INFORMATION RETRIEVAL

I. Pengertian Information Retrieval

Information Retrieval merupakan bagian dari *computer science* yang berhubungan dengan pengambilan informasi dari dokumen-dokumen yang didasarkan pada isi dan konteks dari dokumen-dokumen itu sendiri. Berdasarkan referensi dijelaskan bahwa Information Retrieval merupakan suatu pencarian informasi yang didasarkan pada suatu *query* yang diharapkan dapat memenuhi keinginan *user* dari kumpulan dokumen yang ada. Beberapa pengertian Information Retrieval dari berbagai sumber, antara lain :

Information Retrieval adalah "studi tentang sistem pengindeksan, pencarian, dan mengingat data, khususnya teks atau bentuk tidak terstruktur lainnya."[virtechseo.com]

"Information Retrieval adalah seni dan ilmu mencari informasi dalam dokumen, mencari dokumen itu sendiri, mencari metadata yang menjelaskan dokumen, atau mencari dalam database, apakah relasional database itu berdiri sendiri atau database hypertext jaringan seperti Internet atau intranet, untuk teks , suara, gambar, atau data " [Wikipedia]

Information Retrieval adalah "bidang di persimpangan ilmu informasi dan ilmu komputer. Berkutat dengan pengindeksan dan pengambilan informasi dari sumber informasi heterogen dan sebagian besar-tekstual. Istilah ini diciptakan oleh Mooers pada tahun 1951, yang menganjurkan bahwa diterapkan ke "aspek intelektual" deskripsi informasi dan sistem untuk pencarian (Mooers, 1951). " [Hersh, 2003]

Informasi atau data yang dicari dapat berupa berupa teks, image, audio, video dan lain-lain. Koleksi data teks yang dapat dijadikan sumber pencarian juga dapat berupa pesan teks, seperti e-mail, fax, dan dokumen berita, bahkan dokumen yang beredar di internet. Dengan jumlah dokumen koleksi yang besar sebagai sumber pencarian, maka dibutuhkan suatu sistem yang dapat membantu user menemukan dokumen yang relevan dalam waktu yang singkat dan tepat.

Di teknologi informasi terdapat istilah data retrieval, selain information retrieval. Dua hal ini sangatlah berbeda. Data retrieval secara umum menentukan dokumen yang tepat dari suatu koleksi data, yang isi dokumen tersebut mengandung keyword di dalam query user, tidak akan pernah cukup untuk memenuhi kebutuhan informasi user. Berbeda dengan data retrieval, user dari sistem Information Retrieval lebih memperhatikan dalam mendapatkan (*retrieve*) informasi melalui subyek, daripada retrieve data berdasarkan query yang diberikan, karena user tidak mau tahu bagaimana proses yang sedang berlangsung.

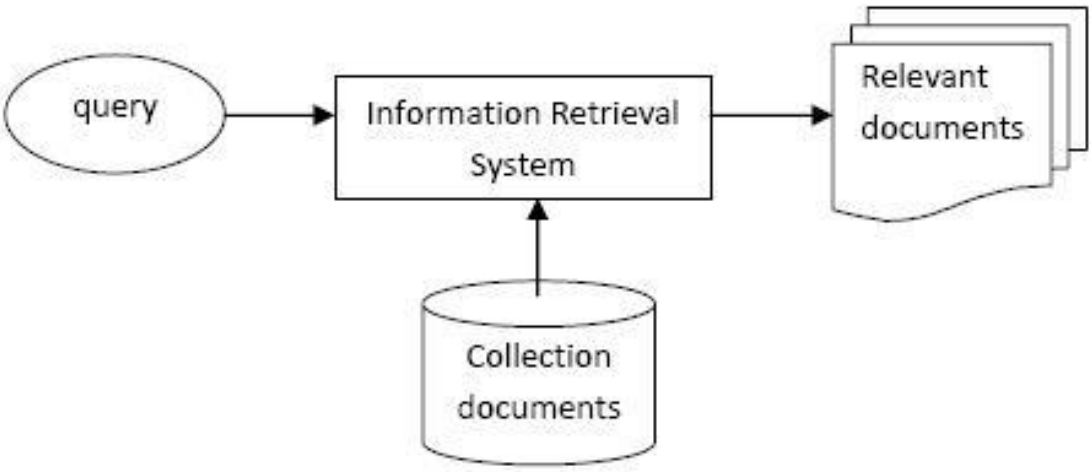
Perbedaan Information Retrieval dan Data Retrieval

Information Retrieval	**Data Retrieval**
Berhubungan dengan text bahasa umum yang tidak selalu terstruktur dan ada kemungkinan memiliki kerancuan arti	Berhubungan dengan data, yang mana semantik strukturnya sudah terdefinisikan
Informasi yang diambil mengenai subyek atau topic	Isi dokumen/data mengandung bagian dari keyword
Semantik sering kali hilang	Semantik terdefinisi dengan baik
Kesalahan kecil masih bisa ditorensi	Kesalahan kecil/tunggal dari suatu obyek menunjukkan kegagalan

Model yang terdapat dalam *Information Retrieval* terbagi dalam 3 model besar, yaitu:

1. *Set-theoretic models,* model merepresentasikan dokumen sebagai himpunan kata atau frase. Contoh model ini ialah *standard Boolean model* dan *extended Boolean model*.
2. *Algebratic model,* model merepresentasikan dokumen dan *query* sebagai vektor atau matriks *similarity* antara vektor dokumen dan vektor *query* yang direpresentasikan sebagai sebuah nilai skalar. Contoh model ini ialah *vector space model* dan *latent semantic indexing* (LSI).
3. *Probabilistic model,* model memperlakukan proses pengembalian dokumen sebagai sebuah *probabilistic inference*. Contoh model ini ialah penerapan teorema bayes dalam model probabilistik.

Proses dalam Information Retrieval dapat digambarkan sebagai sebuah proses untuk mendapatkan *relevant documents* dari *collection documents* yang ada melalui pencarian *query* yang diinputkan user.



Proses yang terjadi di dalam Information Retrieval System terdiri dari 2 bagian utama, yaitu *Indexing subsystem*, dan *Searching subsystem* (*matching system*). Proses *indexing* dilakukan untuk membentuk basisdata terhadap koleksi dokumen yang dimasukkan, atau dengan kata lain, *indexing* merupakan proses persiapan yang dilakukan terhadap dokumen sehingga dokumen siap untuk diproses. Proses *indexing* sendiri meliputi 2 proses, yaitu *document indexing* dan *term indexing*. Dari *term indexing* akan dihasilkan koleksi kata yang akan digunakan untuk meningkatkan performansi pencarian pada tahap selanjutnya. Tahap-tahap yang terjadi pada proses *indexing* ialah:

1. Word Token

   Yaitu mengubah dokumen menjadi kumpulan *term* dengan cara menghapus semua karakter dalam tanda baca yang terdapat pada dokumen dan mengubah kumpulan *term* menjadi *lowercase*.

2. Stopword Removal

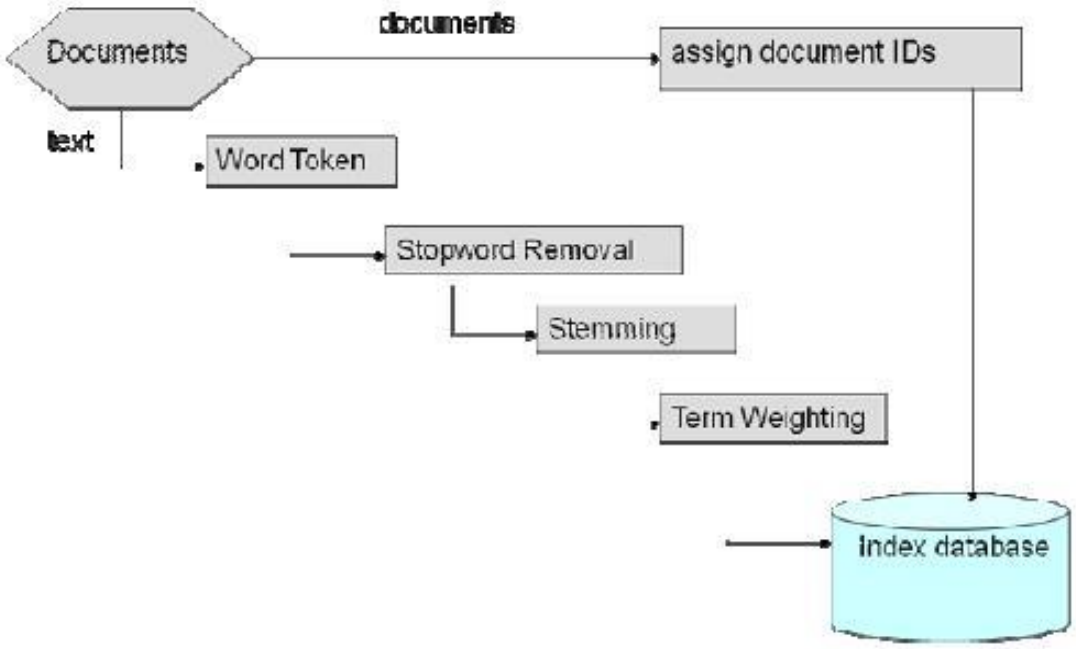
   Proses penghapusan kata-kata yang sering ditampilkan dalam dokumen seperti: *and, or, not* dan sebagainya.

