BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Digital library adalah sebuah koleksi objek digital yang dapat berupa teks, materi visual, ataupun materi audio, yang disimpan dalam media elektronik. Salah satu keunggulan yang ditawarkan *digital library* adalah dalam hal pencarian. Dalam melakukan pencarian dalam *digital library* diperlukan penerapan teknik tertentu supaya hasil pencarian sesuai dengan apa yang dicari. Salah satunya adalah dengan menerapkan algoritma *stemming* dalam pencarian dokumen di dalam *digital library*.

Stemming adalah proses pemotongan imbuhan dari suatu kata ke bentuk asal atau kata dasarnya [7]. Algoritma *stemming* merupakan fitur penting bagi sistem pengindeksan dan pencarian, karena dapat meningkatkan kemampuan *recall* dengan secara otomatis mengubah suatu kata ke bentuk dasarnya [6]. Algoritma *stemming* juga dapat meningkatan *index compression* yang merupakan hal penting karena dapat menjaga agar penggunaan tempat penyimpanan seminimal mungkin.

Masalah yang kemudian muncul adalah bagaimana menentukan algoritma stemming yang baik dari segi *recall* dan *index compression*. Menurut jurnal karya W. B. Frakes salah satu faktor penentu baik tidaknya suatu algoritma stemming adalah dengan menghitung *stemmer strength* [5], semakin baik *stemmer strength* suatu algoritma maka semakin baik pula nilai *recall* dan *index compression*, walaupun memiliki kelemahan karena dapat menguarangi nilai dari *precission*. Parameter-parameter *stemmer strength* yaitu: ukuran rata-rata *word per conflation class*, *index compression factor*, jumlah kata yang berbeda antara kata masukan dan kata keluaran (*stem*), rata-rata huruf yang dihilangkan untuk membentuk *stem*, dan terakhir nilai *median* dan *mean modified hamming distance* antara kata masukan dan *stem.*

Parameter-parameter *stemmer strength* tersebut di atas akan diuji pada algoritma stemming untuk Bahasa Indonesia. Algoritma stemming pada Bahasa Indonesia diantaranya adalah Ahmad Yusoff Sembok (1996) yang sebenarnya dibuat untuk bahasa melayu, Nazief & Adriani (1996) merupakan algoritma stemming awal untuk Bahasa Indonesia, Idris (2001) merupakan pengembangan dari algoritma Ahmad Yusoff Sembok untuk Bahasa Indonesia, Vega (2001), Arifin & Setiono (2002) merupakan penyederhanaan dari Nazief & Adriani, dan Confix-stripping (2005) merupakan pengembangan dari Nazief & Adriani. Algoritma stemming pada Bahasa Indonesia yang dipilih pada penelitian ini adalah Algoritma vega yang dibuat untuk meningkatkan performa *retrieval* dokumen dalam hal ini *recall* dan *precission,* dan algoritma confix-stripping yang memiliki nilai kompresi lebih baik dari Arifin & Setiono.

Berdasarkan jurnal karya Lusiano Marga N. yang berjudul “Analisis Penggunaan Algoritma Stemming Vega Pada Information Retrieval System”, Algoritma Vega adalah algoritma stemming yang hanya menggunakan aturan morfologi tanpa menggunakan kamus kata dasar dalam menentukan *stem*. Keakuratan *stem* dengan kata dasar bukanlah hal yang ditonjolkan oleh Algoritma Vega, akan tetapi algoritma ini dibuat untuk meningkatkan performa *retrieval* dokumen dalam hal ini *recall* dan *precission* [9].

Berdasarkan jurnal karya Asriko Adipathy yang berjudul "Analisis Dan Implementasi Perbandingan Stemming Dengan Menggunakan Algoritma Jelita Asian Dan Algoritma Arifin & Setiono Pada Information Retrieval", Algoritma Confix-Stripping dimana penulis di sini menyebutnya sebagai Algoritma Jelita Asian, memiliki nilai kompresi yang lebih baik dari pada algoritma Arifin & Setiono. Algoritma Confix-Stripping adalah algoritma stemming yang selain menggunakan aturan morfologi juga menggunakan kamus kata dasar dalam menentukan stem [1].

