Evaluasi Sistem Information Retrieval

Memahami metrik evaluasi untuk mengukur efektivitas sistem pencarian informasi

Tujuan Pembelajaran

Memahami Metrik Dasar

Menguasai konsep Precision dan Recall sebagai fondasi evaluasi IR

Menerapkan Metrik Lanjutan

Menggunakan F1-Score, MAP, dan NDCG untuk evaluasi komprehensif

Evaluasi Sistem IR

Mampu menganalisis dan membandingkan performa berbagai sistem IR

Pada akhir pembelajaran ini, mahasiswa akan memiliki kemampuan untuk mengevaluasi sistem Information Retrieval menggunakan berbagai metrik standar industri. Kemampuan ini sangat penting dalam pengembangan dan optimasi mesin pencari modern.

Mengapa Evaluasi IR Penting?

Sistem Information Retrieval yang efektif harus dapat mengembalikan dokumen yang relevan dengan query pengguna. Tanpa evaluasi yang sistematis, kita tidak dapat:

- Mengukur seberapa baik sistem memenuhi kebutuhan pengguna
- Membandingkan performa berbagai algoritma pencarian
- Mengidentifikasi area yang perlu diperbaiki
- Memvalidasi perubahan atau optimasi sistem

Metrik evaluasi memberikan cara objektif dan terukur untuk menilai kualitas hasil pencarian, memungkinkan pengembangan sistem IR yang lebih baik dan responsif terhadap kebutuhan pengguna.

Confusion Matrix dalam IR

Sebelum memahami metrik evaluasi, kita perlu mengenal konsep dasar klasifikasi dokumen dalam sistem IR:

True Positive (TP)

Dokumen relevan yang berhasil dikembalikan oleh sistem

False Positive (FP)

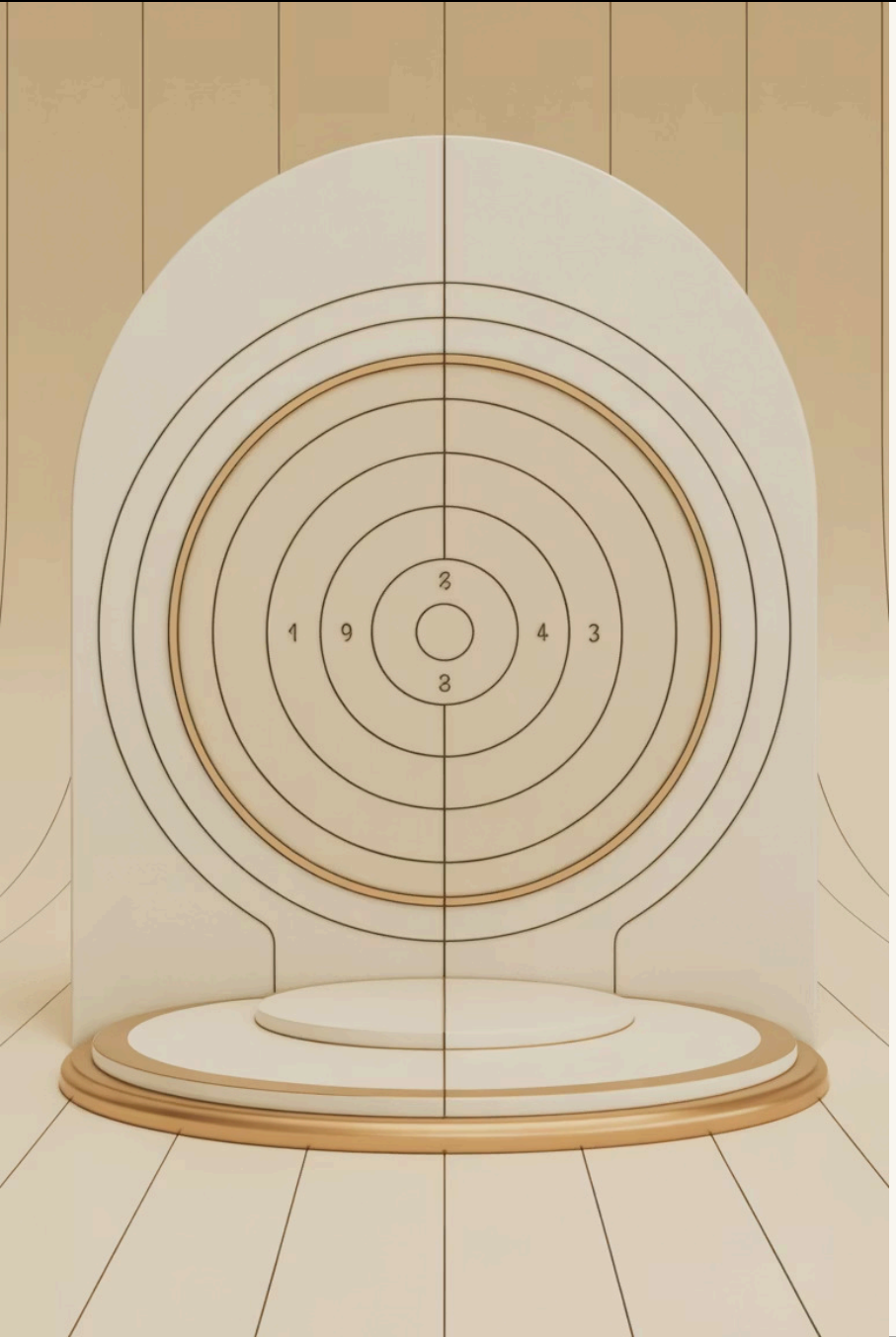
Dokumen tidak relevan yang dikembalikan sistem

