Perancangan Pengindeks Kata pada Dokumen Teks menggunakan Aplikasi Berbasis Web

Herny Februariyanti
Program Studi Sistem Informasi, Universitas Stikubank
email: herny@unisbank.ac.id

Abstrak

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 proses pengindekan data yang. Proses indek kata merupakan salah satu tahapan pada penyiapan basis data untuk keperluan operasi temu kembali informasi. Pengindekan meliputi proses menghilangkan noise, dimana noise pada kalimat diantaranya adalah : imbuhan, angka dan stop word.

Proses indek juga meliputi pengelompokan kata menurut maknanya atau biasa disebut clustering. Dari hasil proses indek dihasilkan basis data yang siap untuk di query untuk diambil informasinya.

Dalam text preprocessing ada beberapa langkah yang perlu dilakukan untuk mendapatkan teks yang bebas derau (noise) atau bebas kata-kata yang tidak bermakna. Selain membebaskan dari derau, text preprocessing juga mengembalikan kata menjadi kata dasar atau root word.Langkah-langkah dalam Text preprocessing dalam bahasa Indonesia adalah :Proses Filtering, Proses Tokenizing, Proses Stemming.

Kata kunci : indek kata, temu kembali

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 proses pengindekan data.

Proses indek kata merupakan salah satu tahapan pada penyiapan basis data untuk keperluan operasi temu kembali informasi. Pengindekan meliputi proses menghilangkan noise, dimana noise pada kalimat diantaranya adalah : imbuhan, angka dan stop word.

Proses indek juga meliputi pengelompokan kata menurut maknanya atau biasa disebut clustering. Dari hasil proses indek dihasilkan basis data yang siap untuk di query untuk diambil informasinya.

Proses indek bisa memakan waktu yang lama tergantung dari besar korpus yang akan diindek, sehingga pada aplikasi pengindek berbasis web diperlukan teknik khusus sehingga proses dapat dilakukan selama mungkin.

Metode yang akan digunakan dalam penelitian ini terdiri dari langkah-langkah sebagai berikut:

1. Obyek Penelitian

   Obyek penelitian dari penelitian ini adalah kata-kata bahasa Indonesia berimbuhan.

2. Data yang diperlukan

   Merupakan data yang mendukung dalam penelitian ini meliputi data primer dan data sekunder.

Data primer :Data yang diperoleh langsung dari kamus besar bahasa Indonesia.

Data Sekunder : Data yang diperoleh dengan membaca dan mempelajari referensi mengenai stemming kata dan pemrograman berbasis web.

3. Teknik Pengumpulan Data

Pengumpulan data dimaksudkan agar mendapatkan bahan-bahan yang relevan, akurat dan reliable. Maka teknik pengumpulan data yang dilakukan dalam penelitian ini adalah sebagai berikut :

Observasi : Dengan melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan kemampuan stemming kata.

Studi Pustaka : Dengan pengumpulan data dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalan dalam penelitian ini.

PENGERTIAN INTERNET

Internet dalam bahasa inggris merupakan singkatan "International Networking ". Pengertian internet secara umum adalah jaringan komputer yang ada di seluruh dunia di mana setiap komputer memiliki alamat (internet Address) yang dapat digunakan untuk mengirim data atau informasi. Dalam hal ini komputer yang dulunya berdiri sendiri menjadi dapat berhubungan langsung dengan host – host atau komputer – komputer yang lainnya. Bentuk data dapat ditransmisikan melalui internet mencakup teks, suara, udara, video, piranti lunak.

Menurut Ause (1997), internet merupakan sekumpulan jaringan yang saling terhubung dengan jaringan lain menggunakan bahasa yang dikenal dengan TCP/IP.

Sedangkan menurut Ellisworth (1995), internet adalah jaringan komunikasi digital yang menghubungkan jaringan – jaringan yang lebih kecil dari banyak negara di seluruh dunia. Internet menggunakan protokol standar yang disebut TCP/IP.

Dari beberapa pengertian internet di atas dapat ditarik kesimpulan bahwa internet adalah merupakan suatu jaringan komunikasi digital global yang dapat menembus batas geografis dan menghubungkan banyak komputer di berbagai negara dengan menggunakan suatu bahasa atau protokol standar yang dikenal dengan nama TCP/IP.