Berdasarkan penjelasan yang telah dipaparkan, maka Algoritma Vega dan Algoritma Confix-Stripping akan dianalisis dan dibandingkan pada dokumen uji.

Dokumen sample uji yang dipakai adalah *digital library* Sarihusada yang berisi tutorial-tutorial seputar ibu dan anak saat masih di dalam kandungan hingga balita. Dengan dilakukannya analisis dan perbandingan kedua algoritma ini agar diketahui algoritma terbaik yang dapat menghasilkan *stem* (kata dasar) dari kata berimbuhan pada Bahasa Indonesia pada *digital library* tersebut.

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah disebutkan di atas dapat ditemukan masalah yang dapat dirumuskan dalam satu rumusan masalah , yaitu bagaimana melakukan analisis dan perbandingan Algoritma Stemming Vega dan Confix-stripping pada Bahasa Indonesia dengan menentukan kemampuan *recall* dan *index compression*.

1.3. Maksud dan Tujuan

Penelitian ini memiliki maksud dan tujuan yaitu:

1.3.1. Maksud

Berdasarkan permasalahan yang diteliti, maksud dilakukannya penelitian ini adalah untuk melakukan analisis dan perbandingan Algoritma Stemming Vega dan Algoritma Stemming Confix-stripping pada Bahasa Indonesia dengan menghitung parameter ukuran rata-rata *word per conflation class*, *index compression factor*, jumlah kata yang berbeda antara kata masukan dan kata keluaran (*stem*), rata-rata huruf yang dihilangkan untuk membentuk *stem*, dan terakhir nilai *median* dan *mean modified hamming distance* antara kata masukan dan *stem*.

1.3.2. Tujuan

Tujuan yang ingin dicapai dari penelitian ini adalah untuk mengetahui algoritma terbaik yang dapat menghasilkan *stem* (kata dasar) dari kata berimbuhan dalam hal *recall* dan *index compression* pada Bahasa Indonesia khususnya pada *digital library* Sarihusada.

1.4. Batasan Masalah

Agar pembahasan penelitian terfokuskan pada lingkup masalah yang diinginkan, maka ada batasan masalah yang diterapkan, yaitu :

