STEMMING DOKUMEN TEKS BAHASA INDONESIA MENGGUNAKAN ALGORITMA PORTER

Oleh :
Lasmedi Afuan
Prodi Teknik Informatika, Fakultas Sains dan Teknik, Universitas Jenderal Soedirman
Jl. Mayjen Sungkono Blater Km 5. Purbalingga
Email: lasmedi.afuan@unsoed.ac.id

ABSTRAK

Informasi merupakan hal yang sangat mudah didapatkan dan diakses. Tetapi terkadang informasi yang diperoleh tidak sesuai dengan apa yang diinginkan pengguna. Diperlukan sistem yang dapat membantu mencari informasi yang dibutuhkan secara efektif dan efisien. Sistem informasi ini sering kali disebut dengan istilah sistem temu kembali informasi (STKI). Pada STKI salah satu tahapan yang sangat penting adalah tahap *Stemming*. Tahapan ini merupakan tahapan mentransformasikan kata dalam sebuah kalimat atau dokumen ke kata dasarnya. Pada penelitian ini, akan dijelaskan proses *Stemming* pada kalimat bahasa indonesia dengan menggunakan algortitma porter untuk mendapat *root word* dari kata dalam dokumen teks. Tahapan yang ada pada algoritma porter diterjemahkan menjadi koding program PHP. Kamus kata dasar dan *stoplist* disimpan di MySql. Pada proses stemming dilakukan tidak kata perkata, akan tetapi langsung stemming pada dokumen. Sehingga proses stemming yang dilakukan lebih cepat dan efektif.

Kata kunci: Sistem temu kembali informasi, root word, algoritma porter, php, mysql

A. PENDAHULUAN

Teknologi informasi dan komunikasi pada era sekarang mengalami perkembangan pesat. Orang berlomba memanfaatkan TIK, TIK terutama *internet* telah digunakan sebagai alat untuk mengakses dan mendapatkan informasi. Permasalahan yang sering kali muncul dalam mengakses dan mendapatkan informasi adalah memilih informasi yang tepat sesuai dengan keinginan *user*. Untuk mengatasi masalah pencarian informasi, maka munculah sistem temu kembali informasi (STKI). STKI memungkinkan pengguna untuk mencari informasi yang tersimpan didalam dokumen secara efektif dan efisien. Efektif berarti user mendapatkan dokumen yang relevan dengan query yang diinputkan. Efisien berarti waktu pencarian yang sesingkat-singkatnya (Agusta, 2009).

Salah satu tahapan yang sangat penting dalam STKI adalah proses *stemming*. Stemming merupakan salah satu tahapan *text pre-processing* pada

STKI. *Stemming* mentrasformasikan kata-kata dalam dokumen menjadi kata akarnya (*root word*) atau kata dasar atau proses penghilangan imbuhan kata. Pada makalah ini, penulis akan menjelaskan tahapan *stemming* dokumen teks menggunakan algoritma porter. Proses *Stemming* pada dokumen bahasa indonesia sedikit lebih kompleks, karena pada dokumen bahasa indonesia harus menghilangkan imbuhan-imbuhan untuk mendapatkan kata dasarnya.

B. METODOLOGI PENELITIAN

Metode Penelitian yang digunakan dalam penelitian *stemming* kalimat bahasa indonesia menggunakan algoritma *porter* antara lain :

