Prosiding
ANNUAL RESEARCH SEMINAR 2016
6 Desember 2016, Vol 2 No. 1

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Perbandingan Algoritma *Breadth First Search* dan *Depth First Search* Sebagai *Focused Crawler*

Doddy Teguh Yuwono, Abdul Fadlil, Sunardi
Magister Teknik Informatika
Universitas Ahmad Dahlan
Yogyakarta, Indonesia
doddy.zha09@gmail.com

***Abstrak* - Perkembangan dunia internet dan kebebasan dari sesorang untuk membuat web mengakibatkan peningkatan jumlah penyebaran informasi, dokumen, ataupun artikel dengan sangat pesat. Hal tersebut menjadi suatu permasalahan untuk seseorang melakukan pencarian data yang relevan dan dibutuhkan dari web misalnya web pariwisata. *Web Crawler* diperlukan khusus diperuntukkan bagi pengguna internet mencari halaman yang relevan. Web Crawler adalah suatu program yang melakukan proses scanning ke halaman-halaman internet untuk dibuat indexnya dan mendukung sebuah *search engine*. Berbeda dengan crawler yang dipakai oleh search engine pada umumnya bertujuan mengumpulkan semua halaman Web sebanyak mungkin, sedangkan focused crawler dapat memberikan halaman web sesuai dengan topik yang dimaksud. Focused crawler secara selektif menelusuri dan mengambil halaman web yang relevan. Penelitian ini dilakukan untuk memperoleh perbandingan antara Algoritma Breadth First Search (BFS) dan Depth First Search (DFS) sebagai Algoritma pencarian serta didukung dengan Naïve Bayes Classifier (NBC) untuk menilai perbandingan dari kedua Algoritma pencarian tersebut, diharapkan dengan kombinasi ini menghasilkan klasifikasi yang tinggi dan maksimal.**

Kata Kunci : Focused crawler, search engine, breadth first search, depth first search, naïve bayes classsifier.

I. LATAR BELAKANG

Pesatnya dunia internet dan kebebasan membuat web mengakibatkan web bekembang dengan sangat pesat dan menjadi permasalahan untuk melakukan pencarian data ataupun informasi yang memang dibutuhkan dari halaman web. Halaman web dapat dikatakan kaya apabila konten-konten yangdapat dilihat ataupun tersedia sehingga terdapat akses terhadap konten tersebut oleh pengguna. Konten pada halaman web bisa terdiri dari artikel maupun sebuah dokumen yang diunggah pada halaman web tersebut. Pemindaian suatu halaman web dapat difokuskan dengan mencari dokumen apa saja yang tersimpan dalam web tersebut yang dapat diperoleh dan diakses secara publik.

Web *crawler* merupakan bagian dari komponen yang terdapat dalam search engine. Web *crawler* atau dikenal dengan istilah Web Spider atau Web Robot memiliki fungsi atau tugas utama dalam melakukan penjelajahan dan pengambilan informasi dari halaman web yang terdapat di internet. Hasil pengumpulan situs web selanjutnya akan diindeks sehingga mempermudah pencarian informasi. Secara singkat, proses dari web *crawler* dimulai dengan memberikan URL awal sebagai benih (*seed*) penelusuran ke dalam sebuah antrian (*queue*). Selanjutnya crawler akan men-download halaman web berdasarkan URL yang dimaksud. Setelah disimpan di dalam koleksi, halaman web yang diperoleh akan diurai (*parsed*) untuk diekstrak outgoing link yang belum dikunjungi dan kembali dimasukkan ke dalam queue. Proses terus berlangsung sampai antrian URL kosong atau kondisi berhenti terpenuhi. Crawling strategy merupakan variasi algoritma saat pemilihan URL pada *queue*. Algoritma yang bisa digunakan yaitu Algoritma BFS dan DFS.

Berbeda dengan crawler biasa, focused crawler atau biasa disebut topical crawler secara selektif menelusuri dan mengambil halaman web yang relevan dengan topik tertentu. Untuk menentukan link yang akan ditelusuri, focused crawler dipandu oleh sebuah classifier yang akan mengklasifikasi relevansi topic halaman web. Metode yang bisa digunakan dalam klasifikasi yaitu *naive bayes, algoritma genetika, neural network* dan *support vector machine*. *Naïve Bayes classifier* dipilih sebagai metode pengklasifikasian karena rumusannya sederhana tetapi handal. Mengenai implementasi focused crawler pada web Pariwisata agar terlebih dahulu mengetahui perbedaan antara focused crawler dengan crawler biasa dalam menjaring URL baru.

