PERBANDINGAN ALGORITMA STEMMING NAZIEF & ADRIANI, PORTER DAN ARIFIN SETIONO UNTUK DOKUMEN TEKS BAHASA INDONESIA

Oppie Rezalina

Program Studi Teknik Informatika
Fakultas Teknik
Universitas Muhammadiyah Jember

Email : oppierezalina@gmail.com

ABSTRAK

Pencarian informasi berupa dokumen teks atau yang dikenal dengan istilah Information Retrieval (IR) merupakan proses pemisahan dokumen-dokumen yang dianggap relevan dari sekumpulan dokumen yang tersedia. Salah satu bagian penting dari Information Retrieval adalah proses stemming. Stemming adalah proses mereduksi kata berimbuhan menjadi kata dasar. Dengan proses stemming, kata yang dimasukkan ke dalam index adalah dalam bentuk umum, sehingga dapat menghasilkan dokumen yang lebih relevan. Terdapat beberapa algoritma stemming yang memiliki kelebihan dan kekurangan masing-masing. Pada penelitian ini penulis ingin mengetahui performansi dari algoritma stemming Nazief & Adriani, Porter dan Arifin Setiono pada pencarian kata dasar yang terdapat dalam abstrak sebuah jurnal dengan implementasi yang dilakukan yaitu memberikan suatu kemudahan dalam hal stemming dokumen teks berbahasa Indonesia serta analisis algoritma yang tingkat akurasinya besar dan membutuhkan waktu yang cepat. Parameter yang diuji berupa kecepatan dan ketepatan dari masing-masing algoritma. Dari hasil pengujian sistem dapat disimpulkan bahwa algoritma Nazief & Adriani memiliki tingkat keakurasian paling tinggi dengan prosentase 0,1% lebih akurat dibandingkan algoritma Arifin Setiono dan 0,9% lebih akurat dibandingkan dengan algoritma Porter. Begitu juga dalam hal kecepatan, algoritma Nazief & Adriani masih lebih cepat menyelesaikan proses stemming dibandingkan dengan dua algoritma lainnya. Sebagai saran untuk pengembang berikutnya kamus kata dasar diharapkan lebih lengkap dan melakukan pengembangan terhadap morfologi pada algoirtma porter stemmer untuk memperoleh akurasi yang lebih besar.

Kata Kunci: Stemming, Nazief & Adriani, Porter, Arifin Setiono.

I. Latar Belakang

Pencarian informasi berupa dokumen teks atau yang dikenal dengan istilah Information Retrieval (IR) merupakan proses pemisahan dokumen-dokumen yang dianggap relevan dari sekumpulan dokumen yang tersedia. Salah satu bagian penting dari Information Retrieval adalah proses stemming. Stemming adalah proses mereduksi kata berimbuhan menjadi kata dasar. Stemming sangat berguna untuk proses indexing maupun searching di dalam Information Retrieval. Dengan proses stemming, kata yang dimasukkan ke dalam index adalah dalam bentuk umum, sehingga dapat menghasilkan dokumen yang lebih relevan. Metode stemming adalah salah satu cara yang digunakan untuk mengubah kata untuk menemukan akar kata dengan menerapkan aturan morfologi bahasa yang baik dan benar. Proses stemming dilakukan dengan menghilangkan semua imbuhan (affiks) baik yang terdiri dari awalan (prefiks) sisipan (infiks) maupun akhiran (suffiks) dan kombinasi awalan dan akhiran (konfiks).

