TEMU KEMBALI INFORMASI

JULIO ADISANTOSO
Departemen Ilmu Komputer IPB

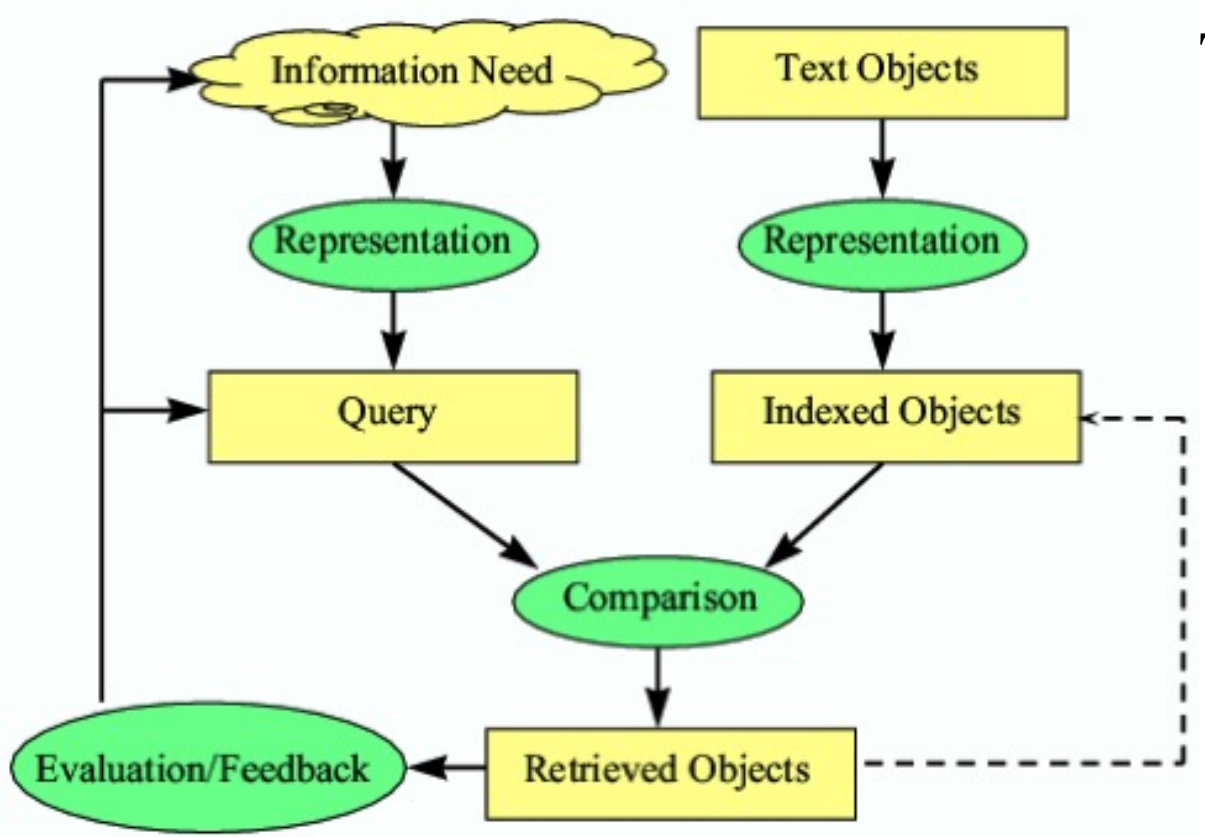
Pertemuan 3

IR MODEL

Mengapa Model?

1. Banyak pengembangan teknologi IR seperti web search, translator system, spam filter, dsb membutuhkan teori dan percobaan.
2. Percobaan menggunakan data empiris dengan berbagai situasi dibutuhkan agar teknologi IR yang dikembangkan dapat sesuai dengan yang diharapkan oleh user
3. Model dapat membantu menjelaskan teori dan hasil percobaan dengan lebih terstruktur dan mudah.

