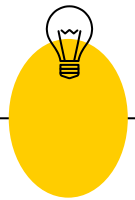
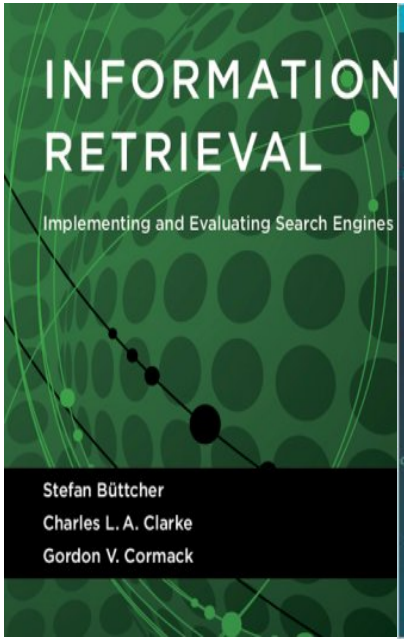
Sistem Temu Kembali Informasi

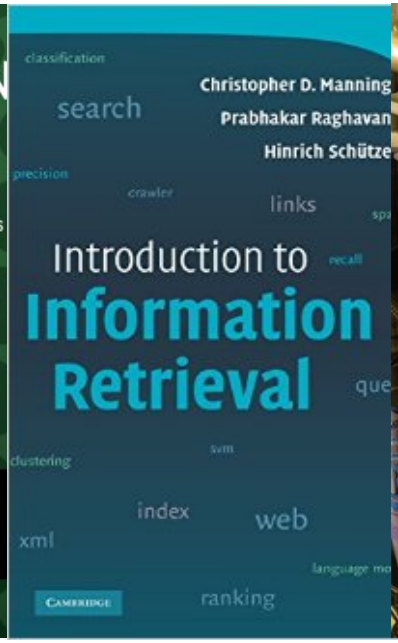
"Klastering Dokumen dengan K-Means"

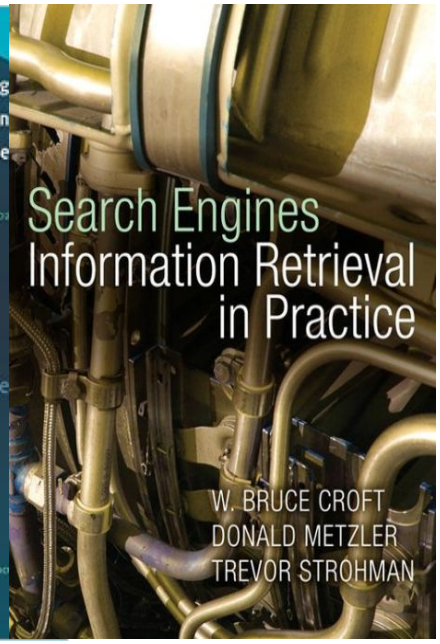


TIM Dosen STKI

Buku Penunjang & Literatur







Course Outline

- Clustering Text
- Algoritma Clustering
- Partitional algorithm
- K means clustering

Clustering

- **Definisi:** Proses pengelompokan text dokumen yang *memiliki kesamaan topik*.
- **Tujuan:** *Mempartisi text dokumen menjadi beberapa kelompok* dimana text dokumen dalam kelompok yang sama adalah mempunyai *kemiripan satu sama lain berdasarkan frekuensi kemunculan term.*
- **Teknik clustering** text dokumen ini merupakan teknik yang lebih spesifik dari pengorganisasian **unsupervised dokumen**, ekstraksi topik otomatis serta pengambilan dan penyaringan informasi secara cepat dari objek data yang **tidak terstruktur**.

Isu Pada Clustering

- Representasi Dokumen
  - Ruang vektor ?
- Ukuran Kesamaan Jarak
- Penentuan titik pusat cluster (centroid)
- Banyaknya kelas
  - Tetap /
  - Tergantung pada data

Apa yang Membuat dokumen Berhubungan

- Ideal : kesamaan semantik
- Praktis : kesamaan statistik
  - Menggunakan ukuran kesamaan Cosine
  - Dokumen sebagai vektor

