Analisis Dan Perbandingan Stemming Teks Bahasa Indonesia Menggunakan Algoritma Vega Dan Algoritma Confix-Stripping

Helmi Agustian[1], Tati Harihayati[2]

Teknik Informatika – Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung

Email : helmi_mail@yahoo.com[1], tharihayati@yahoo.com[2]

ABSTRAK

Algoritma *stemming* merupakan sebuah algoritma yang dapat mengubah kata berimbuhan ke bentuk kata dasarnya. Penerapan algoritma stemming salah satunya pada sistem pencarian dokumen, dengan meningkatkan kemampuan pada segi *recall* dan *index compression.* Pengukuran parameter-parameter *stemmer strength* dapat mewakili kemampuan algoritma *stemming* dalam hal *recall* dan *index compression* [2]. Algoritma vega berdasarkan pengukuran parameter-parameter *stemmer strength* yaitu ukuran rata-rata *word per conflation class*, *index compression factor*, jumlah kata yang berbeda antara kata masukan dan *stem*, rata-rata huruf yang dihilangkan untuk membentuk stem, dan terakhir nilai *median* dan *mean modified hamming distance* antara kata masukan dan *stem*, terhadap dokumen-dokumen sample uji menunjukan kecenderungan untuk selalu mengubah kata berimbuhan lebih sering dari pada algoritma confix-stripping. Sedangkan algoritma confix-stripping menunjukan kecenderungan untuk mengubah beberapa kata berimbuhan menjadi satu kata dasar yang sama lebih banyak dari pada algoritma vega.

Kata Kunci : *Stemming, Vega, Confix-stripping*

1. PENDAHULUAN

Stemming merupakan suatu proses untuk menemukan kata dasar dari sebuah kata bentukan atau kata berimbuhan. Dengan menghilangkan semua imbuhan baik yang terdiri dari awalan, sisipan, akhiran dan kombinasi awalan akhiran pada kata turunan. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi Bahasa Indonesia yang baik dan benar. Stemming adalah alat dasar pemrosesan teks yang digunakan untuk *text retrieval* secara efisien dan efektif.

Imbuhan pada Bahasa Indonesia lebih kompleks bila dibandingkan dengan imbuhan pada Bahasa Inggris. Karena seperti yang telah disebutkan di atas bahwa imbuhan pada Bahasa Indonesia terdiri dari awalan, sisipan, akhiran, bentuk perulangan dan kombinasi awalan akhiran. Imbuhan-imbuhan yang melekat pada suatu kata harus dihilangkan untuk mengubah bentuk kata tersebut menjadi bentuk kata dasarnya.

Proses *stemming* biasanya juga disebut *conflation*, yang digunakan pada *search engine* untuk ekspresi *query* dan *indexing* dan untuk *permasalahan natural language processing*. Proses *conflation* dapat dilakukan secara manual maupun otomatis. Program untuk *conflation* otomatis ini disebut sebagai "*stemmer*". *Stemmer* digunakan di *information retrieval system* untuk mengurangi ukuran file indeks. Term bisa di-*stemming* pada saat *indexing* maupun *searching*. Keuntungan dari *stemming* saat *indexing* adalah effisiensi dan kompresi file indeks. Sehingga bisa dikatakan dengan *stemming* mampu meningkatkan performansi *information retrieval system*.

Ada beberapa cara yang dilakukan *stemmer* untuk menjalankan proses *stemming* diantaranya: Table look-up, affix removal stemmers, successor variety dan n-gram stemmer. Pada penelitian ini akan dilakukan analisis dan perbandingan algoritma vega dan confix-stripping yang sama-sama merupakan metode *affix removal stemmers*.

