

BAB I

PENDAHULUAN

Pada bab ini akan dijelaskan mengenai beberapa hal yang melatarbelakangi penelitian ini di lakukan, rumusan masalah dari persoalan yang ada, batasan permasalahan yang dilakukan dalam penelitian, tujuan dan relevansi atau manfaat penelitian terhadap perkembangan solusi dari permasalahan yang diangkat, serta sistematika penulisan yang dipergunakan.

1.1. Latar Belakang

Informasi yang sangat banyak di *internet* membuat orang tidak sulit untuk mencari informasi yang diinginkan, namun relefan atau tidaknya informasi tersebut menjadi permasalahan yang harus diperoleh solusinya. Oleh karena itulah mesin pencari (*search engine*) sangat dibutuhkan. Alasan seseorang memilih *search engine* mana yang digunakan adalah dari segi kecepatan dan keakuratannya. Pengambilan web (*web crawler*) merupakan salah satu bagian mesin pencari terpenting. Karena bertugas untuk mengumpulkan informasi yang akan diolah. *Web crawler* adalah sebuah aplikasi yang berfungsi untuk mengambil semua informasi yang diperlukan dengan cara menjelajahi halaman-halaman situs di internet yang kemudian disimpan di dalam *database*[1,2]. Sebelum di simpan dalam *database*, data ini akan lebih dahulu di proses agar lebih mudah saat pencarian, penambahan, pengubahan dan penghapusan data.

Desain dan pengimplementasian *web crawler* bahasa Indonesia telah dilaksanakan dan dipublikasikan sebelumnya oleh Arifin(2011) pada tugas akhir Teknik Informatika Universitas Trunojoyo Madura. Penelitian ini diprakarsai oleh proyek search engine khusus untuk bahasa Indonesia yang biasa disebut RISE (*Research Indonesian Search Engine*). *Web crawler* ini mampu mengambil *website* khusus berbahasa Indonesia dengan *regular expression* dan langsung disimpan kedalam *database*. Sebelum proses penyimpanan ke *database*, aplikasi *web crawler* ini mengapus *tag* dan *stopword*[3]. Namun masih banyak yang harus di sempurnakan, salah satunya yaitu di ekstraksi informasi(*web data extraction*) agar data yang akan di proses selanjutnya menjadi lebih akurat dalam perhitungan pembobotan kata.

Salah satu cara untuk ekstraksi informasi yaitu dengan pencarian pola otomatis berdasarkan pencocokan pohon (*tree matching*) (Liu, 2007). Pada tugas akhir ini mencoba untuk mengimplementasikan algoritma *simple tree matching* pada aplikasi utama *web crawler*. Sehingga konten yang tersimpan bisa berupa informasi penting.

1.2. Perumusan Masalah

Dari latar belakang yang telah di tulis sebelumnya dapat di rumuskan permasalahan yaitu:

- a. Bagaimana membangun *web crawler* bahasa indonesia dengan ekstraksi informasi
- b. Bagaimana mendapatkan *content* dari halaman *web* yang di *download*.

1.3. Tujuan dan Manfaat

Tujuan dari penelitian ini adalah:

- a. Menerapkan ekstraksi informasi dalam membangun *web crawler* bahasa indonesia
- b. Membersihkan konten yang sudah terdownload
Manfaat/luaran dari penelitian ini antara lain:
 - a. *Web crawler* mampu menghimpun *website* bahasa indonesia
 - b. Mendapatkan hasil isi *website*/konten yang bersih .

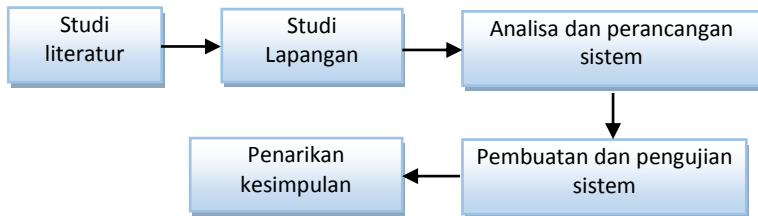
1.4. Batasan Masalah

Agar tidak terlalu melebar dan arah dari tujuan dari penelitian ini jelas maka dapat diambil batasan masalah yang di uraikan sebagai berikut:

1. Aplikasi *web crawler* ini memproses *file* berextension *html*
2. Protocol yang digunakan dalam *web crawler* ini adalah *http*
3. Tidak melayani *autentifikasi proxy*
4. Aplikasi hanya mengekstrak halaman web pada bagian *body*
5. Halaman *website* pembanding harus sejajar kedudukannya.

