SKRIPSI

FOCUSED WEB CRAWLING PADA LINGKUNGAN HADOOP

JOVANKA HELEN MARADENIA

NPM: 2014730029

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2018

UNDERGRADUATE THESIS

FOCUSED WEB CRAWLING ON HADOOP ENVIRONMENT

JOVANKA HELEN MARADENIA

NPM: 2014730029

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2018

LEMBAR PENGESAHAN

FOCUSED WEB CRAWLING PADA LINGKUNGAN HADOOP

JOVANKA HELEN MARADENIA

NPM: 2014730029

Bandung, 3 Juli 2018

Menyetujui,

Pembimbing

Gede Karya, M.T., CISA, IPM

Ketua Tim Penguji

Luciana Abednego, M.T.

Anggota Tim Penguji

Chandra Wijaya, M.T.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

***FOCUSED WEB CRAWLING* PADA LINGKUNGAN HADOOP**

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 3 Juli 2018

JOVANKA HELEN MARADENIA
NPM: 2014730029

ABSTRAK

World Wide Web merupakan sebuah wadah informasi yang sangat luas dan dapat diakses dimana saja. Penggunaan umum dari *web crawler* yaitu terkait *search engine*. *Search engine* menggunakan *web crawler* untuk mengumpulkan informasi apa saja yang ada pada halaman *web* publik di internet dan melihat *trend* pasar atas dasar pencarian kata. Ketika pengguna internet mengetik topik pencarian pada *search engine*, *search engine* dapat mengembalikan halaman *web* yang relevan. *Focused web crawler* merupakan *web crawler* yang memiliki tugas untuk menyimpan halaman yang relevan dengan topik pencarian. *Focused web crawler* menghitung nilai relevansi antara halaman *web* dan topik yang pengguna ingin cari. Sehingga ketika pengguna ingin mencari sebuah topik, *web crawler* akan mengembalikan halaman *web* yang relevan dengan apa yang pengguna inginkan.

Jumlah halaman *web* yang disimpan pada saat *crawling* dapat berukuran sangat besar. Untuk itu diperlukan wadah penyimpanan yang besar. Hadoop merupakan sebuah *framework* yang menangani data yang berukuran raksasa. Komponen utama dari Hadoop yaitu *Hadoop Distributed File System* (HDFS), yang dapat menghubungkan beberapa komputer agar dapat saling bekerja sama dalam menyimpan dan mengolah suatu data. Sehingga jika ada salah satu komputer yang mati, data tetap terjaga karena HDFS membuat replika data pada masing-masing komputer. Hadoop sendiri memiliki HBase(Hadoop Database) yang berbasis NoSQL(*Not Only SQL*). Hasil *crawling* kemudian disimpan pada tabel penyimpanan HBase yang berjalan pada Hadoop.

Tahapan *crawling* dimulai dengan penelusuran halaman *web* menggunakan algoritma pencarian *Breadth-First Search* (BFS). BFS dimulai dari akar (atau halaman *web*) dan mengunjungi *node* tetangga terlebih dahulu sebelum mengunjungi *node* tetangga pada level berikutnya. Sedangkan untuk menghitung nilai relevansi menggunakan *Vector Space Model*. *Vector Space Model* merepresentasikan halaman *web* dan topik sebagai vektor, yang kemudian jarak antar vektor tersebut disimpan sebagai nilai relevansi. Untuk meningkatkan nilai relevansi maka digunakan algoritma *stemming* untuk menghapus imbuhan kata pada bahasa Indonesia, meskipun masih banyak kata dalam bahasa Indonesia yang tidak dapat di-*stemming* menggunakan aturan *stemming* Bahasa Indonesia. Perangkat lunak yang dibangun yaitu antarmuka berupa situs induk dan agen *crawler*. Situs induk dibangun agar pengguna dapat memasukkan informasi *url* yang ingin di-*crawl* dan dapat memantau hasil dan status *url* tersebut. Sedangkan agen *crawler* bertugas untuk menelusuri *url* yang disimpan sebelumnya oleh pengguna.

