Temu-Kembali Informasi

2018

02: Arsitektur Search Engine

Husni

Outline

- Arsitektur Software
- Proses Pembuatan Indeks (*Indexing*)
- Proses Pencarian (*Searching*)

Arsitektur Software

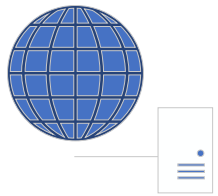
- Arsitektur Software merujuk ke struktur tingkat tinggi dari suatu sistem perangkat lunak.
- Struktur ini diperlukan untuk menjelaskan tentang sistem perangkat lunak.
- Setiap struktur terdiri dari elemen perangkat lunak, hubungan di antara mereka, dan properti dari elemen dan relasi tersebut.
- [Wikipedia]

Arsitektur *Search Engine*

- Arsitektur perangkat lunak dapat ditentukan pada berbagai tingkat abstraksi, juga disebut pandangan (*view*).
- Tampilan fungsional tingkat tinggi dapat diadopsi untuk memperlihatkan apa yang dilakukan mesin telusur, bukan bagaimana cara penerapannya.
- Arsitektur perangkat lunak mesin pencari harus memenuhi dua persyaratan: efektivitas dan efisiensi.
- Efektivitas mengacu pada kualitas retrieval, efisiensi terhadap kecepatan pengambilan.
- Persyaratan lain bergantung kepada dua kategori ini.
- Contoh: Skalabilitas menuntut efisiensi; kesegaran (*freshness*) hasil meningkatkan efektifitas dan membutuhkan efisiensi.

Arsitektur Search Engine

- Mesin pencari pada dasarnya menerapkan dua proses: pengindeksan dan pencarian, di atas lapisan penyimpanan.
- Pengindeksan adalah proses latar belakang untuk menyiapkan data yang akan dicari untuk pencarian yang efisien, dan memperbaruinya.
- Pencarian menawarkan antarmuka pengguna untuk pengiriman query, dan mengimplementasikan pemrosesan query, pemeringkatan, dan presentasi hasil.
- Lapisan penyimpanan mengimplementasikan model data untuk menyimpan dokumen, indeks, dan log sehingga dimungkinkan pencarian terdistribusi dan paralel.
- Silakan lihat arsitektur awal Google yang dijelaskan dalam "*The Anatomy of a Large-scale Hypertextual Web Search Engine*" karya [Sergey Brin dan Larray Page, 1998]



Proses Indexing

Data Storage

Proses Pencarian (*Search*)

Akuisisi

Konversi ke plain text dan unified encoding

Proses Indexing

Data Storage

Proses Pencarian (*Search*)

Proses Indexing: Akuisisi

- Dalam langkah akuisisi, dokumen dikoleksi, disiapkan dan disimpan.
  - Kadang koleksi dokumen yang telah ada (*existing*) dapat digunakan
  - Lebih sering koleksi dokumen perlu dibangun dari awal.
- Komponen:
  - Crawler
  - Converter
  - Document Store

Proses Indexing: Akuisisi - Crawler

Crawler (perayap) menemukan dan mengambil dokumen. Terdapat beberapa varian berbeda:

- Web crawler
  - Eksploitasi *hyperlink* untuk menemukan halaman web
  - Apa tantangan untuk perayapan web?
- Site crawler
  - Perayap web yang dibatasi untuk situs tertentu (misal Domain)
- Focused crawler / topical crawler
  - Mengakuisisi dokumen yang sesuai dengan topik, genre, atau kriteria lain yang ditentukan sebelumnya; membuang yang lain tetapi masih mengeksploitasi hyperlink untuk pencarian
  - Membutuhkan pengelompok (*classifier*) dokumen untuk mengidentifikasi dokumen yang cocok
  - Contoh: penelusuran akademis, penelusuran berita, penelusuran bisnis, penelusuran pekerjaan, dll.
- Document crawler
  - Memindai direktori lokal, email, database, dll.
  - Contoh: pencarian perusahaan, pencarian desktop.

Proses Indexing: Akuisisi - Crawler

Crawler (perayap) menemukan dan mengambil dokumen. Terdapat beberapa varian berbeda:

- **Web crawler**
  - Eksploitasi hyperlink untuk menemukan halaman web
  - Kebijakan eksplorasi (misal hindari jebakan *spider*), identifikasi duplikat (misalnya, normalisasi URL), kebijakan kunjungan-ulang/perbaruan, kesopanan, paralelisasi
- **Site crawler**
  - Perayap web yang dibatasi untuk situs tertentu (mis., Domain)
- **Focused crawler / topical crawler**
  - Mengakuisisi dokumen yang sesuai dengan topik, genre, atau kriteria lain yang ditentukan sebelumnya; membuang yang lain tetapi masih mengeksploitasi hyperlink untuk pencarian
  - Membutuhkan pengelompok (*classifier*) dokumen untuk mengidentifikasi dokumen yang cocok
  - Contoh: penelusuran akademis, penelusuran berita, penelusuran bisnis, penelusuran pekerjaan, dll.
- **Document crawler**
  - Memindai direktori lokal, email, database, dll.
  - Contoh: pencarian perusahaan, pencarian desktop.

Proses Indexing: Akuisisi - Crawler

- Beberapa situs web menginformasikan tentang pembaruan (update) melalui umpan web (misalnya menggunakan RSS atau Atom).
- Untuk situs semacam itu, crawler mungkin akan memilih untuk berlangganan umpan saja, yaitu memeriksa umpan (*feed*) tersebut untuk mengetahui adanya pembaruan.

Proses Indexing: Akuisisi - Converter

Di mesin pencari, converter menyeragamkan dokumen sebagai berikut:

- Reformat/Ekstraksi Teks
  - Dokumen datang dalam berbagai format (misal HTML, PDF, DOC)
  - Langkah pemrosesan lanjutan mengharuskan format input yang seragam (misal *plain text*)
  - Mengekstraksi teks biasa dari dokumen biner adalah *lossy* (misal formatnya hilang)
  - Mengekstraksi format teks juga relevan, untuk langkah lebih lanjut
  - Definisi dari format dokumen spesifik mesin pencari sangat berguna
- Normalisasi *Encoding*
  - Dokumen teks biasa datang dalam berbagai pengkodean (misal ASCII, Unicode)
  - Langkah pemrosesan lanjutan memerlukan *encoding* input seragam (misal Unicode)
  - Spesifikasi pengkodean tidak dapat dipercaya, pengkodean harus dideteksi
  - Pengkodean dokumen sering tidak valid, harus diperbaiki
  - Kesalahan merambat; saat terlihat dalam hasil pencarian, mesin pencari disalahkan.

Proses Indexing: Akuisisi - Document Store

Document store mengelola dokumen-dokumen yang diperoleh:

- Dokumen asli disimpan untuk memungkinkan pemrosesan ulang saat *progress* berlangsung
  - Mengapa dokumen diduplikasi di lokal?
- Dokumen yang dikonversi disimpan untuk *caching* (yaitu efisiensi)
- Meta data mengenai dokumen disimpan (misal asal, tanggal crawl, dll.)
- Riwayat versi dari setiap dokumen dapat disimpan ketika mereka di*crawl* lagi
- Skala sering mendesak bagi database dokumen terdistribusi.

Proses Indexing: Akuisisi - Document Store

Document store mengelola dokumen-dokumen yang diperoleh:

- Dokumen asli disimpan sehingga memungkinkan pemrosesan ulang selama *progress* berlangsung
  - Akses lebih cepat untuk pemrosesan ulang
  - Dokumen asli mungkin tidak selalu tersedia
- Dokumen yang dikonversi disimpan untuk pen-cache-an (yaitu efisiensi)
- Meta data tentang dokumen disimpan (misal asal, tanggal crawl, dll.)
- Riwayat versi dari setiap dokumen dapat disimpan saat dokumen tersebut dihimpun ulang
- Skala sering perlu untuk database dokumen terdistribusi.

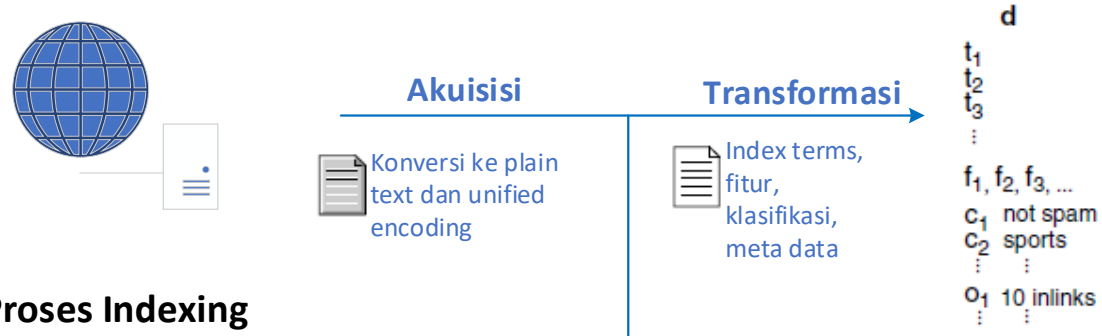
Akuisisi

Konversi ke plain text dan unified encoding

Proses Indexing

Data Storage

Proses Pencarian (*Search*)



Proses Indexing

Data Storage

Proses Pencarian (*Search*)

Proses Indexing: Transformasi Teks

Langkah transformasi mengekstraksi dari teks dokumen kunci yang dapat dicari dalam indeks. Ada dua jenis kunci (*key*):

- Term Indeks (*term*)
  - Kata atau frase dari suatu teks dokumen
  - Tujuannya adalah untuk mewakili isi dokumen
  - Semua term indeks dari semua dokumen digabungkan membentuk kosakata
- Fitur
  - Fitur adalah properti individu terukur dari suatu dokumen
  - Tujuannya adalah mewakili dokumen untuk klasifikasi
  - Set fitur yang berbeda cocok untuk tujuan klasifikasi yang berbeda
  - Contoh *goal* klasifikasi: spam, bahasa, genre, . . .
  - Fitur dan label kelas turunan dapat disimpan.

Proses Indexing: Transformasi Teks

Langkah transformasi mengekstraksi dari teks dokumen kunci yang dapat dicari dalam indeks. Ada dua jenis kunci berbeda:

- Term Indeks (term)
  - Kata atau frase dari suatu teks dokumen
  - Tujuannya adalah untuk mewakili isi dokumen
  - Semua term indeks dari semua dokumen digabungkan membentuk kosakata
- Fitur
  - Fitur adalah properti terukur individu dari suatu dokumen
  - Tujuannya adalah mewakili dokumen untuk klasifikasi
  - Set fitur yang berbeda cocok untuk tujuan klasifikasi yang berbeda
  - Contoh goal klasifikasi: spam, bahasa, genre, . . .
  - Fitur dan label kelas turunan dapat disimpan.
- Komponen:
  - Segmenter
  - Stopping
  - Stemmer / Lemmatizer
  - Link Extraction
  - Information Extraction
  - Classification

Proses Indexing: Transformasi Teks - Segmenter

Segmentasi berarti memecah dokumen ke dalam bagian-bagian penyusunnya. Segmentasi dapat dibedakan pada 2 tingkat:

- Segmentasi Halaman (*page*)
  - Analisis kode HTML dari suatu halaman web berkaitan dengan strukturnya
  - Ekstraksi konten utama vs. iklan, navigasi, tajuk (header), footer, dll.
  - Ekstraksi struktur teks dan pemformatan teks
  - Seringkali HTML dari suatu halaman web tidak valid; itu harus diproses
- Segmentasi Teks
  - Analisis teks polos laman web berkaitan dengan linguistik dan unit teks struktural, seperti kata, kalimat, paragraf
  - Tokenisasi mengubah string teks menjadi urutan token, dimana token dapat berupa kata atau tanda baca. Contoh:
    - Tokenisasi (pemisahan) spasi putih: token dipisahkan oleh karakter spasi
    - Definisi sederhana kata: token yang berupa string alfanumerik
    - Mengapa definisi-definisi ini tidak cukup?
  - Kata berhuruf kecil digunakan sebagai kandidat term Indeks.