Pengujian di dalam penelitian ini menggunakan parameter penggujian yang dijelaskan di dalam jurnal karya Frakes W.B. dan Fox C.J. [2]. Parameter-parameter ini digunakan untuk mengukur *stemmer strength*, dimana penting dalam kasus algoritma penghilangan imbuhan, karena dapat memprediksi kemampuan *recall* dan *precision* serta *index compression* dari algoritma tersebut. *Stemmer strength* adalah suatu ukuran bagaimana sebuah *stemmer* dapat merubah sebuah kata menjadi *stem*. Paramter pengukuran yang digunakan untuk mengukur *stemmer strength* diantaranya adalah, ukuran rata-rata word per conflation class, index compression factor, jumlah kata yang berbeda antara kata masukan dan kata keluaran (stem), rata-rata

huruf yang dihilangkan untuk membentuk stem, dan terakhir nilai median dan mean modified hamming distance antara kata masukan dan stem.

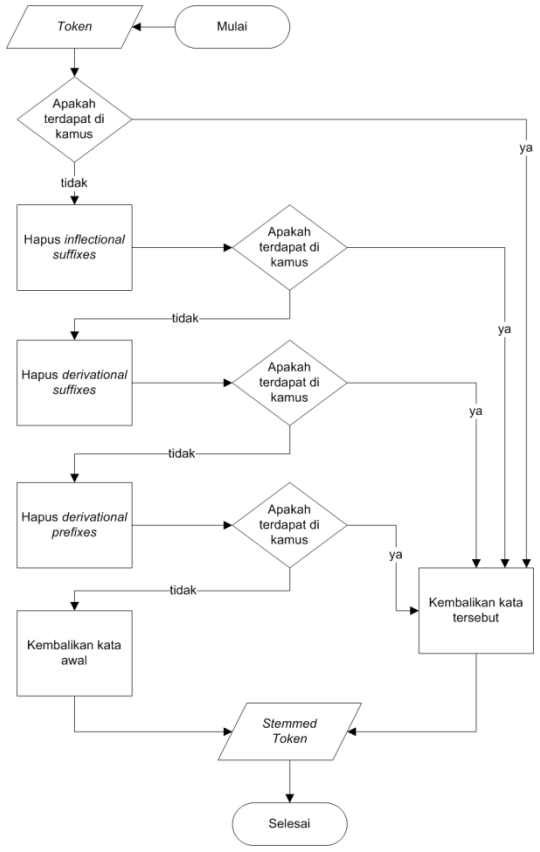
1.1. Algoritma Vega

Algoritma vega merupakan algoritma stemming yang diperkenalkan oleh Vinsensius Berlian Vega pada tahun 2001. Algoritma vega tidak menggunakan kamus data dalam melakukan pengecekan kata dasar.

Secara garis besar tahapan dalam algoritma vega berturut-turut adalah pembuangan partikel, pembuangan kata kepunyaan, pembuangan circumfix, pembuangan awalan, dan terakhir pembuangan akhiran.

1.2. Algoritma Confix-stripping

Algoritma stemming Confix-stripping (CS) adalah sebuah metode stemming pada Bahasa Indonesia yang diperkenalkan oleh Jelita Asian yang merupakan pengembangan dari metode stemming yang dibuat oleh Nazief dan Adriani [1]. Algoritma ini menggunakan kamus data untuk melakukan pengecekan kata dasar. Secara garis besar tahapan algoritma confix-stripping dapat dijelaskan seperti Gambar 1.



Gambar 1 Langkah-langkah algoritma confix-stripping

1.3. Parameter-parameter Uji

Pengujian di dalam penelitian ini menggunakan parameter penggujian yang dijelaskan di dalam jurnal karya Frakes W.B. dan Fox C.J. [2]. Parameter pertama adalah ukuran rata-rata word per conflation class, yaitu menunjukan jumlah rata-rata kata yang menghasilkan *stem* yang sama dari *corpus*, dengan rumus sebagai berikut:

$$\boldsymbol{wcc = \frac{n}{s}}$$

Dimana :

wcc = Word per conflation class

n = Jumlah kata unik sebelum stemming

s = Jumlah kata unik hasil stemming

Suatu algoritma stemming menginginkan nilai per conflation class yang lebih besar dari algoritma lainnya, untuk menunjukan algoritma tersebut lebih baik.

