MIND Journal | ISSN (p): 2528-0015 | ISSN (e): 2528-0902 | Vol. 2 | No. 2 | Halaman 61 - 69
DOI : https://doi.org/10.26760/mindjournal Juli 2017

Perbandingan Metode *Breadth First Search* dan *Backlink* pada *Web Crawler*

JASMAN PARDEDE, ASEP NANA HERMANA, GALIH SWARGHANI

Teknik Informatika, Fakultas Teknologi Industri
Institut Teknologi Nasional (ITENAS) Bandung

Email : jasmanpardede78@gmail.com

ABSTRAK

Dalam sebuah search engine terdapat beberapa komponen penting yang salah satunya adalah crawler / web crawler. Crawler adalah sebuah komponen dalam search engine yang berfungsi untuk mencari semua link pada setiap halaman dimana hasil pengumpulan alamat web selanjutnya akan diindeks. Crawler bekerja dengan menggunakan algoritma pencarian yang beragam, diantaranya adalah Breadth First Search dan Backlink. Breadth first search merupakan algoritma untuk melakukan pencarian secara berurutan dengan mengunjungi setiap simpul secara preorder. Backlink memanfaatkan tautan yang berada disitus lain dan mengarah ke situs tertentu. Adapun hasil dari uji aplikasi yaitu dengan membandingkan kedua metode tersebut dengan cara melihat performa pengambilan URL terbanyak pada Detik.com dan Kompas.com. Metode breadth first search secara performa lebih baik dibandingkan dengan metode backlink, dalam pengujian crawling, perbedaan jumlah url mencapai 25,17 pada website detik.com dan 28,94% pada website Kompas.com.

***Kata kunci**: Web Crawler, Breadth First Search, Backlink.*

ABSTRACT

In a search engine, there are several important components, one of which is a crawler / web crawler. Crawler is a component in the search engine used to find all the links on each page and then collecting the results will then be indexed web address. Crawler works by using search algorithms are diverse, including the Breadth First Search and backlinks. Breadth first search is an algorithm to do the search berurutandengan visiting each node in a preorder. While backlinks utilize our site link contained another and leads to a specific site. The results of the test application is by comparing the two methods by looking at the URL retrieval highest performance on Detik.com and Kompas.com. Breadth first search method is a better performance than the backlink methods, in testing crawling, different numbers on the website url detik.com reached 25.17 and 28.94% in Kompas.com website.

***Keywords**: Web Crawler, Breadth First Search, Backlink.*

1. PENDAHULUAN

Search engine atau dalam secara istilah disebut mesin pencari merupakan sebuah sistem yang memungkinkan pengguna untuk mencari data di internet hanya dengan mengetikan sebuah kata kunci (*keyword*). Dalam sebuah *search engine* terdapat beberapa komponen penting yang salah satunya adalah *crawler* | *web crawler*. *Crawler* dalam sebuah *search engine* berfungsi untuk mencari semua link pada setiap halaman. *Web crawler* melakukan tiga yaitu proses *scraping*, *parsing* dan *indexing.* Proses *scraping* dibantu menggunakan *library* jsoup pada saat proses *scraping* berlangsung. Proses *parsing* yaitu proses saat pemisahan kata atau kode berdasarkan *element* yang sudah ditentukan. Hasil pengumpulan situs *web* selanjutnya akan di *index*. Proses *indexing* yaitu dengan melakukan *split* berdasarkan tanda "/". *Indexing* dilakukan untuk menentukan judul berita, dan melakukan pengelompokkan terhadap data yang sudah didapatkan berdasarkan alur metode. *Crawler* bekerja dengan menggunakan algoritma pencarian yang beragam, diantaranya adalah *Breadth First Search* dan *Backlink*. Pengimplementasian *web crawler* pada penelitian ini dilakukan dengan cara membandingkan dua metode yaitu *Breadth First Search* (BFS) dengan *Backlink* untuk melakukan pencarian dan pengarsipan pada sebuah *website*.

