ISSN 2541-1004

PENGEMBANGAN ALGORITMA STEMMING BAHASA INDONESIA DENGAN PENDEKATAN *DICTIONARY BASE STEMMING* UNTUK MENENTUKAN KATA DASAR DARI KATA YANG BERIMBUHAN

Ahmad Fikri Zulfikar
Teknik Informatika, Fakultas Teknik, Universitas Pamulang
Email: dosen00386@unpam.ac.id

ABSTRAK

Information Retrieval adalah studi tentang sistem pengindeksan, pencarian, dan mengingat data, khususnya teks atau bentuk tidak teratur lainnya. Pada proses tahap *indexing* didalam *information retrieval* terdapat proses *stemming* yaitu proses mengubah suatu kata bentukan menjadi kata dasar. Proses *stemming* sangat tergantung kepada bahasa dari kata yang akan di *stemm.* Hal ini dikarnakan, dalam melakukan proses *stemming* harus mengaplikasikan aturan morfologikal dari suatu bahasa. Bahasa Indonesia memiliki kata berimbuhan yang lebih kompleks dibandingkan dengan Bahasa Lainnya. Terdapat dua pendekatan dalam proses *stemming* yaitu: *Light Stemming* dan *Dictionary Base Stemming*. pendekatan dengan proses *Dictionary Base Stemming* dapat memberikan solusi untuk men-*stemm* kata berimbuhan dalam Bahasa Indonesia, karena menggunakan struktur morfologi untuk mengekstrak kata berimbuhan menjadi kata dasar (*root word*). pada penelitian ini akan menggunakan pendekatan proses *Dictionary Base Stemming* untuk mengembangkan Algoritma *Stemming* Bahasa Indonesia dalam menentukan kata dasar (*root word*) pada kata yang berimbuhan dengan Metode Pengembangan Sistem OOAD.

Kata kunci: Information Retrieval, Stemming, Dictionary Base, Bahasa Indonesia, dan OOAD

1. PENDAHULUAN

Untuk proses *stemming* Bahasa Indonesia ini sebelumnya sudah pernah diteliti oleh beberapa peneliti diantaranya: Algoritma *Stemming* Nazief & Andriani, Algoritma *Stemming* Arifin & Setiono dan Algortima *Stemming* Vegas (Asian & Williams, 2005, p. 2).

Dari pernyataan diatas, pendekatan dengan proses *Dictionary Base Stemming* dapat memberikan solusi untuk men-*stemm* kata berimbuhan dalam Bahasa Indonesia, karena menggunakan struktur morfologi untuk mengekstrak kata berimbuhan menjadi kata dasar (*root word*).

Pada penelitian ini penulis akan menggunakan pendekatan proses *Dictionary Base Stemming* untuk mengembangkan Algoritma *Stemming* Bahasa Indonesia dalam menentukan kata dasar (*root word*) pada kata yang berimbuhan.

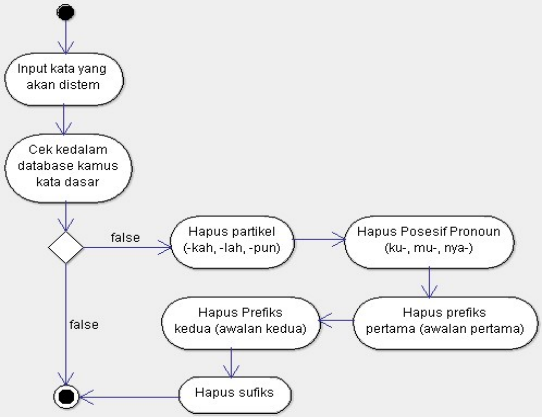
MASALAH

a. Bahasa Indonesia memiliki kata berimbuhan yang kompleks dibandingkan dengan bahasa yang lainya, karena memiliki imbuhan yang variatif seperti prefiks (awalan), sufiks (akhiran), konfiks (kombinasi awalan-akhiran), dan infiks (sisipan).
b. Belum adanya aplikasi *prototype* untuk menguji ketepatan penentuan kata dasar (*root word*) dari kata berimbuhan yang Berbahasa Indonesia.