True Negative (TN)

Dokumen tidak relevan yang tidak dikembalikan

False Negative (FN)

Dokumen relevan yang terlewatkan sistem



Precision: Mengukur Ketepatan

Definisi

Precision mengukur proporsi dokumen relevan dari semua dokumen yang dikembalikan sistem.

$$Precision = \frac{TP}{TP + FP}$$

Precision menjawab pertanyaan: "Dari semua dokumen yang dikembalikan, berapa banyak yang benar-benar relevan?"

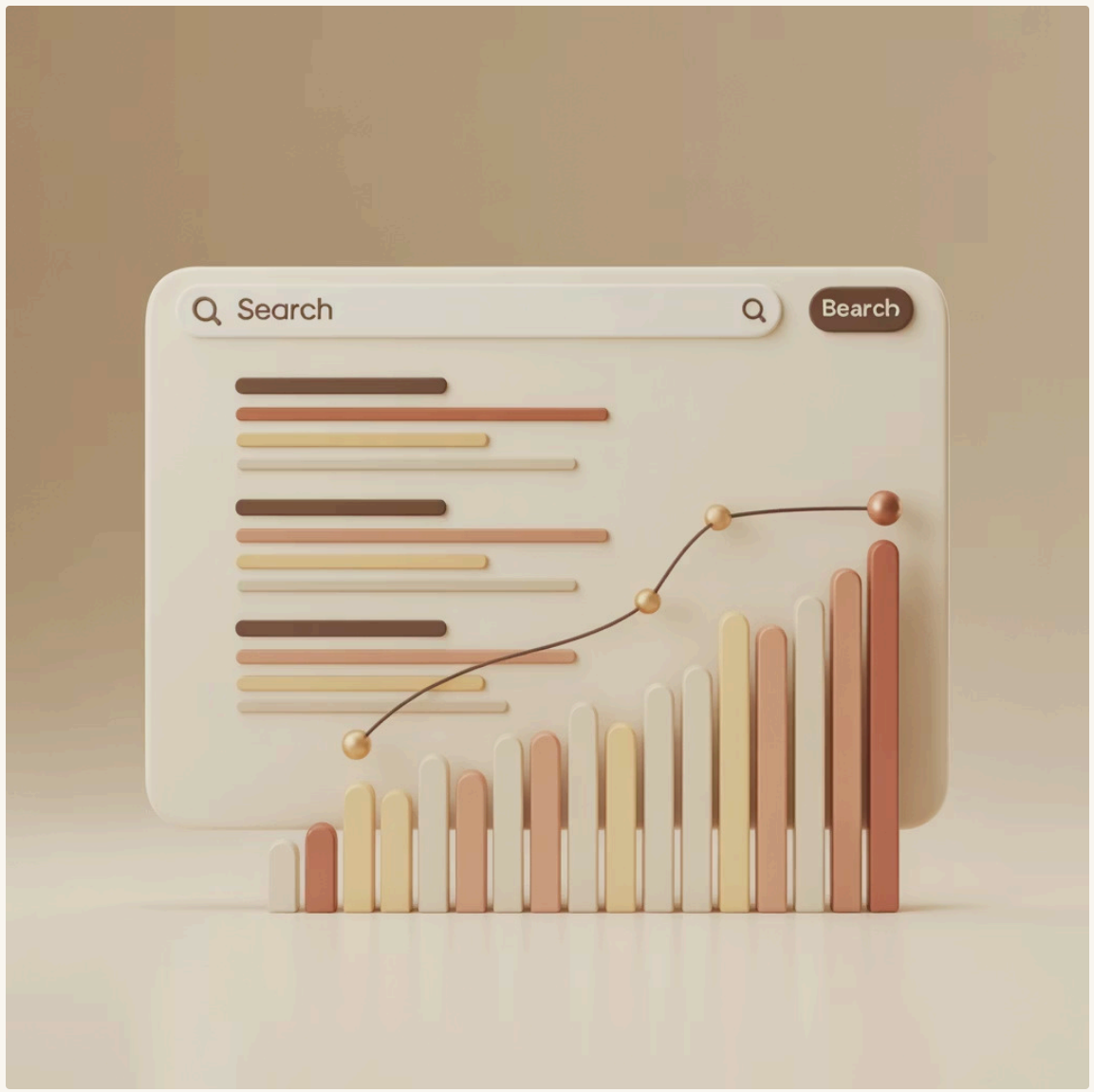
Contoh Kasus

Sistem mengembalikan 100 dokumen untuk query "machine learning". Dari 100 dokumen tersebut, 80 dokumen relevan dan 20 tidak relevan.

Precision = 80/100 = 0.80 atau 80%

Ini berarti 80% dari hasil pencarian adalah dokumen yang relevan.

Recall: Mengukur Kelengkapan



Contoh Kasus

Dalam koleksi database terdapat 150 dokumen relevan untuk query "machine learning". Sistem hanya berhasil menemukan 80 dokumen.

Recall = 80/150 = 0.53 atau 53%

Ini berarti sistem hanya menemukan 53% dari semua dokumen relevan yang ada.

Definisi

Recall mengukur proporsi dokumen relevan yang berhasil dikembalikan dari semua dokumen relevan yang ada.

$$Recall = \frac{TP}{TP + FN}$$

Recall menjawab pertanyaan: "Dari semua dokumen relevan yang ada, berapa banyak yang berhasil ditemukan sistem?"

Trade-off Precision dan Recall

Precision dan Recall memiliki hubungan terbalik yang penting untuk dipahami dalam desain sistem IR:

Precision Tinggi

Sistem sangat selektif, hanya mengembalikan dokumen dengan confidence tinggi

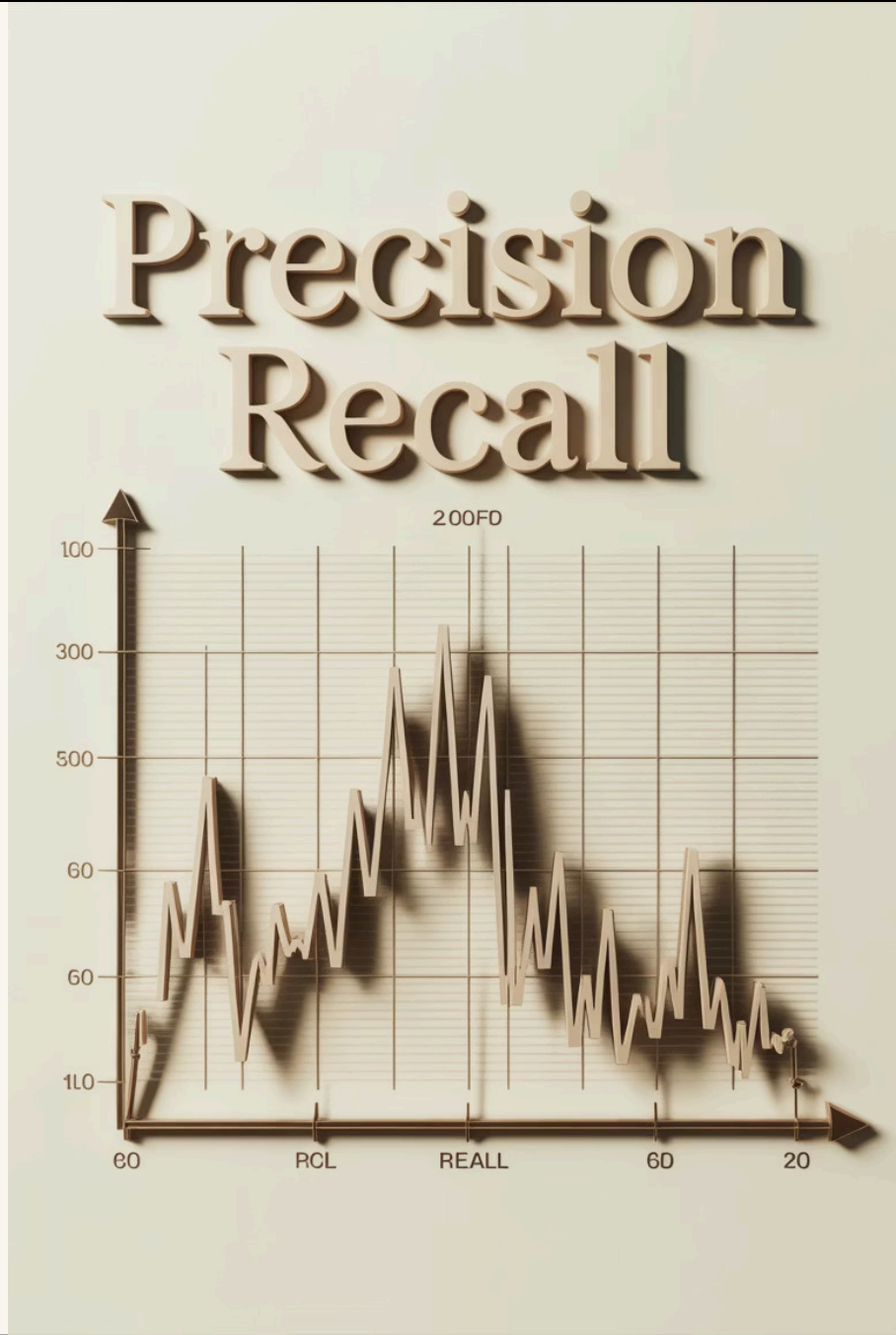
Konsekuensi: Banyak dokumen relevan terlewatkan (Recall rendah)

Recall Tinggi

Sistem mengembalikan banyak dokumen untuk memastikan semua relevan tertangkap

Konsekuensi: Banyak dokumen tidak relevan ikut masuk (Precision rendah)

Keseimbangan optimal bergantung pada kebutuhan aplikasi. Sistem medis mungkin memprioritaskan Recall tinggi, sementara mesin pencari web fokus pada Precision.



F1-Score: Harmonic Mean

F1-Score menggabungkan Precision dan Recall menjadi satu metrik yang seimbang, memberikan ukuran tunggal untuk performa sistem IR.

Formula F1-Score

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1-Score adalah harmonic mean dari Precision dan Recall. Menggunakan harmonic mean memberikan penalti lebih besar ketika salah satu metrik sangat rendah.

Karakteristik F1-Score:

- Nilai berkisar dari 0 hingga 1 (atau 0% hingga 100%)
- Mencapai nilai maksimum ketika Precision = Recall
- Sensitif terhadap nilai ekstrem rendah
- Cocok untuk membandingkan sistem dengan trade-off berbeda

Contoh Perhitungan

Sistem A:

Precision = 0.80, Recall = 0.60

F1 = 2 × (0.80 × 0.60)/(0.80 + 0.60) = 0.686

Sistem B:

Precision = 0.70, Recall = 0.70

F1 = 2 × (0.70 × 0.70)/(0.70 + 0.70) = 0.700

Meskipun Sistem A memiliki Precision lebih tinggi, Sistem B memiliki F1-Score lebih baik karena lebih seimbang.

