ANALISIS DAN IMPLEMENTASI CORPUS-BASED STEMMING DENGAN MENGGUNAKAN CO-OCCURRENCE DARI VARIASI KATA

Dimas Aryo Kunto Wibisono[1], Yanuar Firdaus A.w.[2], Angelina Prima Kurniati[3]

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

Abstrak

Di zaman globalisasi belakangan ini, informasi tentunya menjadi hal yang sangat penting bagi manusia. Dengan informasi yang dapat berupa teks, gambar, ataupun suara, manusia dapat menjawab segala macam bentuk pertanyaan yang muncul unutk memecahkan masalah yang dihadapinya.

Untuk mendapatkan informasi tersebut, manusia dapat menggunakan berbagai macam search engine seperti Google, Yahoo, Altavista, ataupun yang lainnya. Di dalam search engine tersebut terdapat suatu proses pengembalian kata dasar yang disebut stemming. Untuk menghasilkan pencarian dokumen yang akurat, tentunya proses stemming tersebut harus baik.

Banyak terdapat algoritma stemming yang telah dikembangkan, salah satunya adalah Algoritma Nazief Adriani yang merupakan algoritma stemming unutk Bahasa Indonesia. Algoritma ini adalah algoritma terbaik untuk struktur morfologi Bahasa Indonesia. Namun, di dalam pengimplementasiannya masih terdapat beberapa kekurangan, salah satunya adalah penyamarataan makna variasi kata.

Oleh karena itulah digunakan metode corpus-based stemmer yang tidak tergantung bahasa untuk menghindari penyamarataan makna variasi kata tersebut. Pada pemgimplementasiannya, metode yang menggunakan statistik co-occurrence dari variasi kata ini dapat meningkatkan akurasi dari sistem Information Retrieval.

Kata Kunci : Search Engine, Stemming, Algoritma Nazief Adriani, Corpus-Based Stemmer, Co-occurrence, Information Retrieval

Abstract

Lately, information has became very important thing for human. They can answer all kind of questions to solve their problem by information that can be in the form of text, picture, or voice. They can use all kind of search engine like Google, Yahoo, Altavista, etc. to get that information. There are some process inside that search engine. One of those process is stemmning. Stemming is a process to return original form of word variants. In order to get best return hits, the stemmer process must be good.

There are many stemming algorithm that has been developed. One of them is Nazief & Adriani Algorithm that is Indonesian stemming. This algorithm is the best algorithm for Indonesian language structure. However, there are still some shortcomings in its implementation. One of them is leveling the meaning of the word variants.

Therefore, writer uses corpus-based stemmer which language independent to avoid those problem, including leveling the meaning of the word variants. Using statistics of word variants, this method can enhance accuracy of the Information Retrieval System.

Keywords : Search Engine, Stemming, Algoritma Nazief Adriani, Corpus-Based Stemmer, Co-occurrence, Information Retrieval

1. Pendahuluan

1.1 Latar Belakang

Information Retrieval (IR) secara umum merupakan suatu teknik untuk menemukan informasi di dalam kumpulan dokumen atau di dalam media-media lainnya dengan memberikan *query* atau *keyword* berupa teks, suara, gambar atau bentuk-bentuk lainnya. Penerapannya yang paling sering dijumpai adalah *search engine* atau mesin pencari.

Untuk meningkatkan jumlah dokumen yang diperoleh salah satunya adalah dengan menggunakan pemotong kata berimbuhan (*stemmer*). *Stemmer* merupakan salah satu alat bantu paling sederhana dalam bidang IR. *Stemmer* digunakan untuk mendapatkan kata dasar atau bentuk yang lebih umum dari suatu kata sehingga mengurangi variasi kata pada dokumen-dokumen. Dengan demikian dokumen yang diinginkan akan semakin banyak diperoleh.

Sebagai contoh, dokumen yang mengandung kata-kata berimbuhan "pendapat", "pendapatan", "didapat" dan sebagainya akan dirujuk oleh satu kata dasar yang sama yaitu "dapat". Namun beberapa kata berimbuhan yang mempunyai kata dasar yang sama, memiliki makna yang berbeda, sehingga kurang tepat apabila menyamakan seluruh variasi kata tersebut kepada kata dasarnya dengan menggunakan *stemmer*. Contohnya kata "pendapat" dengan "pendapatan". Meskipun memiliki kata dasar yang sama, keduanya memiliki makna yang sangat berbeda.