Breadth First Search merupakan algoritma yang melakukan pencarian secara berurutan dengan mengunjungi setiap simpul secara *preorder*. *Backlink* memanfaatkan tautan yang berada disitus lain dan mengarah ke situs tertentu. Inti dari penelitian ini adalah merancang sebuah aplikasi yang memanfaatkan konsep *crawler* untuk melakukan pencarian artikel pada sebuah situs tertentu dan membandingkan hasil penerapan dari kedua konsep algoritma pencarian untuk mengetahui mana yang lebih handal.

2. METODOLOGI PENELITIAN

Web crawler diaplikasikan sebagai sistem pengarsipan berupa alamat *web*, judul berita dan konten artikel dari *web* yang di *crawling. Web crawler* memiliki tiga proses pada saat melakukan *crawling,* yaitu proses pembacaan halaman (*scraping*), proses pemisahan kata (*parsing*), dan proses pengindeksan (*indexing*). Hasil akhir *web crawler* di implementasikan sebagai sistem pengarsipan berupa alamat *web,* judul berita, dan konten artikel. Proses *crawler* ini dilakukan pada *website* Detik.com dan Kompas.com.

Proses *scraping* merupakan proses paling awal saat melakukan *crawling* untuk mendapatkan html berdasarkan input objek yang diingin kan user. Penguraian dilakukan berdasarkan *method* yang digunakan dalam penentuan setiap *tag html, body* dan head, dengan pembacaan "<" dan">". Pembacaan berdasarkan *tag html,* dimana pada halaman web yang berupa sekumpulan kode dibaca sesuai dengan penguraian *tag html, head* dan *body.*
Proses *parsing* merupakan proses kedua dalam melakukan *crawling* proses tersebut yaitu pemisahan kata untuk menentukan sebuah *element* yang diambil. Element yang diambil yaitu element "a href" untuk pengambilan *link*. Selanjutnya setelah ditentukan sebuah *element* untuk pemisahan, maka setiap pemisahan kata dilakukan berdasarkan *element* tersebut.

Proses pengambilan data pada kata yang telah dipisah berdasarkan *element* "a href":

<a href= "https://news.detik.com/berita/d-3379142/bnpb-banjir-kembali-terjadi-di-ntb">

Hasil proses *parsing:*

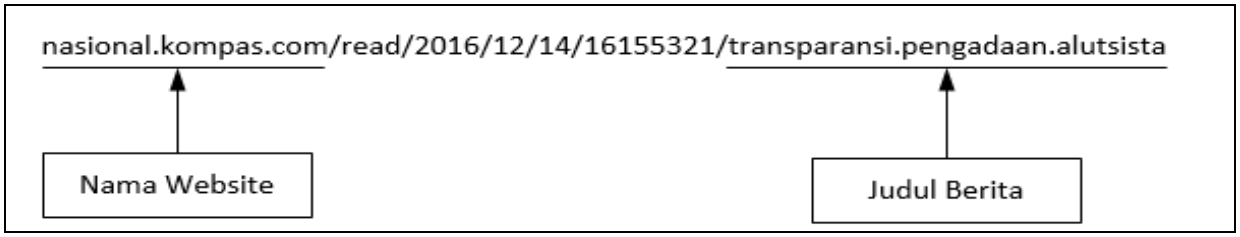
https://news.detik.com/berita/d-3379142/bnpb-banjir-kembali-terjadi-di-ntb

Proses *indexing* merupakan proses terakhir dalam melakukan *crawling*, dan proses dimana metode *breadth first search* dan *backlink* diterapkan. Pada *indexing* dilakukan *split* terhadap web yang sedang di *crawl, split* yang dilakukan berupa pemisahan tanda "/". Kemudian sistem melakukan pengecekkan kata yang telah di *split*. Pengecekkan kata dari tahap *splitting,* untuk menentukan judul berita, dan dilakukan pengelompokan berdasarkan alur metode *breadth first search* dan *backlink.*

Breadth first search crawling menguji tiap link pada sebuah halaman sebelum memproses ke halaman 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 [2]. Saat melakukan crawling menggunakan pengurutan berdasarkan *backlink*, maka pada saat mengunjungi halaman web, setiap *hyperlink* yang ada dilihat merujuk ke halaman web yang mana. Kalau halaman web yang dirujuk belum ada di list, maka halaman tersebut dimasukkan ke dalam list dan *backlink*nya diberi nilai 1[2].