Proses Indexing: Transformasi Teks - Segmenter

Segmentasi berarti memecah dokumen ke dalam bagian-bagian penyusunnya. Segmentasi dapat dibedakan pada 2 tingkat :

- Segmentasi Halaman
  - Analisis kode HTML dari suatu halaman web berkaitan dengan strukturnya
  - Ekstraksi konten utama vs. iklan, navigasi, tajuk (header), footer, dll.
  - Ekstraksi struktur teks dan pemformatan teks
  - Seringkali HTML dari suatu halaman web tidak valid; itu harus diproses.
- Segmentasi Teks
  - Analisis teks polos laman web berkaitan dengan linguistik dan unit teks struktural, seperti kata, kalimat, paragraf
  - Tokenisasi mengubah string teks menjadi urutan token, di mana token dapat berupa kata atau tanda baca. Contoh:
    - Tokenisasi (pemisahan) spasi putih: token dipisahkan oleh karakter spasi
    - Definisi sederhana kata: token yang berupa string alfanumerik
    - Tanda baca dan *adjectivization* menggunakan tanda hubung tidak dipisahkan oleh spasi
    - Kontraksi dan karakter khusus diabaikan.
  - Kata berhuruf kecil digunakan sebagai kandidat term Indeks.

Proses Indexing: Transformasi Teks - Stopping

Stopping, juga *stop word removal*, membuang kata tertentu dari kumpulan term indeks suatu dokumen, atau seluruh kosakata. Kandidat dari *stop words*:

- **Kata Fungsi**

  Kata-kata yang memiliki sedikit makna leksikal, ambigu, hanya melayani tujuan gramatikal, atau menentukan sikap atau suasana hati. Contoh: dari, untuk, bagi, kepada.

- **Kata yang Sering**

  Kata-kata yang paling sering muncul dari suatu bahasa, atau dalam kumpulan dokumen. Contoh: "Wikipedia" muncul di setiap halaman Wikipedia.

- **Kata Spesifik Domain**

  Kata-kata yang tidak mendiskriminasikan dalam domain pencarian yang diberikan. Contoh: "belajar" dapat diabaikan dalam domain pendidikan, terlepas dari frekuensinya.

Keuntungan dari stopping adalah ukuran indeks yang tereduksi, kecepatan pemrosesan query yang lebih cepat, dan reduksi noise potensial dalam retrieval.

Sisi minusnya adalah bahwa kotak-pojok mungkin tidak cukup tercakup. Contoh: pencarian “to be or not to be”.

Tergantung pda model temu-kembali, stopping dapat/tidak dapat mempengaruhi efektifitas dari retrieval. Stopping pada teks biasanya lebih konservatif daripada terhadap query.

Proses Indexing: Transformasi Teks - Stemmer/Lemmatizer

Stemming bertujuan untuk mengurangi istilah indeks infleksi ke batang umum. Contoh: "statistics" juga harus cocok dengan "statistic" dan "statistical". Dua pendekatan untuk stemming:

- *(Heuristic) Stemming*

  Penghapusan afiks berbasis aturan (yaitu sufiks, prefiks, dan infiks). Contoh: “worker ”, “megavolt”, “un-bloody-likely”. Heuristik naif: memotong kata pada huruf ke-4.

- *Lemmatization*

  Pemetaan kata ke bentuk akarnya, bahkan jika dieja berbeda. Contoh: “saw” dan “see”.

Sisi positif dari *stemming*/*lemmatization* adalah peningkatan peluang untuk menemukan dokumen ketika menggunakan tata bahasa yang berbeda atau kata-kata turunan yang berbeda dari query.

Apa masalah yang terkait dengan *stemming* agresif?

Proses Indexing: Transformasi Teks - Stemmer/Lemmatizer

Stemming bertujuan untuk mengurangi istilah indeks infleksi ke batang umum. Contoh: "statistics" juga harus cocok dengan "statistic" dan "statistical". Dua pendekatan untuk *stemming*:

- *(Heuristic) Stemming*

  Penghapusan afiks berbasis aturan (yaitu sufiks, prefiks, dan infiks). Contoh: “worker ”, “megavolt”, “un-bloody-likely”. Heuristik naif: memotong kata pada huruf ke-4.

- *Lemmatization*

  Pemetaan kata ke bentuk akarnya, bahkan jika dieja berbeda. Contoh: “saw” dan “see”.

Sisi positif dari *stemming*/*lemmatization* adalah peningkatan peluang untuk menemukan dokumen ketika menggunakan tata bahasa yang berbeda atau kata-kata turunan yang berbeda dari query.

Kelemahan adalah konfidasi kata-kata yang tidak berhubungan. Contoh: “university”, “universe”, dan “universal” distem menjadi “univers” oleh stemmer umum. Kata hasil stemming mungkin bukan yang asli. *Lemmatization* membutuhkan sumber daya yang mahal.

Efektivitas tergantung pada bahasa (misalnya, bahasa China vs. bahasa Arab). Pendekatan alternatif: **ekspansi Query**.

Proses Indexing: Transformasi Teks - Link Extraction

Ekstraksi tautan dan teks tautan dari dokumen. Ini melayani dua tujuan:

- Analisis tautan (*link*)

  Hyperlink menginduksi graf antar halaman web. Analisis tautan melintasi graf ini untuk mengidentifikasi laman web otoritatif. Algoritma untuk ini termasuk PageRank dan HITS.

- Augmentasi teks dengan teks jangkar (*anchor*)

  Teks yang ditemukan di halaman web mungkin tidak cukup untuk mendeskripsikan isinya.

  Contoh: halaman produk atau halaman yang hanya menampilkan gambar. Hyperlink sering menyertakan teks, dan teks sebelum dan sesudah hyperlink dapat membenarkan tautan. Teks jangkar ini ditambahkan ke teks yang diekstrak dari halaman yang ditautkan untuk diindeks juga.

Proses Indexing: Transformasi Teks - Ekstraksi Informasi

Ekstraksi Informasi bertujuan untuk mengidentifikasi term indeks yang lebih kompleks dengan menggunakan teknologi pemrosesan bahasa alami (secara komputasi mahal):

- Frasa benda

  Frasa yang memiliki kata benda sebagai kata dasar, yaitu kata benda dan kata apa pun yang memodifikasinya. Contoh: "Rumah kuning dijual.", "Saya ingin skate board".

- Entitas bernama (*named entities*)

  Kata atau frasa yang menunjuk sesuatu (misalnya, tempat, seseorang, organisasi, dll.).

- Resolusi *Coreference*

  Coreferences, yaitu *anaphora* dan *cataphora*, adalah ekspresi yang merujuk ke belakang atau ke depan di dalam teks. Menyelesaikannya penting untuk pemahaman teks, namun, salah satu masalah yang paling sulit dari pemrosesan bahasa alami

- Deteksi Relasi

  Ekstraksi hubungan antara entitas bernama yang disebutkan dalam teks. Contoh: "Bill hidup di Amerika Serikat".

- Ekstraksi informasi Semi-Terstruktur

  Ekstraksi tabel, kutipan, referensi, komentar, dll.

Proses Indexing: Transformasi Teks - Klasifikasi

Klasifikasi menerapkan pembelajaran mesin untuk mengidentifikasi jenis dokumen tertentu serta mengkategorikannya. Ini didasarkan pada fitur yang diekstraksi saat transformasi. Tujuan klasifikasi umum adalah:

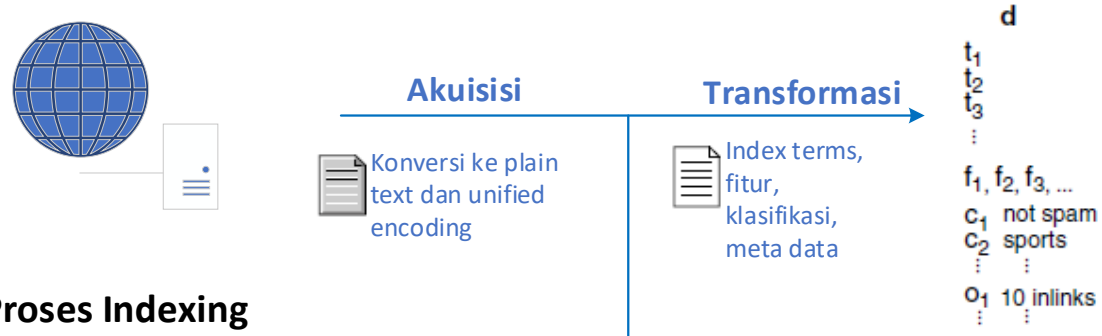
- **Deteksi Spam/malware**

  Menentukan apakah situs/halaman web mencoba menumbangkan peringkat mesin pencari (spam), atau membahayakan penggunanya (*malware*).

- **Deteksi Bahasa**

  Menentukan bahasa (utama) dari suatu halaman web.

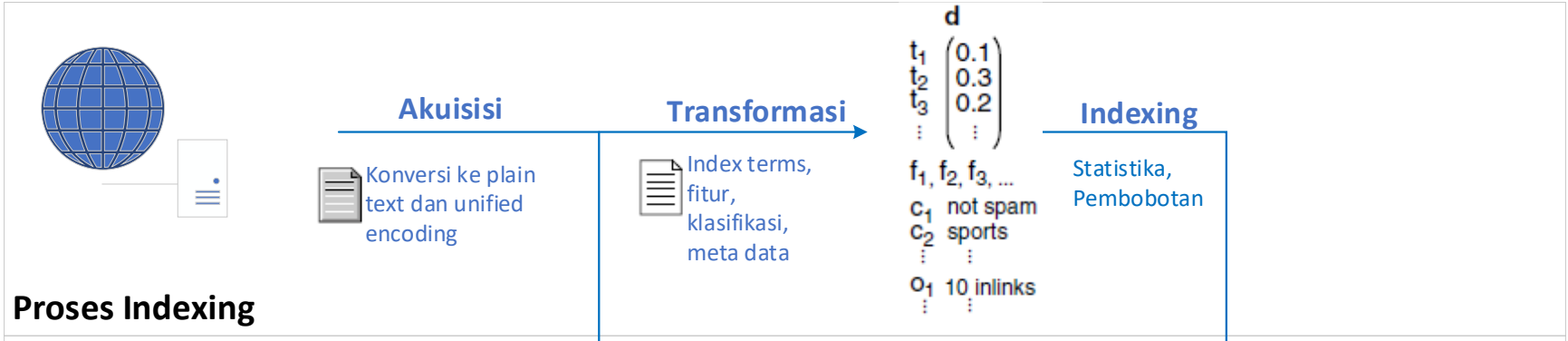
- **Kategorisasi Topik/Analisis Klaster**
  - Menentukan topik dokumen, baik dengan menugaskan deskriptor subjek yang telah ditentukan sebelumnya (mis. Olahraga, politik, teknologi, dll.), Mencakup topik yang penting bagi pengguna atau dalam domain pencarian, atau dengan mengidentifikasi topik menggunakan analisis klaster, di mana kelompok yang diidentifikasi harus diberi label. Topik dapat tumpang tindih dan membentuk hierarki.
- **Ketegorisasi Genre**
  - Menentukan genre laman web (misalnya, beranda pribadi, papan pesan, blog, toko, dll.). Tidak ada daftar genre web yang disepakati bersama. Genre tumpang tindih dan membentuk hierarki, dan dapat bergantung pada domain pencarian.



