Temu-Kembali Informasi

2018

02: Arsitektur Search Engine

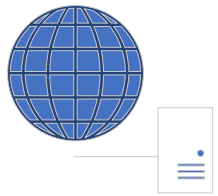
Versi Ringkas ++

Arsitektur Software

- Arsitektur Software merujuk ke struktur tingkat tinggi dari suatu sistem perangkat lunak.
- Struktur ini diperlukan untuk menjelaskan tentang sistem perangkat lunak.
- Setiap struktur terdiri dari elemen perangkat lunak, hubungan di antara mereka, dan properti dari elemen dan relasi tersebut.
- [Wikipedia]

Contoh 1:

Arsitektur Search Engine



Proses Indexing

Data Storage

Proses Pencarian (*Search*)

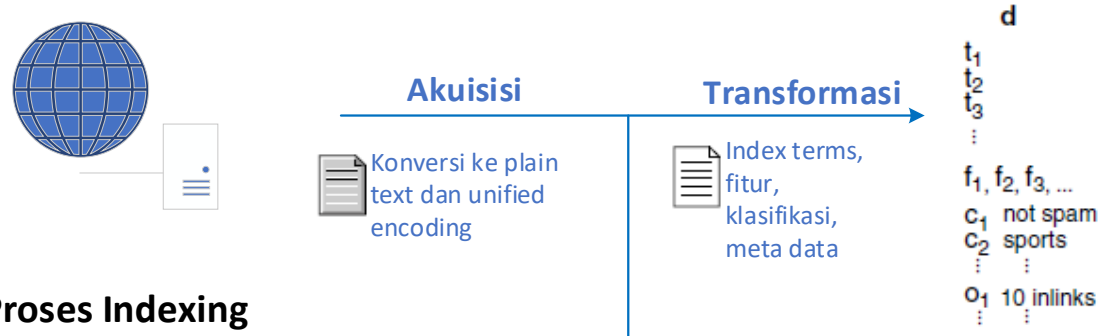
Akuisisi

Konversi ke plain text dan unified encoding

Proses Indexing

Data Storage

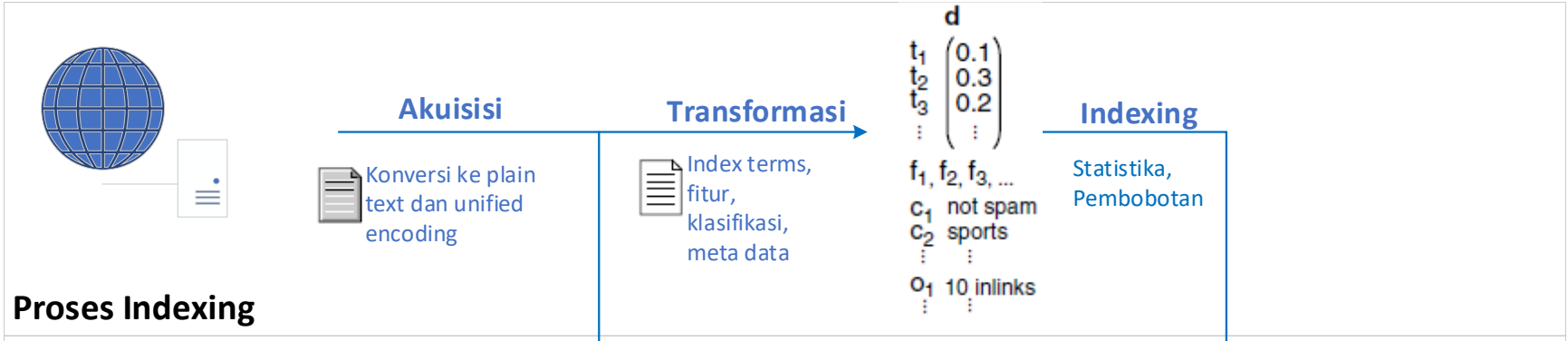
Proses Pencarian (*Search*)



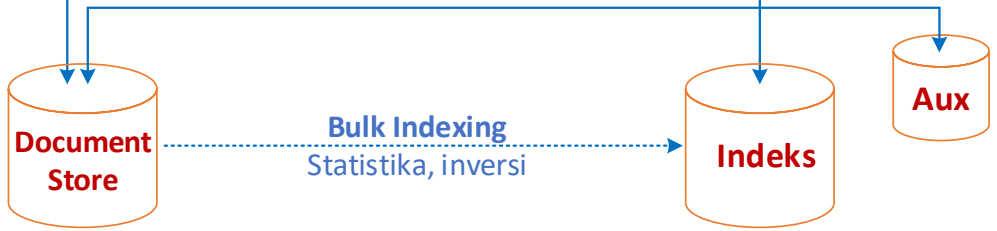
Proses Indexing

Data Storage

Proses Pencarian (*Search*)

