ANALISIS DAN IMPLEMENTASI PERBANDINGAN STEMMING MENGGUNAKAN ALGORITMA NAZIEF & ADRIANI DENGAN ALGORITMA AHMAD YUSOFF SEMBOK PADA INFORMATION RETRIEVAL

Yonissa Qamilla Intan Atazsu[1], Angelina Prima Kurniati[2], Moch Arif Bijaksana[3]

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

Abstrak

Stemming dalam sistem Information Retrieval digunakan untuk membatasi varian bentuk kata yang berbeda menjadi bentuk dasarnya, sehingga nantinya dapat meningkatkan kemampuan sistem dalam menemukan dokumen relevan sesuai query yang ada. Sistem Information retrieval dikatakan ideal apabila dapat menemukan informasi yang relevan saja. Term yang ada pada query ataupun kumpulan informasi memiliki banyak varian morfologi sehingga term-term yang memiliki bentuk beda tidak akan dianggap ekivalen oleh sistem. Dalam tugas akhir ini, dibuat sebuah sistem Information retrieval yang mengimplementasikan teknik stemming dengan menggunakan algoritma Nazief & Adriani dan algoritma Ahmad Yusoff Sembok.

Algoritma Nazief & Adriani dan algoritma Ahmad Yusoff Sembok merupakan algoritma stemming untuk teks berbahasa indonesia. Perbedaan kedua algoritma ini ada pada proses penghilangan imbuhannya (affixes). Untuk Algoritma Nazief & Adriani, proses penghilangan imbuhan dilakukan pada akhiran (suffixes) terlebih dahulu. Sedangkan pada algoritma Ahmad Yusoff Sembok penghilangan imbuhan terlebih dahulu dilakukan pada awalan (prefixes).

Pada tugas akhir ini dilakukan analisis pengaruh penerapan stemming menggunakan algoritma Nazief&Adriani dan algoritma Ahmad Yusoff Sembok. Hasil penelitian menunjukkan bahwa dengan stemming sistem mampu mereduksi term yang dihasilkan sehingga mampu mengurangi ukuran index. Dari sudut pandang performansi sistem, bisa dikatakan penerapan algoritma Nazief&Adriani lebih baik dibanding algoritma Ahmad Yusoff Sembok dimana nilai recall selalu lebih besar daripada nilai precision, dan nilai recall pada algoritma Nazief&Adriani lebih besar dari nilai recall pada algoritma Ahmad Yusoff Sembok.

Kata Kunci : Information Retrieval, Stemming, Algoritma Nazief & Adriani, Algoritma Ahmad Yusoff Sembok

Abstract

Stemming in Information Retrieval system is used to limit the different variant forms of the word into its basic form, so that later may improve the ability of the system in finding relevant documents according to existing queries. Information retrieval system is said to be ideal if it can find relevant information only. Terms that exist in the query or the collection of information has many morphological variants so that terms which have a different form will not be deemed equivalent by the system. In this thesis, created an Information retrieval system that implements a stemming technique using the algorithm and the algorithm Nazief & Adriani Sembok Ahmad Yusoff.

Nazief&Adriani algorithm and the algorithm is an algorithm Ahmad Yusoff Sembok stemming for Indonesian language text. The second difference of this algorithm is in the process of removal imbuhannya (affixes). For Algorithm Nazief & Adriani, affixes removal process performed on the suffix (suffixes) in advance. While the algorithm is the removal of affixes Ahmad Yusoff Sembok first performed on the prefix (prefixes).

In this final analysis was performed using the influence of the application of stemming algorithms and algorithms Nazief & Adriani Sembok Ahmad Yusoff. The results showed that the system is able to reduce the term stemming generated so as to reduce the size of the index. From the standpoint of system performance, to say the application of the algorithm Nazief & Adriani more better than algorithm Ahmad Yusoff Sembok recall where the value is always greater than the value of precision and recall values in the algorithm Nazief & Adriani greater than the value of a recall on Ahmad Yusoff Sembok algorithm.

