Penerapan Teori Graf dan Web Crawler dalam Pemodelan *World Wide Web*

Farhan Makarim (13515003)[1]
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
[1]*farhan_makarim@student.itb.ac.id*

***Abstrak*—*Website* atau sering juga disebut Web adalah kumpulan-kumpulan halaman yang menampilkan berbagai macam informasi, baik dalam bentuk teks, data, gambar diam ataupun bergerak, data animasi, suara, video maupun gabungan dari semuanya untuk kemudian membentuk rangkaian bangunan yang saling berkaitan dimana masing-masing kontennya dihubungkan dengan jaringan halaman atau *hyperlink*. Proses perancangan *website* dapat dimodelkan dengan menggunakan teori graf yang diantaranya adalah graf berarah, graf tidak berarah dan graf *bipartite*. Pengkajian tentang graf ini bermanfaat untuk menentukan algoritma *searching, crawling* serta *ranking* yang mana ketiga algoritma tersebut digunakan pada mesin pencari (*search engine*). Makalah ini akan membahas beberapa teori tentang graf dan algoritma *searching* yang digunakan untuk merancang website yang *powerful* yang lebih cepat terdeteksi oleh mesin pencari.**

***Kata Kunci — Website*, Graf, *World Wide Web*, Crawler**

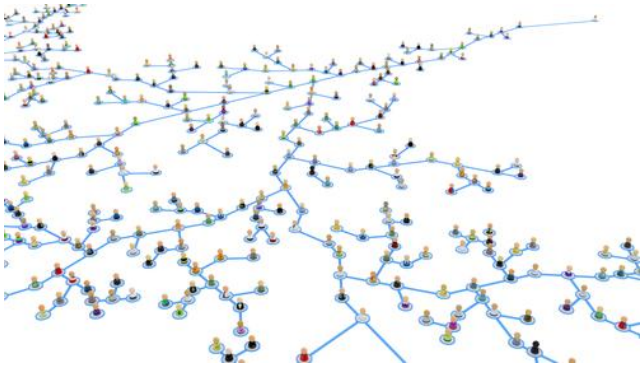
I. PENDAHULUAN

World Wide Web (WWW) merupakan sebuah sistem informasi global yang terdiri dari halaman web yang memiliki alamat dan memiliki keterkaitan dengan halaman web yang lain menggunakan *hyperlink*. Halaman web ini dapat diakses dengan menggunakan jaringan internet. *World Wide Web* diciptakan oleh Tim Berners-Lee dan beberapa peneliti lain di laboratorium energi tinggi CERN Geneva pada tahun 1989. Pada awalnya, WWW berfungsi sebagai sarana berbagi informasi antar peneliti di laboratorium tersebut. Sekarang, WWW telah berkembang menjadi repositori informasi global dan sebuah media komunikasi baru. *World Wide Web* menyediakan ruang bagi kita untuk berbagi informasi secara global.

World Wide Web memiliki karakteristik yang unik, diantaranya adalah sifatnya yang terdesentralisasi, tidak memiliki pusat dan perkembangannya yang cepat. Berbeda dengan jaringan lainnya (seperti telepon), WWW tidak memiliki struktur yang terencana oleh sistem. Jaringan *hyperlink* yang terbentuk antar halaman web merupakan kumpulan aksi yang tidak terkoordinasi dari individu-individu pengguna *website*. Sistem yang terdesentralisasi ini merupakan keunggulan yang dimiliki oleh *website*. Karena sistem ini, pengguna dapat menambahkan informasi baru ke dalam *website* secara mudah. Hal inilah yang menyebabkan proses tukar informasi yang terjadi antar web menjadi sangat cepat dan masif.

Meski demikian, Keunggulan tersebut juga menimbulkan permasalahan baru. Akibat strukturnya yang besar dan terus bertambah dengan sangat cepat, proses pencarian (*searching*) informasi menjadi semakin sulit. Oleh karena itu, diperlukan sebuah strategi untuk melakukan pencarian di dalam *World Wide Web*. Strategi pencarian ini dapat dikembangkan setelah kita mengetahui bagaimana strukutr relasi yang menghubungkan antar halaman web.

Struktur relasi pada *World Wide Web* dapat dimodelkan sebagai sebuah graf berarah. Graf *World Wide Web* dapat dinyatakan dengan formulasi G = (V,E), dimana V adalah himpunan tidak kosong dari halaman web dan E adalah himpunan pasangan yang menunjukan bahwa ada *hyperlink* dari halaman web satu ke halaman web yang lainnya.