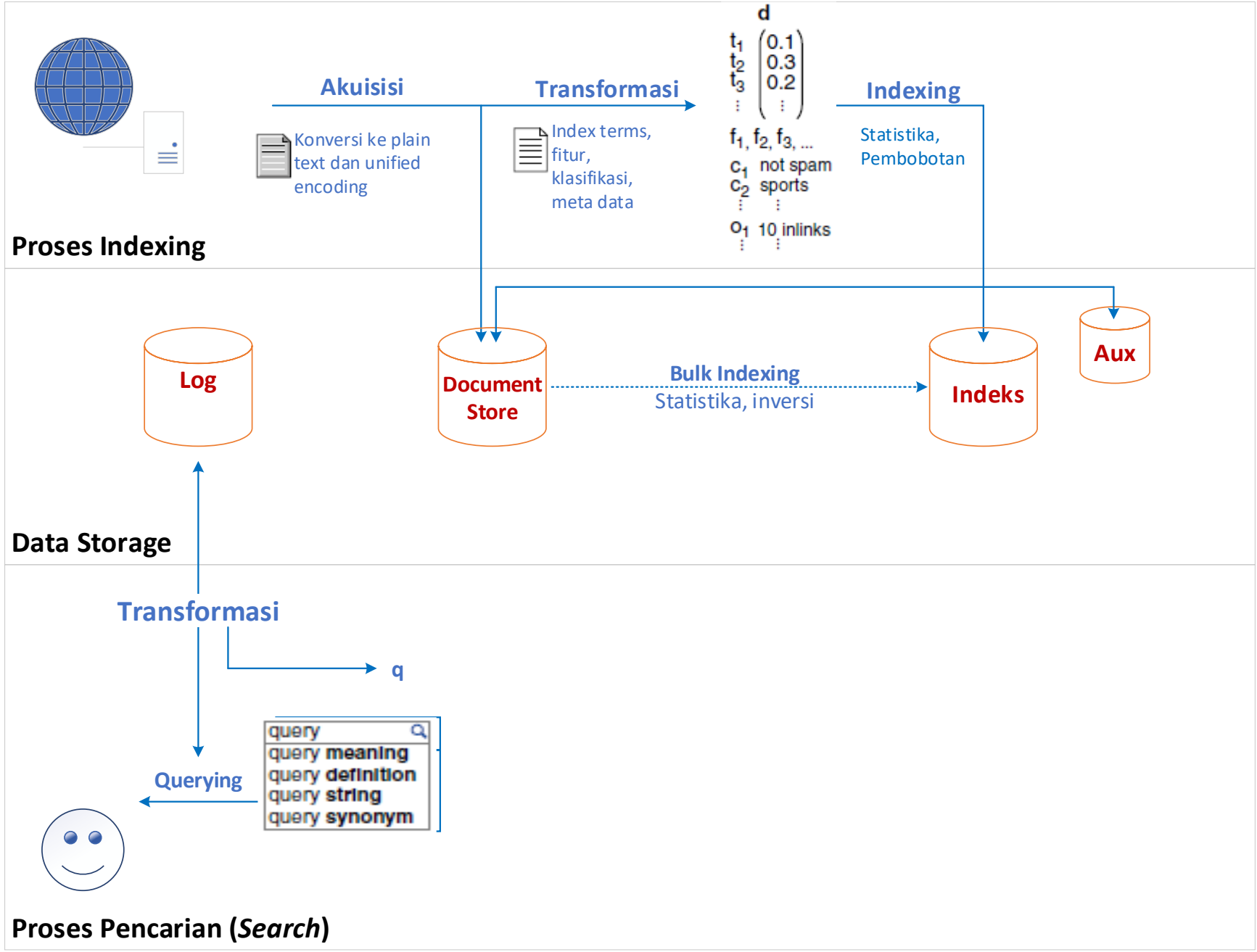


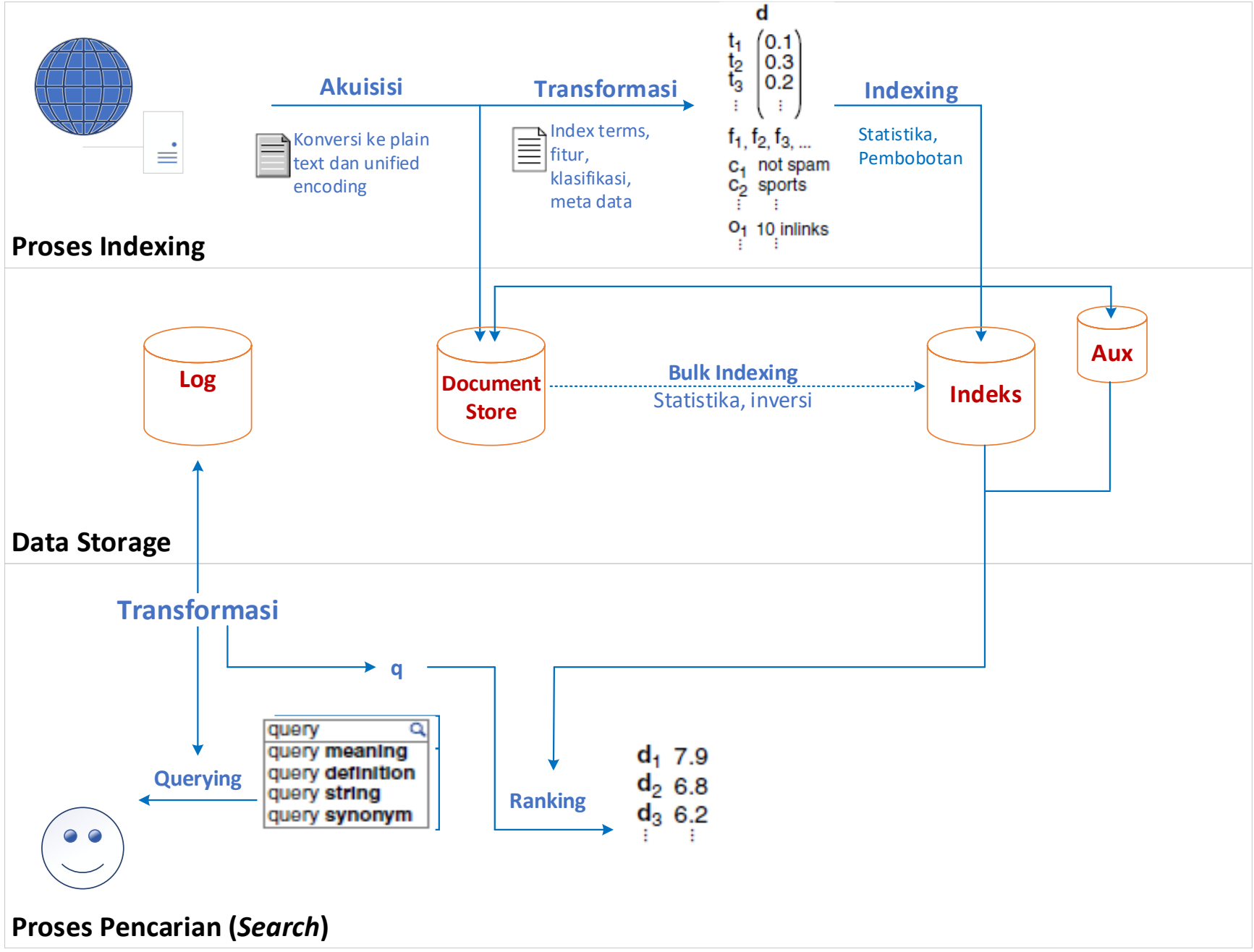
Proses Indexing

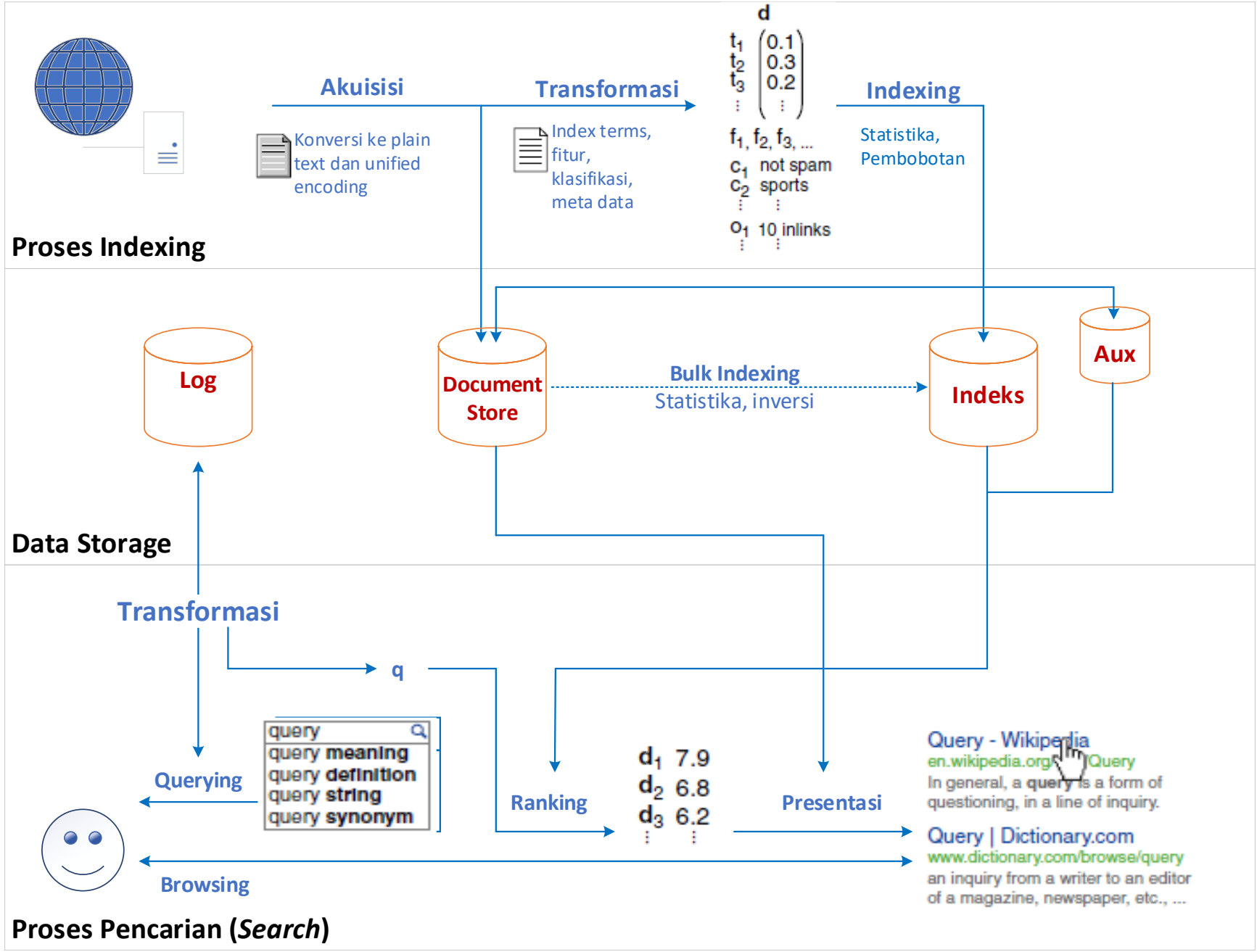


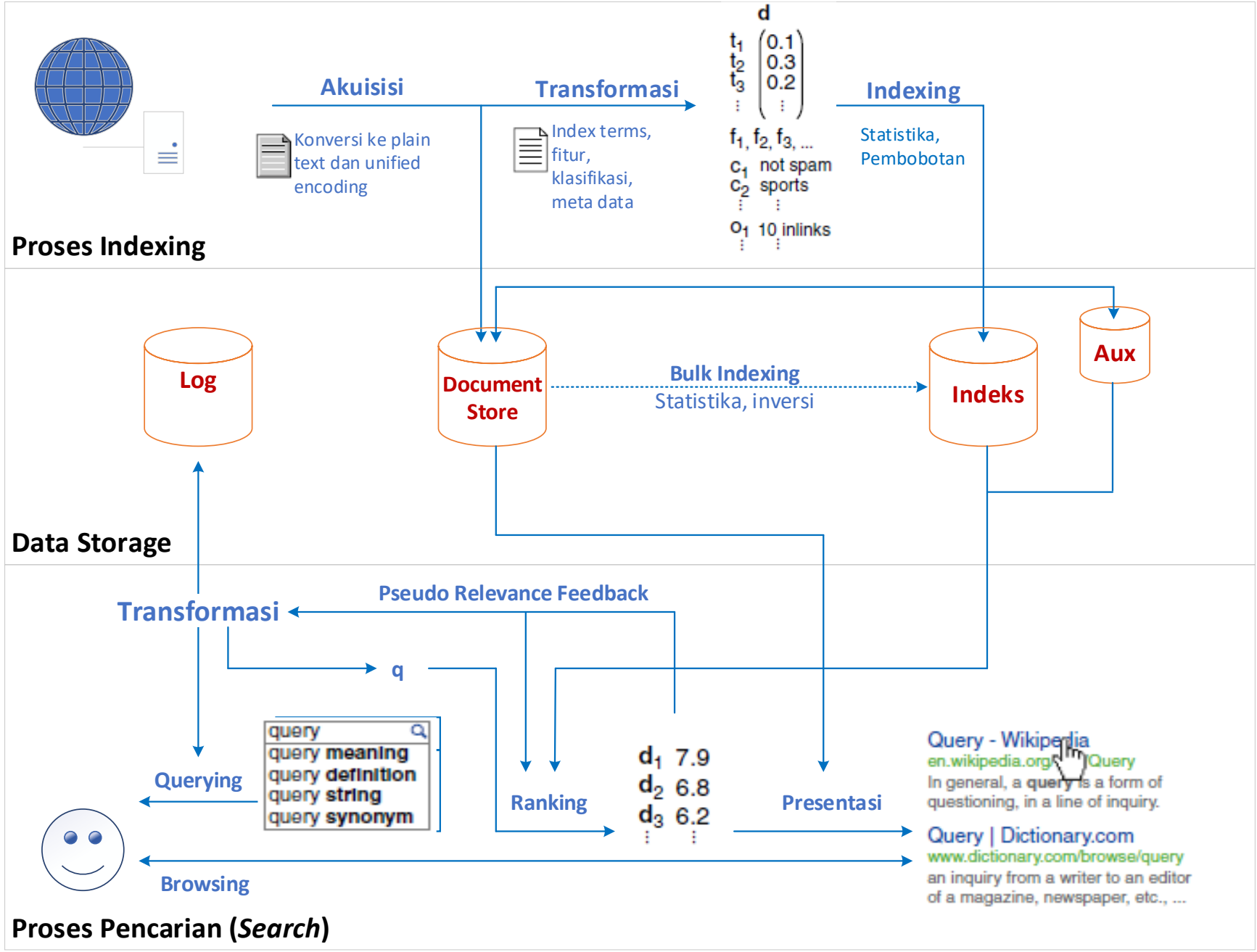
Data Storage

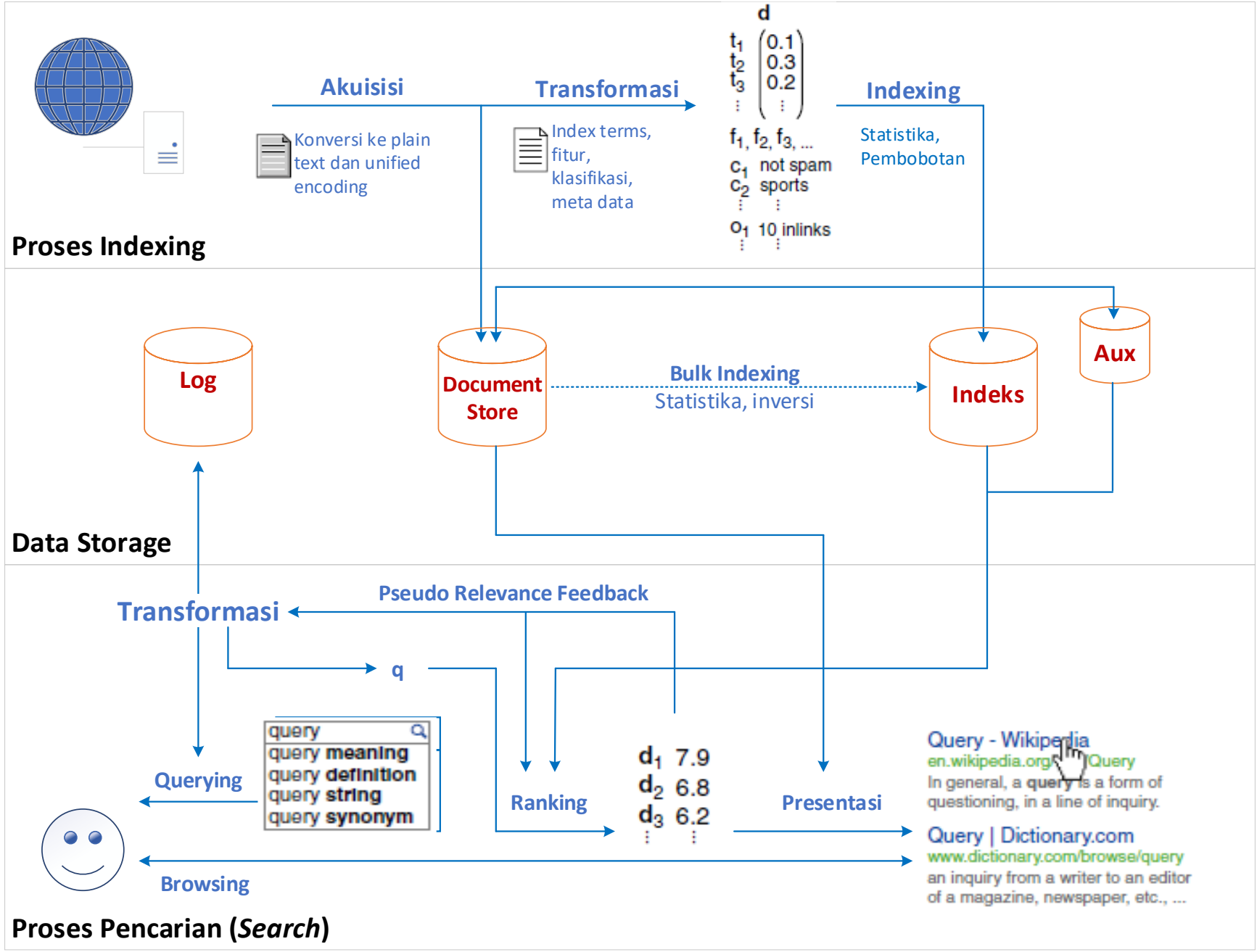