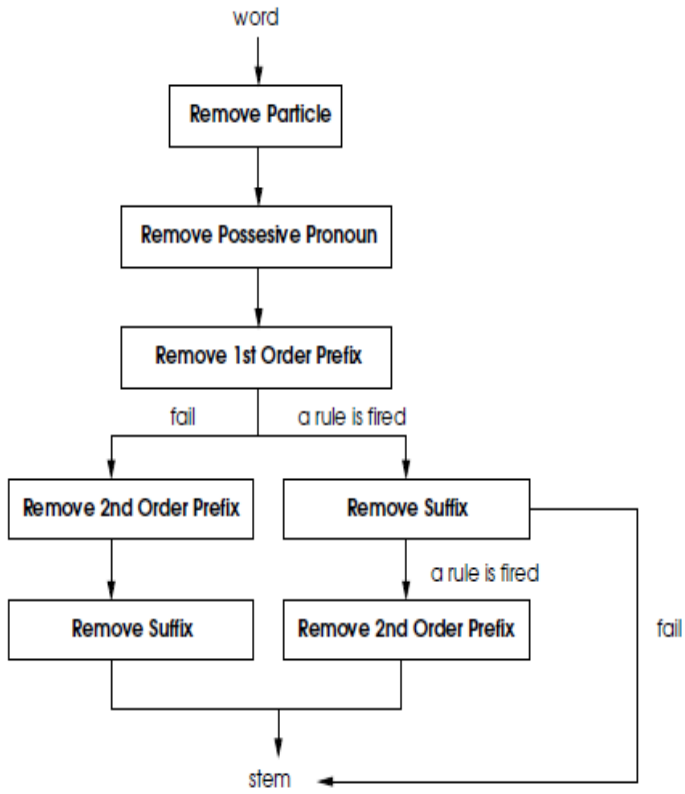
1. Studi Pustaka

   Studi pustaka dilakukan terkait dengan pengumpulan literatur, pustaka mengenai algoritma porter, serta studi mengenai imbuhan kata dalam bahasa indonesia. Selain itu juga pada metode ini dikumpulkan kata dasar bahasa indonesia, yang kemudian akan dijadikan sebagai kamus kata dasar.
2. Ujicoba

   Tahapan ujicoba dilakukan mencoba aplikasi *stemming* menggunakan dokumen teks.

C. HASIL DAN PEMBAHASAN

Stemming merupakan proses yang memetakan bentuk varian kata menjadi kata dasarnya (Fadillah, 2003). Dalam pengembangan aplikasi *stemming* dokumen teks berbahasa indonesia menggunakan bahasa pemrograman PHP dan MySql sebagai DBMS (*database management system*). Tahapan Stemming algoritma *porter* dapat dilihat pada gambar 1.



Gambar 1. Algoritma Porter (Fadillah)

Berdasarkan gambar 1, Adapun langkah-langkah algoritma pada algoritma *Porter* adalah sebagai berikut (Agusta, 2009):

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 *root word*
   b. Hapus awalan kedua. Kemudian kata akhir diasumsikan sebagai *root word*.

Terdapat 5 kelompok aturan pada Algoritma Porter untuk Bahasa Indonesia ini (Agusta, 2009). Aturan tersebut dapat dilihat pada Tabel 1 sampai Tabel 5.

Tabel 1. Aturan Untuk Inflectional Particle

Akhiran	Replacement	Measure Condition	Additional Condition	Contoh
-kah	NULL	2	NULL	bukukah
-lah	NULL	2	NULL	pergilah
-pun	NULL	2	NULL	bukupun

Tabel 2. Aturan Untuk Inflectional Possesive Pronoun

Akhiran	Replacement	Measure Condition	Additional Condition	Contoh
-ku	NULL	2	NULL	bukuku
-mu	NULL	2	NULL	bukumu
-nya	NULL	2	NULL	bukunya

Tabel 3. Aturan Untuk First Order Derivational Prefix

Awalan	Replacement	Measure Condition	Additional Condition	Contoh
meng-	NULL	2	NULL	mengukur → ukur
meny-	S	2	V…*	menyapu → sapu
men-	NULL	2	NULL	menduga → duga
mem-	P	2	V…	memaksa→ paksa
mem-	NULL	2	NULL	membaca→ baca
me-	NULL	2	NULL	merusak → rusak
peng-	NULL	2	NULL	pengukur→ ukur
peny-	S	2	V…	penyapu → sapu
pen-	NULL	2	NULL	penduga → duga
pem-	P	2	V…	pemaksa → paksa
pem-	NULL	2	NULL	pembaca → baca
di-	NULL	2	NULL	diukur → ukur
ter-	NULL	2	NULL	tersapu → sapu
ke-	NULL	2	NULL	kekasih → kasih

