Implementasi Graf pada Metode Crawling dan Indexing di dalam Mesin Pencari Web

Fauzan Muhammad Rifqy 13513081[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]fauzanrifqy@students.itb.ac.id

***Abstrak*—Mesin Pencari adalah sebuah program yang mampu melakukan pencarian data atau informasi yang terdapat dalam laman web yang ada di sebuah jaringan. Search Engine memudahkan kita dalam mengakses informasi yang kita butuhkan. Cukup dengan memberikan input kata kunci tentang informasi yang ingin kita dapatkan, Mesin Pencari mampu memberikan daftar laman web yang mungkin memiliki informasi yang sedang kita cari. Biasanya laman web ditampilkan dalam bentuk daftar terurut berdasarkan peringkat laman web tersebut dalam jaringan. Dari daftar tersebut kita dapat langsung diarahkan ke laman web yang tertera. Dari berbagai macam mesing pencari, beberapa mesin pencari yang paling populer ialah Google Search, Yahoo! Search, dan Bing. Namun bagaimana sebenarnya kerja Web Search Engine dibalik itu semua? Salah satu metode yang dipakai ialah Crawling dan Indexing yang dapat diimplementasikan dengan graf. Dengan metode tersebut mesin pencari akan melakukan pencarian informasi pada tiap laman web dengan algoritma tertentu dan mengembalikan hasil pencariannya disesuaikan dengan input atau masukan yang kita berikan disesuaikan dengan informasi yang dibutuhkan.**

***Kata Kunci*—Mesin Pencari, Crawling, Indexing, Google, Yahoo, Bing.**

I. PENDAHULUAN

Di era teknologi informasi seperti saat ini, kebutuhan untuk mendapatkan akses informasi dengan cepat dan tepat menjadi sangat penting. Seiring terus bertumbuhnya teknologi informasi, data yang disimpan setiap harinya berkembang begitu cepat. Maka kita perlu mempunyai sebuah metode untuk mengolahnya, menyusun setiap informasi sehingga ketika kita membutuhkannya, kita dapat segera menemukannya dengan mudah.

Satu hal yang diperlukan untuk memungkinkan kita mendapatkan sebuah informasi dengan cepat dan tepat ialah, Search Engine. Web Search Engine memungkinkan kita mengolah informasi sehingga menghasilkan informasi-informasi yang kita butuhkan.

Maka hal penting yang harus dimiliki oleh sebuah Search Engine ialah seberapa cepat ia dapat menyajikan hasil informasi yang kita cari, dan apakah hasil yang ia berikan sesuai.

Lalu bagaimana Web Search Engine dapat melakukan hal tersebut? Salah satu metode yang digunakan dalam Web Search Engine ialah metode Crawling dan Indexing.

Crawling merupakan proses pencarian data informasi pada laman web dan menghasilkan laman web yang memiliki informasi yang kita cari.

Indexing merupakan penyortiran laman web yang telah kita temukan dari metode crawling sebelumnya sehingga ketika kita memerlukan akses ke suatu informasi akan mampu memberikan jawaban langsung secara cepat.

Metode crawling menerapkan graf dalam implementasinya, yaitu dengan menghubungkan tiap laman web dan linknya dengan sebuah sisi. Laman web yang memiliki link tertuju pada dirinya paling banyak akan dijadikan perhatian utama untuk kemudian di tampilkan di hasil pencarian.

II. DASAR TEORI GRAF

2.1. Definisi Graf

Graf terdiri atas set yang dilambangkan dengan V serta sekumpulan E. Setiap elemen V dinamakan simpul (vertex) dan setiap elemen E dinamakan sisi (edge). V dan E merupakan sebuah pasangan himpunan dimana pada umumnya graf ditulis dalam bentuk notasi

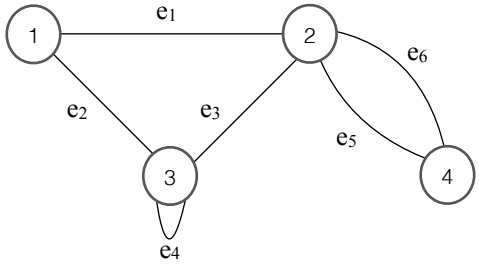
$G = (V,E)$

Himpunan V tidak boleh kosong, artinya paling sedikit sebuah graf hanya memiliki satu buah simpul, sedangkan himpunan E boleh kosong.

