Akurasi dalam Pencarian pada *Search Engines*

Gita Desrianti 18209030
Program Studi Sistem dan teknologi Informasi
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
18209030@std.stei.itb.ac.id

***Abstrak*—Keberadaan informasi yang jumlahnya tidak terhitung terutama yang beredar di dunia maya serta kebutuhan manusia akan penggunaan informasi-informasi tersebut mendorong berkembangnya suatu ilmu untuk melakukan pencarian informasi atau information retrieval.**

Salah satu bentuk information retriever yang sering digunakan adalah web search engine yang memfasilitasi pengguna untuk melakukan pencarian suatu informasi dari seluruh halaman html dan dokumen lainnya yang ada di world wide web sesuai dengan query yang dimasukkan oleh pengguna yang biasanya berupa kata-kata kunci dari informasi yang diinginkan.

Performansi dari sebuah search engine secara umum diukur dari tingkat presisi dan recall search engine itu sendiri. Presisi adalah perbandingan jumlah informasi relevan yang di-retrieve dan ditemukan oleh search engine dengan seluruh jumlah informasi yang di-retrieve oleh search engine. Recall adalah perbandingan jumlah informasi relevan yang berhasil di-retrieve oleh search engine dengan seluruh jumlah informasi yang relevan dengan query yang dimasukkan oleh pengguna.

Performansi ini tentu saja sangat dipengaruhi oleh proses yang bekerja pada search engine. Terdapat tiga perintah utama yang mendasari kerja suatu search engine yaitu web crawler, indexing, dan searching.

Ketiga proses ini tentu saja memiliki kelemahan, sensitivitas, dan tingkat error tertentu yang sangat mempengaruhi tingkat akurasi dalam pencarian pada search engine itu sendiri. Karena ketiga perintah ini bersifat saling lepas maka tingkat akurasi pencarian suatu search engine adalah hasil perkalian dari tingkat efektifitas, efisiensi, serta sensitivitas dari masing-masing proses.

***Kata Kunci*—Information Retrieval, Search Engine, *Presisi*, Recall.**

I. PENDAHULUAN

Saat ini, jumlah informasi yang tersimpan di dunia maya semakin banyak dan tidak terhitung. Untuk mengetahui dan menggunakan informasi-informasi tersebut, informasi yang dibutuhkan tersebut harus dicari dari sejumlah informasi yang ada. Sistem pencarian informasi dari sejumlah dokumen yang ada dipelajari dalam suatu ilmu yaitu *information retrieval*.

Salah satu bentuk sistem pencarian informasi atau *information retrieval* untuk informasi-informasi yang ada di dunia maya adalah *search engines*. *Search engine* menyimpan berbagai dokumen html sehingga ketika pengguna memasukkan *query*, *search engine* ini akan me-*retrieve* dokumen-dokumen yang ada dan mengurutkannya berdasarkan tingkat relevansinya dengan *query* yang dimasukkan pengguna.

Pada makalah ini akan dibahas penghitungan tingkat akurasi suatu pencarian pada *search engines* berdasarkan *query*-nya. Ilmu yang diterapkan pada perhitungan tingkat akurasi ini adalah probabilitas dan statistika.

II. *INFORMATION RETRIEVAL*

Dari http://en.wikipedia.org/wiki/Information_retrieval *information retrieval* didefinisakan sebagai ilmu untuk mencari dokumen, informasi pada *dokumen*, atau metadata pada dokumen, serta mencari database dan *world wide web* yang berelasi/ berhubungan terdapat overlap dalam penggunaan istilah *data retrieval*, *document retrieval*, dan *text retrieval*, tetapi masing-masingnya memiliki literature, teori, praksis, dan teknologi tersendiri. IR merupakan disiplin ilmu yang berbasiskan *computer science*, matematika, ilmu kepustakaan, ilmu informasi, arsitektur informasi, psikologi kognitif, linguistik, dan statistik.

> **Information retrieval** (**IR**) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis, and technologies. IR is interdisciplinary, based on computer science, mathematics, library science, information science, information architecture, cognitive psychology, linguistics, and statistics.

Kebutuhan akan *information retriever* di kehidupan kita sejak awal sangat tinggi. Pada awalnya, informasi-informasi tersebut hanya didokumentasikan dan diberi indeks atau alamat sehingga lebih teratur dan memudahkan pencarian. Namun, dengan semakin

banyaknya informasi yang ada, dokumentasi dan pemberian indeks saja tidak cukup untuk mencari suatu informasi di suatu kumpulan informasi yang jumlahnya sangat banyak, apalagi jika harus mencari isi dari dokumen-dokumen yang berisi informasi tersebut, bukan hanya judul, atau kategori yang cenderung lebih mudah dicari. Pada awalnya, information retrieval ini hanya terbatas pada suatu lingkup kecil. Namun, kebutuhan informasi yang semakin meninggi memperluas lingkup pencarian bahkan saat ini, suatu information retriever mampu mencari informasi dari seluruh dunia.