Proses Pencarian (*Search*)

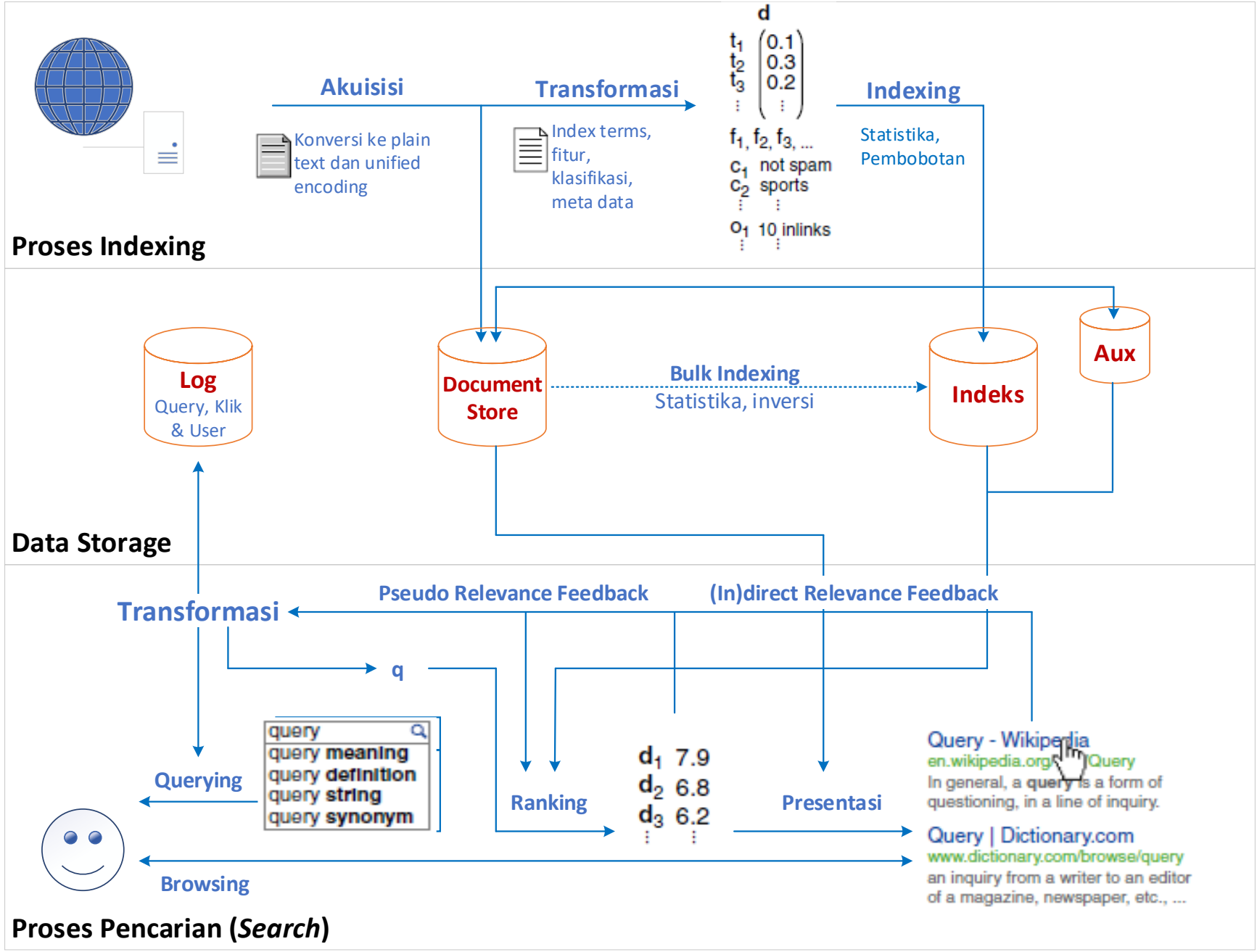

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
q
query
query meaning
query definition
query string
query synonym
Querying
Proses Pencarian (Search)

