

STUDI PERBANDINGAN ALGORITMA - ALGORITMA *STEMMING* UNTUK DOKUMEN TEKS BAHASA INDONESIA

Manase Sahat H Simarankir

Program Studi Teknik Komputer, Politeknik META Industri Cikarang
Cikarang TechnoPark Building Jalan Inti 1 Blok C1 No.7 Lippo Cikarang Bekasi
Email : manasemalo@politeknikmeta.ac.id

ABSTRAK

Stemming merupakan proses untuk memetakan berbagai variasi morfologikal dari kata menjadi bentuk dasar yang sama, dengan menghilangkan semua imbuhan baik yang terdiri dari awalan, sisipan, akhiran dan kombinasi dari awalan dan akhiran pada kata berimbuhan. Pada penelitian ini digunakan dua algoritma *stemming* berbasis kamus dan dua algoritma *stemming* menggunakan aturan imbuhan. Pengujian dilakukan menggunakan 100 dokumen teks Bahasa Indonesia yang sudah ditentukan sebelumnya. Hasil pengujian yang dilakukan menunjukkan bahwa waktu proses *stemming* yang paling cepat terdapat pada algoritma Vega, akurasi yang paling tinggi terdapat pada algoritma Nazief dan Adriani, *overstemming* yang paling sedikit terdapat pada algoritma Nazief dan Adriani karena nilainya lebih rendah, *understemming* yang paling sedikit terdapat pada algoritma Nazief dan Adriani karena nilainya lebih rendah. Dari pengujian yang dilakukan menunjukkan bahwa algoritma yang paling baik terdapat pada algoritma Nazief dan Adriani.

Kata Kunci: *stemming*, *dictionary based*, *purely rule based*, *overstemming* dan *understemming*.

1. PENDAHULUAN

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. Bertambahnya jumlah dokumen teks yang dapat diakses di internet diikuti dengan meningkatnya kebutuhan pengguna akan perangkat pencarian informasi yang efektif dan efisien (Agusta, 2009). Efektif berarti user mendapatkan dokumen yang relevan dengan *query* yang diinputkan, efisien berarti waktu pencarian yang sesingkat-singkatnya.

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata. *Stemming* merupakan proses yang memetakan bentuk varian kata menjadi kata dasarnya (Tala, 2004). 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. Untuk *stemming* dalam bahasa Indonesia, terdapat dua jenis metode *stemming* yang sudah ada, yaitu algoritma *stemming* yang berbasis kamus (*dictionary based*) dan algoritma *stemming* yang berbasis non-kamus (*purely rule based*). Dalam *stemming* Bahasa Indonesia, digunakan beberapa teknik pendekatan seperti *stemming* yang dilakukan oleh *stemmer* Tala, Porter, Arifin dan Setiono, Nazief dan Adriani, Vega. Sebenarnya hampir tidak ada persetujuan umum mengenai keefektifan dari teknik-teknik pendekatan tersebut. Masalah lain dari *stemming* adalah ketergantungan dari beberapa teknik *stemming* tersebut pada kamus yang luas (*comprehensive dictionary*).

2. TINJAUAN PUSTAKA

2.1 Imbuhan Bahasa Indonesia

Bahasa Indonesia adalah bahasa yang memiliki morfologi yang berbeda dengan bahasa lainnya. Bahasa Indonesia memiliki imbuhan-imbuhan yang beraneka ragam. Sering kali sebuah kata dasar atau bentuk dasar perlu diberi imbuhan untuk dapat digunakan dalam pertuturan (Chaer A, 2011). Imbuhan ini dapat mengubah makna, jenis dan fungsi sebuah kata dasar atau bentuk dasar menjadi kata lain yang fungsinya berbeda dengan kata dasar atau bentuk dasarnya. Imbuhan mana yang harus digunakan tergantung pada keperluan penggunaannya di dalam pertuturan. Imbuhan yang ada dalam bahasa Indonesia adalah :