Beberapa masalah yang sering ditemui dalam crawling yaitu kebijakan situs web yang tidak memperbolehkan situsnya

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

dikunjungi atau terus menerus mengunjungi web yang sama (*spider trap*). Dalam penelitian ini dilakukan mekanisme untuk menanggulanginya dengan menghentikan penelusuran jika ditemui kendala tersebut. Adapun perhitungan performansi sistem dilakukan berdasarkan akurasi, presisi (*precision*), dan waktu proses. Akurasi merupakan persentase ketepatan kelas prediksi dengan kelas aktual hasil klasifikasi sistem. Precision merupakan persentase dokumen relevan dengan dokumen yang muncul dalam pencarian. Waktu proses merupakan selisih waktu mulai dengan waktu selesai. *Precision* dibagi menjadi dua, yaitu *precision* ya dan tidak sesuai dengan hasil klasifikasi URL terhadap sistem yang mengakses halaman web.

II. KAJIAN PUSTAKA

2.37. *Kajian Terdahulu*

Pada penelitian Ganguly tahun 2014 membahas tentang penerapan content block segmentation sebagai focus crawling pada suatu web. Objek penelitiannya meliputi penerapan content block segmentation yang berfokus terhadap keyword url. [1]

Radu tahun 2014 yang melakukan penelitian pada focus crawler terhadap huruf romawi dan bahasa romawi. [2]

Kamil Çaliskan dan Rifat Ozcan tahun 2013 objek penelitian Fokus crawlers, tujuannya untuk mengambil halaman hanya yang berkaitan dengan subjek tertentu dari juta halaman web di internet. Penelitian ini membuat fokus crawler untuk memprediksi apakah halaman berkaitan dengan target dari subjek yang diinginkan atau tidak. [3]

Pada penelitian Sunarya tahun 2012 objek penelitiannya Menggunakan algoritma fish search. Dengan metode ini focus crswler bekerja dengan terlebih dahulu mengecek relevansi node awal dengan topik, kemudian dilanjutkan menelusuri node anak. Algoritma fish search memerlukan waktu proses yang relatif lama. Pemrosesan terhadap setiap web yang dikunjungi membuat sistem berjalan diatas waktu 3 menit. [4]

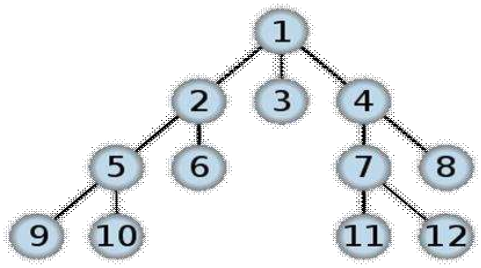
Menurut Kamus Besar Bahasa Indonesia tahun 2002 halaman 43 : [5]

"Analisis adalah penguraian suatu pokok atas berbagai bagiannya dan penelaahan bagian itu sendiri serta hubungan antar bagian untuk memperoleh pengertian yang tepat dan pemahaman arti keseluruhan".

Menurut Komarudin tahun 2001 halaman 53 : [6]

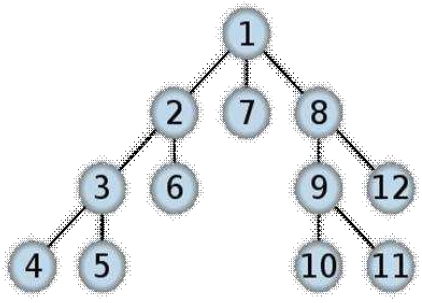
"Analisis adalah kegiatan berfikir untuk menguraikan suatu keseluruhan menjadi komponen sehingga dapat mengenal tanda-tanda komponen, hubungannya satu sama lain dan fungsi masing-masing dalam satu keseluruhan yang terpadu".

Algoritma BFS merupakan salah satu algoritma pencarian yang menguji tiap link pada sebuah halaman sebelum memproses kehalaman berikutnya. Jadi, algoritma ini menelusuri tiap link pada halaman pertama dan kemudian menelusuri tiap link pada halaman pertama pada link pertama dan begitu seterusnya sampai tiap level pada link telah dikunjungi. Alur dari Algoritma BFS dapat dilihat pada Gambar 1.



