

PENERAPAN *FOCUSED CRAWLING* PADA SITUS BERITA *ONLINE*

Aad Miqdad Muadz Muzad¹, Faisal Rahutomo², Imam Fahrur Rozi³

^{1,2,3}Teknik Informatika, Teknologi Informasi, Politeknik Negeri Malang

¹aadmiqdad@gmail.com, ²faisal.polinema@gmail.com, ³imam.rozi@polinema.ac.id

Abstrak

Dalam beberapa penelitian data berita sangat dibutuhkan sebagai objek penelitiannya. Data berita dapat diperoleh dengan mudah dari situs berita *online* dengan cara mengunjungi situs berita yang diinginkan dan mengambil berita tersebut satu per satu. Dikarenakan data berita yang dibutuhkan tidak sedikit jumlahnya maka cara tersebut tidak efektif dan efisien. Maka untuk memudahkan proses pencarian data berita dibutuhkan suatu program yang dapat melakukan pencarian data berita secara keseluruhan. *Web crawler* merupakan suatu proses yang digunakan dalam mesin pencarian atau *search engine* untuk menelusuri atau merayapi halaman dari suatu website guna mencari informasi yang diinginkan. Dengan menerapkan metode *Depth-first crawling* dan *Focused crawling*, *crawler* atau program *crawling* dapat melakukan penelusuran dari halaman ke halaman sesuai *node* yang sudah ditentukan dan dapat fokus menelusuri halaman yang merujuk pada konten berita. Hasil dari program *crawling* ini berupa data berita yang memiliki tipe data XML dan JSON. Selain itu penelitian ini juga menghasilkan data berita sejumlah 150.466 data berita. Data tersebut dapat digunakan dalam penelitian lain yang membutuhkan data berita sebagai objek penelitiannya.

Kata kunci : *Depth-first Crawling, Focused Crawling, Web Crawling*

1. Pendahuluan

Dalam beberapa penelitian suatu data berita sangat dibutuhkan. Contohnya dalam pembuatan ekstraksi berita untuk keleidoskop berita tahunan, analisis komputasi kemunculan dan kepunahan kosakata bahasa Indonesia berdasarkan *corpus* berita *online*, klasifikasi kategori berita, pembuatan aplikasi berita yang direkomendasikan, dan masih banyak lagi. Data berita akan menjadi sangat penting. Karena objek penelitian yang dipakai adalah data berita.

Munculnya situs berita *online* saat ini tentu sangat memudahkan untuk mendapatkan berita. Di situs berita *online* terdapat banyak sekali kategori berita yang diinginkan. Pengunjung tinggal memilih berita yang diinginkan dan mengambilnya. Akan tetapi banyaknya data berita yang dibutuhkan untuk suatu penelitian tentu sangat merepotkan apabila pengunjung harus memilih dan mengambil berita tersebut satu persatu.

Olah karena itu dibutuhkan suatu cara untuk memudahkan dalam pencarian data berita yang tentu tidak sedikit jumlahnya, yaitu dengan membangun mesin pencarian berita. Dengan cara menerapkan *focused crawling* pada situs berita *online* untuk mendapatkan data berita yang diinginkan. *Focused crawling* merupakan proses untuk melakukan penelusuran ke suatu *website* dengan berfokus pada halaman-halaman yang dirasa penting untuk ditelusuri.

2. Web Crawler

Web crawler juga sering disebut sebagai *Web Spider* atau *Web Robot*. *Web crawler* adalah suatu proses untuk melakukan *crawl* atau merayapi suatu laman atau halaman informasi dari suatu *website*. Tidak hanya merayapi, *web crawler* juga mengambil halaman informasi dari *website* tersebut.

Fungsi utama dari *web crawler* adalah untuk melakukan penelusuran atau merayapi informasi dari suatu *website*. Akan tetapi *web crawler* juga dapat berfungsi untuk proses memvalidasi kode *HTML* semua *website*, untuk mencari data khusus seperti mengumpulkan alamat *email* atau nomor telepon, untuk membuat salinan *website* yang diinginkan, dan masih banyak lagi fungsi lain dari *web crawler*.