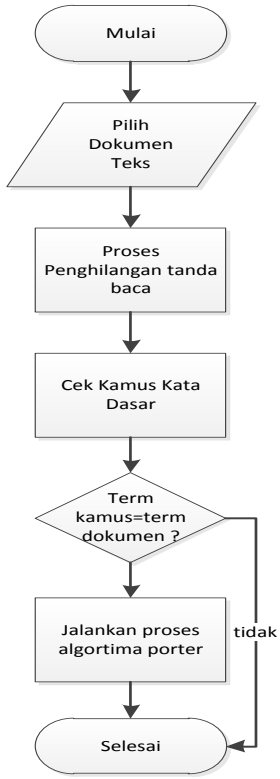
Tabel 4. Aturan Untuk Second Order Derivational Prefix

Awalan	Replacement	Measure Condition	Additional Condition	Contoh
ber-	NULL	2	NULL	*berlari →lari*
bel-	NULL	2	*Ajar*	*belajar →ajar*
be-	NULL	2	*k*er*	*bekerja →kerja*
per-	NULL	2	NULL	*perjelas →jelas*
pel-	NULL	2	*Ajar*	*pelajar →ajar*
pe-	NULL	2	NULL	*pekerja →kerja*

Tabel 5. Aturan Untuk Derivational Suffix

Akhiran	Replacement	Measure Condition	Additional Condition	Contoh
-kan	NULL	2	Prefix bukan anggota {ke, peng}	tarikkan → tarik, mengambilkan → ambil
-an	NULL	2	prefix bukan anggota {di, meng, ter}	makanan → makan, perjanjian → janji
-i	NULL	2	prefix bukan anggota{ber, ke, peng}	Tandai → tanda, mendapati → dapat

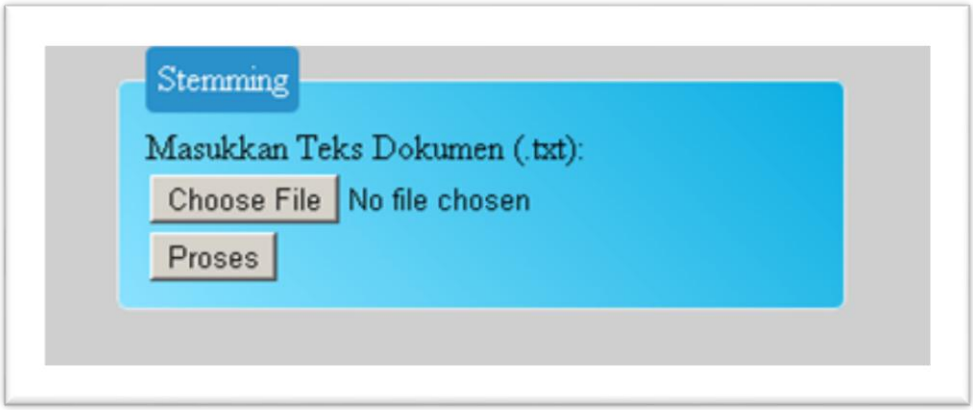
Flowcart dari proses stemming dokumen teks menggunakan algoritma *Porter* dapat dilihat pada gambar 2.



Gambar 2. Flowchart aplikasi

Gambar 2 merupakan tampilan flowchart dari aplikasi stemming dokumen teks menggunakan algoritma Porter. Pada tahap awal, dilakukan proses upload dokumen teks, kemudian dilakukan proses penghilangan tanda baca pada dokumen teks. Setelah itu, dilakukan proses pengecekan setiap kata dalam dokumen ke kamus kata dasar, jika ada maka ubah kata menjadi kata dasar,

jika tidak maka kata pada dokumen ditulis sebagai kata dasarnya. Tampilan dari aplikasi *stemming* dapat dilihat pada gambar 3.