Gambar 1. Representasi hubungan antara satu web dengan web yang lain yang membentuk World Wide Web.
Sumber Gambar :
https://media.licdn.com/mpr/mpr/shrinknp_800_800/AAEAAQAAAAAAAAPvAAAAJGYzNDJkNTU4LTBlZWEtNDkyNi04YjYyLTdhMDc0N2EzNDJlMg.jpg
Diakses : 05 Desember 2016 Pukul 03:40 WIB

Struktur graf pada *World Wide Web* pertama kali dimodelkan menggunakan teori graf. Meskipun demikian, beberapa penelitian mencoba untuk mengungkap struktur

graf *World Wide Web* yang sesungguhnya dengan menjelajahi *World Wide Web* menggunakan algoritma *searching*. Hasil penelitian ini menunjukann bahwa secara makroskopis struktur graf pada web memiliki sifat yang unik.

Sifat pencarian mesin pencari (*search engine*) yang berbasis pada tiga algoritma utama (yaitu *searching*, *crawling* dan *ranking*) membuat proses permodelan graf yang hanya menggunakan teori tentang *directed graph* (graf berarah) menjadi kurang *powerful*. Dalam makalah ini, akan digunakan juga konsep tentang graf *bipartite* yang membuat relasi antar website pada *World Wide Web* menjadi lebih jelas dan lebih terstruktur sehingga proses pencarian oleh mesin pencari akan berjalan lebih cepat dan lebih efektif.

II. TERMINOLOGI MENGENAI GRAF

Berikut ini adalah beberapa terminologi graf serta algoritma pencarian yang digunakan pada makalah ini. Sebuah graf berarah didefinisikan sebagai pasangan yang terdiri dari himpunan simpul V yang tidak kosong dan himpunan sisi E. Graf yang tidak memiki sisi disebut sebagai graf kosong (*null graph*) sedangkan graf yang memiliki sisi ke semua simpul disebut sebagai graf lengkap.

Sisi E didefinisikan sebagai pasangan simpul (V1,V2), dimana V1 disebut juga simpul awal dan V2 disebut dengan simpul akhir. Derajat masuk V didefinisikan sebagai jumlah sisi (jumlah sisi yang masuk ke simpul V), sementara itu derajat keluar V didefinisikan sebagai jumlah sisi (jumlah sisi yang keluar dari simpul V).

Graf bipartite (*bipartite graph*) adalah graf yang himpunan simpulnya dapat dipisah menjadi dua himpunan bagian V_1 dan V_2 sedemikian sehingga setiap sisi pada graf menghubungkan simpul di V_1 ke sebuah simpul di V_2.

III. PERMODELAN GRAF *WORLD WIDE WEB*

A. Model Graf Small World Network (Watt-Strogats)

Secara formal, *Small World Network* adalah sebuah graf yang mayoritas simpulnya tidak saling bertetangga satu sama lain, namun setiap pasangan simpulnya memiliki panjang lintasan antar dua buah simpul sembarangan (L) berbanding lurus dengan algoritma dari jumlah simpul pada graf(n).

Berbagai macam jaringan yang terdapat di alam dapat dimodelkan menggunakan graf ini, contohnya seperti jaringan kabel transmisi listrik, jaringan syaraf pada cacing dan graf pada jejaring sosial.

Pada model graf ini, sekumpulan simpul-simpul cenderung untuk membentuk kelompok-kelompok yang saling terhubung satu sama lain. Model graf ini cocok untuk memodelkan *world wide web*, karena sekumpulan halaman web yang saling berkaitan cenderung membentuk sebuah kelompok halaman web yang disebut dengan *website*.

Selain itu, salah satu alasan pemilihan model graf *small world network* sebagai model graf *world wide web* adalah bukti eksperimen yang dilakukan oleh Albert Jeong dan Barabasi pada tahun 1999. Berdasarkan penelitian yang mereka lakukan, diameter graf *world wide web* dapat ditentukan dengan persamaan **d = 0.35 + 2.06 log(n)** dengan d adalah nilai maksimal panjang lintasan antar dua buah simpul seberang, dan n adalah jumlah simpul pada graf. Berdasarkan persamaan tersebut, terlihat bahwa nilai d berbanding lurus dengan log(n). Hal ini mengindikasikan sifat "*Small World Network*" yang dimiliki oleh graf *World Wide Web.*

Berdasarkan persamaan di atas, dengan memasukan nilai estimasi jumlah halaman web yang ada di *World Wide Web* (N = 8 x 10^8), maka d adalah 18,59. Jadi, berdasarkan teori ini, kita bisa menjelajahi *World Wide Web* cukup dengan maksimal 19 klik pada *hyperlink.*

Terdapat dua karakteristik yang terdapat di dalam model graf *Small World Network*, yaitu *characteristic path length* (L) dan *clustering coefficient* (C).

