



Hak cipta dan penggunaan kembali:

Lisensi ini mengizinkan setiap orang untuk menggubah, memperbaiki, dan membuat ciptaan turunan bukan untuk kepentingan komersial, selama anda mencantumkan nama penulis dan melisensikan ciptaan turunan dengan syarat yang serupa dengan ciptaan asli.

Copyright and reuse:

This license lets you remix, tweak, and build upon work non-commercially, as long as you credit the origin creator and license it on your new creations under the identical terms.

BAB I

PENDAHULUAN

1.1 Latar Belakang

Information retrieval atau disingkat dengan IR adalah menemukan bahan (dokumen) dari dokumen terstruktur (biasanya teks) yang memenuhi kebutuhan informasi pada ruang lingkup besar (biasanya disimpan di komputer). Di era teknologi ini, pencarian dan pengambilan informasi dilakukan dengan mengakses informasi via *online* menggunakan *web search engine*. Bidang *information retrieval* juga mendukung pengguna dalam *browsing* atau penyaringan koleksi dokumen atau pengolahan lebih lanjut seperangkat dokumen yang diambil (Manning, 2009).

Dalam mendapatkan informasi dalam dunia maya, rata-rata pengguna *search engine* menggunakan 2,4 kata pada satu kali *query* pencarian. Dengan kata lain, *query* pencarian yang dimasukkan oleh pengguna mesin pencarian rata-rata terdiri dari dua sampai tiga kata. Pengeksekusian *query* dilakukan berdasarkan kata kunci yang diberikan oleh pengguna. Namun, pengguna tidak selalu memasukkan kata kunci pencarian dengan memasukkan kata yang tepat (Spink & Jansen, 2004).

Banyaknya informasi yang terdapat dalam *internet*, aplikasi pencarian informasi menjadi hal yang dibutuhkan oleh penjelajah dunia maya untuk mendapatkan informasi sesuai yang dibutuhkan. Saat ini terdapat 634 juta situs *internet* yang ada di dunia maya (Noor, 2013). Aplikasi *search engine* menjadi

aplikasi yang sangat berguna bagi pengguna *internet* untuk mendapatkan informasi.

Dalam perkembangan dunia *information retrieval*, *web search engine* menggunakan algoritma dan teknik perhitungan dalam mengeksekusi *query*. *Information retrieval* memiliki beberapa proses untuk menginterpretasikan *query* antara lain *word token*, *stop word removal*, *word boundaries*, dan *stemming* (Zhu, 2010).

Stemming adalah salah satu cara yang digunakan untuk meningkatkan performa IR dengan cara mentransformasi kata-kata dalam sebuah dokumen teks ke kata dasarnya. Algoritma *stemming* untuk bahasa yang satu berbeda dengan algoritma *stemming* untuk bahasa lainnya. Proses *stemming* pada teks berbahasa Indonesia lebih rumit atau kompleks karena terdapat variasi imbuhan yang harus dibuang untuk mendapatkan kata dasar dari sebuah kata (Agusta, 2009).

Pada tahun 2012, Stephani Marcelli Djojo melakukan penelitian yang berjudul *Web Search Engine Berbasis Algoritma Stemming Nazief-Adriani dan Operator Pencarian untuk Meningkatkan Relevansi Hasil Pencarian (Studi Kasus: Situs Universitas Multimedia Nusantara)*. Penelitiannya telah berhasil menunjukkan proses dan hasil *stemming* dan mengimplementasikannya pada *website* Universitas Multimedia Nusantara. Hasil *stemming* berhasil dilakukan di mana kata yang memiliki imbuhan dapat diubah menjadi kata dasar. Namun, proses *stemming* algoritma *stemming* Nazief-Adriani tersebut masih terdapat kesalahan dalam hasil *stemming* dari beberapa kata berimbuhan yang diberikan.

Algoritma *stemming* kata pada bahasa Indonesia dengan performa yang paling baik saat ini (memiliki jenis kesalahan *stemming* yang paling sedikit) adalah algoritma *stemming Enhanced Confix Stripping* (ECS) (Arifin, 2009). Algoritma *stemming* ECS ini merupakan algoritma perbaikan dari algoritma *stemming Confix Stripping* (CS). Perbaikan yang dilakukan oleh *stemming* ECS adalah perbaikan beberapa aturan pada tabel acuan pemenggalan imbuhan. Selain itu, algoritma *stemming* ECS juga menambahkan langkah pengembalian akhiran jika terjadi penghilangan akhiran yang seharusnya tidak dilakukan (Tahitoe, 2010).