Selain masalah perbedaan makna di atas, terdapat juga masalah yang terkait dengan jenis koleksi dokumen yang dapat mempengaruhi makna kata. Misalnya, kata "membintangi" dan "bintang". Pada jenis koleksi dokumen astronomi kata "membintangi" tidak mempunyai makna yang sama dengan kata "bintang". Sebaliknya pada jenis koleksi dokumen perfilman kedua kata ini bermakna sama yaitu permainan film.

Permasalahan tersebut melatarbelakangi penulis mengangkat *corpus-based stemmer*, yang selanjutnya akan disebut *stemmer* SVC (*Statistics of Variant Co-occurrence*), yang tidak tergantung bahasa untuk menghindari penyamarataan makna variasi kata khususnya pada ahasa Indonesia, karena pada Bahasa Indonesia terdapat banyak variasi kata yang berakar pada kata dasar yang sama, namun memiliki perbedaan makna. Di samping itu, penulis mencoba memperbaiki efektifitas penggunaan *stemmer* Indonesia yang sudah ada dengan menggunakan statistik *co-occurrence* dari variasi kata.

1.2 Perumusan Masalah

Dengan mengacu pada latar belakang masalah diatas, maka permasalahan yang akan dibahas dan diteliti adalah :

1. Bagaimana konsep pengumpulan seluruh bentuk kata-kata unik di dalam sebuah koleksi dokumen menjadi sebuah *equivalence class*.
2. Bagaimana pemodelan dan penyimpanan koleksi dokumen pada sistem aplikasi.
3. Bagaimana membangun aplikasi *corpus-based stemmer* yang mampu meningkatkan akurasi makna kata dengan menggunakan *co-occurrence* dari variasi kata.
4. Bagaimana performansi aplikasi terhadap tingkat akurasi makna kata yang dicapai dilihat dari parameter *recall*, *precision*, dan *F-Measure* antara *baseline stemmer* dibandingkan dengan *corpus based stemmer*.

1.3 Tujuan

Berdasarkan rumusan masalah di atas, maka tujuan dari tugas akhir ini adalah :

1. Memodelkan dan menyimpan koleksi dokumen pada sistem aplikasi.
2. Membangun aplikasi *corpus-based stemmer* yang mampu meningkatkan akurasi makna kata dengan menggunakan *co-occurrence* dari variasi kata.
3. Menganalisis performansi aplikasi terhadap tingkat akurasi makna kata yang dicapai dilihat dari parameter *recall*, *precision*, dan *F-Measure* antara *baseline stemmer* dibandingkan dengan *corpus based stemmer*.

Batasan masalah yang digunakan dalam penelitian ini antara lain :

1. Koleksi dokumen yang digunakan adalah dokumen teks berbahasa Indonesia.
2. Sebagai *baseline stemmer* digunakan Algoritma Nazief Adriani yang di dalamnya telah menggunakan kamus.
3. Aplikasi yang dirancang bersifat *stand alone* atau berdiri sendiri.

1.4 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam penyelesaian tugas akhir ini dengan langkah kerja sebagai berikut :

1. **Studi Literatur**
   Tahap ini merupakan tahap persiapan yang meliputi pengumpulan bahan-bahan pustaka dengan melakukan studi literatur sebagai referensi Tugas Akhir yang meliputi *Information Retrieval*, *corpus analysis*, Algoritma Nazief Adriani, dan topik lainnya yang mendukung penyusunan Tugas Akhir ini. Bahan pustaka ini akan digunakan sebagai dasar teori penyusunan Tugas Akhir.

2. **Pemahaman Sistem**
   Memahami sistem aplikasi yang akan dibangun yang meliputi pemodelan dan penyimpanan koleksi dokumen pada sistem aplikasi dan pemodelan lingkungan perangkat lunak.

3. **Analisis dan Perancangan Aplikasi**
   Menjabarkan *requirement*, serta analisis dan desain perangkat lunak yang akan dibangun dengan mengacu pada hasil pemahaman sistem dan studi literatur yang telah diperoleh sebelumnya. Daftar kebutuhan sistem, desain proses, desain model data, dan desain antar muka aplikasi didefinisikan pada tahap ini.

4. **Implementasi Sistem**
   Pembuatan perangkat lunak yang sesuai dengan analisis perancangan dimulai dengan pengumpulan bentuk kata-kata unik yang ada di dalam koleksi dokumen, kemudian membangun sejumlah kelompok kata-kata yang identik dengan menggunakan Algoritma Nazief Adriani, dilanjutkan dengan memperbaiki kelompok kata-kata tersebut dengan menggunakan metode *corpus-based stemming*.