Proses penerapan breadth first search dan backlink dalam indexing memerlukan kebutuhan data utnuk membantu saat proses *crawling* dengan *breadth first search* dan *backlink.* Kebutuhan data diperoleh dari *website* Kompas.com dan Detik.com, dengan didapatkan kesimpulan, yaitu :

1. Data kode judul, berupa inisialisasi kode yang digunakan untuk penentuan judul yang diambil dari kata atau bagian terakhir pada alamat web yang setiap kata atau bagiannya dipisah oleh "/" seperti yang ditampilkan pada Gambar 1.
2. Data *element* konten artikel yang ditampilkan pada Tabel 1, merupakan sebuah *element* yang digunakan untuk proses pengambilan artikel. Penentuan *element* ini ditentukan berdasarkan sebuah *element* yang menjadi wadah untuk konten artikel yang dimiliki oleh website yang menjadi objek utama.
3. Data Atribut Rel pada *backlink* yang ditampilkan pada Tabel 2, merupakan sebuah kode untuk mendapatkan URL *backlink,* penentuan atribut ini ditentukan berdasarkan sebuah *element* yang menjadi wada untuk penentuan URL yang dapat mengakses URL *root.*



Gambar 1/ Kode Judul Berita

Tabel 1. Data Element Konten Artikel

No	Alamat Web	*Element* Konten Artikel
1.	Kompas.com	div.kcm-read-text
2.	Detik.com	div.text_detail

Tabel 2 Data atribut backlink

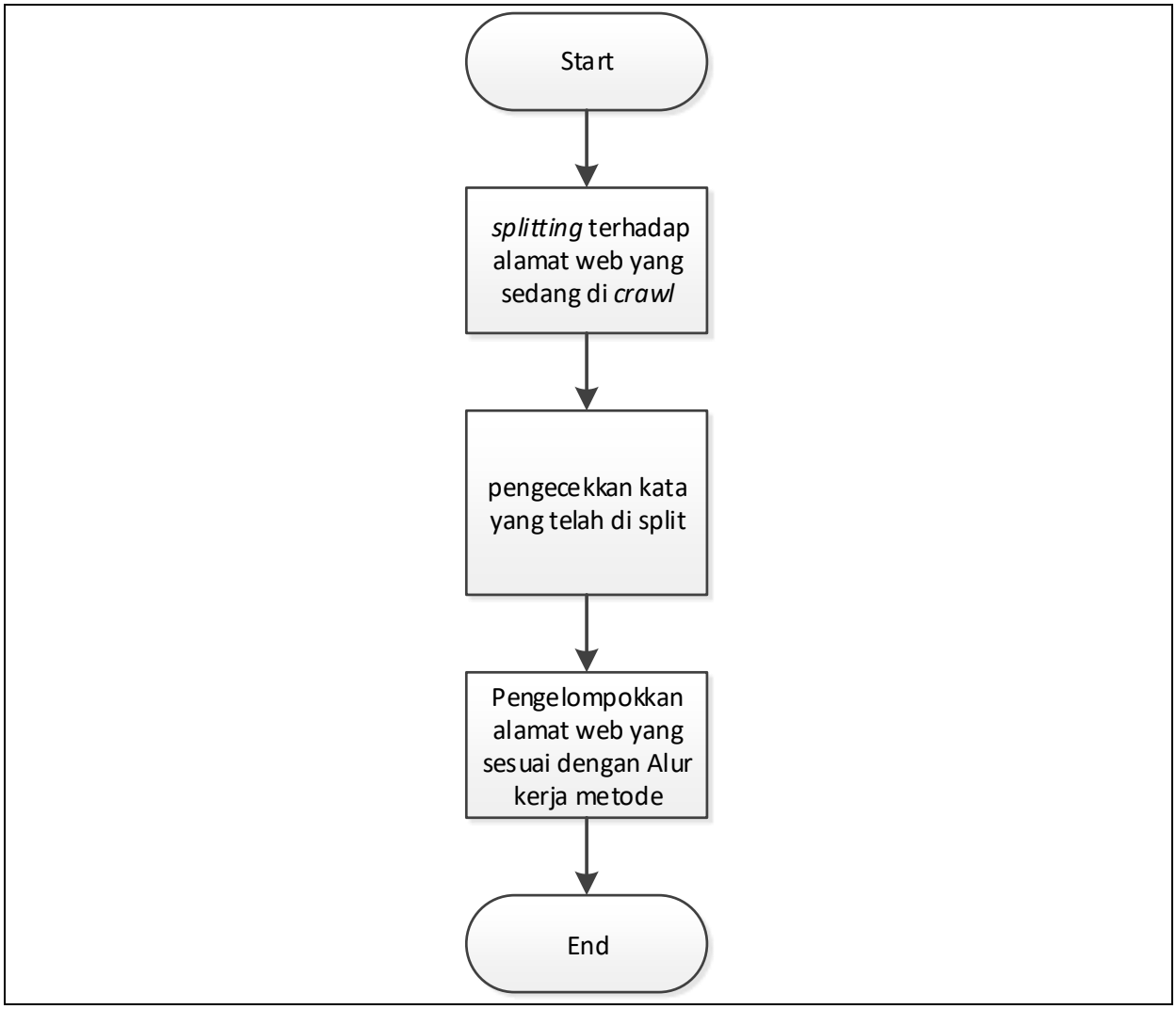
No	Alamat Web	Atribut *Backlink*
1.	Kompas.com	Rel =”dofollow”
2.	Detik.com	Rel =”dofollow”

