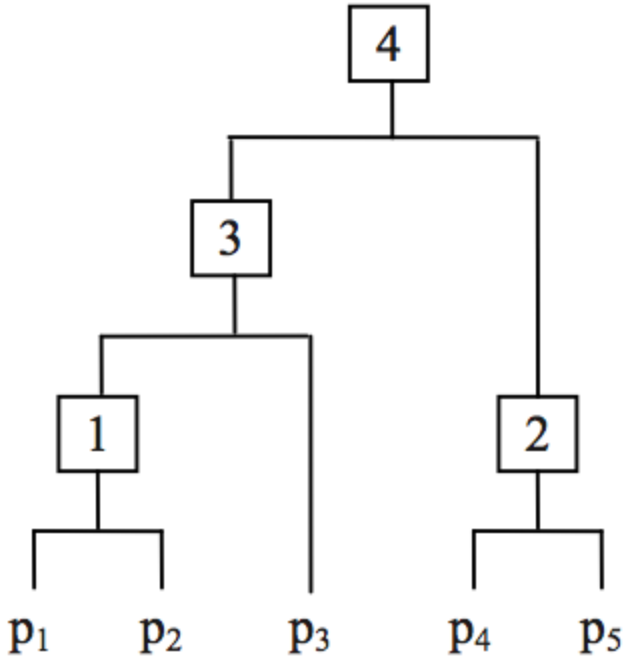


(d) group-average: average of all similarities

Presentasi Hierarchical Clustering

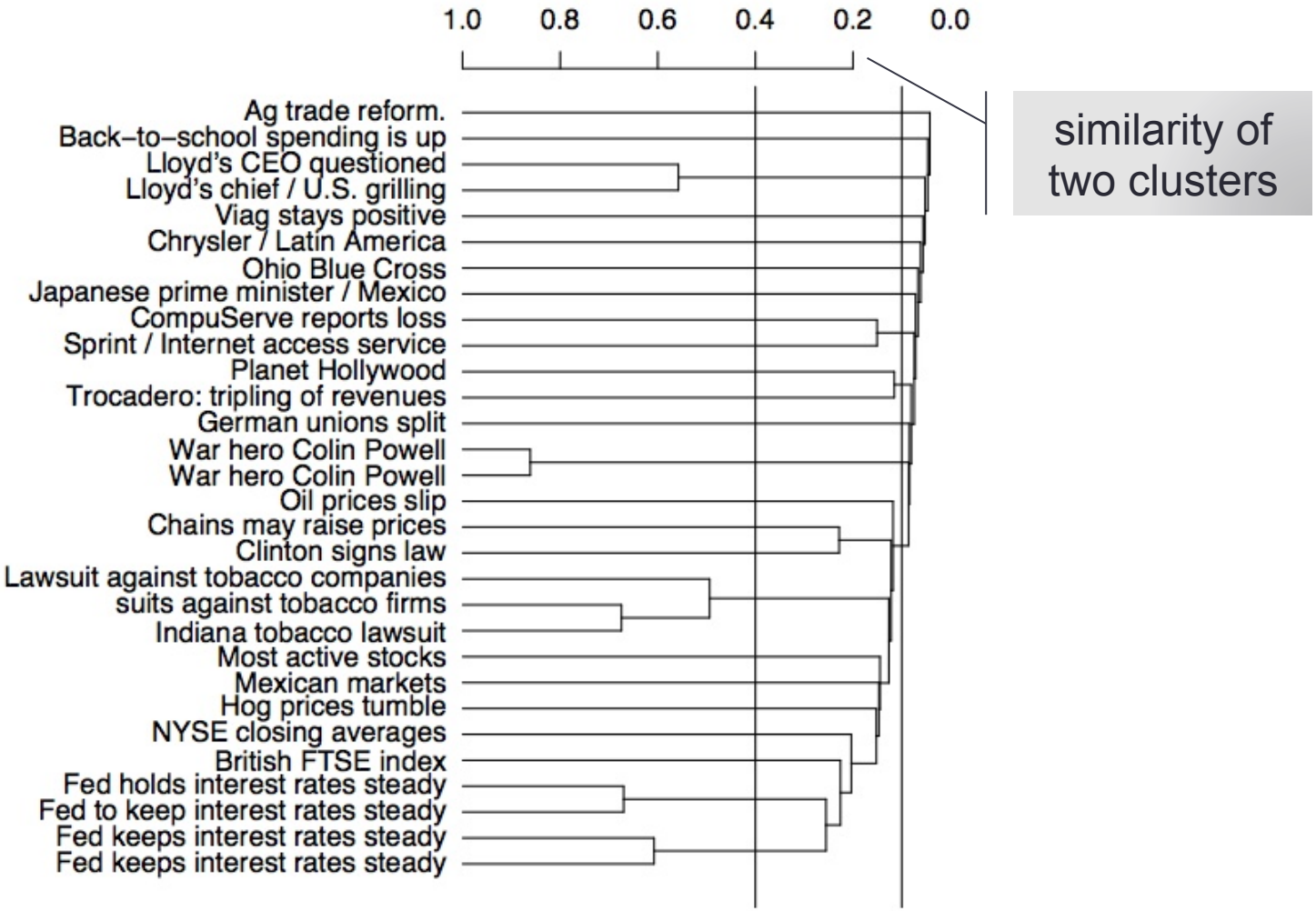


(A). Nested clusters

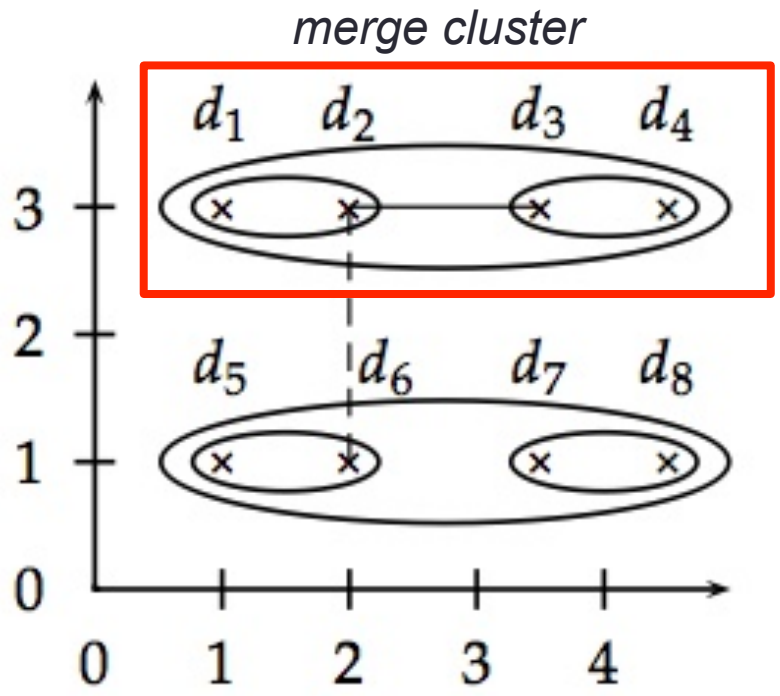


(B) Dendrogram

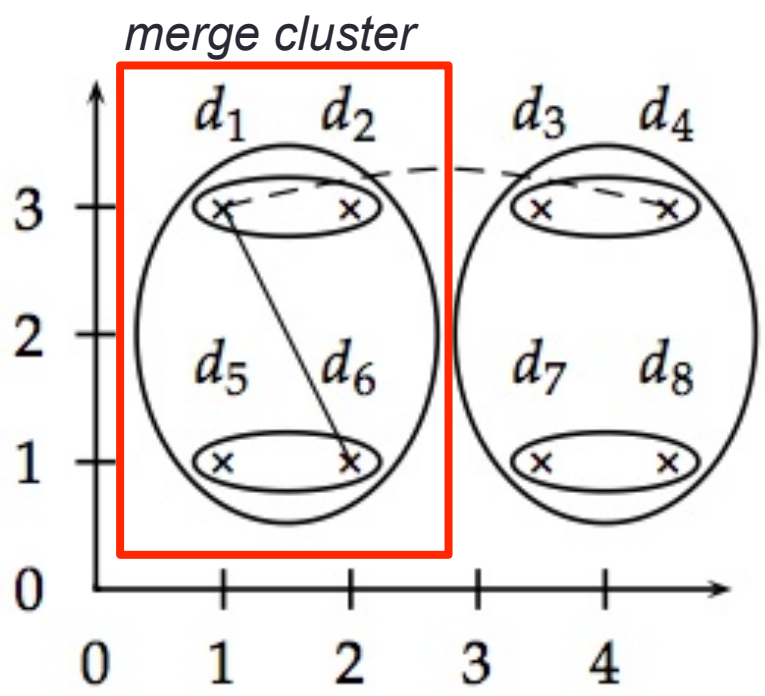
Contoh Dendogram



Single Link dan Complete Link



Single Link



Complete Link

Hierarchical Clustering

```
SIMPLEHAC(d_1, ..., d_N)
 1  for n ← 1 to N
 2  do for i ← 1 to N
 3       do C[n][i] ← SIM(d_n, d_i)