5. **Pengujian Sistem**
   Pengujian perangkat lunak dilakukan dengan mengukur parameter *recall*, *precision*, dan *F-Measure* berdasarkan dokumen yang dihasilkan sistem antara Algoritma Nazief Adriani dibandingkan dengan Algoritma Nazief Adriani yang telah dimodifikasi dengan metode *corpus-based stemming*.

6. **Analisis Hasil**

   Mengevaluasi dan menganalisis tingkat keakuratan hasil informasi yang dihasilkan sistem berdasarkan nilai setiap parameter yang diperoleh dari skenario uji terhadap modifikasi Algoritma Nazief Adriani. Berdasarkan analisis yang diperoleh kemudian dirumuskan sebuah kesimpulan terhadap performa dan kinerja sistem.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari hasil pengamatan selama perancangan, implementasi, dan proses uji coba perangkat lunak yang dilakukan, dapat diambil kesimpulan sebagai berikut :

1. Sistem IR yang dibangun mampu memenuhi kebutuhan sebagai aplikasi *search engine* dengan mampu menyimpan koleksi dokumen pada sistem aplikasi.

2. Nilai rata-rata *recall* antara Metode Nazief Adriani dengan Metode Nazief Adriani – Corpus-Based Stemmer tidak mengalami penurunan. Hal ini dikarenakan dokumen relevan yang terambil oleh kedua metode tersebut selalu sama. Selain itu, dokumen relevan yang terambil oleh Metode Nazief Adriani relatif berada pada peringkat papan atas sehingga tidak memungkinkan hilangnya dokumen relevan tersebut setelah diperbaiki oleh Metode Corpus-Based Stemmer.

3. Hasil pengimplementasian Metode Corpus-Based Stemmer dengan menggunakan *co-occurrence* dari variasi kata pada Algoritma Nazief Adriani dapat memastikan terjadinya peningkatan akurasi makna kata dilihat dari meningkatnya nilai rata-rata *precision* dan *F-measure*. Dibandingkan dengan hasil yang diperoleh dengan menggunakan Algoritma Nazief Adriani, dengan bantuan Metode Corpus-Based Stemmer, terjadi peningkatan nilai rata-rata *precision* sebesar 4.66% dan *F-measure* sebesar 2.93%.

5.2 Saran

Berikut merupakan beberapa saran untuk pengembangan sistem di masa yang akan datang, berdasar pada hasil perancangan, implementasi, dan uji coba yang telah dilakukan, yaitu:

1. Pada sistem bisa ditambahkan fitur untuk menghitung waktu pencarian.

2. Pada sistem bisa ditambahkan fungsi *crawling* ke *website*, jadi aplikasinya berupa *online website*.

Referensi

[1] Amelia Putri, Yosi. 2009. *Stemming Untuk Teks Berbahasa Indonesia dan Pengaruhnya Dalam Kategorisasi*. Departemen Teknik Informatika Institut Teknologi Telkom Bandung.

[2] Asian, Jelita, dkk. 2004. *A Testbed for Indonesian Text Retrieval*.

[3] C.J. van Rijsbergen. 1979. *Information Retrieval*. Diperoleh dari : http://www.dcs.gla.ac.uk/Keith/Preface.html

[4] Fakultas Ilmu Komputer Universitas Indonesia. 2008. *Information Retrieval (Perolehan Informasi)*. Universitas Indonesia.

[5] Nakov, Preslav. 2003. *Design and Evaluation of Inflectional Stemmer for Bulgarian*.

[6] Porter, M. 1980. *Overview*.

[7] Xu, Jinxi dan W. Bruced Croft. 1994. *Corpus-Based Stemming using Co-occurrence of Word Variants.*

[8] ____, 1998. Part of Speech Tagging (Pos-Tagging). Diperoleh dari : http://www.ltg.ed.ac.uk/software/pos/ (23 Oktober 2009)

[9] ____, *Information Retrieval*. Diperoleh dari : http://en.wikipedia.com/information_retrieval/ (23 Oktober 2009)

[10] ____, *Stemming Errors*. Diperoleh dari : http://www.lancs.ac.uk/ug/oneillel/stemmer/general/stemmingerrors.htm/ (23 Oktober 2009)