Characteristic path length adalah ukuran seberapa jauh dua buah simpul terpisah. Untuk menghitungnya, pertama hitung jarak lintasan rata-rata dari sebuah simpul v ke seluruh simpul lainnya.

$$d = \frac{\sum_{d \neq w} d(v,w)}{|V(G)| - 1}$$

Setelah panjang lintasan rata-rata dari seluruh simpul, hitung median dari seluruh d, nilai median ini adalah *characteristic path length.* Secara sederhana, *characteristic path length* adalah panjang lintasan rata-rata untuk mencapai sembarang pasangan simpul pada suatu graf.

Sementara itu, *clustering coefficient (C)* adalah ukuran seberapa besar kecenderungan sebuah simpul untuk membentuk kelompok. Nilai C adalah perbandingan jumlah sisi yang terbentuk antar tetangga simpul n dengan jumlah sisi maksimum yang terbentuk antar tetangga simpul n.

Berdasarkan karakterisitik tersebut, Adamic telah membuktikan bahwa sifat *Small World Network* dari graf *World Wide Web* berdasarkan definisi dari Wats dan

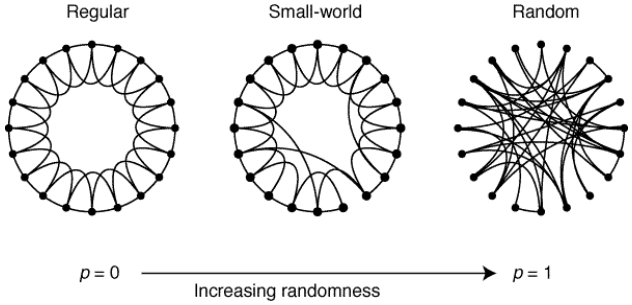
Strogatz. Mereka mendefinisikan sifat-sifat dari graf *Small World Network* sebagai berikut :

1. Nilai *clustering coefficient (C)* pada *Small World Network* lebih besar daripada graf acak Erdos Renyi dengan jumlah simpul yang sama.

2. Nilai *characteristic path length* (L) pada *Small World Network* kurang lebih sama kecilnya dengan L pada graf acak Erdos Renyi dengan jumlah simpul yang sama.

Adamic melakukan penelitian dengan menggunakan data uji dengan sekitar 50.000.000 (lima puluh juta) halaman web dan 269.794 *website*. Data ini dikumpulkan menggunakan *crawler* Alexa. Alexa sendiri merupakan web yang salah satunya berfungsi memeringkat *website* di seluruh dunia. Dengan menganalisis graf dengan *website* berdomain .edu , dia menemukan bahwa nilai L graf .edu adalah 4,062 , sementara itu nilai L graf acak Erdos Renyi adalah 4,048. Nilai C pada garf .edu adalah 0,156 , sementara itu nilai C pada graf acak (*random graph*) Erdos Renyi adalah 0,0012. Berdasarkan hasil penelitian tersebut, dapat disimpulkan bahwa graf *world wide web* memiliki karakteristik *Small World Network*, sesuai dengan definisi yang ditetapkan oleh Watts & Strogatz.

Watts & Strogatz menjelaskan mengenai prosedur pembentukan graf *Small World Network* yang disebut dengan *Random Rewiring Procedure*. Pertama, buah graf dengan n simpul, setiap simpul terhubung dengan d buah simpul yang terdekat. Jumlah sisi yang terbentuk pada graf ini adalah m = (n x d) / 2. Graf ini juga disebut sebagai graf teratur.



Gambar 2. Perbandingan antara graf teratur, graf Small-world dan random graph Erdos Renyi.
Sumber_Gambar : : http://www.nature.com/nature/journal/v393/n6684/images/393440aa.eps.2.gif
Diakses : 09 Desember pukul 06:07 WIB

Dengan kata lain, permodelan graf *World Wide Web* menggunakan "*Small World Network"* lebih baik daripada permodelan menggunakan graf acak Erdos-Renyi karena permodelan ini telah didukung menggunakan bukti data empiris. Meskipun demikian, model graf ini masih belum dapat menjelaskan sifat dinamis dari *world wide web*. Model ini tidak menjelaskan proses penghapusan atau penciptaan halaman web dan *hyperlink* baru. Jumlah simpul dan sisi pada model ini selalu tetap.

