BAB I

PENDAHULUAN

1.1 Latar Belakang

Dalam sistem temu kembali informasi (*Information Retrieval*), algoritma *stemming* digunakan untuk mengurangi perbedaan bentuk dari suatu kata dengan mengembalikannya ke dalam bentuk kata dasar sehingga proses temu kembali menjadi efektif.

Stemming adalah proses pemotongan (pembuangan) imbuhan (*affix*), baik prefiks maupun sufiks, dari sebuah term untuk mendapatkan kata dasar (*root* atau *stem*) dari kata berimbuhan. Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma *stemming* untuk bahasa lainnya, hal ini dikarenakan perbedaan morfologi pada masing-masing bahasa. Beberapa algoritma *stemming* untuk Bahasa Indonesia telah dikembangkan sebelumnya, diantaranya algoritma Nazief & Adriani (1996) dan algoritma Porter (Tala, 2003).

Algortima Nazief & Adriani pertama kali dikembangka oleh Bobby Nazief dan Mirna Adriani (Universitas Indonesia, 1996). Algoritma Nazief dan Adriani dikembangkan berdasarkan aturan morfologi Bahasa Indonesia yang mengelompokkan imbuhan menjadi awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan gabungan awalan akhiran (*confixes*). Algoritma ini menggunakan kamus kata dasar dan mendukung *recoding*, yakni penyusunan kembali kata-kata yang mengalami proses *stemming* berlebih. Kamus kata dasar tersebut memegang

peranan sangat penting untuk mendapatkan hasil *stemming* yang baik[1]. Kamus kata dasar tersebut dibutuhkan untuk memeriksa apakah kata dasar yang melalui proses *stemming* benar dan ditemukan pada kamus saat proses *stemming* dilakukan.

Algoritma Porter dikembangkan pertama kali dikenalkan oleh F.M. Porter (1980) untuk teks ber Basaha Inggris. Algoritma Porter untuk Bahasa Indonesia dikembangkan dengan berdasarkan Algoritma Porter yang dikembangkan oleh W. B. Frakes (Tala, 2003). Algortima ini menghilangkan imbuhan-imbuhan yang ada berdasarkan aturan morfologi dalam Bahasa Indonesia, tanpa menggunakan kamus. Dalam penelitiannya Fadillan Z Tala (2003) mengatakan, untuk korpus yang berkembang dan dalam jumlah yang besar, ketergantungan pada kamus akan menurunkan kemampuan sistem dalam jangka panjang.

Masing-masing *stemmer* memiliki kelebihan dan kekurangannya masing-masing. Efektifitas algoritma *stemming* dapat diukur berdasarkan beberapa parameter, seperti kecepatan proses, keakuratan, dan kesalahan. Untuk itu sebuah penelitian untuk membandingkan efektifitas algoritma Nazief dan Adriani dengan algoritma Porter untuk proses *stemming* pada teks ber-Bahasa Indonesia, sehingga akhirnya akan diketahui algoritma manakah yang lebih cepat, lebih akurat atau yang lebih banyak melakukan kesalahan *stemming*.

1.2 Perumusan Masalah

Bagaimana tingkat efektifitas dan efisiensi algoritma Nazief & Adriani dibanding algoritma Porter pada teks ber Bahasa Indonesia untuk meningkatkan performa sistem sistem temu kembali (*Information Retrieval*).

---

[1] Jiwa Malem Marsya dan Taufik Fuadi Abidin, *Analisa dan Evaluasi Afiks stemming untuk Bahas Indonesia*, Seminar Nasional dan ExpoTeknik Elektro 2011, ISSN 2088-9984

1.3 Batasan Masalah

Untuk memudahkan dalam penelitian ini, maka diperlukan beberapa pembatasan pokok bahasan, diantaranya:

1. Pengujian dilakukan pada dokumen atau teks ber Bahasa Indonesia.
2. Algoritma yang diuji hanya Nazief & Adriani dan Porter.
3. Aplikasi dibangun berbasis web dengan menggunakan bahasa pemrograman PHP dan database MySQL.
4. Dokumen uji diambil dari berita online pada situs detik.com yang diakses pada tanggal 27 September 2015.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk menganalisis dan membandingkan algoritma *stemming* Nazief & Adriani dan Porter pada teks ber Bahasa Indonesia untuk mendukung performa sistem temu kembali informasi (*Information Retrieval*).