$V = \{v_1,v_2,v_3,\ldots,v_n\}$, yaitu himpunan tidak kosong dari simpul-simpul,

$E = \{e_1,e_2,e_3,\ldots,e_n\}$, yaitu himpunan sisi yang menghubungkan sepasang simpul.

Berikut adalah contoh graf :



Gambar 2.1 graf dengan 4 simpul dan 6 sisi

Pada gambar diatas sisi e_5 dan e_6 dinamakan sisi-ganda (*multiple edges* atau *paralel edges*) karena kedua sisi tersebut menghubungkan dua simpula yang sama, yaitu pada kasus diatas menghubungkan simpul 2 dan simpul 4. Sedangkan e_4 pada gambar diatas dinamakan gelang atau kalang (*loop*) karena ia berawal dan berakhir di simpul yang sama yaitu di simpul 3.

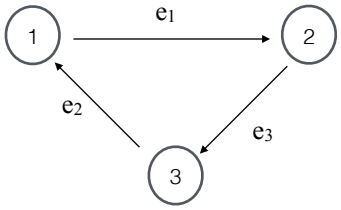
2.2. Jenis-Jenis Graf

Berdasarkan ada tidaknya gelang atau sisi ganda pada suatu graf, maka graf digolongkan menjadi dua jenis :

1. Graf sederhana (*simple graph*).
   Graf yang tidak mengandung gelang maupun sisi-ganda dinamakan graf sederhana.
2. Graf tak sederhana (unsimple graph).
   Graf yang mengandung gelang maupun sisi-ganda dinamakan graf tak sederhana.

Berdasarkan orientasi arah pada sisi, maka secara umum graf dibedakan menjadi 2 jenis yaitu :

1. Graf tak-berarah (*undirected graph*).
   Graf yang sisinya tidak mempunyai orientasi arah disebut graf tak-berarah. Contoh graf tak berarah ditunjukan pada gambar 2.1
2. Graf berarah (*directed graph*).
   Graf yang setiap sisinya diberikan orientasi arah disebut graf berarah.



Gambar 2.2 graf berarah

2.3 Terminologi Graf

1. Ketetanggaan
   Dua buah simpul dikatakan bertetangga bila keduanya terhubung langsung oleh sebuah sisi.
2. Bersisian
   Untuk sembarang sisi e = (v_1,v_2) dikatakan e bersisian dengan simpul v_1 dan v_2.
3. Simpul Terpencil
   Simpul terpencil ialah simpul yang tidak mempunyai sisi yang bersisian dengannya.
4. Graf Kosong
   Graf Kosong ialah graf yang sisinya merupakan himpunan kosong.
5. Derajat
   Derajat suatu simpul adalah jumlah sisi yang bersisian dengan simpul tersebut.
6. Siklus atau Sirkuit
   Lintasan yang berakhir dan berawal pada simpul yang sama disebut sirkuit atau siklus.
7. Terhubung
   Dua buah simpul v_1 dan simpul v_2 disebut terhubung jika terdapat lintasan atau sisi dari v_1 ke v_2.
8. Upagraf
   Misalkan G = (V,E) adalah sebuah graf. G_1 = (V_1,E_1) adalah upagraf dari G jika V_1 merupakan himpunan bagian dari V dan E_1 merupakan himpunan bagian dari E.
9. Graf Berbobot
   Graf berbobot adalah graf yang setiap sisinya memiliki nilai(bobot).
10. Lintasan Euler
    Ialah lintasan yang melalui masing-masing sisi dalam graf tepat satu kali.
11. Sirkuit Euler
    Ialah sirkuit yang melewati masing-masing sisi tepat satu kali.
12. Lintasan Hamilton
    Ialah lintasan yang melalui tiap simpul di dalam graf tepat satu kali.
13. Sirkuit Hamilton
    Ialah sirkuit yang melalui tiap simpul di dalam graf tepat satu kali, kecuali simpul asal (sekaligus simpul akhir) yang dilalui dua kali.