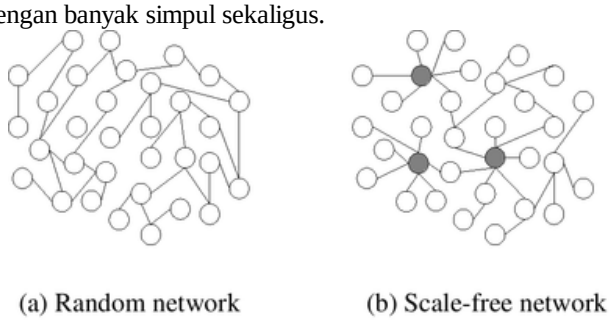
B. Model graf "Scale Free Network" (Barabasi, Albert)

Seperti yang telah dijelaskan, model graf sebelumnya masih belum mencukupi dalam memodelkan graf *Word Wide Web.*

Berdasarkan kelemahan ini, Barabasi & Albert menyusun sebuah model graf "*Scale Free Network"* , yang memiliki karakteristik sebagai berikut :

1. Jaringan graf dapat berkembang dengan penambahan simpul-simpul baru.
2. Simpul yang baru ditambahkan cenderung membuat sisi dengan simpul yang banyak sisi.

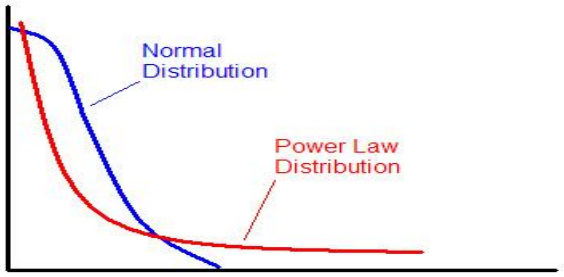
Ciri dari graf "*Free Scale Network"* adalah persebaran derajat pada setiap simpulnya yang tidak merata. Terdapat simpul yang berperan sebagai *hub*. Simpul ini terhubung dengan banyak simpul sekaligus.



Gambar 3.
Perbandingan antara Random Graph Erdos Renyi dan Free Scale Network Graph Barabasi & Albert.
Sumber_Gambar : https://upload.wikimedia.org/wikipedia/commons/thumb/7/77/Scale-free_network_sample.png/400px-Scale-free_network_sample.png
Diakses pada : 09 Desember 2016 pukul 06:42 WIB

Kemunculan simpul *hub* ini disebabkan oleh kecenderungan simpul baru untuk membentuk sisi dengan simpul yang memiliki simpul terbanyak.

Distribusi derajat sisi pada simpul *Scale Free Network Graph* mengikuti hokum pangkat. Probabilitas sebuah simpul memiliki derajat k berbanding lurus dengan $(1/k^n)$, dimana n adalah sebuah konstanta.



Gambar 4 Perbandingan antara Normal distribution dan power law distribution.
Sumber_Gambar : http://farm4.static.flickr.com/3325/3428809345_7ac3dddda6.jpg?v=0
Diakses pada : 09 Desember pukul 07:00 WIB

Grafik tersebut menunjukan bahwa mayoritas simpul memiliki jumlah sisi yang sedikit, namun ada simpul *hub* yang memilki jumlah sisi yang banyak.

Proses pembentukan graf *Scale Free Network* dapat dijelaskan menggunakan mekanisme *Preferential Attachment Model*. Pertama, buat sebuah graf kosong dengan n buah simpul. Lalu, setiap satuan waktu, tambahkan sebuah simpul baru, dengan m buah sisi. Probabilitas sebuah simpul baru membentuk sisi dengan simpul i adalah

$$p = \frac{d_i}{\sum_j d_j}$$

Dimana d_i adalah derajat simpul i, dan sigma d_j adalah jumlah derajat seluruh simpul pada graf. Dengan persamaan di atas, simpul baru akan memiliki kecenderungan untuk membentuk sisi dengan simpul yang memiliki derajat tertinggi.

IV. ANALISIS STRUKTUR GRAF WORLD WORLD WIDE WEB BERDASARKAN DATA EMPIRIS.

Pada bab sebelumnya telah dijelaskan beberapa permodelan *world wide web* menggunakan teori graf. Namun, untuk mendapatkan gambaran nyata mengenai struktur graf *world wide web*, kita harus menjelajahi seluruh halaman web di dalam *world wide web*, mengunduh setiap halamannya dan menganilis keterkaitan antar halaman-halaman web tersebut.