Keywords : Information Retrieval, Stemming, Algoritma Nazief&Adriani algorithm, Ahmad Yusoff Sembok algorithm

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Perangkat pencarian informasi dapat dikatakan sedang dalam keadaan diproses adalah ketika *user* memasukkan sebuah *query* ke dalam sistem. *Query* tersebut dapat dimisalkan pada sebuah *string* pencarian *web* pada mesin pencari. Ada beberapa *string* yang dapat langsung cocok namun belum tentu relevan. Saat ini kebutuhan pengguna akan perangkat pencarian informasi semakin meningkat dan jumlah dokumen teks yang dapat diakses juga semakin banyak, hal ini dapat mengakibatkan user semakin sulit menemukan dokumen yang relevan dengan query yang diinputkan. Sistem *Information Retrieval* yang ideal adalah sistem dimana dapat menemukan informasi yang relevan sesuai permintaan pengguna. Indikator yang lazim dipakai untuk menilai relevansi hasil pencarian suatu dokumen adalah kesesuaian antara *query* yang diberikan dan dokumen yang diperoleh. Akan tetapi, *term-term* yang terdapat di dokumen dan di *query* sering memiliki banyak varian morfologik, sehingga pasangan *term* seperti "memakan" ,"dimakan" dan "makan" tidak akan dianggap ekivalen oleh sistem tanpa suatu bentuk *Natural Language Processing* (NLP). [10]

Pada beberapa kasus, varian morfologik dari *term-term* memiliki interpretasi semantik yang sama dan dapat dianggap ekivalen oleh sistem. Jika dicari suatu dokumen dengan judul "baca buku" dengan menggunakan *query* "membaca", dokumen yang dimaksud tidak akan pernah terdapat dalam hasil pencarian. Dengan *stemming*, *term* seperti "membaca" dan "dibaca" akan dianggap memiliki interpretasi yang sama yaitu menjadi *term* "baca" sehingga antara *term* pada dokumen *index* dengan *query* bisa cocok. Dengan begini pencarian dokumen akan berhasil.

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan [4]. Dalam *Information Retrieval*, algoritma *stemming* digunakan untuk mengurangi perbedaan bentuk dari suatu kata dengan mengembalikannya ke dalam bentuk kata dasar. Hal ini bertujuan untuk meningkatkan kemampuan sistem dalam menemukan dokumen relevan sesuai *query* yang ada sehingga proses temu kembali menjadi lebih efisien. Algoritma *Stemming* untuk bahasa satu dengan bahasa lainnya berbeda. Untuk *stemming* Bahasa Indonesia, ada beberapa teknik pendekatan seperti Jelita Asian (2005), Arifin & Setiono (2002), Nazief & Adriani (1996), Ahmad Yusoff Sembok (1996), Vega (2001), dan Idris (2001). Teknik *stemming* dikembangkan untuk alasan mereduksi *term* menjadi bentuk dasarnya. *Term* yang ada pada dokumen dan *query* memiliki banyak varian morfologik maka akan sulit *term-term* tersebut dianggap ekivalen. Namun dalam beberapa kasus tertentu varian morfologik *term-term* memiliki interpretasi semantik yang sama dan dapat dikategorikan ekivalen.

Dalam tugas akhir ini, penulis menganalisis 2 algoritma stemming untuk bahasa indonesia yaitu Algoritma Nazief & Adriani dan algoritma Ahmad Yusoff Sembok. Perbedaan kedua algoritma ini ada pada proses penghilangan imbuhan

pada kata berbahasa Indonesia. Untuk Algoritma Nazief & Adriani, proses penghilangan imbuhan dilakukan pada akhiran (*suffixes*) terlebih dahulu. Tetapi pada algoritma Ahmad Yusoff Sembok penghilangan imbuhan dilakukan di awalan (*prefixes*). Persamaan dari kedua algoritma tersebut adalah keduanya sama-sama menggunakan kamus untuk pengecekan kata dasar. Dengan persamaan dan perbedaan tersebut itulah mengapa penulis memilih kedua algoritma ini sebagai topik tugas akhir.

Algoritma Ahmad Yusoff Sembok dikembangkan sebagai pendekatan baru dalam *stemming*. *Stemmer* ini tidak sebaik *stemmer* lain namun ada beberapa kata yang tidak dapat diproses pada *stemmer* lain seperti Nazief&Adriani namun dapat diproses oleh Ahmad Yusoff Sembok. Maka dari itu saya memilih algoritma Ahmad Yusoff Sembok sebagai pembanding algoritma Nazief&Adriani yang sudah lebih dulu dikenal.[1]

Stemming yang diimplementasikan pada tugas akhir ini digunakan pada *Information Retrieval*. Kemudian dianalisis pengaruh proses *stemming* tersebut terhadap proses *Information Retrieval* sehingga dapat disimpulkan teknik *stemming* yang terbaik atau yang paling cocok untuk digunakan pada proses *Information Retrieval* untuk teks berbahasa Indonesia. Selain itu, penulis membandingkan performa dan tingkat keakuratan untuk masing-masing algoritma. Mengacu pada 2 konsep algoritma yang diimplementasikan pada proses *stemming* untuk tugas akhir ini maka *stemming* untuk teks berbahasa Indonesia ini diharapkan dapat mendukung proses *Information Retrieval*.