- a. Partikel (*Particle*) -lah, -kah, -pun.
Contoh : pergi + lah = pergilah, buku + kah = bukukah, buku + pun = bukupun.
- b. Kata ganti kepunyaan (*Possive Pronoun*) -ku, -mu, -nya.
Contoh : buku + mu = bukumu, buku + ku = bukuku, buku + nya = bukunya.
- c. Awalan (*Prefix*) ber-, per-, me-, di-, ter-, ke-, se-, dan pe-.
Contoh : ber + main = bermain, me + lempar = melempar, di + makan = dimakan, ter + ambil = terambil, ke + pasar = kepasar, se + banyak = sebanyak, pe + malu = pemalu.
- d. Akhiran (*Suffix*) -kan, -an, -i, dan -nya.
Contoh : dengar + kan = dengarkan, akhir + i = akhiri, bentuk + nya = bentuknya
- e. Sisipan (*Infix*) -el-, -em, dan -er-.
Contoh : telunjuk, gemuruh, gerigi.
- f. Imbuhan gabung (*Confix*) ber-kan, ber-an, per-kan, per-i, me-kan, me-i, memper-, memper-kan, memper-i, di-kan, di-i, diper-, diper-kan, diper-i.
Contoh : ber + dasar + kan = berdasarkan, ber + datang + an = berdatangan, per + satu + kan = persatukan, memper + tanya + kan = mempertanyakan.

2.2 Stemming

Stemming merupakan suatu proses yang terdapat dalam sistem IR (*Information retrieval*) yang mentransformasi kata-kata yang terdapat dalam suatu dokumen ke kata-kata akarnya (*root word*) dengan menggunakan aturan-aturan tertentu (Agusta, 2009). *Stemming* adalah proses pemetaan dan penguraian berbagai bentuk (*variants*) dari suatu kata menjadi bentuk kata dasarnya (*stem*). Proses *stemming* secara luas sudah digunakan di dalam *Information retrieval* (pencarian informasi) untuk meningkatkan kualitas informasi yang didapatkan. Kualitas informasi yang dimaksud misalnya untuk mendapatkan hubungan antara varian kata yang satu dengan yang lainnya. Sebagai contoh kata “diculik”, “menculik” (melakukan tindakan menculik) dan “penculik” (orang yang menculik) yang semula mengandung arti yang berbeda dapat distem menjadi sebuah kata “culik” yang memiliki arti yang sama sehingga kata-kata diatas saling berhubungan. Selain itu *stemming* juga dapat digunakan untuk mengurangi ukuran dari suatu ukuran *index file*. Misalnya dalam suatu deskripsi terdapat varian kata “memberikan”, “diberikan”, “memberi” dan “diberi” hanya memiliki akar kata (*stem*) yaitu “beri”. Proses *stemming* pada teks berbahasa Indonesia berbeda dengan *stemming* pada teks berbahasa Inggris. Pada teks berbahasa Inggris, proses yang diperlukan hanya proses menghilangkan sufiks (akhiran). Sedangkan pada teks berbahasa Indonesia, selain sufiks, prefiks, dan konfiks juga dihilangkan.

2.3 Penelitian Terkait

- a. Penelitian yang dilakukan oleh Ledy Agusta (Agusta, 2009) dengan judul: “Perbandingan algoritma *stemming* Porter dengan algoritma Nazief dan Adriani Untuk *stemming* Dokumen Teks Bahasa Indonesia”. Penelitian ini membandingkan efektifitas dari algoritma *stemming* Porter dengan Algoritma Nazief dan Adriani. Untuk membandingkan performa masing-masing algoritma *stemming*, dibuat program sederhana untuk memproses dokumen teks inputan sehingga diketahui hasil *stem*, waktu proses, presisi dari hasil *stemming* dokumen tersebut. Uji Coba algoritma dilakukan pada 30 dokumen teks Bahasa Indonesia dengan ukuran dokumen yang bervariasi. Hasil uji coba dokumen teks yang dilakukan menunjukkan bahwa :
 - Proses *stemming* dokumen teks berbahasa Indonesia menggunakan Algoritma Porter membutuhkan waktu yang lebih singkat dibandingkan dengan *stemming* menggunakan Algoritma Nazief dan Adriani.
 - Proses *stemming* dokumen teks berbahasa Indonesia menggunakan Algoritma Porter memiliki presentase keakuratan (presisi) lebih kecil dibandingkan dengan *stemming* menggunakan Algoritma Nazief dan Adriani.
- b. Penelitian yang dilakukan oleh Mardi Siswo Utomo (Utomo, 2013) dengan judul: ”Implementasi Stemmer Tala pada Aplikasi berbasis web”. Pada penelitian ini dilakukan evaluasi pada 1000 kata terbanyak pada algoritma *stemming* Tala. Kata-kata tersebut telah bebas dari stopword/stopwords. Kesalahan karena *overstemming* paling banyak terjadi, dari 1000 kata terdapat 177 kata yang salah karena *overstemming*. Kemudian untuk bahasa asing tidak terjadi perubahan karena akhiran dan

