Implementasi Stemmer Tala pada Aplikasi Berbasis Web

Mardi Siswo Utomo
Program Studi Teknik Informatika, Universitas Stikubank
email : mardiutomo@gmail.com

Abstrak

Stemming adalah proses untuk mencari kata dasar pada suatu kata. Pada analisa temu kembali informasi imbuhan merupakan bagian dari informasi yang tidak bermakna, seperti halnya stop word. Sehingga imbuhan harus dihilangkan untuk mempercepat proses pengindekan dan proses query. Proses stemming dapat dilakukan dengan 2 cara yaitu dengan menggunakan kamus dan menggunakan aturan-aturan imbuhan. Untuk mendapatkan tingkat kebenaran hasil yang tinggi biasaya digunakan kamus seperti yang diperkenalkan oleh Nazief dan Adriani, stemmer bahasa melayu oleh Ahmad, Yuso, dan Sembok. tetapi teknik ini membutuhkan waktu komputasi yang tinggi karena ada proses pengambilan data pada database. Sedang untuk aplikasi yang lebih sederhana dan tidak membutuhkan akurasi yang tinggi teknik aturan imbuhan sangat mudah untuk diimplementasikan dan tidak membutuhkan waktu komputasi yang tinggi.

Stemmer tala merupakan adopsi dari algoritma stemmer bahasa inggris terkenal porter stemmer. Stemmer ini menggunakan rule base analisis untuk mencari root sebuah kata. Stemmer ini sama sekali tidak menggunakan kamus sebagai acuan, seperti halnya stemmer ahmad,vega dan jelita. Proses stemming bahasa Indonesia menggunakan algoritma berbasis aturan mempunyai tingkat kesalahan tinggi, sehingga dapat mempengaruhi akurasi hasil akhir. Walaupun demikian performa stemming berbasis aturan relatif stabil dengan jumlah dokumen yang berkembang.

Kata Kunci : Stemming, Stemmer Bahasa Indonesia, Tala

PENDAHULUAN

Hampir setiap aplikasi termasuk berbasis web dengan pengelolaan basis data membutuhkan proses temu kembali informasi. Pada proses temu kembali selain query dan umpan balik pengguna terlebih dahulu dilakukan pengindekan pada data yang ada, proses pengindekan data berbasis teks akan membutuhkan proses stemming.

Stemming adalah proses untuk mencari kata dasar pada suatu kata. Pada analisa temu kembali informasi imbuhan merupakan bagian dari informasi yang tidak bermakna, seperti halnya stop word. Sehingga imbuhan harus dihilangkan untuk mempercepat proses pengindekan dan proses query.

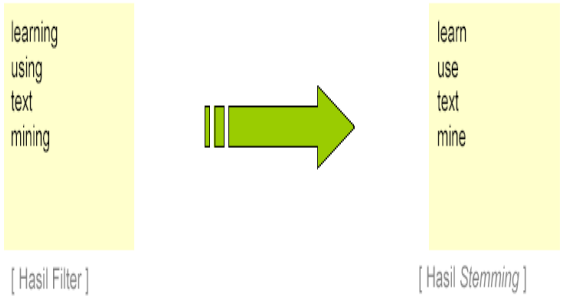
Proses stemming dapat dilakukan dengan 2 cara yaitu dengan menggunakan kamus dan menggunakan aturan-aturan imbuhan. Untuk mendapatkan tingkat kebenaran hasil yang tinggi biasaya digunakan kamus seperti yang diperkenalkan oleh Nazief dan Adriani, stemmer bahasa melayu oleh Ahmad, Yuso, dan Sembok. tetapi teknik ini membutuhkan waktu komputasi yang tinggi karena ada proses pengambilan data pada database. Sedang untuk aplikasi yang lebih sederhana dan tidak membutuhkan akurasi yang tinggi teknik aturan imbuhan sangat mudah untuk diimplementasikan dan tidak membutuhkan waktu komputasi yang tinggi.

Algoritma stemmer berbahasa indonesia tanpa menggunakan kamus diperkenalkan oleh Vega VB dan Bressan S, dengan menghilangkan

imbuhan-imbuhan pada kata-kata berbahasa Indonesia berimbuhan. Selain Vega, Tala juga memperkenalkan porter like stemmer untuk bahasa Indonesia. Tala menggunakan rule base untuk menghilangkan imbuhan kata.

