

# Web Mining & Link Analysis

Mata Kuliah Web Information Retrieval (3 SKS)

# Tujuan Pembelajaran



## Konsep Dasar

Memahami konsep fundamental Web Mining dan Link Analysis sebagai fondasi analisis struktur web modern



## Algoritma PageRank & HITS

Menjelaskan mekanisme kerja algoritma PageRank dan HITS dalam menentukan relevansi dan pentingnya halaman web



## Klikstream

Memahami penerapan analisis klikstream dalam Web Usage Mining untuk mengungkap pola perilaku pengguna



# Chapter 1: Pengenalan Web Mining

Web Mining adalah teknik canggih untuk mengekstraksi pengetahuan berharga dari data web yang sangat besar, kompleks, dan terus berubah secara dinamis. Dengan pertumbuhan eksponensial konten web, Web Mining menjadi kunci untuk menemukan pola tersembunyi dan informasi bermakna.

1

## Web Content Mining

Ekstraksi informasi dari konten halaman web seperti teks, gambar, dan multimedia

2

## Web Structure Mining

Analisis struktur link dan hubungan antar halaman web untuk memahami topologi web

3

## Web Usage Mining

Menganalisis pola interaksi pengguna dengan website melalui log server dan data klikstream

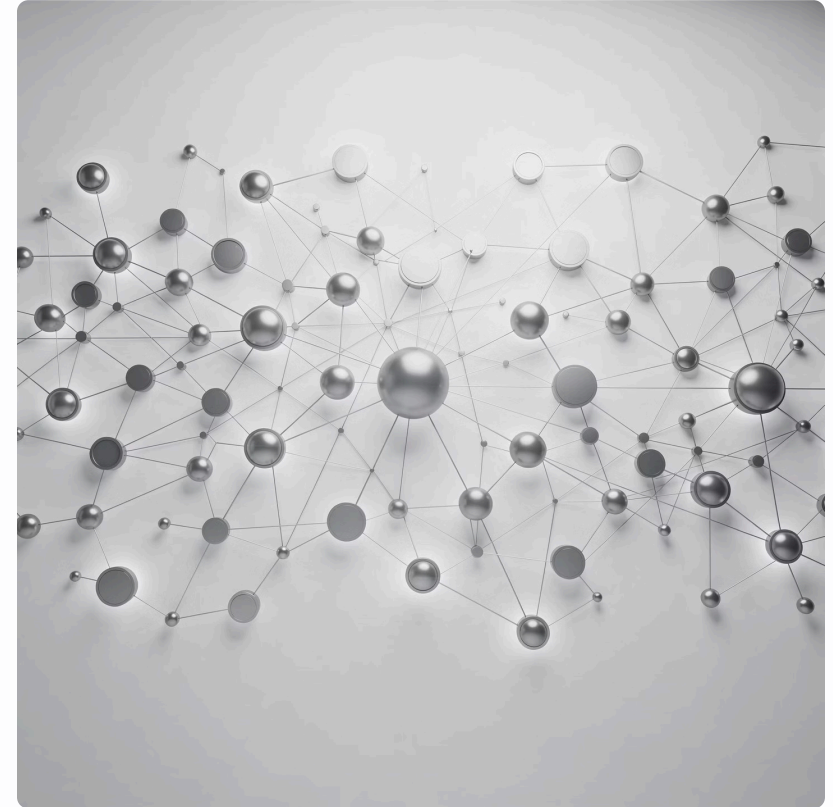
Fokus pembelajaran kita akan mendalam pada Web Structure Mining dan Web Usage Mining, yang menjadi dasar fundamental untuk memahami Link Analysis dan analisis klikstream dalam konteks pencarian informasi modern.

# Web sebagai Graph

Web dapat dipahami sebagai struktur graph raksasa yang terdiri dari miliaran halaman yang saling terhubung melalui hyperlink. Setiap halaman web bertindak sebagai node (simpul), sementara hyperlink yang menghubungkan satu halaman ke halaman lain berperan sebagai edge (sisi) dalam graph.

Konsep krusial dalam pemodelan ini adalah bahwa link bukan sekadar navigasi teknis, melainkan representasi dari "endorsement" atau rekomendasi implisit. Ketika sebuah halaman memberikan link ke halaman lain, ini dapat diinterpretasikan sebagai bentuk kepercayaan, relevansi, atau otoritas yang diberikan halaman sumber kepada halaman tujuan.

- Node merepresentasikan halaman web individual
- Edge merepresentasikan hyperlink yang menghubungkan halaman
- Arah link menunjukkan aliran informasi dan rekomendasi
- Struktur graph memungkinkan analisis matematis untuk ranking



# Chapter 2: Algoritma Link Analysis

Dalam era informasi digital, memberikan hasil pencarian yang relevan dan berkualitas tinggi kepada pengguna merupakan tantangan fundamental. Link Analysis hadir sebagai solusi dengan mengukur pentingnya atau otoritas sebuah halaman web berdasarkan struktur link yang menghubungkannya dengan halaman lain dalam ekosistem web.



