

PERTEMUAN 1

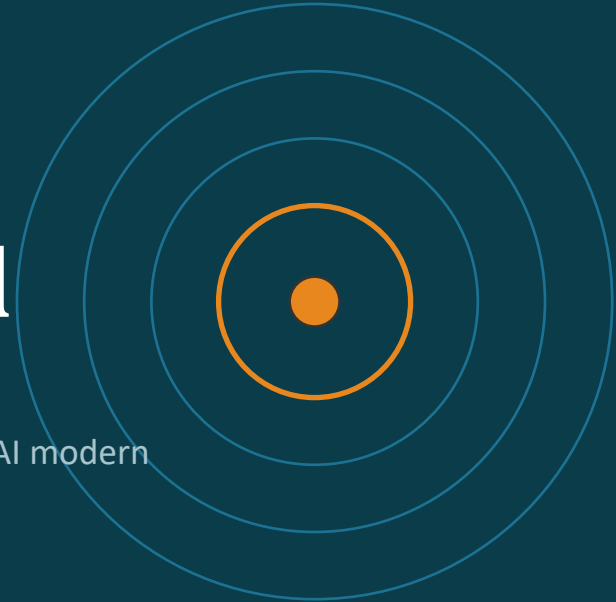
Pengantar Information Retrieval

Konsep, sejarah, arsitektur sistem, dan peran IR dalam ekosistem AI modern

IF261224 • 3 SKS • Semester Gasal

Program Studi Informatika — Fakultas Ilmu Komputer

UPN “Veteran” Jawa Timur



Peta Perkuliahan Pertemuan 1

1

Definisi & ruang lingkup

Apa yang dipelajari dan tidak dipelajari dalam IR

2

Sejarah singkat IR

Dari gagasan Memex sampai pencarian berbasis LLM

3

IR, data retrieval, QA

Membedakan tiga cara berbeda menemukan jawaban

4

Arsitektur sistem IR

Dua pipeline: indexing (offline) dan query (online)

5

IR di ekosistem AI

Mesin pencari, RAG, agen cerdas, dan kasus di Indonesia

Capaian Pembelajaran Pertemuan 1

Sub-CPMK-1 — turunan dari CPMK-07.1 dan CPL-TF-07

Sub-CPMK-1

Mahasiswa mampu menjelaskan konsep, sejarah, ruang lingkup, dan arsitektur sistem Information Retrieval serta posisinya dalam ekosistem AI modern.

Ranah kognitif: C2 (memahami)

Indikator penilaian

1

Ketepatan menjelaskan definisi dan ruang lingkup IR

2

Ketepatan membedakan IR, data retrieval, dan question answering

3

Ketepatan menguraikan komponen arsitektur sistem IR

Penilaian pertemuan ini bersifat formatif (bobot 0%) melalui kuis singkat dan partisipasi diskusi.

Masalah yang Ingin Dipecahkan

Data yang melimpah tidak otomatis berarti informasi akan mudah ditemukan

VOLUME

Miliaran dokumen

Koleksi tumbuh jauh lebih cepat daripada kemampuan manusia membacanya. Pencarian manual tidak lagi mungkin.

KECEPATAN

Ratusan milidetik

Pengguna mengharapkan hasil hampir seketika. Sistem harus mencari tanpa memindai seluruh koleksi.

RELEVANSI

Sepuluh hasil teratas

Yang dinilai pengguna bukan seberapa banyak hasil, melainkan seberapa tepat hasil di urutan pertama.

Information Retrieval adalah rekayasa untuk memenuhi ketiganya sekaligus: skala besar, waktu tanggap (*response time*) singkat, dan hasil yang relevan.

Apa Itu Information Retrieval?

Definisi klasik Manning, Raghavan & Schütze (2008)

“Menemukan materi, umumnya berupa dokumen, yang bersifat tidak terstruktur dan memenuhi kebutuhan informasi, dari dalam koleksi besar yang tersimpan pada komputer.”