1.2 Perumusan masalah

Berdasarkan latar belakang masalah yang dikemukakan diatas, penulis merumuskan bahwa masalah-masalah yang akan diselesaikan yaitu:

1. Bagaimana mengimplementasikan Algoritma Nazief & Adriani dan Algoritma Ahmad Yusoff Sembok dalam proses *stemming* pada teks berbahasa Indonesia?
2. Bagaimana perbandingan antara algoritma Nazief&Adriani dan algoritma Ahmad Yusoff Sembok jika ditinjau dari pengaruhnya terhadap *recall, precision, non-interpolated average precision, factor kompresi index (icf),* rata-rata jumlah *term* dalam suatu *conflation class (wc),* serta tingkat keakuratan hasil *term* yang telah di-*stemming*?
3. Bagaimana pengaruh proses *stemming* yang telah diimplementasikan terhadap proses *Information Retrieval* untuk teks berbahasa Indonesia?

Adapun batasan masalah pada Tugas Akhir ini adalah :

1. Teks yang digunakan untuk *document collection* merupakan berkas berita teks berbahasa Indonesia dengan *query* dan *relevance judgments* yang telah ditentukan sebelumnya yang didapat dari hasil riset *research group* Laboratorium Data Mining Centre (DMC).
2. Pengujian dilakukan secara *offline.*
3. Dokumen merupakan dokumen *free text (unstructured text).*
4. Parameter tingkat keakuratan algoritma *stemming* berdasarkan pada nilai *stem* yang di *stemming* dengan benar.

5. Parameter tingkat kekuatan *stemmer* (*stemmer strength)* dalam mereduksi index *term* berdasarkan pada analisis *icf* (*Index Compression Factor)* dan *wc (Number Of Word Per Conflation Class).*
6. Parameter tingkat performansi terhadap *Information Retrieval* berdasarkan *precision* & *recall* dan *non-interpolated average precision.*

1.3 Tujuan

Tujuan yang ingin dicapai dalam pengerjaan Tugas Akhir ini adalah sebagai berikut :

1. Melakukan implementasi dari algoritma *stemming* yang dipilih yaitu Algoritma Nazief & Adriani dan Algoritma Ahmad Yusoff Sembok.
2 Melakukan perbandingan antara algoritma Nazief&Adriani dan algoritma Ahmad Yusoff Sembok yang ditinjau dari segi *recall, precision, non-interpolated average precision, factor kompresi index (icf),* rata-rata jumlah *term* dalam suatu *conflation class (wc),* serta tingkat keakuratan hasil *term* yang telah di-*stemming*.
3 Menganalisis pengaruh dari masing-masing algoritma *stemming* yang diimplementasikan terhadap *Information retrieval*.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah yang akan dilakukan dalam Tugas Akhir ini adalah :

1. Studi Literatur
   Tahap ini akan melakukan pencarian referensi-referensi dan materi yang ada di internet serta memahami dan mempelajarinya sehingga dapat digunakan untuk menyelesaikan permasalahan dalam tugas akhir ini. Pencarian referensi berkaitan dengan pembangunan aplikasi search engine berbasis Web, *Information Retrieval* dan lebih mendalam mengenai algoritma stemming Nazief & Adriani dan Ahmad Yusoff Sembok.
2. Pengumpulan Data
   Tahap ini akan melakukan pengumpulan data berupa dokumen berita Bahasa Indonesia. Penulis juga membuat sekumpulan dokumen uji untuk membantu dalam skenario pengujian.
3. Analisis Kebutuhan Sistem
   Tahap ini akan melakukan analisis sistem dan menentukan kebutuhan dari sistem yang akan dibangun seperti kebutuhan fungsional sistem, spesifikasi perangkat lunak dan perangkat keras yang digunakan, serta pemodelan sistem yang akan dibangun.
4. Perancangan Sistem
   Tahap ini akan melakukan perancangan sistem dan perangkat lunak serta menerapkan Algoritma Nazief & Adriani dan Ahmad Yusoff Sembok untuk stemmingnya.
5. Implementasi dan Pengujian
   Tahap ini akan melakukan implementasi hasil perancangan dan pengujian terhadap performansi sistem yang telah dibangun. Pada proses ini dilakukan 3 skenario pengujian yakni menguji pengaruh penerapan kedua algoritma terhadap performansi *Information Retrieval*, menguji *stemmer strength* dari

algoritma *stemming dan* menganalisis performansi dari tiap-tiap algoritma yang diimplementasikan.