Salah satu *information retriever* yang paling umum digunakan adalah *web search engines* yang mampu mencari informasi di seluruh halaman web di dunia maya.

Interaksi suatu *information retriever* dengan pengguna adalah dengan meminta kata kunci dari informasi yang akan dicari dari sekumpulan informasi yang ada. Kata-kata kunci yang dimasukkan ini disebut dengan *query* dan kemudian *information retriever* ini akan memunculkan informasi-informasi yang relevan dari informasi yang ada dengan melihat kecocokan *query* dengan informasi yang ada.

III. *Search Engines*

Menurut Wikipedia (http://en.wikipedia.org/wiki/Web_search_engine), "A **web search engine** is designed to search for information on the World Wide Web and FTP servers. The information may consist of web pages, images, information and other types of files". *Search engines* adalah mesin yang didesain untuk mencari informasi yang terdapat pada world wide web dan FTP. Informasi yang dicari bisa saja berbentuk halaman web, gambar, informasi, dan berbagai macam jenis file lainnya.

Search engines bekerja dengan cara menyimpan berbagai halaman web yang di-*retrieve* dari berbagai html. Ketika pengguna memasukkan suatu kata kunci atau rangkaian kata kunci, *search engines* akan mencari halaman-halaman web yang memiliki kecocokan dengan kata kunci tersebut baik dari alamatnya, judulnya, atau pun isinya.

Kualitas suatu *search engines* ditentukan dari kemampuannya mencari informasi yang dibutuhkan serta kemampuannya me-*ranking* informasi-informasi yang di-*retrieve* dari yang paling relevan sampai ke yang paling tidak relevan.

Terdapat banyak web search engine yang dapat diakses secara mudah saat ini, misalnya google, bing, msn, yahoo, dll. Pada dasarnya, berbagai web sudah memiliki search engine sendiri namun perbedaan antara search engine satu dengan search engine lainnya adalah batasan dokumen/ informasi yang di-retrieve, jumlah dokumen/ informasi yang di-retrieve, serta performansinya dalam melakukan pencarian sesuai dengan query yang dimasukkan oleh pengguna.

Bukan hanya mencari semua informasi yang di-retrieve, beberapa search engine bahkan memiliki beberapa fitur untuk memudahkan pengguna dalam melakukan pencarian. Misalnya, tombol "I'm feeling lucky" pada google untuk memberikan kemudahan pada pencarian yang cenderung pasti atau dimana suatu halaman merupakan kemungkinan terbesar jawaban dari query yang dimasukkan. Dengan tombol ini, hasil pencarian terhadap suatu query tidak ditampilkan kepada user tetapi search engine ini akan secara otomatis mengarahkan untuk masuk ke halaman web tersebut. Namun, probabilitas suatu hasil pencarian tersebut harus cukup besar atau google akan menampilkan kembali seluruh kemungkinan informasi yang relevan dengan dokumen tersebut.

Selain itu, di beberapa *search engine*, setiap kali suatu query dimasukkan mesin akan mengecek kembali query tersebut dan menyesuaikan. Apabila terjadi kesalahan penulisan, mesin akan menampilkan hasil pencarian dari kata yang mungkin yang terdekat dengan query yang kita masukkan.

Terdapat juga pengelompokkan jenis informasi yang di-*retrieve* misalnya gambar, halaman web, peta, dll sehingga memudahkan pengguna untuk melakukan pencarian sesuai dengan kategori yang diinginkan.

IV. Performansi *Retrieval* Suatu *Search Engines*

Setiap *search engine* memiliki kemampuan untuk me-*retrieve* informasi-informasi pada berbagai halaman html yang mereka simpan pada suatu database. Tingkat performansi suatu *search engine* secara umum diukur dengan dua parameter yaitu *recall* dan *precision*.

Precision

Presisi adalah probabilitas jumlah dokumen yang relevan dari semua dokumen yang di-*retrieve* dibandingkan dengan jumlah keseluruhan dokumen yang di-*retrieve*. Secara matematis, presisi dapat dituliskan sebagai:

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Pada penerapannya pada *search engine*, jumlah presisi ini sangat sulit diperhitungkan karena jumlah informasi yang sangat banyak dan jumlah informasi yang di-retrieve yang selalu bertambah dan di-update setiap waktunya. Peng-update-an ini menambah jumlah dokumen yang di-retrieve dan mungkin menambah jumlah dokumen yang relevan. Namun, penambahan ini cenderung menurunkan tingkat presisi suatu *search engine* karena jumlah dokumen yang harus di-retrieve pasti jauh lebih banyak dibandingkan dengan dokumen yang relevan.