3. ANALISIS DAN PEMBAHASAN

Indexing web crawler yang dibangun dilakukan dengan dua cara yaitu dengan metode *breadth first search* dan *backlink.* Pada metode *breadth first search* dan *backlink* dilakukan proses dimana sistem mengambil *element* dari URL root untuk pengambilan URL. Sistem mengambil semua atribut “href” untuk mendapatkan URL pada alamat *web* dan melakukan pengecekan *link*. Jika pada saat pengecekkan *link* terdapat duplikasi, maka link yang didapat akan dilewati oleh sistem dan melakukan kembali pengecekkan *link*. Jika pada alamat *web* tidak terdapat *link* lagi, sistem melanjutkan ke proses pengumpulan data berupa URL, judul berita dan konten artikel. Dan jika masih ada *link* yang harus dikunjungi, maka sistem melakukan proses *breadth first search* atau *backlink* kembali.

Pengecekan *link* berupa pencocokkan *link* yang sudah diambil dengan link *root*. Proses terakhir untuk *breadth first search* yaitu sistem menampilkan URL yang diambil berdasarkan urutan pengambilan dari alur *breadth first search.* Sedangkan proses terakhir untuk *backlink* yaitu sistem menampilkan URL backlink yang berupa URL yang mengandung rel "*dofollow*" yang berarti URL tersebut adalah URL *backlink.*

Pada alur kerja *web crawler* terdapat 3 proses tahapan, yaitu:
Proses *web crawler* ini yaitu proses *crawling* secara keseluruhan dari mulai sistem terkoneksi dengan *library* (*jsoup*) untuk memulai proses pembacaan parameter html pada halaman *web*, membaca halaman *web* semua parameter html yang ada pada suatu halaman *web*, dan pada proses terakhir melakukan proses *parsing* menggunakan metode *breadth first search* dan *backlink*, yang dimana hasilnya ditampilkan berupa URL, judul berita dan konten artikel.



Gambar 4. Flowchart Proses *Indexing*

Pada proses *indexing* dilakukan 3 tahapan, yaitu :

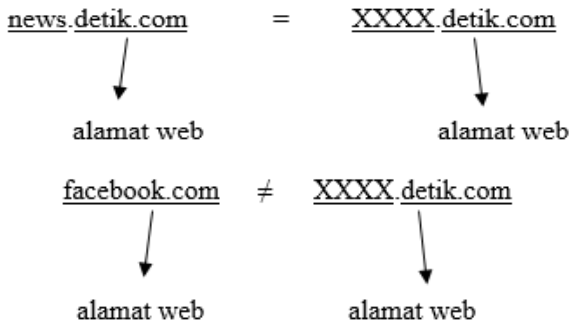
1. Melakukan *spliting* terhadap web yang sedang di *crawl, split* yang dilakukan berupa pemisahan tanda "/".