Pengujian fungsional perangkat lunak dilakukan untuk menguji fitur-fitur pada aplikasi situs induk dan agen *crawler*. Berdasarkan hasil pengujian fungsional, dapat disimpulkan bahwa aplikasi berhasil dibangun dan seluruh fungsi dapat berjalan dengan baik. Pengujian eksperimen dilakukan untuk melihat performa agen *crawler* pada saat dijalankan pada lingkungan terdistribusi Hadoop. Pengujian dilakukan dengan menggunakan empat komputer. Dari hasil pengujian yang didapatkan semakin banyak komputer dan agen *crawler* dipakai, maka *crawling* semakin cepat. Semakin banyak komputer yang aktif, waktu pencarian semakin kecil.

Kata-kata kunci: *Focused Web Crawler*, *Breadth-First Search*, *Vector Space Model*, *Stemming*, Data Besar, HBase, Hadoop

ABSTRACT

World Wide Web is a very wide and accessible information container. Common use of web crawlers is related to search engines. Search engines use web crawlers to collect information on a public web page on the internet and see market trends based on word search. When an internet user types a search topic on a search engine, the search engine can return the relevant web pages. Focused web crawlers are web crawlers that have the task of storing pages that are relevant to the search topic. Focused web crawlers calculate the value of the relevance between the web page and the topic that the user wants to search. So when a user wants to search a topic, the web crawler will return a web page that relevant to what the user wants.

The number of web pages saved at crawling can be very large. For that we need a large storage container. Hadoop is a framework that handles massive data. The main component of Hadoop is HDFS (Hadoop Distributed File System), which can connect multiple computers to work together to store and process data. So if one of the computer is dead, the data is maintained because HDFS makes replicas of data on each computer. Hadoop itself has HBase (Hadoop Database) with NoSQL (Not Only SQL) based. The crawling results are then stored on HBase table that runs on Hadoop.

The crawling stages begin with a web page search using the Breadth-First Search (BFS) algorithm. It starts at the tree root (or web page) and explores the neighbor nodes first, before moving to the next level neighbors. While to calculate the value of relevance is using Vector Space Model. Vector Space Model represents the web page and topic as vectors, where the vector distance is then stored as a value of relevance. To increase the value of relevance, then used the stemming algorithm to remove the word idioms in Indonesian language, although there are still many words in Indonesian that can not be stem using the rules of Indonesian stemming. The result of software development is website interface and the crawler agent. The website is built so that users can enter the url information they want to crawl and can monitor the results and status of the url. While the crawler agent task is to crawl the url that was previously stored by the user.

Functional software testing is performed to test the features on the website and crawler agent application. Based on the results of functional testing, it can be concluded that the application successfully built and all functions can run well. Experimental tests are performed to see the performance of crawler agents when run on Hadoop distributed environment. The test was performed using four computers. From the test results obtained when more computer and crawler agents added, then the crawling is much more faster. And the more computer is active, the smaller the search time.

Keywords: Focused Web Crawler, Breadth-First Search, Vector Space Model, Stemming, Big Data, HBase, Hadoop

Dipersembahkan kepada diri sendiri dan Orangtua penulis

KATA PENGANTAR

Puji dan syukur penulis panjatkan kepada Tuhan Yesus Kristus atas segala rahmat dan berkatNya penulis dapat menyelesaikan skripsi yang berjudul "*Focused Web Crawling* pada Lingkungan Hadoop". Dalam proses penyusunan skripsi, penulis banyak mendapat kesempatan untuk menambah ilmu dengan mempelajari hal-hal baru, serta mendapatkan banyak bantuan baik secara langsung maupun tidak langsung dari berbagai pihak. Oleh karena itu, pada kesempatan ini penulis ingin mengucapkan terima kasih kepada:

1. Orangtua penulis yaitu Leo Lubis dan Julis Hutagalung, serta adik-adik penulis yaitu Jessica Lubis dan Jennifer Lubis yang selalu memberikan dukungan secara jasmani maupun rohani.
2. Bapak Gede Karya, M.T., CISA, IPM selaku dosen pembimbing yang telah memberi arahan dan masukan selama penyusunan skripsi.
3. Ibu Luciana Abednego, M.T. dan Bapak Chandra Wijaya, M.T. selaku dosen penguji yang telah memberikan kritik dan saran untuk skripsi.
4. Sahabat penulis yang sudah menemani dan membantu perkuliahan selama empat tahun yaitu Devi Siman dan Carissa Tobing.
5. Rekan-rekan kuliah yang telah memberikan bantuan dalam penyusunan skripsi yaitu Melinda Abianti dan Kevin Pratama.