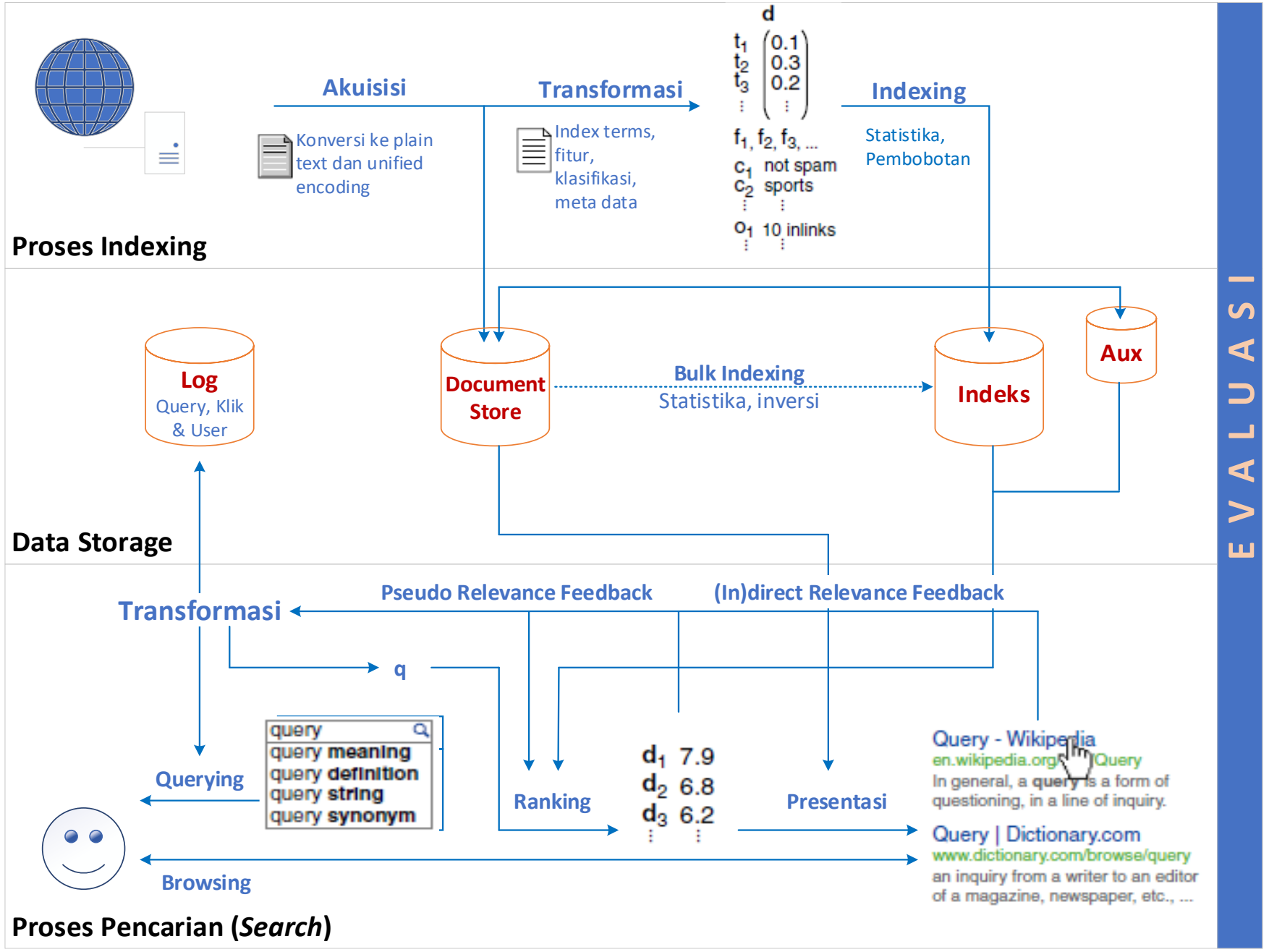

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Proses Pencarian (Search)

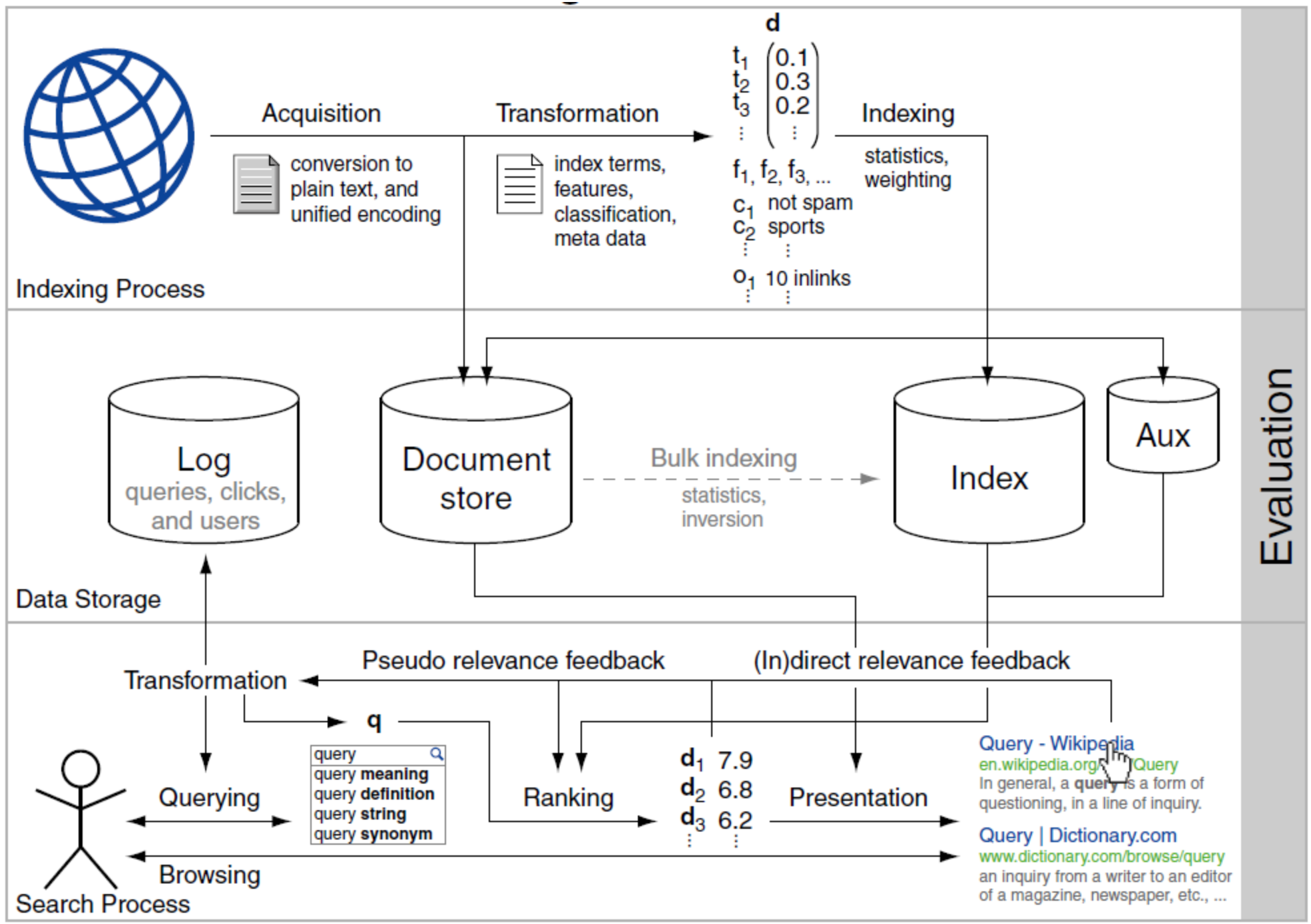

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentasi
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...
Browsing
Proses Pencarian (Search)


Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
Pseudo Relevance Feedback
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentasi
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...
Browsing
Proses Pencarian (Search)


Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
Pseudo Relevance Feedback
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentasi
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...
Browsing
Proses Pencarian (Search)


Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Query, Klik & User
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
Pseudo Relevance Feedback
(In)direct Relevance Feedback
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentasi
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...
Browsing
Proses Pencarian (Search)


Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Query, Klik & User
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data Storage
Transformasi
Pseudo Relevance Feedback
(In)direct Relevance Feedback
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentasi
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...
Browsing
Proses Pencarian (Search)
EVALUASI