III. Mesin Pencari Web

3.1 Definisi Search Engine

Mesin Pencari merupakan sebuah program komputer yang dirancang untuk mencari laman web yang terdapat dalam jaringan dengan tujuan mengarahkan kita pada informasi yang kita butuhkan di laman web.

3.2 Sejarah Search Engine

Pada tahun 1960, Ted Nelson menciptakan projek dinamakan projek Xanadu dan dalam tiga tahun berhasil membuat sebuah hiperteks yang berisi hiperlink. Saat ini kita kenal sebagai kata dengan link. Diawal perkembangan internet, Tim Berners-Lee membuat sebuah situs web yang berisi daftar link situs web yang ada di internet. Seiringnya perkembangan situs web yang terus bertambah, situs-situs ini sudah tidak dapat lagi membuat daftar situs.

Utilitas pencari pertama kali yang digunakan dalam mencari situs di internet adalah Archie yang dibuat tahun 1990 oleh Alan Emtage, Bill Heelan, dan J. Peter Deutsh, mereka merupakan mahasiswa ilmu komputer Universitas McGill, Amerika Serikat. Archie dirancang untuk melakukan pencarian pada semua file yang ada dalam sebuah direktori, sehingga pada dasarnya archie masih melakukan pencarian file di internet, bukan berupa pencarian kata kunci yang kita kenal saat ini.

Pada tahun 1991, Mark Mccahil dari Universitas Minnesota menciptakan program search yang mampu melakukan indeksasi pada teks dokumen, program tersebut dinamakan Gopher. Akibatnya seluruh situs gopher menjadi situs web setelah terciptanya *World Wide Web.*

Pada tahun 1994 muncul mesin pencarian yang berbeda

dari mesin pencarian sebelumnya, yaitu Web Crawler. Mesin Pencarian Web Crawler mamungkinkan penggunanya mencari informasi dengan menggunakan kata kunci tertentu yang muncul di laman web mana saja. Bermula dari sinilah standarisasi mesin pencari yang ada saat ini.

Sejak saat itu mulai bermunculan mesin pencari diantaranya adalah WebCrawler, HotBot, Excite, Infoseek, Inkotomi, dan AltaVista.

SEO (Search Engine Optimization) merupakan proses yang dilakukan secara sistematis bertujuan meningkatkan volume dan kualitas trafik kunjungan melalui mesin pencari menuju situs web tertentu dengan memanfaatkan mekanisme kerja atau algoritma mesin pencari tersebut.

Tujuan SEO ialah menempatkan sebuh laman web pada urutan teratas atau setidaknya pada halaman pertama hasil pencarian. Dan tentunya laman web yang muncul pada urutan teratas akan memiliki kemungkinan dikunjungi lebih banyak dibandingkan laman web lainnya yang tidak ada di halaman pertama hasil pencarian.

Pada mulanya mesin pencari menggunakan algoritma pencarian yang didasarkan sepenuhnya pada informasi yang diberikan oleh pemilik laman web, sehingga pemilik lama web dapat memanipulasi serangkaian kata kunci yang tidak sesuai pada laman webnya agar mesin pencari dapat menempatkan situsnya pada urutan yang tidak sesuai. Hal ini menyebabkan hasil pencarian menjadi tidak akurat dan tidak berkualitas.

Larry Page dan Sergey Brin, dua mahasiswa dari jurusan Ilmu Komputer Universitas Stanford berusaha menyelesaikan permasalahan tersebut dengan membangun Backrub, sebuah mesin pencari yang mengurutkan situs atau laman web dengan perhitungan matematika.