Aplikasi untuk *web crawler* sebenarnya sudah banyak, akan tetapi inti dari *web crawler* secara keseluruhan sama, yaitu:

- Mengunduh halaman *website*.
- Mem-*parsing* halaman *website* tersebut dan menelusuri semua *link*.
- Untuk *link* yang sudah ditelusuri, ulangi proses di atas, yaitu men-*download* halaman dan mem-*parsing* halaman tersebut untuk menelusuri semua *link*. Kedalaman untuk menelusuri *link* dapat disesuaikan dengan kebutuhan.

2.1 Focused Crawling

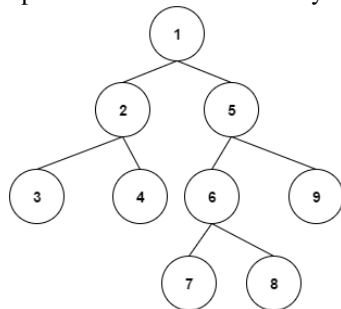
Dengan *focused crawling* memungkinkan *crawler* menyeleksi halaman-halaman web yang relevan dengan topik tertentu yang telah didefinisikan sebelumnya sehingga *crawler* tidak mencari seluruh web secara mendalam. *Focused crawling* memanfaatkan aturan-aturan keputusan berdasarkan pada analisis isi, struktur *link* dan teks, untuk menjaga agar *crawler* fokus pada topik tertentu.

Penerapan *focused crawling* digunakan agar proses penelusuran dapat terfokus. Artinya pada saat proses penelusuran program *crawling* hanya akan menelusuri *link* yang terkait dengan halaman konten berita. Sebagai contoh dalam suatu halaman terdapat banyak sekali *link*, seperti *link* untuk ke halaman home, *link* untuk pindah ke kategori lainnya, *link* konten berita, dsb. Dengan menggunakan *focused crawling*, *link-link* tersebut tidak akan ditelusuri semua, hanya *link* yang terkait halaman konten berita.

2.2 Depth-first Crawling

Sesuai penjelasan sebelumnya, bahwa dasar dari *web crawler* salah satunya adalah menelusuri semua *link* yang ada pada suatu *website*. Untuk dapat melakukan hal itu, maka dapat digunakan salah satu metode dalam kecerdasan buatan atau *artificial intelligence*. Yaitu *depth-first search*. Karena digunakan untuk proses *crawling*, maka diartikan sebagai *depth-first crawling*.

Depth-first crawling akan menelusuri dari *node* pertama kemudian dilanjutkan ke *node* paling kiri pada level berikutnya sampai tidak terdapat *node* paling dalam. Jika penelusuran sudah mencapai *node* paling dalam, maka akan dilakukan penelusuran mundur atau *backtracking* untuk melakukan pencarian ke *node* berikutnya.



Gambar 1 Depth-first Crawling

3. Implementasi Program Crawling

Situs berita *online* yang digunakan dalam implementasi program *crawling* ini adalah sebagai berikut:

- www.kompas.com
- www.tempo.co
- www.metrotvnews.com
- www.merdeka.com
- www.republika.co.id

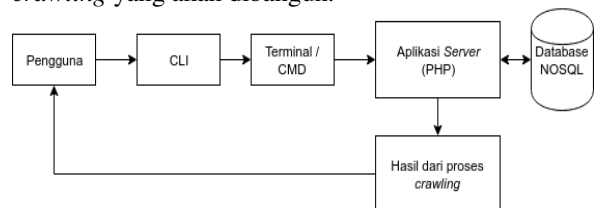
- www.viva.co.id
- www.tribunnews.com

Sedangkan untuk kategori berita yang digunakan adalah sebagai berikut:

- Teknologi
- Otomotif
- Bisnis ekonomi
- Nasional
- Olahraga
- Bola
- Lifestyle
- Travel

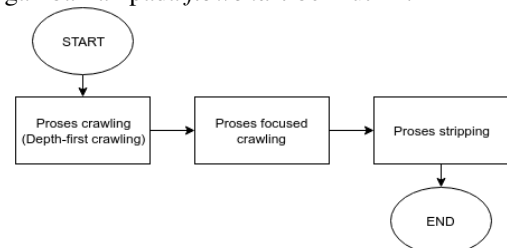
Berita yang akan dilakukan pencarian atau *di-crawling* adalah berita pada tahun 2015 (periode Juli sampai dengan Desember).