Gambar 1. Alur pencarian algoritma Breadth First Search

Algoritma DFS, pencarian dalam algoritma ini dimulai dari node yang paling kiri sampai node yang ada pada setiap level selesai dilalui. Jika pada level yang paling dalam solusi yang diharapkan belum diperoleh, maka pencarian akan dilanjutkan pada node yang berada disebelah kanan node awal. Node yang berada di kiri dapat dihapus dari memori agar memori yang digunakan menjadi lebih sedikit. Jika pada level yang paling dalam belum juga ditemukan solusi dari pencarian, maka pencarian akan dilanjutkan pada level sebelumnya. Demikian seterusnya proses akan terus berlangsung sampai ditemukannya solusi. Jika pada level dan node dari algoritma pencarian ini menumakan solusi, maka proses backtracking (penelusuran untuk mendapatkan jalur yang diinginkan) tidak diperlukan. Alur dari algoritma DFS dapat dilihat pada Gambar 2.



Gambar 2. Alur pencarian Algoritma Depth First Search

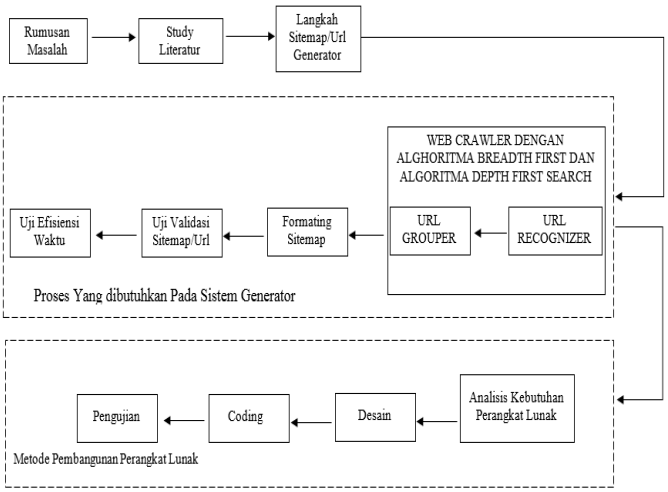
ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Naïve Bayes Classifier (NBC) merupakan pengklasifikasi dengan probabilitas sederhana. Teorema dalam statistika untuk menghitung peluang, *Bayes Optimal Classifier* menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. NBC mengasumsikan ketidaktergantungan (independent) yang tinggi dari suatu pengklasifikasian. Keuntungan penggunaan NBC adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (*training data*) yang kecil sebagai penentu estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena pada NBC yang diasumsikan sebagai variable independent, maka hanya variasi dari suatu variable dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi. Rumus dari *Naïve Bayes Classifier* dapat dilihat pada persamaan (1).

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\,p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}. \quad (1)$$

III. METODOLOGI PENELITIAN

2.38. *Desain Penelitian*



Gambar 3. Desain Penelitian

Penjelasan Desain Penelitian pada Gambar 3 adalah sebagai berikut :

- Rumusan Masalah

  Beberapa permasalahan yang menjadi minat peneliti untuk menjadikan penelitian ini sebagai penelitian adalah menganalisis perbedaan antara algoritma pencarian Breadth First Search dan Algoritma Depth First Search sebagai focused crawler, pendalaman proses pembuatan sitemap, kriteria sitemap yang dapat diterima oleh mesin pencari dan efisiensi metode *Search Engine Optimization* menggunakan sitemap terhadap waktu pengenalan mesin pencari pada sebuah website.

- Studi Literatur

  Setelah dilakukan perumusan masalah, maka dilakukan pencarian teori, konsep, dan sumber Penelitian lain yang dapat dijadikan landasan teori untuk penelitian, hal ini dilakukan agar penelitian mempunyai landasan yang kuat. Tahapan studi literatur pada penelitian ini dilakukan dengan cara mempelajari literature yang meliputi konsep pencarian dengan menggunakan algoritma Breadth First Search dan Algoritma Depth First Search yang diimplementasikan pada halaman sebuah website. Selain itu pada tahap ini dilakukan pencarian referensi tentang URL *recognizing*, pola pengelompokan URL dan metode *Search Engine Optimization* menggunakan *sitemap*.