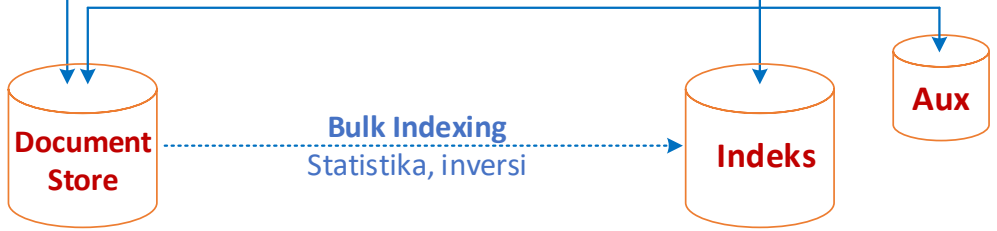
Proses Indexing

Data Storage

Proses Pencarian (*Search*)



Proses Indexing



Data Storage

Proses Pencarian (*Search*)

Proses Indexing: Indexing

Tahapan *indexing* membuat struktur data indeks yang diperlukan untuk *retrieval* yang cepat dari dokumen-dokumen yang ditransformasi.

Struktur data yang paling umum digunakan adalah indeks terbalik (*inverted index*).

Struktur data lainnya termasuk *suffix array* dan *signature file*.

Komponen:

- Statistika dokumen
- Pembobotan (*weighting*)
- Inversi
- Distribusi

Proses Indexing: Indexing - Statistik Dokumen

Kumpulkan dan simpan informasi tambahan tentang dokumen, diperlukan untuk pemrosesan Query online yang cepat, misalnya untuk menilai (menskor) dan memberi peringkat (*ranking*) dokumen saat menyertakan informasi yang hanya tersedia pada waktu Query.

Informasi tersebut mungkin termasuk :

- Frekuensi term per dokumen, topik dan genrenya
- Frekuensi dokumen per term
- Posisi term per dokumen
- Panjang dokumen

Struktur data yang digunakan adalah penyimpanan kunci-nilai (yaitu *hashmap*).

Proses Indexing: Indexing - Pembobotan (*Weighting*)

Untuk setiap term indeks suatu dokumen, hitung bobot yang menunjukkan pentingnya dengan memperhatikan isi dokumen, memungkinkan untuk menskor dokumen berkaitan dengan Query.

Skema pembobotan umum:

- *Term frequency* (tf)

  Logaritma dari jumlah kemunculan suatu term di dalam dokumen.
- *Inverse Document frequency* (idf)
  - Frekuensi dokumen (df): Jumlah dokumen yang mengandung suatu term
  - Logaritma dari jumlah dokumen dibagi dengan df .
- tf idf

  Salah satu skema pembobotan term paling dikenal dalam IR.
- BM25

  Mirip dengan tf idf , tetapi menghasilkan kinerja retrieval lebih baik.

Pra-komputasi bobot term dan menyimpannya dalam indeks atau struktur data tambahan mempercepat penskoran dokumen. Skema pembobotan lain dihitung berdasarkan informasi tentang Query.

Proses Indexing: Indexing - Inversi

Inversi berarti mengubah data dokumen-term menjadi data term-dokumen untuk membuat struktur data indeks terbalik.

- ***Bulk indexing***

  Membuat indeks dengan memproses semua dokumen yang ditransform. Ini terjadi secara offline; Setelah siap, indeks yang sedang digunakan diganti dengan yang baru.

- ***Index update***

  Perbarui indeks yang ada saat ada kehadiran dokumen baru. Ini terjadi secara online, sementara indeks sedang digunakan.

Proses Indexing: Indexing - Distribusi

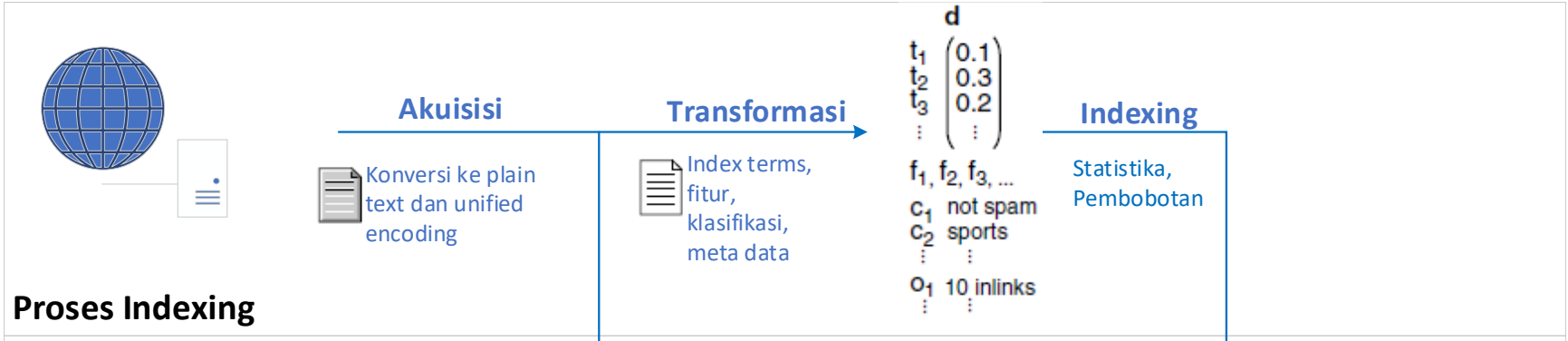
Distribusi indeks (juga disebut *sharding* atau *partitioning*) pada banyak mesin dan pusat data untuk mendukung paralelisasi.

- Distribusi dokumen
  - Membagi koleksi; indeks lebih kecil berupa sub-koleksi pada mesin-mesin berbeda
- Distribusi Term
  - Membagi indeks dari seluruh koleksi berdasarkan term
  - Mesin berbeda melayani term-term berbeda
- Replikasi
  - Salinan (bagian dari) indeks pada banyak lokasi
- Apa alasan men*sharding* berdasarkan dokumen, term dan replikasi?

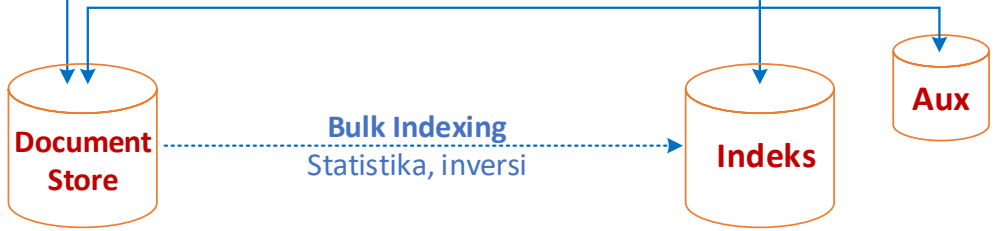
Proses Indexing: Indexing - Distribusi

Distribusi indeks (juga disebut *sharding* atau *partitioning*) pada banyak mesin dan pusat data untuk mendukung paralelisasi.

- Distribusi Dokumen
  - Membagi koleksi; indeks lebih kecil berupa sub-koleksi pada mesin berbeda
  - Memungkinkan paralelisme untuk pengindeksan dan pemrosesan Query
  - Indeks yang lebih kecil sering lebih cepat, apalagi jika ada *cache*
- Distribusi Term
  - Membagi indeks dari seluruh koleksi berdasarkan term
  - Mesin berbeda melayani term-term berbeda
  - Tidak semua mesin harus memproses setiap Query
- Replikasi
  - Salinan (bagian dari) indeks pada banyak lokasi
  - Mengurangi penundaan selama pemrosesan kueri
  - *Fault tolerance*

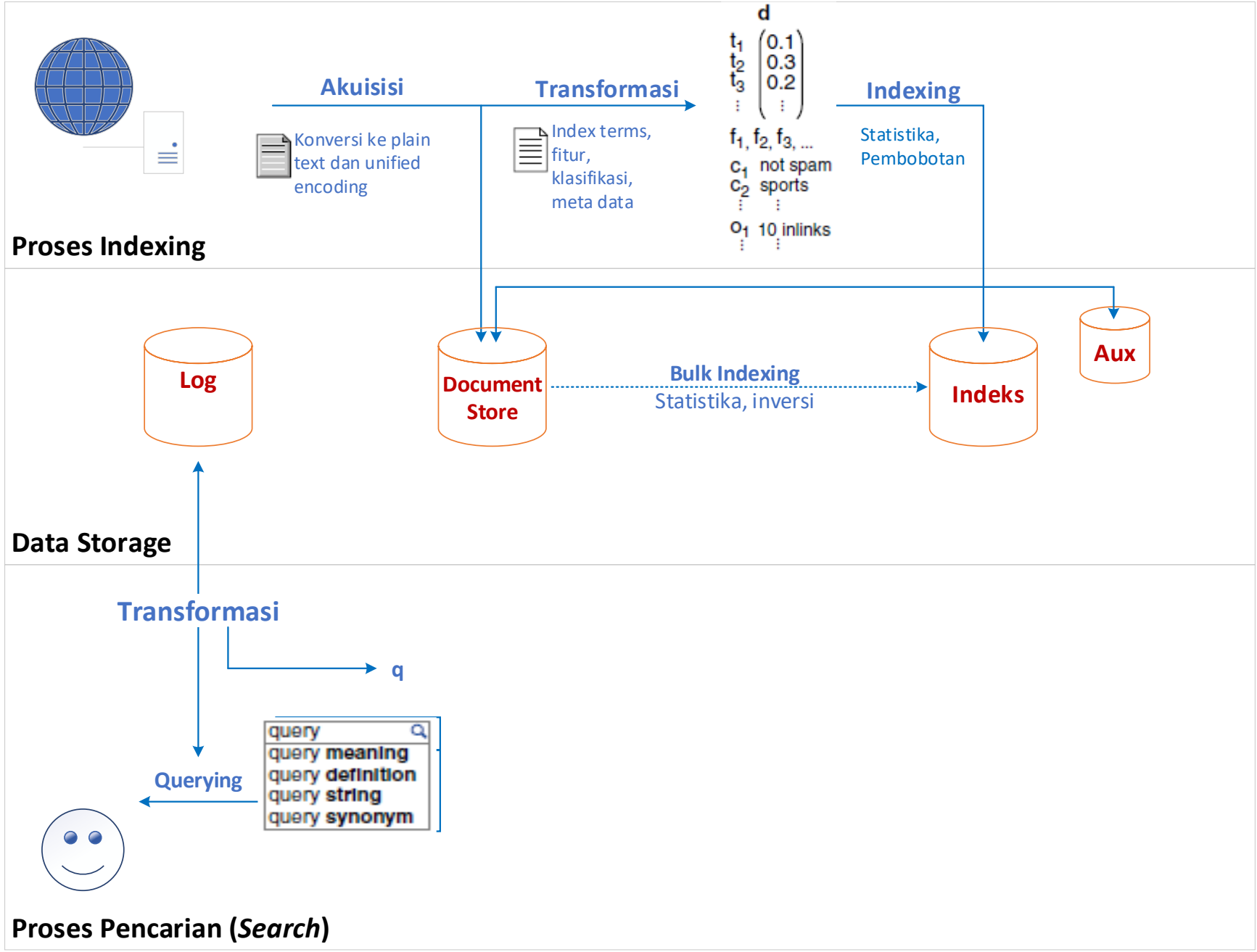


Proses Indexing



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)

Proses Pencarian: Interaksi Pengguna

Interaksi pengguna termasuk antarmuka pengguna yang ditawarkan oleh mesin pencari, skenario penggunaan yang diimpikan, dan implementasinya.

