

## KOM341

### Temu Kembali Informasi

---

KULIAH #9

- Text Clustering (Ch.16 & 17)

### Clustering

---

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

JAS - DEPT. ILMU KOMPUTER IPB 2

### Yahoo! Hierarchy

---

www.yahoo.com/Science

JAS - DEPT. ILMU KOMPUTER IPB 3

### Scatter/Gather: Cutting, Karger, and Pedersen

---

JAS - DEPT. ILMU KOMPUTER IPB 4

### Vivisimo (<http://clusty.com/>)

➔ <http://search.yippy.com/>

JAS - DEPT. ILMU KOMPUTER IPB 5

### 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).

JAS - DEPT. ILMU KOMPUTER IPB 6

## 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?

## Apa yang membuat dokumen berhubungan?

- Ideal : kesamaan semantik
- Praktis : kesamaan statistik
  - Menggunakan ukuran kesamaan Cosine
  - Dokumen sebagai vektor
  - Untuk beberapa algoritme, 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

## Partitioning Algorithms

- Metode partisi: susun partisi  $n$  dokumen ke dalam  $K$  kelompok
- Formulasi masalah:
  - Diketahui koleksi dokumen dan nilai  $K$
  - Dapatkan partisi  $K$  kelompok dokumen yang mengoptimumkan partisi dengan kriteria tertentu:
    - Globally optimal: exhaustively enumerate all partitions
    - Effective heuristic methods: K-means and K-medoids algorithms

## K-means

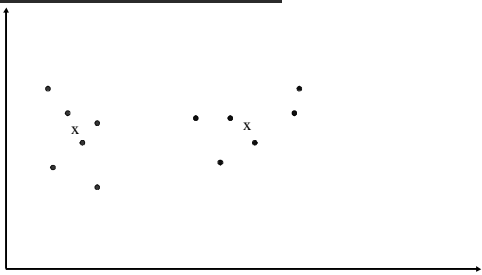
- Asumsikan tiap dokumen sebagai vektor bernilai bilangan riil
- Kelompokkan dokumen berdasarkan centroid pada suatu cluster  $c$  :
 
$$\bar{\mu}(c) = \frac{1}{|c|} \sum_{\vec{x} \in c} \vec{x}$$
- Penempatan elemen pada clusters berdasarkan jarak terhadap centroid dari cluster yang ada (similarities)

## Algoritme K-means

- Pilih  $K$  dokumen secara acak  $\{s_1, s_2, \dots, s_K\}$  sebagai "seed".
- Lakukan iterasi:
  - Untuk setiap dokumen  $d_i$ , masukkan di ke cluster  $c_j$  sehingga jarak( $x_i, s_j$ ) adalah minimum.
  - Perbaiki centroid tiap cluster  $c_j$ 

$$s_j = \mu(c_j)$$

### Contoh K-means (K=2)



JAS - DEPT. ILMU KOMPUTER IPB

13

### Kapan Iterasi Berhenti?

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

JAS - DEPT. ILMU KOMPUTER IPB

14

### Memilih Seed

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

- Contoh
 

A	B	C
○	○	○
○	○	○
D	E	F

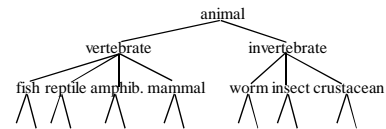
- Jika mulai dengan B dan E sebagai centroid, maka akan konvergen ke {A,B,C} dan {D,E,F}
- Jika mulai dengan D dan F sebagai centroid, maka akan konvergen ke {A,B,D,E} dan {C,F}

JAS - DEPT. ILMU KOMPUTER IPB

15

### Hierarchical Clustering

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

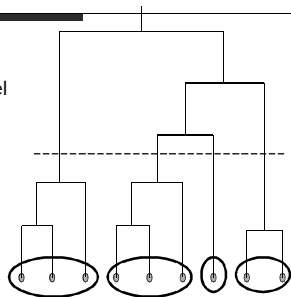


JAS - DEPT. ILMU KOMPUTER IPB

16

### Dendrogram

Cluster diperoleh dengan memotong dendrogram pada level tertentu.