B. *Recall*

Recall adalah probabilitas dokumen relevan yang di-*retrieve* dibandingkan dengan jumlah dokumen-dokumen yang relevan. Secara matematis *recall* dituliskan sebagai:

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Recall merupakan ukuran sensitivitas suatu *information retriever* dalam menemukan informasi-informasi yang sebenarnya relevan dengan *query* yang dimasukkan.

Hal ini sangat berhubungan dengan kemampuan *information retriever* ini untuk menyimpan dokumen-dokumen dan informasi-informasi lainnya khususnya informasi yang relevan dengan informasi yang dibutuhkan.

V. CARA KERJA *SEARCH ENNGINE*

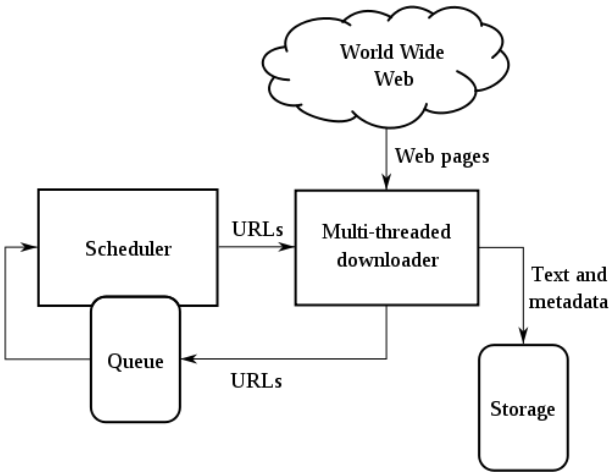
Pada dasarnya, semua *search engine* beroperasi dengan basic yang sama yaitu 3 buah perintah dasar. Ketiga perintah ini adalah *web crawling*, pemberian indeks (*indexing*), dan pencarian (*searching*).

Setiap operasi ini merupakan elemen yang penting dan krusial pada setiap *search engine* untuk meningkatkan performansinya, yaitu *recall* dan presisi.

Web Crawling

Web crawler adalah program computer yang melakukan penelusuran terhadap world wide web dan menyimpan informasi-informasi yang ada pada suatu storage dengan teratur.

Web crawler bekerja dengan arsitektur seperti gambar dibawah ini.



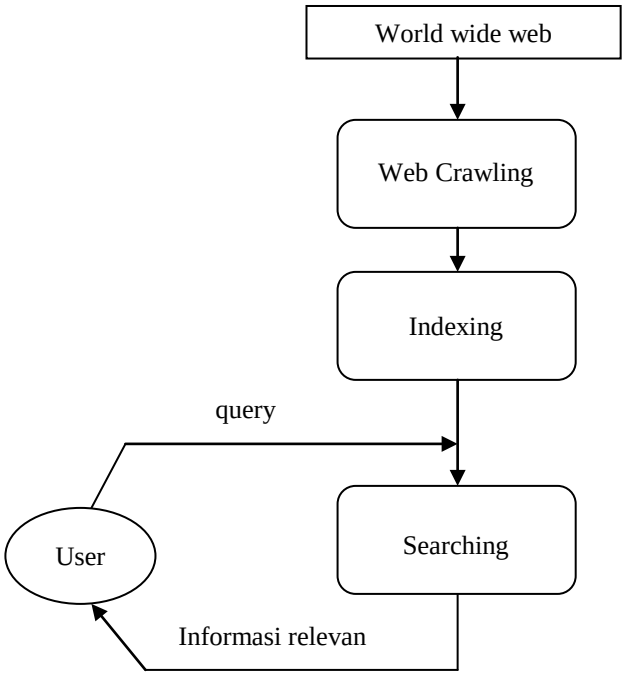
(http://en.wikipedia.org/wiki/Web_crawler#Web_crawler_architectures).

Indexing

Setelah data disimpan pada storage, *search engine* akan melakukan *indexing* yaitu mengumpulkan, mengelompokkan, dan menyimpan data dengan lebih teratur sehingga dapat memfasilitasi proses pencarian pada *search engine* oleh pengguna dengan lebih cepat dan efektif. (http://en.wikipedia.org/wiki/Index_%28search_engine%29).