1

Tidak terstruktur

Teks bebas tanpa skema baku, berbeda dari tabel basis data

2

Kebutuhan informasi

Yang dicari pengguna, belum tentu sama dengan kata yang diketik

3

Koleksi besar

Skala menuntut struktur data khusus, bukan pemindaian berurutan

4

Menemukan

Bukan sekadar mencocokkan, tetapi mengurutkan menurut relevansi

Ruang Lingkup Information Retrieval

Batas bidang: apa yang menjadi fokus dan apa yang bersinggungan

Inti bidang IR

- Representasi dan pengindeksan dokumen
- Pemodelan kebutuhan informasi menjadi query
- Pencocokan dan perangkingan menurut relevansi
- Evaluasi efektivitas dan efisiensi sistem
- Pemanfaatan umpan balik dan perilaku pengguna
- Temu kembali web, multimedia, dan lintas bahasa

Bersinggungan, bukan fokus utama

- Basis data relasional dan optimasi kueri SQL
- *Natural language processing* (parsing, terjemahan)
- Data mining pada data terstruktur
- Perancangan antarmuka dan *user experience*
- Infrastruktur penyimpanan terdistribusi

Kolom sebelah kanan tetap dibutuhkan sebagai penopang, IR modern dibangun di atas keduanya.

Perjalanan Singkat Information Retrieval

1945

Gagasan Memex

Vannevar Bush, “As We May Think”: mesin penelusur asosiatif

1957

Pembobotan otomatis

H. P. Luhn: frekuensi kata sebagai penanda isi dokumen

1960-an

Cranfield & SMART

Paradigma evaluasi dan Vector Space Model (Salton)

1994

Okapi BM25

Model probabilistik yang menjadi baseline fungsi perankingan berbasis kata kunci sampai hari ini

1998

PageRank

Struktur tautan web sebagai sinyal kualitas halaman

2013

Word embedding

Word2Vec: makna kata sebagai vektor terdistribusi

2019–20

BERT, DPR, RAG

Transformer untuk ranking; retrieval mensuplai konteks ke generator

2023–

Pencarian generatif

Antarmuka percakapan dan agen pencari berbasis LLM

Dua kartu terakhir menjadi fokus paruh kedua semester (pertemuan 10–15).

IR, Data Retrieval, dan Question Answering

Tiga cara berbeda menemukan jawaban dari data

Aspek	Data Retrieval	Information Retrieval	Question Answering
Objek data	Terstruktur (tabel, skema)	Tidak terstruktur (teks bebas)	Teks bebas atau basis pengetahuan
Bentuk query	Bahasa formal (SQL)	Kata kunci atau bahasa alami	Pertanyaan bahasa alami
Kecocokan	Eksak (benar atau salah)	Parsial (berderajat relevansi)	Semantik (memahami maksud)
Keluaran	Himpunan baris	Daftar dokumen terurut	Jawaban ringkas beserta sumber
Ukuran berhasil	Kebenaran hasil	Precision, recall, nDCG	Ketepatan dan kesetiaan jawaban
Contoh	SELECT pada basis data akademik	Google, pencarian di Tokopedia	Asisten berbasis LLM dengan RAG

Ketiganya bukan pengganti satu sama lain, namun digabungkan dalam satu blok pada sistem modern.

Empat Konsep Inti

Kosakata dasar yang dipakai sepanjang semester

1

Dokumen

Unit informasi yang dikembalikan sistem: artikel, halaman web, paragraf, produk, bahkan potongan (chunk) teks.

2

Koleksi / Korpus

Himpunan seluruh dokumen yang dapat dicari. Ukurannya menentukan pilihan struktur data dan algoritma.

3

Query

Ungkapan kebutuhan informasi dalam bentuk yang dipahami sistem — seringkali pendek, ambigu, dan tidak lengkap.

4

Relevansi

Derajat kesesuaian dokumen dengan kebutuhan informasi pengguna, bukan sekadar kemunculan kata pada dokumen.

Relevansi diberi warna berbeda karena ia satu-satunya konsep yang tidak dapat dihitung secara pasti — dan justru itulah inti persoalan IR.

Mengapa Relevansi Sulit Diukur?

Empat sumber kesulitan yang menjadi alasan keberadaan bidang IR

Subjektif

Dua orang dengan query identik dapat menilai dokumen yang sama secara berbeda.

Bergantung konteks

Lokasi, waktu, riwayat penelusuran, dan tugas pengguna mengubah apa yang dianggap relevan.