3. Stemming

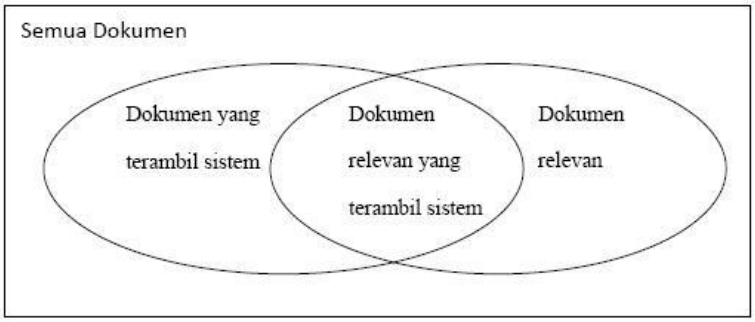
   Proses mengubah suatu kata bentukan menjadi kata dasar.

4. Term Weighting

   Proses pembobotan setiap *term* di dalam dokumen.



Search subsystem (matching) merupakan proses menemukan kembali informasi (dokumen) yang relevan terhadap *query* yang diberikan. Tidak semua dokumen yang diambil (*retrieved*) oleh system merupakan dokumen yang sesuai dengan keinginan user (*relevant*). Gambar dibawah ini menunjukkan hubungan antara dokumen relevan, dokumen yang terambil oleh system, dan dokumen relevan yang terambil oleh system:



II. **Pengukuran Performansi Information Retrieval System**

Nilai performansi dari aplikasi IR menunjukkan keberhasilan dari suatu IRS dalam mengembalikan informasi yang dibutuhkan oleh *user*. Untuk mengukur performansi

dari IRS, digunakan koleksi uji. Koleksi uji terdiri dari tiga bagian, yaitu koleksi dokumen, *query*, dan *relevance judgement*. Koleksi dokumen adalah kumpulan dokumen yang dijadikan bahan pencarian oleh sistem. *Relevance judgement* adalah daftar dokumen-dokumen yang relevan dengan semua *query* yang telah disediakan. Parameter yang digunakan dalam performansi sistem, antara lain :

1. *Precision* (ketepatan)

   Precision ialah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen yang terambil oleh sistem baik relevan maupun tidak relevan.

   precision = Jumlah dokumen yang relevan dengan query dan terambil.
   jumlah seluruh dokumen yang terambil

2. *Recall* (kelengkapan)

   Recall ialah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen relevan yang ada dalam koleksi dokumen (terambil ataupun tak terambil sistem).

   recall = Jumlah dokumen yang relevan dengan query dan terambil sistem.
   jumlah seluruh dokumen relevan dalam koleksi dokumen

3. *Interpolate Average Precision* (IAP)

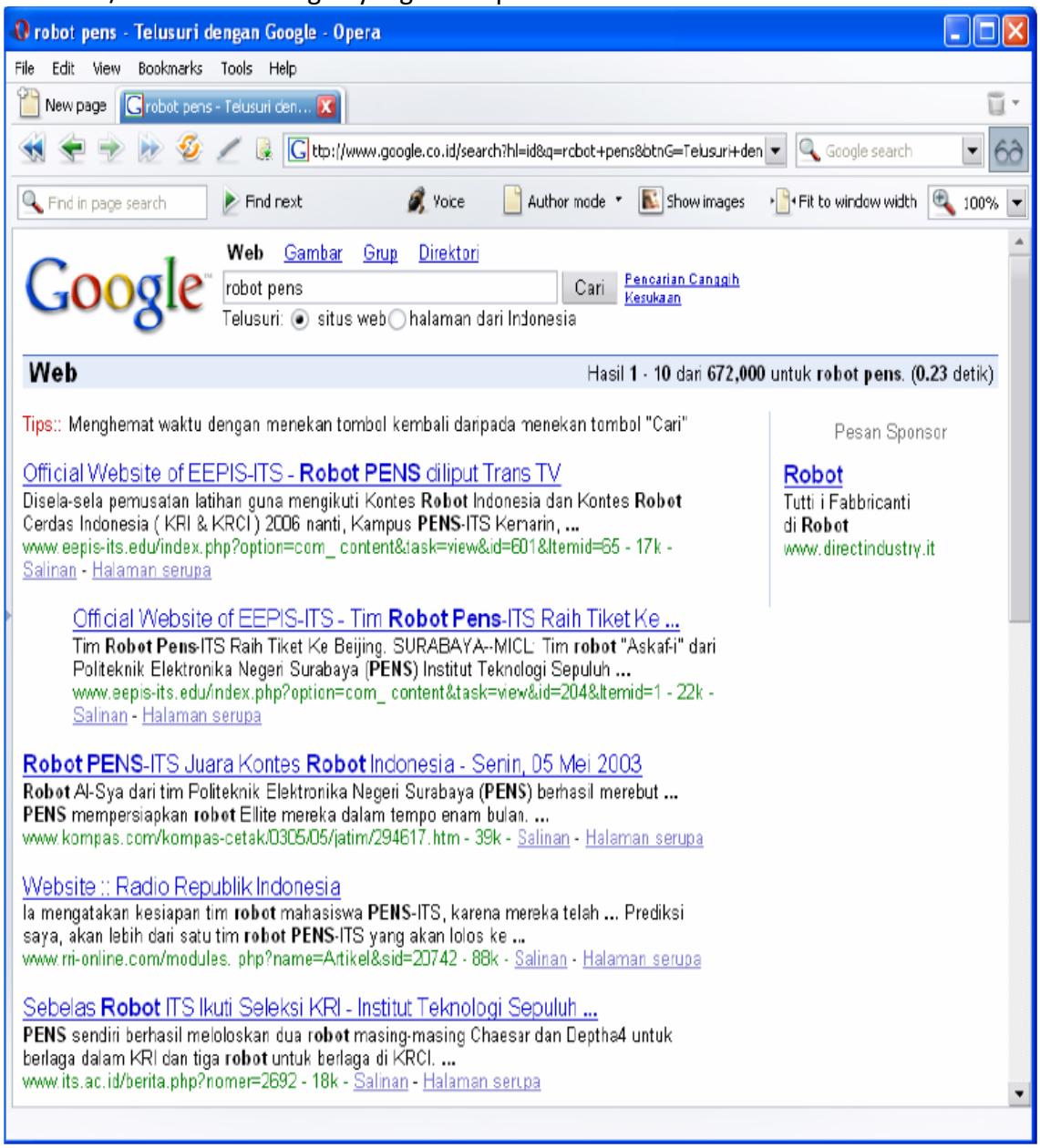
   Pengukuran performansi dengan mempertimbangkan aspek keterurutan atau rangking dapat dilakukan dengan melakukan interpolasi antara *precision* dan *recall*. IAP akan mencatat semua Semua dokumen yang relevan dan urutan dokumen tersebut pada hasil *IRS* dan menghitung nilai *precision*nya.

   Nilai *precision* untuk semua titik ditentukan oleh perubahan nilai *recall* yang terjadi. Nilai *precision* berubah pada saat nilai recall berubah naik. *Precision* disatu titik *recall* tertentu adalah maksimal *precision* untuk semua titik *recall* yang lebih kecil dari titik tersebut. Sebagai contoh, suatu IRS mendapatkan 10 dokumen berdasarkan suatu *query* dengan urutan sebagai berikut D1, D2, D3, D4, D5, D6, D7, D8, D9, dan D10. Dokumen yang relevan dalam koleksi dokumen berdasar *query* tersebut ialah D2, D4, D7, D13, dan D20, maka nilai *precision* dari sistem tersebut ialah 3/10 = 0.3, sedangkan nilai *recall* nya ialah 3/6 = 0.5.