Searching

Setelah informasi dikelompokkan dan disimpan dengan teratur, proses *search engine* yang terakhir adalah melakukan pencarian terhadap informasi yang dibutuhkan berdasarkan kata kunci yang dimasukkan.



VI. PEMBAHASAN

Masing-masing dari ketiga perintah ini sangat mempengaruhi nilai presisi dan *recall* dari suatu *search engine*.

Perintah pertama yaitu *web crawling* sangat mempengaruhi jumlah informasi yang di-*retrieve* dan jumlah informasi yang relevan dengan *query* yang diminta oleh pengguna. Secara gamblang, *web crawling* ini merupakan unsure terpenting dalam menentukan tingkat *recall* dan presisi dari suatu *search engine* karena baik *recall* maupun presisi sangat bergantung pada jumlah informasi, baik yang relevan maupun informasi yang di-*retrieve*.

Perintah kedua yaitu *indexing* merupakan cara untuk meningkatkan performansi suatu search engine terutama bila dilihat dari efektifitas dan efisiensinya. Dengan indexing, maka informasi-informasi yang memiliki kesamaan akan dikelompokkan sehingga secara tidak langsung nilai *recall* dan presisi dapat ditingkatkan. Karena dengan pengelompokan ini, jumlah informasi yang

di-*retrieve* akan lebih sedikit namun jumlah informasi yang relevan diharapkan akan lebih banyak dengan tingkat relevansi yang tinggi.

Perintah ketiga yaitu pencarian atau *searching*. Pada operasi inilah proses me-*retrieve* berlangsung. Jumlah dokumen yang di-*retrieve,* jumlah dokumen yang relevan, dan penentuan jumlah dokumen yang memiliki relevansi dengan informasi yang diinginkan pengguna dilakukan. Secara tidak langsung, proses *searching* inilah yang menentukan relevansi suatu informasi dengan informasi yang dicari oleh pengguna.

Seperti pada proses-proses lainnya, pada proses *information retrieval* pada *search engine* terdapat juga kebocoran atau beberapa error yang terjadi sehingga menurunkan performansi dari suatu *search engine* tersebut. Dari ketiga perintah yang mendasari suatu *search engine*, masing-masingnya memiliki galat terhadap akurasi pencarian pada suatu search engine pada setiap perintahnya.

Pada saat web crawling, galat yang mungkin terjadi disebabkan oleh:

1. Adanya www yang tidak ditelusuri
2. Adanya www yang tidak berhasil di-*download* untuk diberi indeks dan disimpan dalam storage
3. Adanya halaman URL yang tidak masuk pada queue
4. Adanya halaman yang tidak tersimpan pada saat storage.

Sedangkan pada saat indexing, terdapat kemungkinan galat dari kesalahan pengelompokkan suatu www atau pada saat penyimpanan hasil pemberian indeks itu kedalam storage. Tentu saja kesalahan pengelompokan ini sangat mempengaruhi terutama tingkat sensitivitas dari suatu *search engine* sendiri atau dengan kata lain tingkat *recall* pada *information retrieval* pada *search engine* tersebut.

Pada saat pencarian, terdapat dua sebab tidak relevannya hasil dokumen yang di-retrieve:

1. Kesalahan atau ketidaksesuaian input *query* oleh user pada saat melakukan pencarian.
2. Error yang disebabkan oleh search engine itu sendiri

Error yang disebabkan oleh search engine sendiri misalnya, kesalahan pembacaan query masukan user. Selain itu, apabila pada query terdapat beberapa kata kunci, search engine akan memprioritaskan kata-kata tertentu dibandingkan kata-kata lain (misalnya tidak secara khusus mencari semua informasi yang mengandung kata-kata preposisi pada query, dll). Kemungkinan kesalahan dapat disebabkan oleh kesalahan peletakkan prioritas kata kunci dari search engine itu sehingga hasil yang diberikan kurang relevan dengan hasil yang diharapkan. Kemungkinan ketiga adalah kesalahan pada pencarian, misalnya terlewatnya suatu halaman web yang relevan atau terikutnya suatu halaman web yang tidak relevan. Kemungkinan lainnya adalah salah peletakkan ranking relevansi pada informasi yang di-*retrieve*.

Kesalahan terakhir pada dasarnya tidak mempengaruhi tingkat presisi dan *recall* dari suatu search engine. Karena secara kuantitas, pengukuran presisi dan *recall* akan sama. Namun, dari sekian banyak informasi yang di-*retrieve* pengguna *search engine* cenderung hanya melihat beberapa informasi dalam bentuk halaman web atau dokumen yang menurut *search engine* yang digunakan merupakan informasi-informasi paling relevan diantara informasi-informasi yang relevan lainnya. Jadi, pada penggunaannya, kesalahan pengurutan ranking tingkat relevansi suatu dokumen/ informasi juga akan mempengaruhi tingkat akurasi dari suatu *search engine*.

