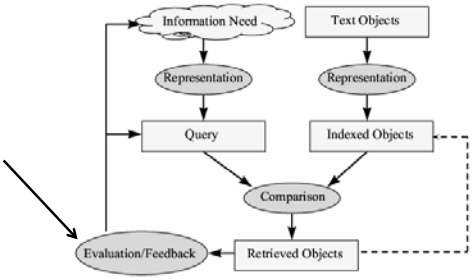
KOM341
Temu Kembali Informasi

KULIAH #5
- Evaluasi IR

Proses Temu-Kembali



Contoh query : sby

Contoh query: flu burung

Evaluasi IR

Indikator yang dapat diukur:
- ☐ Seberapa cepat dia meng-indeks
  - ■ Banyaknya dokumen/jam
  - ■ Terkait dengan ukuran dokumen
- ☐ Seberapa cepat dia mencari
  - ■ Fungsi dari ukuran indeks

Indikator yang sulit diukur:
- ☐ Seberapa "bahagia" user ??
  - ■ Kesesuaian jawaban hasil query
  - ■ Ketepatan
- ☐ Seberapa "mudah" user ??
  - ■ Menyusun query

Versi user

- ☐ Sangat tergantung setting.
- ☐ Web mesin pencari:
  - ■ User mendapatkan apa yg diinginkan dan kembali melanjutkan pencarian
- ☐ Situs eCommerce:
  - ■ User mendapatkan apa yang diinginkan dan membuat pemesanan
- ☐ Umumnya adalah mengukur kepuasan user.

Kepuasan user

- Mengukur relevansi hasil query.
- Masalah: bagaimana mengukur relevansi?
- Tiga elemen:
  - Koleksi dokumen
  - Kumpulan query
  - Pasangan query-dokumen relevan dan tidak relevan (relevance judgments)

Evaluasi IR

- Kebutuhan informasi diterjemahkan ke dalam bentuk query.
- Relevansi lebih kepada kesesuaian dengan kebutuhan informasi, bukan pada query.
- Contoh kebutuhan informasi: Saya mencari informasi tentang jenis kacang yang dapat meningkatkan kolesterol dan menimbulkan resiko pada jantung.
- Query: ***kacang kolesterol resiko jantung***

Standard test collections

- The Cranfield collection. Collected in the United Kingdom starting in the late 1950s, it contains 1398 abstracts of aerodynamics journal articles, a set of 225 queries, and exhaustive relevance judgments of all (query, document) pairs.
- Text Retrieval Conference (TREC) by The U.S. National Institute of Standards and Technology (NIST). these test collections comprise 6 CDs containing 1.89 million documents and relevance judgments for 450 information needs.

Standard test collections

- NII Test Collections for IR Systems (NTCIR). Similar sizes to the TREC collections, focusing on East Asian language and cross-language information retrieval, wherequeries are made in one language over a document collection containing documents in one or more other languages. http://research.nii.ac.jp/ntcir/data/data-en.html
- Cross Language Evaluation Forum (CLEF). This evaluation series has concentrated on European languages and cross-language information retrieval. http://www.clef-campaign.org/

Standard test collections

- Reuters-21578 and Reuters-RCV1. Reuters-21578 collection: 21578 newswire articles. RCV1 : Reuters Corpus Volume 1 (RCV1), consisting of 806,791 documents.
- 20 Newsgroups. This is another widely used text classification collection, collected by Ken Lang. It consists of 1000 articles from each of 20 Usenet newsgroups, it contains 18941 articles.

Evaluation of retrieval sets

- Precision:
  - rasio dokumen yang di-retrieve adalah relevan
  - P(relevant|retrieved)

$$\text{Precision} = \frac{\#(\text{relevant items retrieved})}{\#(\text{retrieved items})}$$

- Recall:
  - rasio dokumen relevan yang di-retrieve
  - P(retrieved|relevant)

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}$$

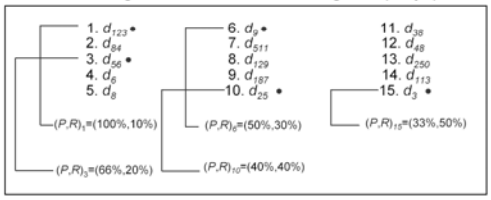
Evaluation of retrieval sets

	Relevant	Not Relevant
Retrieved	tp	fp
Not Retrieved	fn	tn

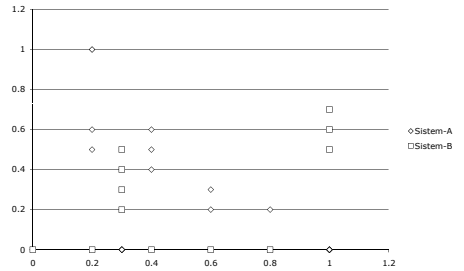
- Precision = P = tp/(tp + fp)
- Recall = R = tp/(tp + fn)
- Accuracy = (tp + tn)/(tp + fp + fn + tn)

Contoh

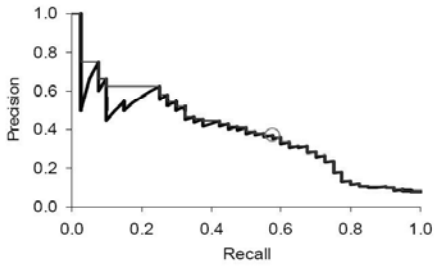
- $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
  - Ten relevant documents
- A ranking of the documents for the given query q



