

WEB CRAWLER MULTI-THREAD TERDISTRIBUSI UNTUK PENGHIMPUNAN DAN EKSTRAKSI HALAMAN WEB BERITA OLAHRAGA BERGANTUNG STRUKTUR

Muhammad Andri Rahmawanto

(13.04.1.1.1.00056)

Pembimbing 1 : Sigit Susanto Putro, S. Kom., M. Kom.

Pembimbing 2 : Yoga Dwitya Pramudita, S. Kom., M. Cs.

ABSTRAK

Web crawler merupakan program yang digunakan suatu *search engine* untuk menghimpun informasi yang diperlukan dari internet secara otomatis sesuai dengan aturan yang telah ditetapkan oleh pengguna. *Web crawler* umumnya menerapkan teknik pemrograman *single-thread* dan belum menerapkan deteksi pembaruan *web*, jadi kecepatan dalam proses *crawling* belum maksimal. Sedangkan pada saat ini teknologi informasi berkembang pesat, maka perlu adanya suatu *web crawler* yang memiliki kecepatan tinggi dalam proses *crawling* dan menyediakan informasi yang terbaru untuk memenuhi kebutuhan informasi penggunanya.

Penelitian ini menerapkan teknik pemrograman *multi-thread* terdistribusi dalam pengembangan *web crawler* untuk menghasilkan *web crawler* yang lebih cepat dalam proses *crawling* berita olahraga. Selain itu, penelitian ini juga menerapkan deteksi pembaruan halaman *website* untuk mendapatkan informasi terbaru. Proses *crawling* atau ekstraksi isi halaman sangat bergantung dengan struktur dari *website* tersebut.

Hasil dari penerapan konsep pemrograman *multi-thread* dapat meningkatkan kecepatan proses *crawling* dibandingkan *single-thread* sebesar 122,95 detik. Kemudian penerapan deteksi pembaruan *web* kecepatan proses *crawling* menurun yakni 6,27 detik. Tetapi keunggulan penerapan deteksi pembaruan *web* data yang tersimpan tidak sama. Begitupun pada penerapan distribusi *database*, kecepatan proses *crawling* mengalami penurunan sebesar 24,76 detik pada server 1 dan 23,92 detik pada server 2.

Kata kunci: *web crawler*, *multi-thread*, deteksi pembaruan isi halaman *website*, sistem terdistribusi.

KATA PENGANTAR



Alhamdulillah segala puji syukur kami panjatkan kehadirat Allah SWT, karena atas limpahan rahmat dan hidayah-Nya penulis dapat menyelesaikan skripsi yang berjudul:

WEB CRAWLER MULTI-THREAD TERDISTRIBUSI UNTUK PENGHIMPUNAN DAN EKSTRAKSI HALAMAN WEB BERITA OLAHRAGA BERGANTUNG STRUKTUR

Skripsi ini disusun sebagai persyaratan kelulusan pada Program Studi Strata-1 Teknik Informatika Fakultas Teknik Universitas Trunojoyo Madura.

Dalam penyusunan skripsi ini penulis menyadari telah mendapat banyak saran, petunjuk, bimbingan serta dorongan dari berbagai pihak yang merupakan pengalaman berharga bagi penulis. Oleh karena itu dengan segala hormat dan kerendahan hati perkenankanlah penulis mengucapkan terima kasih kepada:

1. Allah SWT, yang telah memberikan rahmat, taufiq serta hidayah-Nya kepada penulis dalam penyelesaian skripsi ini, dan semoga sholawat serta salam tetap tercurah kepada Nabi Muhammad SAW.
2. Kedua orang tuaku Fadholi dan Sutartilah serta keluarga tercinta yang banyak memberikan dukungan dan motivasi, serta do'a yang selalu dipanjatkan kepada Allah SWT. Semoga kita selalu mendapat Ridho dan kasih sayang dari Allah SWT.
3. Bapak Sigit Susanto Putro, S. Kom., M. Kom. dan Bapak Yoga Dwitya Pramudita, S. Kom., M. Cs. yang telah membantu membuka wawasan dan membantu penulis menyelesaikan skripsi ini.
4. Bapak Husni, S. Kom., M. Kom. yang selalu memberikan motivasi dan dukungan kepada penulis untuk menyelesaikan skripsi ini.
5. Seluruh bapak / ibu dosen Teknik Informatika yang telah mendukung dan memberikan bimbingan saya dalam menyelesaikan skripsi ini.
6. Saudara-saudaraku UKMFT-TOFATEK yang telah memberikan kebersamaan dan keluarga baru di kampus tercinta.

7. Saudara-saudaraku seperjuangan INCREMENT terima kasih atas kebersamaannya selama masa perkuliahan, kalian adalah saudara seperjuangan yang terbaik.
8. Terimakasih untuk sahabatku Surya Ade Putra dan Nanda Yana Mardika yang telah membantu dalam segala hal untuk penyelesaian skripsi ini.
9. Semua pihak yang tidak mungkin penulis sebutkan satu persatu yang telah membantu sehingga skripsi ini dapat diselesaikan.

Dalam penyusunan skripsi ini penulis menyadari masih banyak kekurangan, dikarenakan keterbatasan ilmu dan pengalaman yang dimiliki. Untuk itu kritik dan saran yang membangun sangat penulis harapkan agar lebih baik lagi kedepannya.

Akhir kata, semoga apa yang telah dihasilkan dalam skripsi ini dapat bermanfaat bagi kemajuan ilmu pengetahuan dan teknologi khususnya bidang minat Sistem Terdistribusi.