III. PENERAPAN APLIKASI INFORMATION RETRIEVAL

A. Searching Text melalui Web Search Engine

Keyword dimasukkan oleh user untuk pencarian informasi yang diinginkan pada Search Engine, yang mana informasi yang didapatkan mengandung relevansi/keterkaitan dengan yang diharapkan



B. *Information retrieval di Perpustakaan*

Perpustakaan adalah salah satu institusi pertama yang mengadopsi sistem IR untuk mendapatkan informasi. Pada umumnya, sistem yang digunakan di

perpustakaan pada awalnya dikembangkan oleh institusi akademis dan kemudian oleh produsen komersil. Pada generasi pertama, sistem pada dasarnya terdiri dari suatu otomatisasi dari teknologi sebelumnya (seperti kartu katalog) dan memungkinkan pencarian berdasar judul dan nama pengarang. Pada generasi kedua , kemampuan pencarian ditambahkan dengan pencarian berdasarkan pokok utama, dengan kata kunci, dan tambahan lagi fasilitas kueri kompleks. Pada generasi ketiga, yang sekarang ini yang sedang menyebar, fokusnya adalah meningkatkan antarmuka grafis, format elektronik, fitur *hypertext*, dan sistem arsitektur terbuka.

C. CBIR(Content Based Image Retrieval) Technology

Retrieval berdasarkan kategori konten dan warna. Dimana user mendeskripsikan image apa yang akan dicari dengan cara memilih kategori misalnya jenis image, Negara, tahun pembuatan dsb.

http://www.simoskow.sch.id/kreativitas/B10.pdf (accessed date: 30 Mei 2011 at 00.10 am)

INI TAMBAHAN BAHAN DARI TAON KEMAREN, SIAPA TAU BISA TAMBAH WAWASAN, MAKASIH YE BUK

Information Retrieval

Pendahuluan

Sejak tahun 1940, permasalahan dalam penyimpanan dan penemuan kembali informasi (*information retrieval*) telah menarik perhatian. Dalam pernyataan yang sederhana, untuk mendapatkan informasi yang akurat dan dapat diakses secara cepat adalah sangat sulit. Efek dari hal ini adalah mengabaikan penemuan informasi yang relevan, yang akhirnya mengarah pada duplikasi pekerjaan dan usaha. Dengan adanya komputer, banyak pemikiran yang diberikan dalam penggunaan komputer untuk menyediakan sistem retrieval yang cerdas dan cepat. Di perpustakaan misalnya, yang memiliki banyak permasalahan mengenai penyimpanan dan penemuan kembali informasi seperti *cataloguing* dan administrasi umum, telah berhasil diambil alih oleh komputer. Walaupun demikian, permasalahan retrieval yang efektif belum sepenuhnya terpecahkan.

Pada prinsipnya, penyimpanan dan penemuan kembali adalah sederhana. Andaikan terdapat penyimpanan dokumen dan seseorang merumuskan pertanyaan yang jawabannya adalah satu set dokumen yang memuaskan kebutuhan informasi yang dinyatakan dalam pertanyaannya tersebut, dia dapat memperoleh dokumen tersebut dengan membaca seluruh dokumen yang terdapat pada penyimpanan tersebut, menahan dokumen yang relevan dan membuang yang lainnya. Dalam beberapa

hal, hal ini mendasari penemuan kembali yang sempurna. Tetapi, solusi ini benar-benar tidak dapat dilaksanakan. Seorang user tidak mempunyai waktu atau tidak ingin meluangkan waktunya membaca seluruh dokumen yang ada, terlepas dari kenyataan bahwa secara fisik dia tidak akan mungkin melakukannya.

Ketika komputer berkecepatan tinggi menjadi tersedia untuk pekerjaan kualitatif, banyak pemikiran yang menyatakan bahwa komputer akan mampu membaca keseluruhan dokumen untuk mengintisarikan dokumen yang relevan. Hal tersebut akan menjadi nyata yaitu penggunaan ***natural language text*** pada sebuah dokumen tidak hanya menyebabkan permasalahan input dan penyimpanan tetapi juga tidak terpecahkannya permasalahan intelektual dalam karakteristik isi dokumen. Dapat dibayangkan perkembangan perangkat keras di masa depan memungkinkan input dan penyimpanan dengan natural language. Tetapi karakterisasi pada *software* yang berusaha meniru *human* proses pada proses 'reading' adalah permasalahan yang tentu saja sangat lekat. Secara lebih rinci, 'reading' melibatkan usaha dalam mengintisarikan informasi sintatik dan semantik, dari teks dan penggunaannya untuk memutuskan apakah setiap dokumen relevan atau tidak dengan permintaan tertentu. Kesulitan tidak hanya dalam mengetahui bagaimana mengintisarikan informasi tetapi juga bagaimana menggunakannya untuk memutuskan keterkaitannya. Kemajuan yang lamban dari *linguistik* modern pada *front-semantic* dan kegagalan mesin penerjemah yang menarik perhatian menunjukkan bahwa permasalahan tersebut belum sepenuhnya terpecahkan.

Tujuan dari strategi *retrieval* otomatis adalah untuk memperoleh kembali semua dokumen yang relevan dan pada saat yang sama bisa terambil beberapa dokumen yang tidak relevan. Ketika karakterisasi dari sebuah dokumen terpecahkan, harusnya dokumen tersebut direpresentasikan secara relevan dengan sebuah *query*, hal tersebut memungkinkan dokumen diperoleh kembali sebagai respon dari query tersebut.

Dalam cara ini pegindeksan manual mempunyai karakteristik tradisional, ketika memberikan indeks ke dokumen, pengindeks akan berusaha mencarikan dokumen yang diminta oleh user sesuai dengan indeksnya. Secara implisit pengindeks membangun query untuk dokumen yang relevan. Ketika pengideksan dilakukan secara otomatis, hal tersebut diasumsikan bahwa dengan mendorong teks dari sebuah dokumen *(query)* pada analisis otomatis yang sama, hasilnya akan berupa penyajian dari isi dokumen, dan jika dokumen berkaitan dengan *query*, sebuah prosedur komputasional akan menunjukkan hal ini.

Dengan cukup beralasan, hal tersebut memungkinkan manusia untuk menetapkan keterkaitan sebuah dokumen dalam sebuah *query*. Agar komputer melakukan hal ini, kita harus membangun sebuah model dalam keputusan relevan yang dapat terukur. Hal tersebut sangat menarik untuk dicatat bahwa banyak riset mengenai *information retrieval* dapat ditunjukkan untuk dikaitkan dengan aspek yang berbeda dari model seperti itu.

Information Retrieval System

Information Retrieval System dimulai dengan hal-hal pada sisi input. Masalah yang utama di sini adalah untuk mendapatkan suatu penyajian dari setiap dokumen dan *query* yang sesuai untuk menggunakan suatu komputer. Kebanyakan *retrieval system* berbasis komputer hanya menyimpan suatu penyajian dokumen (*query*) yang berarti bahwa teks suatu dokumen hilang satu kali ketika telah diproses untuk kepentingan peningkatan penyajiannya. Suatu penyajian dokumen bisa dianggap penting, sebagai contoh daftar kata-kata yang disadap, dibandingkan mempunyai komputer yang memproses bahasa yang alami. Pendekatan alternatif adalah untuk mempunyai suatu bahasa tiruan dimana di setiap semua dokumen dan query dapat dirumuskan.

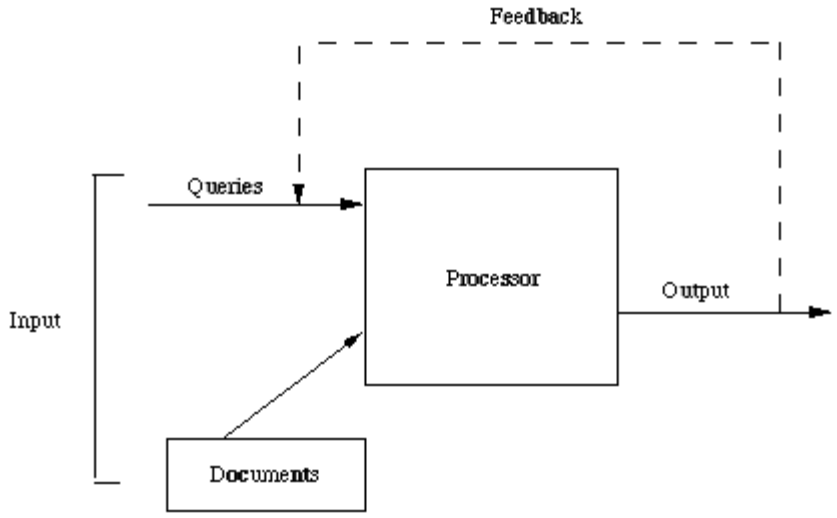


Figure 1.1. A typical IR system

Ada beberapa bukti untuk menunjukkan bahwa ini dapat efektif. Tentu saja ini mensyaratkan bahwa seorang pengguna akan diajar untuk menyatakan kebutuhan informasinya dalam bahasa.