Algoritma-algoritma stemming memiliki kelebihan dan kekurangannya masing-masing. Terdapat penelitian sebelumnya mengenai Perbandingan Algoritma Stemming Porter dan Algoritma Stemming Adriani Nazief Untuk Stemming Dokumen Teks Bahasa Indonesia yang menganalisis perbandingan pada dokumen berbahasa Indonesia. Berdasarkan hasil penelitian tersebut, disimpulkan bahwa algoritma porter lebih baik dalam hal kecepatan waktu namun memiliki kelemahan dalam hal keakuratan. Penelitian lainnya menyebutkan bahwa algoritma Arifin Setiono digunakan karena memiliki kelebihan dalam hal mengatasi Overstemming yaitu jika kata tidak ditemukan setelah penghapusan maka algoritma ini kemudian mencoba untuk mengembalikan semua kombinasi yang dihapus untuk mendapatkan kata yang valid.

Dalam penelitian ini akan dilakukan analisis performansi pada dokumen teks dengan menggunakan metode stemming dengan membandingkan dari tiga algoritma stemming Nazief & Adriani, Porter dan Arifin setiono yang

nantinya akan diterapkan pada dokumen berbahasa Indonesia. Parameter yang akan diuji yaitu kecepatan dan ketepatan dari ketiga algoritma yang berpengaruh pada presentasi algoritma yang di implementasikan.

Berdasarkan uraian latar belakang masalah yang dikemukakan, maka dapat dirumuskan beberapa masalah sebagai berikut:

1. Bagaimana performansi algoritma Nazief & Adriani dalam *stemming* teks berbahasa Indonesia.
2. Bagaimana performansi algoritma Porter dalam *stemming* teks berbahasa Indonesia.
3. Bagaimana performansi algoritma Arifin Setiono dalam *stemming* teks berbahasa Indonesia.
4. Bagaimana hasil perbandingan algoritma Nazief & Adriani, Porter dan algoritma Arifin Setiono untuk proses *stemming* teks berbahasa Indonesia.

Tujuan dari penelitian ini adalah untuk mengetahui performansi berupa kecepatan dan ketepatan dari algoritma Nazief & Adriani, Porter dan algoritma Arifin Setiono dengan metode *stemming* dan membandingkannya.

Adapun batasan masalah dari penelitian ini yaitu:

1. Dokumen yang digunakan adalah dokumen berbahasa Indonesia.
2. Parameter yang akan di hasilkan pada aplikasi ini adalah kecepatan dan ketepatan dari ketiga algoritma.
3. Kamus sebagai pembanding kata yang di *stemming* berupa kata dasar yang sesuai dengan Kamus Besar Bahasa Indonesia.
4. Taham *filtering* dihilangkan..

Sedangkan metodologi yang digunakan pada penelitian ini adalah sebagai berikut.

1. Metode studi literatur yaitu pengumpulan data yang di lakukan melalui membaca dan mempelajari reeferensi-referensi berupa jurnal ilmiah, skripsi dan buku. Fasilitas internet yang di pergunakan untuk media sebagai pencari data atau informasi yang di publikasikan di dunia maya yang berkaitan dengan obyek penelitian.
2. Analisis dan Perancangan. Pada perancangan sistem dilakukan perancangan antarmuka pengguna (*user*). Pada perancangan hasil, form hasil menampilkan tingkat waktu perhitungan dan hasil deteksi kata dasar isi dari abstrak penelitian yang telah diproses.
3. Implementasi Sistem

Dalam implementasi sistem ini, yang dapat dilakukan pada *stemming* dengan metode Nazief & Adriani, Porter dan Arifin Setiono untuk menganalisa ketepatan dalam menentukan kata dasar dan waktu yang dibutuhkan dalam mencari kata dasar dari suatu kata.

4. Pengujian

Tahap pengujian merupakan tahap yang ingin mengetahui kesesuaian sistem yang telah dibangun. Yang dilakukan dalam tahap pengujian ini adalah mengevaluasi *system* menggunakan *black box testing* dan mengevaluasi pengguna berdasarkan *user experience*.