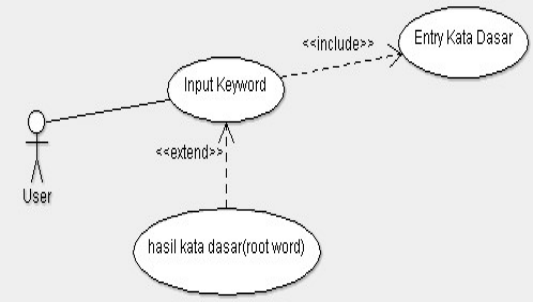
2. METODE PENELITIAN

Pada tahap ini dibahas tentang perancangan sistem *stemming* Bahasa Indonesia menggunakan pendekatan *Dictionary Base Stemming* dengan model pengembangan sistem *Object Orientied Analys Design* (OOAD) dengan notasi *Unified Modeling Language* (UML) yang terdiri dari *Activity Diagram, Use Case Diagram, Class Diagram,* dan *Sequence Diagram.*

a. Activity Diagram



b. Usecase Diagram



TEKNIK ANALISA DATA

Data yang digunakan adalah aturan morfologi imbuhan Bahasa Indonesia yang dikumpulkan penulis dari beberapa buku tentang bahasa indonesia, antara lain:

a. Buku Tata Bahasa Baku Bahasa Indonesia (Alwi, 1998, p. 78)
b. Pembentukan kata Bahasa Indonesia (Kridalaksana, 1996, pp. 98-106)

Kamus yang dugunakan dalam algoritma *stemming* ini, adalah kamus kumpulan kata dasar Bahasa Indonesia yang diambil dari kamus elektronik Bahasa Indonesia-Inggris dan kamus elektronik Bahasa Indonesia-Belanda. Dengan dibuatkan aplikasi sederhana untuk proses *stemming* Bahasa Indonesia, dapat ditentukan ketepatan dalam menentukan kata dasar dari kata yang berimbuhan dengan menggunakan pendekatan *dictionary base stemming*.

Berikut ini adalah rancangan table hasil algoritma *stemmer* tersebut:

Tabel 15 Rancangan Tabel Hasil Uji Coba *stemming*

No.	Inputan	Hapus partikel	Hapus PP	Hapus awalan 1	Hapus awalan 2	Hapus Akhiran	Kata dasar
99	x-30x	x-30x	x-30x	x-30x	x-30x	x-30x	x-30x
99	x-30x	x-30x	x-30x	x-30x	x-30x	x-30x	x-30x

Tabel 16 Rancangan Tabel Hasil Validasi *stemming*

Katagori Hasil *Stemmer*	Total	Persentase(%)
x-30-x	99	99%
x-30-x	99	99%

TEKNIK ANALISIS PENGGUNAAN METODE DALAM SISTEM

Pada teknik penggunaan metode dalam sistem ini berisi tentang metode-metode yang diterapkan dalam proses *stemming* Bahasa Indonesia untuk menentukan kata dasar dari kata yang berimbuhan (Alhanin & Juzaiddin, 2011).

a. Metode *Dictionary Base Stemming* adalah pendekatan struktur morfologi pada suatu bahasa yang terdapat aturan imbuhan untuk mengekstrak kata dasar. Pola aturan imbuhan pada Bahasa Indonesia dalam berbagai variatif seperti prefiks, sufiks, konfiks, dan infiks dengan segmentasi sebagai berikut:

Tabel 3 Kombinasi Aturan Awalan-Akhiran

Awalan	Akhiran yang tidak diizinkan
be-	-i
di-	-an
ke-	-i, -kan
me-	-an
se-	-i, -kan

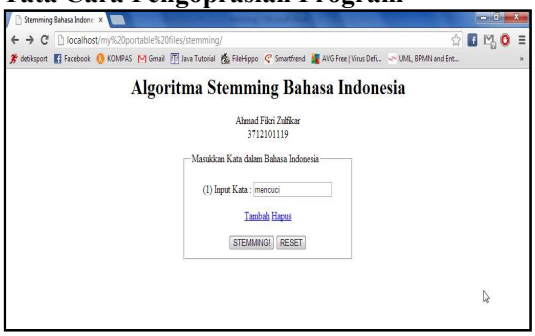
Tabel 4 Cara menentukan tipe awalan te-

Following Characters				Tipe Awalan
Set 1	Set 2	Set 3	Set 4	
"-r-"	"-r-"	-	-	None
"-r-"		-	-	ter-luluh
"-r-"	not (vowel or "-r-")	"-er-"	vowel	Ter
"-r-"	not (vowel or "-r-")	"-er-"	not vowel	ter-
"-r-"	not (vowel or "-r-")	not "-er-"	-	Ter
not (vowel or "-r-")	"-er-"	Vowel	-	None
not (vowel or "-r-")	"-er-"	not vowel	-	Te

Tabel 5 Jenis Awalan berdasarkan tipe awalannya

Tipe Awalan	Awalan yang harus dihapus
di-	di-
ke-	ke-
se-	se-
te-	te-
ter-	ter-
ter-luluh	Ter

Tata Cara Pengoprasian Program



3. PEMBAHASAN

Hasil table uji coba dari 30 kata yang berimbuhan yang diproses oleh algoritma *Stemming* Bahasa Indonesia dengan pendekatan *Dictionary Base Stemming* ditunjukkan pada Tabel 6.

