Kinerja Web Crawler Menggunakan Metode Html Dom
The Performance Of The Web Crawler Using Html Dom Method

Muhammad Takdir Muslihi

Jurusan Teknik Elektro , Fakultas Teknik, Universitas Fajar
Jl. Prof. Dr. Abdurrahman Basalamah II No.25, Makassar, 90123, Telp/Fax: 0411-4460084

E-mail: mtakdir.muslihi@gmail.com

Abstrak – Penelitian yang dilakukan untuk menerapkan *Web Crawler* ke dalam sistem informasi untuk mencari dan memperlihatkan daftar link dari website yang isinya relevan terhadap data dosen di dalam sistem informasi akademik. Algoritma yang akan digunakan adalah algoritma BFS, Backlink dan Pagerank. Hasil dari penelitian ini adalah Web Crawler sukses mendapatkan 187 daftar link yang berkaitan dengan dosen pada sistem informasi akademik.Dilihat dari sisi algoritma yang digunakan dalam web crawler, algoritma Pagerank dan BFS bekerja lebih efisien dalam menemukan link dalam sebuah website dibandingkan algoritma BFS.
.

Kata Kunci: Web Crawler, Sistem Informasi Akademik

Abstract *– Currently, the courses are still looking forthe supporting documents of accreditation especially of lecturers data manually one by one so it takes a long time. This research aimed to apply Web Crawler into the information system used to investige an demonstrate link list of the websites of which the content was relevant to the lectures' data in the accreditation information system. The algorithm used were the BFS, Backlink and PageRank algorithms. The result of the research indicated that Web Crawler had successfully discovered 187 link lists which were related to the lecturers in the accreditation information systems. Viewed from the side of the algorithms which was used in the Web Crawler, the PageRank and BFS algorithms had worked more efficiently to discover the link in a web site compared to the Backlink algorithm.*

Keywords: *Web Crawler,Academic Information System*

PENDAHULUAN

Informasi atau pun berita seperti informasi dosen atau mahasiswa di sebuah perguruan tinggi biasanya tersebar secara acak di World Wide Web (WWW) dan tidak terintegrasi dengan sistem akademik. Beberapa penelitian seperti yang dilakukan oleh Hasbullah (2013), menghasilkan sebuah sistem Informasi akademik yang mengintegrasikan data-data dari sistem informasi yang telah ada yaitu sistem informasi akademik dengan menggunakan teknologi web service. Namun dalam penelitian tersebut data yang dientegrasikan masih terbatas pada nilai dan angka di bidang akademik mahasiswa, belum mencakup informasi lain yang berupa teks.

Dari masalah tersebut perlu dikembangkan sistem yang mampu memperoleh informasi berupa teks secara otomatis dari website yang berkaitan terhadap perguruan tinggi tersebut. Penelitian yang akan dilakukan adalah untuk menerapkan metode web crawler ke dalam sebuah sistem informasi untuk mencari dan dan memperlihatkan dokumen online yang informasinya relevan terhadap data-data di dalam sistem informasi akademik (Muslihi & Hutomy, 2012). Tujuan dari penelitian ini adalah agar sistem informasi akademik yang ada saat ini memiliki fungsi untuk menampilkan daftar dokumen secara cepat dari website.

METODOLOGI PENELITIAN

Jenis penelitian ini adalah penelitian eksperimental yang bersifat aplikatif sehingga dari perumusan masalah dapat dilakukan dengan metode studi pustaka, metode pengumpulan data lapangan dan pembuatan aplikasi berbasis web. Pengumpulan data lapangan adalah dengan menyalin data dari database pada server yang bertempat di fakultas teknik Universitas X. Hasil data dari penelitian ini akan dianalisis dengan metode metrik Harvest-rate. Data *Harvest-rate* merepresentasikan pecahan halaman Web yang ditelusuri yang memenuhi target ***#r*** dalam keseluruhan halaman *#p* yang diperoleh.