TRANSMISSION CONTROL PROTOCOL

Internet beroperasi menggunakan satu set protokol yang mengontrol dan mengarahkan data di dalam jaringan. Protokol – protokol ini disebut sebagai TCP/IP.

Jaringan besar yang menyusun internet memberikan peluang bagi penggunanya supaya dapat saling berkomunikasi dengan menggunakan dua protokol yaitu TCP dan IP.

Protokol TCP/IP adalah suatu tipe protokol yang di gunakan untuk melakukan komunikasi data dan informasi di internet. Sedangkan protokol sendiri adalah suatu kesatuan prosedur atau bahasa yang memungkinkan 2 atau lebih sistem yang berbeda dapat saling berkomunikasi. Protokol ini merupakan suatu protokol terbuka dimana protokol ini dapat di terapkan dan menghubungkan berbagai sistem tanpa memandang spesifikasi ataupun tipe mesin komputer yang digunakan.

Dalam membawa suatu informasi pada internet merupakan tanggung jawab TCP, di mana TCP memenggal informasi menjadi paket – paket yang berisi data untuk ditransfer dan di susun ulang di tempat tujuan. Lalu IP bertugas memastikan pengiriman data yang akurat ke alamat yang benar.

TCP/IP terdiri dari beberapa layer. Berikut merupakan fungsi dari masing – masing layer TCP/IP adalah :

Physical Layer : Bagian ini berfungsi melewatkan data yang di kirim melalui media fisik seperti konektor dan kabel.

Data Link Layer : Bagian ini berfungsi mempaketkan data ke dalam bentuk frame.

Internet Protokol ; Berfungsi meroute data antar sistem.

TCP: TCP berfungsi meneruskan data dari link layer dan mengubahnya ke dalam bentuk paket.

Application and Service : Bagian ini berfungsi meneruskan paket ke software aplikasi yang biasa digunakan oleh user.

UNIFORM RESOURCE LOCATOR (URL)

URL adalah suatu sarana yang digunakan untuk menentukan lokasi informasi pada suatu web server. URL merupakan cara standar untuk menentukansitus atau halaman pada internet.

URL sama halnya dengan alamat dalam surat biasa yang terdiri dari kode pos dan alamat serta nomor jalan. Begitu juga dengan URL, URL memberikan informasi yang tersedia melalui internet dengan cara standar yang mana menentukan elemen internet seperti lokasi server, dokumen, file dan lain – lainnya.

Format umum URL adalah sebagai berikut :

Protokol_transfer :// nama_host / path / nama_file

Contoh : http :// www.amazon.com/ buku / index.html

Internet yang sangat besar merupakan interkoneksi, terdistribusi, tempat yang tidak seragam dan URL menstandarkan dari keseragaman ini.

DOMAIN NAME SYSTEM (DNS)

Dalam dunia internet, kita bisa masuk ke host – host apapun dengan 2 cara. Cara pertama dan paling efisien adalah dengan mengetik alamat internet protokol atau IP address dari host yang ingin kita tuju. Walaupun ini merupakan cara yang paling efisien tetapi bukan cara yang paling praktis.

Cara yang kedua yaitu yang paling praktis adalah mengakses ke host dengan mengetik nama host yang kita tuju, misalnya www.hotmail.com.

Kebanyakan host IP akan mempunyai cara kedua baik IP address berbentuk numeric maupun nama untuk tetap menjaga kestabilan peningkatan dari nama – nama baru yang semakin bertambah di internet maka dibuatlah DNS (Domain Name System).