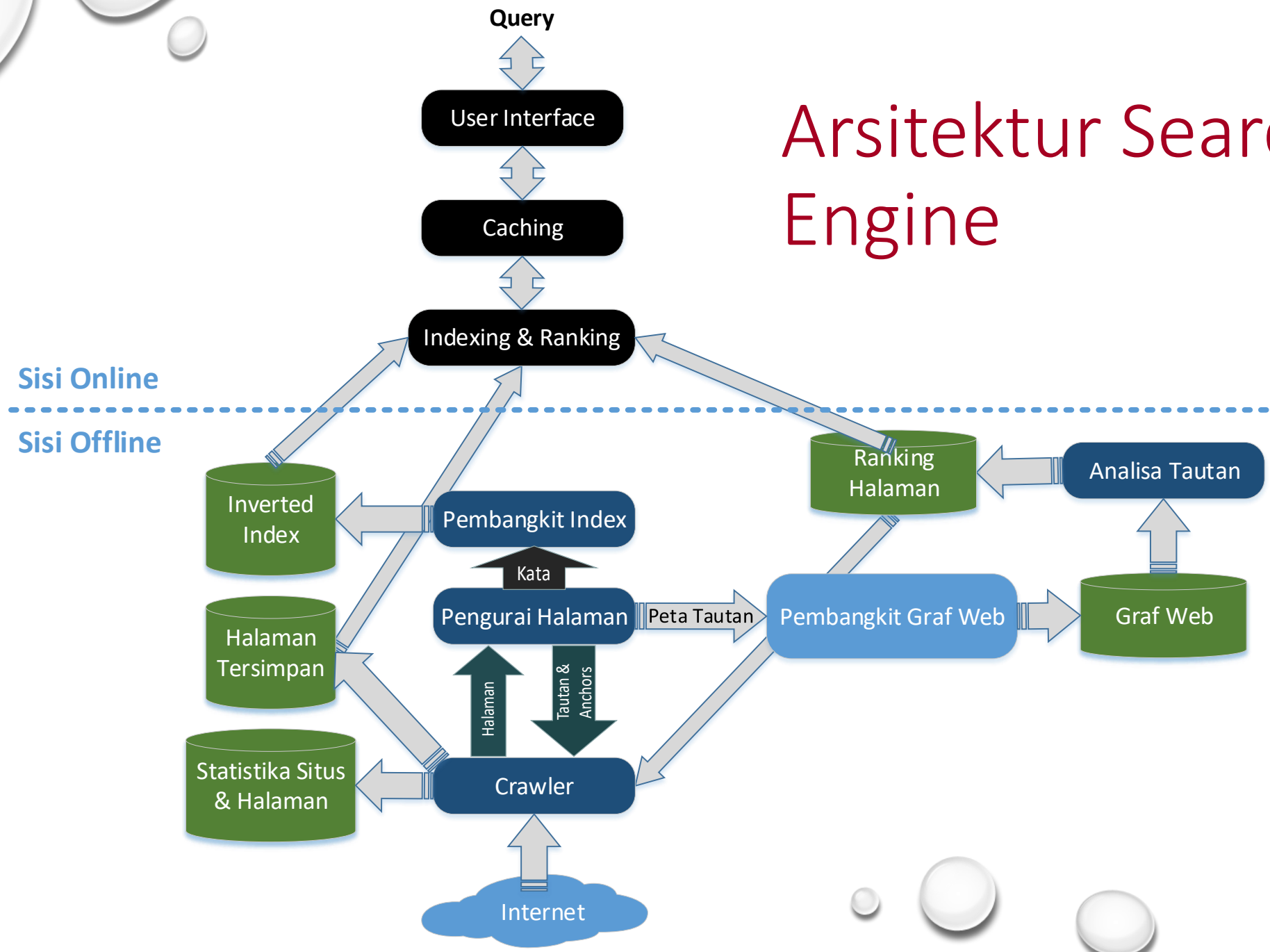

Acquisition
conversion to plain text, and unified encoding
Transformation
index terms, features, classification, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
statistics, weighting
Indexing Process
Log
queries, clicks, and users
Document store
Bulk indexing
statistics, inversion
Index
Aux
Data Storage
Evaluation
Transformation
Pseudo relevance feedback
(In)direct relevance feedback
q
query
query meaning
query definition
query string
query synonym
Querying
Ranking
d1 7.9
d2 6.8
d3 6.2
Presentation
Browsing
Search Process
Query - Wikipedia
en.wikipedia.org/wiki/Query
In general, a query is a form of questioning, in a line of inquiry.
Query | Dictionary.com
www.dictionary.com/browse/query
an inquiry from a writer to an editor of a magazine, newspaper, etc., ...

Contoh 2:

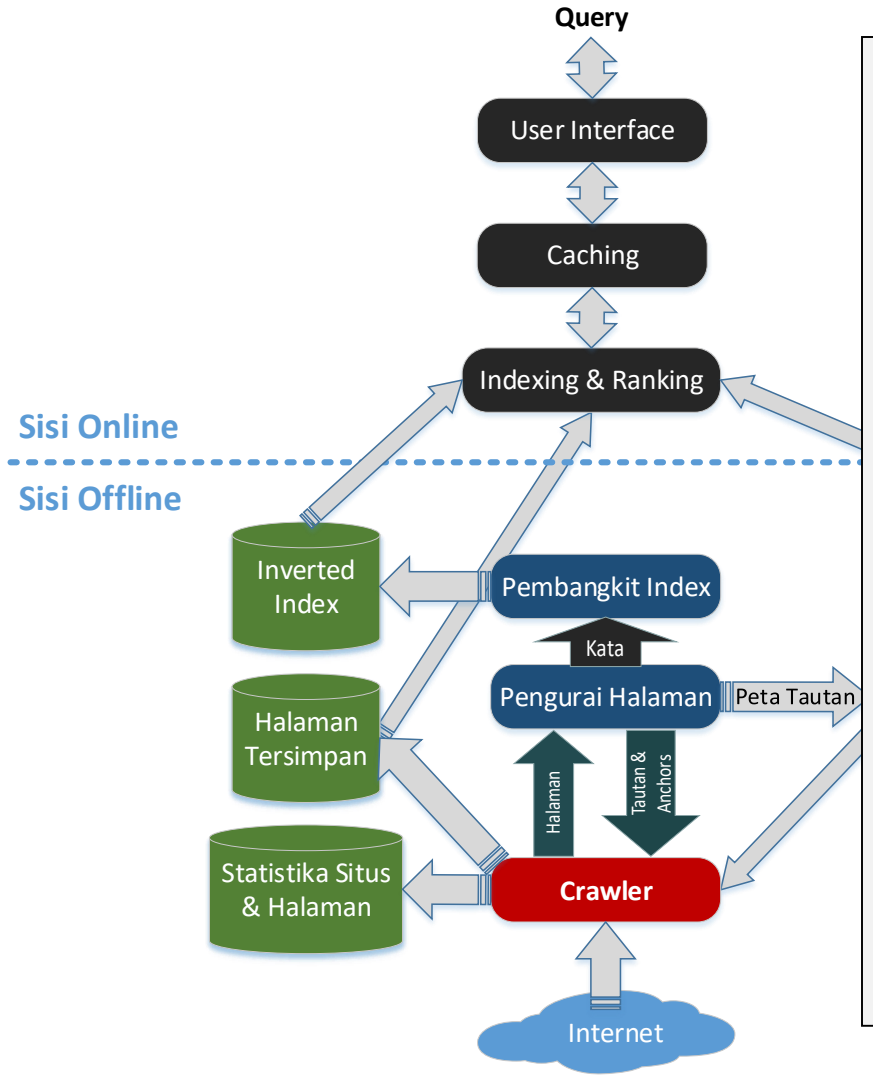
Arsitektur Search Engine

dari Microsoft Research

Arsitektur Search Engine

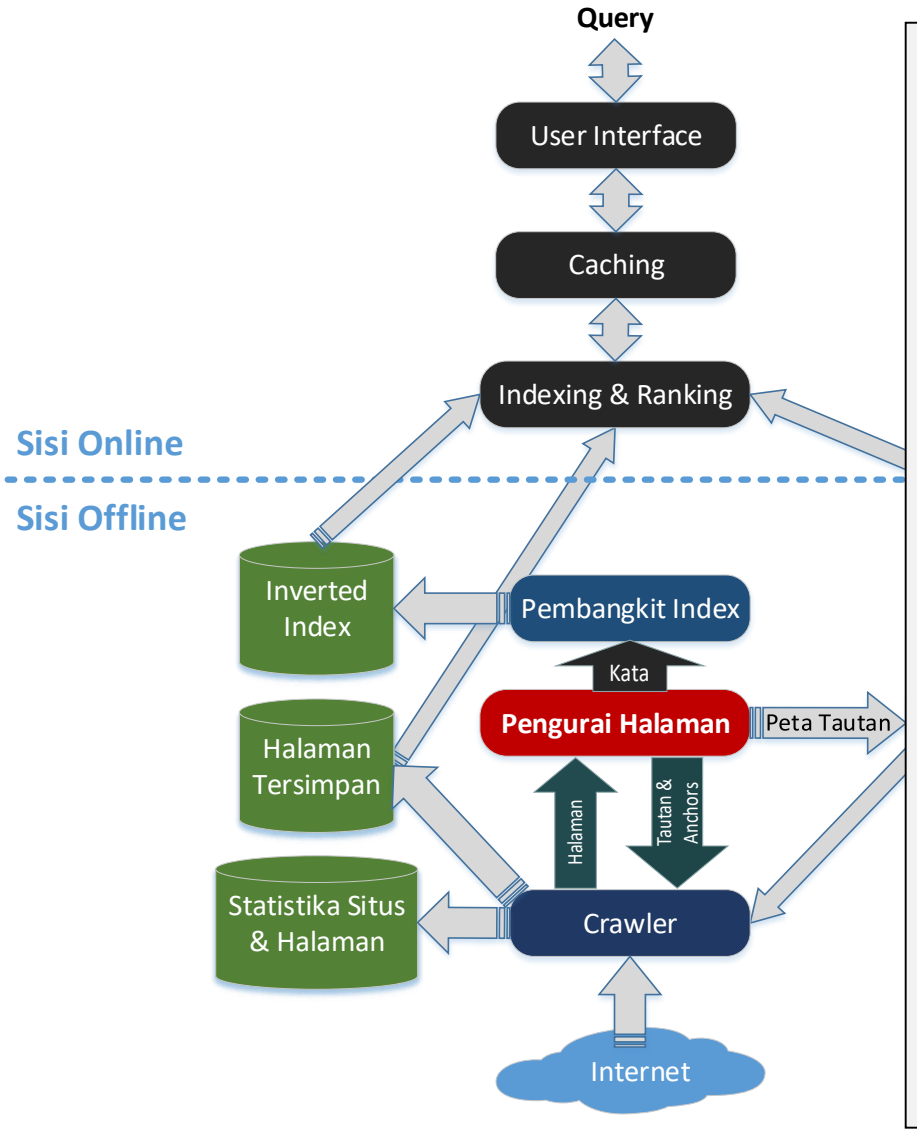


Arsitektur: Crawler



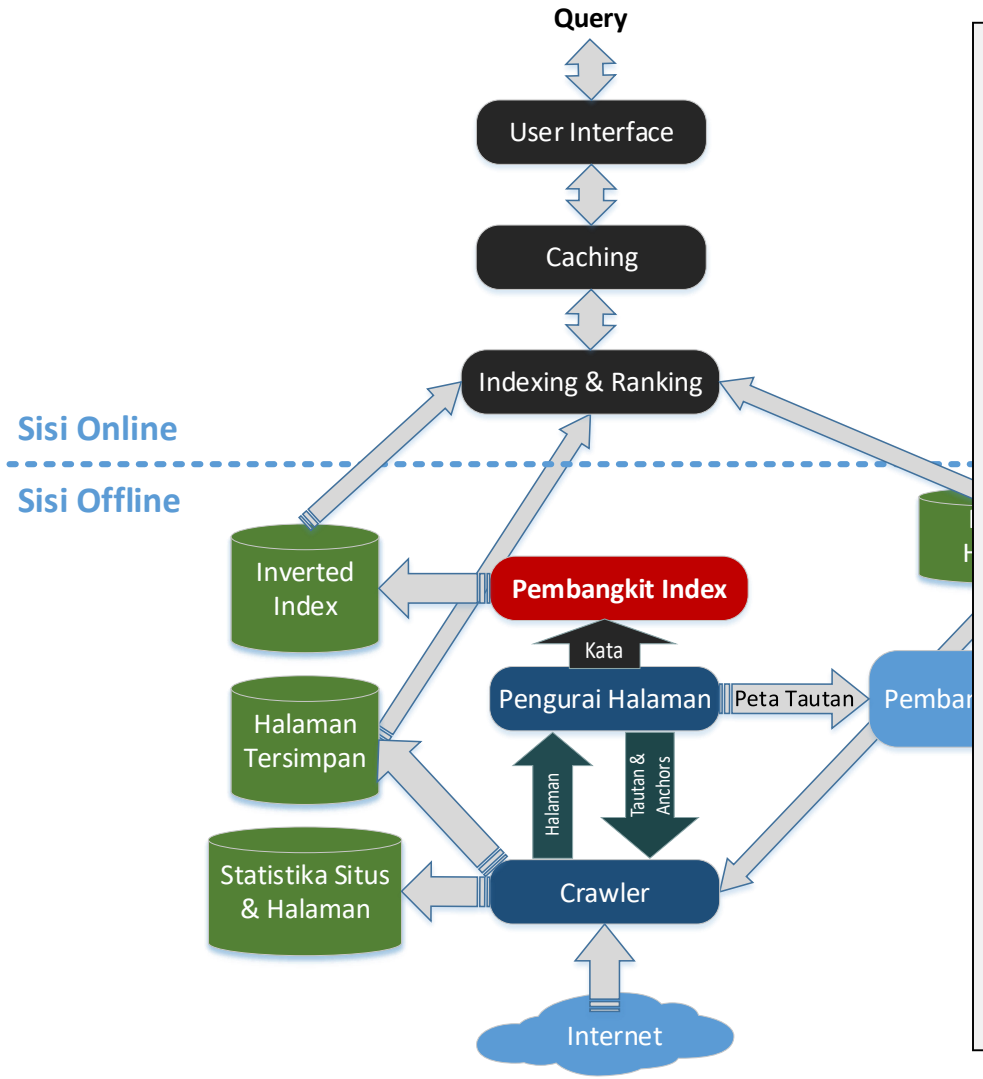
- Fungsi
  - Mengambil (Fetch) halaman web dengan mengikuti hyperlink
  - Me-refresh halaman secara periodik
- Masalah Inti
  - Bandwidth & storage terbatas vs. volume data sangat besar
  - Frekuensi update halaman
- Solusi
  - Prioritaskan crawling berdasarkan pada ranking halaman dan statistik lain

Arsitektur: Page Parser

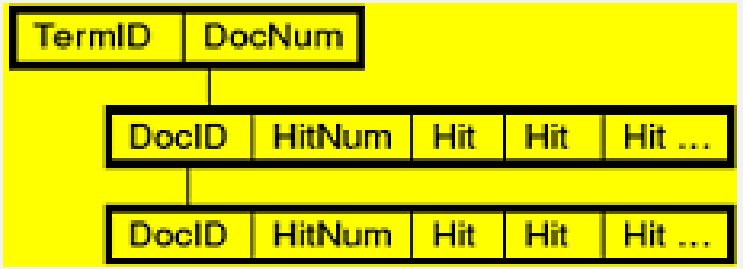


- Fungsi
  - Mengekstrak aliran data untuk indexing
    a. Title: kata-kata dalam <title>…</title>
    b. URL
    c. Body
      - Teks Anchor
      - Teks Plain
      - H1…6
      - Bold, Italic, etc
      - Large, Medium, Small
  - Membangun peta link parsial
  - Mengirim hyperlink yang ditemukan ke crawler
- Masalah Inti
  - Fitur apa yang akan diekstrak?

Arsitektur: Index Builder



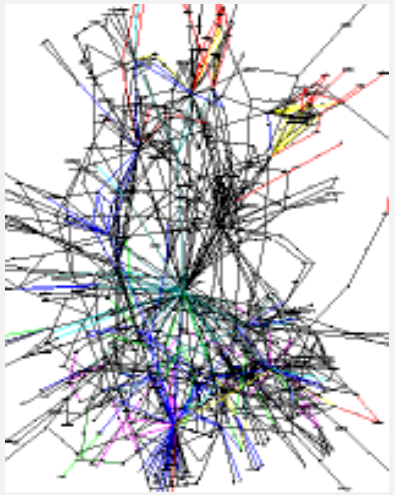
- Fungsi
  - Membangun inverted index berdasarkan pada data halaman yang telah diparse



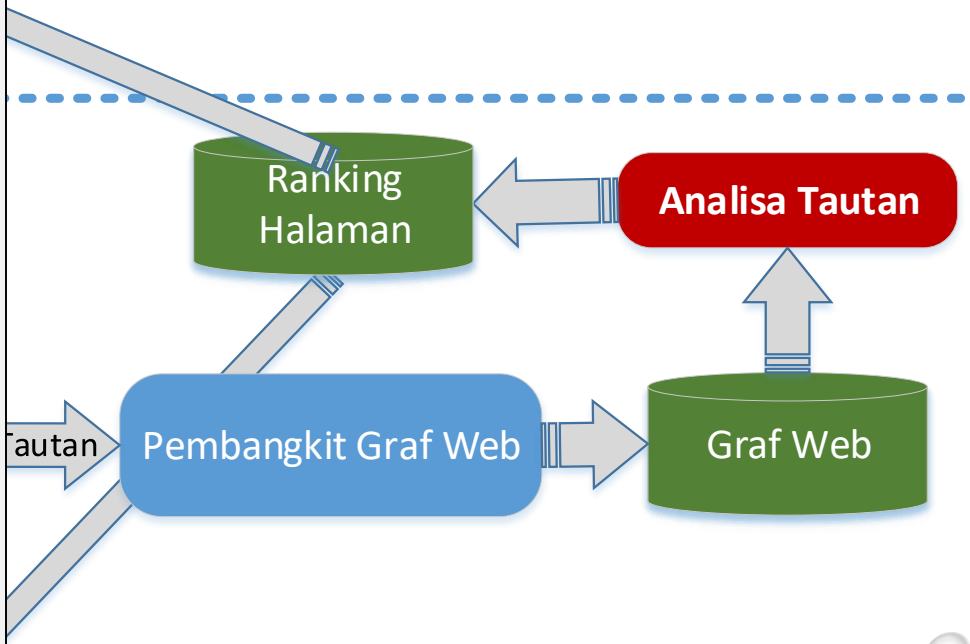
- Masalah Inti
  - Efisiensi vs. memory terbatas & terdistribusi
- Solusi
  - Indexing terdistribusi
  - Partisi berdasarkan dokumen, bukan partisi berdasarkan term