II. Landasan Teori

Information Retrieval

Information Retrieval (IR) adalah ilmu pencarian informasi dari sejumlah data yang sudah hilang karena terlalu banyaknya data yang ada. Ilmu ini dipopulerkan oleh Vannevar Bush pada tahun 1945 dan implementasinya mulai dikenalkan pada tahun 1950-an. Pada tahun 1990-an, sudah banyak teknik dan metode dari in*form*atioan retrieval yang dikembangkan dan dipakai. Tujuan dari sistem IR adalah untuk memenuhi kebutuhan informasi pengguna dengan me-*retrieve* semua dokumen yang mungkin relevan, pada waktu yang sama me-*retrieve* sesedikit mungkin dokumen yang tidak relevan.

Sistem IR yang baik memungkinkan pengguna menentukan secara cepat dan akurat apakah isi dari dokumen yang diterima memenuhi kebutuhannya. Tujuan yang harus dipenuhi adalah bagaimana menyusun dokumen yang telah didapatkan tersebut ditampilkan terurut dari dokumen yang memiliki tingkat relevansi tinggi ke tingkat relevansi yang lebih rendah. Penyusunan dokumen tersebut disebut sebagai perangkingan dokumen.

Stemming

Stemming adalah suatu proses pencarian bentuk dasar dari tiap kata yang berada pada suatu dokumen teks, selain untuk memperkecil jumlah indeks yang berbeda dari suatu dokumen, juga untuk melakukan pengelompokan kata-kata lain yang memiliki kata dasar dan arti yang serupa namun memiliki bentuk atau form yang berbeda karena mendapatkan imbuhan yang berbeda dengan menerapkan aturan morfologi bahasa Indonesia yang baik dan benar.

Proses *stemming* dilakukan dengan menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*preffixes*) sisipan (*infixes*) maupun akhiran (*suffixes*), *stemming* dilakukan atas dasar asumsi bahwa kata-kata yang memiliki stem yang sama memiliki makna dasar yang sama.

Teknik *stemming* dapat dikategorikan menjadi 3 yaitu berdasarkan aturan dalam bahasa tertentu, berdasarkan kamus, dan berdasarkan kemunculan bersama. Salah satu tujuan utama dilakukan proses *stemming* adalah meningkatkan efisiensi dengan cara memilah isi dokumen menjadi unit-unit kecil yang akan menjadi penciri misalnya berupa kata, frase atau kalimat. Terdapat beberapa algoritma dalam *stemming*, antara lain:

1. Algoritma Nazief & Adriani

Algoritma Nazief & Adriani

Algoritma Nazief dan Adriani memiliki tiga komponen, yaitu: pengelompokan imbuhan, urutan penggunaan aturan (*rule*) dan kamus (*dictionary*). Kamus akan dicek setiap penerapan aturan *stemming* berhasil diidentifikasi, dan apabila *stemming* berhasil menemukan akar kata maka algoritma akan mengembalikan kata dalam kamus dan algoritma berhenti.

Langkah-langkah *stemming* algoritma Nazief & Adriani:

1) Kata yang akan di*stemm* dicari dalam kamus. Jika ditemukan maka dianggap kata tersebut adalah akar kata sehingga kata tersebut di*return* dan algoritma *stop* di sini.
2) Hilangkan imbuhan infleksi (“-lah”, “-kah”,”-ku”,”-mu” dan “-nya”). Jika ini berhasil dan jika akhiran adalah partikel (“-lah” atau “-kah”) langkah ini dilanjutkan dengan menghilangkan imbuhan possesive (“-ku”, “-mu” dan “-nya”).
3) Hilangkan imbuhan derivasi (“-i” atau “-an”).
   Jika berhasil, lanjutkan ke langkah 4, jika tidak lakukan hal berikut ini:
   a. Jika “-an” dibuang dan huruf terakhir dari kata adalah “-k”, maka “-k” juga dibuang dan pergi ke langkah 4.
   b. Penghilangan akhiran “-i”, “-an” dan “-kan” dibatalkan.
4) Penghilangan awalan dengan berbagai variasi.
   Jika semua langkah telah ditempuh dan tidak berhasil, maka kembalikan kata asli yang belum di*stemm*.

