# Search Engine

**Fauzan Azmi**
*azmifauzan@gmail.com*
*http://www.azmifauzan.net*



## Apa itu Search Engine

Mesin pencari web atau yang lebih dikenal dengan istilah web search engine merupakan program komputer yang dirancang untuk mencari informasi yang tersedia didalam dunia maya [4]. Berbeda halnya dengan direktori web (seperti dmoz.org) yang dikerjakan oleh manusia untuk mengelompokkan suatu halaman informasi berdasarkan kriteria yang ada, web search engine mengumpulkan informasi yang tersedia secara otomatis.

## Cara Kerja *Search Engine*

Mesin pencari web bekerja dengan cara menyimpan hampir semua informasi halaman web, yang diambil langsung dari *www*. Halaman-halaman ini diambil secara otomatis. Isi setiap halaman lalu dianalisis untuk menentukan cara mengindeksnya (misalnya, kata-kata diambil dari judul, subjudul, atau *field* khusus yang disebut meta tag). Data tentang halaman web disimpan dalam sebuah database indeks untuk digunakan dalam pencarian selanjutnya. Sebagian mesin pencari, seperti Google, menyimpan seluruh atau sebagian halaman sumber (yang disebut cache) maupun informasi tentang halaman web itu sendiri.

Ketika seorang pengguna mengunjungi mesin pencari dan memasukkan *query*, biasanya dengan memasukkan kata kunci, mesin mencari indeks dan memberikan daftar halaman web yang paling sesuai dengan kriterianya, biasanya disertai ringkasan singkat mengenai judul dokumen dan terkadang sebagian teksnya.

Mesin pencari lain yang menggunakan proses *real-time*, seperti Orase, tidak menggunakan indeks dalam cara kerjanya [4]. Informasi yang diperlukan mesin tersebut hanya dikumpulkan jika ada pencarian baru. Jika dibandingkan dengan sistem berbasis indeks yang digunakan mesin-mesin seperti Google, sistem *real-time* ini unggul dalam beberapa hal seperti informasi selalu mutakhir, (hampir) tak ada *broken link*, dan lebih sedikit sumberdaya sistem yang diperlukan (Google menggunakan hampir 100.000 komputer, Orase hanya satu.). Tetapi, ada juga kelemahannya yaitu pencarian lebih lama rampungnya.

## Komponen utama dalam *Search Engine*

Sebuah search engine memiliki beberapa komponen agar dapat menyediakan layanan utamanya sebagai sebuah mesin pencari informasi. Komponen tersebut antara lain [2] :

1. *Web Crawler*
2. *Indexing System*
3. *Search System*

### *Web Crawler*

*Web crawler* atau yang dikenal juga dengan istilah *web spider* bertugas untuk mengumpulkan semua informasi yang ada di dalam halaman web. *Web crawler* bekerja secara otomatis dengan cara memberikan sejumlah alamat website untuk dikunjungi serta menyimpan semua informasi yang terkandung didalamnya. Setiap kali *web crawler* mengunjungi sebuah website, maka dia akan mendata semua link yang ada dihalaman yang dikunjunginya itu untuk kemudian di kunjungi lagi satu persatu.

Proses *web crawler* dalam mengunjungi setiap dokumen web disebut dengan web crawling atau spidering. Beberapa websites, khususnya yang berhubungan dengan pencarian menggunakan proses spidering untuk memperbaharui data data mereka. Web crawler biasa digunakan untuk membuat salinan secara sebhagian atau keseluruhan halaman web yang telah dikunjunginya agar dapat dip roses lebih lanjut oleh system pengindexan. Crawler dapat juga digunakan untuk proses pemeliharaan sebuah website, seperti memvalidasi kode html sebuah web, dan crawler juga digunakan untuk memperoleh data yang khusus seperti mengumpulkan alamat e-mail.

*Web crawler* termasuk kedalam bagian *software agent* atau yang lebih dikenal dengan istilah program *bot*. Secara umum *crawler* memulai prosesnya dengan memberikan daftar sejumlah alamat website untuk dikunjungi, disebut sebagai *seeds.* Setiap kali sebuah halaman web dikunjungi, *crawler* akan mencari alamat yang lain yang terdapat didalamnya dan menambahkan kedalam daftar *seeds* sebelumnya.