Arsitektur: Link Analysis

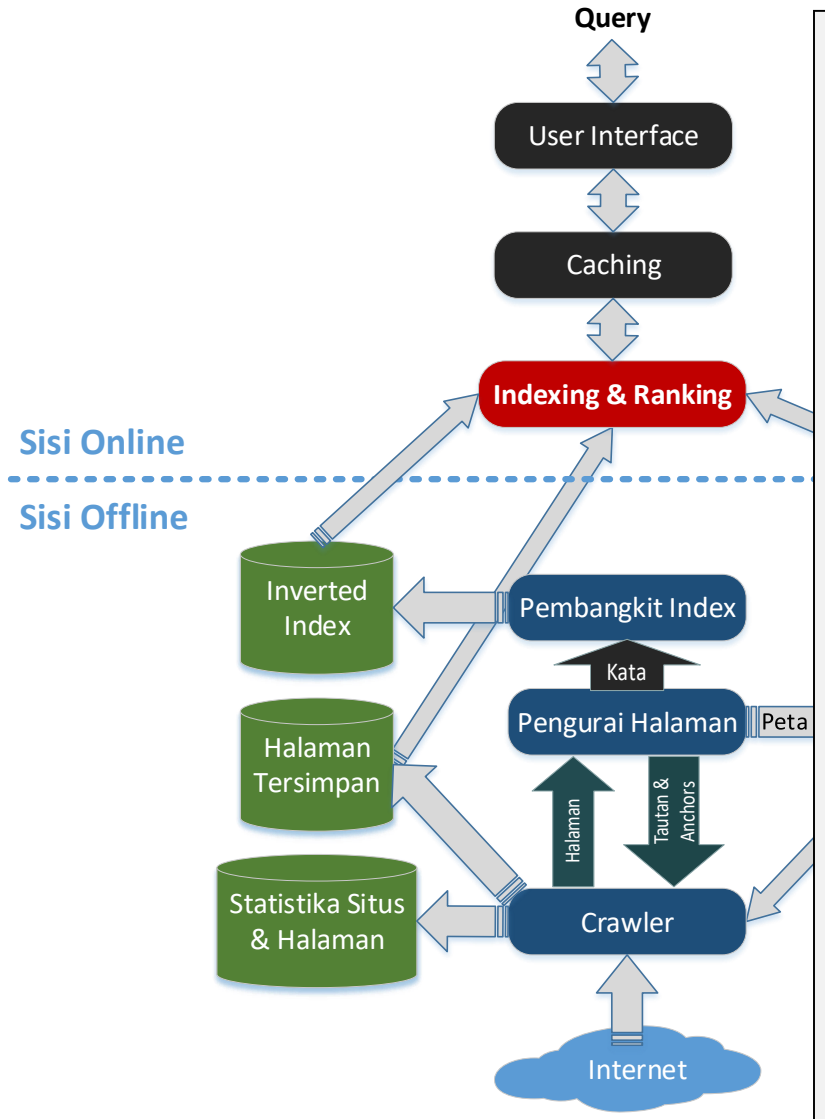
- Fungsi
  - Mengukur kualitas atau otoritas dari suatu halaman berdasarkan pada graf link



- Masalah Inti
  - Algoritma yang efisien pada graf raksasa
  - Link-spam?
  - Apakah hanya link analysis cara untuk menentukan qualitas dari halaman?

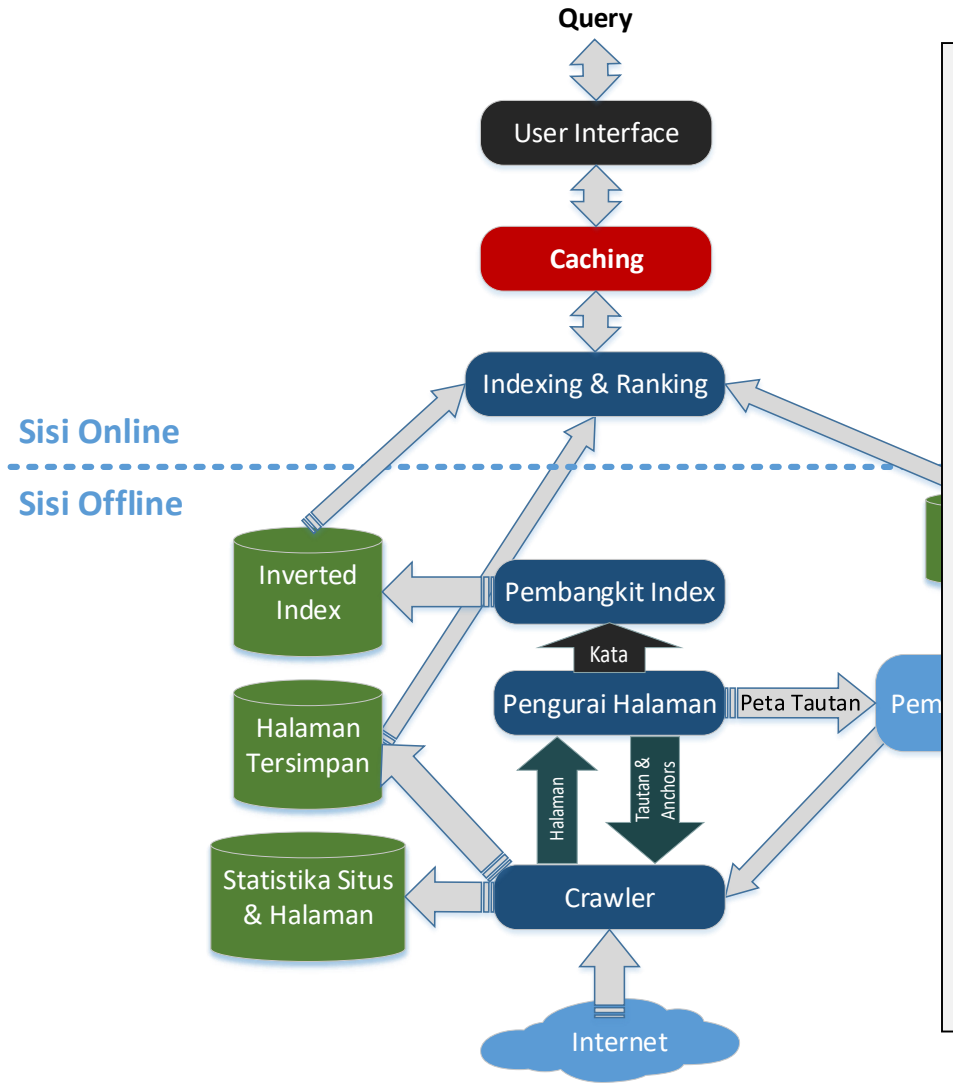


Arsitektur: Indexing & Ranking



- Masalah utama dalam komunitas IR dan telah dikaji puluhan tahun
- Fungsi
  - Indexing: dengan cepat menemukan halaman yang mengandung term query
  - Ranking: mengurutkan halaman sesuai dengan relevansi terhadap query
- Masalah Inti
  - Kinerja: inverted list untuk suatu term hot mungkin ratusan megabyte.
  - Akurasi: fungsi ranking dengan ratusan parameter:
    - Teks Anchor
    - Ranking halaman
    - Term proximity
    - TF*IDF
    - …
- Solusi
  - Kinerja: Top-K query & index pruning
  - Akurasi: Tuning atau learning?