Stemming

Proses stemming adalah proses untuk mencari root dari kata hasil dari proses filtering. Pencarian root sebuah kata atau biasa disebut degan kata dasar dapat memperkecil hasil indeks tanpa harus menghilangkan makna. Filtering adalah proses pengambilan kata-kata yang dianggap penting atau mempunyai makna. Ada dua pendekatan pada proses stemming yaitu pendekatan kamus dan pendekatan aturan. Beberapa penelitian juga telah dilakukan untuk stemmer bahasa Indonesia baik untuk pendekatan kamus ataupun pendekatan aturan. Ahmad,Vega, Jelita dan Tala mereka masing-masing mempunyai algoritma yang berbeda dalam melakukan proses stemmer pada dokumen bernahasa Indonesia. Gambar 1 merupakan gambaran dari hasil proses stemming dalam bahasa inggris, pada gambar tersebut diperlihatkan kata asal *learning* dirubah menjadi kata dasarnya yaitu *learn*. Kemudian kata *using* dikembalikan ke bentuk dasar menjadi *use*. Tetapi kata *text* merupakan kata dasar sehingga tidak dirubah.



Gambar 1. Contoh proses stemming bahasa inggris

Stemmer Bahasa Indonesia

Dalam penelitian oleh ahmad dkk (1996) , dijelaskan bahawa penggunaan kamus sangat memegang peranan penting untuk melakukan pencarian kata dasar dalam bahasa melayu. Tetapi dalam penelitian Tala dijelaskan untuk korpus yang berkembang dan dalam jumlah yang besar, ketergantungan pada kamus akan menurunkan kemampuan sistem dalam jangka panjang (Tala, 2004). Tala memilih menggunakan komputasi dalam pencarian kata dasar dengan menggunakan algoritma berbasis aturan.

Stemmer Bahasa Indonesia Tala

Struktur pembentukan kata dalam Bahasa Indonesia adalah sebagai berikut:

[awalan-1] + [awalan-2] + dasar + [akhiran] + [kepunyaan] + [sandang]

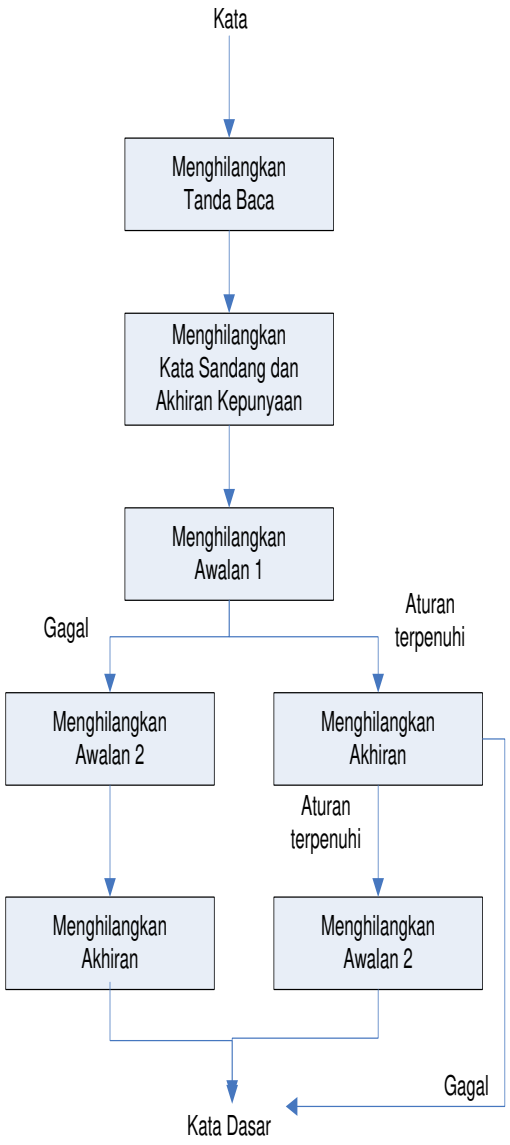
Masing-masing bagian tersebut (yang dalam kotak bisa ada atau tidak), digabungkan dengan kata dasar membentuk kata berimbuhan.

Stemmer tala merupakan adopsi dari algoritma stemmer bahasa inggris terkenal porter stemmer. Stemmer ini menggunakan rule base analisis untuk mencari root sebuah kata. Stemmer ini sama sekali tidak menggunakan kamus sebagai acuan, seperti halnya stemmer ahmad,vega dan jelita

Pada stemmer Tala terdapat 5 langkah utama dengan 3 langkah awal dan 2 langkah pilihan, langkah-langkah tersebut sbb:

1. Menghilangkan partikel
2. Menghilangkan kata sandang dan kepunyaan.
3. Menghilangkan awalan 1
4. Jika suatu aturan terpenuhi jalankan sbb :
   a. Hilangkan Akhiran
   b. Jika suatu aturan terpenuhi, hilangkan awalan 2. Jika tidak proses stemming selesai
5. Jika tidak ada aturan yang terpenuhi jalankan sbb :
   a. Hilangkan awalan 2.
   b. Hilangkan Akhiran
   c. Proses stemming selesai.