DNS merupakan database yang terdistribusi yang mengandung nama host dan informasi IP address serta nama semua domain yang ada di internet. Sebuah nama yang merupakan host dari sebuah server ada pada setiap domain. Misalnya .com yang mengandung semua informasi yang berhubungan DNS tentang domain tersebut. Nama – nama domain yang mempunyai level tinggi (top level domain) dapat di lihat pada tabel 1:

Tabel 1 Macam-macam Domain Name Server

Top level domain	Deskripsi	Contoh
.com	commercial	Microsoft.com Compaq.com
.gov	government	Whitehouse.gov Senate.gov
.mil	military	Army.mil Navy.mil
.edu	education	Umich.edu UMN.edu
.net	Network service	InterNIC.net
		Earthlink.net

PHP

PHP adalah bahasa server-side scripting yang menyatu dengan HTML untuk membuat halaman web yang dinamis. Maksud dari server-side scripting adalah sintaks dan perintah – perintah yang diberikan akan sepenuhnya dijalankan di server tetapi disertakan pada dokumen HTML. Pembuatan web ini merupakan kombinasi antara PHP sendiri sebagai bahasa pemrograman dan HTML sebagai pembangun halaman web.

PHP merupakan software open source (gratis) dan mampu lintas platform, yaitu dapat digunakan dengan sistem operasi dan web server apapun. PHP mampu berjalan di Windows dan beberapa versi Linux. PHP juga dapat di bangun

sebagai modul pada web server Apache dan sebagai binary yang dapat berjalan sebagai CGI.

Keunggulan dari server-side antara lain: (Sutarman,2003)

a. Tidak di perlukan kompabilitas browser atau harus mengunakan browser tertentu, karena serverlah yang akan mengerjakan script PHP. Hasil yang di kirim kembali ke browser umumnya berupa teks atau gambar saja.
b. Dapat memanfaatkan sumber aplikasi yang dimiliki oleh server, misalnya koneksi ke database.
c. Script tidak dapat dilihat dengan fasilitas view HTML source.

MySQL (My Structured Query Language)

MySQL adalah sebuah program pembuat database yang bersifat open source, artinya siapa saja boleh menggunakan dan tidak dicekal (Nugroho, 2004).

My SQL sebenarnya produk yang berjalan pada platform Linux. Karena sifatnya yang open soure, dia dapat dijalankan pada semua platform baik Windows maupun Linux. Selain itu MySQL juga merupakan program pengakses database yang bersifat jaringan sehingga dapat digunakan untuk aplikasi multi user (banyak pengguna). Saat ini database MySQL telah digunakan hampir oleh semua programmer database, apalagi dalam pemrograman web.

Kelebihan dari MySQL adalah ia menggunakan bahasa query standar yang dimiliki SQL (Structured Query Language). SQL adalah suatu bahasa permintaan yang terstruktur yang telah distandarkan untuk semua program pengakses database seperti Oracle, SQL Server dan lain - lain.

Sebagai sebuah program penghasil database, MySQL tidak dapat berjalan sendiri tanpa adanya sebuah aplikasi lain (interface). MySQL dapat di dukung oleh hampir semua program aplikasi baik yang open source seperti PHP maupun tidak, yang ada pada platform Windows seperti Visual Basic, Delphi, dan lainnya.

DIAGRAM ARSITEKTUR INFORMASI

Gerret (2002) mengusulkan sejumlah model visual untuk menggambarkan arsitektur informasi. Konsep yang mendasari usulan Garret adalah :

a. Sistem menunjukkan jalur (*paths*) kepada pemakai.
b. Pemakai berjalan sepanjang jalur melalui sejumlah aksi (*actions*)
c. Aksi tersebut menyebabkan sistem menghasilkan sejumlah hasil (*results*)

Meskipun model visual yang diusulkan oleh Garret sudah dapat digunakan dalam menggambarkan arsitektur informasi, akan tetapi model tersebut mempunyai kelemahan dimana diagram yang digunakan tidak dapat menunjukkan relasi antara kelompok informasi dengan proses yang dibutuhkan untuk menghasilkan informasi tersebut.

Dengan mendasarkan pada konsep yang disajikan oleh Garret maka penulis mengusulkan model visual yang dapat menghubungkan kelompok informasi dengan proses yang diperlukan untuk menghasilkan halaman web tersebut. (Edhi Nugroho, 2003)

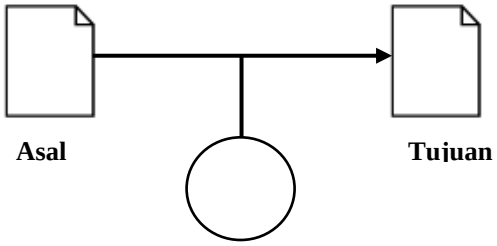
Informasi yang berisi informasi statis digambarkan sebagai sebuah halaman seperti diperlihatkan pada Gambar 1.a Apabila Informasi mempunyai informasi yang lebih rinci maka kelompok informasi tersebut dapat digambarkan dengan menggunakan komponen pada Gambar 1.b

