KULIAH 7

MODEL PROBABILITAS

2 November 2006

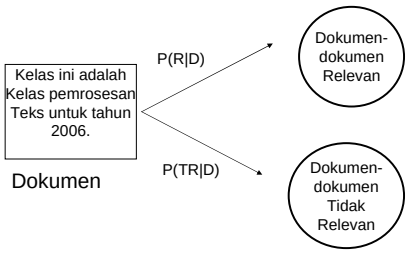
Bab 2: Grossman

Bab 7: Baeza-Yates & Ribeiro-Neto

Model Probabilitas

- Model probabilitas digunakan untuk memperoleh suatu set jawaban bagi suatu query.
- Bila ada suatu query, maka ada suatu set jawaban yang *ideal*
- Deskripsi dari suatu set jawaban ideal dimodelkan dalam bentuk probabilitas
  - Properti dari set jawaban ideal itu apa saja?

Model Probabilitas Dasar



Model Probabilitas Dasar

- Asumsikan relevansi suatu dokumen adalah independen dari dokumen-dokumen lain di dalam koleksi.
- Urutkan dokumen-dokumen sesuai nilai probabilitas relevansinya.

Prinsip Urutan Probabilitas

- Jika ada suatu query *q* dan suatu dokumen *dj*, model probabilitas mencoba untuk memperkirakan probabilitas bahwa pemakai akan menemukan dokumen *dj* relevan.
- Set jawaban yang ideal disebut sebagai *R* dan mestinya menunjukkan nilai probabiltas relevan yang maksimal. Dokumen-dokumen pada set *R* diramalkan sebagai relevan.
- Bagaimana cara menghitung probabilitas?

Model Probabilitas

- Tahapan penghitungan probabilitas dari suatu dokumen
  - Peroleh suatu set dokumen awal
  - Pemakai meneliti apakah dok ini adalah relevan (biasanya hanya 10-20 dok teratas)
  - Sistem IR menggunakan informasi ini untuk memperbaiki deskripsi dari set jawaban ideal
  - Dengan mengulangi proses ini, diharapkan deskripsi dari set jawaban ideal akan meningkat
- Pada awalnya perlu untuk menebak deskripsi dari suatu set jawaban ideal.

Beberapa Dasar Probabilitas

Bayes' Rule:

P(B|A) = P(A|B) * P(B) / P(A)

$$p(R \mid x) = \frac{p(x \mid R)\, p(R)}{p(x)}$$

$$p(NR \mid x) = \frac{p(x \mid NR)\, p(NR)}{p(x)}$$

Urutan

- Urutan dengan Probabilitas dihitung dengan:
  - *sim(q,dj) = P(dj relevant-to q) / P(dj non-relevant-to q)*
  - Ini adalah kemungkinan bahwa dokumen dj adalah relevan
  - Dengan menggunakan kemungkinan akan memperkecil probabilitas dari suatu penilaian yang salah
- Definisi:
  - *P(R | vec(dj))* : probabilitas bahwa suatu dok relevan
  - P(¬R | vec(dj)) : probabilitas dok itu tidak relevan

Urutan

- $sim(d_j,q) = P(R \mid vec(dj)) / P(\neg R \mid vec(dj))$
  $= \frac{[P(vec(d_j) \mid R) * P(R)]}{[P(vec(d_j) \mid \neg R) * P(\neg R)]}$

  $\sim \frac{P(vec(d_j) \mid R)}{P(vec(d_j) \mid \neg R)}$

- $P(vec(d_j) \mid R)$: probabilitas dari pemilihan dokumen secara acak d_j dari set dokumen yang relevan R

Urutan

- $sim(dj,q) \sim \frac{P(vec(dj) \mid R)}{P(vec(dj) \mid \neg R)}$

  $\sim \frac{[\Pi\ P(k_i \mid R)]}{[\Pi\ P(k_i \mid \neg R)]}$

- $P(k_i \mid R)$: probabilitas kata indeks k_i ada di suatu dokumen yang dipilih secara acak dari set dokumen yang relevan R.

Urutan Awal

- Probabilitas $P(k_i \mid R)$ dan $P(k_i \mid \neg R)$?
- Perkirakan berdasarkan asumsi:
  - $P(k_i \mid R) = 0.5$
  - $P(k_i \mid \neg R) = \frac{n_i}{N}$

    dimana n_i adalah jumlah dok yang berisi k_i
  - Gunakan perkiraan ini untuk memperoleh urutan awal
  - Lakukan perbaikan pada urutan awal

Memperbaiki Urutan Awal

- Asumsikan
  - V : set dari dok yang diperoleh pada awal proses
  - V_i : subset dari dok yang diperoleh yg berisi k_i
- Evaluasi kembali perkiraan:
  - $P(k_i \mid R) = \frac{V_i}{V}$
  - $P(k_i \mid \neg R) = \frac{n_i - V_i}{N - V}$
- Ulangi secara rekursif.

Contoh Model Probabilitas

- Nilai probabilitas kata pada dokumen dj

Kata	P(k|R)	P(k|¬R)
k1	0.8	0.4
k2	0.6	0.1
k3	0.2	0.9
k4	0.9	0.6

P(R) = 0.1 ; P(¬R) = 0.9

Sim(dj,q) = $\frac{0.8 * 0.6 * 0.2 * 0.9 * 0.1}{0.4 * 0.1 * 0.9 * 0.6 * 0.9} = \frac{0.00784}{0.01944} = 0.403$