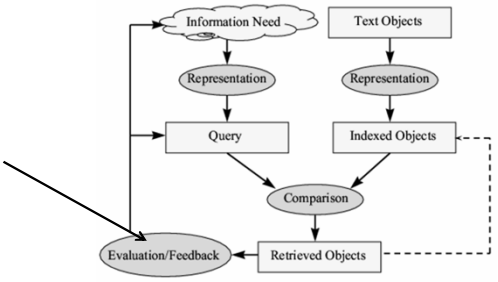
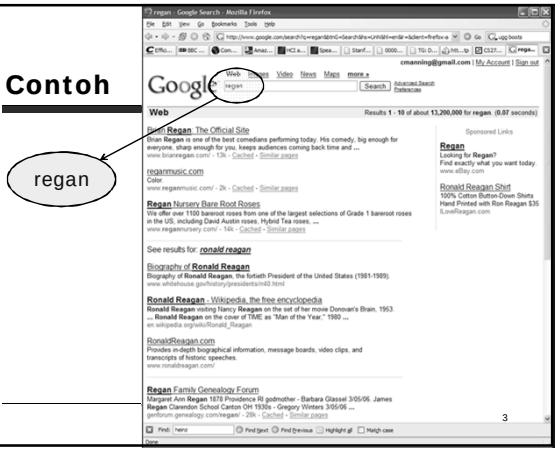
KOM341
Temu Kembali Informasi

KULIAH #6
- Relevance feedback
- Query expansion

Proses Temu-Kembali



Contoh

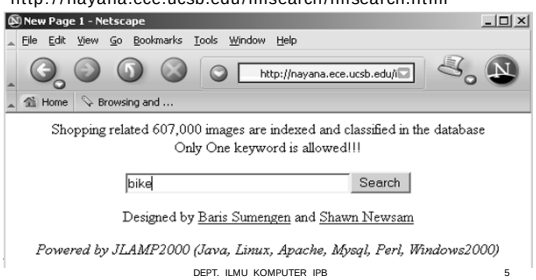


Relevance Feedback

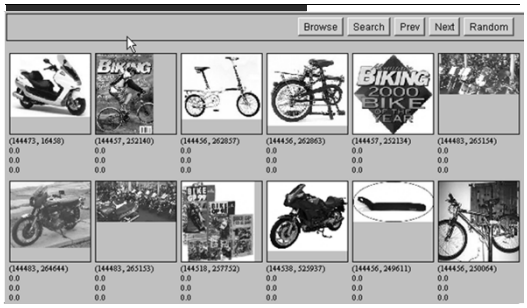
- Relevance feedback: user memberi feedback pada dokumen hasil yang dianggap relevan
  - User memberikan query pendek dan sederhana
  - User memberi tanda pada dokumen yang dihasilkan sebagai relevan dan tidak relevan.
  - IRs menghitung dan memperbaiki query berdasarkan feedback dari user tadi.
  - Dilakukan berulang sesuai dengan banyaknya iterasi yang diinginkan.
- Ide: sulit memformulasikan query yang baik ketika tidak tahu tentang koleksi yang ada.

Contoh

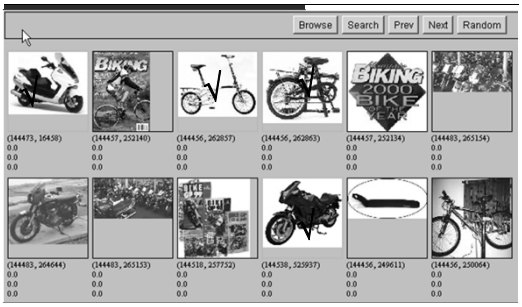
Image search engine
http://nayana.ece.ucsb.edu/imsearch/imsearch.html



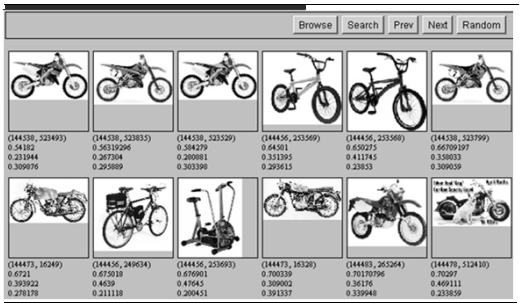
Hasil Query Awal



Relevance Feedback



Hasil Setelah RF



Relevance Feedback

- Kita dapat mengubah query berdasarkan pada relevance feedback dan menerapkan vector space model.
- Gunakan hanya dokumen yang ditandai.
- Relevance feedback dapat meningkatkan recall dan precision