Algoritma tersebut dinamakan Page Rank, prinsipnya ialah laman web yang paling banyak di link oleh laman web lain. Nilai Page Rank juga akan semakin tinggi jika kualitas link yang mengarah kepadanya dinilai baik.

Hingga pada tahun 1998 Page dan Brin mendirikan Google yang merupakan versi tingkat lanjut dari Backrub yang dalam waktu singkat mampu memperoleh kepercayaan publik karena dapat menyajikan kualitas pencarian yang akurat dan cepat.

3.3 Cara Kerja Mesin Pencari

Dengan sebuah link, search engine menelusuri setiap laman web yang saling terhubung, jutaan laman web terkoneksi dan disimpan untuk selanjutnya ditelusuri kata kunci yang akan dicari.

Laman web disimpan dalam hard drives untuk selanjutnya di proses ketika kita melakukan pencarian. Untuk menyelesaikan pencarian tersebut, mesin memegang miliaran laman web yang dapat diakses dalam sepersekian detik.

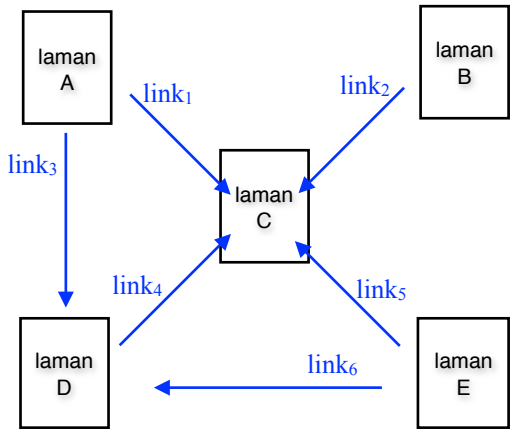
Pada Sebuah mesin pencari, relevansi tidak berarti hanya pencocokan kata pada sebuah laman web saja, banyak faktor yang perlu dipertimbangkan dalam meningkatkan akurasi dari kata kunci yang dicari.

Saat ini mesin pencari menafsirkan seberapa penting sebuah laman web dengan mengukur popularitasnya. Dan hingga saat ini cara tersebut terbukti cukup berhasil dalam prakteknya.

IV. Implementasi Graf di dalam Mesin Pencari

Sebuah mesin pencari diawal pencariannya akan melakukan sebuah metode yang disebut dengan crawling, metode ini menggunakan prinsip implementasi dari graf, yaitu dengan mengasumsikan laman web sebagai elemen himpunan simpul dilambangkan dengan V (vertex) dan link sebagai elemen himpunan sisi berarah dilambangkan dengan E (edge).

Berikut Ilustrasi laman web pada metode crawling :



Gambar 4.1 Ilustrasi laman web dalam graf

Selanjutnya setelah kita menelusuri setiap link yang ada pada laman web, kita akan mendapatkan jutaan laman page yang saling terhubung oleh sebuah link.

Data ini kemudian di simpan dalam ratusan komputer sebelum nantinya dilakukan pencarian pada masing-masing laman web.

Pertanyaan selanjutnya adalah bagaimana sebuah search engine mengatur dan mengurutkan laman web yang penting untuk di tampilkan di awal hasil pencarian?

Pertama-tama mesin pencari akan melakukan pencocokan kata kunci dengan laman web. Mesin pencari akan mengumpulkan laman web yang mengandung kata yang kita cari. Di tahap ini kita telah meminimalisir jutaan laman web yang tidak memiliki kata kunci yang kita cari.

Namun untuk mendapatkan akurasi yang tinggi tentang informasi yang diinginkan oleh pengguna, mesin pencari perlu memilih kembali laman web yang ditelusurinya dari tahap satu, karena ketika kita mengetikkan “harimau”, kita tidak menginginkan sebuah web yang hanya menuliskan kata harimau, namun yang kita inginkan adalah deskripsi tentang hewan tersebut ataupun gambar dan video tentang “harimau”.