Akhir kata, penulis menyadari bahwa skripsi ini tidak lepas dari kekurangan. Namun penulis berharap skripsi ini dapat memberikan kontribusi baik untuk penelitian atau pembelajaran selanjutnya.

Bandung, Juli 2018

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

Pada bab ini akan dibahas mengenai latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

1.1 Latar Belakang

World Wide Web merupakan kumpulan web yang menyediakan berbagai macam informasi. Jumlah web saat ini sangatlah banyak dan akan semakin bertambah. *Web crawler* (atau dikenal juga sebagai *web spider, web robot, bot, crawl* atau *automatic indexer*), merupakan perangkat lunak yang dengan menggunakan metode tertentu melakukan penjelajahan *web* dan membuat indeks dari data yang dicari. Penggunaan umum dari *web crawler* yaitu terkait dalam *search engine*. *Search engine* menggunakan *web crawler* untuk mengumpulkan informasi apa saja yang ada pada halaman *web* publik di internet. Tugas *web crawler* yaitu mengunjungi halaman *web* beserta *link* didalamnya dan menyimpan isinya.

Tujuan utama dari *web crawler* adalah mengumpulkan data. Ketika pengguna internet mengetikkan kata pencarian pada *search engine*, *search engine* dapat mengembalikan halaman *web* yang relevan. Dengan jumlah halaman internet yang sangat banyak dan akan terus bertambah, sulit untuk menemukan halaman internet yang relevan dengan apa yang pengguna inginkan. Ada berbagai metode *crawling*, yaitu *incremental crawling*, *deep crawling*, *distributed crawling*, *focused crawling*, dan lain-lain.

Salah satu masalah dalam *crawling* adalah mencari isi yang sesuai dengan kriteria tertentu (*focused*). *Focused web crawling* digunakan untuk mencari halaman *web* yang paling relevan dengan apa yang pengguna ingin cari. *Focused Crawler* hanya mengunduh halaman web yang relevan dengan topik yang ingin dicari dan menghindari mengunduh halaman yang tidak ada kaitannya dengan topik tersebut. Halaman yang relevan disimpan pada wadah pengimpanan dan halaman yang tidak relevan dibuang. Ketika pengguna ingin mencari sebuah topik, *web crawler* akan mengembalikan halaman *web* yang dinilai relevan dengan apa yang pengguna inginkan.

Hadoop merupakan sebuah *framework* yang menangani data yang banya dan berukuran sangat besar. Hadoop juga menghubungkan beberapa komputer agar dapat saling bekerja sama dalam menyimpan dan mengolah suatu data, sehingga jika ada salah satu komputer yang mati, maka data tetap dapat diakses. Hal ini disebut dengan HDFS (*Hadoop Distributed File System*). HDFS memiliki *database* yang berjalan diatasnya yaitu HBase. HBase merupakan *database* yang berbasis NoSQL(*Not Only SQL*) dan berorientasi kolom yang dapat memproses data dengan skala yang besar secara interaktif. Hasil yang didapatkan melalui metode *focused web crawling* akan disimpan ke dalam tabel HBase yang berjalan diatas Hadoop.

Jumlah halaman *web* yang disimpan pada saat *crawling* dapat berjumlah sangat banyak. Untuk itu diperlukan wadah penyimpanan yang besar. Oleh karena itu, pada skripsi ini dikembangkan *focused web crawler* pada lingkungan Hadoop.