- Proses Sitemap Generator

  Pada tahap ini akan dikaji hal-hal yang berkaitan dengan sitemap generator. Secara umum proses yang harus dilalui untuk membuat perangkat lunak sitemap generator adalah tahap pengumpulan url (menggunakan web crawler dengan Algoritma *Breadth First Search*, Algoritma *Depth First Search*, dan URL *recognizer*), tahap url grouping, tahap pembuatan sitemap dari data url yang sudah difilter dan dikelompokan, dan validasi sitemap yang telah dibuat. Setelah proses yang disebutkan sebelumnya dilewati dengan sukses, dilakukan pembuktian efisiensi waktu pengenalan search engine yang terdapat pada metode *Search Engine Optimization* (SEO) berbasis sitemap.

- Analisis Kebutuhan Sistem

  Pada tahapan ini dilakukan perumusan kebutuhan perangkat lunak dari *sitemap generator* yang akan dibangun. Didalam tahap analisis ini segala masalah dan anomali yang terjadi perlu diperhatikan seperti analisis data yang dibutuhkan, proses pengenalan url, pengelompokan url dan penyusunan data hingga menjadi sebuah sitemap yang valid.

- Desain Sistem

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Pada tahap desain sistem, hasil dari analisis kebutuhan perangkat lunak disampaikan ke dalam model desain perangkat lunak yang meliputi pemodelan kebutuhan, arsitektur perangkat lunak dan tampilan antarmuka perangkat lunak. Tahap desain ini digunakan untuk menentukan rancangan fungsi, rancangan data (basis data) dan antarmuka yang akan diimplementasikan pada perangkat lunak sitemap generator.

- *Coding*

  Tahap coding dilakukan untuk menterjemahkan sistem yang telah dimodelkan kedalam bahasa pemrograman. Bahasa pemrograman yang digunakan peneliti adalah PHP dengan didukung MySQL sebagai database.

- Pengujian dan Evaluasi

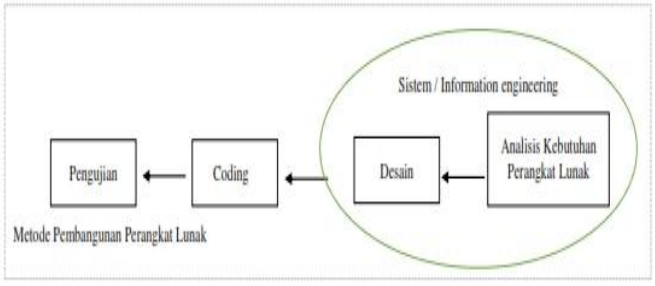
  Tahapan akhir adalah proses pengujian. Teknik pengujian yang digunakan peneliti adalah teknik pengujian Kotak Hitam (*Black Box)*, teknik pengujian ini digunakan untuk lebih memfokuskan pengujian dari sisi fungsionalitasnya.

2.39. Metode Pengembangan Perangkat Lunak

Metode pengembangan perangkat lunak yang digunakan untuk merancang dan membangun sebuah perangkat lunak Web Crawler sebagai Sitemap Generator adalah model proses Sequential Linear , adapun model proses Sequential Linear yang dikembangkan oleh Roger S. Pressman adalah sebagai berikut :

"Model Proses yang digunakan adalah model Sequential Linear (Pressman, 2001. h.28). Model ini adalah model klasik yang bersifat sistematis, berurutan dalam membangun software" [5]

Urutan dari setiap fase dalam pengembangan perangkat lunak dengan model Sequential Linear dapat dilihat pada Gambar 4.



Gambar 4. Fase-fase dalam Model Sequential Linear

Pada tahap ini, akan dilakukan analisis terhadap masalah dan anomali yang timbul dalam proses crawling data sebuah website. Proses tersebut dilakukan untuk menyesuaikan data yang didapat, dengan kebutuhan sistem yang diperlukan oleh sitemap generator.

- Analisis Permasalahan

  Pada tahap ini, akan dilakukan analisis terhadap masalah dan anomali yang timbul dalam proses crawling data sebuah website. Proses tersebut dilakukan untuk menyesuaikan data yang didapat, dengan kebutuhan sistem yang diperlukan oleh sitemap generator.

- Design

  Pada tahap perancangan ini hal yang dilakukan adalah menterjemahkan syarat dari setiap kebutuhan yang diperlukan oleh perangkat lunak ke sebuah perancangan perangkat lunak sebagai model yang akan dibuat sebelum tahap coding. Fokus dari proses ini adalah arsitektur perangkat lunak, representasi tampilan antarmuka, detail algoritma dan struktur data.