Bangkalan, 30 Juli 2017

Penulis

DAFTAR ISI

LEMBAR PENGESAHAN	iii
HALAMAN PERNYATAAN ORISINALITAS	v
HALAMAN PERSEMBAHAN.....	vii
ABSTRAK.....	ix
KATA PENGANTAR	xi
DAFTAR ISI.....	xiii
DAFTAR GAMBAR	xv
DAFTAR TABEL.....	xvi
BAB I.....	1
1.1. Latar Belakang.....	1
1.2. Perumusan Masalah.....	4
1.3. Tujuan dan Manfaat.....	4
1.4. Batasan Masalah.....	4
1.5. Metodologi Penelitian	5
1.5.1 Studi Pustaka	5
1.5.2 Analisa dan Perancangan Sistem.....	5
1.5.3 Implementasi Sistem	5
1.5.4 Uji Coba Sistem	5
1.5.5 Analisa dan Evaluasi	5
1.5.6 Dokumentasi.....	5
1.6. Sistematika Penulisan	5
BAB II.....	7
2.1. <i>Web Crawler</i>	7
2.2. <i>Single-thread</i> dan <i>Multi-thread</i>	8
2.3. Struktur <i>Website</i>	8
2.4. Arsitektur Distribusi	9
2.5. Replikasi Basis Data Terdistribusi.....	9
2.6. Teknik Multi Master Replicated.....	10
2.7. Deteksi Pembaruan <i>Website</i>	10
2.8. Penelitian Sebelumnya	11
BAB III	15
3.1 Deskripsi Sistem.....	15
3.2 Arsitektur <i>Single-thread</i>	16
3.3 Arsitektur <i>Multithread</i>	17
3.4 Rancangan Database.....	20
3.4.1 Struktur Tabel.....	20
3.5 Rancangan Antar Muka	22

3.6 Skenario Uji Coba Sistem.....	22
BAB IV	27
4.1 Kebutuhan Perangkat Keras (<i>Hardware</i>)	27
4.2 Kebutuhan Perangkat Lunak (<i>Software</i>).....	27
4.3 Struktur <i>Web</i> yang akan di-crawl	27
4.3.1 Struktur Halaman Utama <i>Web</i>	27
4.3.2 Struktur Halaman Isi Berita.....	28
4.4 Pembuatan Sistem.....	30
4.4.1 <i>Bot Crawler Class</i>	30
4.4.2 <i>Database Class</i>	33
4.4.3 <i>Crawler Class Single-thread</i>	36
4.4.4 <i>Crawler Class Multi-thread</i>	38
4.4.5 <i>Content Cleaner Class</i>	39
4.4.6 <i>Main Class</i>	39
4.5 Data Hasil <i>Crawling</i>	42
4.6 Implementasi Sistem.....	42
4.6.1 Tampilan Halaman Utama.....	43
4.6.2 Tampilan Baca Berita	43
4.6.3 Tampilan Hapus Berita.....	44
4.7 Uji Coba Sistem.....	44
4.7.1 Uji <i>Web Crawler Single-Thread</i> (Non-Deteksi Pembaruan) ..	45
4.7.2 Uji <i>Web Crawler Multi-Thread</i> (Non-Deteksi Pembaruan) ..	45
4.7.3 Uji <i>Web Crawler Multi-Thread</i> + Deteksi Pembaruan <i>Web</i>	46
4.7.4 Uji <i>Web Crawler Multi-Thread</i> Terdistribusi + Deteksi Pembaruan <i>Web</i> 1.....	47
4.7.5 Uji <i>Web Crawler Multi-Thread</i> Terdistribusi + Deteksi Pembaruan <i>Web</i> 2.....	48
4.8 Analisa Hasil Uji Coba.....	49
BAB V	51
5.1 Kesimpulan.....	51
5.2 Saran	51
LAMPIRAN.....	55

DAFTAR GAMBAR

Gambar 2. 1 Arsitektur Web Crawler [11].....	7
Gambar 2. 2 Sebelum Perubahan.	11
Gambar 2. 3 Setelah Perubahan.	11
Gambar 3. 1 Arsitektur Web Crawler Multi-thread Terdistribusi.	15
Gambar 3. 2 Arsitektur Web Crawler Single-thread.	16
Gambar 3. 3 Arsitektur Web Crawler Multi-thread.	17
Gambar 3. 4 Flowchart Bot Crawler.	19
Gambar 3. 5 Rancangan Tampilan Pengguna.	22
Gambar 3. 6 Skema Jaringan Komputer Laboratorium Sistem Terdistribusi.	23
Gambar 4. 1 Tampilan Kolom Berita Terbaru Okezone Olahraga.....	28
Gambar 4. 2 Tampilan Halaman Isi Berita Okezone Olahraga	29
Gambar 4. 3 Data Hasil Crawling.	42
Gambar 4. 4 Halaman Utama.	43
Gambar 4. 5 Tampilan Baca Berita.....	44
Gambar 4. 6 Tampilan Hapus Berita.....	44

DAFTAR TABEL

Tabel 2. 1 Analisa Penelitian Sebelumnya.	12
Tabel 3. 1 Struktur Tabel Berita.	20
Tabel 3. 2 Struktur Tabel Console.	21
Tabel 3. 3 Skenario Pengujian Sistem.	23
Tabel 4. 1 Format Tanggal Masing – masing Web.	32
Tabel 4. 2 Uji Coba 1 Single-thread.	45
Tabel 4. 3 Uji Coba 1 Multi-thread	46
Tabel 4. 4 Uji Coba 1 Multi-thread Update.	46
Tabel 4. 5 Uji Coba Server 1.	47
Tabel 4. 6 Uji Coba Server 2.	47
Tabel 4. 7 Uji Coba 1 Multi-thread Update Terdistribusi.	48
Tabel 4. 8 Uji Coba 2 Multi-thread Update Terdistribusi.	49
Tabel 4. 9 Tabel Perbandingan Hasil Ujicoba.	49