1.5 Manfaat Penelitian

Adapun manfaat luaran dalam penelitian ini untuk pembaca adalah :

1. Mengetahui *stemmer* mana yang lebih efektif pada teks ber Bahasa Indonesia untuk mendukung performa sistem temu kembali.
2. Menjadi referensi berbagai penelitian berkaitan tentang *Information Retrieval*, khususnya *Stemming*.

1.6 Kerangka Pikiran

Dalam upaya untuk meningkatkan performa sistem temu kembali informasi (*information retrieval*), salah satu proses yang harus di lakukan adalah *stemming*, yaitu proses mengubah suatu kata menjadi bentuk dasarnya.

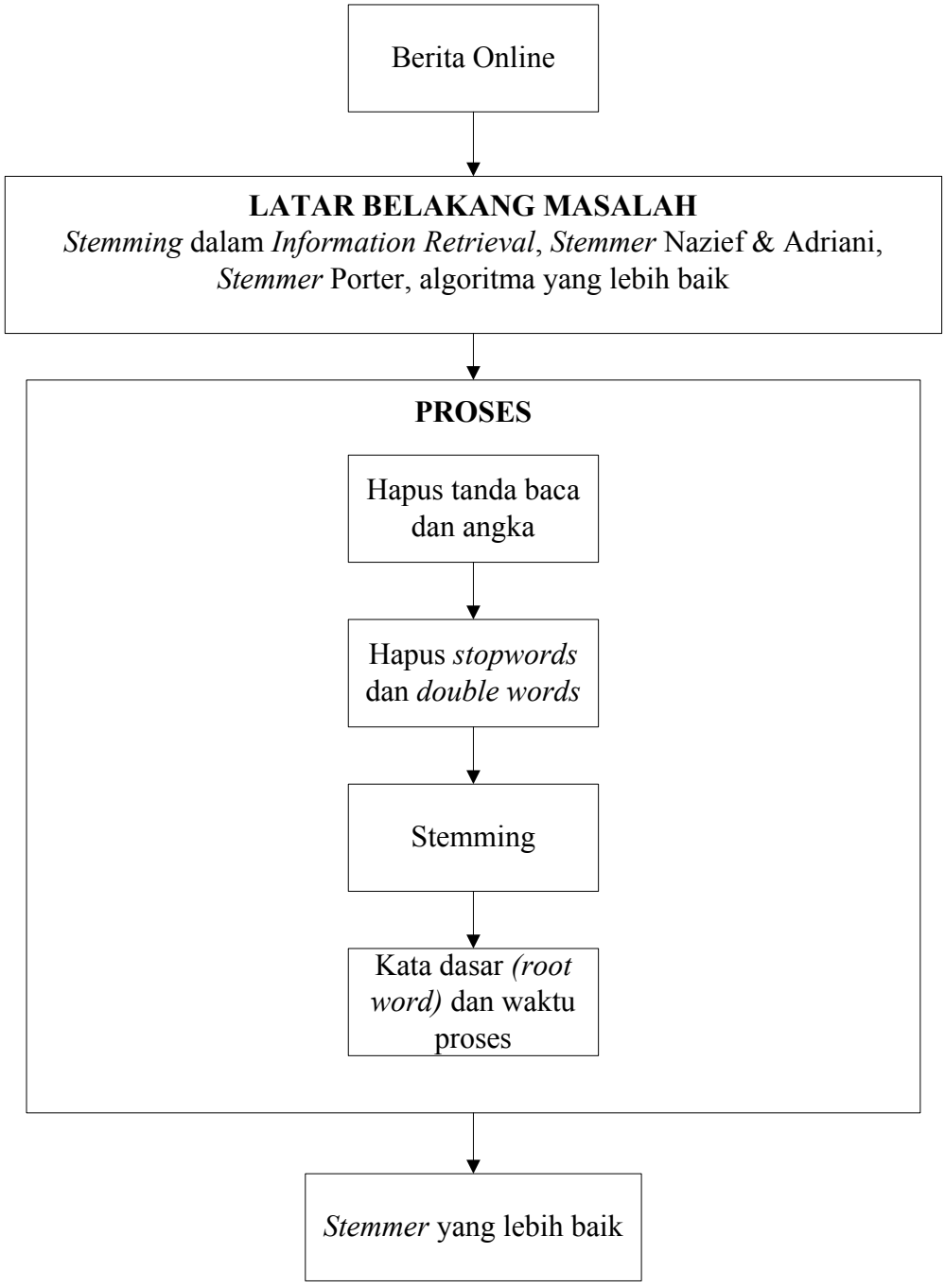
Algortima Nazief & Adriani dan algoritma Porter merupakan dua dari sekian banyak algoritma *stemming* yang dikembangkan untuk Bahasa Indonesia. Kedua algoritma ini memiliki karakteristik yang berbeda, proses penghilangan imbuhan (*stripping*) pada algoritma Nazief & Adriani menggunakan kamus kata dasar sebagai acuan dalam penentuan suatu kata dasar. Algoritma ini juga mendukung proses *recoding*. Sedangkan proses *stripping* pada algoritma Porter hanya mengacu pada aturan morfologi pada Bahas Indonesia.