Reformulasi Query

- Berdasarkan feedback dari user
- Berdasarkan informasi yang diperoleh dari sekumpulan dokumen awal yang diperoleh
- Berdasarkan pada informasi global dari koleksi dokumen

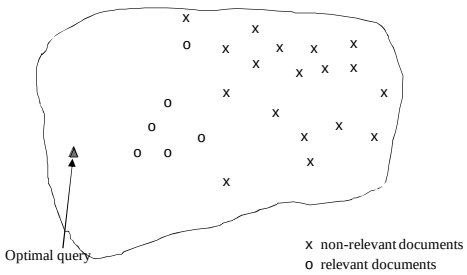
Rocchio Algorithm

- Implementasi RF berdasarkan vector space model.
- Memaksimumkan *sim* (Q, C_r) - *sim* (Q, C_{nr})
- Optimal query:

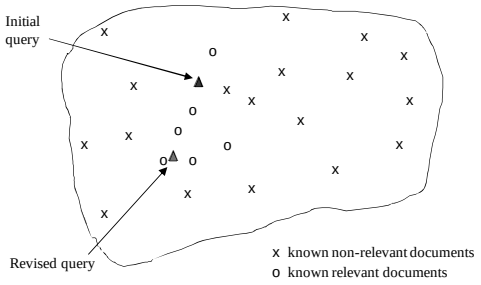
$$\vec{Q}_{opt} = \frac{1}{|C_r|}\sum_{\vec{d}_j \in C_r}\vec{d}_j - \frac{1}{N-|C_r|}\sum_{\vec{d}_j \notin C_r}\vec{d}_j$$

- Q_{opt} = optimal query; C_r = dok. relevan; N = ukuran koleksi
- Tidak realistik: kita tidak tahu dok. Yang relevan.

Best Query



Relevance Feedback



Rocchio 1971 Algorithm

- Praktis menggunakan:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

- q_m = query yang dimodifikasi; q_0 = query awal; α,β,γ: bobot yang dipilih; D_r = *vektor dok relevan yg diketahui*; D_{nr} = vektor tdk relevan yg diketahui
- Query baru mendekati dokumen relevan, dan menjauhi dokumen yang tidak relevan
- Bobot istilah dapat menjadi negatif
  - Bobot istilah yang negatif dihilangkan (dibuat 0)

Contoh

- Misal diketahui:

	tani	gagal	panen	hama	banjir
$\vec{d}_1(R)$	1	10	19	0	2
$\vec{d}_2(TR)$	4	0	12	8	20
$\vec{d}_3(R)$	7	4	1	3	8
$\vec{d}_4(R)$	9	5	2	1	2
$\vec{q}$	0	0	5	10	2

- Misalkan : α=1, β=¾, γ=¼

Contoh

- $\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$
- $\vec{q}'$ = (0 0 5 10 2) + ¾ (1/3) [(1 10 19 0 2) + (7 4 1 3 8) + (9 5 2 1 2)] – ¼ (4 0 12 8 20)
  = (0 0 5 10 2) + (4¼ 4¾ 5½ 1 3) – (1 0 3 2 5)
  = (3¼ 4¾ 7½ 9 0)
- Similarity (dot product)
  - $sim(d_1, q) = 99$ $sim(d_1, q') = 193¼$ naik
  - $sim(d_2, q) = 180$ $sim(d_2, q') = 175$ turun
  - $sim(d_3, q) = 51$ $sim(d_3, q') = 76¼$ naik
  - $sim(d_4, q) = 24$ $sim(d_4, q') = 77$ naik

Evaluasi RF

- Gunakan q_0 dan hitung grafik P/R
- Gunakan q_m dan hitung grafik P/R
- Bandingkan.

Pseudo Relevance Feedback

- Blind relevance feedback
- Metode untuk analisis lokal secara otomatis:
  - Menggunakan metode relevance feedback tanpa input eksplisit dari user.
  - Pseudo Relevance Feedback
  - Hanya asumsikan dokumen yang diperoleh pada top n adalah relevan, dan gunakan untuk membentuk query yang baru.
  - Query expansion diperbolehkan berisi kata-kata yang berkaitan dengan kata-kata pada query.