Algoritma clustering

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

Partitional 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

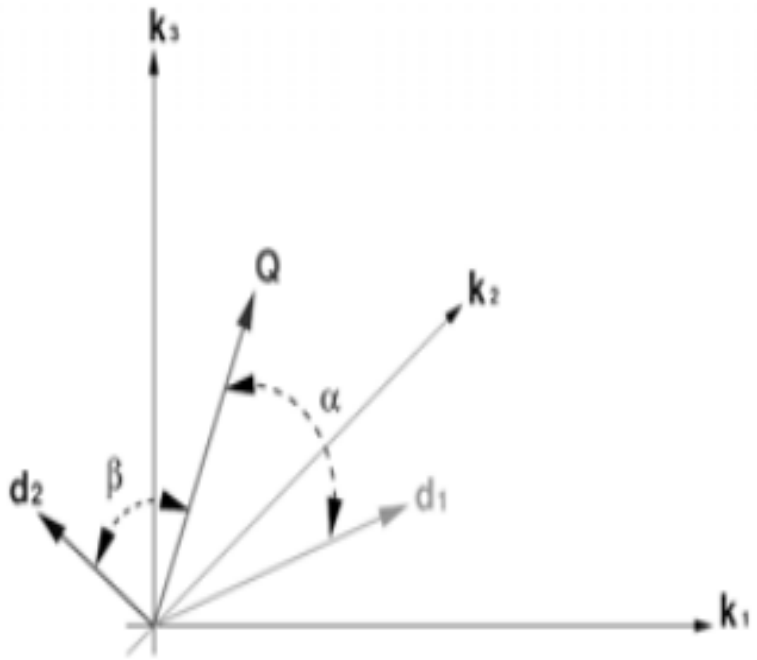
- Prinsip utama dari k-menas adalah menyusun ***k*** buah prototype / pusat massa (*centroid*) / rata-rata (*mean*) dari sekumpulan data berdimensi ***n***.
- Teknik ini mensyaratkan nilai ***k*** sudah diketahui sebelumnya (*a priori*).
- Algoritma k-means **dimulai** dengan **pembentukan prototipe cluster** di awal kemudian secara iteratif prototipe cluster ini diperbaiki hingga konvergen (tidak terjadi perubahan yang signifikan pada prototipe cluster).

Tahapan k-means

1) Inisialisasi titik pusat Cluster.
2) Masukan setiap dokumen ke cluster yang paling cocok berdasarkan ukuran kedekatan dengan centroid. (**Similarity Measurement**).
3) Setelah semua dokumen masuk ke cluster. Hitung ulang centroid cluster berdasarkan dokumen yang berada di dalam cluster tersebut.
4) Jika centroid tidak berubah (dengan treshold tertentu) maka stop. Jika tidak, kembali ke langkah 2.

Similarity Measurement

Ukuran kesamaan antara vektor $\mathbf{d_i}$ dan $\mathbf{q}$.



Teknik similarity measurement

- ***Euclideance Distance.***
- Manhattan Distance.
- Cosine similarity.
- Mahalanobis Distance.

Euclideance Distance

- *rumus*: $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|^2$

- *Dimana d adalah dokumen yang akan diproses, Nilai vektor pada jarak Euclidean dikatakan mirip jika jarak antar dokumen mendekati 0.*

Contoh Kasus K-means

- Ada 4 buah dokumen sebagai berikut:
  - D1: Mata uang rupiah
  - D2: Penyakit mata
  - D3: Pencurian uang adalah penyakit masyarakat
  - D4: Penyakit kekurangan uang
- Kelompokan 4 dokumen tersebut menjadi 2 cluster, ketentuan:
  - Menggunakan teknik k-means clustering
  - Similarity measure dengan Euclideance distance
  - Ditentukan Centroid D1 dan D4
  - Term weighting menggunakan TFIDF
  - Preprocessing: tokenization, stopword removal, stemming

Pembahasan

Dokumen 1	Mata uang rupiah
Dokumen 2	penyakit mata
Dokumen 3	pencurian uang adalah penyakit masyarakat
Dokumen 4	penyakit kekurangan uang

Preprocessing: *tokenization, stopword removal, stemming*

Hasil:

mata
uang
rupiah
sakit
curi
masyarakat
kurang

Term Weighting

	D1	D2	D3	D4	DF	IDF
mata	1	1	0	0	2	0,30103
uang	1	0	1	1	3	0,124939
rupiah	1	0	0	0	1	0,60206
sakit	0	1	1	1	3	0,124939
curi	0	0	1	0	1	0,60206
masyarakat	0	0	1	0	1	0,60206
kurang	0	0	0	1	1	0,60206

TFIDF

	D1	D2	D3	D4
mata	0,30103	0,30103	0	0
uang	0,12494	0	0,124939	0,124939
rupiah	0,60206	0	0	0
sakit	0	0,124939	0,124939	0,124939
curi	0	0	0,60206	0
masyarakat	0	0	0,60206	0
kurang	0	0	0	0,60206

Proses cluster

- D1 sebagai titik pusat kluster 1, dan D4 sebagai titik pusat kluster 2
- *rumus*: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|^2$
- Contoh:
  d(x,y) = (|0,30103-0,30103|^2) + (|0,124939 - 0|^2) + (|0,60206 - 0|^2) + (|0 -0,124939 |^2) + (|0-0|^2) + (|0-0|^2) + (|0-0|^2)
- **d(x,y) = 0,627451679**

	C1 (D1)	C2 (D4)
D1	0	0,911691
D2	0,627451679	0,68462
D3	1,092546313	1,042798
D4	0,911691402	0

Hasil

	C1 (D1)	C2 (D4)
D1	V	-
D2	V	-
D3	-	V
D4	-	V

Cluster 1 : D1, D2	
Cluster 2 : D3,D4	

Studi Kasus 2
Clustering dengan algoritma K-Means

- Ada 4 dokumen sebagai berikut :

