IMPLEMENTASI MODIFIKASI ALGORITMA ENHANCED CONFIX STRIPPING STEMMER PADA TEKS BAHASA INDONESIA

Noverdy Anggara[1], Ade Romadhony[2], Mahmud Dwi Suliiyo[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Stemming merupakan proses pembentukan kata dasar dari kata-kata yang telah mendapatkan modifikasi dalam penggunaannya. Penggunaan kata yang terdapat pada kalimat terstruktur diantaranya sudah mendapat imbuhan yang terdiri dari awalan, akhiran ataupun sisipan. Stemming merupakan bagian dari Preprocessing, yaitu fase terakhir setelah Tokenization dan Stoplist Removal. Proses stemming berbeda dalam tiap bahasa karena dalam pembentukan kata memiliki perbedaan pada tiap bahasa. Pada bahasa Indonesia, ada beberapa algoritma yang dapat dipakai dalam proses stemming, diantaranya Algoritma Arifin-Setiono, Algoritma Nazief-Adriani dan Algoritma Enhanced Confix Stripping (ECS) Stemmer. Algoritma ECS adalah algortima perbaikan dari algoritma Confix Stripping (CS) Stemmer.

Pada tugas akhir ini akan diajukan implementasi dan modifikasi algoritma Enhanced Confix Stripping Stemmer pada teks berbahasa Indonesia. Algoritma Enhanced Confix Stripping Stemmer memiliki kekurangan dan keterbatasan dalam menangani kata yang memiliki sisipan dan kata yang memiliki huruf akhir seperti akhiran. Modifikasi algortima Enhanced Confix Stripping Stemmer dirancang untuk melakukan perbaikan terhadap kelemahan tersebut. Dari hasil pengujian akan terlihat perbedaan tingkat akurasi antara algoritma Ehnanced Confix Stripping Stemmer dan modifikasi algoritma Ehnanced Confix Stripping Stemmer, akan dibuktikan bahwa skema yang telah dimodifikasi dapat menghasilkan nilai akurasi yang lebih tinggi.

Kata Kunci : Stemming, Enhanced Confix Stripping Stremmer, Imbuhan, Preprocessing.

Abstract

Stemming is the process of forming the base of the words that have been getting modifications in its use. The use of words contained in the sentence structure of which have got Affixes include prefixes, suffixes, or infixes. Stemming is part od the Preprocessing, which is the last phase after Tokenization and Stoplist Removal. Stemming process is different in each language because the word has a different formation in each language. In Indonesia, there are several algorithms that can be used in the process of stemming, such like Arifin-Setiono Algorithm, Nazief-Adriani Algorithm and Enhanced Confix Stripping Stemmer (ECS) Stemmer. ECS algorithm is a refinement of the Confix Stripping Stemmer algorithm.

This final project will be presented and the implementations of ECS Stemmer algorithm and its modifications to the Indonesian language text. Enhanced Confix Stripping Stemmer Algorithm have drawbacks and limitations in dealing with infixes and few letter in th end part of word that such as suffix. Modifications of Enhanced Confix Stripping Stemmeris designed to repair the weakness. From the test result will be seen the difference between the accuracy of the Enhanced Confix Stripping Stemmer algorithm and its modification, it will be proven that the scheme has been modified to produce a higher accuracy value.

Keywords : Stemming, Enhanced Confix Stripping Stemmer, Affixes, Preprocessing.

I. PENDAHULUAN

I.1 Latar belakang masalah

Information Retrieval System merupakan sistem pengambilan informasi yang dapat diimplementasikan pada pencarian kata pada isi dan konteks dokumen. Dalam me*retrieve* berbagai jenis dokumen dilakukan *preprocessing* untuk mengambil informasi yang ada. *Preprocessing* sendiri terdiri dari *tokenization, Stoplist Removal,* dan *Stemming.*

Stemming merupakan suatu teknik untuk mentranformasi kata-kata dalam sebuah dokumen teks menjadi bentuk kata dasar[1]. Proses *stemming* berbeda dalam tiap bahasa karena pada setiap bahasa yang digunakan di berbagai negara memiliki aturan-aturan yang berbeda dalam penggunaan kata berimbuhan[4]. Bahasa Perancis memiliki perbedaan aturan penggunaan tata bahasa dengan bahasa Arab. Pada bahasa Indonesia terdapat komplekstias pada variasi imbuhan yang menjadi titik fokus pada pembentukan kata dasarnya. Algoritma Stemming yang digunakan pertama kali untuk menstemming bahasa Indonesia adalah Algoritma Nazief-Adriani (1996), mengacu pada algoritma Porter Stemmer yang digunakan pada bahasa Inggris. Algoritma stemming mengalami perkembangan untuk meminimalisir kekurangan-kekurangan yang ada, setelah Algoritma Nazief-Adriani selanjutnya ada algoritma Vega (2001), algoritma Arifin-Setiono (2002) dan algoritma Confix Stripping Stemmer (2007). Yang diangkat pada tugas akhir ini adalah algoritma Enhanced Confix Stripping Stemmer (2008), merupakan algoritma perbaikan dari Confix Stripping Stemmer.