awalan tidak dikenali oleh sistem. Selanjutnya kesalahan juga terjadi pada nama orang/istilah/singkatan.

Kebanyakan kata dengan akhiran 'i' akan terpotong oleh sistem huruf 'i' terakhirnya, karena tidak ada mekanisme pendeteksi apakah 'i' tersebut akhiran atau bagian dari kata. Kesalahan juga terjadi pada kesalahan ketik / masukan kata, susunan imbuhan yang salah atau imbuhan asing secara statistik, dari 1000 kata tersebut terdapat 256 kata yang mengalami kesalahan pencarian kata dasar, sehingga tingkat keberhasilannya adalah 74,4 %.

- c. Penelitian yang dilakukan oleh Muhammad Latif Rozi (Rozi, 2013) dengan judul: "Implementasi dan Analisis Perbandingan Algoritma *Stemming* Nazief dan Adriani dengan Algoritma *Stemming* Vega dalam *Information Retrieval System*". Untuk mengetahui hasil dari algoritma *stemming* Nazief dan Adriani dan algoritma *stemming* Vega, maka dibuat *information retrieval system* yang dapat memperlihatkan performansi yang dihasilkan kedua algoritma tersebut. Pengujian dilakukan dengan menggunakan parameter waktu, akurasi, *Word Conflation Class*, dan *Index Compression Factor*.

Word Conflation Class adalah ukuran rata-rata kumpulan kata yang sudah melalui proses *stemming*, terlepas kata dasar hasil proses *stemming* benar atau salah. *Index Compression Factor* adalah seberapa besar jumlah kata-kata unik didalam dokumen dapat tereduksi dengan adanya proses *stemming*. Kemudian kedua algoritma *stemming* tersebut akan diterapkan di dalam *Information Retrieval System*. Berdasarkan pengujian yang dilakukan, maka diperoleh hasil sebagai berikut :

- Algoritma Nazief dan Adriani memiliki waktu proses yang lebih besar daripada algoritma Vega. Rata-rata waktu proses *stemming* pada Algoritma Nazief dan Adriani sebesar 2,87 detik sedangkan rata-rata waktu proses Algoritma Vega sebesar 0,04 detik.
- Algoritma Nazief dan Adriani memiliki akurasi yang lebih besar dari pada algoritma Vega. Rata-rata akurasi Algoritma Nazief dan Adriani sebesar 82,84% sedangkan rata-rata akurasi Algoritma Vega sebesar 60,37%.
- Algoritma Vega memiliki nilai WCC dan ICF yang relatif lebih besar daripada algoritma Nazief dan Adriani. Perbedaan nilai WCC dan ICF antara algoritma Nazief dan Adriani dengan algoritma Vega tidak terlalu besar. Rata-rata nilai WCC dan ICF Algoritma Nazief dan Adriani masing-masing sebesar 1,06 dan 0,057. Sedangkan nilai WCC dan ICF Algoritma Vega masing-masing sebesar 1,064 dan 0,06.
- Untuk *query* yang menjurus ke istilah dalam bidang tertentu, hasil *Information Retrieval System* antara algoritma Nazief dan Adriani dan algoritma Vega relatif sama.

3. METODE PENELITIAN

3.1 Pengumpulan Data

Data yang mendukung dalam penelitian ini diambil dari kamus besar bahasa Indonesia. Data juga diperoleh dengan membaca dan mempelajari referensi maupun dokumen yang berhubungan langsung dengan permasalahan dalam penelitian juga koleksi dokumen yang digunakan.