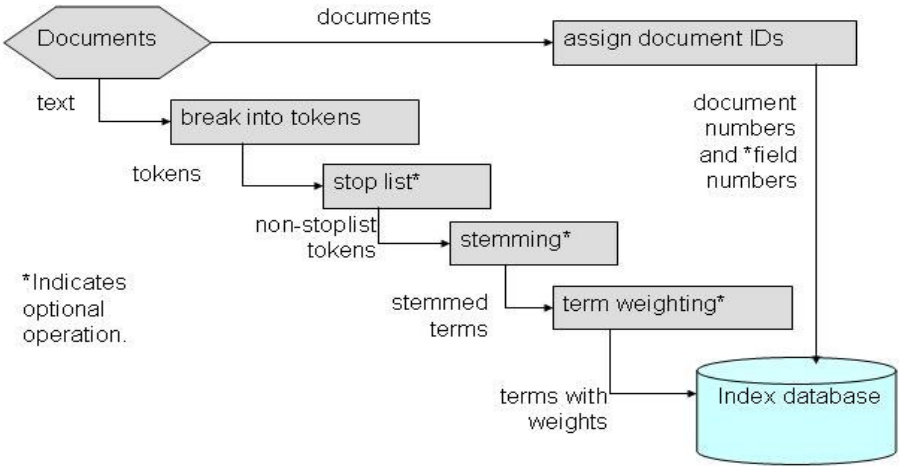
Dalam melakukan prosesnya, web crawler juga mempunyai beberapa persoalan yang harus mampu di atasinya. Permasalahan tersebut mencakup [3]:

- Halaman mana yang harus dikunjungi terlebih dahulu.
- Aturan dalam proses mengunjungi kembali sebuah halaman.
- Performansi, mencakup banyaknya halaman yang harus dikunjungi.
- Aturan dalam setiap kunjungan agar server yang dikunjungi tidak kelebihan beban.
- Kegagalan, mencakup tidak tersedianya halaman yang dikunjungi, *server down, timeout,* maupun jebakan yang sengaja dibuat oleh webmaster.
- Seberapa jauh kedalaman sebuah website yang akan dikunjungi.
- Hal yang tak kalah pentingnya adalah kemampuan *web crawler* untuk mengikuti perkembangan teknologi web, dimana setiap kali teknologi baru muncul, *web crawler* harus dapat menyesuaikan diri agar dapat mengunjungi halaman web yang menggunakan teknologi baru tersebut.

Proses sebuah *web crawler* untuk mendata *link – link* yang terdapat didalam sebuah halaman web menggunakan pendekatan *regular expression*. *Crawler* akan menelurusi setiap karakter yang ada untuk menemukan *hyperlink* tag html (<a>). Setiap hyperlink tag yang ditemukan diperiksa lebih lanjut apakah tag tersebut mengandung atribut *nofollow rel*, jika tidak ada maka diambil nilai yang terdapat didalam attribute *href* yang merupakan sebuah link baru.

***Indexing system***

*Indexing system* bertugas untuk menganalisa halaman web yang telah tersimpan sebelumnya dengan cara mengindeks setiap kemungkinan term yang terdapat di dalamnnya. Data term yang ditemukan disimpan dalam sebuah database indeks untuk digunakan dalam pencarian selanjutnya.




*Indexing System [1]*

*Indexing system* mengumpulkan, memilah dan menyimpan data untuk memberikan kemudahan dalam pengaksesan informasi secara tepat dan akurat. Proses pengolahan halaman web agar dapat digunakan untuk proses pencarian berikutnya dinakamakan *web indexing*. Dalam implementasinya *index system* dirancang dari penggabungan beberapa cabang ilmu antara lain ilmu bahasa, psikologi, matematika, informatika,

fisika, dan ilmu komputer.

Tujuan dari penyimpanan data berupa indeks adalah untuk performansi dan kecepatan dalam menemukan informasi yang relevan berdasarkan inputan user. Tanpa adanya indeks, *search engine* harus melakukan *scan* terhadap setiap dokumen yang ada didalam database. Hal ini tentu saja akan membutuhkan proses sumber daya yang sangat besar dalam proses komputasi. Sebagai contoh, indeks dari 10.000 dokumen dapat diproses dalam waktu beberapa detik saja, sedangkan penulusuran secara berurutan setiap kata yang terdapat di dalam 10.000 dokumen akan membutuhkan waktu yang berjam lamanya. Tempat tambahan mungkin akan dibutuhkan di dalam computer untuk penyimpanan indeks, tapi hal ini akan terbayar dengan penghematan waktu pada saat pemrosesan pencarian dokumen yang dibutuhkan.