Program *crawling* ini nantinya akan berjalan menggunakan CLI (*Command Line Interface*). Hal ini dipilih karena menjalankan program melalui *website* dirasa memiliki kekurangan yaitu adanya *time out*, baik pada kode program itu sendiri, pada *web server* atau pada yang mengaksesnya. Sehingga tidak cocok menjalankan program yang harus aktif untuk jangka waktu yang relative lama. Berikut adalah gambaran umum dari program *crawling* yang akan dibangun.



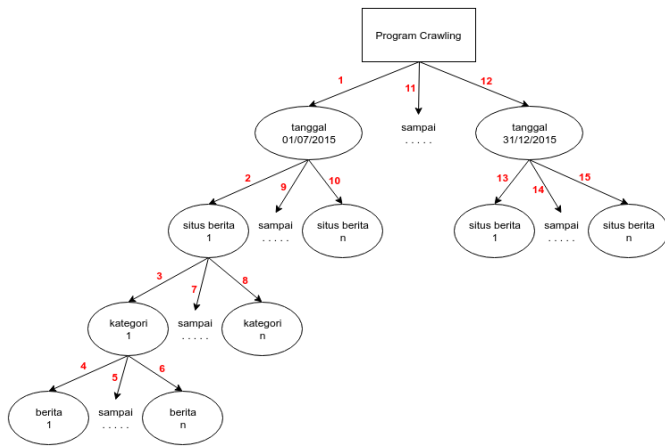
Gambar 2 Gambaran Umum Program Crawling

Untuk dapat melakukan proses *crawling* dibutuhkan tiga fungsi dasar dalam program *crawling*, berikut tiga fungsi dasar yang digambarkan pada *flowchart* berikut ini.