Query $\neq$ kebutuhan

Pengguna sering tidak tahu cara merumuskan apa yang sebenarnya dicari.

Berubah seiring waktu

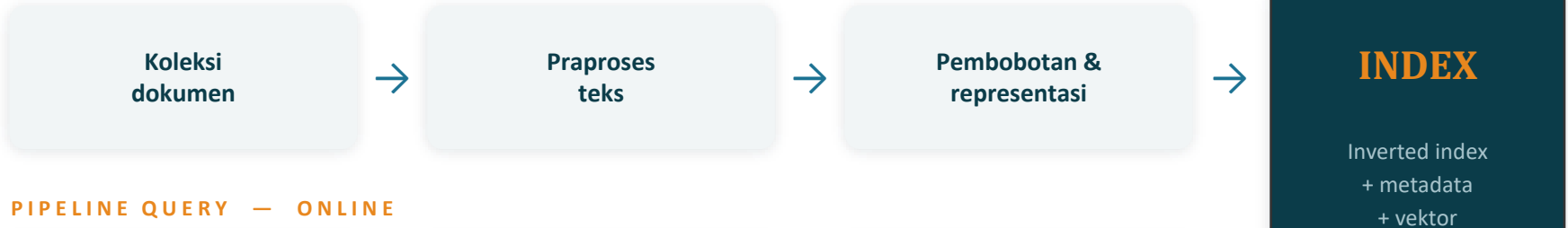
Query “pemilu” bermakna berbeda pada tahun yang berbeda.

Topik Diskusi: sebutkan satu pencarian yang pernah gagal memberi Anda hasil yang tepat. Sumber kegagalannya yang mana?

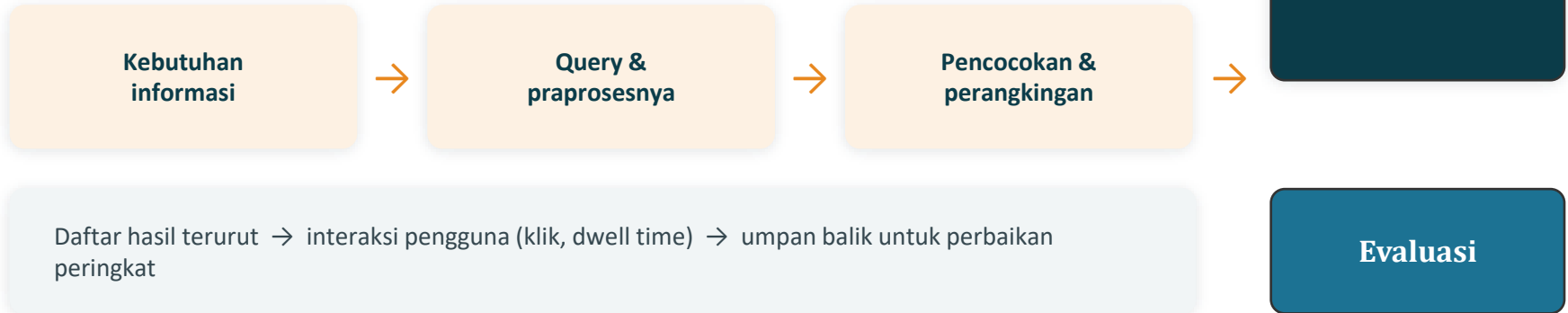
Arsitektur Umum Sistem Information Retrieval

Dua pipeline yang bertemu pada satu struktur data: index

PIPELINE INDEXING — OFFLINE



PIPELINE QUERY — ONLINE



Seluruh mata kuliah ini pada dasarnya adalah pembahasan mendalam atas setiap kotak pada diagram di atas.

Pipeline Indexing (Offline)

Dikerjakan sekali di muka dan diperbarui berkala — boleh lambat, harus menyeluruh

1

Akuisisi dokumen

Crawling web, ekspor basis data, unggahan pengguna, atau feed data. Menentukan cakupan koleksi.

Mg 9

2

Parsing & ekstraksi

Mengubah HTML, PDF, atau DOCX menjadi teks polos beserta metadata (judul, penulis, tanggal).

Mg 2

3

Praproses teks

Tokenisasi, case folding, penghapusan stopword, stemming atau lemmatization.