Pseudo Relevance Feedback

- Ambil top n dokumen
- Dari semua kata-kata pada dokumen tsb., ambil top t kata
- Urutan kata-kata menunjukkan cara kata-kata tersebut diurutkan:
  - n (banyaknya dokumen yang berisi kata t)
  - f (jumlah kemunculan kata t)
  - n * idf
  - f * idf

Pseudo Relevance Feedback

- Contoh: Top 3 dokumen:
  - D1 : A, B, B, C, D
  - D2 : C, D, E, E, A, A
  - D3 : A, A, A
  - Asumsikan idf dari A=1, B=1, C = 1, D=2, E = 2
- Rank:

kata	n	f	n * idf	f * idf
A	3	6	3	6
B	1	2	1	2
C	2	2	2	2
D	2	2	4	4
E	1	2	4	8

Query Expansion

Query Expansion

- Banyak kaitan dengan RF:
  - QE merupakan suatu teknik umum untuk memperbaiki query sehingga dapat memperoleh hasil yang lebih baik.
  - Idenya adalah mengubah query sehingga lebih dekat ke dokumen yang relevan.
  - Cara mengubahnya : menambah, membuang, atau mengubah bobot kata pada query.
- RF vs QE
  - Pada RF, user memberikan input tambahan (relevant/tidak-relevant) pada dokumen, yang digunakan untuk membobot kembali kata-kata pada dokumen
  - Pada QE, user memberikan tambahan input (kata yg baik/tidak baik) pada kata atau frase.

Metode Reformulasi Query

- Global methods
  - QE menggunakan thesaurus atau WordNet
  - QE melalui thesaurus otomatis
  - Teknik mirip koreksi ejaan
- Local/basic methods
  - Relevance feedback
  - Pseudo relevance feedback
  - Indirect relevance feedback

Thesaurus

- Suatu thesaurus memberikan informasi tentang synonym dan kata-kata serta frase yang secara semantik berkaitan.
- Misal (http://thesaurus.reference.com):
  - **market**
  - Part of Speech: verb
  - Definition: package and sell goods
  - Synonyms: advertise, barter, display, exchange, merchandise, offer for sale, retail, vend, wholesale
  - Antonyms: buy

Ekspansi Query dgn Thesaurus

- Tidak memerlukan input dari user
- Untuk setiap kata ***t*** pada suatu query, ekspansi query dengan sinonim dan kata lain ***t*** dari thesaurus.
- Bobot kata-kata tambahan dapat lebih kecil daripada kata-kata pada query awal.
- Biasanya meningkatkan recall.
- Banyak digunakan pada bidang ilmu pengetahuan / teknik

Wordnet

- http://www.cogsci.princeton.edu/~wn/
- Suatu database yang detil berisi hubungan semantik antara kata- kata dalam bahasa Inggris.
- Kira- kira berisi 144,000 kata dalam bahasa Inggris.
- Kata benda, sifat, kerja, dan keterangan dikelompokkan menjadi 109,000 set sinonim yang disebut synsets.

Hubungan Pada WordNet Synset

- Antonym: front → back
- Attribute: benevolence → good (noun to adjective)
- Pertainym: alphabetical → alphabet (adjective to noun)
- Similar: unquestioning → absolute
- Cause: kill → die
- Holonym: chapter → text (part-of)
- Meronym: computer → cpu (whole-of)
- Hyponym: tree → plant (specialization)
- Hypernym: fruit → apple (generalization)

QE menggunakan WordNet

- Tambahkan sinonim pada synset yang sama.
- Tambahkan hiponim untuk memasukkan kata-kata khusus.
- Tambahkan hipernim untuk membuat query lebih umum.
- Tambahkan kata-kata lain yang berkaitan untuk memperluas query.

QE menggunakan WordNet

- Contoh query awal :
  information system
- WordNet (synonym):
  - information : message, content, subject matter, substance
  - system : group, grouping
- Query expansion:
  information message system group

Tipe Ekspansi Query

- Global Analysis: (statis; dari semua dokumen dalam koleksi)
  - Controlled vocabulary
  - Manual thesaurus
  - Automatically derived thesaurus (kemunculan secara statistik)
  - Based on query log mining (umum di web)
- Local Analysis: (dynamic)
  - Analisis dokumen yang terambil

Controlled Vocabulary

Automatic Thesaurus Generation

- Membuat thesaurus secara otomatis dengan menganalisis dokumen dalam koleksi
- Dua pendekatan utama:
  - Berdasarkan kemunculan kata
  - Berdasarkan hubungan gramatikal
- Kemunculan kata lebih robust, sedangkan hubungan gramatikal lebih akurat.

Co-occurrence Thesaurus

- Cara paling sederhana adalah menghitung kesamaan antar kata (term-term similarities) in C = AA^T dimana A adalah matrik term-document.
- $w_{i,j}$ = (normalized) weighted count (t_i , $\mathbf{d}_j$)