Proses IR



Tiga proses dasar IR:

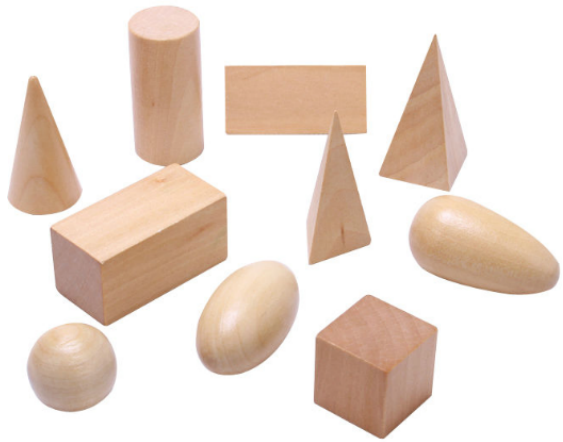
1. representasi isi dokumen,
2. representasi informasi yang dibutuhkan user (disebut query)
3. membandingkan kedua representasi tersebut

Pemodelan IR

- IR terdiri atas 4 komponen yang dinotasikan masing-masing sebagai $[D, Q, F, R(d_j, q)]$
- Keterangan:
  - D adalah kumpulan dokumen (korpus)
  - Q adalah query
  - F adalah representasi isi dokumen dan query
  - $R(d_j, q)$ adalah fungsi yang membandingkan representasi isi dokumen d_j dengan query q
- Bentuk model IR tergantung pada:
  - Bagaimana representasi isi dokumen dan query
  - Bagaimana fungsi $R(d_j, q)$

IR Model

- Exact match atau best match model
  - Boolean Model
  - Region Model
- Statistical model
  - Vector space model
  - Probabilistic model
  - Latent semantic model
- Linguistic and knowledge-based models



Boolean Model

Keuntungan:

- Mudah diimplementasikan dan membutuhkan komputasi yang tidak rumit
- User mudah menyusun query dengan menggunakan operator logika, misalnya OR untuk menyatakan hubungan sinonim, AND untuk frasa
- Query dapat ditulis lebih mudah dipahami (tidak ambigu)

Kekurangan:

- Sulit untuk menyusun query yang kompleks
- Tidak ada pemeringkatan kesesuaian antara dokumen dengan query
- Tidak mengenal pembobotan

Contoh Query Boolean

Contoh query:

hero AND (**angel** OR NOT **man**)

Formulasi query :

$= [k_4 \wedge \{k_2 \vee \neg k_5\}]$

$= [(0\ 1\ 0\ 1) \wedge \{(1\ 0\ 0\ 0\} \vee \neg\ (0\ 0\ 1\ 0)\}]$

$= (0\ 1\ 0\ 1)$

Hasil query (tidak ada urutan):

d_2 dan d_4

	d1	d2	d3	d4
affirm	0	1	0	0
angel	1	0	0	0
conqueror	0	0	0	1
hero	0	1	0	1
man	0	0	1	0
pallid	1	0	0	0
play	0	0	1	0
tragedy	0	0	1	0
unveil	0	1	0	0
uprise	0	1	0	0
wan	1	0	0	0
worm	0	0	0	1

Region Model

- Merupakan pengembangan dari Boolean Model
- Dokumen terkelompok dalam beberapa bagian, biasanya ditandai oleh *tag* dalam format XML
- Menggunakan sedikitnya 2 operator dalam query: `CONTAINING` atau `CONTAINED_BY`
- Contoh mencari semua baris dimana Hamlet berkata ”farewell”:

```
(<LINE> CONTAINING farewell) CONTAINED_BY
(<SPEECH> CONTAINING (<SPEAKER> CONTAINING
Hamlet))
```