Mg 2

4

Pembobotan & representasi

Menghitung TF-IDF atau menghasilkan vektor embedding untuk setiap dokumen.

Mg 4, 11

5

Penyimpanan index

Membangun inverted index, kompresi postings, dan struktur pencarian vektor.

Mg 3, 11

Label di kanan menunjukkan pertemuan tempat setiap tahap dibahas secara mendalam.

Pipeline Query (Online)

Dieksekusi setiap kali pengguna menekan Enter — harus selesai dalam hitungan milidetik

1

Penerimaan query

Menangkap masukan pengguna beserta konteks: lokasi, bahasa, perangkat, riwayat sesi.

Mg 5

2

Pemahaman query

Praproses yang konsisten dengan indexing, koreksi ejaan, ekspansi, dan penulisan ulang query.

Mg 5

3

Retrieval kandidat

Mengambil beberapa ratus hingga ribuan dokumen kandidat secara cepat dari index.

Mg 3, 11

4

Perangkingan

Menghitung skor relevansi, lalu menyusun ulang kandidat teratas dengan model yang lebih mahal.

Mg 4–7, 12

5

Penyajian & umpan balik

Menampilkan hasil beserta cuplikan, merekam klik, dan menyalurkannya kembali ke sistem.

Mg 7, 15

Pola dua tahap — retrieval cepat lalu perangkingan mahal — adalah rancangan standar mesin pencari modern.

Menelusuri Satu Query

Contoh: seorang mahasiswa mengetik “banjir surabaya kemarin”

Kebutuhan informasi

Ingin tahu lokasi dan dampak banjir terbaru di Surabaya — tidak seluruhnya tertulis dalam query.

Query terproses

Token: [banjir] [surabaya] [kemarin]. Kata “kemarin” diterjemahkan menjadi penyaring rentang tanggal.

Retrieval kandidat

Index memberikan seluruh dokumen yang memuat “banjir” dan “surabaya” — misalnya beberapa ribu artikel.

Perangkingan

Dokumen diberi skor berdasarkan bobot term, kesegaran waktu, kredibilitas sumber, dan kedekatan kata.

Penyajian

Sepuluh artikel teratas ditampilkan bersama cuplikan yang memuat kata pencarian dalam sorotan.

Umpan balik

Klik pengguna pada hasil ketiga menjadi sinyal bahwa peringkat masih dapat diperbaiki.

Perhatikan: query hanya tiga kata, tetapi sistem membuat sejumlah keputusan tersembunyi sebelum hasil muncul.

Peran IR dalam Ekosistem AI Modern

Kemunculan LLM tidak menggantikan IR — justru menjadikannya semakin penting

Mesin pencari

Bentuk klasik dan masih dominan. Kini diperkaya pemahaman semantik dan ringkasan hasil.

Retrieval-Augmented Generation

Retrieval mensuplai fakta mutakhir ke LLM sehingga jawaban dapat dirujuk dan diverifikasi.

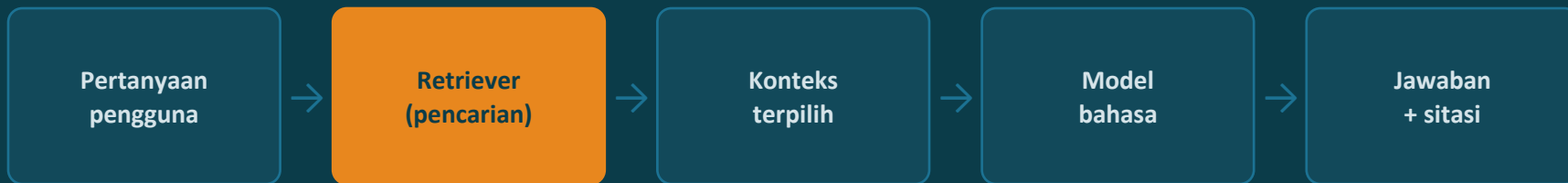
Agen cerdas

Agen memanggil pencarian berulang kali sebagai alat untuk menyelesaikan tugas bertahap.

Model bahasa hanya mengetahui apa yang ada dalam bobotnya. Untuk menjawab pertanyaan tentang dokumen tertentu, data terbaru, atau informasi internal organisasi, model tersebut harus dipasangkan dengan sistem temu kembali yang baik.