Algoritma Enhanced Confix Stripping (ECS) Stemmer dapat digunakan untuk melakukan stemming pada dokumen teks bahasa Indonesia[3]. Algoritma *ECS Stemmer* memiliki beberapa kelemahan, diantarannya keterbatasan dalam me*nstemming* kata yang memiliki sisipan, lalu kekurangan mengenai overstemming. Oleh sebab itu, dalam Tugas Akhir ini, diajukan modifikasi perbaikan terhadap algoritma *ECS Stemmer* untuk mengatasi kelemahan tersebut sehingga mendapatkan tingkat akurasi yang lebih baik.

I.2 Perumusan masalah

Permasalahan-permasalahan yang akan diteliti pada tugas akhir ini antara lain:

Bagaimana menghasilkan tingkat akurasi yang lebih baik dari algoritma-algoritma *stemming* untuk bahasa Indonesia sebelumnya dengan memodifikasi algoritma Enhanced Confix Stripping Stemmer?

Adapun batasan-batasan masalah pada Tugas Akhir ini antara lain :

a. Teks yang digunakan sebagai dokumen uji merupakan novel yang menggunakan bahasa Indonesia yang baku.
b. Kamus yang menjadi acuan adalah KBBI.
c. Teks uji yang digunakan sudah dialihkan menjadi bertipe .txt
d. Sistem yang akan dibangun menggunakan bahasa pemrograman java.

I.3 Tujuan

Mengacu pada masalah-masalah diatas, tujuan Tugas Akhir ini adalah :

Menghasilkan tingkat akurasi yang lebih baik dari algoritma-algoritma *stemming* untuk bahasa Indonesia sebelumnya dengan memodifikasi algoritma Enhanced Confix Stripping Stemmer

I.4 Hipotesa

Dengan menambahkan skema baru pemotongan imbuhan dan aturan tambahan tabel pemotongan imbuhan pada Modifikasi Algoritma Enhanced Confix Stripping Stremmer untuk *stemming* pada teks bahasa Indonesia akan menghasilkan akurasi yang tinggi (rata-rata lebih dari 90% dari 4 novel dokumen uji) daripada Algoritma Enhanced Confix Stripping Stemmer murni tanpa modifikasi.

I.5 Metodologi Penyelesaian

Metodologi penyelesaian masalah yang akan dilakukan pada penelitian Tugas Akhir ini adalah sebagai berikut :

1. Studi literatur

   Tahap ini akan melakukan pencarian referensi dan materi yang ada, berupa paper, jurnal internasional dan buku. Memahami dan mempelajari referensi tersebut untuk menyelesaikan permasalahan dalam tugas akhir ini. Pencarian referensi meliputi studi pustaka tentang:

   a. Stemming
   b. Enhanced Confix Stripping Stremmer
   c. Aturan pemenggalan imbuhan bahasa Indonesia
   d. Algoritma Confix Stripping Stemmer

2. Pengumpulan data

   Mengumpulkan dataset aturan pemenggalan imbuhan pada teks bahasa Indonesia, mengumpulkan dokumen uji, stoplist dan kamus acuan berdasarkan KBBI.

3. Analisis dan perancangan sistem

   Melakukan analisis *stemming* pada teks bahasa Indonesia, merancang sistem menggunakan data yang sudah dikumpulkan.

4. Implementasi dan pembangunan sistem

   Melakukan pengimplementasian sistem, akan dituangkan kedalam sebuah program yang bisa melakukan proses *stemming* pada teks bahasa Indonesia secara akurat.

5. Pengujian sistem dan analisa hasil

   Melakukan pengujian terhadap program yang sudah dirancang sedemikian rupa agar mendapatkan hasil yang selanjutnya akan dianalisis.

6. Pengambilan kesimpulan dan penyusunan laporan

   Melakukan pengambilan kesimpulan dari hasil penelitian dan melakukan penyusunan laporan Tugas Akhir.

I.6 Sistematika Penulisan

Tugas akhir ini disusun dengan sistematika penulisan sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini dibahas mengenai latar belakang, rumusan masalah, batasan masalah, tujuan, metodologi penelitian, dan sistematika penulisan Tugas Akhir ini.