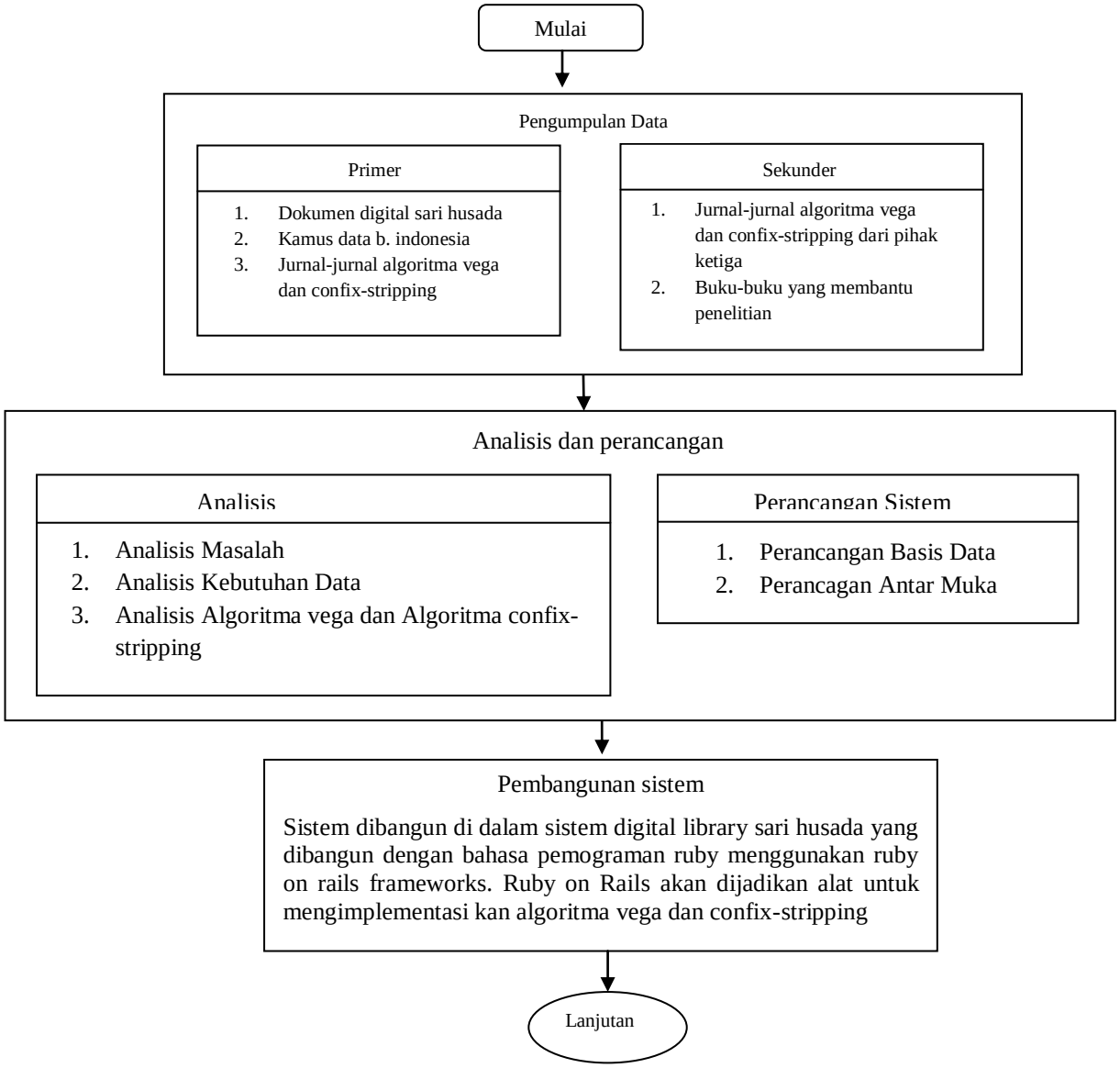
1. Dokumen-dokumen yang dipakai untuk melakukan penelitian adalah dokumen berformat pdf pada *digital library* Sarihusada.
2. Sistem dibangun menggunakan ruby on rails.
3. Penelitian berpusat pada proses stemming tidak sampai mengimplementasikan ke proses pembuatan indeks ataupun proses pencarian.
4. Parameter uji yang dilakukan antar lain: ukuran rata-rata *word per conflation class*, *index compression factor*, jumlah kata yang berbeda antara kata masukan dan kata keluaran (*stem*), rata-rata huruf yang dihilangkan untuk membentuk *stem*, dan terakhir nilai *median* dan *mean modified hamming distance* antara kata masukan dan *stem*