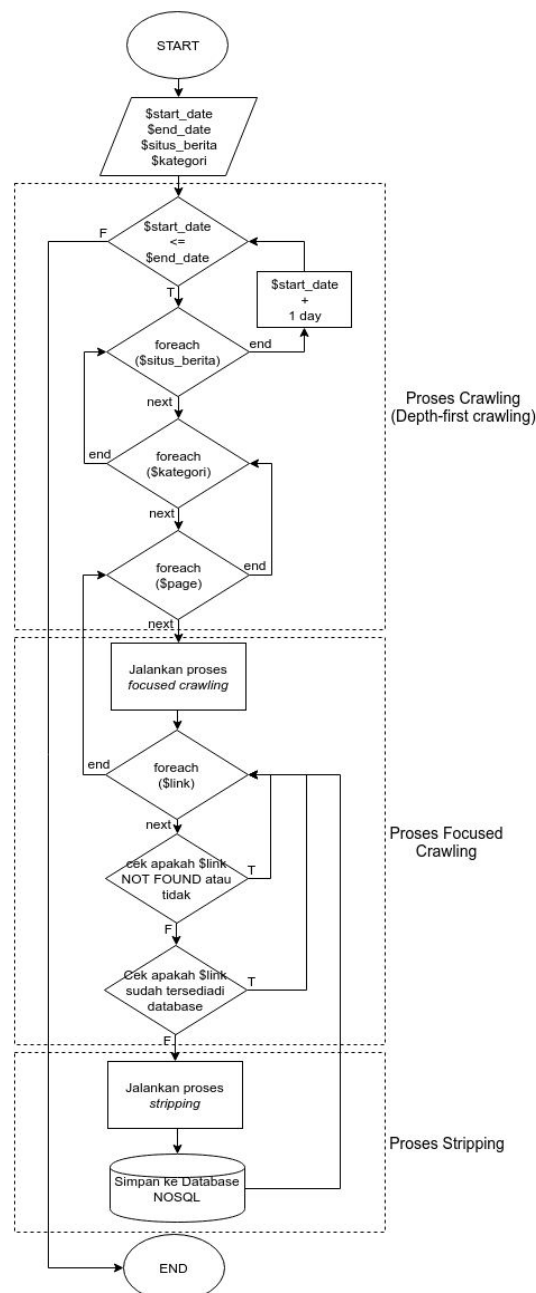


Gambar 3 Flowchart dari Proses Dasar Program Crawling

Metode *depth-first crawling* digunakan dalam proses penelusuran dari *node* sampai ke *node* akhir (halaman berita). Mulai dari *node* *root* (awal program), turun ke *node* tanggal berita yang akan ditelusuri, kemudian ke situs berita apa yang akan ditelusuri, setelah itu ke kategori berita apa yang akan ditelusuri, dan akhirnya sampai ke *node* akhir atau halaman dari konten berita yang dicari. Untuk lebih jelasnya dapat dilihat pada Gambar 4.



Gambar 4 Proses Crawling Menggunakan Metode Depth-first Crawling

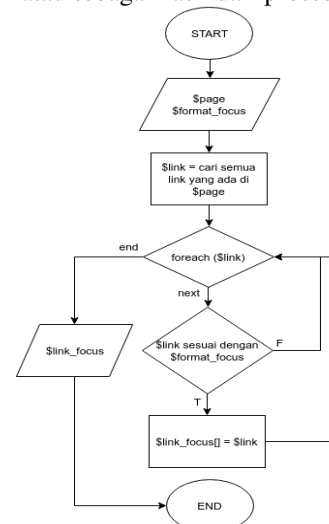


Gambar 5 Flowchart dari Program Crawling

Flowchart pada Gambar 5 menjelaskan secara detail proses *crawling*. Dimana pertama harus memasukkan tanggal mulai, tanggal selesai, situs berita *online*, dan kategori. Kemudian masuk ke proses perulangan dari tanggal awal sampai tanggal selesai. Kemudian masuk ke perulangan situs berita *online* dan perulangan kategori. Karena situs berita *online* terdapat halaman (*page*) maka dibutuhkan perulangan *page*. Barulah masuk ke proses *focused crawling* dimana flowchart detailnya terdapat pada Gambar 6.

Hasil dari *focused crawling* tersebut dimasukkan ke proses perulangan untuk membaca tiap *link* hasil dari *focused crawling*. Kemudian *link* tersebut dicek apakah tersedia atau tidak dan juga dicek apakah *link* tersebut sudah ada di *database* atau belum. Barulah masuk ke proses pembersihan kode HTML atau *stripping*. Detail flowchart dari proses *stripping* terdapat pada Gambar 7.

Hasil dari proses *stripping* tersebutlah yang akan disimpan ke *database*. Dan data inilah yang digunakan atau sebagai hasil dari proses *crawling*.

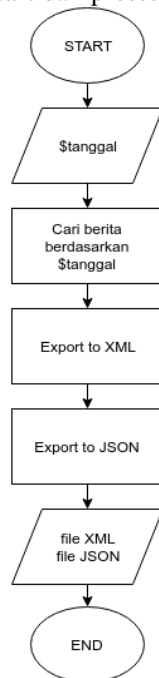


Gambar 6 Flowchart dari Proses Focused Crawling



Gambar 7 Flowchart dari Proses Stripping

Sedangkan untuk proses *export* yang digunakan untuk meng-*export* data berita yang sudah di-*crawling* berupa data XML dan JSON agar dapat diolah untuk kebutuhan lain. Masukan dari proses ini adalah tanggal dari berita yang akan di-*export* kemudian masuk ke proses *export to XML* dan juga proses *export to JSON*. Gambar 8 merupakan *flowchart* dari proses *export*.