BAB II LANDASAN TEORI

Pada bab ini dibahas mengenai teori-teori yang digunakan dalam penyusunan Tugas Akhir. Teori yang terdapat pada bab ini mencakup pengertian *Stemming*, *preprocessing*, algoritma Enhanced Confix Stripping Stemmer.

BAB III PERANCANGAN SISTEM

Pada bab ini dibahas mengenai langkah-langkah dalam mengidentifikasi *Stemming*, perancangan sistem (*user interface*).

BAB IV IMPLEMENTASI DAN ANALISIS

Pada bab ini dibahas mengenai implementasi modifikasi algoritma Enhanced Confix Stripping Stemmer, uji coba terhadap program yang telah dibuat, dan melakukan analisis terhadap hasil yang didapat dari implementasi.

BAB V PENUTUP

Pada bab ini berisi kesimpulan yang didperoleh dari pembuatan Tugas Akhir ini dan saran yang mungkin dapat berguna dalam penelitian selanjutnya.

V. KESIMPULAN DAN SARAN

V.1 Kesimpulan

Dari uji coba yang telah dilakukan dan menganalisis hasil pengujian terhadap implementasi dan modifikasi algortima Enhanced Confix Stripping Stemmer ini dapat diambil beberapa kesimpulan antara lain :

a. Algortima Enhanced Confix Stripping Stemmer dapat digunakan untuk proses Stemming dokumen text berbahasa Indonesia. Tetapi beberapa kata tidak mampu di*stemming* dikarenakan kata gabung, bahasa asing, salah tulis dan kata tidak tertera dalam kamus.
b. Modifikasi algoritma ECS Stemmer memiliki akurasi yang lebih tinggi dibandingkan algoritma ECS murni dengan perbedaan akurasi sebesar 0,1% sampai dengan 5,0% karena secara umum kelemahan yang ada pada ECS murni mampu ditutup pada penerapan modifikasi ECS Stemmer.
c. Diperlukan algoritma tambahan untuk kata yang mengacu pada dua kata atau lebih walaupun sudah dapat ter*stemming.*
d. Kamus kata dasar sebagai acuan pencari kata dasar memiliki peranan penting dalam penerapan algoritma stemming apapun. Semakin lengkap kamus kata dasar semakin banyak kata dapat ter*stemming*.

V.2 Saran

Saran untuk penelitian lebih lanjut tentang permasalahan yang diangkat pada tugas akhir ini, untuk kata yang berakhir pada dua kata atau lebih, gunakan algoritma tambahan seperti algoritma *connected component* atau algoritma lain yang dapat menyelesaikan permasalahan tersebut.

DAFTAR PUSTAKA

[1] Agusta, Ledy. 2009. **Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia**. Bali, Indonesia : Fakultas Teknologi Informasi, Universitas Kristen Satya Wacana

[2] Ardiansyah, Shandy. 2012. **Implementasi dan Optimasi Algoritma Nazief Adriani pada Dokumen Teks Bahasa Indonesia.** Bandung, Indonesia.

[3] Arifin, A.Z. dan A.N. Setiono. 2002. **Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering**. *Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA)*, Teknik Elektro, Institut Teknologi Sepuluh Nopember.

[4] Arifin, A.Z., I.P.A.K. Mahendra dan H.T. Ciptaningtyas. 2009. ***Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language***, *Proceeding of International Conference on Information & Communication Technology and Systems (ICTS).*

[5] Asian, J. 2007. ***Effective Techniques for Indonesian Text Retrieval***. *PhD Thesis*. School of Computer Science and Information Technology RMIT University Australia.

[6] Manning, Christopher D., Prabhakar Raghavan dan Hinrich Schütze. 2009. ***An Introduction to Information Retrieval***. Cambridge: Cambridge University

[7] Rimaldo, M.Y. 2012. **Sistem Pendeteksi Plagiat Dokumen Teks Menggunakan Algoritma Smith-Waterman serta Algoritma Arifin dan Setiono**. Teknik Informatika, Institut Teknologi Telkom.

[8] Wicaksono, Y. A. 2012. **Analisis dan Implementasi Algoritma Rabin-karp dan Algoritma Nazief-Adriani pada Sistem Pendeteksi Plagiat Dokumen Teks Berbahasa Indonesia**. Teknik Informatika, Institut Teknologi Telkom.

[9] Xu, J. and Croft, W. B. 1998. ***Corpus-based stemming using cooccurrence of word variants***. ACM Transactions on Information Systems, 16 (1), pp. 61-81.