Membandingkan 2 sistem IR



Kurva R-P

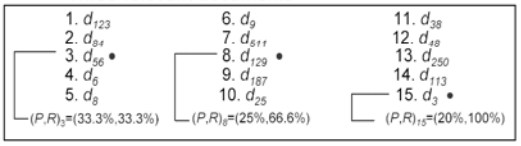


Interpolasi

- Nilai R-P tiap query berbeda sehingga sulit membandingkan antar metode.
- Perlu dilakukan interpolasi.
- Cara interpolasi:
  - Menghubungkan titik
  - Menghubungkan titik maksimum
  - Menghubungkan titik minimum
  - Menghubungkan titik rata-rata

Contoh

- $R_q=\{d_3, d_{56}, d_{129}\}$
  - Three relevant documents



Interpolasi

- Interpolated Precisions at standard recall levels

$$\overline{P}(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

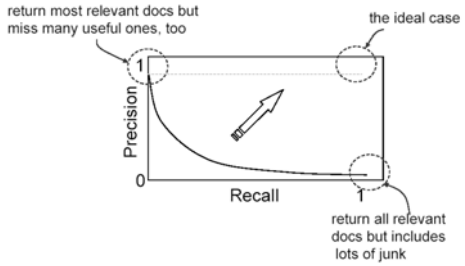
  - the *j*-th standard recall level (e.g., r_5 is recall level 50%)
- Example 3.3 (cont.)

$(P,R)_3$=(33.3%,33.3%)
$(P,R)_6$=(25%,66.6%)
$(P,R)_{15}$=(20%,100%)

$$\overline{P}_i(r_j) = \max_{r_j \le r \le r_{j+1}} P_i(r)$$

Precision	Recall
33.3%	0%
33.3%	10%
33.3%	20%
33.3%	30%
25%	40%
25%	50%
25%	60%
20%	70%
20%	80%
20%	90%
20%	100%

Trade-off



Precision Rata-rata

- Ada k buah query $\{q_1, q_2, \ldots, q_k\}$
- Untuk query tertentu, hitung titik P/R untuk setiap dokumen yang relevan, pada titip recall standar.
- Hitung rata-rata Precision setiap query pada setiap titik recall yang standar.

$$P(r) = \frac{1}{k}\sum_{j=1}^{k} P_j(r)$$

Contoh

Query	q1	q2	q3
Dokumen relevan	d1, d8, d10, d120, d15	d8, d9, d25, d40, d78, d85, d88, d100	d7, d10, d12, d20
Hasil query (sesuai ranking)	d10, d50, d8, d19, d100, d30, d15, d80, d92, d65	d100, d90, d32, d65, d78, d25, d88, d95, d62, d120	d10, d15, d90, d7, d95, d12, d120, d30, d20, d100

MAP: Mean average precision

- Nilai Precision rata-rata diperoleh dari top k dokumen, setiap kali suatu dokumen yang relevan diperoleh
- Menghindari interpolasi, tidak menggunakan titik recall yang tetap.
- MAP untuk koleksi query adalah rata-ratanya.

$$\mathrm{MAP}(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j}\mathrm{Precision}(R_{jk})$$

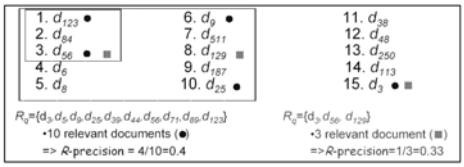
MAP: Contoh

1. d_{123} • (*P*=1.0)
2. d_{84}
3. d_{56} • (*P*=0.66)
4. d_6
5. d_8
6. d_9 • (*P*=0.5)
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} • (*P*=0.4)
11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 • (*P*=0.3)

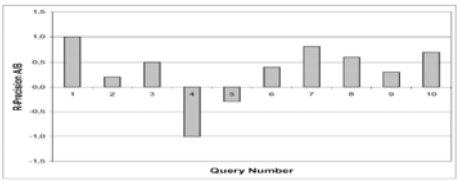
(1.0+0.66+0.5+0.4+0.3)/5=0.57

R-Precision

– Generate a single value summary of ranking by computing the precision at the *R*-th position in the ranking
  - Where *R* is the total number of relevant docs for the current query



Precision histograms



- A positive $RP_{A/B}(i)$ indicates that the algorithm *A* is better than *B* for the *i*-th query and vice versa

F-measure

- The weighted harmonic mean of precision and recall:

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- Balanced F-measure : bobot P sama dengan bobot R, artinya α=1/2 atau β = 1, sehingga

$$F_{\beta=1} = \frac{2PR}{P+R}$$

Assessing relevance

- Menentukan relevansi dokumen terhadap suatu query menggunakan pooling dari beberapa ahli.
- Menilai hasil menggunakan statistik Kappa:

$$kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

dimana P(A):proporsi banyaknya penilai yang setuju, P(E): persetujuan yang merupakan kebetulan

Contoh

		Judge 2 Relevance: Yes	No	Total
Judge 1 Relevance	Yes	300	20	320
	No	10	70	80
	Total	310	90	400

- P(A) = (300+ 70)/400 = 370/400 = 0.925
- Pooled marginals
  - P(nr) = (80+90)/(400+400) = 170/800 = 0.2125
  - P(r) = (320+ 310)/(400+ 400) = 630/800 = 0.7878
- $P(E) = P(nr)^2 + P(r)^2 = 0.2125^2 + 0.7878^2 = 0.665$
- Statistik Kappa: (0.925−0.665)/(1− 0.665) = 0.776
- Kesimpulan nilai Kappa:
  - Kappa ≥ 0.8 = persetujuannya baik
  - 0.67 ≤ Kappa < 0.8 = persetujuan yang fair
  - Kappa < 0.67 = ditolak