2. Algoritma Porter

Algoritma kedua yang digunakan adalah algoritma Porter. Adapun langkah – langkah algoritma ini adalah sebagai berikut:

1. Hapus *particle*.
2. Hapus *possesive pronoun*.
3. Hapus awalan pertama. Jika tidak ada lanjutkan ke langkah 4a, jika ada cari maka lanjutkan ke langkah 4b.
4. a. Hapus awalan kedua, lanjutkan ke langkah 5a.
   b. Hapus akhiran, jika tidak ditemukan maka kata tersebut diasumsikan sebagai *root word*. Jika ditemukan maka lanjutkan ke langkah 5b.
5. a. Hapus akhiran. Kemudian kata akhir diasumsikan sebagai kata dasar.
   b. Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *root word*.

Pada Porter Stemmer untuk Indonesia perlu ditambahkan beberapa aturan dalam algoritma agar memberikan hasil yang lebih maksimal dan untuk mempermudah proses stem maka dibuatlah beberapa kamus kecil, antara lain sebagai berikut :

1. Kamus kata dasar yang dilekati partikel, untuk menyimpan kata dasar yang memiliki suku kata terakhir (partikel infleksional) serta kata tersebut tidak mendapat imbuhan apapun.Seperti: masalah.
2. Kamus kata dasar yang dilekati partikel berprefiks untuk menyimpan kata dasar yang memiliki suku kata terakhir (partikel infleksional) dan mempunyai prefiks. Seperti: menikah.
3. Kamus kata dasar yang dilekati kata ganti milik, untuk menyimpan kata dasar yang memiliki suku kata terakhir (kata ganti infleksional) serta kata dasar tersebut tidak mendapatkan imbuhan apapun. Seperti: bangku.
4. Kamus kata dasar yang dilekati kata ganti milik berprefiks, untuk menyimpan kata dasar yang memiliki suku kata terakhir (kata ganti infleksional) dan mempunyai prefiks. Seperti: bersuku.
5. Kamus kata dasar yang dilekati prefiks pertama, untuk menyimpan kata dasar yang memiliki suku kata pertama (prefiks derivasional pertama) serta kata dasar tersebut tidak mendapatkan imbuhan apapun. Seperti: median.
6. Kamus kata dasar yang dilekati prefiks pertama bersufiks, untuk menyimpan kata dasar yang memiliki suku kata pertama (prefiks derivasional pertama) dan mempunyai sufiks derivasional.Seperti: terapan.
7. Kamus kata dasar yang dilekati prefiks kedua, untuk menyimpan kata dasar yang memiliki suku kata pertama (prefiks derivasional kedua) serta kata dasar tersebut tidak mendapatkan imbuhan apapun. Seperti: percaya.
8. Kamus kata dasar yang dilekati prefiks kedua bersufiks, untuk menyimpan kata dasar yang memiliki suku kata pertama (prefiks derivasional) dan mempunyai sufiks derivasional. Seperti: perasaan.

Kamus kata dasar yang dilekati sufiks, untuk menyimpan kata dasar yang memiliki suku kata terakhir (sufiks derivasional). Seperti: pantai.

3. Algoritma Arifin Setiono

Algoritma Arifin Setiono merupakan algoritma yang digunakan untuk pencarian kata dasar pada dokumen teks dengan teknik *stemming*. Input dari algoritma ini adalah dokumen teks yang diproses sehingga menghasilkan output berupa kata dasar. Algoritma Arifin Setiono mengasumsikan bahwa setiap kata memiliki dua awalan dan tiga akhiran, yaitu:

[AW1] + [AW2] + KD + [AK3] + [AK2] + [AK1]

Dimana AW = awalan, KD = kata dasar dan AK = akhiran (Hamzah, 2006).