 4       I[n] ← 1  (keeps track of active clusters)
 5  A ← []  (assembles clustering as a sequence of merges)
 6  for k ← 1 to N − 1
 7  do ⟨i, m⟩ ← arg max_{⟨i,m⟩: i≠m ∧ I[i]=1 ∧ I[m]=1} C[i][m]
 8       A.APPEND(⟨i, m⟩)  (store merge)
 9       for j ← 1 to N
10       do C[i][j] ← SIM(i, m, j)
11            C[j][i] ← SIM(i, m, j)
12       I[m] ← 0  (deactivate cluster)
13  return A
```

Contoh Single Link HAC

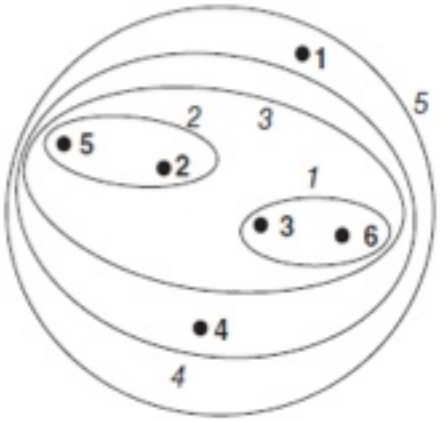
- Contoh diberikan data sebagai berikut:

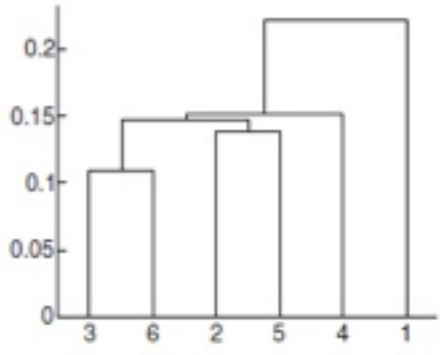
Titik	x	y
a1	0.40	0.53
a2	0.22	0.38
a3	0.35	0.32
a4	0.26	0.19
a5	0.08	0.41
a6	0.45	0.30

	a1	a2	a3	a4	a5	a6
a1	0	0.24	0.22	0.37	0.34	0.23
a2	0.24	0	0.15	0.2	0.14	0.25
a3	0.22	0.15	0	0.15	0.28	0.11
a4	0.37	0.2	0.15	0	0.29	0.22
a5	0.34	0.14	0.28	0.29	0	0.39
a6	0.23	0.25	0.11	0.22	0.39	0

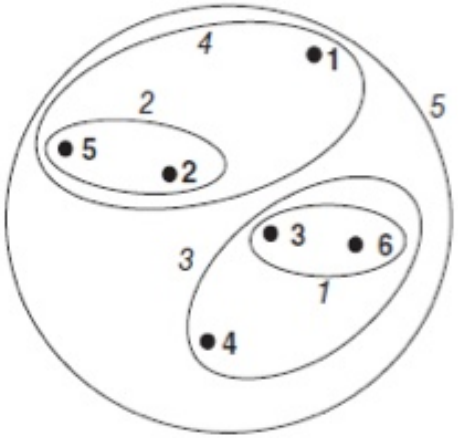
Contoh HAC

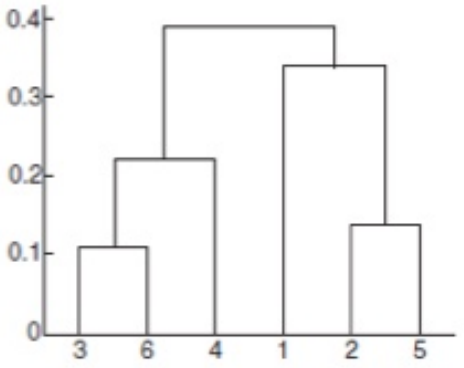
- Single Link





- Complete Link





(a) Complete link clustering.