1.2 Rumusan Masalah

Berdasarkan latar belakang terdapat rumusan masalah yang akan dikaji seperti berikut:

1. Bagaimana menentukan kriteria pencarian sehingga proses *crawling* menjadi lebih fokus?
2. Bagaimana mekanisme penyimpanan informasi hasil *crawling* yang telah dikelompokkan berdasarkan kriteria fokus pada lingkungan *file* sistem terdistribusi Hadoop?

1.3 Tujuan

Berdasarkan rumusan masalah di atas, terdapat tujuan yang akan dicapai sebagai berikut:

1. Menentukan kriteria pencarian sehingga proses *crawling* menjadi lebih fokus.
2. Memahami mekanisme penyimpanan informasi hasil *crawling* yang telah dikelompokkan berdasarkan kriteria fokus pada lingkungan *file* sistem terdistribusi Hadoop.

1.4 Batasan Masalah

Untuk memenuhi tujuan yang disampaikan sebelumnya, maka akan dirancang dan diimplementasikan hal-hal berikut:

1. Halaman web yang akan di-*crawling* menyimpan konten berupa teks berbahasa Indonesia. Pemilihan bahasa Indonesia perlu dibatasi karena analisis konten yang memerlukan algoritma *stemming* bergantung pada bahasanya.
2. Topik masukan untuk proses *crawling* tidak memiliki duplikasi kata. Hal tersebut dilakukan agar tidak mengganggu nilai vektor topik yang kemudian digunakan untuk perhitungan nilai relevansi menggunakan *Vector Space Model*.

1.5 Metodologi

Berikut langkah-langkah yang dilakukan dalam pembuatan skripsi:

1. Studi pustaka mengenai *focused web crawling*, lingkungan *big data* Hadoop, dan J2EE.
2. Menganalisis algoritma pencarian halaman *web* menggunakan *Breadth-First Search*.
3. Menganilisis algoritma perhitungan nilai relevansi menggunakan *Vector Space Model* dan *Porter Stemming*.
4. Merancang kebutuhan penyimpanan data hasil *crawling* pada lingkungan Hadoop.
5. Merancang fitur dan antarmuka perangkat lunak.
6. Mengimplementasikan penyimpanan data hasil *crawling* dan sistem perangkat lunak *focused web crawling*, yang terdiri atas situs induk dan agen *crawler*.
7. Menguji sistem perangkat lunak *focused web crawling* baik fungsional maupun performa pada lingkungan terdistribusi Hadoop.

1.6 Sistematika Pembahasan

Sistematika penulisan dalam skripsi ini adalah sebagai berikut:

- Bab 1 Pendahuluan

  Bab ini membahas tentang latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan.

- Bab 2 Landasan Teori

  Bab ini membahas tentang teori-teori yang mendukung pembuatan skripsi yang telah dilakukan mengenai *web crawler*, *focused web crawler*, *Vector Space Model*, *Porter Stemming*, Hadoop, dan J2EE.

- Bab 3 Analisis

  Bab ini membahas tentang deskripsi masalah, analisis masalah pada proses *crawling*, perhitungan nilai relevansi menggunakan *Vector Space Model* dan usulan solusinya. Analisis kebutuhan perangkat lunak yang meliputi diagram *use case* untuk kebutuhan fungsional, tabel penyimpanan, dan diagram kelas sederhana untuk menganalisis kelas-kelas yang akan dibuat.

- Bab 4 Perancangan

  Bab ini membahas tentang rancangan antarmuka, rancangan penyimpanan tabel pada basis data, diagram kelas rinci dari perangkat lunak yang akan dibangun dan *pseudocode* dari fungsi yang diterapkan pada agen *crawler*.

- Bab 5 Implementasi dan Pengujian Perangkat Lunak

  Bab ini membahas tentang lingkungan implementasi, implementasi perangkat lunak, implementasi basis data, pengujian fungsional terhadap perangkat lunak, pengujian eksperimen terhadap perangkat lunak yang berjalan diatas HDFS (*Hadoop Distributed File System*), dan kesimpulan hasil pengujian.

- Bab 6 Kesimpulan dan Saran

  Bab ini membahas tentang kesimpulan akhir dari penelitian dan saran untuk melakukan pengembangan penelitian lebih lanjut.