## Struktur Link

Menganalisis pola hubungan antar halaman



## Perhitungan Skor

Menentukan pentingnya halaman secara objektif



## Hasil Relevan

Menyajikan informasi terbaik kepada pengguna

Dua algoritma yang telah merevolusi dunia pencarian informasi adalah **PageRank** dan **HITS**. Keduanya menggunakan pendekatan matematis yang berbeda namun saling melengkapi dalam menentukan kualitas dan relevansi halaman web.

# PageRank: Inti Algoritma Google

PageRank, yang diciptakan oleh Larry Page dan menjadi fondasi kesuksesan Google, menerapkan konsep inovatif "random surfer model" untuk menentukan pentingnya halaman web. Bayangkan seorang pengguna internet yang secara acak mengklik link dari satu halaman ke halaman lain, atau terkadang melompat (teleport) ke halaman yang sama sekali tidak terhubung.

01

---

## Random Surfer Model

Pengguna mengikuti link dengan probabilitas tertentu atau teleport ke halaman acak

02

---

## Distribusi Probabilitas

Menghitung kemungkinan pengguna berada di halaman tertentu setelah banyak klik

03

---

## Link Masuk Berkualitas

Halaman mendapat skor tinggi jika menerima link dari halaman yang sudah penting

04

---

## Teleportasi

Mengatasi masalah dead-end dan spider trap dengan faktor damping (biasanya 0.85)

Formula dasar PageRank:  $PR(A) = (1-d) + d \times \sum (PR(Ti)/C(Ti))$ , di mana  $d$  adalah damping factor,  $Ti$  adalah halaman yang memberikan link ke  $A$ , dan  $C(Ti)$  adalah jumlah link keluar dari  $Ti$ .



# Visualisasi PageRank

"PageRank menghitung pentingnya halaman berdasarkan kualitas dan kuantitas link masuk, mencerminkan struktur demokratis web di mana setiap link adalah 'suara' kepercayaan"

## Prinsip Kerja

- Setiap halaman memiliki skor PageRank awal yang sama
- Skor didistribusikan melalui link keluar
- Iterasi berlanjut hingga konvergensi tercapai
- Halaman dengan banyak link berkualitas mendapat skor tertinggi

## Karakteristik

- Independen dari query pencarian
- Dihitung secara global untuk seluruh web
- Stabil dan sulit dimanipulasi
- Menjadi salah satu dari 200+ faktor ranking Google



# HITS (Hyperlink-Induced Topic Search)

HITS (Hyperlink-Induced Topic Search) adalah algoritma yang dikembangkan oleh Jon Kleinberg dengan pendekatan yang berbeda dari PageRank. HITS memisahkan peran halaman web menjadi dua kategori fundamental: **Hub** dan **Authority**, menciptakan ekosistem saling mendukung dalam struktur web.

## Authority (Otoritas)

Halaman yang mengandung informasi berkualitas tinggi dan menjadi referensi utama dalam topik tertentu. Authority mendapat banyak link masuk dari Hub yang relevan.

- Sumber informasi berkualitas
- Banyak link masuk dari Hub
- Contoh: Wikipedia, jurnal akademik

## Hub (Penghubung)

Halaman yang berfungsi sebagai direktori atau kolektor, memberikan banyak link keluar menuju Authority yang berkualitas dalam domain tertentu.

- Mengumpulkan link ke Authority
- Banyak link keluar berkualitas
- Contoh: direktori topik, portal berita

HITS sangat cocok untuk pencarian topik spesifik dan analisis komunitas web karena dapat mengidentifikasi struktur otoritas dalam domain tertentu. Algoritma ini menghitung skor Authority dan Hub secara iteratif hingga konvergen.



# Perbandingan PageRank & HITS

## Ruang Lingkup

**PageRank:** Bersifat global, dihitung untuk seluruh web secara independen dari query pencarian

**HITS:** Bersifat query-dependent, dihitung untuk subset halaman relevan dengan topik tertentu

## Stabilitas

**PageRank:** Lebih stabil dan tahan terhadap manipulasi spam link karena bersifat global

**HITS:** Lebih rentan terhadap manipulasi link lokal dan struktur komunitas kecil

## Kompleksitas

**PageRank:** Komputasi lebih efisien karena dihitung sekali untuk seluruh web

**HITS:** Memerlukan komputasi ulang untuk setiap query, lebih intensif secara komputasi

## Penerapan

**PageRank:** Ideal untuk ranking global dan mesin pencari umum seperti Google

**HITS:** Sangat baik untuk pencarian topik spesifik dan analisis struktur komunitas web

Dalam praktiknya, sistem pencarian modern sering mengkombinasikan kekuatan kedua algoritma ini bersama dengan ratusan faktor lain untuk menghasilkan hasil pencarian yang paling relevan dan berkualitas bagi pengguna.