Sistem *retrieval on-line* memungkinkan pengguna untuk mengubah permintaannya selama pencarian. Dengan demikian, diharapkan peningkatan *retrieval* berikutnya dapat berjalan. Prosedur seperti itu biasanya dikenal sebagai 'umpan balik'. Sebuah contoh suatu sistem *retrieval on-line* yang canggih adalah sistem MEDLINE (Mc Carn dan Leiter).

Kemudian bagian yang kedua setelah input adalah *processor* yang merupakan bagian dari sistem retrieval yang terkait dengan proses retrieval. Proses dapat melibatkan struktur informasi dalam beberapa cara yang sesuai, seperti penggolongannya. Ini juga akan melibatkan penyelenggaraan fungsi retrieval yang nyata, yang akan melaksanakan strategi pencarian sebagai jawaban atas suatu *query*. Dalam diagram, dokumen-dokumen telah ditempatkan dalam suatu kotak terpisah untuk menekankan fakta bahwa mereka bukan hanya input tetapi dapat digunakan sepanjang proses *retrieval*, sedemikian sehingga strukturnya lebih tepat dilihat sebagai bagian dari proses *retrieval*.

Bagian akhir dari *Information Retrieval* adalah output, yang pada umumnya merupakan satu set kutipan atau angka-angka dokumen. Dalam sebuah sistem operasional cerita berakhir di sini. Bagaimanapun juga, di suatu sistem percobaan ini meninggalkan evaluasi untuk dilaksanakan.

Perspektif terhadap *Information Retrieval*

Bagian ini tidak dimaksudkan untuk membuat sebuah usaha pada suatu jumlah yang menyeluruh dan lengkap dari pengembangan *Information Retrieval* yang historis. Setidak-tidaknya, ini tidak akan bisa meningkatkan jumlah yang diberikan oleh Cleverdon dan Salton. Walaupun *Information Retrieval* dapat dibagi lagi dalam beberapa cara, ini tampaknya bahwa ada tiga area riset utama yang diantaranya menyusun sebuah porsi materi yang dipertimbangkan. Area tersebut adalah **isi analisa, struktur informasi, dan evaluasi.** Dengan singkat, isi analisa mempunyai kaitan dengan gambaran isi dokumen dalam suatu bentuk yang cocok untuk proses komputer. Struktur informasi mempunyai kaitan dengan pemanfaatan hubungan antar dokumen untuk meningkatkan efisiensi dan efektivitas strategi retrieval. Evaluasi mempunyai kaitan dengan pengukuran dari efektivitas retrieval.

Luhn menggunakan jumlah frekuensi kata-kata dalam teks dokumen untuk menentukan kata-kata yang mana yang cukup penting untuk mewakili atau mengenali dokumen di komputer (lebih detil tentang hal ini di bab yang berikutnya). Inilah yang disebut 'kata kunci' yang akan digunakan untuk memperoleh untuk masing-masing dokumen. Sebagai tambahan, frekwensi kejadian dari kata-kata ini

dapat juga digunakan di badan teks untuk menandai suatu derajat tingkat arti. Ini menyediakan suatu rencana penimbang sederhana untuk 'kata kunci' pada setiap daftar dan membuat ketersediaan suatu wakil dokumen dalam wujud sebuah 'uraian kata kunci yang dihargai'.

Dalam posisi ini, mungkin saja menyenangkan untuk menguraikan penggunaan ‘kata kunci'. Hal ni telah menjadi praktek umum didalam literatur *Information Retrieval* untuk mengacu pada materi deskriptif yang disadap dari teks sebagai kata kunci atau terminologi. Materi seperti itu biasanya merupakan hasil beberapa proses seperti pengumpulan bersama-sama dari varian analisis yang berbeda pada kata yang sama. Dalam tulisan ini, kata kunci dan istilah akan digunakan secara berbeda.

Penggunaan informasi statistik tentang distribusi kata-kata di dokumen dimanfaatkan lebih lanjut oleh Maron, Kuhns dan Stiles yang memperoleh asosiasi statistik antar kata kunci. Asosiasi ini menyajikan suatu basis untuk pembangunan suatu kamus sebagai bantuan kedalam *retrieval*. Banyak akhir-akhir ini riset dibawa bersama-sama dengan penerbitan pada *Statistical Association Methods for Mechanized Documentation* (Stevens).

Sparck Jones telah melanjutkan pekerjaan ini menggunakan ukuran asosiasi antar kata kunci berdasar pada frekwensi *co-occurrence*, yaitu frekuensi dimana dua kata kunci terjadi bersama-sama dalam dokumen yang sama. Dia telah menunjukkan bahwa kata-kata terkait seperti itu dapat digunakan secara efektif untuk meningkatkan daya ingat, sehingga dapat meningkatkan proporsi dari dokumen yang relevan yang didapat kembali. Secara menarik, awal gagasan Luhn masih sedang dikembangkan dan banyak metode otomatis pada karakterisasi yang didasarkan pada awal pekerjaannya.

Istilah struktur informasi mencakup secara rinci suatu organisasi informasi logis, seperti perwakilan dokumen, untuk kepentingan informasi retrieval. Pengembangan di struktur informasi telah diperbaharui. Alasan yang utama untuk kelambatan pengembangan di area informasi retrieval ini adalah bahwa dalam jangka waktu panjang tak seorangpun menyadari komputer itu tidak akan memberi suatu waktu retrieval yang bisa diterima dengan suatu dokumen besar kecuali jika beberapa struktur logis dibebankan. Sesungguhnya, pemilik *data-base* besar masih segan untuk mencoba teknik organisasi baru yang berjanji *retrieval* lebih baik dan lebih cepat. Kelambatan untuk mengenali dan mengadopsi teknik baru pada umumnya karena keterbatasan bukti eksperimental yang mendukung. Eksperimen yang lebih awal dengan sistem dokumen retrieval biasanya mengadopsi suatu organisasi serial file, dimana, walaupun efisien ketika sejumlah besar *query* diproses secara serempak di dalam suatu mode *batch*, membuktikan tidak cukup jika masing-masing *query* memerlukan waktu respon yang singkat. Organisasi

yang populer diadopsi untuk mengganti file tersebut. Baru-baru ini eksperimen telah mencoba untuk mempertunjukkan keunggulan dari *file clustered* untuk *retrieval on-line.*

Organisasi dari file ini diproduksi oleh suatu metode klasifikasi otomatis. Good dan Fairthorne adalah di antara yang pertama menyatakan bahwa penggolongan otomatis mungkin terbukti bermanfaat di dalam dokumen *retrieval*. Beberapa tahun kemudian ada eksperimen serius dilaksanakan pada dokumen yang di-*cluster* (Doyle; Rocchio). Semua eksperimen sejauh ini telah ada dalam skala kecil, sejak *clustering* hanya masuk ke dalam dirinya sendiri ketika skala ditingkatkan.

Evaluasi sistem retrieval telah terbukti sangat sulit. Senko dalam suatu survei yang sempurna mengatakan: 'Tanpa sebuah evaluasi system, keraguan menjadi area yang paling menyusahkan di dalam *Information Retrieval*...'. Disamping pekerjaan yang dipelopori sempurna dan dilaksanakan oleh Cleverdon di dalam area ini, dan di samping banyak ukuran efektivitas yang telah diusulkan oleh Robertson, suatu teori evaluasi umum tidak pernah muncul.