Gambar 3. Tampilan Awal Proses Stemming

Pada gambar 3, merupakan tampilan awal dari aplikasi *stemming* dokumen. Pada tampilan awal ini, terdapat inputan untuk memilih dokumen teks yang akan dilakukan proses *stemming*. Jika sudah memilih klik tombol **proses** untuk memulai proses stemming. Tampilan hasil dari proses stemming dokumen teks dapat dilihat pada gambar 4.

oses analyzing adalah proses analisa dari hasil proses tagging sehingga diketahui seberapa jauh tingkat keterhubungan antar kata-kata dan antar dokumnen yang ada

ks Awal (Hilangkan tanda baca): teks.txt

oses analyzing proses analisa hasil proses tagging diketahui seberapa jauh tingkat keterhubungan antar kata kata antar dokumnen ada

lasil Proses Stemming Dokumen Teks "teks.txt"

No	Kata	Partikel	Posessive Pronoun (PP)	Awalan 1	Awalan 2	Akhiran	Kata Dasar
1	Proses	Proses	Proses	Proses	Proses	Proses	Proses
2	analyzing	analyzing	analyzing	analyzing	analyzing	analyzing	analyzing
3	proses	proses	proses	proses	proses	proses	proses
4	analisa	analisa	analisa	analisa	analisa	analisa	analisa
5	hasil	hasil	hasil	hasil	hasil	hasil	hasil
6	proses	proses	proses	proses	proses	proses	proses
7	tagging	tagging	tagging	tagging	tagging	tagging	tagging
8	diketahui	diketahui	diketahui	ketahui	ketahui	ketahu	ketahu
9	seberapa	seberapa	seberapa	seberapa	berapa	berapa	berapa
10	jauh	jauh	jauh	jauh	jauh	jauh	jauh
11	tingkat	tingkat	tingkat	tingkat	tingkat	tingkat	tingkat
12	keterhubungan	keterhubungan	keterhubungan	terhubungan	terhubungan	terhubung	terhubung
13	antar	antar	antar	antar	antar	antar	antar
14	kata	kata	kata	kata	kata	kata	kata
15	kata	kata	kata	kata	kata	kata	kata
16	antar	antar	antar	antar	antar	antar	antar
17	dokumnen	dokumnen	dokumnen	dokumnen	dokumnen	dokumnen	dokumnen
18	ada	ada	ada	ada	ada	ada	ada

'aktu Stemming : 8.048956155777 detik

Gambar 4. Hasil Proses Stemming

Hasil Pengujian

Pengujian (dokumen teks)

```
Proses analyzing adalah proses analisa dari hasil  proses tagging
sehingga diketahui seberapa jauh tingkat keterhubungan antar kata-
kata dan antar  dokumen yang ada
```

Setelah dilakukan proses stemming pada dokumen teks

```
Proses analyzing proses analisa hasil  proses tagging ketahu berapa
jauh tingkat terhubung antar kata kata antar  dokumen ada
```

D. KESIMPULAN DAN SARAN

1. Presisi pada Proses *stemming* masih belum mencapai hasil yang maksimal, hal ini bukan karena aplikasi yang tidak benar, akan tetapi kamus kata dasar yang masih belum terlalu lengkap.
2. Aplikasi Stemming dokumen bahasa indonesia ini, sementara hanya bisa membaca dokumen dengan ekstensi (.txt), sehingga aplikasi ini masih harus disempurnakan agar bisa membaca berbagai format dokumen.

DAFTAR PUSTAKA

Agusta Ledy, 2009. *Perbandingan Algoritma Stemming Porter Dengan Algoritma,*

Fadillah Z. Tala, *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia,* Netherland, Universiteit van Amsterdam

Lancaster, F.W. 1979. *Information Retrieval Systems: Characteristics, Testing, and Evaluation,* 2nd Edition, John Wiley, New York.

Nazief & Adriani,2009. *Untuk Stemming Dokumen Teks Bahasa Indonesia.* KNSI. Bali .