Pembuatan aplikasi web adalah dengan merancang aplikasi web bernama *Web crawler*. *Web crawler* adalah sebuah perangkat lunak yang digunakan untuk

menjelajah serta mengumpulkan halaman-halaman *web* yang selanjutnya diindeks oleh mesin pencari (Zuliarso, 2009). Hasil pengumpulan situs Web selanjutnya akan diindeks oleh mesin pencari sehingga mempermudah pencarian informasi di Internet.Ide dasarnya sederhana dan mirip dengan saat menjelajahi halaman *web*site dengan menggunakan *browser*. Bermula pada pointawal berupa sebuah link alamat *web*site dan dibuka pada *browser*, lalu *browser* melakukan *reques*t dan men-*download* data dari *webserver* melalui protokol HTTP. Setiap *hyperlink* yang ditemui pada konten yang tampil akan dibuka lagi pada windows/tab *browser* yang baru, demikian proses terus berulang. Sebuah *web crawler* mengotomatisasikan pekerjaan itu (Giles & Council, 2010).

Dalam langkah pertama, sebuah *webcrawler* mengambil URL dan mengunduh halaman dari Internet berdasarkan URL yang diberikan. Seringkali halaman yang diunduh disimpan ke sebuah file atau ditempatkan di basisdata. Dengan menyimpan halaman web, maka *crawler* atau program yang lain dapat memanipulasi halaman itu untuk diindeks (dalam kasus mesin pencari) atau untuk pengarsipan untuk digunakan oleh pengarsip otomatis (Rosmala & Syafei, 2012).

Tahap kedua, *Web crawler* memparsing keseluruhan halaman yang diunduh dan mengambil link-link ke halaman lain. Tiap link dalam halaman didefinisikan dengan sebuah penanda HTML yang serupa dengan yang ditunjukkan disini :"<AHREF="http://example">Link</A>". Setelah *crawler* mengambil link dari halaman, tiap link ditambahkan ke sebuah daftar untuk di-*crawl*.

Langkah ketiga dari *web crawling* adalah mengulangi proses. Berikut adalah rancangan algoritma dasar webcrawler:

```
    enqueue(url queue, starting url)
    while (not empty(url queue))
    url = dequeue(url queue)
    page = crawl page(url)
    enqueue(crawled pages, (url,
page))
    url list = extract urls(page)
    foreach u in url list
    enqueue(links, (url, u))
    if (u/∈url queue and (u,-
)/∈crawled pages)
    enqueue(url queue, u)
    reorder queue(url queue)
```

Deskripsi fungsi :

enqueue(queue, element) : menambahkan elemen di ujung dari queue.

Dequeue (queue) : menghilangkan elemen di awal queue dan memberikan ke program yang memanggil. Reorder_queue : mengurutkan queue menggunakan informasi yang ada di tautan.

Dalam penelitian kali ini akan diusulkan sistem yang menerapkan webcrawler kemudian diimplementasikan ke dalam sistem informasi akademik. Akan diuji 2 teknik dalam mendownload halaman web, yang pertama adalah dengan mendownload dokumen dalam bentuk file htmldan akan disimpan ke dalam database. Teknik kedua yang akan diuji adalah adalah dengan menggunakan teknik Document Object Model (DOM) data yang akan disimpan dalam database adalah berupa konten teks saja.

Dalam penelitian ini akan digunakan dan diuji 3 algoritma dalam webcrawler. Yang pertama adalah BFS yang merupakan dasar algoritma webcrawler, kemudian yang kedua adalah algoritma backlink yang menghitung jumlah link yang menuju pada sebuah url, dan yang ketiga adalah algoritma pagerank (Gupta, 2012)..

Pada algoritma BFS, tidak ada pengurutan lagi. Antrian link yang akan di-crawl telah didefinisikan pada baris kedelapan dan kesembilan pseudocode.