Selain itu, tala membagi imbuhan menjadi 5 cluster / kelompok. Alur proses dari algoritma Tala diperlihatkan pada gambar 2.



Gambar 2. Proses stemming algoritma Tala (Tala 2004:7)

Proses menghilangkan partikel

Pada proses ini dokumen dibersihkan dari partikel / tanda baca. Selain tanda baca dalam proses ini juga dihilangkan semua angka serta kata-kata yang tidak bermakna (stopword). Stopword yang diketahui disimpan dalam tabel basis data stopword kemudian untuk semua kata yang ada dalam tabel tersebut akan dihilangkan. Isi tabel basis data stopword diambil dari daftar stopword Tala (Tala 2004).

Masukan untuk proses stemming adalah kata hasil dari tokenizing. Tanda baca dan angka sudah dihilangkan sebelum dilakukan tokenizing. Kemudian data stopword tersimpan dalam tabel basis data, proses menghilangkan stopword akan lebih cepat dilakukan sekaligus melalui perintah Query. Sehingga Stopword akan dihilangkan setelah proses stemming selesai dilaksanakan pada semua dokumen. Proses menghilangkan stopword dibahas pada pemrosesan indek artikel.

Proses menghilangkan kata sandang dan kepunyaan

Pada proses ini dokumen melalui perlakuan untuk menghilangkan kata sandang dan kepunyaan. Proses ini dibagi dalam 2 cluster proses yang harus diproses secara urut. Algoritma 1 adalah algoritma yang digunakan untuk menghilangkan kata sandang dan kepunyaan.

Algoritma 1 *Pseudocode untuk menghilangkan kata sandang*

```
// Aturan cluster 1
$str= ganti("lah "," ") pada $str;
$str= ganti("kah "," ") pada $str;
$str= ganti("pun "," " ) pada $str;
// Aturan cluster 2
$str= ganti("nya "," ") pada $str;
$str= ganti("ku "," ") pada $str;
$str= ganti("mu "," " ) pada $str;
```

Menghilangkan awalan 1

Pada proses ini dokumen melalui perlakuan untuk menghilangkan awalan, stemmer Tala melokalisasi awalan 1 dalam 1 cluster proses yang harus diproses secara urut. Algoritma 2 adalah algoritma yang digunakan untuk menghilangkan awalan 1.

Algoritma 2 *Algoritma untuk menghilangkan awalan 1.*

```
// Aturan cluster 3
$str= ganti(" meng”," ") pada $str;
$str= ganti(" menya"," s") pada $str;
$str= ganti(" menyi"," s") pada $str;
$str= ganti(" menyu"," s") pada $str;
$str= ganti(" menye"," s) pada $str;
$str= ganti(" menyo"," s") pada $str;
$str= ganti(" meny"," s") pada $str;
$str= ganti(“ men"," " ) pada $str;
$str= ganti(" mema"," p") pada $str;
$str= ganti(" memi"," p") pada $str;
$str= ganti(" memu"," p) pada $str;
$str= ganti(" meme"," p) pada $str;
$str= ganti(" memo"," p) pada $str;
$str= ganti(" mem"," " ) pada $str;
$str= ganti(" me"," " ) pada $str;
$str= ganti(" peng"," " ) pada $str;
$str= ganti(" penya"," s") pada $str;
$str= ganti(" penyi"," s") pada $str;
$str= ganti(" penyu"," s") pada $str;
$str= ganti(" penye"," s) pada $str;
$str= ganti(" penyo"," s") pada $str;
$str= ganti(" peny"," s) pada $str;
$str= ganti(" pen"," " ) pada $str;
$str= ganti(" pema"," p") pada $str;
$str= ganti(" pemi"," p") pada $str;
$str= ganti(" pemu"," p) pada $str;
$str= ganti(" peme"," p) pada $str;
$str= ganti(" pemo"," p) pada $str;
$str= ganti(" pem"," " ) pada $str;
$str= ganti(" di"," ") pada $str;
$str= ganti(" ter"," " ) pada $str;
$str= ganti(" ke"," " ) pada $str;
```

Menghilangkan awalan 2

Pada proses ini dokumen melalui perlakuan untuk menghilangkan awalan, stemmer Tala melokalisasi awalan 2 dalam 1 cluster proses yang harus diproses secara urut. Algoritma 3 adalah Algoritma yang digunakan untuk menghilangkan awalan 1.