Faktor utama yang harus diperhatikan pada saat pembangunan *index system* antara lain [4] :

1. *Merge Factors*

   Index system harus dapat membedakan pada saat sebuah indeks baru ditambahkan kedalam database, apakah indeks tersebut menambahkan data baru atau hanya memperbaharui data yang lama.
2. *Storage Techniques*

   Bagaimana menyimpan sebuah indeks itu sendiri, apakah indeks disimpan dalam bentuk data terkompresi atau harus di saring terlebih dahulu.
3. *Index Size*

   Berapa banyak ukuran yang harus disiapkan untuk dapat menampung semua indeks dokumen web.
4. *Lookup Speed*

   Kecepetan pencarian indeks didalam database juga perlu diperhitungkan, karena indeks yang di simpan dalam jumlah yang sangat banyak.
5. *Maintenance*

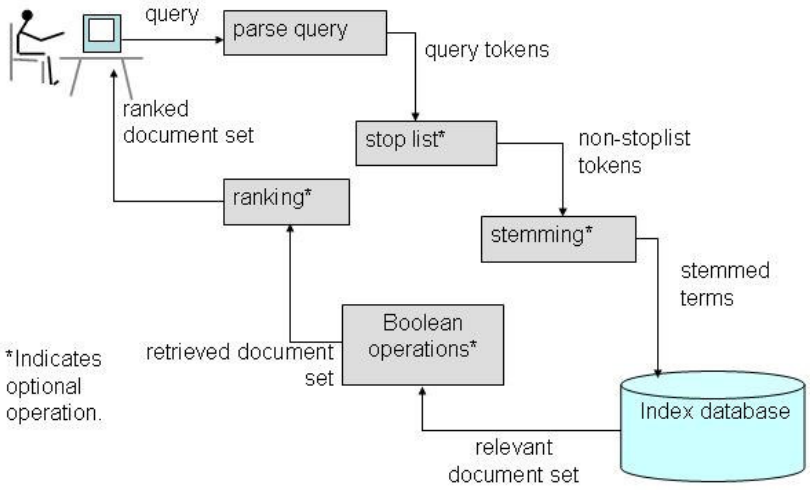
Bagaimana *index system* dapat memelihara data indeks yang sudah tersimpan.

6. *Fault Tolerance*

   Fakator kegagalan juga harus diperhitungkan, baik itu berupa kegagalan perangkat keras, maupun kegagalan yang disebabkna oleh system itu sendiri.

***Search system***

Search system inilah yang berhubungan langsung dengan pengguna, meyediakan hasil pencarian informasi yang diinginkan. Ketika seorang pengguna mengunjungi mesin pencari dan memasukkan kata pencarian biasanya dengan beberapa kata kunci, search system akan mencari data dari indeks database, data yang cocok kemudian akan ditampilkan, biasanya disertai ringkasan singkat mengenai judul dokumen dan terkadang sebagian teksnya.




*Search system [1]*

## Referensi

[1] Firdaus, Yanuar. 2008 : *Text Processing Methods*. Bandung : IT Telkom. available at : http://www.ittelkom.ac.id/staf/yfa/

[2] Firdaus, Yanuar. 2008 : *Web Search*. Bandung : IT Telkom. available at : http://www.ittelkom.ac.id/staf/yfa/

[3] Firdaus, Yanuar. 2008 : *Web Search 2*. Bandung : IT Telkom. available at : http://www.ittelkom.ac.id/staf/yfa/

[4] Wikipedia. *PageRank*. http://en.wikipedia.org/wiki/Index_(search_engine). Diakses tanggal 25 juli 08.

[5] Wikipedia. *Web Crawler*. available at : http://en.wikipedia.org/wiki/WebCrawler. diakses tanggal 25 juli 08.

## Biografi Penulis

**Fauzan Azmi**, Lahir di Padang, 21 Oktober 1985. Menyelesaikan Program Diploma 3 pada jurusan Teknik Informatika di STT Telkom, Bandung pada tahun 2006. Saat ini sedang melanjutkan pendidikan S1 di institusi yang sama. Informasi lebih lanjut mengenai penulis dapat ditemukan pada websitenya http://www.azmifauzan.net