Di masa lalu telah ada banyak debat tentang kebenaran evaluasi berdasar pada pertimbangan keterkaitan yang disajikan dengan perbuatan salah manusia. Cuadra dan Katter memperkirakan keterkaitan yang terukur pada suatu skala nomor urut, dalam rank-ordering menunjukkan bahwa posisi suatu dokumen pada skala seperti itu dipengaruhi oleh variabel eksternal yang umumnya tidak dikendalikan di laboratorium. Lesk dan Salton sesudah itu menunjukkan bahwa suatu *dichotomous* skala di suatu dokumen adalah yang relevan atau tidak relevan, ketika diperlakukan ke suatu kemungkinan kesalahan tertentu, tidak berlaku hasil mendapatkan evaluasi yang berkaitan dengan ketepatan (proporsi dokumen yang didapat kembali adalah relevan) dan *recall* (proporsi dari dokumen yang relevan yang didapat kembali). Sekarang ini efektivitas retrieval kebanyakan masih diukur dalam kaitan dengan ketepatan dan daya ingat atau oleh ukuran berdasarkan *thereon*. Masih ada ketidak cukupan perawatan statistik yang mengungkapkan bagaimana uji signifikansi dapat digunakan. Maka, setelah beberapa dekade riset di dalam area ini pada dasarnya hanya mempunyai ketepatan dan daya ingat, dan hipotesis yang menyatakan, 'Didalam sistem tunggal, asumsi bahwa suatu urutan *sub-searches* untuk pertanyaan tertentu adalah dibuat order; pesanan yang logis dari ketepatan menurun yang diharapkan, dan kebutuhan itu dinyatakan dalam pertanyaan, ada suatu hubungan kebalikan antara daya ingat dan ketepatan, jika hasil sejumlah pencarian berbeda dirata-ratakan.' (Cleverdon).

Efisiensi dan Efektivitas

Sebagian besar riset dan pengembangan di dalam informasi retrieval diarahkan pada peningkatan efisiensi dan efektivitas *retrieval*. Efisiensi biasanya diukur dalam kaitan dengan sumber daya komputer digunakan seperti *core*, *backing store,* dan *CPU time*. Untuk mengukur efisiensi suatu mesin yang berjalan sendiri merupakan suatu kesulitan. Setidak-tidaknya, haruslah diukur bersama dengan keefektifan memperoleh beberapa gagasan untuk manfaat dalam kaitan dengan biaya unit. Efektivitas itu biasanya diukur dalam kaitan dengan ketepatan dan daya ingat. Ketepatan itu sendiri menjadi perbandingan dari jumlah dokumen *retrieval* yang relevan dengan total jumlah dokumen *retrieval*, dan daya ingat menjadi perbandingan dari jumlah dokumen retrieval yang relevan kepada total jumlah dokumen yang relevan (kedua-duanya didapat kembali dan tidak didapat kembali). Alasan untuk penekanan dua ukuran ini adalah acuan yang sering dibuat ke efektivitas *retrieval*. Ini akan mencukupi sampai kita menjangkau bab itu untuk berpikir tentang efektivitas *retrieval* dalam kaitan dengan ketepatan dan daya ingat. Sebelum dapat menghargai evaluasi pengamatan yang perlu dipahami adalah apa yang menimbulkan pengamatan itu.

Indexing

Suatu bahasa indeks adalah bahasa yang digunakan untuk menguraikan dokumen dan permintaan. Unsur-Unsur dari bahasa indeks adalah terminologi indeks, yang mungkin diperoleh dari teks dokumen untuk diuraikan, atau mungkin dengan bebas. Bahasa indeks dapat diuraikan menjadi *pre-coordinate* atau *post-coordinate*, yang pertama menunjukkan bahwa terminologi dikoordinir ketika mengindeks dan ketika dalam pencarian. Secara lebih rinci, dalam indeks *pre-coordinate* suatu kombinasi logis tentang segala terminologi indeks mungkin digunakan sebagai suatu label untuk mengidentifikasi suatu kelas dokumen, sedangkan di dalam indeks *post-coordinate* kelas yang sama akan dikenali pada waktu pencarian dengan mengombinasikan kelas dokumen berlabel dengan terminologi indeks individu.

Bahasa indeks yang muncul dari algoritma *conflation* dapat dijelaskan sebagai indeks dengan kosakata yang tak terkendalikan, *post-coordinate* dan merupakan turunan. Kosa kata terminologi indeks pada tahap evolusi kumpulan dokumen hanya merupakan satuan dari semua *conflation* kelas nama.

Ada banyak kontroversi tentang macam bahasa index yang mana yang terbaik untuk pencarian kembali dokumen. Perdebatan utama adalah tentang apakah indeks otomatis sebaik atau lebih baik daripada indeks manual. Masing-masing bisa dilakukan pada berbagai tingkatan kompleksitas.

Bagaimanapun, sepertinya terbukti dalam keduanya, indexing otomatis dan manual, menambahkan kompleksitas dalam wujud kendali yang lebih terperinci. Pesan adalah kosa kata tak terkendalikan berdasar pada bahasa alami untuk mencapai efektivitas pencarian kembali yang dapat diperbandingkan dengan kosa kata dengan kendali rumit.

Mungkin bukti yang paling substansial untuk indexing otomatis telah keluar dari SMART Project (1966). Salton baru-baru ini meringkas kesimpulan nya:'... pada rata-rata prosedur indeks yang paling sederhana yang mengidentifikasi dokumen yang diinginkan atau kueri oleh satu set terminologi, tertimbang atau tak tertimbang, diperoleh dari dokumen atau teks kueri adalah juga yang paling efektif'. Rekomendasinya harus jelas, analisa teks otomatis perlu menggunakan terminologi tertimbang diperoleh dari kutipan dokumen yang panjangnya sedikitnya satu dokumen abstrak.

Dokumen representatif yang digunakan oleh SMART project lebih canggih dari pada sekedar daftar batang yang diintisarikan oleh *conflation*. Tidak ada keraguan, dibanding format kata biasa metode ini lebih efektif (Carroll dan Debruyn). Pada puncaknya, the SMART project ini menambahkan indeks tertimbang, di mana suatu istilah index mungkin adalah beberapa kelas konsep melalui penggunaan berbagai kamus.

1. Motivasi

Information Retrieval (IR) adalah kegiatan yang berhubungan dengan penyajian, penyimpanan, pengorganisasian, dan pengaksesan ke materi informasi. Pengorganisasian dan penyajian dari materi informasi perlu menyediakan akses yang gampang kepada pengguna informasi yang dipilihnya. Sayangnya, kebutuhan informasi pemakai bukanlah masalah yang sederhana. Sebagai contoh pemakai memerlukan hal ini dalam konteks Internet :

> Mencari semua halaman yang berisi informasi tenis regu pada perguruan tinggi yang memenuhi syarat: (1) berada di universitas di Amerika Serikat dan (2) mengikuti turnamen tenis NCAA. Supaya cocok dengan keinginan pengguna, maka halaman harus meliputi informasi tentang rangking nasional regu pada tiga tahun terakhir dan email atau nomor telepon dari pelatih regu.

Yang jelas, uraian yang menyangkut kebutuhan informasi pemakai ini tidak bisa digunakan secara langsung untuk mencari informasi dengan menggunakan antarmuka Web seperti mesin pencari. Sebagai

gantinya, pemakai harus menterjemahkan informasi ini ke dalam *query* yang dapat diproses oleh mesin pencari (atau sistem IR).

Dalam format umumnya, terjemahan ini menghasilkan satu kata kunci (atau bisa juga *query*) yang meringkas uraian dari informasi yang diperlukan pemakai. Dengan *query* yang dibutuhkan pemakai, IR sistem akan mendapat informasi yang mungkin dicari pemakai atau mungkin informasi yang berhubungan dengannya. Penekanannya adalah pada pencarian kembali informasi sebagai lawan perolehan kembali data.

***Information Retrieval* dan *Data Retrieval*.**

Pencarian data dalam konteks sistem IR, sebagian besar berisi tentang menentukan dokumen mana yang sesuai kueri pemakai dalam koleksi dokumen, lebih sering terjadi ketidakcocokan kebutuhan informasi yang diinginkan pemakai. Sesungguhnya, pemakai dari suatu sistem IR lebih memperhatikan *information retrieval* tentang suatu bahasan daripada pencarian data yang memenuhi *query* pengguna. Bahasa pencarian data lebih mengarah pada pencarian semua obyek yang memenuhi kondisi seperti yang ada di dalam ungkapan reguler atau pada suatu ungkapan relational aljabar. Dengan begitu untuk pencarian data, satu kesalahan obyek pada ribuan object pencarian itu berarti kegagalan. Sedangkan untuk sistem *information retrieval*, object yang didapat bisa saja tidak akurat, boleh terjadi kesalahan kecil yang mungkin tidak diperhatikan. Alasan utama yang menyebabkan perbedaan ini adalah *information retrieval* itu pada umumnya berhubungan dengan bahasa sehari-hari yang tidak selalu tersusun baik dan bisa menimbulkan kerancuan secara semantik. Pada sisi lain, sistem pencarian data (seperti suatu database relational) berhadapan dengan data yang mempunyai arti semantik dan struktur yang tersusun dengan baik.