Algoritma 3 *adalah Algoritma yang digunakan untuk menghilangkan awalan 1.*

```
// Aturan cluster 4
$str= ganti(" ber"," ") pada $str;
$str= ganti(" bel"," ") pada $str;
$str= ganti(" be"," ") pada $str;
$str= ganti(" per"," " ) pada $str;
$str= ganti(" pel"," ") pada $str;
$str= ganti(" pe"," " ) pada $str;
```

Menghilangkan akhiran.

Pada proses ini dokumen melalui perlakuan untuk menghilangkan awalan, stemmer Tala melokalisasi akhiran dalam 1 cluster proses yang harus diproses secara urut. Algoritma 4 adalah algoritma yang digunakan untuk menghilangkan awalan 1.

Algoritma 4 *Algoritma untuk menghilangkan awalan 1.*

```
// Aturan cluster 5
$str= ganti("kan "," ") pada $str;
$str= ganti("an "," ") pada $str;
$str= ganti("i "," ") pada $str;
```

Setelah 5 tahap dilalui maka kata sudah dianggap telah menjadi root atau kata dasar. Menurut Tala kata dasar pada bahasa Indonesia terdiri paling sedikit 2 kata, sehingga sebelum dilakukan penggantian / penghilangan awalan, akhiran ataupun partikel diperhatikan panjang huruf yang tersisa. Jumlah huruf yang akan diproses minimal 2 + (panjang imbuhan yang akan dihilangkan) + 2 (spasi, untuk depan dan belakang kata).

Hasil pengukuran

Pada proses stemming dilakukan evaluasi pada 1000 kata terbanyak. Kata-kata tersebut telah bebas dari stopword / stopwords. Hasil evaluasi stemmer tala diperlihatkan pada tabel 1.

Pada tabel diperlihatkan 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. Kemudian

kesalahan juga terjadi pada nama orang / istilah / singkatan. Kebayakan 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 %.

Tabel 1. Hasil evaluasi kesalahan stemming

Jenis Kesalahan	Contoh	Hasil Stemmer	Seharusnya	Byk
Nama orang, tempat, istilah, singkatan	Januari, Februari, Melitus	Januar, Februar, Litus	-Tetap-	47
Bahasa asing	Recycle, Dump	-tetap-	-sudah benar-	-
Kesalahan kata, susunan imbuhan, imbuhan kata asing	Dikelompokkan, defisiensi, prevalensi	-tetap-	lompok k, defisien s, prevale ns	26
Kata terlalu banyak dipotong (overstemming)	Metode, Pemberian, Bayi	Tode, ian, bay	Metode, beri, bayi	177
Kata terlalu sedikit dipotong (understemming)	Dipengaruhi, menjalani, keterlambatan,	Ngaruh, jalani, terlambat	Aruh, jalan, lambat	6

KESIMPULAN

Berdasarkan hasil penelitian dari bab sebelumnya maka dapat disimpulkan beberapa hal sebagai berikut :

Proses stemming bahasa Indonesia menggunakan algoritma berbasis aturan mempunyai tingkat kesalahan tinggi, sehingga dapat mempengaruhi akurasi hasil akhir. Walaupun demikian performa stemming berbasis aturan relatif stabil dengan jumlah dokumen yang berkembang.

Penelitian ini menggunakan corpus yang relatif kecil (abstrak), dapat diteliti lebih lanjut pada corpus yang lebih besar lagi misalnya isi artikel, skripsi, tesis atau desertasi, untuk melihat kualitas hasil pengukuran.

Penggunaan algoritma stemming bahasa Indonesia berbasis kamus dan aturan dapat meningkatkan kualitas indek.

DAFTAR PUSTAKA

Murhadin, E. (2003). *PHP Programming Fundamental dan MySQL Fundamental,* http://ikc.cbn.net.id/umum/endy-php.php

Nugroho, B. (2004). *PHP & MySQL Dengan Editor Dreamweaver MX,* Andi, Yogyakarta

PHP-Nuke. (n.d.). *PHP-Nuke* . Retrieved July 29, 2013, from http://www.phpnuke.org

Pressman, R. S. (1997). *Software engineering: a practitioner's approach* (4th ed.). New York: McGraw-Hill.

Prothelon’s. (2005). *Web Desain, PHP Programming, Language Learning*, http://prothelon.com/mambo/tutorial

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. (2011). *Design and Implementation of Document Similarity for web-based Medical Journal Management*

Vega, V.B. dan Bressan S. (2004). *Stemming Indosenian without a dictionary,* Ganome Institute dan National University of Singapore