3.2 Metode yang diusulkan

Metode yang diusulkan untuk penelitian ini menggunakan pendekatan algoritma *stemming* berbasis kamus dan algoritma *stemming* menggunakan aturan imbuhan untuk dokumen teks Bahasa Indonesia.

3.3 Tahap *preprocessing*

Langkah *preprocessing* ini digunakan untuk menghilangkan bagian-bagian yang tidak diperlukan yang terdapat pada sebuah dokumen dimana hal ini akan menjadi *noise* pada proses selanjutnya, selain itu langkah *preprocessing* ini sendiri berfungsi sebagai parameter *input* algoritma *stemming*.

3.4 Algoritma *stemming*

Dalam penelitian ini akan digunakan algoritma *stemming* Nazief dan Adriani, algoritma Arifin dan Setiono, algoritma Vega dan algoritma Tala untuk mendapatkan kata dasar dalam bahasa Indonesia. Masing-masing algoritma *stemming* akan diujikan dengan dataset yang sama dan hasilnya akan dievaluasi.

3.5 Eksperimen dan pengujian dataset

Untuk memudahkan dalam melakukan pengujian maka dilakukan testing model terhadap aplikasi *stemming* yang sudah dibuat sebelumnya. Selanjutnya akan dilakukan evaluasi terhadap hasil *stemming*. Evaluasi dilakukan untuk mengetahui waktu proses *stemming* nilai akurasi dari masing-masing algoritma dan juga kata-kata yang mengalami *overstemming* dan *understemming*. Adapun tahapan pengujian yang dilakukan dapat dilihat pada gambar 1 dan gambar 2.

Gambar 1. Halaman *input* dokumen pengujian

No	Inputan Kata	Kata Dasar
1	analisis	analisis
2	berorientasi	orientasi
3	objek	objek
4	analisis	analisis
5	berorientasi	orientasi
6	objek	objek
7	menurut	turut
8	whitten	whitten
9	et	et
10	al	al
11	adalah	ada
12	pendekatan	dekat
13	pemodelan	podel
14	objek	objek
15	selama	selama
16	analisis	analisis
17	dan	dan
18	desain	desain
19	sistem	sistem

Jumlah Kata : 19
Waktu : 0.014 detik

Gambar 2. Halaman *output* hasil pengujian

4. HASIL DAN PEMBAHASAN

4.1 Koleksi Dokumen

Koleksi dokumen yang digunakan untuk pengujian adalah dokumen sampel sebanyak 100 dokumen yang sudah dibuat dengan format berekstensi *.txt*. Kata-kata dalam dokumen tersebut diperoleh dari isi artikel-artikel dan berita, baik artikel tentang kesehatan, artikel teknik, artikel keilmuan maupun berita-berita dari media elektronik. Total kata keseluruhan pada 100 dokumen adalah 25.937 kata.

4.2 Kamus

Pada penelitian ini, kamus kata dasar diambil dari daftar kata dasar pada Kamus Besar Bahasa Indonesia (KBBI) luring CHM V1.5 yang diunduh dari *ebsoft.web.id*. Total kata dasar dalam kamus adalah sebanyak 31.590 kata dasar.

4.3 Relevance judgments

Pada penelitian ini, sebelum dilakukan evaluasi tentang performansi hasil *stemming* baik waktu proses, akurasi kata maupun kata-kata yang mengalami *overstemming* dan *understemming* dalam setiap dokumen maka dibuat *relevance judgments* per kata untuk melihat seberapa relevannya hasil *stemming* terhadap *relevance judgments* yang dibuat secara pengetahuan manusia. *Relevance judgments* yang dibuat adalah berupa kata dasar saja seperti pada Tabel 1.