(b) Complete link dendrogram.

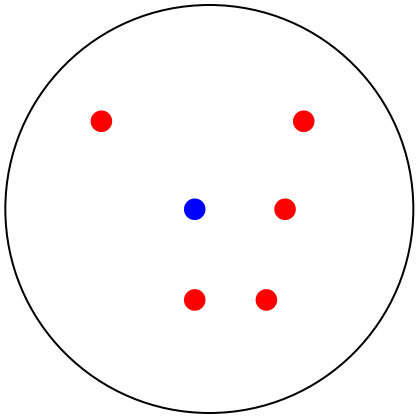
Evaluasi Clustering

- Purity : rasio antara class dominan dalam cluster c_j dan ukuran cluster ω_j

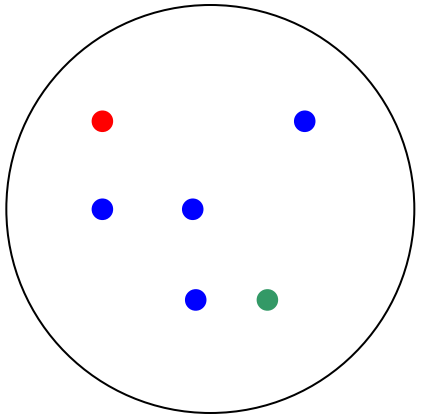
$$\text{purity}(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

- Dimana:
  - $\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ adalah himpunan cluster
    - ω_K ➔himpunan dokumen dalam ω_K.
  - $C = \{c_1, c_2, ..., c_J\}$ adalah himpunan class
    - c_j ➔himpunan dokumen dalam c_j.
- Clustering buruk jika nilai purity mendekati 0, dan baik jika nilai purity mendekati 1.

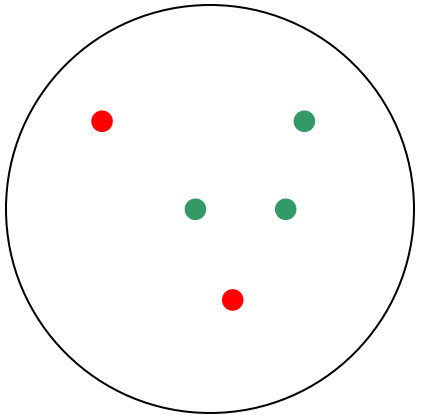
Contoh Purity



Cluster I



Cluster II



Cluster III

Purity = 1/17 * (5+4+3) = 12/17

Rand Index

- Evaluasi lain adalah menghitung prosentase terhadap keputusan benar dalam tiap cluster.
  - setiap cluster terdiri dari N(N-1)/2 pasangan dokumen
- Dua dokumen dalam cluster yang sama jika dan hanya jika mereka serupa.
  - True Positive (TP)=a: dua dokumen dalam satu cluster yang sama.
  - True Negative (TN)=d: dua dokumen tidak serupa berada di dua cluster yang berbeda.
  - False Positive (FP)=b: dua dokumen tidak serupa berada di cluster yang sama.
  - False Negative (FN)=c: dua dokumen serupa berada di dua cluster berbeda.

Rand Index

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	a	b
Different classes in ground truth	c	d

$$RI = \frac{A + D}{A + B + C + D}$$

Contoh

- Berdasar slide 38

class\cluster	v1	v2	v3	Jumlah
u1	5	1	2	8
u2	1	4	0	5
u3	0	1	3	4
Jumlah	6	6	5	n=17

$$a = \sum_{i,j} \binom{n_{ij}}{2}$$

$$b = \sum_{i} \binom{n_{i\bullet}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$$

$$c = \sum_{j} \binom{n_{\bullet j}}{2} - \sum_{i,j} \binom{n_{ij}}{2}$$

$$a + b + c + d = \binom{n}{2}$$

Contoh

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	10+6+3=19	28+10+6-19=25
Different classes in ground truth	2*15+10-19=21	136-65=71

$$RI = (19 + 71)/(19 + 25 + 21 + 71) \approx 0.66$$

TERIMA KASIH.