

WEB INFORMATION RETRIEVAL

Pertemuan 1

Pengantar Web Information Retrieval

Definisi • Ruang Lingkup • Traditional IR vs Web IR • Search Lifecycle • Relevance

Dr. Ahmad Muhammad Thantawi, MTA, MOS, MCE

SEARCH IS A DECISION SYSTEM

 web information retrieval

Web Information Retrieval — Overview

Definisi, query, indexing, ranking, dan search engine...

Result 2

Dokumen relevan lain yang cocok dengan query pengguna.

Result 3

Dokumen relevan lain yang cocok dengan query pengguna.



QUERY → MATCH → RANK → ANSWER

Capaian Pembelajaran Pertemuan 1

Mahasiswa memahami konsep dasar Information Retrieval dan konteksnya pada lingkungan web.

1

Sub-CPMK 1

Menjelaskan definisi, ruang lingkup, tujuan, dan karakteristik Web Information Retrieval serta membedakannya dari IR tradisional.



Indikator Keberhasilan

1. Menjelaskan istilah IR, query, document, relevance, ranking.
2. Membedakan database retrieval, traditional IR, dan Web IR.
3. Memetakan alur pencarian web secara konseptual.



Bukti Belajar

- Diskusi pemantik
- Analisis SERP
- Latihan information need → query
- Case method
- Quiz formatif

Posisi Pertemuan 1 dalam Alur Mata Kuliah

Pertemuan ini membangun bahasa konsep sebelum masuk ke arsitektur, crawling, indexing, dan ranking.



Logika pembelajaran

Pahami masalah pencarian → pahami mesin pencari → pelajari cara menemukan dokumen → bangun indeks → beri skor/ranking → ukur kualitas hasil.

Mengapa pencarian web bukan sekadar “mencari kata”?

Satu query pendek dapat mewakili kebutuhan informasi yang berbeda.

“**“jaguar” dapat berarti mobil, hewan, tim olahraga, atau sistem operasi lama. Mesin pencari harus menebak maksud pengguna.**

Contoh ambiguitas query

Ambiguity

Satu kata, banyak makna

Scale

Miliaran halaman

Freshness

Web terus berubah

Trust

Tidak semua sumber berkualitas

Evolusi Pencarian Informasi

Dari katalog statis menuju pencarian web, semantic search, dan AI-assisted search.



Apa itu Information Retrieval (IR)?

IR berfokus pada menemukan informasi yang paling relevan dari koleksi besar berdasarkan kebutuhan pengguna.

INFORMATION RETRIEVAL

Proses merepresentasikan, menyimpan, mengorganisasi, mencari, dan memberi peringkat dokumen agar informasi relevan dapat ditemukan secara efisien.

Query

Representasi kebutuhan pengguna

Document

Unit informasi yang dicari

Relevance

Seberapa sesuai hasil dengan kebutuhan

Ranking

Urutan hasil berdasarkan estimasi relevansi

Tujuan Utama Sistem IR

Bukan menemukan semua dokumen—melainkan membantu pengguna menemukan informasi yang berguna dengan cepat.



Relevant

Hasil sesuai kebutuhan informasi



Efficient

Respons cepat pada koleksi besar



Usable

Mudah dipahami & dinavigasi



Trustworthy

Kualitas, keamanan, dan konteks

Ukuran keberhasilan akhirnya: apakah pengguna mendapatkan jawaban atau dokumen yang berguna?

Apa yang Membuat Web IR Berbeda?

Web IR menggabungkan prinsip IR dengan karakteristik unik web: skala, struktur link, dinamika, dan perilaku pengguna.

Skala

Koleksi sangat besar dan tersebar

Link graph

Hyperlink membawa sinyal otoritas

Dinamika

Konten berubah terus-menerus

Heterogen

HTML, PDF, gambar, video, forum, produk

WEB IR

IR principles

+

Web signals

+

User context

CRAWL

INDEX

RANK

Tujuannya tetap sama: estimasi relevansi terbaik pada lingkungan yang jauh lebih kompleks.