Dari berbagai kemungkinan kesalahan yang terjadi pada suatu pencarian pada *search engine*, maka tingkat akurasi dari suatu *search engine* bergantung pada tingkat efisiensi pada setiap prosesnya yaitu *web crawling*, *indexing*, dan *searching*.

Meskipun proses dilakukan secara berkelanjutan, namun setiap prosesnya independen dengan proses lain. Jadi, untuk mengetahui akurasi total dari keseluruhan proses pada *search engine*, perlu dicari probabilitas seluruh proses (*web crawling, indexing,* dan *searching*) berjalan dengan baik atau dengan kata lain probabilitas dari irisan keberhasilan proses *web crawling,* digabung dengan keberhasilan proses *indexing,* dan keberhasilan proses *searching.*

Karena ketiga proses bersifat independen atau saling lepas, maka tingkat akurasi dari suatu *search engine* dapan dituliskan secara matematis sebagai:

$$\eta = \alpha \, . \, \beta \, . \, \gamma$$

dengan η adalah tingkat akurasi *search engine*
α adalah efektifitas *web crawler*
β adalah efektifitas dan efisiensi mesin indeks
γ adalah kemampuan / sensetivitas mesin pencari

Selain dari segi jumlah, kemampuan pengurutan ranking pada *search engine* juga sangat mempengaruhi tingkat akurasi pada pencarian oleh pengguna karena umumnya, pengguna cenderung hanya melihat halaman-halaman paling awal dari hasil pencarian karena *search engine* yang digunakan tentu saja menampilkan informasi-informasi paling relevan di halaman awal.

Terdapat 8 hal dasar yang umumnya digunakan oleh *search engine* untuk mendasari pengurutan tingkat relevansi suatu informasi pada halaman web.

1. Keyword pada domain name
   Domain name menjadi prioritas dari pencarian relevansi suatu pencarian karena suatu domain name dianggap mewakili keseluruhan isi dari website itu sendiri. Selain itu, domain name adalah hal pertama yang ditelusuri oleh suatu search engine.
2. Keyword pada nama file
   Seperti halnya pada halaman web, nama file pada dokumen yang ada adalah hal yang pertama kali ditelusuri pada dokumen sebelum isi dokumen itu.
3. Keyword pada page title

Setelah domain name dan nama file, keyword pada page title juga sangat mempengaruhi pada pencarian informasi yang relevan dengan *query*

4. Keyword pada headline
   Headline terutama pada berita merupakan gambaran umum dari isi informasi.
5. Keyword yang relevan pada meta tag
6. Keyword pada page content
   Isi informasi dari suatu halaman dapat ditelusuri pada page content
7. Meta tags atau tag-tag HTML khusus yang digunakan untuk menggambarkan keseluruhan dari isi website
8. Link Popularity
   Dengan tingginya popularitas suatu halaman web, *search engine* akan menganggap bahwa halaman tersebut cenderung relevan dengan kebanyakan informasi yang diharapkan.

VI. KESIMPULAN

Dari pembahasan diatas dapat diperoleh beberapa kesimpulan, yaitu :

1. *Information retrieval* adalah ilmu untuk mencari suatu informasi tertentu dari keseluruhan informasi yang ada.
2. *Search engine* adalah salah satu contoh *information retriever.*
3. Performansi suatu *search engine* dapat dilihat dari tingkat presisi dan *recall* dari hasil pencarian
4. *Search engine* bekerja dengan tiga perintah dasar yaitu *web crawling, indexing,* dan *searching.*
5. Tingkat akurasi dari suatu *search engine* dihitung dengan mengalikan tingkat efisiensi dan efektivitas pada proses *web crawling, indexing,* dan *searching*.
6. 8 hal yang mendasari perankingan tingkat relevansi pada *search engine* adalah keyword pada domain name, pada nama file, pada page title, pada headline, keyword yang relevan pada meta tag, keyword pada page content, meta tags, dan link popularity.

REFERENSI

http://en.wikipedia.org/wiki/Index_%28search_engine%29
http://en.wikipedia.org/wiki/Information_retrieval
http://en.wikipedia.org/wiki/Web_crawler#Web_crawler_architectures
http://en.wikipedia.org/wiki/Web_search_engine
http://www.kumpulancara.com/2009/08/8-trik-memperoleh-ranking-pada-search.html#axzz18KCnFqv9

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 17 Desember 2010

ttd

Gita Desrianti
18209030