Pada algoritma backlink, pengurutan dilakukan berdasarkan nilai backlink setiap url. Backlink adalah jumlah link dari keseluruhan web yang ditelusuri yang menuju padalink tersebut [8] . Pseudocode dari algoritma backlink adalah berikut ini:

foreach*u* in url queue

backlink count[*u*] = number of terms (-,*u*) in links

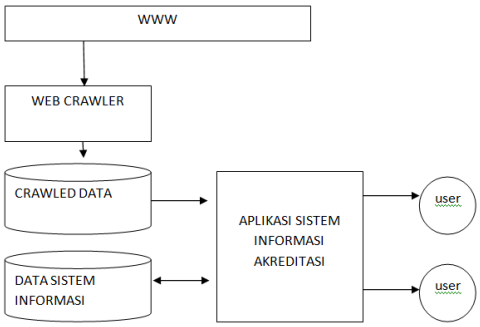
sort url queue by backlink count[*u*].

Pada algoritma pagerank, pengurutan berdasarkan nilai pagerank pada link tersebut. Mendapatkan Nilai pagerank dari dari setiap link adalah dengan memanggil fungsi bernama CalculatePagerank , dimana di dalam fungsi tersebut dirumuskan sebagai berikut:

$$IR[u] = (1 - 0.9) + 0.9 \sum_i \frac{IR[v_i]}{c_i} \quad (1)$$

Dimana (*vi, u*) ∈ links dan *ci* adalah jumlah link dalam halaman *vi* . Setelah itu pengurutan dilakukan

berdasarkan nilai pagerank yang telah diperoleh. Berikut bagan sistem yang diusulkan:



Gambar 1. Flow Chart Sistem Usulan

Pada gambar 1, Data dari www akan didownload oleh web crawler yang dibangun dengan menggunakan bahasa pemrograman php. Setelah itu hasil *crawl* akan disimpan ke dalam mysql database. Kemudian data hasil *crawl* akan disinkronkan dengan data yang telah ada pada sistem informasi akademik, dalam kasus ini adalah data nama dosen. Proses sinkronisasinya adalah setiap nama dosen akan menjadi kata kunci untuk pencarian informasi pada data hasil *crawl* menggunakan mysql query. Setelah diperoleh, hasil pencarian itu akan ditampilkan bergandengan dengan nama dosen.

HASIL DAN PEMBAHASAN

Berikut ini adalah sampel data hasil Web Crawler pada situs eng.unhas.ac.id/id/informatika dengan batas url yang dicrawl adalah 100 url.

id	URL	Title	Contents
164	http://eng.unhas.ac.id/geologi	Program Studi Teknik Geologi	Program Studi Teknik Geologi ...
165	http://eng.unhas.ac.id/pertambangan	Program Studi Teknik Pertambangan ...	Program Studi Teknik Pertambangan ...
166	http://eng.unhas.ac.id/kelautan	Program Studi Teknik Kelautan	Program Studi Teknik Kelautan ...
167	http://eng.unhas.ac.id/fakultas	FAKULTAS TEKNIK	FAKULTAS TEKNIK $(function...
168	http://eng.unhas.ac.id/s2elektro	\n FAKULTAS TEKNIK	"-//W3C//DTD HTML 4.0 Transitional//EN\" \"http:/...
169	http://eng.unhas.ac.id/s2mesin	\\n FAKULTAS TEKNIK	\n\"-//W3C//DTD HTML 4.0 Transitional//EN\\\" \\\...
170	http://eng.unhas.ac.id/informatika/id/page/2/Profi...	Program Studi Teknik Informatika ...	Program Studi Teknik Informatika ...
171	http://eng.unhas.ac.id/informatika/id/page/2	Program Studi Teknik Informatika ...	Program Studi Teknik Informatika ...
172	http://eng.unhas.ac.id/informatika/en/page/2	Program Studi Teknik Informatika ...	Program Studi Teknik Informatika ...

Gambar 2. Data Hasil Web Crawler

Gambar 2 menunjukkan database dengan nama crawled data yang menyimpan hasil dari proses web crawler. Field URL pada gambar 3 berisikan daftar link URL dari halaman website yang didownload oleh Web Crawler. Dengan menggunakan teknik DOM, web crawler mengekstrak judul dalam halaman web tersebut kemudian disimpan ke dalam field 'title'. Kemudian web crawler menyimpan konten yang berupa teks dari halaman website tersebut ke dalam field 'Contents'.Kemudian pada Web Crawler yang menggunakan algoritma Pagerank dan backlink, nilai pagerank atau backlink disimpan dalam kolom 'value.' Hasil dari Web Crawler dengan tiga algoritma disimpan pada tiga database yang berbeda.