4. KESIMPULAN

Berdasarkan hasil penelitian dan pembahasan yang disampaikan pada bagian sebelumnya, maka dapat ditarik beberapa kesimpulan seperti berikut:

a. Implementasi dengan pendekatan *Dictionary Base Stemming* Bahasa Indonesia untuk menentukan kata dasar (*root word*) dari kata berimbuhan yang bervariatif seperti: prefiks, sufiks, infiks, dan konfiks, cukup memuaskan dilihat dari hasil uji coba dengan 30 sempel kata berimbuhan Bahasa Indonesia yang sudah ditentukan oleh penulis, dimana katagori dari hasil *stemmer Exact Match* nilai presentasenya sebesar 93.3 % (persen), sedangkan katagori hasil *stemmer Unchange* nilai presentasenya mencapai 6.7% (persen) dikarnakan ketidaktepatan dalam melakukan proses pemenggalan kata yang berulang-ulang, dan katagori hasil *stemmer Spelling Exception* nilai presentasenya 0% (persen).

b. Analisis dan desain sistem pada proses Algoritma *Stemming* Bahasa Indonesia ini dilakukan dengan menggunakan metodologi *OOAD* yang dijabarkan menjadi empat tahap yaitu definisi kebutuhan, perancangan sistem, implementasi serta integrasi dan pengujian sistem. Pada tahap definisi kebutuhan Algoritma *Stemming* Bahasa Indonesia dibutuhkan dan diaplikasikan dibidang *information retrieval system* (infromasi sistem temu kembali) dan komputasi *linguistic*. Pada tahap perancangan ini menggunakan model diagram UML (*Unified Modeling Language*) yang terdiri dari *activity diagram, use case diagram, class diagram, sequence diagram*, *deployment diagram* dan *component diagram*. Tahap implementasi, Sistem untuk proses *stemming* ini berbasis web dengan

menggunakan bahasa PHP dan *database* MySQL. Untuk tahap terakhir yaitu integrasi dan pengujian sistem, sistem yang dibangun diuji dengan menggunakan metode pengujian *White Box* dan *Black Box*.