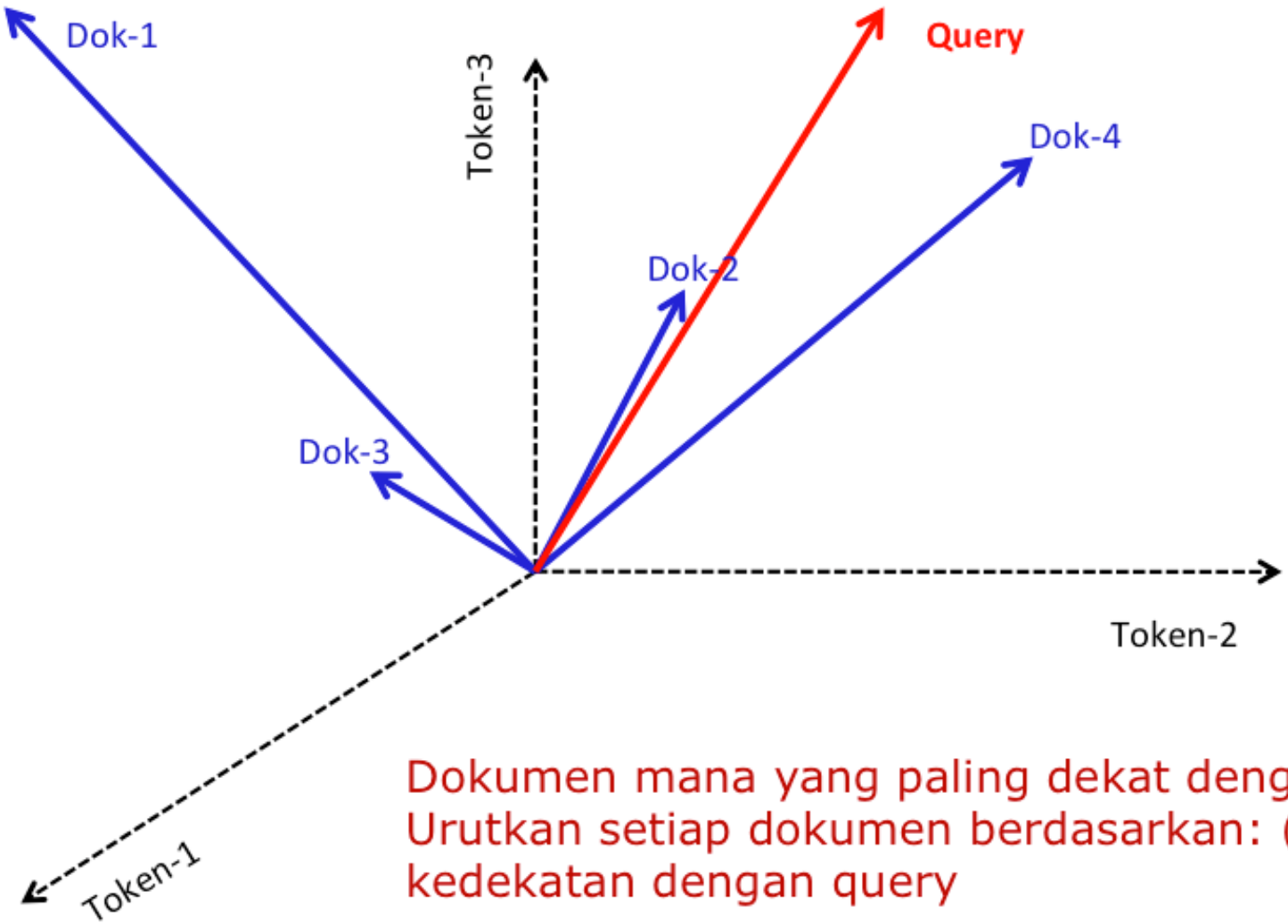
Contoh Dokumen

<ACT>
<TITLE>ACT[103] I[104]</TITLE>
<SCENE>
<TITLE>SCENE[105] I:[106] Elsinore:[107] A[108] platform[109] before[110] the[111] castle:[112]</TITLE>
<STGDIR>FRANCISCO[113] at[114] his[115] post:[116] Enter[117] to[118] him[119] BERNARDO[120]</STGDIR>
<SPEECH>
<SPEAKER>BERNARDO[121]</SPEAKER>
<LINE>Who's[122] there?[123]</LINE>
</SPEECH>
<SPEECH>
<SPEAKER>FRANCISCO[124]</SPEAKER>
<LINE>Nay,[125] answer[126] me:[127] stand;[128] and[129] unfold[130] yourself:[131]</LINE>

Vector Space Model

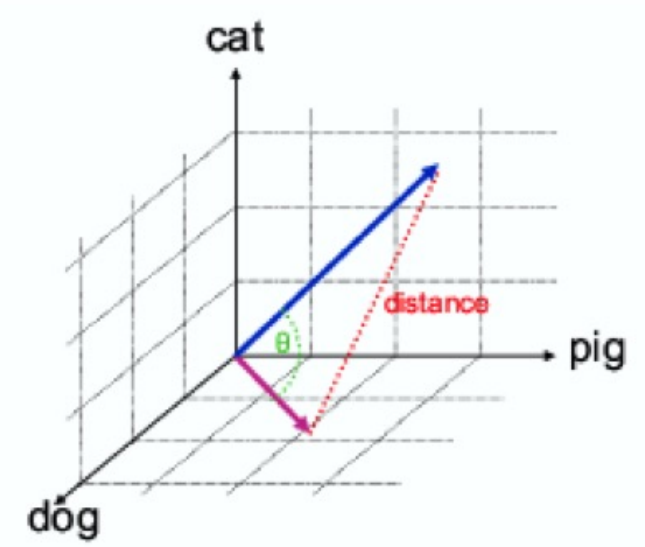
- Dokumen dan query direpresentasikan sebagai vektor dalam ruang berdimensi tinggi
- Memungkinkan partial matching dan pemeringkatan dokumen. Cenderung sebagai best matching
- Dokumen dan query dibandingkan dengan cara membandingkan vektor masing-masing, misalnya menggunakan ukuran jarak antar vektor, atau menggunakan ukuran kemiripan antar vektor.
- Dokumen yang memiliki jarak dekat (atau ukuran kesamaan tinggi) dengan query, dianggap sebagai dokumen yang relevan dengan query

