BAB I

PENDAHULUAN

1.1 Latar Belakang

Teknologi yang berkembang pesat dan menjadi salah satu kebutuhan bagi kita semua. Banyak pengguna *internet* mempublikasikan karya yang mereka buat di *internet* sehingga membuat banyak pilihan akan akses informasi [1].

Sedemikian pesatnya pertambahan jumlah dokumen beserta keanekaragamannya, menyebabkan masalah baru pada saat pencarian dokumen. Salah satu kesulitan tersebut yaitu mendapatkan hasil pencarian yang relevan.

Supaya memperudah kita dalam pengambilan informasi dalam jumlah besar pada *internet* dibutuhkan metode *web crawler*. *Web crawler* merupakan salah satu komponen penting dalam sebuah mesin pencari modern. Selain itu *Web crawler* juga memiliki fungsi penjelajahan dan pengunduhan halaman - halaman *web* yang ada di *internet* [1]. Proses pengunduhan informasi dari *website* disebut *web scraping*.

Dengan melakukan metode *web scraping* pada dokumen *web* dalam bahasa *HTML*, kemudian dokumen tersebut diambil data yang kita butuhkan untuk digunakan untuk pengolahan data atau digunakan untuk keperluan lainnya [2].

Bila ditinjau dari volume dokumen teks yang berada di internet, perpustakaan digital, dan web intranet perusahaan yang sangat besar, suatu sistem yang efisien diperlukan untuk mengekstraksi informasi agar waktu untuk mendapatkan informasi menjadi lebih pendek [3].

Pada kebanyakan mesin pencarian saat ini, respon dari *query* pengguna mengembalikan hasil pencarian dengan menampilkan sebagian dari dokumen (*snippets*). Jika *query* terlalu umum, maka sangat sulit bagi *user* untuk mengidentifikasi dokumen mana yang sesuai. Pengguna diharuskan untuk melihat satu-persatu detil hasil pencarian dokumen untuk mengetahui dokumen mana yang relevan bagi *user*. Dan juga, keterhubungan antar dokumen pada hasil pencarian tidak disediakan [4].

Salah satu alternatif yang dapat menyelesaikan masalah diatas yaitu pengelompokan hasil pencarian secara otomatis ke dalam kelompok - kelompok tematik (*cluster*). Hal ini dapat membantu pengguna dalam mengidentifikasi dokumen hasil pencarian secara spesifik.

Berdasarkan penjelasan diatas mengenai *web crawler, web scraping* dan *clustering* tersebut maka dibuatkan aplikasi sebagai tugas akhir yang berjudul "**Implementasi Algoritma *Breadth First Search* Dan *Lingo* Pada Perangkat Lunak Bantu Pengklasteran Abstrak Paper**".

1.2 Rumusan Masalah

Permasalahan yang menarik ketertarikan penulis dalam hal ini berhubungan dengan hal-hal sebagai berikut :

1. Bagaimana menerapkan algoritma *Breadth First Search* dalam proses pencarian data dari suatu *URL* dan meng-unduhnya kedalam database?
2. Bagaimana menerapkan algoritma *Lingo* dalam proses pengklasteran abstrak dari informasi hasil pencarian?
3. Bagaimana performansi algoritma *Breadth First Search* berdasarkan parameter waktu dalam melakukan *scraping*?
4. Bagaimana performansi algoritma *Lingo* berdasarkan parameter waktu dalam melakukan *clustering* dari informasi hasil pencarian?

1.3 Tujuan Penelitian

Tujuan yang ingin dicapai penulis dari penelitian skripsi ini adalah :

1. Dapat menerapkan algoritma *Breadth First Search* dalam proses pencarian abstrak dari suatu *website* dan meng-unduhnya kedalam database.
2. Dapat menerapkan algoritma *Lingo* dalam perangkat lunak bantu pengklasteran abstrak paper dari informasi hasil pencarian.
3. Dapat menganalisis performansi algoritma *Breadth First Search* berdasarkan parameter waktu dalam melakukan *scraping*.
4. Dapat menganalisis performansi algoritma *Lingo* berdasarkan parameter waktu dalam melakukan *clustering* dari informasi hasil pencarian.

1.4 Batasan Masalah

Supaya tugas akhir ini dapat terarah, maka penulis menentukan batasan-batasan dalam pembuatan perangkat lunak bantu pengklasteran abstrak paper. Adapun batasan - batasan masalah penelitian ini adalah sebagai berikut :

1. Dokumen yang di-*download* merupakan abstrak dokumen.

2. Setiap *URL* hanya akan dikunjungi sekali dan *web crawler* tidak memiliki kemampuan untuk *revisit*.
3. *Crawler* tidak akan meng-*crawl external link* dari *URL seeds*.
4. Dokumen yang digunakan hanya dokumen berbahasa Indonesia.
5. Kata - kata bahasa asing dalam isi dokumen akan dianggap sebagai kata dalam bahasa Indonesia.
6. Kesalahan ketik suatu kata dianggap sebagai suatu kata yang baru.

1.5 Sistematika Penulisan

Untuk memperoleh gambaran yang lebih jelas mengenai pembahasan masalah skripsi ini, maka dalam penulisan dicantumkan sistematika pembahasan terdiri dari lima bab sebagai berikut:

BAB I PENDAHULUAN

Mengurai tentang latar belakang permasalahan, mencoba merumuskan inti permasalahan yang dihadapi, menentukan tujuan dan kegunaan penelitian, yang kemudian diikuti dengan pembahasan masalah, asumsi dan sistematika penulisan.

BAB II LANDASAN TEORI

Membahas berbagai konsep dasar dan teori yang berkaitan dengan tahap penelitian yang dilakukan dan hal – hal yang berguna dalam proses analisis perasalahan serta tinjauan terhadap penelitian – penelitian serupa yang pernah dilakukan sebelumnya.

BAB III ANALISIS DAN PERANCANGAN

Menganalisis masalah dari model penelitian untuk memperlihatkan kerkaitan antara *variable* yang diteliti serta model matematis untuk analisisnya. Dan merancang sistem yang akan diimplementasikan pada tahap selanjutnya.

BAB IV IMPLEMENTASI

Merupakan tahapan yang dilakukan dalam penelitian secara garis besar sejak dari tahap persiapan sampai penarikan kesimpulan, metode dan kaidah yang diterapkan dalam penelitian. Termasuk menentukan cara penggumpulan data, penentuan sampel penelitian dan teknik pengambilannya, serta metode analisis yang akan dipergunakan dalam perangkat lunak yang akan dibangun. Serta melakukan tahap pengujian setelah implementasi selesai.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi tentang pernyataan berupa kesimpulan dari pembahasan perangkat lunak yang dibuat secara keseluruhan dan saran untuk mengembangkan perangkat lunak yang lebih baik untuk ke depannya.