Tabel 1 Perbandingan Metode download

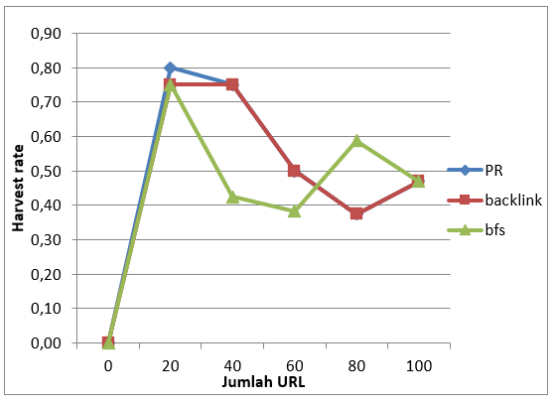
Metode Download	**Jumlah URL**	**Besar data (MB)**	**Waktu Eksekusi (second)**
Html File	100	2,13	59,043
Text Content Only	100	0,74	59,006

Tabel 1 memperlihatkan perbandingan antara dua jenis metode download yang digunakan dalam sistem. Kedua metode download dibatasi hanya pada 100 URL. Metode pertama yaitu dengan mendownload konten website dalama bentuk file HTML menggunakan memori penyimpanan sebesar 2,13 MB dan waktu eksekusinya selama 59,043 detik. Pada metode download yang kedua yaitu dengan mendownload halaman website berupa konten teks saja, metode tersebut menggunakan memori penyimpanan dengan sebesar 0,74 MB dan waktu eksekusinya selama 59,006 detik.

Nama dosen pada sistem informasi akademik akan menjadi acuan untuk sinkronisasi dengan data hasil Web Crawler. Proses sinkronisasi yang disebut dengan *information retrieval* menghasilkan informasi berupa berita , publikasi, dan penelitan masing-masing dosen.

Gambar 3. Tampilan sistem

Gambar 3 adalah tampilan antarmuka sistem pada sistem informasi akademik yang menampilkan daftar link dokumen dari internet yang berkaitan dengan dosen tersebut. Nama dosen menjadi acuan pencarian dokumen pada database hasil web crawler.Informasi yang ditampilkan adalah judul/ title dokumen, link yang menuju ke URL website asli yang memiliki informasi dosen tersebut, dan sebagian isi dokumen dari website yang dituju.



Gambar 4 . Kinerja Web Crawler

Gambar 4 menunjukkan grafik hasil analisis data yaitu nilai harvest rate yang menunjukkan kinerja web crawler. Sumbu x horizontal pada gambar menunjukkan jumlah link yang didownload oleh Web Crawler, sedangkan sumbu y vertikal menunjukkan harvest rate atau kinerja dari web crawler pada sistem informasi akademik. Pada saat jumlah link yang ditelusuri oleh web crawler berjumlah 20, harvest rate ketiga algoritma mencapai 0,80. Pada saat web crawler menelusuri 20 hingga 60 link pada sebuah website, nilai harvest rate algoritma Pagerank dan Backlink sama yaitu 0,75 kemudian menurun hingga 0,50. Sedangkan nilai harvest rate algoritma BFS yaitu berada di sekitar 0,40. Pada saat web crawler menelusuri 80 link, nilai harvest rate algoritma BFS adalah 0,60 . Sedangkan algoritma Pagerank dan Backlink berada pada nilai 0,40. Pada saat web crawler menelusuri 100 link, nilai harvest rate ketiga algoritma adalah 0,48.