Parameter kedua yaitu index compression factor, menunjukan seberapa besar tingkat kompresi indeks dari algoritma stemming, atau seberapa besar tingkat pengurangan dari *index size* setelah dilakukan *stemming*. Apabila jumlah kata-kata di dalam *corpus* (kata unik sebelum stemming) sebanyak n, kemudian setelah dilakukan stemming apabila menghasilkan beberapa kata yang sama dilakukan penghapusan sehingga seluruh hasil stemming merupakan kata unik sebanyak s. Maka untuk menghitung parameter index compression factor (icf) adalah :

$$\boldsymbol{icf = \frac{(n-s)}{n} x\ 100\%}$$

Dimana :

icf = Index compression factor

n = Jumlah kata unik sebelum stemming

s = Jumlah kata unik hasil stemming

Suatu algoritma stemming menginginkan nilai index compression factor yang lebih besar dari algoritma lainnya, untuk menunjukan algoritma tersebut lebih baik.

Parameter ketiga adalah jumlah kata yang berbeda antara kata masukan dan kata keluaran atau stem. Sebuah algoritma stemming dapat menghasilkan kata yang tidak berubah dari kata masukannya. Algoritma stemming yang kuat akan mengubah kata lebih sering dari pada algoritma stemming yang lemah.

Parameter keempat adalah nilai rata-rata huruf yang dihapus, yaitu nilai rata-rata huruf dari kata masukan yang dihapus untuk membentuk sebuah *stem*. Algoritma stemming yang kuat akan menghapus huruf lebih banyak untuk membentuk sebuah *stem*.

Parameter kelima dan keenam berturut-turut adalah nilai mean dan median modified hamming distance. Nilai hamming distance antara dua *string* yang memiliki panjang yang sama didefinisikan sebagai jumlah karakter yang berbeda di posisi yang sama. Nilai hamming distance antara dua string yang memiliki panjang berbeda didefinisikan sebagai perbedaan panjang antara dua string tersebut.

Sebagai contoh terdapat tiga kata, membaca, dibaca, dan membacakan. Apabila sebuah algoritma stemming mengubahnya ke stem 'baca', maka :

1. membaca → beda 3 huruf dengan 'baca'
2. dibaca→ beda 2 huruf dengan 'baca'
3. membacakan → beda 6 huruf dengan 'baca'

Sehingga nilai mean modified hamming distance dapat dihitung sebagai nilai rata-ratanya yaitu sebagai berikut:

(3+2+6) / 3 = 3,66

Didapat nilai mean modified hamming distance sebesar 3,66 kata.

Nilai median modified hamming distance juga dapat diperoleh sebagai nilai tengah dari ketiga nilai tersebut setelah diurutkan. Sehingga didapat nilai median modified hamming distance sebesar 3.

2. ISI PENELITIAN

Hasil pengujian algoritma vega dan algoritma confix-stripping merupakan pengujian terhadap parameter uji pada dokumen-dokumen uji. Dokumen-dokumen yang diambil merupakan dokumen sample dari digital library Sarihusada sebanyak sepuluh dokumen. Kesepuluh dokumen memiliki jumlah kata masukan (kata uji) berupa *corpus* untuk diuji ke masing-masing algoritma yang telah melalui hasil filtering, sebagai berikut :

2.1. Ukuran Rata-rata Word Per Conflation Class

Tabel 1 ukuran rata-rata word per conflation class

Dok No	n	s		wcc	
		Vega	CS	Vega	CS
1	417	407	400	1,02 kata	1,04 kata
2	78	77	77	1.013 kata	1.013 kata
3	86	85	85	1.012 kata	1.012 kata
4	184	182	181	1.01 kata	1.016 kata
5	80	79	78	1.01 kata	1.016 kata
6	462	451	450	1.024 kata	1.027 kata
7	58	57	57	1.017 kata	1.017 kata
8	85	84	83	1.012 kata	1.024 kata
9	115	114	114	1.0087 kata	1.008 kata
10	128	126	124	1,015 kata	1,032 kata
Total	1693	1662	1649	1,018 kata	1,026 kata

Algoritma stemming yang bagus dalam hal stemmer strength menginginkan nilai mean word per conflation class yang lebih besar dari algoritma stemming lainnya. Berdasarkan pengujian pada Tabel 1 parameter mean word per conflation class algoritma confix-stripping lebih baik dengan rata-rata 1,026 kata.