7. Analisis
   Tahap ini akan melakukan analisa hasil pengujian dan pengukuran performansi berdasarkan data yang diuji serta mengambil kesimpulan dari hasil yang telah dianalisa. Pengujian metode dilakukan dengan menggunakan sistem yang sebelumnya telah diimplementasikan pada tahap implementasi.
8. Pembuatan Laporan
   Tahap ini akan melakukan dokumentasi tahap-tahap kegiatan dan hasil yang di dapat ke dalam laporan Tugas Akhir. Di tahap ini akan dijelaskan pula mengenai langkah-langkah secara detail dalam menganalisis kebutuhan dari awal, perancangan sistem, implementasi, pengujian, serta analisisnya.

5. Kesimpulan dan Saran

Kesimpulan yang dapat diambil dari tugas akhir ini antara lain:

1. Rata-rata nilai *recall* dan *non-IAP* yang dihasilkan relatif lebih besar pada algoritma Nazief&Adriani dibandingkan dengan algoritma Ahmad Yusoff Sembok. Hal ini menunjukkan bahwa performansi *information retrieval* yang menggunakan algoritma Nazief&Adriani lebih baik dari pada algoritma Ahmad Yusoff Sembok.
2. Dilihat dari nilai icf dan wc yang dihasilkan algoritma Nazief&Adriani lebih besar dibandingkan dengan yang dihasilkan algoritma Ahmad Yusoff Sembok. Ini berarti kekuatan *(stemmer strength)* dari algoritma Nazief&Adriani lebih tinggi daripada algoritma Ahmad Yusoff Sembok.
3. Algoritma Nazief menghasilkan tingkat keakuratan hasil *stemming* yang lebih baik daripada algoritma Sembok. Ini terlihat dari banyaknya kata yang tidak berhasil di *stemming* oleh algoritma Sembok. Semakin sedikit kata unik setelah *stemming*, maka algoritma tersebut semakin akurat.

5.1 Saran

Saran dari penulis untuk keperluan penelitian lebih lanjut:

1. Gunakan juga dokumen uji yang mempunyai jumlah dokumen sekitar 3000 dokumen. Sepertinya dengan banyaknya dokumen uji yang dipakai, semakin terlihat performansi dari sistem *Information Retrieval* yang dibuat. Dan bisa semakin terlihat bagaimana *stemmer strength* dari algoritma *stemming* yang digunakan.

Daftar Pustaka

[1] Ahmad, Fatimah, Yusoff, M., & Sembok M.T.,Tengku. 1996. "*Experiments with a Stemming Algorithm for Malay Words*". Journal of the American Society for Information Science,47,909-918

[2] Asriko Adipathy. 2010. "*Analisis Algoritma Jelita Asian dan Arifin&setiono untuk Information Retrieval*. IT Telkom, Bandung

[3] Danang Nur Hadianto. 2009. *"Implementasi dan Analisis Algoritma STANS Stemming dalam Information Retrieval Sistem"*. IT Telkom, Bandung.

[4] Fadillah Z. Tala. 2009. "*A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*". Netherland, Universiteit van Amsterdam.

[5] Jelita Asian, Hugh E. Williams, S.M.M. Tahaghoghi. 2005. "*Stemming Indonesian*". School of Computer Science and Information

[6] Nazief, Bobby dan Mirna Adriani. 1996. "*Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia*". Fakulty of Computer Science University of Indonesia.

[7] Othman, A. 1993. "*Pengakar Perkataan Melayu untuk Sistem Capaian Dokumen*". MSc Thesis. National University of Malaysia

[8] Rila Mandala. 2010 "*Peningkatan Performansi Sistem Temu-Kembali Informasi dengan Perluasan Query Secara Otomatis*". Institut Teknologi Bandung, Bandung.

[9] Rini Riandha Asri. 2010. *"Analisis Stemming pada Information Retrieval Sistem dengan Algoritma Porter dan Krovetz"*. IT Telkom, Bandung.

[10] Vega, V. B. 2001. "*Information Retrieval for the Indonesian Language*". Master's thesis, National University Singapore.

[11] Yosi Amelia Putri. 2009. *"Stemming untuk Teks Berbahasa Indonesia dan Pengaruhnya dalam Kategorisasi"*. IT Telkom, Bandung.