Database Retrieval vs Information Retrieval

Keduanya “mencari”, tetapi jenis data, query, dan hasilnya berbeda.

Aspek	Database Retrieval	Information Retrieval
Data	Terstruktur (tabel, kolom)	Dominan tidak terstruktur / semi-terstruktur
Query	Eksak: SQL / kondisi logis	Kata kunci / natural language
Kecocokan	True / false	Derajat relevansi
Output	Record yang memenuhi kondisi	Daftar hasil berperingkat
Contoh	SELECT * FROM mahasiswa WHERE...	“beasiswa sistem informasi jakarta”

Traditional IR vs Web IR

Web IR mempertahankan inti IR tetapi menambahkan problem skala, link, spam, freshness, dan konteks pengguna.

Traditional IR

- Koleksi relatif terkontrol
- Dokumen cenderung stabil
- Metadata lebih konsisten
- Query difokuskan pada isi dokumen
- Evaluasi dilakukan pada test collection
- Contoh: perpustakaan digital, arsip organisasi

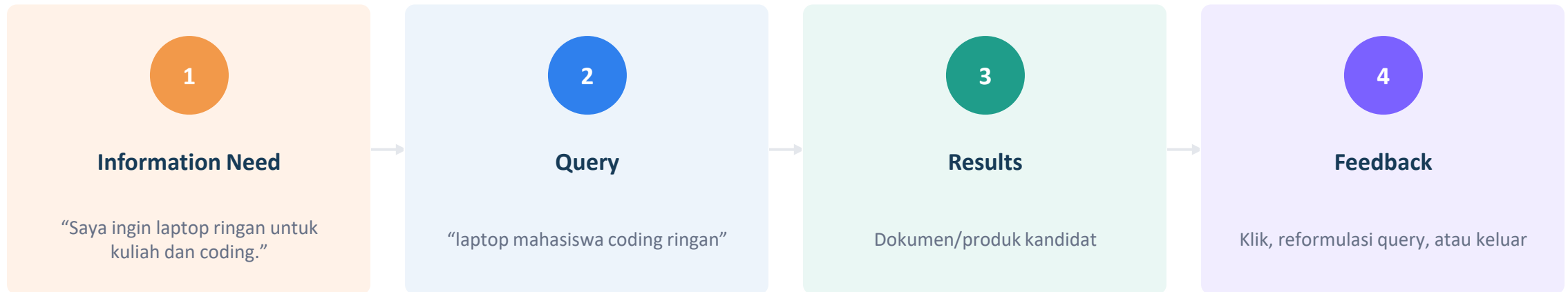
Web IR

- Koleksi sangat besar & terbuka
- Konten berubah cepat
- Link dan struktur web menjadi sinyal
- Spam/SEO manipulation perlu ditangani
- Ranking dipengaruhi konteks dan perilaku
- Contoh: web search, marketplace, portal berita

Inti yang sama: representasi dokumen + query → estimasi relevansi → ranking hasil.

Information Need $\neq$ Query

Pengguna memiliki kebutuhan di kepala; query hanya representasi singkat dari kebutuhan tersebut.



Tantangan IR: gap antara kebutuhan sebenarnya dan query yang diketik.

Apa arti “relevan”?

Relevance bukan sekadar ada kata yang sama; relevansi bergantung pada kebutuhan, konteks, dan tujuan pengguna.

Topical

Apakah isi dokumen membahas topik yang dicari?

Task

Apakah hasil membantu menyelesaikan tugas pengguna?

Context

Apakah sesuai lokasi, waktu, perangkat, bahasa?

Trust

Apakah sumber cukup berkualitas & aman?