2.2. Index Compression Factor

Tabel 2 index compression factor

Dok No	n	s		icf	
		Vega	CS	Vega	CS
1	417	407	400	2.398 %	4.076 %
2	78	77	77	1.282 %	1.28 %
3	86	85	85	1.162 %	1.162 %
4	184	182	181	1.086 %	1.63 %
5	80	79	78	1.25 %	2.5 %
6	462	451	450	2.381 %	2.597%
7	58	57	57	1.724 %	1.724 %
8	85	84	83	1.176 %	2.353 %
9	115	114	114	0.869 %	0.869 %
10	128	126	124	1,562 %	3,125 %
Total	1693	1662	1649	1,831 %	2,598 %

Algoritma stemming yang bagus dalam hal stemmer strength menginginkan nilai index compression factor yang lebih besar dari algoritma stemming lainnya. Berdasarkan pengujian pada Tabel 2 parameter index compression factor algoritma confix-stripping lebih baik sebesar 2,598 %

2.3. Jumlah Kata Yang Berbeda Antara Kata Masukan Dan Keluaran

Tabel 3 jumlah kata yang berbeda antara kata masukan dan keluaran

Dok No	Vega	CS
1	409 kata	365 kata
2	76 kata	65 kata
3	83 kata	72 kata
4	179 kata	72 kata
5	74 kata	61 kata
6	448 kata	413 kata
7	54 kata	49 kata
8	82 kata	65 kata
9	108 kata	88 kata
10	123kata	101 kata
Total	1636 kata	1351 kata

Algoritma stemming yang bagus dalam hal stemmer strength lebih sering merubah kata dari algoritma stemming lainnya, sehingga menghasilkan kata yang berbeda. Berdasarkan pengujian pada Tabel 3 algoritma vega lebih baik dengan total perbedaan 1636 kata dari 10 dokumen sampel uji.

2.4. Nilai Rata-rata Huruf Yang Dihapus

Tabel 4 nilai rata-rata huruf yang dihapus

Dok No	m		n	Vega	CS
	Vega	CS			
1	1662	1450	417	3.985 huruf	3.477 huruf
2	305	265	78	3.91 huruf	3.397 huruf
3	317	289	86	3.686 huruf	3.36 huruf
4	699	604	184	3.798 huruf	3.282 huruf
5	279	227	80	3.487 huruf	2.837 huruf
6	1763	1567	462	3.816 huruf	3.391 huruf
7	206	172	58	3.551 huruf	2.965 huruf
8	295	225	85	3.4705 huruf	2.647 huruf
9	408	319	115	3.547 huruf	2.773huruf
10	486	360	128	3,796 huruf	2,812 huruf
Total	6420	5478	1693	3,792 huruf	3,235 huruf

Algoritma stemming yang bagus dalam hal stemmer strength lebih sering menghapus huruf untuk membentuk sebuah stem. Berdasarkan pengujian pada Tabel 4 algoritma vega lebih baik karena rata-rata menghapus 3,792 huruf dari 10 dokumen uji.