Gambar 8 *Flowchart* dari Proses *Export*

Sesuai dengan yang telah ditentukan sebelumnya, program *crawling* yang dibuat menggunakan CLI. Adapun antarmuka dari program *crawling* dapat dilihat pada Gambar 9 dan untuk antarmuka program *export* terdapat pada Gambar 10.

```

amiq@amiq-pc: /var/www/html/SKRIPSI/v.3.3.0
amiq@amiq-pc:~$ cd /var/www/html/SKRIPSI/v.3.3.0
amiq@amiq-pc:/var/www/html/SKRIPSI/v.3.3.0$ php app.php
#####
# Welcome to Crawling App, please choice menu #
# aadmiqdad@gmail.com #
#####
Enter 'w' to choice webs
Enter 'c' to choice categories
Enter 'ds' to set start date (yyyy/mm/dd)
Enter 'de' to set end date (yyyy/mm/dd)
Enter 'r' to run
Enter 'x' to reset
Enter 'q' to quit
Enter 'h' to help
  
```

Gambar 9 Antarmuka Program *Crawling*

```

amiq@amiq-pc: /var/www/html/SKRIPSI/v.3.3.0
amiq@amiq-pc:~$ cd /var/www/html/SKRIPSI/v.3.3.0
amiq@amiq-pc:/var/www/html/SKRIPSI/v.3.3.0$ php export.php
#####
# Welcome to Crawling App | Export, please choice menu #
# aadmiqdad@gmail.com #
#####
Enter 'd' to set date (yyyy/mm/dd)
Enter 'r' to run
Enter 'x' to reset
Enter 'q' to quit
Enter 'h' to help
  
```

Gambar 10 Antarmuka Program *Export*

```

#####
# 2016/01/01 #
#####
2016/01/01, Kompas.com, Teknologi | PAGE 1
[CRAWLING] http://tekno.kompas.com/read/xml/2016/01/01/17410727/Ini.Kata.Pelaku
[STRIPPING] [SUCCESS]
[ADDED TO DB] [SUCCESS]
[CRAWLING] http://tekno.kompas.com/read/xml/2016/01/01/12454607/Burung.Menunggu
Telur.di.Doodle.Tahun.Baru.Google [SUCCESS]
[STRIPPING] [SUCCESS]
[ADDED TO DB] [SUCCESS]
[CRAWLING] http://tekno.kompas.com/read/xml/2016/01/01/11200007/Apple.Bakal.Bel
ajar.dari.Kesalahan.iPhone. [SUCCESS]
[STRIPPING] [SUCCESS]
[ADDED TO DB] [SUCCESS]
[CRAWLING] http://tekno.kompas.com/read/xml/2016/01/01/08085047/Malam.Tahun.Bar
u.Layanan.WhatsApp.dilaporkan.Tumbang [SUCCESS]
[STRIPPING] [SUCCESS]
[ADDED TO DB] [SUCCESS]
#####
# FINISH #
#####
  
```

Gambar 11 Hasil dari Program *Crawling* Ketika Dijalankan Untuk Proses *Crawling*

Gambar 11 merupakan hasil ketika program *crawling* melakukan proses *crawling*. Penulis melakukan percobaan dengan data *input*-an yaitu, situs berita *online* *kompas.com*, kategori berita teknologi, tanggal mulai 2016/01/01, dan tanggal akhir 2016/01/01.

Program *crawling* menampilkan *link* yang di-*crawling* dan menampilkan status dari proses *crawling*, apakah berhasil atau gagal. Program juga menampilkan status dari proses *stripping* dan memasukkan data kedalam *database*.

4. Hasil Pengujian

Berdasarkan pada hasil pengujian validasi dan pengujian performa yang telah dilakukan oleh penulis, maka proses *crawling* yang dapat dan memungkinkan untuk dilakukan (sebelum bulan Mei 2016) selama enam bulan (dari bulan Juli 2015 sampai dengan bulan Desember 2015) adalah sebagai berikut:

- Kompas.com, kategori Teknologi, Bisnis Ekonomi, Nasional, Olahraga, dan *Travel*.
- Republika.co.id, kategori Teknologi, Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola, *Lifestyle*, dan *Travel*.
- Viva.co.id, kategori Teknologi, Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola, *Lifestyle*, dan *Travel*.
- Tribunnews.com, kategori Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola, *Lifestyle*, dan *Travel*.