Interaksi pengguna dasar:

- *Query submission*
- *Result presentation*

Proses Pencarian: Interaksi Pengguna

Interaksi pengguna termasuk antarmuka pengguna yang ditawarkan oleh mesin pencari, skenario penggunaan yang diimpikan, dan implementasinya.

Interaksi pengguna lanjutan:

- *Query refinement*
- *Result exploration*

Proses Pencarian: Interaksi Pengguna

Interaksi pengguna termasuk antarmuka pengguna yang ditawarkan oleh mesin pencari, skenario penggunaan yang diimpikan, dan implementasinya.

Interaksi pengguna lanjutan:

- *Query refinement*
- *Result exploration*

Komponen:

- *Query Language*
- *Query Transformation*
- *Results Output*

Proses Pencarian: Interaksi Pengguna - Bahasa Query

Bahasa query mendefinisikan sintaks dan semantik dari query yang valid.

Termasuk perintah untuk mempengaruhi pencarian, disebut operator query.

Jenis query umum:

- Query terstruktur
- Query kata-kunci (*keyword*)
- Query pertanyaan
- Query berdasarkan contoh

Operator umum:

- Operator boolean (AND, OR, NOT (atau –))
- Operator lain?

Proses Pencarian: Interaksi Pengguna - Bahasa Query

Bahasa query mendefinisikan sintaks dan semantik dari query yang valid.

Termasuk perintah untuk mempengaruhi pencarian bernama operator query.

Jenis query umum:

- Query terstruktur
- Query kata-kunci (*keyword*)
- Query pertanyaan
- Query berdasarkan contoh

Operator umum:

- Operator Boolean (AND, OR, NOT (or –))
- Quotes / phrasal search (“phrase of text”)
- Field search (title, text, url)
- Wildcards (*, 50..100)
- Site search (site:example.com)

Bentuk paling dasar dari bahasa query adalah kata kunci pencarian.

Hanya sekitar 1% dari Query web berisi operator [White dan Morris 2007]. Search engine tidak bergantung pada pengguna yang menjadi ahli bahasa.

Mesin pencari spesifik domain sering memiliki bahasa query khusus, memungkinkan kendali perilaku retrieval yang sangat halus.

Proses Pencarian: Interaksi Pengguna - Tranformasi Query

Langkah transformasi Query memetakan kata kunci Query ke term indeks, dan menyaring Query dalam upaya lebih memahami maksud pengguna.

- Transformasi Teks

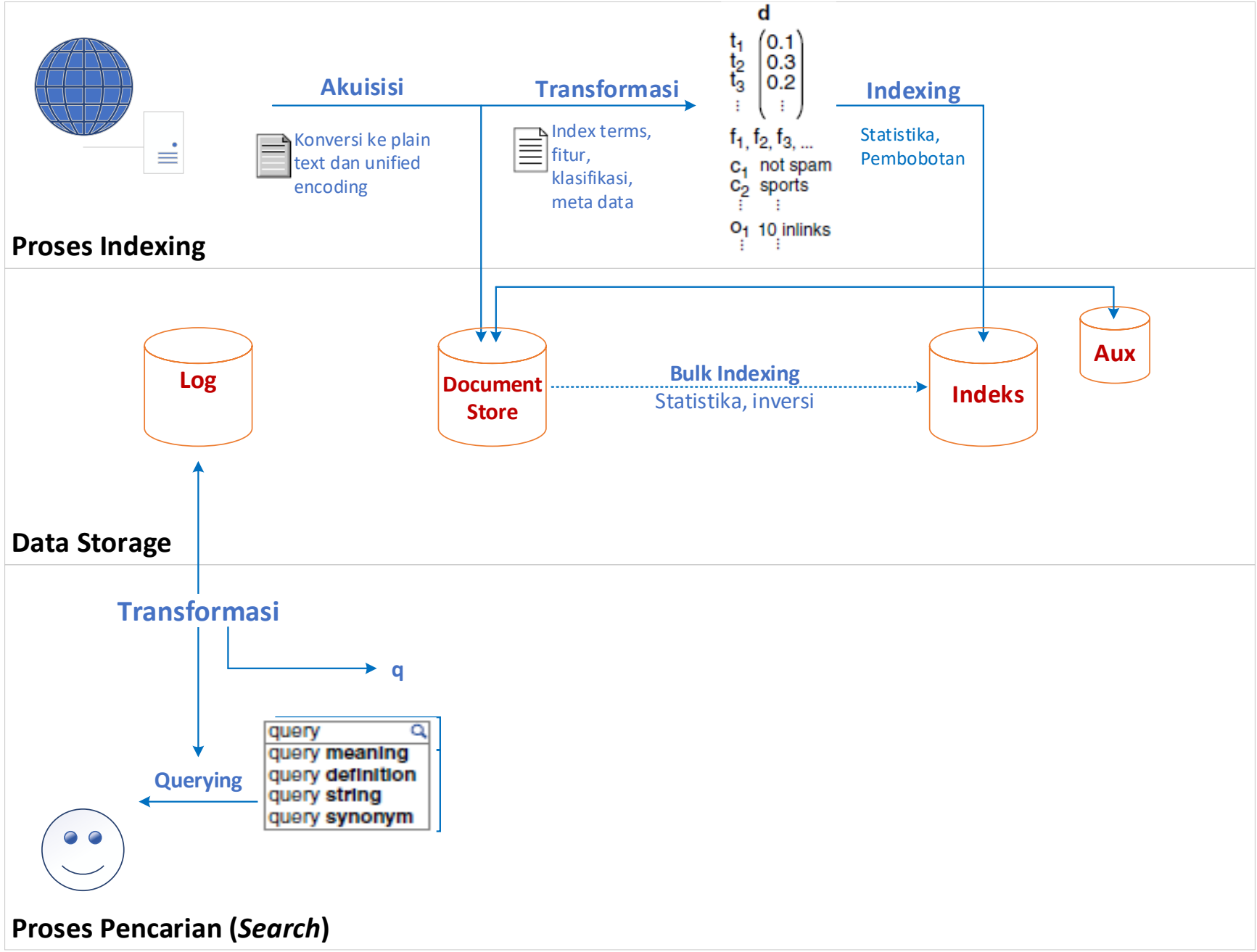
  Menggunakan cara yang sama dengan langkah transformasi teks pada dokumen, terdiri dari *tokenization*, *stoping*, *stemming*, dll, untuk memastikan kesesuaian dengan term indeks.

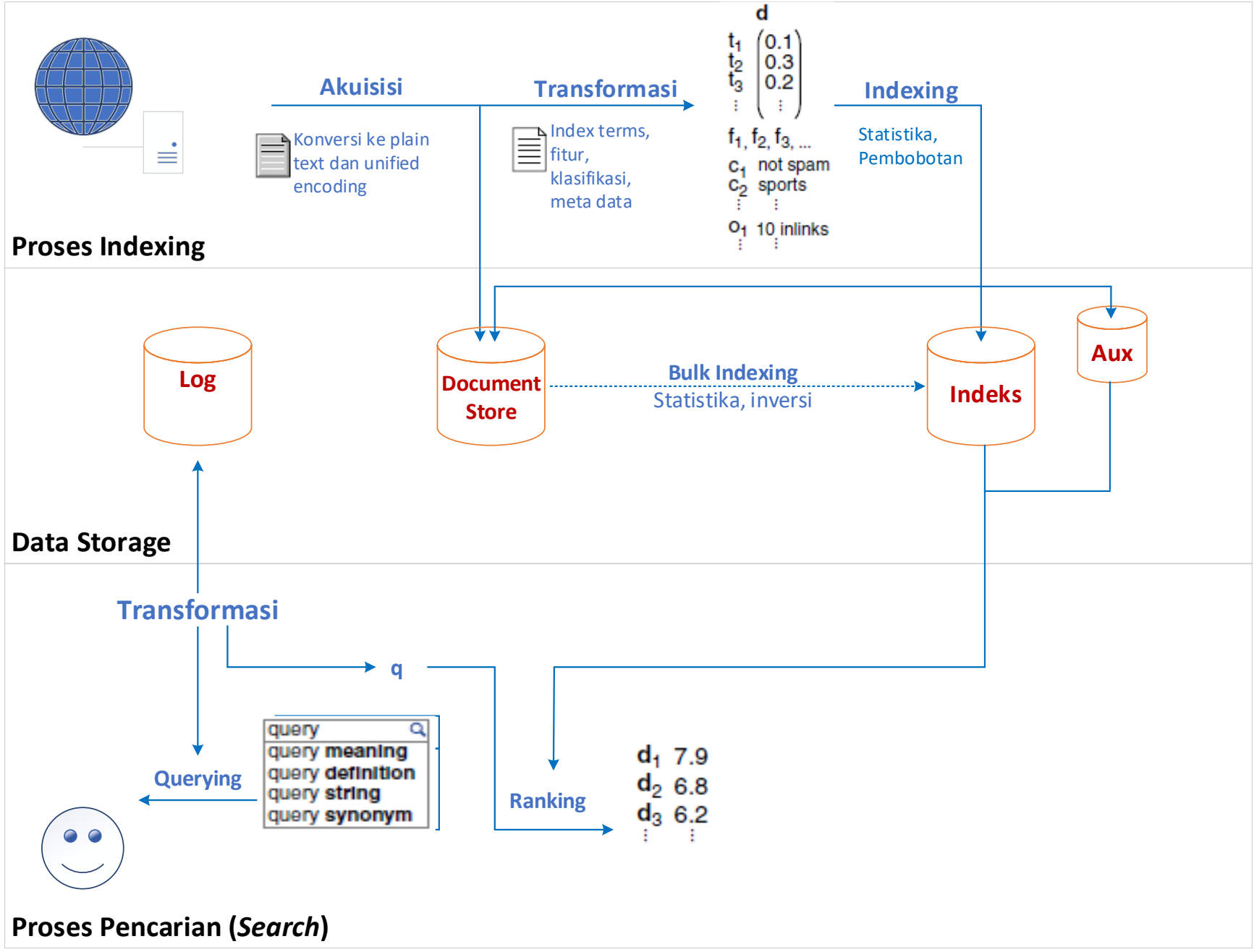
- Pemeriksaan Ejaan dan Anjuran Query

  Memberikan umpan balik pengguna tentang query pada berbagai tingkat urgensi, mulai dari petunjuk kecil ("Apakah maksud Anda ...") hingga penggantian otomatis, bergantung pada keyakinan pada alternatif yang disarankan. Log query digunakan, membuat sistem pencarian berbasis log.

- Ekspansi Query

  Saran penyempurnaan kata, dan term tambahan ditambahkan ke query, membuatnya lebih spesifik. Log query dan ko-okurensi term dalam dokumen dieksploitasi di sini.