Pencarian data, selama menyediakan suatu solusi kepada pemakai sistem database, tidak memecahkan permasalahan dalam informasi yang mencari informasi tentang suatu topik. Untuk bisa memenuhi keefektifan dalam memenuhi kebutuhan informasi pemakai, bagaimanapun juga sistem IR harus menginterpretasikan materi informasi pada koleksi dokumen dan memberikan rangking kerelevanan terhadap *query* pemakai. Penafsiran tentang isi dokumen ini meliputi penggalian informasi semantik dan *syntactic* dari teks dokumen dan menyesuaikannya dengan informasi yang dibutuhkan pemakai. Kesulitannya tidak hanya keharusan tahu tentang bagaimana menggali informasi tetapi harus mengetahui juga tentang bagaimana menggunakannya untuk menentukan kesesuaiannya. Gagasan dari

kesesuaian ini adalah inti dari *information retrieval*. Nyatanya tujuan utama dari sistem IR adalah untuk mendapatkan semua dokumen yang berhubungan dengan query pengguna dan meminimalisir dokumen yang tidak berhubungan dengan *query* pengguna sebisa mungkin.

Berikut ini adalah perbedaan antara data retrieval dan information retrieval dapat dilihat pada table.

	Data Retrieval (DR)	Information Retrieval (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Table Information Retrieval (IR) dan Data Retrieval (DR)

Dalam data retrieval, kita mengecek untuk melihat apakah suatu item ada atau tidak dalam suatu file. Dalam information retrieval, kita akan menemukan item yang cocok secara parsial dengan permintaan dan kemudian memilih satu yang paling cocok di antara beberapa item yang terpilih. Penarikan kesimpulan yang digunakan dalam *data retrieval* adalah secara deduksi, misal aRb dan bRc maka hasilnya adalah aRc. Dalam *information retrieval* biasanya menggunakan penarikan kesimpulan secara induksi, hubungan hanya ditetapkan dengan tingkat kepastian atau ketidakpastian, dan karenanya, kepercayaan kita terhadap kesimpulan yang diambil adalah variabel. Perbedaan ini mengarahkan kita untuk menggambarkan *data retrieval* sebagai *deterministic* sedangkan *informasi*

retrieval sebagai *probabilistic.* Seringkali teorema Bayes dilibatkan untuk menarik kesimpulan dalam *information retrieval*.

Perbedaan yang lain yaitu dalam penggolongannya. Dalam *data retrieval* kita lebih cenderung tertarik pada penggolongan *monothetic,* yaitu penggolongan dengan penggambaran kelas oleh atribut-atribut objek yang diperlukan pada sebuah kelas. Dalam *information retrieval* penggolongan seperti itu tidak bermanfaat, sebenarnya penggolongan *polythetic* lebih sering diinginkan. Dalam penggolongan tersebut setiap individu dalam kelas hanya akan mememiliki sebuah proposi dari seluruh atribut yang dimiliki oleh semua anggota dalam kelas tersebut. Oleh karena itu, tidak ada atribut yang cukup maupun perlu untuk keanggotaan setiap kelas.

Bahasa kueri dalam Data Retrieval

Bahasa kuery untuk *data retrieval* biasanya menjadi tiruan kosa kata dan sintaksis terbatas, dalam *information retrieval* kita lebih memilih untuk menggunakan bahasa alami walaupun ada beberapa perkecualian. Dalam *data retrieval*, *query* biasanya merupakan suatu spesifikasi lengkap dari apa yang diinginkan. Dalam *information retrieval* ini, kuerinya tanpa alternatif dan tidak sempurna. Perbedaan akhir ini memunculkan sebagian dari fakta bahwa dalam *information retrieval*, kita sedang mencari-cari dokumen yang relevan. Tingkat kecocokan dalam *information retrieval* diasumsikan untuk menandai adanya kemungkinan dari keterkaitan menyangkut item tersebut. Satu konsekuensi sederhana dari perbedaan ini adalah *data retrieval* lebih sensitif terhadap kesalahan. Suatu kesalahan dalam pencocokan tidak akan mendapatkan kembali item yang diinginkan yang menandakan suatu kegagalan total sistem. Di dalam *information retrieval*, kesalahan kecil dalam pencocokan biasanya tidak mempengaruhi hasil dari sistem.

Banyak pengembalian informasi sistem otomatis bersifat percobaan. *Information retrieval* sebagian besar bersifat percobaan dan dilanjutkan di laboratorium sedangkan sistem operasional adalah sistem komersil yang meminta pembayaran bagi layanan yang mereka sediakan. Secara alami dua sistem dievaluasi dengan cara yang berbeda. ‘Dunia nyata' sistem *information retrieval* dievaluasi dalam kaitannya dengan 'kepuasan pemakai' dan harga yang ingin dibayarkan oleh pemakai layanan. Sistem IR yang bersifat percobaan dievaluasi dengan membandingkan percobaan perolehan kembali dengan standard khusus untuk tujuan itu.

Information retrieval sebagai pusat dari peradaban teknologi

Selama 20 tahun ini wilayah *information retrieval* telah tumbuh dengan baik melebihi tujuan utamanya yaitu mengurutkan dokumen dan mencari dokumen yang berguna dalam kumpulan koleksi. Sekarang ini, riset dalam *information retrieval* meliputi modeling, penggolongan dokumen, sistem arsitektur, antarmuka, visualisasi data, penyaringan, bahasa, dan lain lain. Di samping kedewasaannya sampai saat ini, *information retrieval* dilihat sebagai lingkup minat sempit yang sebagian besar digeluti oleh tenaga ahli informasi dan pustakawan. Visi yang telah ditanamkan selama bertahun-tahun selain penghamburan yang cepat, antarmuka, *information retrieval* juga ingin dijadikan alat untuk multimedia dan *hypertext* aplikasi. Salah satu yang telah terwujud adalah pengenalan *World Wide Web* pada awal tahun 1990-an.

Web telah menjadi tempat penyimpanan yang universal bagi pengetahuan manusia dan kebudayaan yang membuat pembagian informasi dan ide menjadi mempunyai lingkup yang tidak pernah terjadi sebelumnya. Kesuksesan ini terjadi karena antarmuka standar yang selalu sama tidak tergantung lingkungan kerja yang dipakai untuk menjalankan antarmuka. Hasilnya pengguna terlindungi dari hal-hal detail dari protokol komunikasi, lokasi mesin, dan Sistem Operasi. Lebih jauh lagi, pengguna bisa membuat dokumen berbasis web sendiri dan membuatnya terhubung dengan web dokumen lain tanpa larangan. Ini adalah aspek utama karena ini membuat web menjadi media penyebaran informasi yang bisa diakses oleh setiap orang. Dunia tanpa batas ini telah menarik minat jutaan manuia pada awal perkembangannya. Lebih jauh lagi, ini menyebabkan revolusi pada penggunaan komputer oleh pengguna dan persiapan tugas sehari-harinya. Sebagai contoh home banking dan home shopping telah menjadi populer dan semakin banyak menghasilkan pendapatan.

Meskipun banyak keberhasilan yang telah dicapai, web juga telah menimbulkan masalah sendiri. Menemukan informasi yang berguna pada web bisa saja sangat sulit. Sebagai contoh, pengguna untuk mendapatkan informasi yang diinginkan pengguna dia harus memilih links yang akan menyambung ke halaman lain. Selama links tersebut banyak dan hampir semua tidak dikenal, maka dia akan kesulitan untuk mencari nformasi yang dimaksudkannya tadi. Bagi pengguna yang kurang punya minat pada hal ini bisa menyebabkan dia frustasi. Halangan utamanya adalah banyaknya data model yang ada pada web, dan kebanyakan berkualitas rendah. Kesulitan ini membuat orang yang berminat pada *information retrieval* berusaha memperbaharui teknik pencarian. Hasilnya, *Information Retrieval* telah mendapat tempat yang sama dengan teknologi lain pada pusat peradaban.

2. Konsep dasar

Keefektifan perolehan kembali informasi relevan secara langsung dipengaruhi baik melalui perintah pemakai dan pandangan logis dari dokumen yang diadopsi oleh sistem perolehan kembali.