Untuk melakukan penjelajahan dan pengunduhan setiap halaman web di dalam *world wide web*, digunakanlah program *web crawler*. *Web crawler* adalah program untuk menjelajahi isi *world wide web* secara otomatis dan juga sistematis. *Web crawler* menjelajahi *world wide web* dengan mengikuti strukutr *hyperlink* yang terdapat di dalam halaman web.

Pertama, web crawler diberikan sebuah daftar *Uniform Resouce Locator* (URL) yang harus dikunjungi. Kemudian, *crawler* akan mengunjungi URL tersebut, mengunduh halaman web pada URL, mengidentifikasi seluruh *hyperlink* yang terdapat di dalam halaman web dan memasukan daftar *hyperlink* tersebut ke daftar URL yang harus dikunjungi. Proses ini dilakukan berulang-ulang hingga crawler memutuskan kapan untuk berhenti.

Menurut Trupti, dkk ada beberapa komponen dasar yang dimiliki oleh web crawler, yaitu :

1. *Crawler Frontier*
   Sebuah daftar URL yang akan dikunjungi oleh web Crawler
2. *Page Downloader*
   Program untuk mengunduh halaman web berdasarkan URL yang ada di dalam Crawler Frontier
3. *Web Repository*
   Sebuah tempat penyimpanan halaman web yang telah berhasil diunduh oleh web crawler

Web crawler digunakan oleh peneliti graf *world wide web* untuk menganalisis struktur graf pada *world wide web*.

Selain itu, *web crawler* juga digunakan dalam mesin pencari. *Crawling* (proses pengunduhan halaman web untuk dianalisis lebih lanjut) merupakan proses yang pertama kali dilakukan oleh mesin pencari dalam pencarian informasi di dalam *world wide web*.

V. KESIMPULAN

Struktur *World Wide Web* dapat dimodelkan menjadi sebuah graf berarah, dimana halaman web merupakan simpul pada graf dan *hyperlink* merupakan sisi pada graf. Karena sifat strukturnya yang terdesentralisasi dan acak, graf ini dapat dimodelkan menggunakan model graf acak, seperti graf acak Erdos-Renyi, *Small World Network* Watts-Strogatz, dan *Scale Free Network Graph* (dalam makalah ini hanya dibahas dua tipe graf yang terakhir). Dari ketiga model graf tersebut, model graf *scale free network* lebih cocok untuk memodelkan graf *world wide web* karena sudah dikonfirmasi kebenarannya oleh data empiris.

Sementara itu, struktur graf *world wide web* juga dapat ditentukan menggunakan data empiris. Halaman-halaman web dalam jumlah besar dapat diunduh dari *world wide web* menggunakan *web crawler* untuk dianalisis lebih lanjut.

.

VII. UCAPAN TERIMAKASIH

Penulis mengucapkan puji dan syukur kepada Allah SWT karena atas rahmat dan karunia-Nya penulis dapat menyelesaikan makalah ini. Selain itu, penulis juga mengucapkan banyak terimakasih kepada Bapak Ir. Rinaldi Munir, M.T dan Ibu Harili selaku dosen Matematika Diskrit (IF-2120) yang telah memberikan ilmu yang bermanfaat kepada penulis sebagai dasar dalam penulisan makalah ini.

REFERENSI

[1] Narsigh Deo,Panjah Grupta, "World Wide Web : Graph Theoretic Perspective", University of Central Florida, 2001.
[2] Reka Albert, Hawoong Jeong, Albert-Laszlo Barabasi, "Diameter of world wide web" in nature, vol 401, 1999
[3] Andre Broder, Andrew Tomskins, dkk, "Graph structure in the web,Alta Vista Company,2000.
[4] D.Watts, S. Strogatz, "collective dynamics of 'small world' networks" in nature, volume 393, 1998.
[5] Jungho Choo, "crawling the web : Discovery and Maintenance of Large Scale Web Data", standford university, 2001.
[6] Albert-Laszlo Barabasi, Eric Bonabeau, "Scale Free Netwok", in scientific American 2003.
[7] Trupti V. Udapure, Ravindra D. Kali, Rajesh C. Dharmik, Study of Web Crawler and its different types". In IOSR Journal of Computer Engineering, 2014.

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 2016

Farhan Makarim 13515003