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)

Proses Pencarian: Ranking

Diberikan suatu Query bertransformasi, tahapan Ranking menskor dan meranking dokumen yang diindeks dengan menilai relevansinya terhadap Query.

Langkah ini adalah inti penerapan model *retrieval* yang mendasari mesin telusur. Teori tentang bagaimana relevansi dapat dikuantifikasi.

Model *retrieval* terdiri dari fungsi untuk merepresentasikan dokumen dan Query, dan fungsi untuk memberi peringkat berdasarkan pada representasi.

Representasi dokumen biasanya dihitung secara *offline* dan diindeks sebelumnya(misal bobot term menggunakan tf idf).

Peringkat dihitung *online* untuk setiap Query.

Komponen:

- *Document Scoring*
- *Efficient Document Scoring*
- *Distribution*

Proses Pencarian: Ranking - Penskoran Domumen

Langkah ***Scoring*** menghitung relevansi dokumen yang diindeks ke suatu Query.

Jika t $\in$ V menunjukkan term t dari kamus V term-term indeks, dan Wx : V x X $\rightarrow$ R menunjukkan fungsi pembobotan term, dimana X berupa himpunan dokumen D dan Query Q. Maka bentuk paling dasar dari *scoring* dokumen:

$$\sum_{t \in V} W_q(t).Wd(t)$$

dimana $W_q(t)$ dan $W_d(t)$ adalah bobot term yang menunjukkan pentingnya t bagi query q $\in$ Q dan dokumen d $\in$ D.

Observasi:

- Bobot term $W_d(t)$ telah dihitung dan diindeks.
- Bobot term $W_q(t)$ harus dihiting *on the fly*.
- Term t punya kepentingan, karena itu bobot tidak nol. Perhatikan juga sinonim.
- Mayoritas term dari V tidak signifikan.

Proses Pencarian: Ranking - Penskoran Dokumen Efisien

Menghitung skor dokumen membutuhkan akses indeks. Strategi akses yang mengatur apa yang dapat dicapai dan bagaimana data harus diatur. Dua strategi yang paling menonjol adalah sebagai berikut :

- Penskoran *Document-at-a-time*
  - Prekondisi: dokumen di dalam daftar postingan indeks diurutkan (misal berdasarkan document-ID, bukan kualitas dokumen).
  - Daftar postingan dari Term Query dijelajah secara paralel untuk menskor satu dokumen dalam satu waktu.
  - Skor setiap dokumen langsung lengkap, tetapi peringkatnya hanya di bagian akhir.
  - Biaya disk IO konkuren bertambah mengikuti panjang Query.
- Penskoran *Term-at-a-time*
  - Jelajah postingan satu demi satu (misal pengurutan term berdasarkan frekuensi atau kepentingan).
  - Mempertahankan daftar postingan Query sementara, yang berisi dokumen kandidat.
  - Ketika nilai setiap dokumen terakumulasi, peringkat perkiraan tersedia.
  - Lebih banyak memori utama yang diperlukan untuk mempertahankan postlist sementara.
- Optimasi yang aman dan tidak aman ada, misalnya, untuk menghentikan pencarian lebih awal.

Proses Pencarian: Ranking - Distribusi

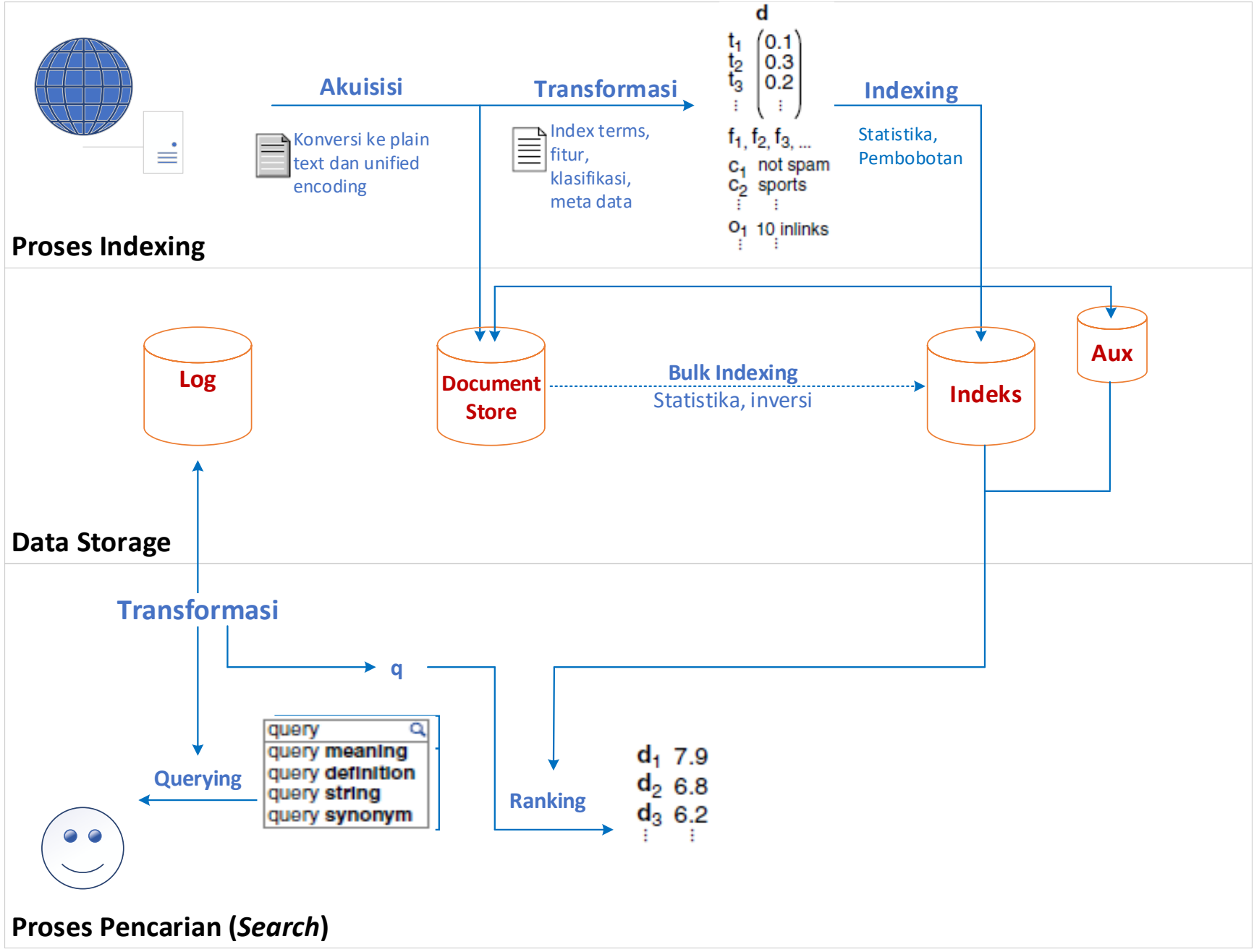
Distribusi pemrosesan kueri bergantung pada indeks.

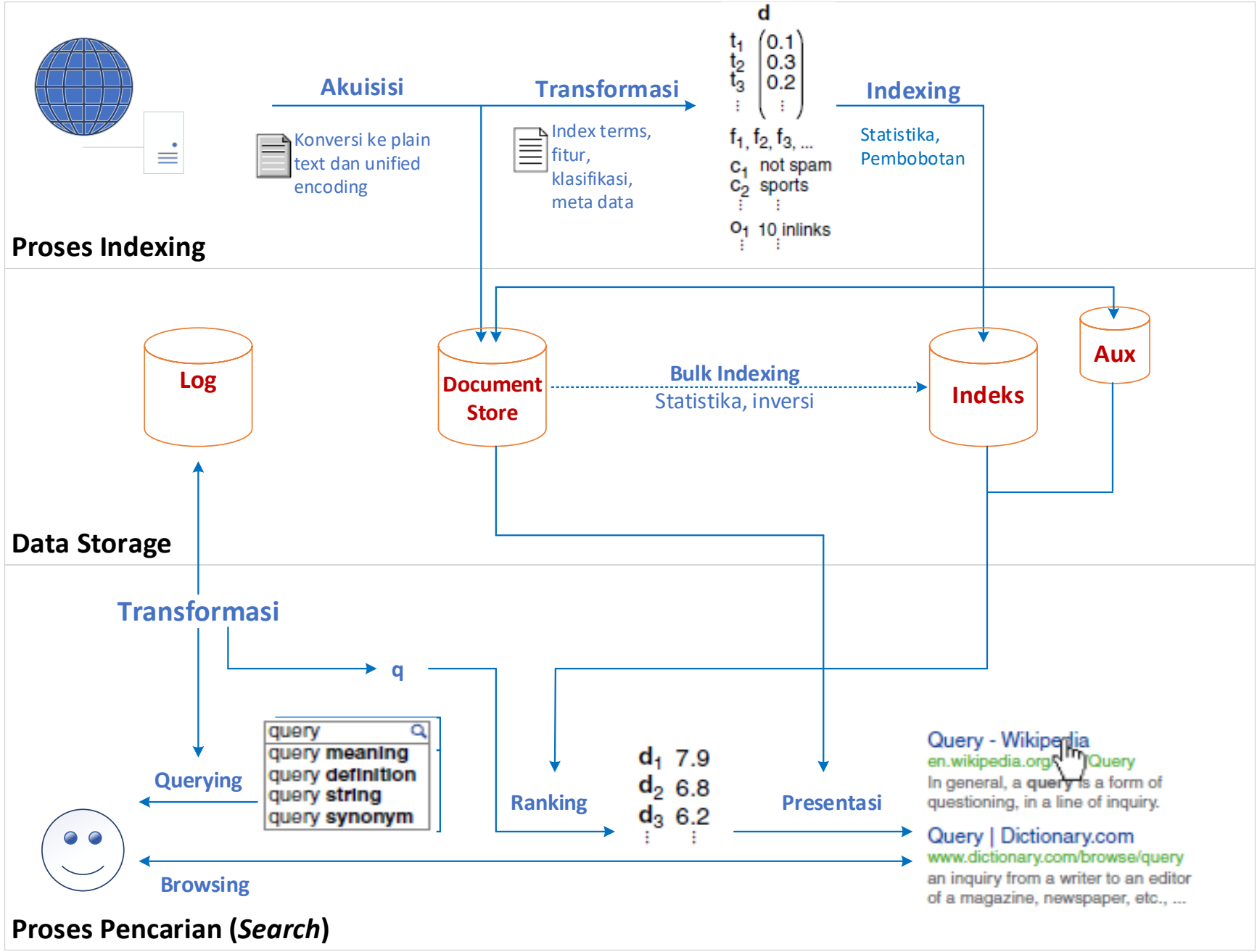
- Query broker / load balancer

  Memutuskan mana shard dan mana salinan yang direplikasi untuk diakses. Menerima dan menggabungkan hasil.

- Cache

  Menyimpan data yang sering digunakan pada tempat dekat (misal dalam memori utama) untuk mengurangi latensi. Cache mungkin termasuk indeks yang berisi dokumen penting yang hanya disimpan di memori utama, hasil pencarian yang sudah ditentukan sebelumnya, daftar postingan sementara, bagian dari daftar postingan, dan penggunaan hirarki cache yang dioptimalkan dari sistem operasi ke cache perangkat keras.