Langkah – langkah algoritma Arifin Setiono dalam proses *stemming* isi dokumen teks adalah sebagai berikut:

1. Lakukan pemeriksaan setiap kata, siapkan variabel p1,p2,s1,s2,s3
2. Pemotangan dilakukan secara berurut, yaitu:
   a. Awalan I, hasil disimpan pada p1
   b. Awalan II, hasil disimpan pada p2
   c. Akhiran I, hasil disimpan dalam s1
   d. Akhiran II, hasil disimpan dalam s2
   e. Akhiran III, hasil disimpan dalam s3

Setiap tahap pemotongan hasil dicek dalam kamus, jika ada dalam kamus algoritma selesai, jika tidak ada proses dilanjutnya ke pemotongan berikutnya. Jika sampai pada langkah 2.e. belum ditemukan dalam kamus, maka dilakukan proses kombinasi. Kata dasar yang dihasilkan dikombinasikan dengan imbuhan-imbuhan dalam 12 kombinasi, yaitu:

a. Kata Dasar
b. Kata Dasar + AK III
c. Kata Dasar + AK III + AK II
d. Kata Dasar + AK III + AK II + AK I
e. AW I + AW II + Kata Dasar
f. AW I + AW II + Kata Dasar + AK III
g. AW I + AW II + Kata Dasar + AK III + AK II
h. AW I + AW II + Kata Dasar + AK III + AKII + AK I
i. AW II + Kata Dasar
j. AW II + Kata Dasar + AK III
k. AW II + Kata Dasar + AK III + AK II
l. AW II + Kata Dasar + AK III + AK II + AK I

III. Pembahasan

Implementasi Sistem

Setelah dilakukan perancangan, maka tahap selanjutnya adalah implementasi sistem ke dalam bentuk program komputer. Bahasa pemrograman yang digunakan adalah PHP dengan menggunakan aplikasi database MYSQL. Aplikasi ini berjalan di komputer dengan sistem operasi Windows 7. Implementasi yang dilakukan yaitu memberikan suatu kemudahan dalam hal stemming teks bahasa Indonesia dengan algoritma Stemming Nazief & Adriani, Porter dan Arifin Setiono serta analisis algoritma yang tingkat akurasinya besar dan membutuhkan waktu yang cepat.

Hasil Pengujian

Uji Coba algoritma dilakukan pada 10 dokumen teks bahasa Indonesia dengan ukuran dokumen yang bervariasi. Tabel hasi pengujian terdiri dari nama dokumen, jumlah data real, jumlah kata dasar dari setiap algoritma yang berhasil di s*temm*. Data real merupakan data uji yang didapat melalui perhitungan manual dengan mencari kata dasar pada dokumen yang sama. Hasil uji coba dokumen teks dapat dilihat pada tabel dibawah.

IV. Kesimpulan dan Saran

Kesimpulan

Dari hasil pengujian sistem dapat disimpulkan bahwa algoritma Nazief & Adriani lebih unggul dalam hal kecepatan dan akurasi dibandingkan dengan dua algoritma lainnya.

Saran

Berdasarkan hasil pengujian algoritma stemming Nazief & Adriani, Porter dan Arifin & Setiono, saran untuk pengembang berikutnya kamus kata dasar diharapkan lebih lengkap dan melakukan pengembangan terhadap morfologi pada algoirtma porter stemmer untuk memperoleh akurasi yang lebih besar dan diharapkan tahap filtering tidak dihilangkan.