2.5. Nilai Mean Modified Hamming Distance

Tabel 5 nilai mean modified hamming distance

No	h		n	Vega	CS
	Vega	Cs			
1	1662	1528	417	3.985 kata	3.664 kata
2	305	280	78	3.91 kata	3.589 kata
3	317	304	86	3.686 kata	3.534 kata
4	699	646	184	3.798 kata	3.51 kata
5	279	239	80	3.487 kata	2.987 kata
6	1763	1641	462	3.816 kata	3.552 kata
7	206	180	58	3.551 kata	3.103 kata
8	295	240	85	3.47 kata	2.823 kata
9	408	333	115	3.547 kata	2.895 kata
10	486	382	128	3,796 kata	2,984 kata
Total	6420	5773	1693	3,792 kata	3,409 kata

Algoritma stemming yang bagus dalam hal stemmer strength memiliki nilai hamming distance lebih besar dari algoritma stemming lainnya, atau memiliki rata-rata hamming distance yang lebih besar untuk sejumlah kata di dalam dokumen uji. Berdasarkan pengujian pada Tabel 5 algoritma vega lebih baik karena memiliki rata-rata hamming distance sebanyak 3,792 kata.

2.6. Nilai Median Modified Hamming Distance

Tabel 6 nilai median modified hamming distance

Dok No	Vega	CS
1	4	4
2	4	3
3	3	3
4	4	3
5	3	3
6	4	3
7	3	3
8	3	3
9	3	3
10	4	3
Rata-rata Total	3,5	3,1

Algoritma stemming yang bagus dalam hal stemmer strength memiliki nilai hamming distance lebih besar dari algoritma stemming lainnya, dapat berarti juga untuk sejumlah kata di dalam dokumen memiliki nilai tengah (median) yang paling besar. Berdasarkan pengujian pada Tabel 6 dari 10 dokumen algoritma vega rata-rata memiliki nilai median yang lebih baik sebesar 3,5.

3. PENUTUP

3.1. Kesimpulan

Berdasarkan hasil pengujian yang telah dilakukan dalam analisis dan perbandingan algoritma vega dan algoritma confix-stripping, dapat disimpulkan bahwa keenam parameter uji tidak menunjukan kekonsistenan dalam menentukan algoritma mana yang lebih baik. Hasil pengujian terhadap sepuluh dokumen uji menunjukan algoritma vega lebih baik pada empat parameter uji yaitu jumlah kata yang berbeda antara kata masukan dan *stem*, nilai rata-rata huruf yang dihapus, nilai mean modified hamming distance, dan nilai median modified hamming distance. Sedangkan algoritma confix-stripping lebih baik pada dua parameter uji lainnya yaitu ukuran

rata-rata word per conflation class dan index compression factor.

Hal di atas berarti kedua algoritma masing-masing memiliki kelebihan untuk kasus tertentu. Algoritma vega memiliki kelebihan dalam mengubah *term* menjadi kata yang berbeda atau baru. Sedangkan algoritma confix-stripping memiliki kelebihan dalam mengurangi jumlah *term*, seperti *mereduksi index size* pada proses *indexing*.

3.2. Saran

Berdasarkan hasil pengujian yang telah dilakukan, maka saran yang diharapkan setelah dilakukannya analisis dan perbandingan algoritma vega dan algoritma confix-stripping pada teks bahasa indonesia adalah sebagai berikut :

1. Melakukan pengujian terhadap stem yang dihasilkan algoritma vega apakah dapat mempengaruhi proses *indexing*.

2. Melakukan analisis performansi algoritma confix-stripping dalam mereduksi ukuran *index size.*

3. Melakukan implementasi terhadap proses pencarian pada *information retrieval*.

4. DAFTAR PUSTAKA

[1] Asian, J., Williams, H.E. and Tahaghoghi, S.M.M. “Stemming Indonesian”. RMIT University, Melbourne, Australia.

[2] Frakes, W.B., Fox, C.J. “Strength and Similarity of Affix Removal Stemming Algorithms”. Computer Science Department. Virginia Tech and James Madison University.

[3] Vega, B. V., dan Bressan, S. 2001. “Indexing the Indonesian web: Language Identification and Miscellaneous Issues”. Tenth International World Wide Web Conference, Hong Kong.