Dokumen 1	Vaksin difteri pemerintah
Dokumen 2	Penyakit difteri
Dokumen 3	Pemberian vaksin terhadap penyakit pada anak
Dokumen 4	Penyakit kekurangan vaksin

1. Langkah 1: Hasil yang diharapkan

- Preprocessing: tokenization, stopword removal, stemming
  - difteri
  - vaksin
  - pemerintah
  - sakit
  - beri
  - anak
  - kurang

2. Langkah 2: TERM WEIGHTING

-

	D1	D2	D3	D4	DF	IDF
difteri	1	1	0	0	2	0,30103
vaksin	1	0	1	1	3	0,124939
pemerintah	1	0	0	0	1	0,60206
sakit	0	1	1	1	3	0,124939
beri	0	0	1	0	1	0,60206
anak	0	0	1	0	1	0,60206
kurang	0	0	0	1	1	0,60206

3. Langkah 3: Hitung TFIDF

	D1	D2	D3	D4
difteri	0,30103	0,30103	0	0
vaksin	0,12494	0	0,12494	0,12494
pemerintah	0,60206	0	0	0
sakit	0	0,12494	0,12494	0,12494
beri	0	0	0,60206	0
anak	0	0	0,60206	0
kurang	0	0	0	0,60206

ITERASI 1 PROSES CLUSTER

- D1 sebagai titik pusat kluster 1, dan D4 sebagai titik pusat kluster 2
- rumus: $d(x,y) = \sum_{i=1}^{n} |x_i - y_i|^2$
- Contoh:
- d(x,y) = (|0,30103-0,30103|2) + (|0,124939 - 0|2) + (|0,60206 - 0|2) + (|0 -0,124939 |2) + (|0-0|2) + (|0-0|2) + (|0-0|2)
- d(x,y) = 0, 627451679

JARAK	C1 (D1)	C2 (D4)
D1	0	0,911691
D2	0,627451679	0,68462
D3	1,092546313	1,042798
D4	0,911691402	0

HASIL ITERASI 1

- C1 (D1) C2 (D4)

- D1 V -
- D2 V -
- D3 - V
- D4 - V

- Cluster 1 : D1,D2
- Cluster 2 : D3,D4

5. iterasi 2 HITUNG CLUSTER BARU

	TFIDF D1	TFIDF D2	C1 Baru
difteri	0,30103	0,30103	0,30103
vaksin	0,12494	0	0.06247
pemerintah	0,60206	0	0,30103
sakit	0	0,12494	0.06247
beri	0	0	0
anak	0	0	0
kurang	0	0	0

	TFIDF D3	TFIDF D4	C2 Baru
difteri	0	0	0
vaksin	0,12494	0,12494	0,12494
pemerintah	0	0	0
sakit	0,12494	0,12494	0,12494
beri	0,60206	0	0.301030
anak	0,60206	0	0.301030
kurang	0	0,60206	0.301030

6. Langkah ke 4 – iterasi 2

HITUNG JARAK DOKUMEN KE CLUSTER BARU

JARAK	C1	C2
D1	0.313725839	0.86055921
D2	0.313725839	0.614886917
D3	0.95603108	0.521399247
D4	0.742643383	0.521399247

Hasil Akhir

- C1 C2
- D1 V -
- D2 V -
- D3 - V
- D4 - V
- Cluster 1 : D1,D2
- Cluster 2 : D3,D4

Karena anggota Cluster dari

Iterasi 1 s/d Iterasi 2 tidak berubah atau tetap,

maka sudah pasti

anggota Cluster 1 = Dokumen 1 dan Dokumen 2,

sedangkan

anggota Cluster 2 = Dokumen 3 dan Dokumen 4

Kesimpulan & Review

Partial Clustering yaitu proses pengelompokan text dokumen yang *memiliki kesamaan topik*.

Tujuan: *Mempartisi text dokumen menjadi beberapa kelompok* dimana text dokumen dalam kelompok yang sama adalah mempunyai *kemiripan satu sama lain berdasarkan frekuensi kemunculan term.*

Kuis (Latihan Soal)

D1	PSIS berburu juara Liga Indonesia
D2	Hasil putusan Sidang Elit Politik
D3	Partai politik berebut suara
D4	Manchester United Juara Liga Inggris
D5	Timnas Indonesia juara Liga AFC

- Kelompokan dokumen tersebut menjadi 2 cluster, dengan ketentuan:
  - Menggunakan teknik k-means clustering.
  - Similarity measure dengan Euclideance distance.
  - Ditentukan Centroid D1 dan D3.
  - Term weighting menggunakan TFIDF.
  - Preprocessing: tokenization, stemming.

Thanks!

Any questions ?