Retrieval adalah Jantung RAG

Kualitas jawaban generatif dibatasi oleh kualitas dokumen yang berhasil ditemukan



Fokus mata kuliah ini

- Dokumen yang salah ambil akan menghasilkan jawaban yang salah, betapapun *powerful*-nya model bahasa.
- Dokumen yang benar tetapi berperingkat rendah akan terpotong oleh batas panjang konteks.
- Karena itu, metrik retrieval — recall dan nDCG — tetap menjadi ukuran utama, bukan pelengkap.

IR di Sekitar Kita: Konteks Indonesia

Enam ranah penerapan yang dapat dijadikan tema final project

Perdagangan elektronik

Pencarian produk pada Tokopedia atau Shopee: menangani salah ketik, sinonim, dan peringkat berbasis perilaku pembeli.

Portal berita

Arsip Kompas, Detik, atau Tempo dengan kebutuhan pencarian yang mengutamakan kesegaran waktu.

Repositori akademik

Garuda, SINTA, dan repositori tugas akhir kampus untuk penelusuran literatur berbahasa Indonesia.

Layanan publik & hukum

JDIH dan portal peraturan: pencarian pasal yang menuntut ketepatan tinggi dan penelusuran rujukan.

Kesehatan digital

Pencarian gejala dan artikel medis yang mensyaratkan sumber terpercaya serta penanganan istilah awam.

Sistem internal kampus

Pencarian dokumen akademik, panduan, dan basis pengetahuan layanan mahasiswa.

Tantangan khas Bahasa Indonesia: imbuhan yang produktif, kata ulang, campur kode, dan singkatan tidak baku pada query.

Peta Perjalanan 16 Minggu

Fondasi klasik dahulu, baru pendekatan modern — keduanya tetap dipakai berdampingan

Minggu 1-7 — Fondasi Klasik

- Pengantar dan arsitektur sistem IR
- Pemrosesan teks dan inverted index
- Vector Space Model dan TF-IDF
- BM25 dan relevance feedback
- Evaluasi sistem IR
- Learning to Rank

Minggu 9-15 — IR Modern

- Web IR, crawling, dan link analysis
- Representasi semantik dan embedding
- Dense retrieval dan vector database
- Re-ranking dan hybrid search
- Retrieval-Augmented Generation
- Etika, bias, dan penerapan sistem

Minggu 8: Evaluasi Tengah Semester (30%)

•

Minggu 16: Final Project Mini Search Engine (40%)

Diskusi Kelas

Kerjakan berpasangan, 10 menit, lalu bagikan ke kelas

1

Sebutkan satu aplikasi yang Anda pakai hari ini yang di dalamnya terdapat sistem IR. Apa yang berperan sebagai dokumen, dan apa yang berperan sebagai query?

2

Menurut Anda, mengapa mesin pencari tidak cukup hanya menampilkan seluruh dokumen yang memuat kata pencarian?

3

Jika Anda diminta membangun mesin pencari untuk repositori tugas akhir kampus, informasi apa saja yang perlu disimpan pada index?

Rangkuman & Persiapan Pertemuan Berikutnya

- IR berurusan dengan data tidak terstruktur, kecocokan parsial, dan hasil yang diurutkan menurut relevansi.
- Relevansi bersifat subjektif dan kontekstual — karena itu evaluasi menjadi bagian tak terpisahkan dari IR.
- Setiap sistem IR terdiri atas dua pipeline: indexing yang berjalan offline dan query yang berjalan online.
- Kehadiran LLM tidak menggantikan IR; retrieval justru menjadi penentu kebenaran jawaban pada sistem RAG.

Persiapan Pertemuan 2

Baca Manning dkk. Bab 1–2. Pasang Python 3, Jupyter, NLTK, dan Sastrawi. Siapkan satu berkas teks berbahasa Indonesia untuk dipakai pada praktikum preproses.

Pustaka pertemuan ini

- [1] Manning, Raghavan & Schütze (2008), Bab 1
- [2] Croft, Metzler & Strohman (2015), Bab 1
- [3] Grainger, Turnbull & Irwin (2025), Bab 1