# Chapter 3: Klikstream dan Web Usage Mining

Klikstream adalah rekaman komprehensif dari jejak digital yang ditinggalkan pengguna saat bernavigasi di web. Setiap klik, setiap perpindahan halaman, setiap interaksi terekam dan membentuk narasi perilaku pengguna yang sangat berharga untuk analisis.

- 1 — Landing Page  
Halaman pertama yang dikunjungi
- 2 — Navigasi  
Urutan halaman yang dijelajahi
- 3 — Interaksi  
Klik, scroll, form submission
- 4 — Konversi/Exit  
Tindakan akhir atau keluar

Data klikstream membantu organisasi memahami perilaku pengguna secara mendalam: pola kunjungan, durasi waktu di setiap halaman, urutan navigasi, titik keluar, dan banyak lagi. Analisis ini menjadi fundamental dalam Web Usage Mining untuk mengoptimalkan pengalaman pengguna dan strategi bisnis digital.





# Manfaat Klikstream



## Personalisasi Konten

Menggunakan pola klikstream untuk menyajikan konten yang disesuaikan dengan preferensi individual pengguna. Sistem rekomendasi dapat memprediksi produk atau artikel yang mungkin menarik berdasarkan riwayat klik sebelumnya.



## Optimasi Navigasi

Mengidentifikasi jalur navigasi yang paling efisien dan titik-titik kesulitan dalam struktur website. Memperbaiki arsitektur informasi berdasarkan perilaku aktual pengguna untuk meningkatkan usability.



## Prediksi Perilaku

Membangun model prediktif untuk mengantisipasi tindakan pengguna berikutnya. Mendukung strategi pemasaran dengan menargetkan pengguna pada waktu dan konteks yang tepat untuk memaksimalkan konversi.

# Studi Kasus Klikstream

## Analisis E-Commerce: Meningkatkan Konversi Penjualan



Sebuah platform e-commerce besar menganalisis data klikstream dari jutaan pengguna untuk mengoptimalkan proses pembelian dan meningkatkan conversion rate.

### Temuan Kunci

- 70% pengguna meninggalkan keranjang belanja di halaman pembayaran
- Halaman produk dengan video mendapat engagement 3x lebih tinggi
- Pengguna yang membaca review menghabiskan 45% lebih banyak

#### Identifikasi Pola

Menganalisis jalur pengguna dari landing hingga checkout, menemukan titik drop-off kritis

#### Implementasi Perbaikan

Menyederhanakan proses checkout, menambah trust signals, memperbaiki loading time

#### Hasil

Conversion rate meningkat 28%, cart abandonment turun 15%, customer satisfaction naik signifikan

# Integrasi Web Mining & Link Analysis

Kekuatan sejati muncul ketika kita mengintegrasikan berbagai teknik Web Mining. Kombinasi analisis link struktural dengan data perilaku pengguna menghasilkan sistem pencarian dan rekomendasi yang jauh lebih powerful dan akurat.



- ❏ **Contoh Nyata:** Google menggabungkan PageRank dengan data klik pengguna (click-through rate, dwell time, bounce rate) untuk menyempurnakan algoritma ranking. Halaman dengan skor PageRank tinggi namun CTR rendah akan diturunkan rankingnya, sementara halaman dengan engagement tinggi akan dipromosikan.

# Kesimpulan & Aksi

## Web Mining & Link Analysis: Fondasi Informasi Digital

Web Mining dan Link Analysis adalah teknologi krusial yang memungkinkan kita mengelola dan menemukan informasi bermakna di tengah lautan data web yang terus berkembang. Tanpa teknik-teknik ini, web modern tidak akan dapat berfungsi secara efektif.

## PageRank & HITS: Dua Pilar Ranking

Algoritma PageRank dan HITS telah merevolusi cara kita menentukan relevansi dan otoritas halaman web. Keduanya menawarkan perspektif berbeda namun komplementer dalam mengukur pentingnya informasi di web.

## Klikstream: Jendela Perilaku Pengguna

Analisis klikstream memberikan wawasan mendalam tentang bagaimana pengguna berinteraksi dengan web, memungkinkan personalisasi, optimasi, dan prediksi yang lebih akurat untuk pengalaman pengguna yang superior.

---

# Saatnya Beraksi!

Jangan berhenti di teori—eksplorasi lebih dalam konsep-konsep ini melalui implementasi praktis. Mulai proyek analisis link sederhana, eksperimen dengan data klikstream, atau bangun sistem rekomendasi mini. Pengalaman langsung adalah guru terbaik dalam memahami Web Mining & Link Analysis!

**Mulai Proyek**

**Eksplorasi Dataset**