Crawler

Spiders, Robot Web, Bots

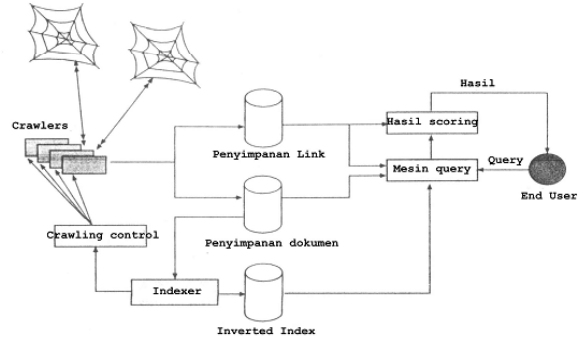
Crawler

- **Crawling** adalah proses pengambilan sejumlah besar halaman Web -- *dengan cepat* -- ke dalam suatu tempat penyimpanan lokal dan mengindeksnya berdasar sejumlah kata kunci.
- Idealnya (untuk ukuran saat ini) program akan mengambil puluhan ribu halaman Web per detik.
- Crawler adalah salah satu komponen utama dalam sebuah *Search Engine*, sebagai aplikasi Information Retrieval modern.

Search Engines

- Berdasarkan Pew Internet Project Report (2002):
  - Lebih dari 25% dari pengguna Internet memperoleh informasi pada Web dengan bantuan search engines.
  - Lebih dari 80% pengguna Internet telah memakai search engine paling sedikit sekali.
- Desain dari sebuah search engine dapat dibedakan dengan banyak kriteria. Perbedaan yang paling penting mungkin antara kedua jenis berikut:
  - Search engine yang mempunyai **fungsi umum** (yang ditujukan untuk menyediakan jawaban umum untuk kemungkinan yang paling besar dari user), dan
  - Search engine dengan **fungsi khusus** yang memfokuskan pada topik atau domain yang spesifik, misal untuk medis, hukum, biotechnology, dan lain sebagainya.

Arsitektur Umum sebuah Search Engine



Tinjauan pada Indexing:

Hubungan Crawler dengan Information Retrieval

- Bagaimana cara mengolah data hasil crawl ke dalam suatu sistem index yang dipakai untuk menjawab *query*.
- Indexing memungkinkan penanganan sejumlah keywords dengan kombinasi Boolean pada query.
- Pendekatan yang dilakukan **tidak seperti** query pada database relational, seperti SQL (bukan exact match).
- Query engine harus dengan mudah mengembalikan semua hasil yang memenuhi persyaratan yang diberikan melalui *"perintah yang ditulis sesukanya"* dalam ranking relevansi yang dapat dipercaya.
- Pada bagian inilah teori IR (Information Retrieval) berperan.

Suplemen

- *Silakan dipelajari sendiri beberapa "catatan pendek" saya yang mencakup:*
  - *Kondisi WWW pada tahun 2002-2003*
    - *Walau sudah lama, tetap terdapat beberapa fakta yang menarik.*
  - *Sejarah*
    - *Lycos, WebCrawler, Alta Vista, HotBot dan Inktomi*
  - *Direktori Topik*
    - *Yahoo.com, About.com, dan Open Project (dmoz.org)*

Mekanisme Dasar Crawling

- Algoritma **SIMPLE_CRAWLER** berikut merupakan versi yang sudah disederhanakan hanya untuk menunjukkan mekanisme dasar proses crawling.
- Algoritma ini pada dasarnya melakukan suatu kunjungan graf (Graph Traversing).
- Input:
  - S_0 dari URL (sering disebut sebagai *seeds*)
- Output:
  - *D* dan *E*, masing-masing adalah tempat penyimpanan dokumen dan link yang saling berkorespondensi.

PROCEDURE Simple_Crawler(S_0, D, E)

```
1.  Q <--- S0
2.  DO WHILE NOT (isQueueEmpty(Q))
3.          u <--- Dequeue(Q)
4.          d(u) <--- Fetch(u)
5.          CALL Store(D, (d(u),u))
6.          L <--- Parse(d(u))
7.          FOR EACH v IN L
8.                  CALL Store(E, (u,v))
9.                  IF NOT (v ∈ D  OR v ∈ Q)
10.                         THEN Enqueue(Q,v).
11.         END FOR
12. END DO
```

Beberapa Tuntutan pada Crawler (#1)

- **Resuming**
  - Di bawah asumsi dari Web yang statik, isi dari *Q* dan *D* benar-benar berfungsi untuk menjelaskankan *state* dari crawler pada saat tertentu. Dalam pemikiran bahwa proses dapat dimulai lagi dengan tidak ada informasi yang hilang jika *Q* dan *D* dipertahankan. Kita juga harus selalu mengingat perbedaan yang penting antara mengunjungi graf (atau solution tree) dan crawling. Suatu graf (atau solution tree) normalnya dianggap sebagai suatu object yang statik selama eksekusi dari algoritma crawling. Kenyataannya, Web adalah dinamik dan berubah terus-menerus baik pada *content*nya maupun *topology*nya. ***(UTS Genap 2009/2010 sampai disini)***