2.1. Pemakai

Pemakai suatu sistem perolehan kembali harus menterjemahkan informasi yang diperlukannya ke dalam query yang terdapat dalam bahasa yang disajikan oleh sistem. Dengan suatu Information Retrieval system, secara normal menyiratkan bahwa satu set kata-kata yang menyampaikan semantik dari kebutuhan informasi. Dengan suatu *Data Retrieval System*, suatu ungkapan *query* (sebagai contoh suatu regular *expression*) digunakan untuk menyampaikan batasan jawaban yang harus dipenuhi. Di kedua kasus di atas, pemakai mencari informasi yang bermanfaat dengan menggunakan *Retrieval Task*.

Sekarang pertimbangkan, jika seorang pemakai mempunyai minat, sebagai contoh pemakai bisa tertarik terhadap dokumen tentang balap mobil secara umum. Dalam situasi ini, pemakai menggunakan sebuah antarmuka interaktif untuk melihat-lihat koleksi dokumen yang berhubungan dengan balap mobil. Ia mungkin tertarik tentang dokumen tentang Formula 1, tentang pabrik mobil, atau tentang `24 Hours Le Mans'. Ketika sedang pembacaan tentang `24 Hours Le Mans`, ia mungkin berkeinginan mengalihkan perhatiannya ke sebuah dokumen yang menyediakan arah ke Le Mans, kemudian berpindah lagi ke dokumen yang meliput pariwisata di Perancis. Dalam situasi ini pemakai tidak sedang *searching*, melainkan sedang mem-*browse* koleksi dokumen. *Browsing* masih merupakan suatu proses dalam mendapat kembali informasi, tetapi sasaran utamanya tidak tergambar jelas di permulaannya dan mungkin saja berubah selama proses interaksi dengan sistem.

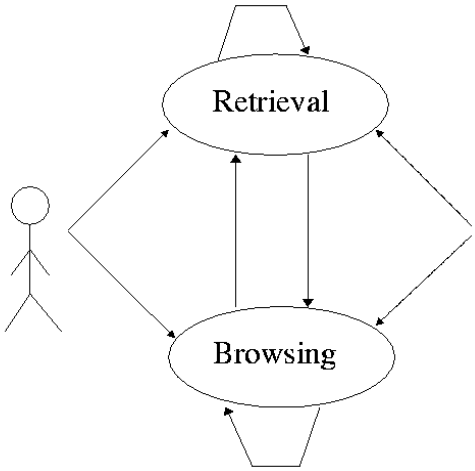


Figure: *Interaction of the user with the retrieval system through distinct tasks.*

Permintaan dari pemakai bisa terdiri dari dua jenis yang berbeda: *information* atau *data retrieval* dan *browsing*. *Classic information retrieval systems* secara normal memungkinkan *information* atau *data retrieval.* Sistem *hypertext* pada umumnya diset untuk menyediakan cara *browsing* yang cepat. Antarmuka web dan perpustakaan digital modern mungkin mencoba mengkombinasikan tugas ini untuk menyediakan kemampuan *retrieval* yang telah meningkat. Bagaimanapun, kombinasi dari *retrieval* dan *browsing* belum merupakan suatu pendekatan yang baik dan bukanlah paradigma yang dominan.

Gambar ini menunjukkan interaksi dari pemakai melalui tugas berbeda yang diidentifikasi. *Data* dan *information retrieval* pada umumnya disajikan oleh sistem pengembalian informasi yang paling modern seperti Web. Lebih lanjut, sistem seperti ini mungkin juga menyediakan beberapa format dalam *browsing* yang masih terbatas. Kombinasi *information* dan *data retrieval* dengan *browsing* saat ini belum umum, namun bisa menjadi populer di masa datang.

Retrieval dan *browsing* terdapat dalam bahasa *World Wide Web*, yaitu ketika pemakai meminta informasi secara interaktif. Alternatif lainnya adalah melakukan perolehan kembali dengan cara permanen dan otomatis yang menggunakan perangkat lunak yang mengambil informasi untuk pemakai. Sebagai contoh, informasi yang berguna untuk seorang pemakai bisa diambil pada waktu tertentu dari suatu jasa layanan berita. Dalam hal ini, dikatakan bahwa sistem IR sedang melaksanakan perolehan

kembali tugas tertentu, yang terdiri dari penyaringan informasi yang relevan untuk kemudian diperiksa oleh pemakai.

2.2. Pandangan Logis tentang Dokumen

Dalam kaitannya dengan pertimbangan historis, dokumen-dokumen dalam suatu koleksi sering diwakili oleh suatu satuan kata kunci atau terminologi index. Kata kunci seperti itu bisa diambil secara langsung dari dokumen teks atau bisa ditetapkan oleh manusia (seperti sering dilaksanakan dalam wilayah ilmu pengetahuan. Bukan masalah apakah kata kunci yang representatif ini diperoleh secara otomatis atau yang dihasilkan oleh suatu spesialis, mereka menyediakan suatu pandangan logis tentang dokumen.

Komputer modern memungkinkan untuk menghadirkan dokumen dengan satuan kata-katanya sendiri. Dalam hal ini sistem perolehan kembali mengadopsi pandangan logis terhadap teks penuh (atau penyajian) tentang suatu dokumen. Tetapi dengan koleksi yang sangat besar, komputer modern mungkin harus mengurangi satuan kata kunci. Ini dapat terpenuhi melalui penghapusan stopwords (seperti artikel dan penghubung), penggunaan pembendungan (yang akan mengurangi kata-kata yang beda bersifat ketatabahasaan), dan identifikasi kata benda kelompok (yang menghapus kata sifat, kata keterangan, dan kata kerja). Lebih lanjut, penekanan terhadap informasi yang dimaksud bisa digunakan. Operasi ini disebut *text operation*. *Text Operation* mengurangi kompleksitas dari penyajian dokumen dan memberikan pandangan yang logis dari itu suatu teks untuk satu set terminologi index.

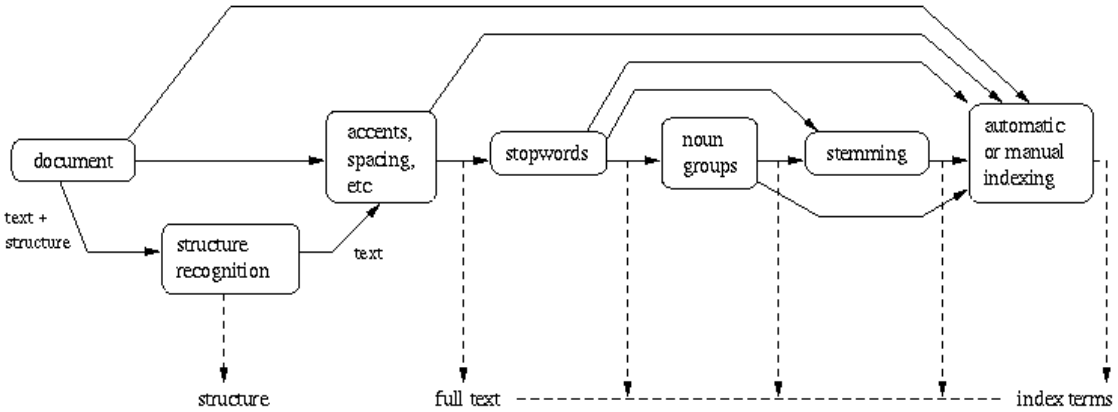


Figure: *Logical view of a document: from full text to a set of index terms.*

Teks yang penuh adalah teks yang jelas dan mempunyai pandangan logis yang lengkap tentang suatu dokumen tetapi pemakaiannya pada umumnya memakan biaya yang tinggi. Sebagian kecil satuan kategori (yang dihasilkan oleh suatu manusia) menyediakan pandangan logis yang ringkas terhadap suatu dokumen tetapi pemakaiannya mungkin mendorong kearah mutu yang lebih rendah. Beberapa pandangan logis bisa diadopsi oleh suatu *Information Retrieval System* seperti di gambar berikut ini. Di samping mengadopsi perwakilan yang manapun, sistem perolehan kembali mungkin juga mengenali struktur internal yang ada dalam suatu dokumen (contohnya bab, bagian, subseksi, dll.). Informasi tersebut bisa bermanfaat dan diperlukan oleh sistem perolehan kembali.

Seperti di gambar di atas, berita secara logis yang mewakili suatu dokumen sebagai rangkaian dimana pandangan logis suatu dokumen mungkin bergeser (secara perlahan) dari suatu penyajian teks penuh kepada suatu penyajian yang lebih tinggi yang ditetapkan oleh manusia.