Penelitian ini memperlihatkan tersedianya sistem yang menampilkan daftar link dari website yang isinya relevan terhadap data dosen di dalam sistem informasi akademik. Penelitian ini juga menemukan bahwa metode download halaman web menggunakan DOM mendownload data lebih sedikit dibandingkan dengan metode download dengan menyimpan halaman html (Ahuja, 2014). Hal ini disebabkan dengan metode DOM halaman yang disimpan adalah hanya konten web yang berupa teks sedangkan metode mendownload halaman file html menyimpan seluruh source code html dari sebuah halaman web, sehingga menggunakan lebih banyak memori unutk penyimpanan data (Mishra dkk., 2011).

Kinerja dari *crawler* yang ideal dikenal metrik *harvest-rate :*

$$hr = \frac{\#\mathrm{r}}{\#\mathrm{p}}, hr \in [0,1] \qquad (2)$$

Harvest-rate merepresentasikan pecahan halaman Web yang ditelusuri yang memenuhi target ***#r*** dalam keseluruhan halaman *#p* yang diperoleh [3]. Dalam penelitian kali ini target **#r** adalah halaman yang menyimpan dokumen yang mengacu pada sistem informasi akademik. Data yang djadikan acuan adalah nama dosen pada sistem informasi akademik. Dari metrik harvest rate, diketahui bahwa algoritma PR bekerja paling baik pada 20 url pertama. Kinerja Algoritma PR dan backlink lebih baik dari pada algoritma bfs pada saat link yang dicrawl 10 % dari jumlah link pada keseluruhan website (Jain & Agrawal, 2014). Algoritma bfs mampu mendownload halaman yang lebih banyak mengandung kata kunci jika Web Crawler dibatasi mendowload 80 link, tapi hal tersebut menjadi tidak efisien Kinerja Algoritma Pagerank dan backlink lebih baik dan efisien dari pada algoritma bfs pada saat link yang dicrawl adalah berjumlah 20 hingga 60 link. Efisien disini maksudnya adalah dengan mendownload lebih sedikit url, web crawler telah mendapatkan banyak link yang mengandung kata kunci dosen

KESIMPULAN

Hasil dari proses web crawler pada penelitian ini disimpan pada database yang bernama *Crawled Data*. Kemudian sistem informasi akademik menampilkan daftar link yang berkaitan dengan dosen yang berasal

dari database tersebut.Algoritma Backlink dan Pagerank memiliki kinerja yang lebih baik dan lebih efisien untuk digunakan sebagai algoritma Web Crawler dengan halaman yang didownload berkisaran 20-60 halaman. Metode download konten berupa teks dari halaman website lebih baik dibandingkan mendowload konten berupa file html karena menggunakan memori lebih kecil.

DAFTAR PUSTAKA

Ahuja M. (2014). Web Crawler : Extracting the Web Data. International Journal of Computer Applications, 13(3):132-137.

Giles C & Councill I. (2010). Measuring the web crawler ethics. Proceedings of the 19th International Conference on World Wide Web - WWW ,19(10): 1101.

Gupta G. (2012). Increasing The Efficiency Of Crawler Using Costumized Sitemap. International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166.

Hasbullah. (2013). *Pengembangan Sistem Informasi Akreditasi yang Terintegrasi Dengan SIAKA FT-UH* (Skripsi). Makassar: Universitas Hasanuddin.

Jain A & Agrawal C. (2014). Sourfey of web crawler algorithm. International Journal of Computing and Business Research , 1(2): 7–14.

Mishra S., Jain A., & Sachan, D. (2011). A Query based Approach to Reduce the Web Crawler Traffic using HTTP Get Request and Dynamic Web Page. International Journal of Computer Applications. 14(3): 8–14.

Muslihi M.T & Hutomy A.(2013). *Pengembangan Sistem Informasi Akreditasi.* (Skripsi). Makassar: Universitas Hasanuddin.

Rosmala D & Syafei R. (2012). Implementasi Web Crawler pada Social Media Monitoring. Jurnal informatika, 2(5):57–68.

Tamara G., Brett W., & Kenny J. (2005). Higher-Order Web Link Analysis Using Multilinear Algebra. Albuquerque: Sandia National Laboratories.

Zuliarso E & Mustofa K. (2009). Crawling Web berdasarkan Ontology. Jurnal Teknologi Informasi DINAMIK, XIV(2): 105–112.