Tabel 6 Hasil Tabel Uji Coba

	Inputan	Hapus partikel	Hapus PP	Hapus prefiks 1	Hapus prefiks 2	Hapus sufiks	Kata dasar	Katagori hasil
1	Merupkan	Merupakan	Merupakan	Rupkan	Rupakan	Rupa	Rupa	Exact Match
2	Menyapu	Menyapu	Menyapu	Sapu	Sapu	Sapu	Sapu	Exact Match
3	Bermain	Bermain	Bermain	Bermain	Main	Main	Main	Exact Match
4	Permainan	Permainan	Permainan	Permainan	Mainan	Main	Main	Exact Match
5	Menyiapkan	Menyiapkan	Menyiapkan	Siapkan	Siapkan	Siap	Siap	Exact Match
6	Mencuci	Mencuci	Mencuci	Cuci	Cuci	Cuci	Cuci	Exact Match
7	Belajar	Belajar	Belajar	Belajar	Ajar	Ajar	Ajar	Exact Match
8	Pelari	Pelari	Pelari	Pelari	Lari	Lari	Lari	Exact Match
9	Mencintai	Mencitai	Mencintai	Cintai	Cintai	Cinta	Cinta	Exact Match
10	Penyayang	Penyayang	Penyayang	Sayang	Sayang	Sayang	Sayang	Exact Match
11	Laki-laki	Laki-laki	Laki-laki	Laki-laki	Laki-laki	Laki-lak	Laki-lak	Unchange
12	Mempunyai	Mempunyai	Mempunyai	Punyai	Punyai	Punya	Punya	Exact Match
13	Melahirkan	Melahirkan	Melahirkan	Lahirkan	Lahirkan	Lahir	Lahir	Exact Match
14	Melupakan	Melupakan	Melupakan	Lupakan	Lupakan	Lupa	Lupa	Exact Match
15	Melakukan	Melakukan	Melakukan	Lakukan	Lakukan	Laku	Laku	Exact Match
16	Melukai	Melukai	Melukai	Lukai	Lukai	Luka	Luka	Exact Match
17	Terlupakan	Terlupakan	Terlupakan	Lupakan	Lupakan	Lupa	Lupa	Exact Match
18	Terlewatkan	Terlewatkan	Terlewatkan	Lewatkan	Lewatkan	Lewat	Lewat	Exact Match
19	Berkelahi	Berkelahi	Berkelahi	Berkelahi	Kelahi	Kelahi	Kelahi	Exact Match
20	Mewarnai	Mewarnai	Mewarnai	Wrnai	Warnai	Warna	Warna	Exact Match
21	Menentukan	Menetukan	Menentukan	Tentukan	Tentukan	Tentu	Tentu	Exact Match
22	Menentang	Menentang	Menentang	Tentang	Tentang	Tentang	Tentang	Exact Match
23	Bernyanyi	Bernyanyi	Bernyanyi	Bernyanyi	Nyanyi	Nyanyi	Nyanyi	Exact Match
24	Mencaci	Mencaci	Mencaci	Caci	Caci	Caci	Caci	Exact Match
25	Berlabuh	Berlabuh	Berlabuh	Berlabuh	Labuh	Labuh	Labuh	Exact Match
26	Bertikai	Bertikai	Bertikai	Bertikai	Tikai	Tikai	Tikai	Exact Match
27	Menjuarai	Menjuarai	Menjuarai	Juarai	Juarai	Juara	Juara	Exact Match
28	Majalah	Majalah	Majalah	Majalah	Majalah	Majalah	Majalah	Exact Match
29	Abu-abu	Abu-abu	Abu-abu	Abu-abu	Abu-abu	Abu-abu	Abu-abu	Unchange
30	Masalah	Masalah	Masalah	Masalah	Masalah	Masalah	Masalah	Exact Match

DAFTAR PUSTAKA

Adriani, M., Nazief, B., Asian, J., & Williams, H. E. (2007). Stemming Indonesia A Confixs Stripping Approach. *ACMTransactions on Asian Language Information Processing, Vol. 6, No. 4*, 13.

Agusta, L. (2009). Comparasi Algoritma Stemming Porter dengan Algoritma Nazief dan Andriani untuk Stemming Dokumen Teks Bahasa Indonesia. *Konfrensi Nasional Sistem dan Informatika.* Bali.

Alhanin, Y., & Juzaiddin, M. (2011). The Enhancement of Arabic Stemming by Using Light Stemming and Dictionary-Based Stemming. *Journal of Software Engineering and Applications*.

Alwi, H. (1998). *Tata Bahasa Baku Bahasa Indonesia.* Balai Pustaka.

Arifin, A. Z., & Setiono, A. N. (2002). Klasifikasi Dokumen Berita Kejadian Berbahasa Indoensia dengan Algoritma Single Pass Clustering. *www.its.ac.id*.

Asian, J., & Williams, H. E. (2005). Stemming Indonesia. *Australia Computer Science Conference.*

Bernstein, A., & Kiefer, C. (2005). Imprecise Queries Using Similarity Joins For Retrieval in Ontologis. *iRDQL.* Winterthurerstr: University Of Zurich.

Fachrurrozi. (2006). *Algoritma dan Pemrograman I.*

Garfield, E. (1997). A Tribute To Calvin N. Mooers, A Pioneer Of Information Retrieval. *The Scientist*.

Kridalaksana, H. (1996). *Pembentukan Kata Dalam Bahasa Indonesia.* Gramedia.

Mandala, R., & Munir, R. (2004). Sistem Stemming Otomatis untuk kata dalam Bahasa Indonesia. *Seminar Nasional Aplikasi Teknologi Informasi.* Yogyakarta.

O'Docherty, M. (2005). *Object Oriented Anlayst and Design* .

Pressman, R. (2010). *Software Engineering : A Practitioner's Approach.* Mc Grew-Hill Companies.

Prof. DR. Sugiyono. (2012). *Metode Penelitian Kuantitatif, Kualitatif dan R&D.* Alfabeta.

Sommerville, I. (2011). *Software Engineering.* Pearson Education.