- Coding

  Tahap coding adalah tahap penterjemahan perancangan perangkat lunak yang telah dibuat sebelumnya kedalam bentuk bahasa yang dimengerti oleh komputer.

- Pengujian

  Pada tahap pengujian, peneliti menjalankan setiap proses dari program untuk menemukan kesalahan-kesalahan yang ada dan mungkin terjadi pada program selain itu peneliti memastikan bahwa hasil akhir yang didapatkan sesuai dengan harapan.

2.40. *Alat dan Bahan*

Pada penelitian ini digunakan alat penelitian berupa perangkat keras dan perangkat lunak sebagai berikut :

- Perangkat keras
  1. OS Windows 8.1
  2. Processor Intel Inside Core i5
  3. RAM 4 GB
  4. Hard disk 500 GB
  5. Monitor LED 19"
  6. Mouse dan keyboard
  7. Modem

- Perangkat lunak
  1. Notepad ++
  2. Appserver (PHP & MySQL)

ISBN : 979-587-626-0 | UNSRI http://ars.ilkom.unsri.ac.id

Bahan penelitian yang digunakan pada penelitian ini meliputi dari *paper, textbook*, dan dokumentasi lainnya yang didapat dari buku-buku maupun mengakses World Wide Web.

IV. HASIL DAN PEMBAHASAN

Hasil yang diharapkan dari penelitian ini adalah implementasi dari Algoritma Breadth First Search (BFS) dan Depth First Search (DFS) serta didukung dengan Naïve Bayes Classifier (NBC) untuk menilai perbandingan dari kedua Algoritma pencarian tersebut, diharapkkan dengan kombinasi ini menghasilkan klasifikasi yang tinggi dan maksimal sebagai *focused Crawl Web.* Selain itu diharapkan dengan penelitian ini dapat memperdalam pemahaman tentang proses pembuatan sitemap, kriteria sitemap yang dapat diterima oleh mesin pencari dan efisiensi metode *Search Engine Optimization* menggunakan sitemap terhadap waktu pengenalan mesin pencarian pada sebuah halaman website.

Pada algoritma BFS dan DFS yang dibandingkan, peneliti mengharapkan hasil berapa lama *focused crawler* mengenali keyword yang dimasukan, berapa lama rata-rata waktu yang diperlukan untuk mengenali keyword yang dimasukan dan Seberapa Akurat hasil yang diperoleh dari penerapan algoritma BFS dan DFS yang dibandingkan.

REFERENSI

[1] Ionut-Gabriel Radu, Traian Rebedea, "A Focused Crawler for Romanian Words Discovery," *RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference,* Vols. 11-13 September 2014, pp. Pages: 1 - 6, DOI: 10.1109/RoEduNet-RENAM.2014.6955323, 2014.

[2] Bireshwar Ganguly, Devashri Raich, "Performance Optimization of Focused Web Crawling Using Content Block Segmentation," *International Conference on Electronic Systems, Signal Processing and Computing Technologies ,* Vols. 9-10 Januari 2014, pp. 365 - 370, DOI: 10.1109/ICESC.2014.69, 2014.

[3] Kamil Çalişkan, Rifat Ozcan, "Comparing classification methods for link context based Focused crawlers," *International Conference on Electronics, Computer and Computation (ICECCO),* Vols. 7-9 November 2013 , pp. 143 - 146, DOI: 10.1109/ICECCO.2013.6718249, 2013.

[4] Yanuar Firdaus A.W., Arie Ardiyanti Suryani, "Analisis dan Implementasi Focused Crawler Menggunakan Algoritma Fish Pada Web Kesehatan," Universitas Telkom, Bandung, 2012.

[5] A. Nugroho, Rekayasa Perangkat Lunak Verorientasi Objek dengan metode USDP, Yogyakarta: Andi, 2010.

[6] H. Inggiantowi, "Perbandingan Algoritma Penelusuran Depth First Search dan Breadth First Search pada Graf serta Aplikasinya," Program Studi Teknik Informatika , Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung, Bandung, 2014.

[7] Badan Pengembangan dan Pembinaan Bahasa, Kemdikbud , "Kamus Besar Bahasa Indonesia (KBBI)," Ebta Setiawan, 2012-2016. [Online]. Available: http://kbbi.web.id/. [Accessed 11 November 2016].