1.5. Metodologi Penelitian

Ada beberapa tahap-tahap dari metodologi yang di terapkan pada penelitian ini. Diantaranya dari studi literatur, studi lapangan, analisa dan perancangan sistem, implementasi dan pengujian, dan yang terakhir penarikan kesimpulan. Berikut adalah alur serta penjelasan dari masing-masing alur tersebut.



Gambar 1.1 Alur metodologi penelitian

1.5.1. Studi Literatur

Di tahap yang pertama ini, dilakukan pembendaharaan teori mengenai *information retrieval*, *search engine*, *web crawler*, *web data extraction*, *partial tree alignment*, *simple tree matching*.

1.5.2. Studi Lapangan

Uji coba program yang telah di implementasikan sebelumnya oleh Arifin pada tahun 2011. Menjalankan *web crawler* pada waktu-waktu tertentu dan uji coba dengan 5-60 *thread* . Proses yang terjadi diantaranya adalah pengambilan web dengan pengecekan bahasa terlebih dahulu dengan *regular expression*. Setelah halaman *web* sudah tersimpan, *web crawler* membersihkan dari *tag-tag html* dan *stoplist* yang ada. Sehingga data bersih yang tersimpan adalah isi web yang bersih. Dan uji coba ekstraksi informasi web untuk menghasilkan bagian-bagian web yang sudah terkelompokkan menurut perbandingan akar dari pohon yang terbentuk dari laman web.

1.5.3. Analisa dan Perancangan Sistem

Menganalisa kebutuhan apa saja yang nantinya dibutuhkan untuk membuat aplikasi di penelitian ini. Baik berupa modul-modul internal program dan perangkat keras (*hardware*) yang dipergunakan. Setelah mendapatkan semua bahan yang dibutuhkan mulailah perancangan sistem.

1.5.4. Pembuatan dan Pengujian Sistem

Mengimplementasikan rancangan sistem yang sudah di buat dalam bentuk fisik sekaligus melakukan pengujian. Perbandingan analisis diberbeda perangkat keras dan setelah itu baru ditarik sebuah kesimpulan.

1.5.5. Penarikan Kesimpulan

Tahap Terakhir dalam penelitian ini adalah penarikan kesimpulan. Beberapa hal penting dapat di catat dan disusun di dalam sebuah laporan akhir.

1.6. Sistematika

Sistematika penulisan karya ilmiah ini dibagi manjadi beberapa bab yang disusun sesuai tahap pengerjaan. Berikut kami jelaskan per poin:

BAB I : PENDAHULUAN

Pada bagian ini terdiri dari : latar belakang, perumusan masalah, batasan masalah, tujuan penelitian, metode penelitian dan sistematika penulisan.

BAB II : TINJAUAN PUSAKA

Pada bagian ini dijelaskan mengenai hal-hal yang melandasi permasalahan yang diangkat sebagai pedoman untuk menyelesaikannya.

BAB III : PERANCANGAN SISTEM

Pada bagian ini akan dijelaskan pemodelan sistem yang akan dibangun. Modul-modul yang dibutuhkan sistem didesain secara lengkap, sehingga tujuan tercapai secara maksimal.

BAB IV : SIMULASI DAN ANALISIS

Pada bab ini menjelaskan tentang implementasi program yang telah dikerjakan. Meliputi penjelasan tentang arsitektur sistem, desain sistem, desain interface, dan menguraikan tentang hasil uji coba serta analisa terhadap sistem.

BAB V : KESIMPULAN

Bab yang berisi kesimpulan dan saran yang dihasilkan dari sistem yang telah diujicobakan.

(Halaman ini sengaja dikosongkan)