3. Masa Lampau, Masa Kini, dan Masa depan

1. Awal Perkembangan

Selama hampir 4000 tahun, manusia telah mengorganisasikan informasi untuk pencarian dan penggunaan kembali. Sebuah contoh khusus adalah daftar isi sebuah buku. Semenjak jumlah informasi berkembang melebihi sebuah buku, menjadi sebuah kebutuhan untuk membangun struktur data khusus untuk menjamin akses yang lebih cepat pada penyimpanan informasi. Sebuah struktur data yang lama dan populer untuk *information retrieval* yang lebih cepat adalah kumpulan kata-kata atau konsep terpilih yang dihubungkan dengan informasi (atau dokumen) yang bersangkutan yaitu indeks. Indeks adalah inti dari setiap sistem *information retrieval* modern. Indeks menyediakan akses data yang lebih cepat dan membuat proses kueri berjalan lebih cepat.

Selama berabad-abad, indeks dibuat secara manual seperti hirarki penggolongan. Dalam kenyataannya, kebanyakan perpustakaan masih menggunakan bentuk ini untuk mengklasifikasi dokumen-dokumen mereka. Hirarki seperti ini pada umumnya sudah dipahami oleh ahli dari bidang ilmu kepustakaan. Lebih jauh lagi, kedatangan komputer telah memungkinkan konstruksi indeks yang

besar secara otomatis. Indeks otomatis membuat permasalahan *information retrieval* lebih kepada sistem itu sendiri dari pada kebutuhan pemakai. Dalam hal ini, adalah penting untuk membedakan anatara dua pandangan permasalahan tentang *information retrieval* yang berbeda, yaitu berbasis komputer atau manual.

2. Information retrieval di Perpustakaan

Perpustakaan adalah salah satu institusi pertama yang mengadopsi sistem IR untuk mendapatkan informasi. Pada umumnya, sistem yang digunakan di perpustakaan pada awalnya dikembangkan oleh institusi akademis dan kemudian oleh produsen komersil. Pada generasi pertama, sistem pada dasarnya terdiri dari suatu otomatisasi dari teknologi sebelumnya (seperti kartu katalog) dan memungkinkan pencarian berdasar judul dan nama pengarang. Pada generasi kedua , kemampuan pencarian ditambahkan dengan pencarian berdasarkan pokok utama, dengan kata kunci, dan tambahan lagi fasilitas kueri kompleks. Pada generasi ketiga, yang sekarang ini yang sedang menyebar, fokusnya adalah meningkatkan antarmuka grafis, format elektronik, fitur *hypertext*, dan sistem arsitektur terbuka.

3. Web dan Perpustakaan Digital

Jika kita pertimbangkan mesin pencarian pada Web sekarang ini, kita bisa menyimpulkan bahwa mesin tersebut melanjutkan penggunaan indeks yang sangat serupa seperti yang digunakan itu oleh pustakawan seabad yang lalu. Kemudian, apa yang telah berubah?

Tiga perubahan dramatis dan pokok telah terjadi dalam kaitannya dengan kemajuan teknologi komputer modern dan popularitas Web. Pertama, lebih murah untuk mempunyai akses kepada berbagai sumber informasi. Hal ini memungkinkan mencapai pengguna lebih luas dari yang mungkin pernah ada sebelumnya. Kedua, keahlian dalam semua macam komunikasi digital menyediakan akses lebih besar ke jaringan. Ini mengindikasikan bahwa sumber informasi tersedia sekalipun terletak sangat jauh dan bahwa akses bisa dilakukan dengan cepat (seringkali, dalam beberapa detik). Ketiga, kebebasan memasang informasi apapun juga yang dinilai bermanfaat telah sangat mendukung ketenaran dari Web. Untuk pertama kali di dalam sejarah, banyak orang mempunyai akses bebas ke suatu medium penerbitan besar.

Pada pokoknya, biaya rendah, akses lebih besar, dan kebebasan penerbitan sudah mengijinkan orang untuk menggunakan Web (dan perpustakaan digital modern) sebagai medium yang sangat interaktif. Demikianlah inter-aktivitas mengijinkan orang menukar pesan, foto, dokumen, perangkat lunak, video, dan 'mengobrol' di dengan nyaman dan biaya rendah. Lebih lanjut, orang-orang dapat melakukannya pada waktu yang mereka sukai (misalnya, kamu dapat membeli suatu buku pada malam hari) yang lebih lanjut meningkatkan kenyamanan layanan. Jadi, inter-aktivitas tinggi menjadi pergeseran pokok dan mutakhir dalam paradigma komunikasi.

Di masa datang, tiga pertanyaan utama perlu dibicarakan. Pertama, meskipun inter-aktivitas tinggi, orang-orang masih mengalami kesulitan (jika tidak mustahil) untuk mendapat kembali informasi yang relevan pada kebutuhan informasi mereka. Jadi, di dalam dunia Web yang dinamis dan perpustakaan digital yang besar, teknik mana yang akan mengijinkan pencarian kembali dengan mutu lebih tinggi? Kedua, dengan permintaan akses yang terus meningkat, tanggapan yang cepat menjadi suatu faktor yang semakin mendesak. Jadi, teknik mana yang akan menghasilkan indeks yang lebih cepat dan kueri yang lebih cepat? Ketiga, mutu dari hasil pencarian kembali sangat dipengaruhi oleh interaksi pemakai dengan sistem. Jadi, bagaimana nantinya suatu pemahaman yang lebih baik menyangkut perilaku pemakai mempengaruhi penyebaran dan disain starategi baru *information retrieval*?

Isu-isu Praktis

Perdagangan elektronik adalah suatu kecenderungan utama pada Web sekarang ini dan telah menguntungkan berjuta-juta orang-orang. Di dalam suatu transaksi elektronik, pembeli pada umumnya harus mengajukan pada penjual beberapa format informasi kredit yang mana yang dapat digunakan untuk membayar produk atau layanan. Dalam format yang paling umum, informasi seperti itu terdiri dari suatu nomor kartu kredit. Bagaimanapun, karena pengiriman nomor kartu kredit melalui Internet bukanlah suatu prosedur aman, data seperti itu pada umumnya dikirim melalui fax. Ini menyiratkan bahwa, sedikitnya dalam permulaan, transaksi antara seorang pemakai baru dan suatu penjual memerlukan pelaksanaan prosedur *off-line* dari beberapa langkah-langkah sebelum transaksi yang sebenarnya dapat berlangsung. Situasi ini dapat diatasi jika data dienkripsi untuk keamanan. Sesungguhnya, beberapa perusahaan dan institusi telah menyediakan beberapa format enkripsi otomatis untuk pertimbangan keamanan.

Bagaimanapun, keamanan bukanlah satu-satunya perhatian. Isu utama yang lain adalah privasi. Seringkali, orang akan menukar informasi sepanjang informasi itu tidak untuk umum. Pertimbangan adalah banyak tetapi salah satunya yang paling umum adalah untuk menghindari penyalahgunaan tentang informasi pribadi oleh pihak ketiga. Jadi, privasi adalah isu yang lain yang mempengaruhi penyebaran Web dan belum dibicarakan dengan baik.

Dua isu lain yang sangat penting adalah hak cipta dan hak paten. Adalah jauh dari kejelasan bagaimana penyebar luasan tentang data melalui Web mempengaruhi hak cipta dan hukum paten di berbagai negara. Ini informasi yang dikirimkannya ? Dan jika bisa dilakukan demikian, apakah hal itu dapat dipertanggungjawabkan jika terjadi suatu penyalahgunaan menyangkut informasi itu (sekalipun bukan sumber informasi).

Referensi

http://www.dcs.gla.ac.uk/keith/chapter1/ch1

http://www.sims.berkeley.edu/~hearts/irbook/1/node2

http://en.wikipedia.org/wiki/informationretrieval

http://kluweronline.com/issn/1386-4564

http://information.net/ir/8-1/paper

http://www.searchtools.com/info/info-retrieval.html

http://www.aaai.org/AITopics/html/info.html

http://www.db.dk/pi/iri/

REFERENSI

http://informationretrievals.wordpress.com/ (accessed date: 9 Mei 2011 at 11:55 am)

http://en.wikipedia.org/wiki/Information_retrieval (accessed date:9 Mei 2011 at 11:59)

http://bungaimuthz.wordpress.com/2011/05/04/apa-itu-information-retrieval/(accessed

date: 9 Mei 2011 at 12:04)

http://puthree91.student.umm.ac.id/download-as-pdf/umm_blog_article_61.pdf(accessed date : 30 Mei 2011 at 00.13)

http://lecturer.eepis-its.edu/~kholid/kuliah/Dasar%20Sistem%20Informasi/day6/Day%206%20-%20Information%20Retrieval.pdf (date: 29 Mei 2011 at 23.35 pm)