Hasil *parsing* berupa sekumpulan alamat web:

1. https://news.detik.com
2. https://finance.detik.com
3. https://hot.detik.com
4. https://inet.detik.com
5. https://facebook.com

2. Pengecekkan kata yang telah di *split*. Pada tahap ini, dilakukan pengecekkan kata dari tahap *splitting,* untuk menentukan judul berita, berdasarkan alur metode *breadth first search* dan *backlink.*

Tahap pengecekkan alamat web:



Gambar 5. pengecekksn alamat web

3. Tahap terakhir, dilakukan pengelompokkan hasil *crawl* yang telah diambil berdasarkan dengan alur setiap metode.

Tahap pengelompokkan alamat web:

1. https://news.detik.com
2. https://finance.detik.com
3. https://hot.detik.com
4. https://inet.detik.com
5. ~~https://facebook.com~~

Hasil *indexing* pada alamat web *breadth first search*:

1. https://news.detik.com
2. https://finance.detik.com
3. https://hot.detik.com
4. https://inet.detik.com

Pada proses *indexing* dengan menggunakan metode *breadth first search* yaitu pengelompokkan berdasarkan alur *breadth first search* yaitu pengambilan URL secara berurutan.

Tahap *parsing* alamat web dengan *backlink*:

1. https://news.detik.com *match* *dofollow*
2. https://finance.detik.com *match* *dofollow*
3. https://hot.detik.com *match* *dofollow*
4. https://inet.detik.com *match* *dofollow*
5. https://facebook.com *not match* *nofollow*

Tahap pengelompokkan alamat web:

1. https://news.detik.com *match* *dofollow*
2. https://finance.detik.com *match* *dofollow*
3. https://hot.detik.com *match* *dofollow*
4. https://inet.detik.com *match* *dofollow*
5. ~~https://facebook.com *not match* *nofollow*~~

Hasil *indexing* pada alamat web *backlink:*

1. https://news.detik.com *match* *dofollow*
2. https://finance.detik.com *match* *dofollow*
3. https://hot.detik.com *match* *dofollow*
4. https://inet.detik.com *match* *dofollow*

Hasil proses *indexing* dengan *breadth first search* dan *backlink* adalah hasil pengelompokkan berdasarkan alur metode *breadth first search* dan *backlink*. Pengelompokkan berdasarkan *breadth first search* yaitu berdasarkan URL yang pertama kali diambil oleh *web crawler* dan secara berurut sampai URL terakhir diambil*.* Pada *backlink* pengelompokkan berdasarkan URL yang mengandung rel "*dofollow*". Sehingga dapat dilihat perbedaan pengambilan pada kedua metode.

Skenario Pengujian

Pengujian yang dilakukan terdiri dari 2 aspek, yaitu pengujian unjuk kerja pengambilan URL dengan menggunakan metode *breadth first search* dan *backlink* pada *website* Detik.com dan Kompas.com dengan kecepatan internetnya 2,5Mbps kondisi tidak stabil. Dan pengujian duplikasi terhadap data yang sudah di *crawl*.

3.1 Pengujian Web Crawler Breadth First Search dan Backlink

Pada tahap ini, merupakan hasil pengujian yang dilakukan untuk melihat kinerja dari *web crawler*. Penilaian kinerja tersebut dilihat berdasarkan jumlah URL yang terambil yang dilakukan *web crawler* dalam satu kali proses.

a. Pengujian Performansi

Tabel 3. Perbandingan Breadth First Search dan Backlink pada Web Detik.com

Pengujian	Alamat web	BFS	Backlink	Pada Tanggal	Jam
		Banyaknya URL	Banyaknya URL		
1	Detik.com	1103 url	884 url	29/12/2017	08.00 WIB
2	Detik.com	982 url	637 url	29/12/2017	08.30 WIB
3	Detik.com	876 url	710 url	29/12/2017	09.15 WIB
4	Detik.com	1022 url	683 url	29/12/2017	09.30 WIB
5	Detik.com	994 url	721 url	29/12/2017	10.00 WIB
6	Detik.com	926 url	821 url	29/12/2017	15.00 WIB
7	Detik.com	899 url	728 url	29/12/2017	15.30 WIB
8	Detik.com	1023 url	637 url	29/12/2017	16.00 WIB
9	Detik.com	1078 url	790 url	29/12/2017	16.30 WIB
10	Detik.com	855 url	690 url	29/12/2017	17.00 WIB