V. Daftar Pustaka

1. Agusta, Ledy. 2009. Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief dan Adriani Untuk Stemming Dokumen Teks Bahasa Indonesia. Fakultas Teknologi Informasi Universitas Kristen Satya Wacana.
2. Asian, Jelita. 2007. Effective Techniques For Indonesian Text Retrieval. Australia: RMIT University.
3. Firmansyah, Arif. 2015. Analisis Perfomansi Algoritma Arifin Setiono Dan Algoritma Porter Untuk Stemming Berbahasa Indonesia. Bandung: Unikom. (Online) http://elib.unikom.ac.id/gdl.php?mod=browse&op=read&id=jbptunikompp-gdl-arieffirma-33911. Diakses terakhir pada 3 Maret 2016.
4. Hamzah, Amir. 2006. Pengaruh Stemming Kata Dalam Peningkatan Unjuk Kerja Document Clustering Untuk Dokumen Berbahasa Indonesia. Jurusan Teknik Informatika, Institut Sains & Teknologi AKPRIND.
5. Maarif, Abdul Azis. 2015. Penerapan Algoritma Tf-Idf Untuk Pencarian Karya Ilmiah. Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro.
6. Pardede, Jasman., dkk. 2013. Implementasi Metode Generalized Vector Space Model Pada Aplikasi Information Retrieval. Jurusan Teknik

Informatika Institut Teknologi Nasional Bandung.
7. Rozi, M Latif., dkk. 2013. Implementasi Dan Analisis Perbandingan Algoritma Stemming Nazief & Adriani Dengan Algoritma Stemming Vega Dalam Information Retrieval System. Fakultas Teknik Informatika Universitas Telkom.

Tabel Hasil Perhitungan Stemming

No	Dokumen	Data Real	Nazief & Adriani		Porter		Arifin & Setiono	
			Kata Dasar	Akurasi	Kata Dasar	Akurasi	Kata Dasar	Akurasi
1	Analisis Kapasitas Simpang Bersinyal Pada Pertigaan Jalan Hayam Wuruk – Jalan Mojopahit Kabupaten Jember	94	69 12.797 detik	73 %	69 34.567 detik	73 %	70 25.4 detik	74 %
2	Aplikasi Pengukuran Kualitas Jasa Sistem Informasi Dengan Logika Fuzzy	80	69 13.093 detik	86 %	69 34.585 detik	86 %	70 17.459 detik	88 %
3	Cropping Plat Nomor Mobil Pada Citra Digital Dengan Metode Mathematical Morphology	85	69 14.247 detik	81 %	68 36.08 detik	80 %	71 19.681 detik	84 %
4	Efektivitas Kekakuan Struktur Bangunan Gedung Terhadap Gempa	119	107 19.403 detik	90 %	104 57.554 detik	87 %	105 37.073 detik	88 %
5	Identifikasi Faktor Usia, Jenis Kelamin dengan Luas Infark Miokard pada Penyakit Jantung Koroner (PJK) Di Ruang ICCU RSD Dr. Soebandi Jember	150	132 29.058 detik	88 %	130 1 menit, 14 detik	87 %	135 54.662 detik	90 %
6	Kinerja Keuangan Berbasis Penciptaan Nilai, Makro Ekonomi Dan Dampaknya Terhadap	89	87 19.695 detik	98 %	87 43.453 detik	98 %	87 30.019 detik	98 %
7	Klasifikasi Penyakit Diabetes Dengan Hidden Naive Bayes	33	29 6.456 detik	88 %	28 13.95 detik	85 %	28 7.362 detik	85 %
8	Pencarian Data Dengan Menggunakan Fungsi Dan Metode Pada Hashing Statis Dan Hashing Dinamis	50	41 9.439 detik	82 %	41 19.415 detik	82 %	41 10.558 detik	82 %
9	Pendekatan Pengujian Regresi untuk Sistem Waktu Nyata, Terdistribusi dan Mempunyai Siklus Hidup Pendek	81	72 12.142 detik	89 %	71 31.876 detik	88 %	70 14.277 detik	86 %
10	Pengaruh Faktor Internal Dan External Diri Sumber Daya Manusia Terhadap Minat Berwirausaha (Studi Pada Mahasiswa Fakultas Ekonomi Universitas Muhammadiyah Jember	89	83 31.156 detik	93 %	82 58.393 detik	92 %	83 31.5663 detik	93 %