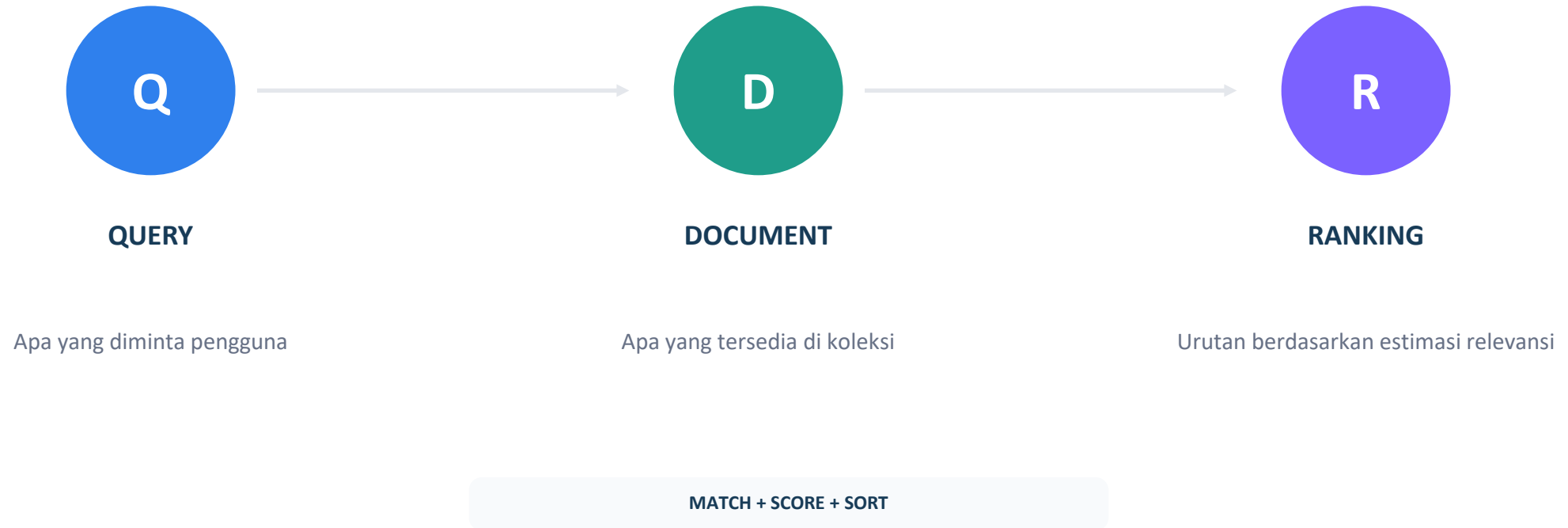
“

Sistem IR mengestimasi relevance; pengguna yang akhirnya menilai apakah hasil benar-benar berguna.