1.5. Metodologi Penelitian

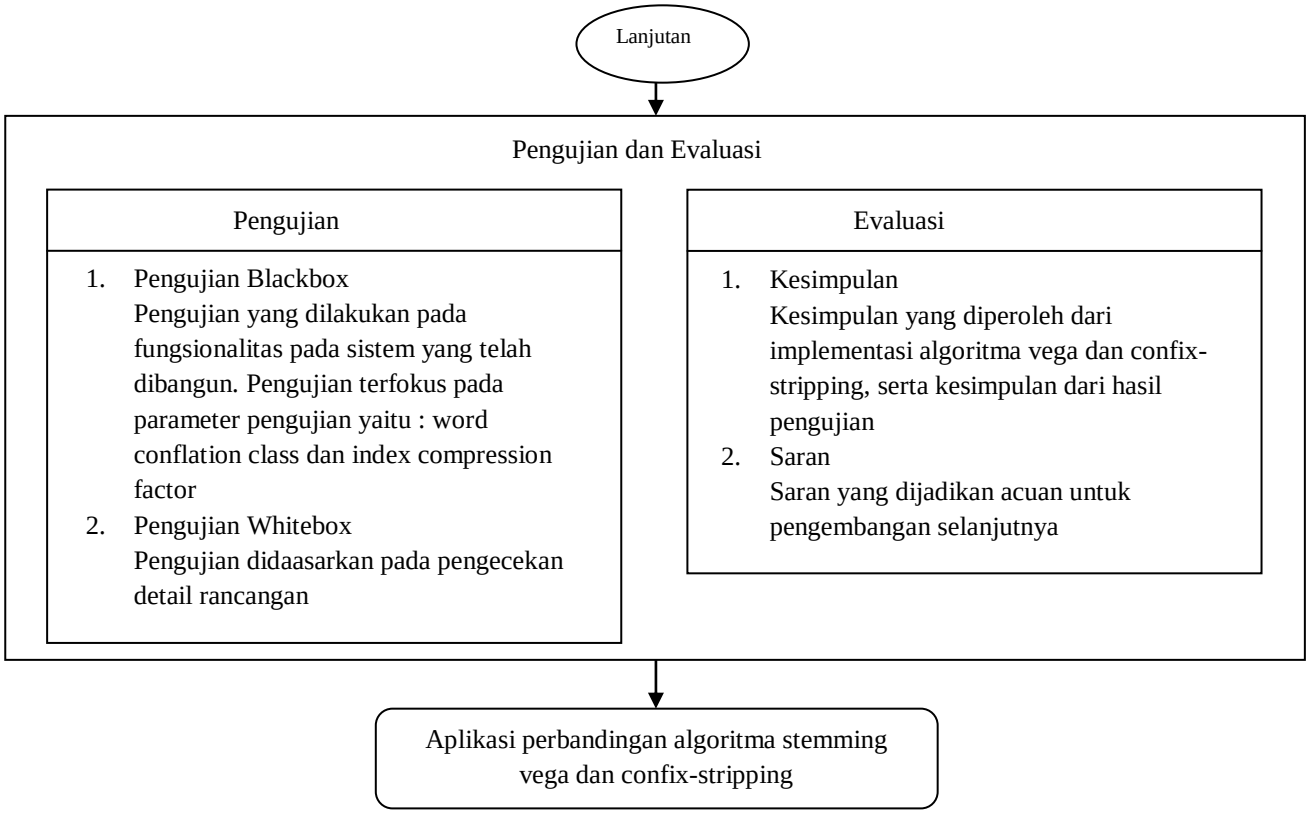
Metodologi penelitian yang digunakan dalam penulisan tugas akhir ini adalah metodologi deskriptif yang bertujuan untuk mengumpulkan informasi aktual secara rinci, melukiskan gejala yang ada, mengindetifikasi masalah atau memeriksa kondisi dan praktik-praktik yang berlaku, membuat perbandingan atau evaluasi.

1.5.1. Alur Penelitian

Diagram pada Gambar 1.1 berikut ini merupakan langkah-langkah yang diambil untuk mendukung proses penelitian agar penelitian dapat berjalan lebih terarah dan sistematis.