Tabel 1. Relevance judgments pada dokumen hasil stemming

No	Inputan Kata	Hasil Stemming	Relevance judgments (kata dasar)
1	judul	judul	judul
2	atmosfer	atmosfer	atmosfer
3	bumi	bumi	bumi
4	akurasi	akurasi	akurasi
5	terperiksa	periksa	periksa
6	melingkupi	lingkup	lingkup
7	adalah	adalah	adalah
8	lapisan	lapis	lapis

4.4 Evaluasi Pengukuran

Evaluasi pengukuran terhadap hasil proses *stemming* dilakukan secara detail mulai dari dokumen pertama sampai dokumen yang terakhir. Masing-masing algoritma *stemming* akan diuji menggunakan 100 dokumen. Adapun pengukuran yang dilakukan adalah sebagai berikut :

- a. Waktu proses
Waktu proses hasil *stemming* akan ditampilkan dalam satuan detik.
- b. Akurasi
$$\text{Akurasi} = \frac{(\text{Kata dasar hasil stemming} = \text{Relevance judgments})}{\text{Jumlah kata dalam dokumen}} \times 100\%$$
- c. *Overstemming* adalah kata yang terlalu banyak dipotong setelah dilakukan proses *stemming* dibandingkan dengan jumlah kata dalam dokumen.
- d. *Understemming* adalah kata yang terlalu sedikit dipotong setelah dilakukan proses *stemming* dibandingkan dengan jumlah kata dalam dokumen.

4.5 Hasil Pengujian

a. Analisa Hasil pengujian

Hasil pengujian 100 dokumen terhadap masing-masing algoritma *stemming* dapat dilihat pada Tabel 2 berikut ini:

Tabel 2. Hasil rata-rata pengujian algoritma stemming

Algoritma	Waktu Proses (detik)	Akurasi (%)	Overstemming (%)	Understemming (%)
Nazief & Adriani	5,147	97,931	1,400	0,397
Arifin & Setiono	15,204	92,099	6,102	1,561
Vega	0,085	63,486	30,332	6,061
Tala	0,22	78,274	19,085	1,706

Dari tabel hasil pengujian diatas terlihat bahwa rata-rata waktu proses *stemming* yang paling singkat terdapat pada algoritma *stemming* Vega hal ini dikarenakan tidak ada pengecekan kata dasar terhadap kamus. Selanjutnya rata-rata nilai akurasi hasil *stemming* yang paling besar terdapat pada algoritma *stemming* Nazief dan Adriani, hal ini dikarenakan pada algoritma tersebut menggunakan kamus kata dasar dan algoritmanya memiliki aturan morfologi imbuhan yang diperbolehkan dan yang tidak diperbolehkan. Untuk *Overstemming* dan *Understemming* yang paling kecil terdapat pada algoritma *stemming* Nazief dan Adriani.

b. Analisa kesalahan hasil *stemming*

Evaluasi kesalahan hasil *stemming* pada setiap algoritma berbeda-beda, kesalahan terjadi karena masing-masing algoritma memiliki karakteristik yang berbeda ketika sebuah kata dilakukan proses *stemming*. Kesalahan tidak hanya terjadi pada kata dasar atau kata berimbuhan saja, tetapi kesalahan juga disebabkan karena salah pengetikan kata dalam dokumen yang digunakan. Selain itu kesalahan juga terjadi pada nama tempat, nama orang dan istilah juga bahasa asing baik pada algoritma yang menggunakan kamus maupun pada algoritma yang menggunakan aturan imbuhan. Berikut kesalahan hasil *stemming* karena pengetikan kata yang salah dalam dokumen.

Tabel 3. Kesalahan hasil *stemming* karena kesalahan pengetikan kata

Contoh	Seharusnya	Hasil <i>stemming</i>	Kata dasar
adaalah	adalah	adaalah	adalah
bartahun	bertahun	bartahun	tahun
diataati	ditaati	diataati	taat
denganya	dengannya	denga	dengan
dikembamngkan	dikembangkan	dikembamngkan	kembang
tetang	tentang	tang	tentang
kuargaku	keluargaku	kuarga	keluarga
lagkah	langkah	lag	langkah
tebuat	terbuat	tebuat	buat
inginya	inginnya	ingi	ingin
kemajuanm	kemajuan	kemajuanm	maju
kemorosotan	kemerosotan	kemorosotan	merosot
menganggab	menganggap	menganggab	anggap

Tabel 4. Kesalahan hasil *stemming* terhadap nama orang, tempat dan istilah

Contoh	Hasil <i>stemming</i>	Seharusnya
asian	asi	tetap
bali	bal	tetap
bekasi	bekas	tetap
lan	l	tetap
yadi	yad	tetap
diego	ego	tetap
abdullah	abdul	tetap
bayi	bay	tetap
bakri	bakr	tetap
betawi	betaw	tetap
bisri	bisr	tetap
budhiman	budhim	tetap
dahlan	dahl	tetap
deputi	deput	tetap
diet	et	tetap
dki	dk	tetap
edi	ed	tetap