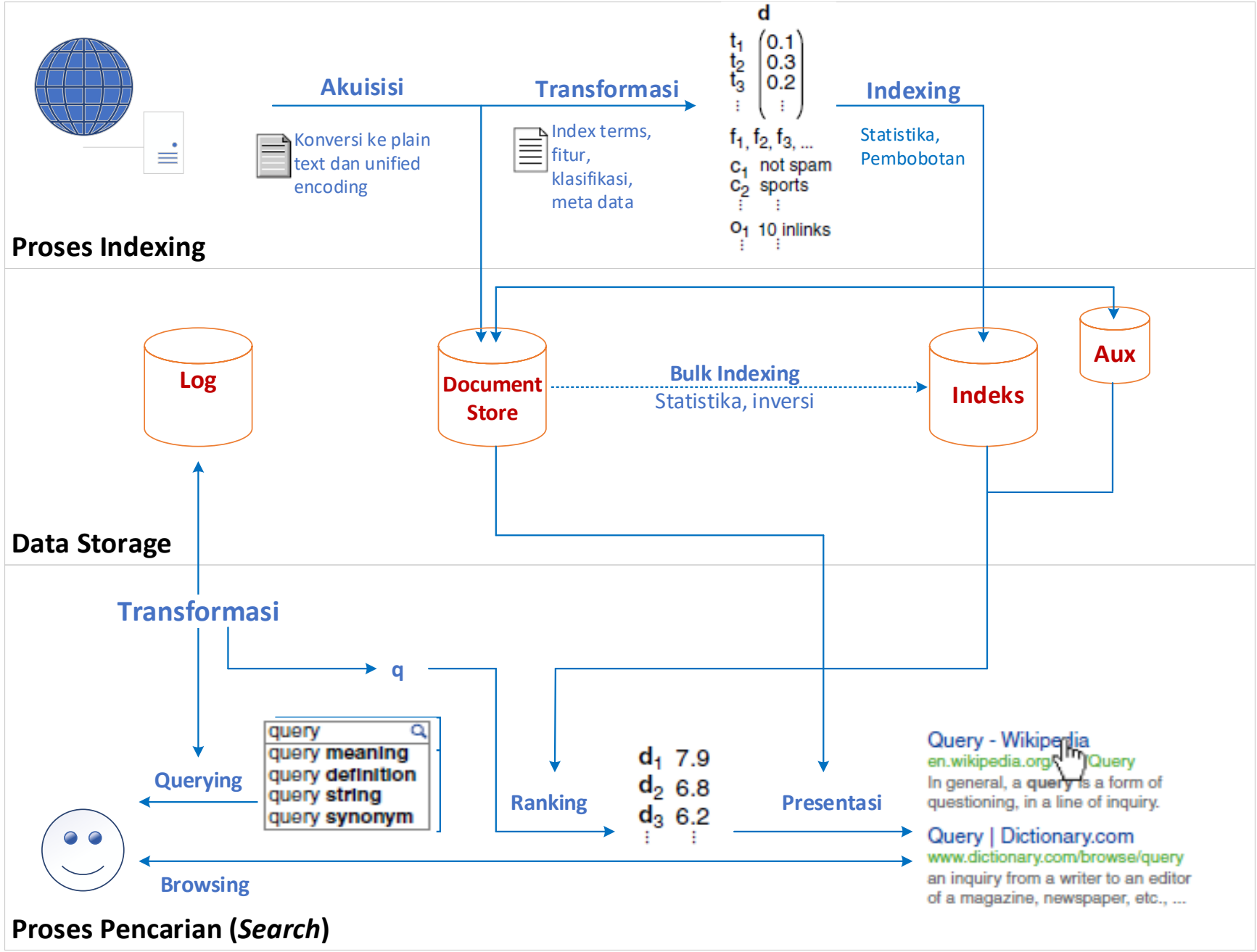

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data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

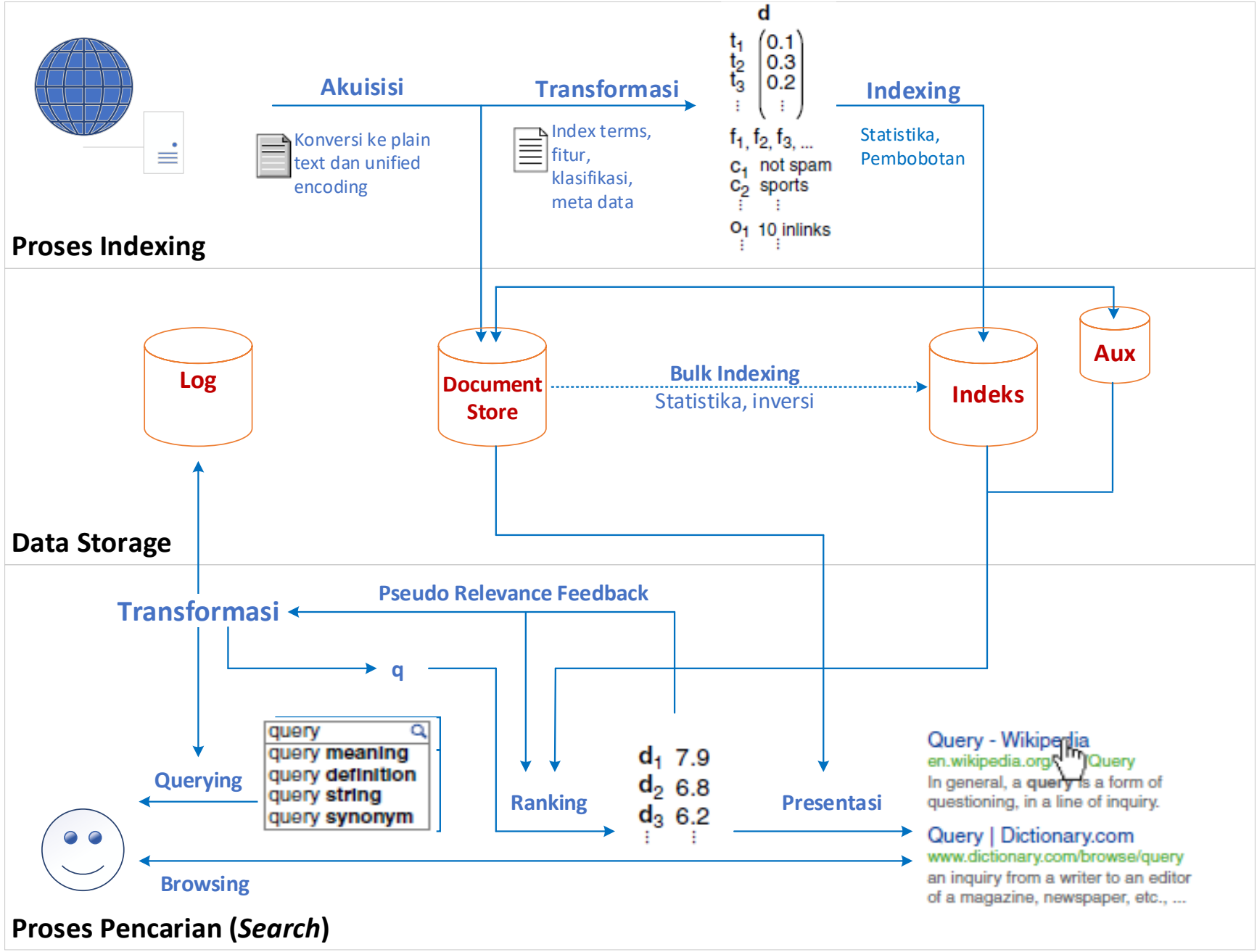

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Document Store
Bulk Indexing
Statistika, inversi
Indeks
Aux
Data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

Proses Pencarian: Interaksi Pengguna - Output Hasil

Langkah Output Hasil menyusun halaman web untuk ditampilkan kepada pengguna. Beberapa langkah retrieval tambahan diperlukan:

- **Snippet generation**
  Mengakses halaman web asli dan mengekstrak kalimat dan frasa yang meringkasnya, tergantung pada Query.
- **Query term highlighting**
  Preprocesses snippet untuk menyorot kata-kata dari query, terlepas dari infleksi mereka.
- **Clustering**
  Secara opsional, kelompokkan himpunan hasil untuk memberikan rangkaian hasil yang lebih beragam.
- **Ad retrieval**
  Mengakses mesin pencari lain yang dirancang khusus untuk pengambilan iklan yang relevan dengan query dari semua iklan yang ditawarkan oleh mitra iklan. Ini berjalan secara paralel dengan retrieval hasil pencarian organik.
- **Other oneboxes**
  Menentukan apakah mesin telusur khusus lainnya dapat memberikan hasil yang relevan. Memberi peringkat onebox ke dalam hasil pencarian sesuai kepentingannya dibandingkan dengan hasil web organik.


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)

Proses Pencarian

Interaksi Pengguna - Query Transformation (Lanj.)

Langkah transformasi Query memetakan kata kunci Query ke term indeks, dan menyaring Query dalam upaya lebih memahami maksud pengguna.

- Transformasi Teks

  Menggunakan cara yang sama dengan langkah transformasi teks dokumen, terdiri dari tokenisasi, Stoping, stemming, dll, untuk memastikan cocok dengan term indeks.

- Periksa Ejaan dan Saran Query

  Memberikan umpan balik pengguna tentang Query pada berbagai tingkat urgensi, mulai dari petunjuk kecil ("Apakah maksud Anda ...") hingga penggantian otomatis, bergantung pada keyakinan pada alternatif yang disarankan. Log Query digunakan untuk membuat sistem pencarian berbasis log.

- Ekspansi Query dan umpan-balik relevansi

  Menyarankan kelengkapan term dan term tambahan ditambahkan ke Query, membuatnya lebih spesifik. Log Query dan term ko-okuren dalam dokumen dieksploitasi di sini.

  Umpan balik relevansi semu menambahkan term ke kueri yang diekstrak dari hasil pencarian teratas.

  Masalah apa yang terlihat dengan umpan balik relevansi semu?

Proses Pencarian

Interaksi Pengguna - Transformasi Query (Lanj.)

Langkah Transformasi Query memetakan kata kunci Query ke term indeks, dan menyaring Query dalam upaya lebih memahami maksud pengguna.