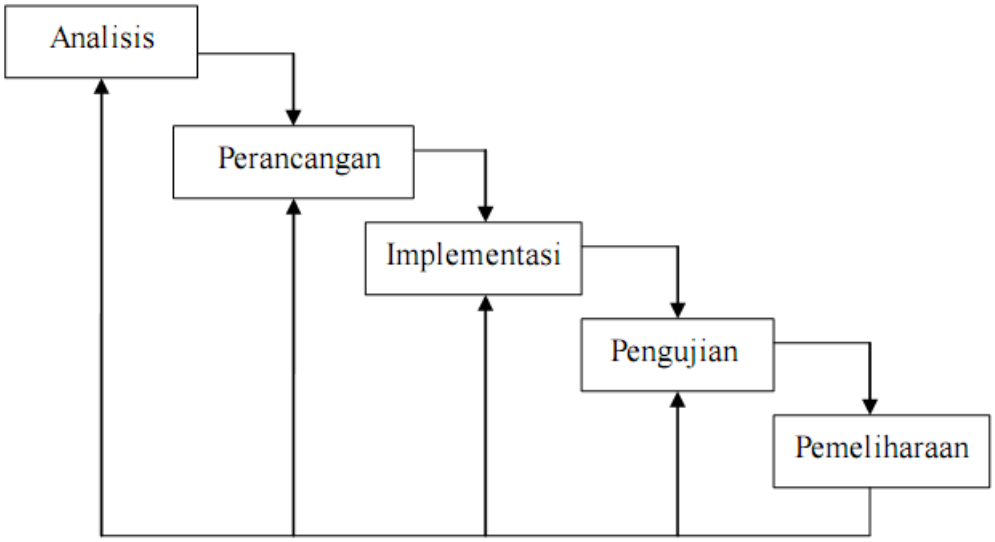
Gambar 1.1 Diagram Alir Penelitian



Gambar 1.1 Diagram Alir Penelitian (lanjutan)

1.5.2. Pembangunan Perangkat Lunak

Dalam membangun perangkat lunak pada penelitian ini, digunakan metode pengembangan perangkat lunak, yaitu dengan menggunakan metode waterfall yang bisa digambarkan seperti Gambar 1.2 berikut :



Gambar 1.2 Waterfall Model

1.6. Sistematika Penulisan

Sistematika penulisan pada penelitian ini disusun untuk memberikan gambaran secara umum mengenai penelitian yang dilakukan. Sistematika penulisan pada penelitian ini sebagai berikut :

BAB 1 PENDAHULUAN

Bab ini menguraikan tentang latar belakang masalah, rumusan masalah, menentukan maksud dan tujuan, batasan masalah, metodologi penelitian serta sistematika penulisan.

BAB 2 LANDASAN TEORI

Bab ini membahas berbagai konsep dasar dan teori - teori yang berkaitan dengan stemming bahasa Indonesia dan algoritma-algoritma stemming. Konsep atau teori yang dijelaskan dimulai dari penjelasan mengenai morfologi bahasa Indonesia, pengertian stemming, pengertian algoritma vega dan algoritma confix-stripping sampai parameter-parameter pengujian yang digunakan.

BAB 3 ANALISIS DAN PERANCANGAN SISTEM

Bab ini menguraikan penjelasan tentang analisis dan perancangan sistem. Analisis sistem berisi analisis masalah, analisis kebutuhan data, analisis algoritma yang terdiri dari algoritma vega dan confix-stripping, analisis kebutuhan nonfungsional, dan analisis kebutuhan fungsional. Perancangan sistem berisi perancangan basis data yang terdiri dari diagram relasi dan struktur table, kemudian perancangan antar muka yang memberikan gambaran antarmuka sistem

BAB 4 IMPLEMENTASI DAN PENGUJIAN

Bab ini berisi tentang implementasi yang bertujuan untuk memastikan bahawa aplikasi telah berhasil diimplementasikan dan dapat berjalan sesuai dengan yang diinginkan, Kemudian disertai hasil pengujian dari aplikasi yang telah dilakukan.

BAB 5 KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan dari hasil penelitian skripsi ini dan saran untuk pengembangan penelitian lebih lanjut.