Konsep kunci

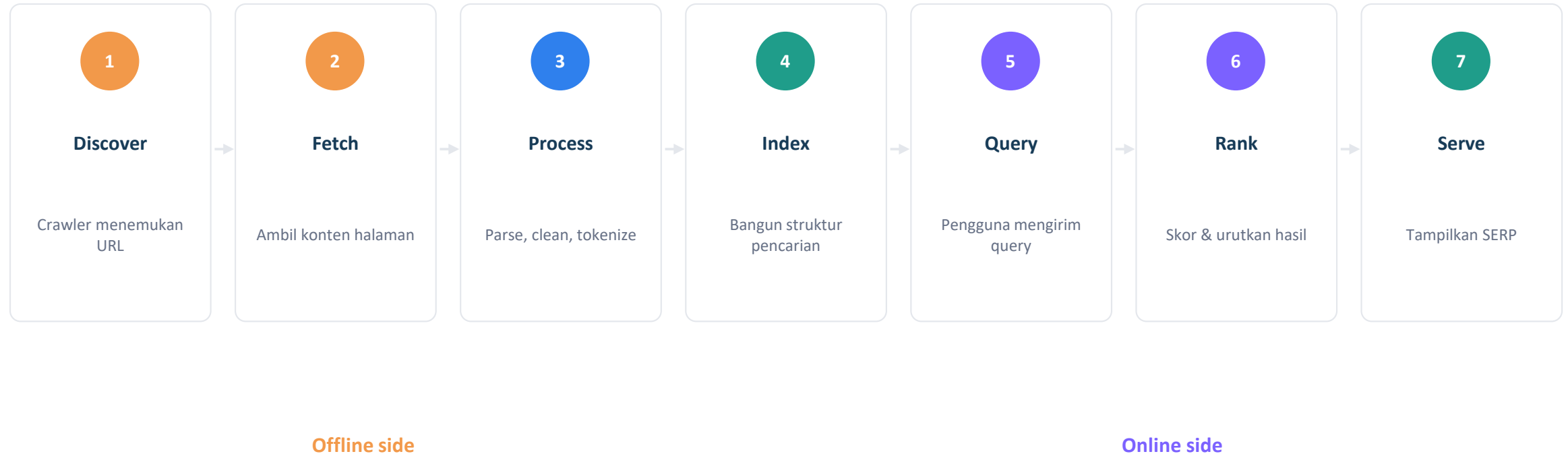
Tiga Objek Utama dalam IR

Semua sistem IR dapat dipahami melalui hubungan query, document, dan ranking.



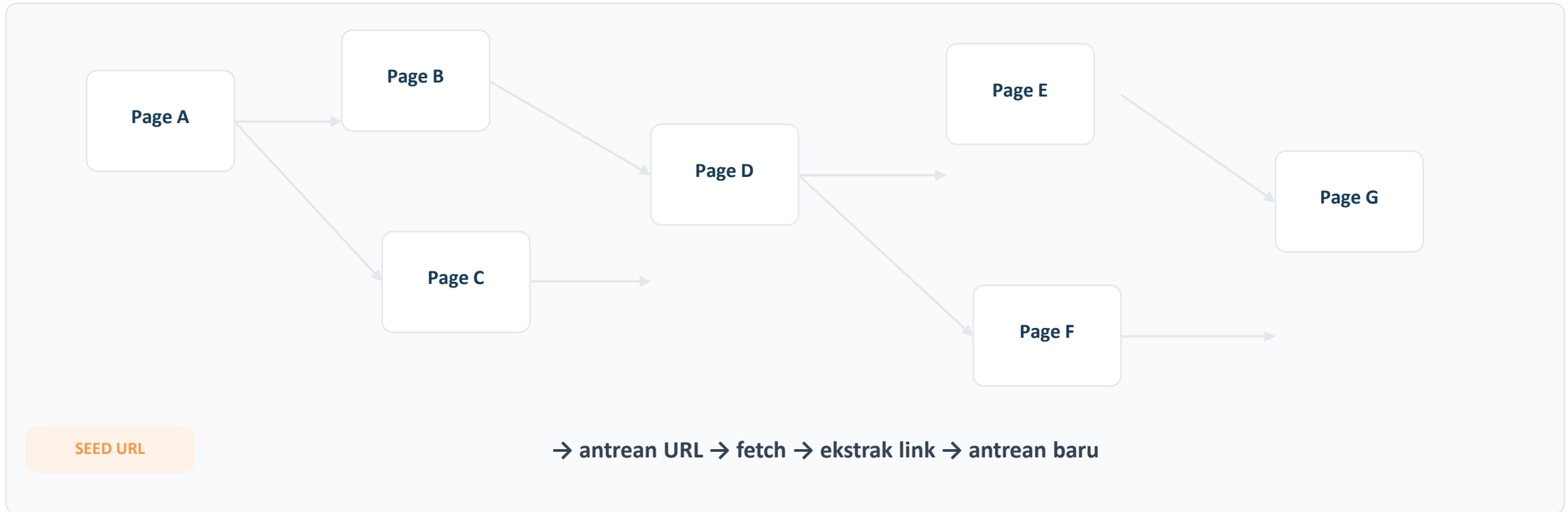
Siklus Hidup Pencarian Web

Pertemuan berikutnya akan membedah setiap komponen; hari ini kita melihat gambaran besarnya.