- Transformasi Teks

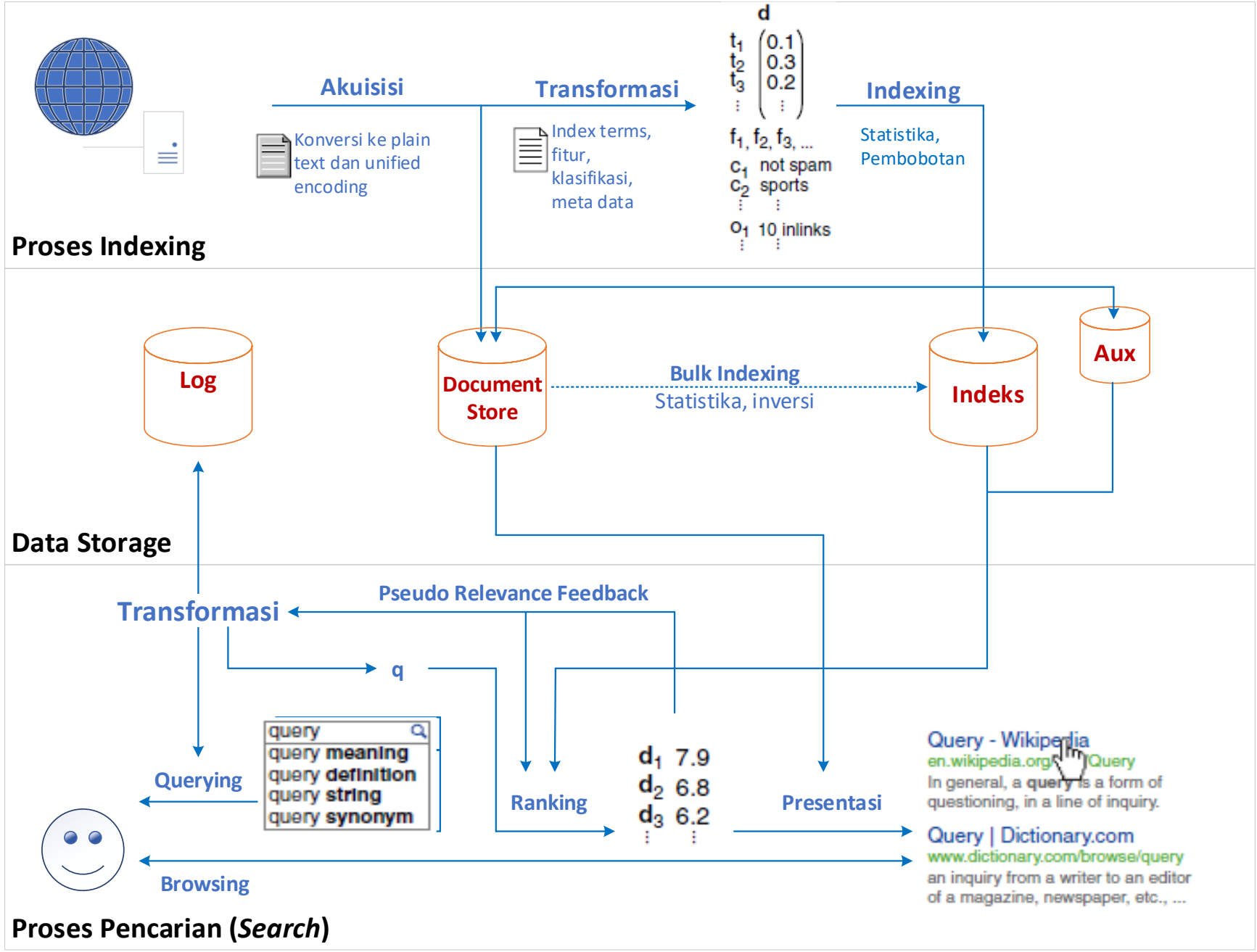
  Menggunakan cara yang sama dengan langkah transformasi teks dokumen, terdiri dari tokenisasi, Stoping, stemming, dll, untuk memastikan cocok dengan term indeks.

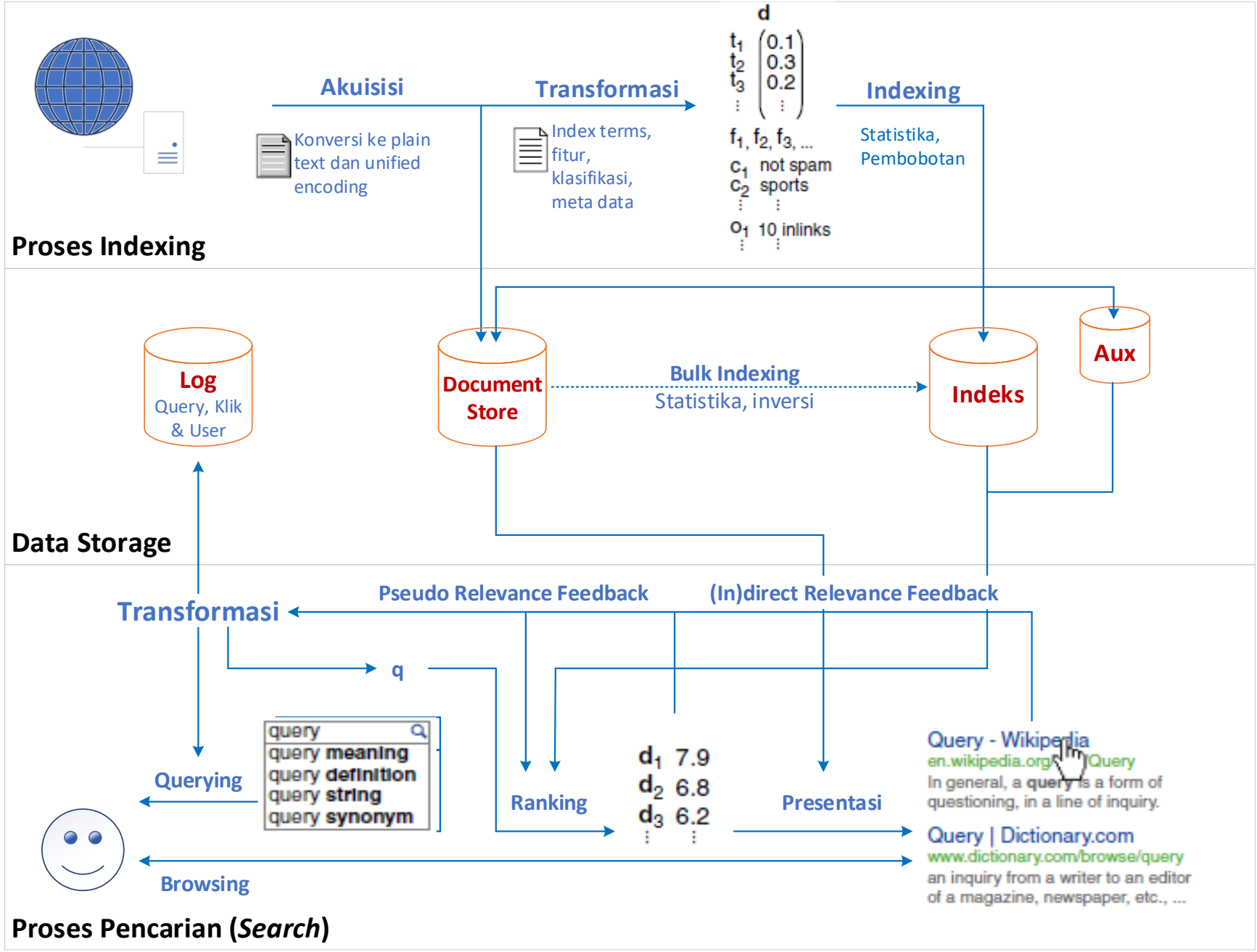
- Periksa Ejaan dan Saran Query

  Memberikan umpan balik pengguna tentang Query pada berbagai tingkat urgensi, mulai dari petunjuk kecil ("Apakah maksud Anda ...") hingga penggantian otomatis, bergantung pada keyakinan pada alternatif yang disarankan. Log Query digunakan untuk membuat sistem pencarian berbasis log.

- Ekspansi Query dan umpan-balik relevansi

  Topik hasil penelusuran teratas diperkuat, yang mungkin berbeda dari maksud pengguna.


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)

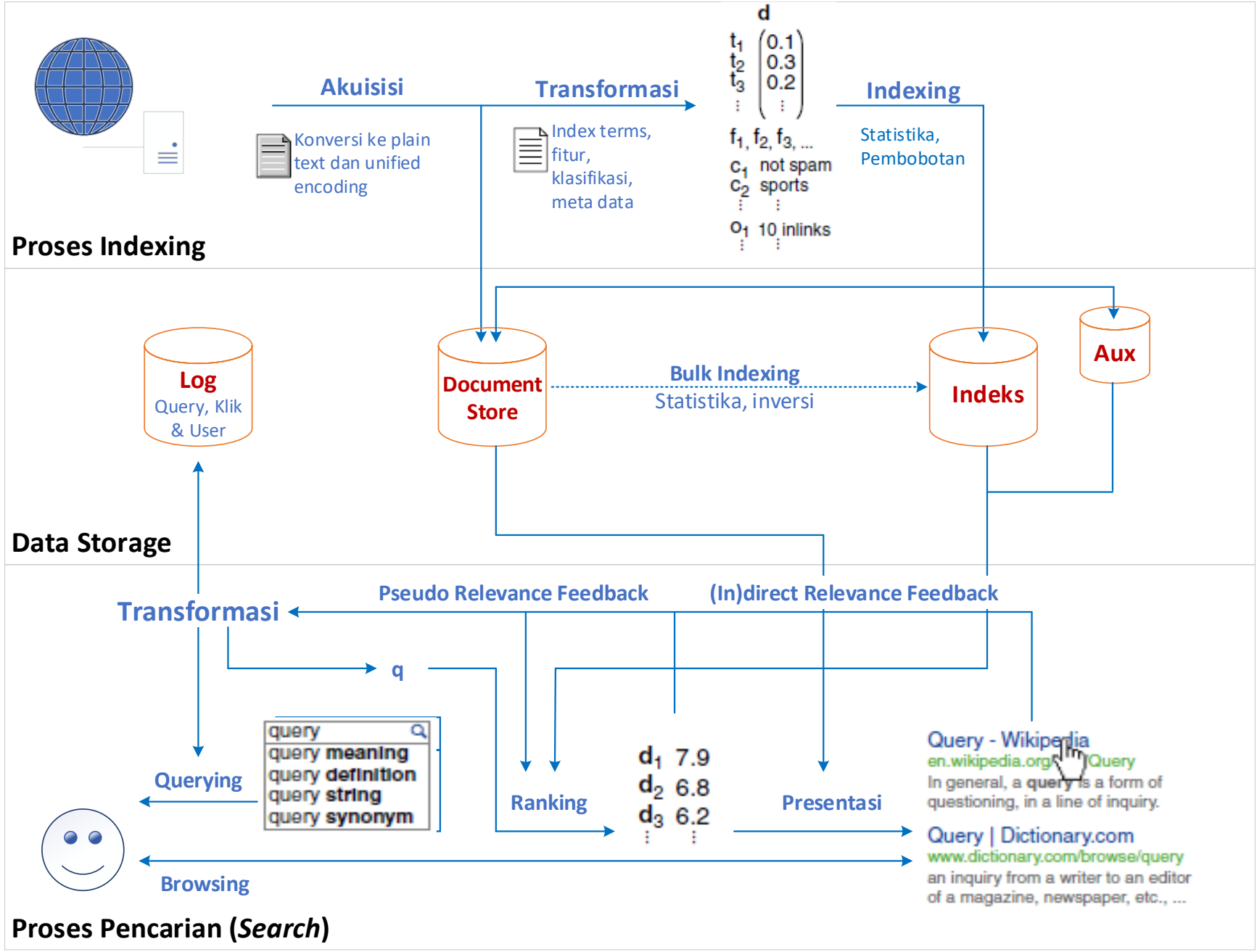

Akuisisi
Konversi ke plain text dan unified encoding
Transformasi
Index terms, fitur, klasifikasi, meta data
d
t1 0.1
t2 0.3
t3 0.2
f1, f2, f3, ...
c1 not spam
c2 sports
o1 10 inlinks
Indexing
Statistika, Pembobotan
Proses Indexing
Log
Query, Klik & User
Document Store
Bulk Indexing
Statistika, inversi
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)