Masing – masing *stemmer* memiliki kelebihan dan kekurangan. Dengan menganalisis dan membandingkan kedua algoritma *stemming* tersebut, maka akan diketahui algoritma mana yang lebih baik.

Dengan menggunakan dokumen uji yang diambil dari situs berita online detik.com, yang diunduh secara otomatis menggunakan perangkat lunak *crawler* kemudian menyimpan indeksnya dalam database dan kontennya dalam file XML. Setiap dokumen akan dihilangkan tanda baca dan angka, kemudian dihilangkan *stopwords* dan *double words* yang membentuk suatu *corpus*. *Corpus* ini yang kemudian akan dilakukan proses *stemming*, sehingga didapat kata dasar (*root word*) dan waktu prosesnya. Hasil yang didapat ini kemudian akan dianalisis secara manual untuk mengetahui apakah kata dasar tersebut benar atau mengalami kesalahan *stemming* (*understemming*, *overstemming*, *unchange* atau *spelling exception*), kemudian dibandingkannya untuk kedua *stemmer* yang diujikan,

sehingga akan diketahui *stemmer* mana yang lebih baik untuk mendukung sistem temu kembali informasi.

Berikut adalah skema implementasi dan analisis algoritma *stemming* pada dokumen ber Bahasa Indonesia.



Gambar 1.1 Skema pikiran implementasi dan analisis algoritma *stemming*

1.7 Sistematika Penulisan

BAB I Pendahuluan, terdiri dari Latar Belakang Pemilihan Judul, Perumusan Masalah, Batasan Masalah, Tujuan Skripsi, Manfaat Skripsi, Kerangka Pikiran dan Sistematika Penulisan.

BAB II Landasan/Dasar Teori, terdiri dari Teori Sistem Temu Kembali (*Information Retrieval*), Teori *Stemming*, Algoritma Nazief & Adriani, Algoritma Porter.

BAB III Metode Penelitian, dalam Bab ini di uraikan cara atau metode yang digunakan untuk pengumpulan data, literatur, analisis dan perancangan sistem, serta implementasi dan analisis hasil.

BAB IV Gambaran Umum Obyek Penelitian, berisi informasi *stemmer* Porter dan Nazief & Adriani, proses pembandingan dari kedua algoritma tersebut.

BAB V Pembahasan, terdiri dari Diagram alir *stemming*, implementasi algoritma *stemming*, dan analisis hasil *stemming*.

BAB VI Penutup, berisi kesimpulan hasil penelitian, sehingga di ketahui algoritma *stemming* mana yang lebih baik untuk mendukung sistem temu kembali informasi (*Information Retrieval*).

LAMPIRAN Berisi daftar pustaka, Daftar *stopwords list* yang digunakan.