Crawler: Menemukan Halaman Web

Crawler menjelajahi web melalui URL dan hyperlink lalu memilih halaman yang perlu diambil.



Indexer: Membuat Web Menjadi “Searchable”

Tanpa indeks, sistem harus membaca semua dokumen setiap kali query masuk.



Query Processor & Ranker

Sistem menafsirkan query, mengambil kandidat dari indeks, memberi skor, lalu mengurutkan hasil.



web crawler tutorial

1

Web Crawler Tutorial

score 0.91

2

How Search Engines Crawl the Web

score 0.83

3

Crawler vs Scraper

score 0.71

4

Intro to Indexing

score 0.58

Ranking = estimasi relevansi, bukan kebenaran absolut.

Anatomi Search Engine Results Page (SERP)

Hasil pencarian adalah antarmuka keputusan: pengguna memindai, membandingkan, lalu memilih.



web information retrieval

[example.org](#) > [course](#) > [ir](#)

Introduction to Web Information Retrieval

Cuplikan isi dokumen yang relevan dengan query pengguna...

[example.org](#) > [course](#) > [ir](#)

Crawler, Indexer and Ranker Explained

Cuplikan isi dokumen yang relevan dengan query pengguna...

[example.org](#) > [course](#) > [ir](#)

Ranking Models for Web Search

Cuplikan isi dokumen yang relevan dengan query pengguna...

Title

Menjelaskan isi halaman dengan ringkas

URL / Source

Memberi konteks sumber dan kredibilitas

Snippet

Potongan teks untuk membantu keputusan klik

Satu Query, Banyak Interpretasi

Query yang sama dapat mengarah pada intent dan hasil yang berbeda.

Query: “python”

Programming

Tutorial Python, docs, package

Snake

Artikel ular python

Movie / culture

Monty Python

Course intent

Kursus Python terdekat/online

Bagaimana mesin menebak intent?

isi query

popularitas konteks

bahasa & lokasi

freshness

riwayat / konteks sesi

Karakteristik Web sebagai Koleksi Dokumen

Web adalah koleksi terbuka—tidak ada satu pengelola yang menjamin struktur dan kualitasnya.

1

Volume

Jumlah halaman sangat besar

2

Velocity

Konten lahir, berubah, hilang

3

Variety

Teks, PDF, video, gambar, produk

4

Redundancy

Duplikasi & near-duplicate

5

Noise

Spam, template, iklan, boilerplate

6

Link graph

Halaman saling terhubung

Tantangan Utama Web Information Retrieval

Kualitas hasil pencarian bergantung pada banyak problem yang tidak muncul pada pencarian database biasa.

1

Freshness

hasil lama bisa cepat usang

2

Spam & manipulation

konten dibuat untuk memengaruhi ranking

3

Ambiguity

query pendek, maksud tidak eksplisit

4

Multilingual

bahasa dan istilah berbeda

5

Authority

sumber populer belum tentu benar

6

Personalization

konteks pengguna dapat mengubah relevansi

Tipe Kebutuhan dalam Pencarian Web

Memahami intent membantu sistem menentukan jenis hasil dan ranking yang tepat.

Informational

““apa itu web crawling””

Mencari pengetahuan / jawaban

Navigational

““wikipedia information retrieval””

Menuju situs/halaman tertentu

Transactional

““download dataset IR””

Melakukan aksi / memperoleh item

Commercial

““best search engine API””

Membandingkan sebelum keputusan

Dari Matching ke Ranking

Sistem modern biasanya menggabungkan banyak sinyal untuk memperkirakan relevansi.