Selanjutnya bagaimana sebuah mesin pencari menentukan laman web yang penting?

Setiap mesin pencari mempunyai faktor-faktor tertentu dalam melakukan pencarian, ada hingga ratusan faktor

sebelum sebuah laman web muncul di hasil pencarian sebagai urutan teratas.

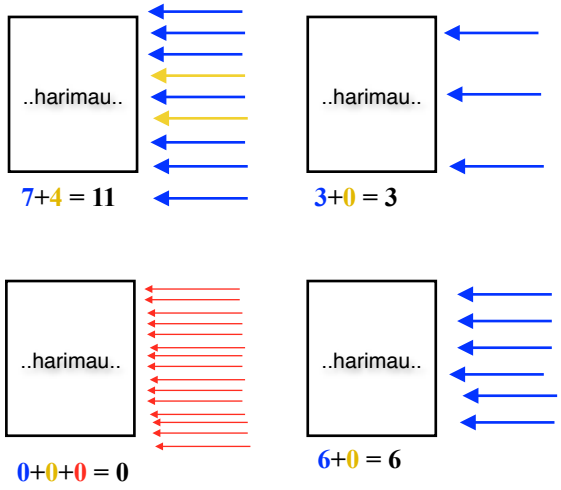
Salah satu metodenya adalah dengan pemanfaatan graf yang telah di ilustrasikan diatas. Laman web yang penting adalah laman web yang paling banyak ditunjuk oleh link dari luar, atau dalam teori graf yaitu simpul ang memiliki nilai derajat sisi (link) berarah masuk yang paling tinggi. Tidak hanya itu, setiap sisi (link) juga akan dinilai seberapa penting sisi tersebut.

Berikut ilustrasi pencariannya :

Misalkan pada pencarian kita mengetikkan kata kunci "harimau" :

Gambar 4.2 Ilustrasi pencarian mesin pencari

Mesin pencari akan mendaptkan laman web yang mengandung kata kunci dari data yang telah dikumpulkan sebelumnya ketika melakukan crawling.



Gambar 4.3 Ilustrasi pencarian mesin pencari sebagai graf berbobot dengan laman web sebagai simpul dan link sebagai sisi berarah dengan nilai tertentu

Setiap laman web akan dinilai berdasarkan derajat sisi (link) yang menunjuk dirinya sendiri dan seberapa besar nilai link tersebut. Jika link yang menunjuk sebuah laman web berasal dari laman web terpercaya maka nilainya akan tinggi, sementara jika berasal dari laman web yang tidak jelas asalnya terlebih jika link tersebut berasal dari spam maka nilai link tersebut akan rendah.

Maka dari ilustrasi diatas mesin pencari akan mengembalikan hasil pencarian laman web pertama bernilai sebelas diurutan pertama.

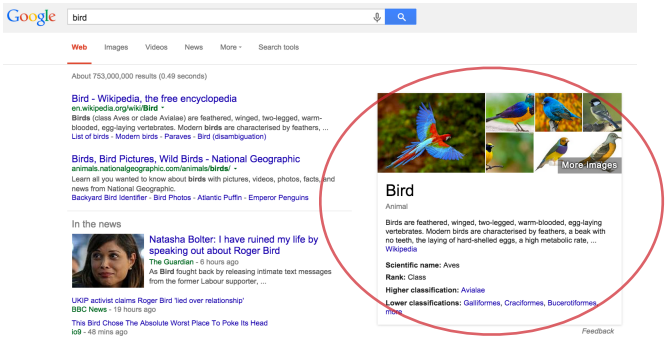
harimau

result *harimau* **:**

Harimau (11)
.....
.....
Harimau (6)
.....
.....