Tabel 5. Kesalahan hasil *stemming* terhadap bahasa asing

Contoh	Hasil <i>stemming</i>	Seharusnya
fireman	firem	tetap
members	s	tetap
medicine	cine	tetap
megaphone	gaphone	tetap

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan pembahasan yang disampaikan pada bab sebelumnya, maka terdapat beberapa kesimpulan sebagai berikut:

1. Untuk algoritma yang menggunakan kamus ditemukan algoritma terbaik dalam proses *stemming* yaitu algoritma Nazief dan Adriani. Hal ini dikarenakan pada algoritma Nazief dan Adriani terdapat penambahan aturan-aturan untuk reduplikasi, penambahan aturan untuk awalan dan akhiran dalam meningkatkan presisi dari setiap kata yang *distemming*.
2. Untuk algoritma yang menggunakan aturan imbuhan ditemukan algoritma terbaik dalam proses *stemming* terdapat pada algoritma Tala. Hal ini dikarenakan dalam algoritma Tala ada aturan *cluster* penggunaan imbuhan.
3. Kesalahan hasil *stemming* terhadap nama tempat, nama orang dan istilah maupun pengetikan kata yang salah dalam dokumen tidak terlalu mempengaruhi peningkatan akurasi. Mengingat jumlah kata yang salah karena hal tersebut tidak terlalu banyak dibanding dengan jumlah total kata keseluruhan.

Ada beberapa hal yang perlu disampaikan agar menjadi bahan pertimbangan untuk melengkapi penelitian dimasa mendatang. Hal-hal tersebut diungkapkan dalam bentuk saran sebagai berikut:

1. Pengujian dilakukan lagi pada jumlah kata yang lebih besar dalam sebuah dokumen seperti isi skripsi, tesis maupun disertasi yang berbahasa Indonesia.
2. Untuk mengurangi tingkat kesalahan perlu dilakukan modifikasi pada algoritma dengan menyempurnakan aturan pemenggalan imbuhan yang digunakan pada setiap algoritma, baik penggunaan partikel, awalan, akhiran maupun sisipan.

6. UCAPAN TERIMAKASIH

Penelitian yang dilakukan oleh penulis di Laboratorium Teknik Komputer Politeknik META Industri Cikarang ini didanai oleh Politeknik Meta Industri Cikarang, Oleh karena itu penulis mengucapkan terimakasih.

7. DAFTAR PUSTAKA

- Agusta, L 2009. 'Perbandingan Algoritma Stemming Porter dengan Algoritma Nazief & Adriani untuk Stemming Dokumen Teks Bahasa Indonesia'. *Konferensi Nasional Sistem dan Informatika*. KNS&I09-036.
- Arifin, A, Z & Setiono, A, N 2002. 'Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering'. Institut Teknologi Sepuluh November (ITS). Surabaya. www.its.ac.id.
- Chaer, A 2011. *Tata Bahasa Praktis Bahasa Indonesia*, Rineka Cipta Jakarta. Indonesia.
- Kamus Besar Bahasa Indonesia (KBBI) luring CHM V1.5, *ebsoft.web.id*. diakses Rabu 25 November 2015.
- Nazief, B & Adriani, M 1996. 'Confix-Stripping: Approach to Stemming Algorithm for Bahasa Indonesia'. Faculty of Computer Science. University of Indonesia.
- Rozi, M, L 2013. 'Implementasi dan Analisis Perbandingan Algoritma Stemming Nazief dan Adriani dengan Algoritma Stemming Vega dalam Information Retrieval System'. Telkom University, Bandung.
- Tala, F, Z 2004. 'A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia', Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands.
- Utomo, M, S 2013. 'Implementasi stemmer Tala pada Aplikasi Berbasis Web', *Jurnal Teknologi Informasi DINAMIK*, Volume 18, No.1.
- Vega, V, B 2001. 'Information Retrieval for the Indonesian Language'. *Master's thesis*, National University Singapore.