Ukuran Jarak vs Kemiripan



Dokumen mana yang paling dekat dengan query?
Urutkan setiap dokumen berdasarkan: (1) jarak, (2) kedekatan dengan query

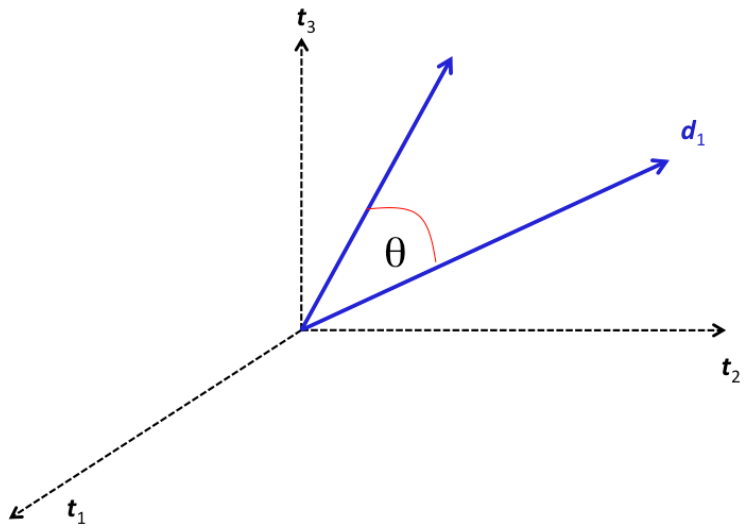
Ukuran Jarak



Ukuran jarak yang sering digunakan adalah Euclidean. Jarak antara vektor dokumen d dengan vektor query q adalah:

$$\delta(d,q) = \sqrt{(d-q)'(d-q)}$$

Ukuran Kemiripan Cosine



Ukuran kemiripan sebagai nilai Cosinus dari sudut θ. Ukuran kemiripan Cosine antara vektor dokumen d dengan vektor query q adalah:

$$sim(d,q) = \frac{d'q}{\sqrt{d'd}\sqrt{q'q}}$$

Urutkan Dokumen Berdasarkan Query!

Query	Petani mengalami gagal panen.
D1	Gagal panen banyak yang terjadi.
D2	Panen raya banyak dilaksanakan.
D3	Jalan raya sering terjadi kecelakaan.
D4	Petani gagal tanam karena mengalami panen yang gagal.

Kata	tf					N/n	idf	Bobot (w)				
	Q	D1	D2	D3	D4			Q	D1	D2	D3	D4
banyak	0	1	1	0	0	4/2=2	0.301	0	0.301	0.301	0	0
dilaksanakan	0	0	1	0	0	4/1=4	0.602	0	0	0.602	0	0
gagal	1	1	0	0	2	4/2=2	0.301	0.301	0.301	0	0	0.602
jalan	0	0	0	1	0	4/1=4	0.602	0	0	0	0.602	0
karena	0	0	0	0	1	4/1=4	0.602	0	0	0	0	0.602
kecelakaan	0	0	0	1	0	4/1=4	0.602	0	0	0	0.602	0
mengalami	1	0	0	0	1	4/1=4	0.602	0.602	0	0	0	0.602
panen	1	1	1	0	1	4/3=1.3	0.125	0.125	0.125	0.125	0.000	0.125
petani	1	0	0	0	1	4/1=4	0.602	0.602	0	0	0	0.602
raya	0	0	1	1	0	4/2=2	0.301	0	0	0.301	0.301	0
sering	0	0	0	1	0	4/1=4	0.602	0	0	0	0.602	0
tanam	0	0	0	0	1	4/1=4	0.602	0	0	0	0	0.602
terjadi	0	1	0	1	0	4/2=2	0.301	0	0.301	0	0.301	0
yang	0	1	0	0	1	4/2=2	0.301	0	0.301	0	0	0.301

Model Lainnya ...

Akan dibahas pada pertemuan selanjutnya ...

TUGAS/PR (sebagai materi diskusi kelas minggu depan):

- Pelajari Extended Boolean
- Kerjakan soal pada Manning *et al* (2008) nomor 2.9, 6.8, 6.9, 6.10, 6.11, 6.19