Selain itu setiap mesin pencari memiliki algoritmanya sendiri dalam menentukan hasil pencariannya. Google sebagai salah satu pemimpin kemajuan mesin pencari memiliki sebuah metode yang dinamakan Knowledge Graph dalam memproses hasil pencarian.

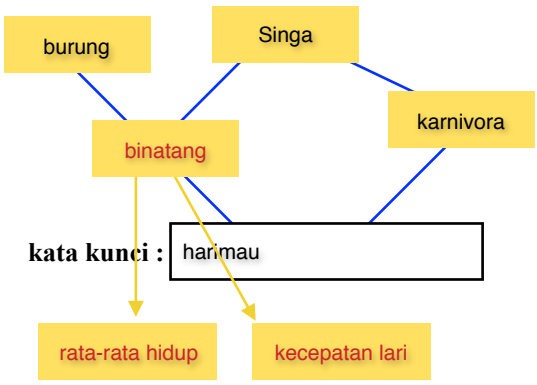
Jika kita mencoba memperhatikan dalam pencarian di mesin pencari google, mesin pencari google dapat memunculkan informasi tentang bojek yang kita cari.



Gambar 4.4 Contoh hasil pencarian pada mesin pencari google yang menggunakan knowledge graph pada lingkaran merah

Knowledge Graph memungkinkan program mengenali kata kunci yang kita cari sebagai sebuah objek yang memiliki keterkaitan dengan objek lain.

Berikut ilustrasi Knowledge Graph :



result *harimau* **:**

Harimau (11)
.....
.....
Harimau (6)
.....
.....

Harimau
adalah seekor binatang
memiliki :
kecepatan : …
rata-rata hidup : …

Gambar 4.5 Ilustrasi Knowledge Graph

V. KESIMPULAN

Untuk meningkatkan akurasi dalam pencarian diperlukan berbagai faktor untuk menentukan sebuah laman web berada di urutan yang sesuai. Metode graf dapat membantu menentukan urutan yang sesuai dengan mencari laman web yang memiliki derajat tertinggi yaitu yang paling banyak memiliki sisi (link) yang menunjuk dirinya juga mempertimbangkan besar atau nilai link yang menunjuk pada dirinya. Laman web dengan nilai tertinggi lah yang akan ditampilkan di urutan teratas hasil perncarian.

VII. ACKNOWLEDGMENT

Puji syukur semoga selalu tercurah kepada Allah SWT Rab Semesta Alam. Rasa terima kasih saya sampaikan khususnya kepada Ibu Dra Harlili, M.Sc. dan Bapak Dr. Ir. Rinaldi Munir, M.T. selaku dosen pengajar mata kuliah Matematika Diskrit. Semoga ilmu yang bapak dan ibu berikan dapat bermanfaat khususnya untuk saya pribadi umunya bagi masyarakat luas. Tak lupa rasa terima kasih juga saya sampaikan kepada teman-teman seperjuangan, mahasiswa Informatika ITB 2013, atas dukungan dan pembimbing diri ini kearah yang lebih baik. Semoga kita dapat mencapai impian kita dan terus berkarya untuk bersama-sama memperbaiki negri dan alam menjadi lebih baik.

REFERENSI

1. Munir, Rinaldi. *Diktat Kuliah IF 2120 Matematika Diskrit*. Bandung: Penerbit Teknik Informatika Institut Teknologi Bandung.
2. http://www.google.com/insidesearch/howsearchworks/crawling-indexing.html, (diakses pada tanggal 8 Desember 2014 pukul 19.00 WIB).
3. http://scanftree.com/Graph-Theory, (diakses pada tanggal 8 Desember 2014 pukul 19.10).
4. http://moz.com/beginners-guide-to-seo/how-search-engines-operate, (diakses tanggal 8 Desember 2014 pukul 19.20)
5. http://mauliaspensagres.wordpress.com/tag/sejarah-search-engine/, (diakses tanggal 8 Desember 2014 pukul 19.20)

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, 8 Desember 2014

ttd

Fauzan Muhammad Rifqy 13513081