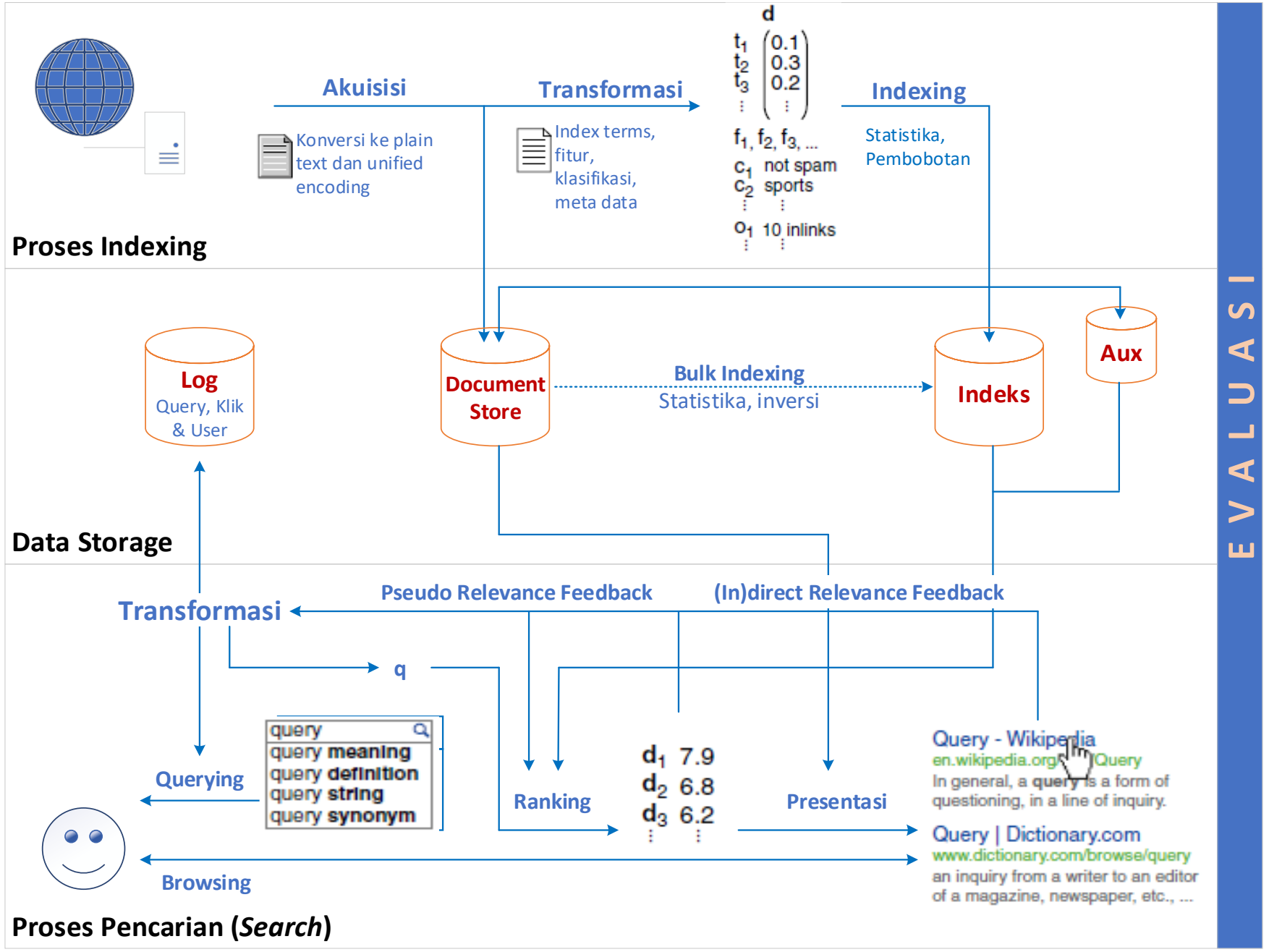
Proses Pencarian: Logging

Catat semua aktivitas pengguna, khususnya Query dan interaksi dengan hasil pencarian.

- **Pemahaman dan Penyaringan Query**
  Log melayani langkah transformasi Query.
- **Umpan-Balik Relevansi (Tak)langsung**
  Log menyimpan langkah transformasi Query dan peringkat dokumen
  Umpan-balik langsung (jarang diterapkan) menanyakan kepada pengguna dokumen mana yang dikembalikan relevan, menggunakan informasi ini untuk menyaring Query dan meranking ulang semua dokumen.
  Umpan-balik tidak langsung mengeksploitasi perilaku pengguna lain pada Query serupa untuk menyaring Query dan melatih algoritma ranking untuk mendapatkan peringkat yang lebih baik di masa mendatang. Data klik hasil dan estimasi waktu *dwell* (tinggal) dianalisis.
- **Personalisasi & Sasaran Iklan**
  Profil pengguna digunakan untuk menyesuaikan hasil pencarian dan iklan dengan konteks dan minat pengguna.
- **Analisis dan Optimisasi Pengalaman Pengguna**
  Perilaku pengguna pada halaman web mesin pencari digunakan untuk mendukung pengguna. Contoh: pewarnaan hasil pencarian.

Log adalah data paling berharga yang dikumpulkan oleh mesin pencari. Mesin pencari khusus untuk pencarian berbasis log mulai menjadi trend beberapa tahun terakhir.


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

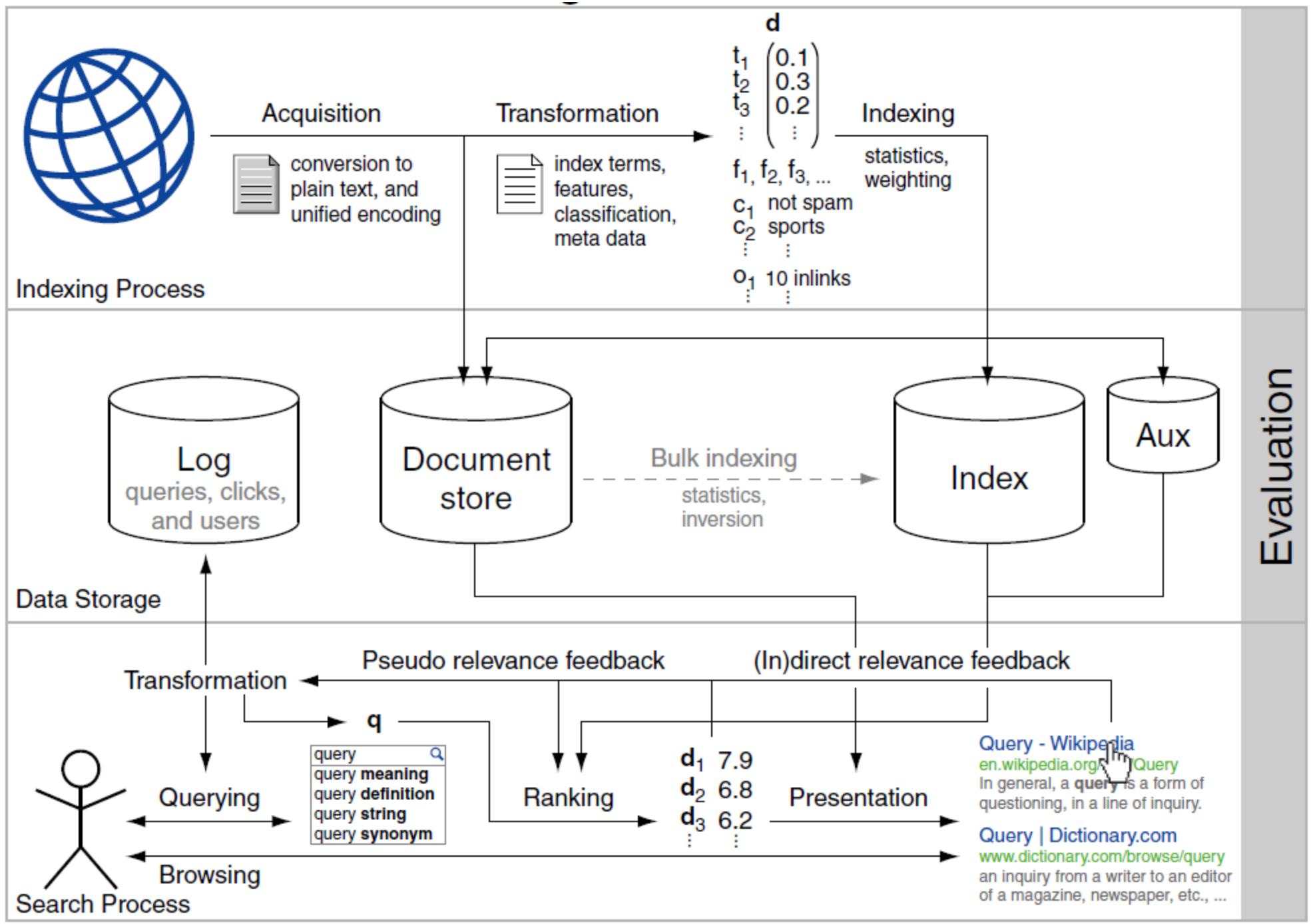
Evaluasi: Ikhtisar

Evaluasi mesin pencari membahas analisis efektivitas dan efisiensinya.

- Analisisi Ranking
  - Definisi dari tujuan
  - Akuisisi penilaian relevansi untuk pasangan Dokumen-Query (misalnya *crowdsourcing*)
  - Teori pengukuran (misalnya, penekanan kuat pada hasil teratas penelusuran web)
  - Analsis Log dari perilaku pencarian yang terekam.
  - Kajian Pengguna dan pengujian A/B.
- Analisis Pengalaman Pengguna
  - Definisi dari tujuan (misal usability, kepuasan pengguna, dll.)
  - Analisis Log dari perilaku pengguna terekam
  - Kajian pengguna dan penguji A/B
- Analisis Kinerja
  - Definisi dari tujuan (misal throughput, response time, dll.)
  - Analisis Log dari perilaku sistem terekam.
  - Eksperimen Lab dan simulasi.

Catatan

- Evaluasi adalah penentuan sistematis Merit, worth, dan signifikansi subjek, menggunakan kriteria yang diatur oleh seperangkat standar.
- Ini dapat membantu untuk memastikan tingkat pencapaian atau nilai sehubungan dengan tujuan dan sasaran yang dicari..
- Tujuan utama evaluasi, selain untuk mendapatkan wawasan tentang inisiatif sebelumnya atau yang sudah ada, adalah untuk memungkinkan refleksi dan membantu dalam identifikasi perubahan masa depan..
- [Wikipedia]


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., ...

Tugas Personal: Membuat Paper Survey

- Mahasiswa TKI2018 diharapkan mampu menulis paper survey yang komprehensif mengenai salah satu bidang Information Retrieval dan mungkin selanjutnya dapat dijadikan topik kajian Skripsi. Paper ini harus mengacu (mensitasi) setidaknya 18 paper 5 tahun terakhir.
- Silakan ambil Template contoh untuk Term Paper yang bagus di sini.
- Pada akhir semester, setiap mahasiswa (mandiri) harus memberikan presentasi mengenai topik yang dipilih, direview secara sistematis dan dituliskan menjadi paper survey tersebut.
- Didiskusikan pekan depan, Siapkan diri anda dengan setidaknya 2-3 sumber (referensi) awal dari topik yang anda pilih.

Tugas Kelompok: Proyek Pemrograman SE

- Buat kelompok 4 s.d 5 mahasiswa
- Membuat Search Engine sederhana
- Ikuti arsitektur yang ada...
- Bahasa pemrograman bebas... Sebaiknya Java atau Python
- Laporan progress 2 mingguan...
- Laporan akhir + Presentasi di bulan Desember 2018