Situs berita di atas memungkinkan digunakan karena semua fungsi berjalan dengan baik dan rata-rata waktu yang dibutuhkan untuk proses *crawling* adalah 1detik/berita dan tidak sampai melebihi 2 detik/berita.

Sedangkan situs berita yang tidak dapat dan tidak dimungkinkan untuk dilakukan proses *crawling* adalah sebagai berikut:

- Kompas.com, kategori Otomotif dan Bola. Karena *tag* HTML pada halaman konten berita berbeda dengan *tag* HTML pada kategori lainnya.
- Tempo.co, karena waktu yang dibutuhkan untuk proses *crawling* cukup lama (3,689111364 detik/berita) dan pada saat

proses *crawling* program sering terhenti karena tidak mendapat respon dari *server*.

- Metrotvnews.com, karena tampilan *website* sudah mengalami pembaharuan desain (*redesign*).
- Merdeka.com, karena waktu yang dibutuhkan untuk proses *crawling* sangat lama sekali (7,183326796 detik/berita).

5. Kesimpulan

Dari penelitian ini dan dari hasil program *crawling* menyimpulkan bahwa untuk dapat membuat mesin pencarian berita dengan proses *crawling* dapat menggunakan metode *focused crawling* dan *depth-first crawling*. Karena kedua metode tersebut sangat cocok digunakan dalam proses *crawling*.

Sedangkan untuk situs berita dan kategori yang memungkinkan dan dapat digunakan dalam proses *crawling* beserta jumlah berita yang didapat adalah sebagai berikut:

- Kompas.com, kategori Teknologi, Bisnis Ekonomi, Nasional, Olahraga, dan *Travel*. Terkumpul 21.809 data berita.
- Republika.co.id, kategori Teknologi, Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola, *Lifestyle*, dan *Travel*. Terkumpul 57.678 data berita.
- Viva.co.id, kategori Teknologi, Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola, *Lifestyle*, dan *Travel*. Terkumpul 37.108 data berita.
- Tribunnews.com, kategori Otomotif, Bisnis Ekonomi, Nasional, Olahraga, Bola,

Lifestyle, dan *Travel*. Terkumpul 33.871 data berita.

Total Berita yang didapat dari hasil penelitian ini adalah 150.466 data berita.

Daftar Pustaka:

- Astuti, K. D., dkk., 2009. *Analisis Performansi Web Crawler Berbasis Backlink, Breadth First Search, dan Pagerank*. Bandung: Universitas Telkom
- Brooks, C. H., *Web Crawling as an AI Project*. US: Department of Computer Science, University of San Francisco
- Budhi, G. S., dkk., 2006. *Aplikasi Web Grabber Untu Mengambil Halaman Web Sesuai Dengan Keyword Yang Diinputkan*. Surabaya: Universitas Kristen Petra
- Kumar, R., dkk., 2014. *Survey Of Web Crawling Algorithms*. India: RGPVUniversity Bhopal
- Nigam, A., 2014. *Web Crawling Algorithms*. India: National Institute of Technology
- Pavalam, S. M., dkk., 2011. *A Survey of Web Crawler Algorithms*. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1
- Rosmala, D. dan Syafei, R. R., 2011. *Implementasi Webcrawler Pada Social Media Monitoring*. Bandung: Institut Teknologi Nasional Bandung
- Sarwosri, dkk., 2009. *Aplikasi Web Crawler Untuk Web Content Pada Mobile Phone*. Surabaya: Institut Teknologi Sepuluh Nopember
- Zuliarso, E., & Mustofa, K., 2009. *Crawling Web Berdasarkan Ontology*. Yogyakarta: Universitas Gadjah Mada