1

Content match

Kecocokan istilah / semantik

2

Link authority

Sinyal struktur hyperlink

3

Freshness

Kebaruan dokumen

4

Quality

Spam, trust, usefulness

5

Context

Lokasi, bahasa, perangkat

6

Behavior

Interaksi agregat pengguna

Ranking = fungsi(sinyal konten, link, kualitas, konteks, dll.)

Web IR Tidak Hanya Google

Prinsip retrieval digunakan di banyak sistem informasi modern.

1

Marketplace

Cari produk + filter + ranking

2

E-commerce help

FAQ & knowledge base search

3

Digital Library

Dokumen ilmiah dan metadata

4

Enterprise Search

Dokumen internal organisasi

5

News / Media

Artikel dan topik terkini

6

AI / RAG

Retrieve context sebelum generasi

Pencarian Portal Kampus

Portal menyimpan berita, kalender akademik, SOP, file PDF, dan halaman program studi. Bagaimana agar mahasiswa cepat menemukan informasi?

 jadwal wisuda 2026

A

Berita wisuda 2024

lama

B

Kalender akademik 2026

mungkin relevan

C

Panduan wisuda 2026.pdf

sangat relevan

D

Halaman alumni

kurang relevan

Tugas sistem IR

Menentukan kandidat dokumen dan mengurutkan yang paling membantu pengguna.

Sinyal yang mungkin

Isi dokumen • tanggal • tipe file • struktur portal • otoritas halaman

Output ideal

Panduan atau jadwal wisuda 2026 muncul di posisi teratas dengan snippet yang jelas.

Pencarian Marketplace: “sepatu lari pria”

Marketplace menggabungkan retrieval dengan filter, ranking bisnis, stok, harga, rating, dan personalisasi.

PRODUCT RunFast Pro Rp899k • 4.8 ★ stok 14	PRODUCT TrailLite 2 Rp749k • 4.7 ★ stok 8	PRODUCT Urban Walk Rp499k • 4.9 ★ stok 22
---	--	--

“ “Paling relevan” belum tentu “paling murah”. Sistem perlu menyeimbangkan relevance, availability, quality, dan business rules.

Diskusi kelas

Latihan: Dari Information Need ke Query

Kerjakan berpasangan selama 7 menit.

A

Mahasiswa mencari regulasi terbaru tentang magang kampus.

Tulis: information need → query awal → query reformulasi

B

Pengguna ingin mencari hotel murah dekat stasiun untuk besok malam.

Tulis: information need → query awal → query reformulasi

C

Staf IT ingin dokumentasi cara mengatasi error “403 forbidden”.

Tulis: information need → query awal → query reformulasi

Pertanyaan refleksi: apa informasi yang hilang ketika kebutuhan diterjemahkan menjadi query?

Rancang Konsep Search untuk Portal Akademik

Kelompok 4–5 orang • 15 menit diskusi • 3 menit presentasi.

- 1 Tentukan 3 tipe pengguna dan kebutuhan informasinya.
- 2 Definisikan jenis dokumen yang akan diindeks.
- 3 Buat 5 contoh query yang realistis.
- 4 Tentukan 4 sinyal ranking yang relevan.
- 5 Jelaskan risiko jika ranking buruk.

OUTPUT: 1 lembar kanvas “User → Query → Documents → Ranking Signals → Success Criteria”

Kesalahpahaman Umum tentang Web IR

Memahami batasannya membantu kita tidak menyederhanakan masalah pencarian.



“Search = Ctrl+F”

Ctrl+F mencari string pada satu dokumen; IR mencari dan meranking banyak dokumen.



“Hasil #1 pasti paling benar”

Ranking hanya estimasi relevansi berdasarkan sinyal dan tujuan sistem.



“Semakin banyak hasil semakin baik”

Pengguna butuh hasil yang tepat, bukan jumlah hasil yang besar.



“Keyword match sudah cukup”

Makna, intent, link, quality, freshness, dan konteks juga penting.

Cek Pemahaman

Jawab singkat berdasarkan konsep. Diskusikan alasan di balik jawaban Anda.