JAS - DEPT. ILMU KOMPUTER IPB

17

### 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

JAS - DEPT. ILMU KOMPUTER IPB

18

## 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

JAS - DEPT. ILMU KOMPUTER IPB

19

## Single Link

- Menggunakan ukuran kesamaan yang terbesar dari tiap pasangan.

$$sim(c_i, c_j) = \max_{x \in c_i, y \in c_j} sim(x, y)$$

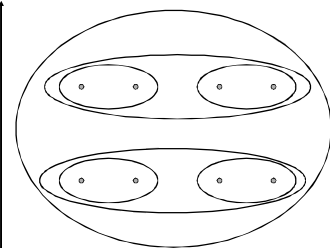
- Setelah  $c_i$  dan  $c_j$  digabung, maka ukuran kesamaan dari cluster yang dihasilkan dengan cluster lainnya,  $c_k$ , adalah:

$$sim((c_i \cup c_j), c_k) = \max(sim(c_i, c_k), sim(c_j, c_k))$$

JAS - DEPT. ILMU KOMPUTER IPB

20

## Single Link



JAS - DEPT. ILMU KOMPUTER IPB

21

## Complete Link

- Menggunakan ukuran kesamaan yang terkecil dari tiap pasangan.

$$sim(c_i, c_j) = \min_{x \in c_i, y \in c_j} sim(x, y)$$

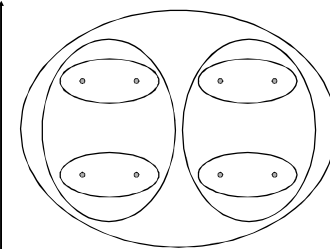
- Setelah  $c_i$  dan  $c_j$  digabung, maka ukuran kesamaan dari cluster yang dihasilkan dengan cluster lainnya,  $c_k$ , adalah:

$$sim((c_i \cup c_j), c_k) = \min(sim(c_i, c_k), sim(c_j, c_k))$$

JAS - DEPT. ILMU KOMPUTER IPB

22

## Complete Link



JAS - DEPT. ILMU KOMPUTER IPB

23

## Average Link

- Menggunakan rata-rata dari pasangan ukuran kesamaan.

$$sim(c_i, c_j) =$$

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

- Merupakan kompromi dari pendekatan single link dan complete link.

JAS - DEPT. ILMU KOMPUTER IPB

24

### Bagaimana Cluster yang Baik ?

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

### Beberapa Ukuran Kesamaan

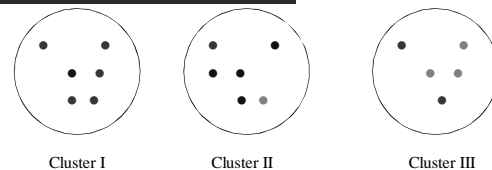
- Inner Product  $|X \cap Y|$
- Dice Coefficient  $\frac{2|X \cap Y|}{|X| + |Y|}$
- Cosine Coefficient  $\frac{|X \cap Y|}{\sqrt{|X|} \cdot \sqrt{|Y|}}$
- Jaccard Coefficient  $\frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$

### 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,  $\pi_1, \pi_2, \dots, \pi_k$  dengan  $n_i$  anggota.
- Purity, rasio antara kelas yang dominan pada cluster  $\pi_i$  dan ukuran cluster  $n_i$

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

### Contoh Purity



Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6  
 Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6  
 Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

### Rand Index (RI)

#anggota	Cluster yang sama	Cluster yang berbeda
Benar Sama	A	C
Benar Berbeda	B	D

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

### Contoh : Single Link

- Dokumen A, B, C, D, dan E mempunyai ukuran kesamaan sebagai berikut

	A	B	C	D	E
A	-	2	7	9	4
B	2	-	10	11	14
C	7	10	-	4	8
D	9	11	4	-	2
E	4	14	8	2	-

**LATIHAN**

---

- Diketahui matrik term-document sbb:

	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

- Dengan ukuran kesamaan Cosine dan Complete Link, lakukan clustering terhadap dokumen.

---

JULIO ADISANTOSO - ILKOM IPB