Mean Average Precision (MAP)

MAP adalah metrik yang mempertimbangkan urutan ranking dokumen dalam hasil pencarian, memberikan bobot lebih pada dokumen relevan yang muncul lebih awal.

01

Average Precision (AP)

Hitung precision pada setiap posisi dokumen relevan, lalu ambil rata-ratanya

02

Mean Average Precision

Hitung AP untuk semua query, lalu ambil rata-rata dari semua AP

Formula MAP

$$AP(q) = \frac{1}{|R_q|} \sum_{k=1}^{n} P(k) \times rel(k)$$

$$MAP = \frac{1}{|Q|} \sum_{q=1}^{|Q|} AP(q)$$

Dimana: *Rq* = jumlah dokumen relevan untuk query q, *P(k)* = precision pada posisi k, *rel(k)* = 1 jika dokumen di posisi k relevan, 0 jika tidak

MAP sangat berguna karena menekankan pentingnya menempatkan dokumen relevan di posisi atas hasil pencarian, sesuai dengan perilaku pengguna yang cenderung hanya melihat hasil teratas.

Normalized Discounted Cumulative Gain (NDCG)

NDCG adalah metrik evaluasi yang mengakomodasi tingkat relevansi bertingkat (graded relevance), bukan hanya biner relevan/tidak relevan.

1

Cumulative Gain (CG)

Menjumlahkan semua skor relevansi dari dokumen yang dikembalikan

$$CG_p = \sum_{i=1}^{p} rel_i$$

2

Discounted Cumulative Gain (DCG)

Memberikan penalti pada dokumen relevan yang muncul di posisi lebih rendah

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$$

3

Normalized DCG (NDCG)

Normalisasi DCG dengan membaginya dengan DCG ideal (IDCG)

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

Perbandingan Metrik Evaluasi

Setiap metrik memiliki kelebihan dan kelemahan tersendiri. Pemilihan metrik harus disesuaikan dengan kebutuhan sistem IR yang dikembangkan.

Metrik	Kelebihan	Kelemahan	Penggunaan Terbaik
Precision	Mudah dipahami, fokus pada kualitas	Tidak mempertimbangkan kelengkapan	Sistem yang memprioritaskan akurasi
Recall	Mengukur kelengkapan hasil	Tidak mempertimbangkan noise	Aplikasi medis, legal research
F1-Score	Keseimbangan precision-recall	Tidak mempertimbangkan ranking	Evaluasi sistem secara keseluruhan
MAP	Mempertimbangkan posisi ranking	Hanya untuk relevansi biner	Mesin pencari web, e-commerce
NDCG	Support graded relevance	Lebih kompleks untuk dihitung	Sistem dengan multi-level relevance

Rangkuman dan Praktik Evaluasi

Prinsip Dasar

Precision mengukur ketepatan, Recall mengukur kelengkapan. Keduanya memiliki trade-off yang perlu dikelola sesuai kebutuhan aplikasi.

Metrik Gabungan

F1-Score memberikan ukuran seimbang. MAP dan NDCG mempertimbangkan urutan ranking untuk evaluasi yang lebih realistis.

Pemilihan Metrik

Pilih metrik berdasarkan karakteristik sistem dan kebutuhan pengguna. Gunakan kombinasi metrik untuk evaluasi komprehensif.

Langkah Praktis Evaluasi IR

1. Definisikan test collection dengan ground truth relevance judgments
2. Jalankan sistem IR pada set query yang representatif
3. Hitung berbagai metrik evaluasi untuk analisis multi-dimensi
4. Bandingkan dengan baseline atau sistem kompetitor
5. Identifikasi area perbaikan dan iterasi pengembangan sistem

Dengan memahami dan menerapkan metrik evaluasi ini, mahasiswa dapat mengembangkan sistem Information Retrieval yang efektif dan terukur kualitasnya.