1

1. Perbedaan utama DB retrieval dan IR adalah ...

2

2. Yang paling tepat menjelaskan relevance adalah ...

3

3. Komponen yang membuat koleksi dokumen searchable adalah ...

4

4. Web IR berbeda dari traditional IR terutama karena ...

5

5. Query “wikipedia web IR” paling dekat dengan intent ...

Gunakan 4 menit • jawab individual • lalu bandingkan dengan teman di sebelah Anda

Kunci & Pokok Jawaban

Gunakan setelah mahasiswa menyelesaikan quiz.

1

IR menghasilkan ranking berbasis derajat relevansi; DB retrieval umumnya exact match/boolean.

2

Kesesuaian hasil dengan kebutuhan informasi dalam konteks tertentu.

3

Indexer + index (mis. inverted index).

4

Skala, dinamika, link graph, heterogenitas, spam, dan konteks pengguna.

5

Navigational intent.

Enam Ide yang Harus Dibawa Pulang

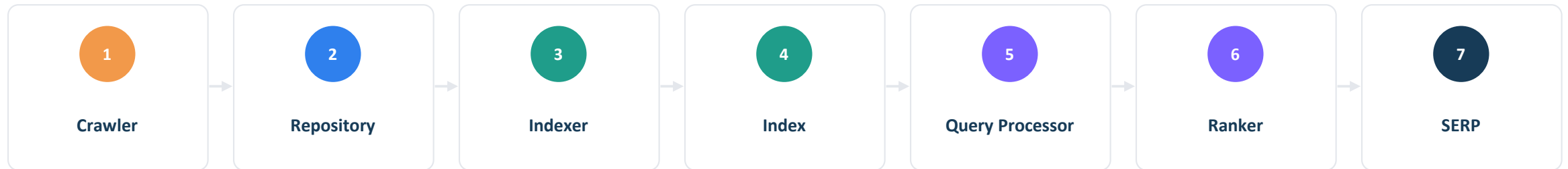
Pertemuan 1 membangun fondasi konseptual untuk memahami sistem pencarian web.

- 1 IR mencari informasi relevan, bukan sekadar string yang sama.
- 2 Query adalah representasi singkat dari information need.
- 3 Relevance bersifat kontekstual dan diestimasi oleh sistem.
- 4 Web IR menambahkan skala, link, dinamika, spam, dan konteks.
- 5 Search lifecycle: discover → fetch → process → index → query → rank → serve.
- 6 Hasil pencarian adalah keputusan ranking, bukan daftar acak.

FOUNDATION COMPLETE → NEXT: HOW THE SYSTEM IS BUILT

Pertemuan 2 — Arsitektur Sistem Web IR

Kita akan membedah komponen: crawler, repository, document processing, indexer, query processor, ranker, dan interface.



Persiapan sebelum kelas

Baca ringkas tentang search engine architecture dan identifikasi komponen yang berjalan offline vs online.

Pertanyaan pembuka

Mengapa mesin pencari tidak melakukan crawling pada saat pengguna menekan tombol Search?

Referensi Utama & Bacaan Lanjutan

Gunakan bab pengantar dan konsep retrieval sebagai bacaan pendamping.

Manning, Raghavan & Schütze

Introduction to Information Retrieval. Cambridge University Press. Fokus: konsep IR, indexing, ranking, evaluasi.

Baeza-Yates & Ribeiro-Neto

Modern Information Retrieval. Addison-Wesley. Fokus: web search, retrieval models, evaluation, web mining.

Croft, Metzler & Strohman

Search Engines: Information Retrieval in Practice. Fokus: search engine pipeline dan ranking.

Dokumentasi & studi kasus

Materi pendukung tentang mesin pencari, web crawling, ranking, dan search UX digunakan sebagai contoh kontekstual.

Catatan: detail algoritma crawling, indexing, ranking, dan evaluasi akan dibahas bertahap pada pertemuan berikutnya.