Pada tahun 2002, Sindo Weekly (sebelumnya bernama Trust) adalah majalah mingguan yang berasal dari Indonesia yang diterbitkan oleh PT Hikmat Makna Aksara dengan slogan "Eksklusif dan Terpercaya". Sindo Weekly adalah majalah berita mingguan umum disajikan dalam format, tampilan dan presentasi yang lebih handal, akurat, dan berimbang (MNC).

Pada tahun 2012 dibuat *website* resmi dari Sindo Weekly, www.sindoweekly-magz.com, sebagai fasilitas akses informasi majalah Sindo Weekly via *online*. Laman situs tersebut berisikan isi atau artikel dari majalah Sindo Mingguan yang diterbitkan setiap pekan. Sesuai dengan tujuan utama dari majalah mingguan ini, yaitu "Eksklusif dan Terpercaya", layanan pengaksesan informasi menjadi sangat penting. Dengan banyaknya berita atau artikel dalam *website* tersebut, dibutuhkan fitur pencarian atau *searching* untuk memudahkan pencarian artikel atau berita. Saat ini sudah terdapat fitur pencarian pada laman

website yang hanya berupa *query* standar, tanpa menggunakan algoritma tertentu untuk meningkatkan keakuratan informasi yang dicari.

Berdasarkan kekurangan yang terjadi pada algoritma *stemming* Nazief dan Adriani, penulis melakukan penelitian hasil *stemming* algoritma *Enhanced Confix Stripping*. Kata hasil *stemming* akan digunakan sebagai kata kunci pencarian pada *search engine* untuk meningkatkan hasil *query* pencarian informasi dengan studi kasus situs majalah mingguan Sindo Weekly.

1.2 Perumusan Masalah

Berdasarkan pada latar belakang yang dicantumkan di atas, berikut rumusan masalah.

1. Bagaimana hasil *stemming* algoritma *Enhanced Confix Stripping* dalam melakukan perubahan kata berimbuhan mejadi kata dasar sebagai perbaikan dari kesalahan yang terjadi pada algoritma *stemming* Nazief-Adriani?
2. Bagaimana mengimplementasikan algoritma *stemming Enhanced Confix Stripping* pada *search engine* untuk pencarian informasi?

1.3 Batasan Masalah

Penelitian ini membahas proses *stemming* menggunakan algoritma *stemming Enhanced Confix Stripping* dan pembuangan kata yang tidak memiliki makna atau *stop word removal* kemudian mengimplementasikan algoritma tersebut ke dalam *web search engine*. *Stemming* yang dilakukan hanya dalam bahasa Indonesia. Proses *stemming* tidak dilakukan jika masukan yang diberikan

tidak menggunakan bahasa Indonesia yang benar atau kata tidak terdapat dalam *database* kamus.

Variabel penelitian menggunakan metode evaluasi Paice, yaitu berdasarkan pada nilai *understemming index* dan *overstemming index*.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mengimplementasikan algoritma *stemming Enhanced Confix Stripping* ke dalam sebuah *web search engine* untuk meningkatkan hasil pencarian.

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah sebagai berikut.

1. Meningkatkan layanan pencarian informasi via *online*, dalam kasus ini pencarian artikel atau berita pada situs Sindo Weekly Magazine.
2. Meningkatkan kinerja *search engine* dengan mengimplementasikan algoritma *stemming* dari fitur *search engine* yang telah ada.

1.6 Sistematika Penulisan

Laporan ini tersusun dari beberapa bab dengan penjelasan sebagai berikut.

1. Bab I: Pendahuluan

Bab ini berisikan tentang latar belakang permasalahan, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan laporan.

2. Bab II: Landasan Teori

Dalam bab ini berisikan teori-teori terkait dengan penelitian ini. Teori-teori yang digunakan antara lain adalah teori mengenai *information retrieval*, algoritma *stemming* (Nazief-Adriani, *Confix Stripping*, *Enhanced Confix Stripping*), *stop word* dan metode evaluasi Paice sebagai variabel perhitungan hasil implementasi algoritma yang digunakan dalam penelitian.

3. Bab III: Analisis dan Perancangan

Bab ini berisikan metodologi penelitian, analisis mengenai proses perancangan sistem, serta desain antarmuka *stemmer* dan *Sindo Search Engine*. Digunakan diagram umum dan *flowchart* untuk memudahkan pemahaman perancangan sistem.

4. Bab IV: Uji Coba dan Pembahasan

Dalam bab ini berisikan hasil uji coba terhadap hasil implementasi algoritma *stemming* ECS beserta analisis pembahasan oleh penulis dari hasil yang diperoleh tersebut.

5. Bab V: Kesimpulan dan Saran

Bab ini berisikan kesimpulan oleh penulis mengenai penelitian serta saran yang dapat diterapkan untuk penelitian selanjutnya.