Tabel 4. Perbandingan Breadth First Search dan Backlink pada Web Kompas.com

Pengujian	Alamat Web	BFS	Backlink	Pada Tanggal	Jam
		Banyaknya URL	Banyaknya URL		
1	Kompas.com	987 url	544 url	29/12/2017	18.00 WIB
2	Kompas.com	844 url	632 url	29/12/2017	18.30 WIB
3	Kompas.com	934 url	722 url	29/12/2017	19.15 WIB
4	Kompas.com	901 url	642 url	29/12/2017	19.30 WIB
5	Kompas.com	877 url	582 url	29/12/2017	20.00 WIB
6	Kompas.com	823 url	782 url	29/12/2017	20.30 WIB
7	Kompas.com	1015 url	689 url	29/12/2017	21.00 WIB
8	Kompas.com	966 url	772 url	29/12/2017	21.30 WIB
9	Kompas.com	1106 url	589 url	29/12/2017	22.00 WIB
10	Kompas.com	867 Url	668 url	29/12/2017	22.30 WIB

Tabel 5. Pengujian Pengulangan Data Crawl Kompas.com dan Detik.com

No	Detik.com	Kompas.com
1	Tidak ada pengulangan	Tidak ada pengulangan
2	Tidak ada pengulangan	Tidak ada pengulangan
3	Tidak ada pengulangan	Tidak ada pengulangan
4	Tidak ada pengulangan	Tidak ada pengulangan
5	Tidak ada pengulangan	Tidak ada pengulangan
6	Tidak ada pengulangan	Tidak ada pengulangan
7	Tidak ada pengulangan	Tidak ada pengulangan
8	Tidak ada pengulangan	Tidak ada pengulangan
9	Tidak ada pengulangan	Tidak ada pengulangan
10	Tidak ada pengulangan	Tidak ada pengulangan

4. KESIMPULAN

Berdasarkan pengujian yang dilakukan, didapatkan kesimpulan metode *breadth first search* secara performa lebih baik dibandingkan dengan metode *backlink* untuk diterapkan pada *web crawler.* Jumlah URL hasil *crawling* metode *breadth first search* pada *website* Detik.com lebih banyak sebesar 25,17% dari *backlink.* Sedangkan pada *website* Kompas.com metode *breadth first search* mendapatkan URL lebih banyak sebesar 28,94% dari *backlink, dan* pengujian pengulangan data aplikasi *web crawler breadth first search* dan *backlink* tidak terdapat pengulangan data.

DAFTAR RUJUKAN

[1] Yusuf, Muhammad. "Apa itu Web Crawler". Muhammad Yusuf Gunadarma. 2012. Web. 23 Oktober 2016.

[2] Sulastri dan Eri Zuliarso, 2010 Aplikasi *Web crawler* Berdasarkan *Breadth First Search* dan *Back-Link*

[3] Kustanto, Cynthia, Mutia S, Ratna, Viqarunnisa, Pocut, "Penerapan Algoritma Breadth-first Search dan Depth-first Search Pada FTP Search Engine for ITB Network", Teknik Informatika, Institut Teknologi Bandung, Bandung.

[4] Munir, Rinaldi. Strategi Algoritmik Diktat Kuliah IF2251. Program Studi Teknik Informatika, Sekolah Teknik Elektro dan Informatika, institut Teknologi Bandung, bandung. 2006.

[5] Maslucha Dewi, 2015 Makalah Algoritma Breadth First Search.

[6] Budi Yuwono, Savio L. Y. Lam, Jerry H.Ying, Dik L. Lee, 1996, *A World Wide Web Resource Discovery System*, in Proceedings of ICDE.

[7] Zebua, Javier., 2010, Aplikasi Pencarian Buku Berbasis Web Semantik Untuk Perpustakaan SMK Yadika 7 Bogor , UniversitasGunadarma.