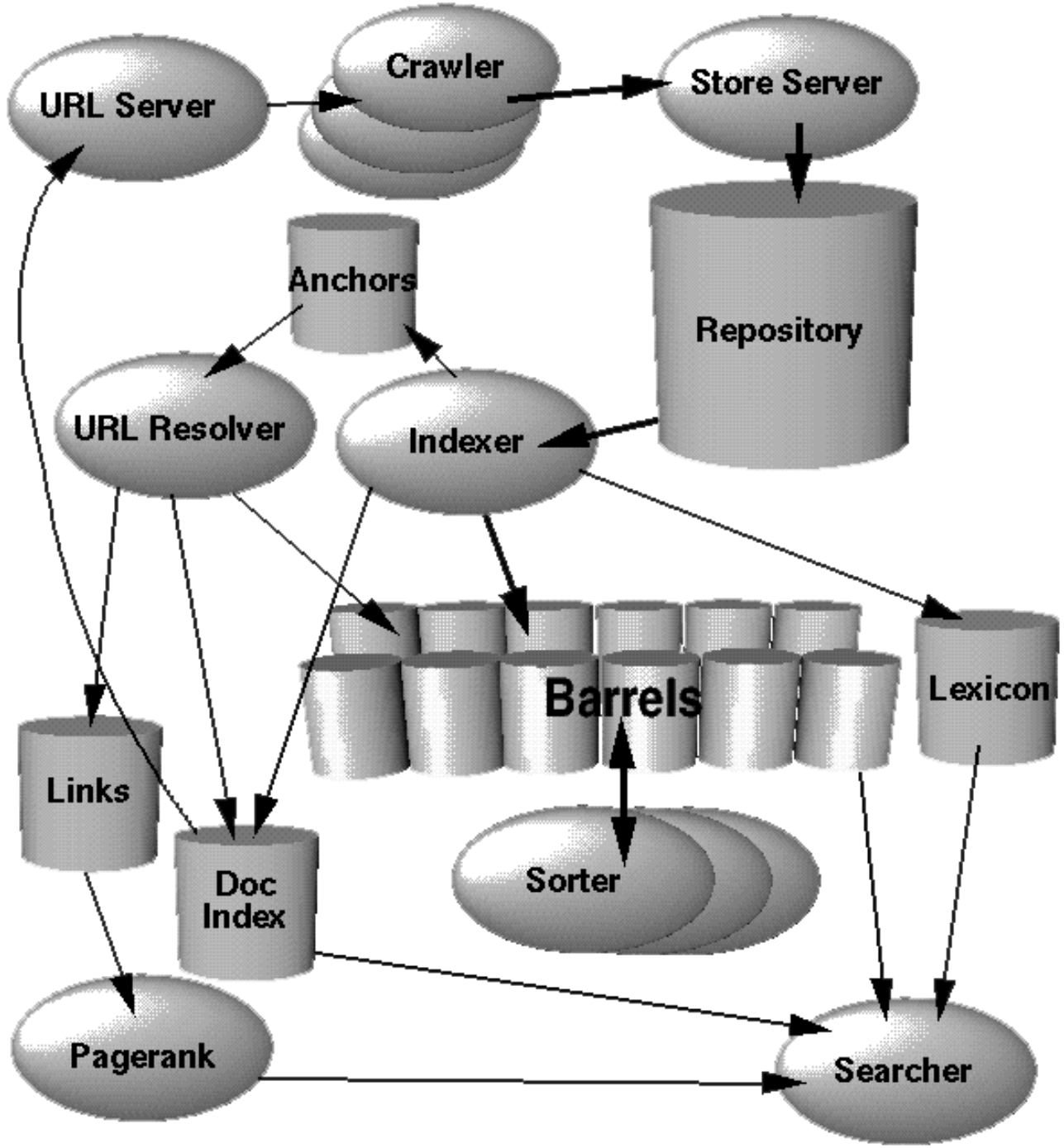
Arsitektur: Caching



- Fungsi
  - Men-cache hasil dari query yang sering untuk menjawab ribuan query per detik dengan waktu respon interaktif
- Masalah Inti
  - Apa yang dicache?
- Solusi
  - Caching banyak level
    - Level Query
    - Level Term

Contoh Lain Arsitektur Search Engine

Search Engine Google



Arsitektur Search Engine

Crawler

Store

Indexer

trash

trash

trash

Algorithms (Programs)

Search Interface

Query

Results

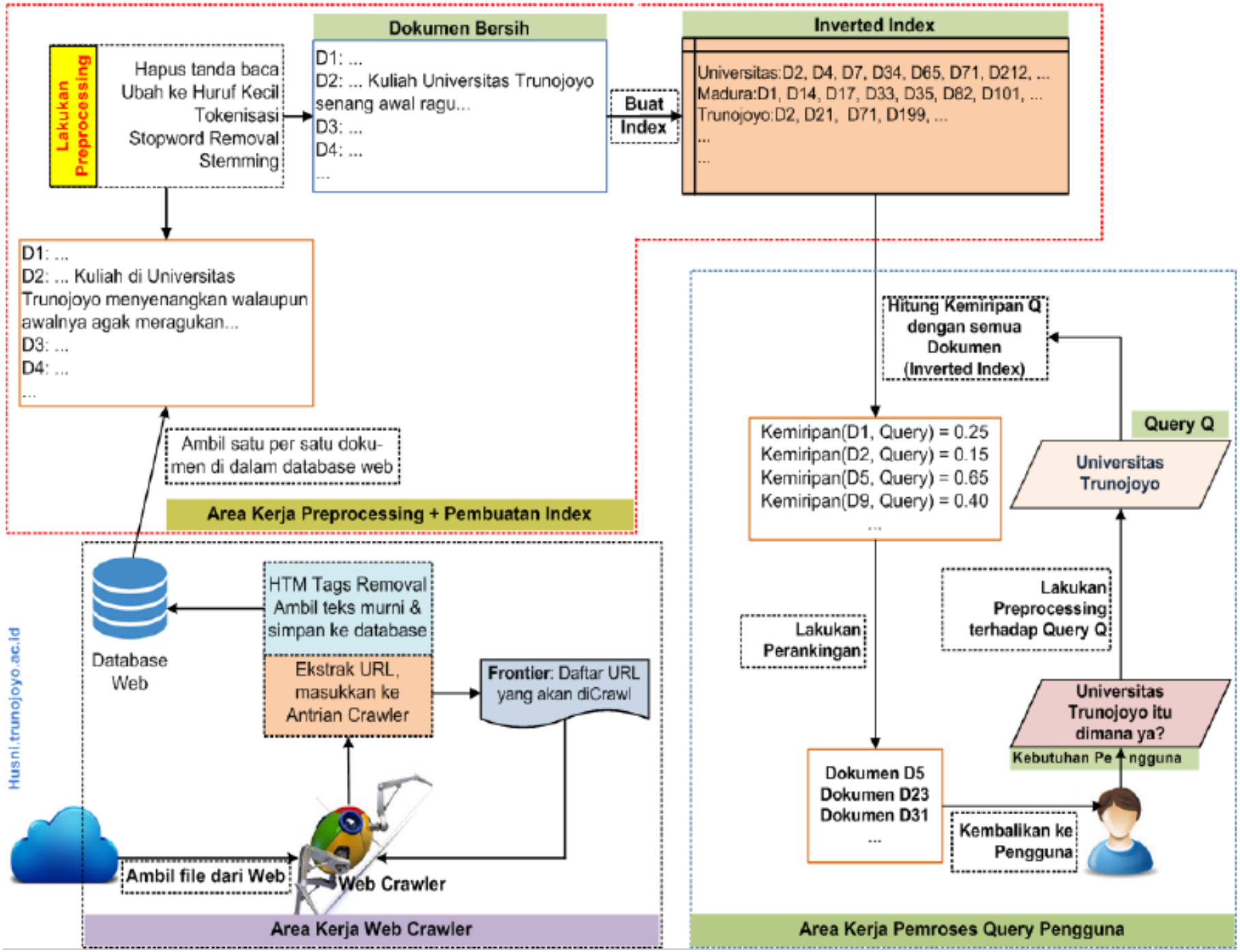
60 Trillion Pages
Or
60 Lakh Crore

indexes

100 Million GB

indexes

Sorted based on Content / Factors


Lakukan Preprocessing
Hapus tanda baca
Ubah ke Huruf Kecil
Tokenisasi
Stopword Removal
Stemming
Dokumen Bersih
D1: ...
D2: ... Kuliah Universitas Trunojoyo senang awal ragu...
D3: ...
D4: ...
...
Buat Index
Inverted Index
Universitas:D2, D4, D7, D34, D65, D71, D212, ...
Madura:D1, D14, D17, D33, D35, D82, D101, ...
Trunojoyo:D2, D21, D71, D199, ...
...
...
D1: ...
D2: ... Kuliah di Universitas Trunojoyo menyenangkan walaupun awalnya agak meragukan...
D3: ...
D4: ...
...
Ambil satu per satu dokumen di dalam database web
Area Kerja Preprocessing + Pembuatan Index
Database Web
HTM Tags Removal
Ambil teks murni & simpan ke database
Ekstrak URL, masukkan ke Antrian Crawler
Frontier: Daftar URL yang akan diCrawl
Husni.trunojoyo.ac.id
Ambil file dari Web
Web Crawler
Area Kerja Web Crawler
Hitung Kemiripan Q dengan semua Dokumen (Inverted Index)
Kemiripan(D1, Query) = 0.25
Kemiripan(D2, Query) = 0.15
Kemiripan(D5, Query) = 0.65
Kemiripan(D9, Query) = 0.40
...
Query Q
Universitas Trunojoyo
Lakukan Preprocessing terhadap Query Q
Lakukan Perankingan
Universitas Trunojoyo itu dimana ya?
Kebutuhan Pengguna
Dokumen D5
Dokumen D23
Dokumen D31
...
Kembalikan ke Pengguna
Area Kerja Pemroses Query Pengguna