(a) **(b)**

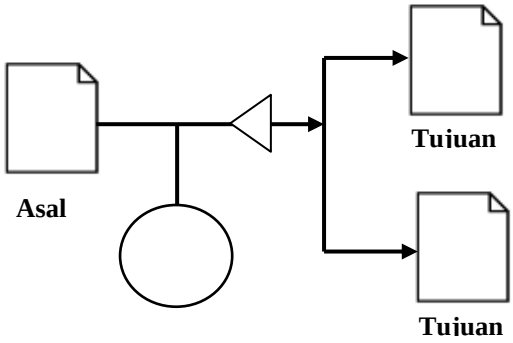
Gambar 1. (a) : Simbol Kelompok Informasi Tunggal; (b) : Simbol Kelompok Informasi Jamak

Kelompok informasi yang berisi informasi dinamis digambarkan seperti kelompok informasi statis tetapi dengan menghubungkan

kelompok informasi tersebut ke proses yang diperlukan untuk menghasilkan kelompok informasi itu. Diagram yang digunakan diperlihatkan pada Gambar 2.a Apabila sebuah proses menghasilkan dua atau lebih kemungkinan hasil maka dapat digunakan tanda segitiga untuk menunjukkan kemungkinan yang muncul. (Gambar 2.b)



Gambar 2. (a) Informasi dinamis yang dihasilkan melalui sebuah proses



Gambar 2.(b) Proses yang menghasilkan kemungkinan dua informasi

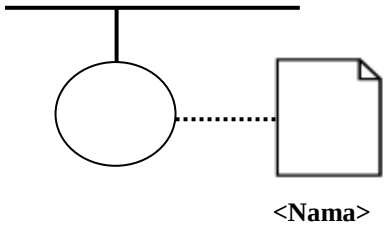
Informasi yang bersifat dinamis seringkali diimplementasikan menggunakan template (pola). Keuntungan dari pemakaian template antara lain :

a. Menyediakan antar muka yang baku.

b. Mempersingkat waktu pengembangan

c. Memudahkan perubahan tampilan informasi.

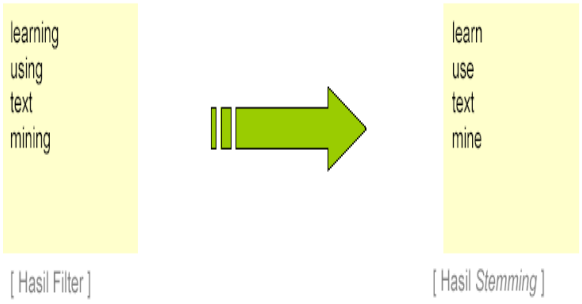
Untuk menggambarkan bahwa sebuah proses menggunakan template maka proses tersebut dihubungkan ke diagram halaman dengan menggunakan sebuah garis putus-putus seperti diperlihatkan pada Gambar 3



Gambar 3. Pemakaian Template

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 4 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 4 Contoh proses stemming bahasa inggris

STEMMER BAHASA INDONESIA

Dalam penelitian oleh ahmad dkk (1996) , dijelaskan bahawa penggunaan kamus sangat memegan 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.

HASIl DAN PEMBAHASAN

Ruang lingkup produk

Sistem ini adalah Rekayasa Perangkat Lunak Komputer berbasis web yang bertujuan untuk melakukan pencarian kata dasar dari sebuah kata. Hal-hal yang diharapkan oleh pengguna agar dapat diwujudkan dalam sistem ini diantaranya adalah hal-hal sebagai berikut :

- Pengguna dapat melakukan proses pencarian kata dasar pada kata yang dimasukan.
- Sistem lain dapat menggunakan fungsi dan prosedur yang digunakan untuk melakukan stemming.
- Aplikasi ini dapat berjalan pada server yang terhubung ke internet ataupun hanya terhubung lokal intranet.

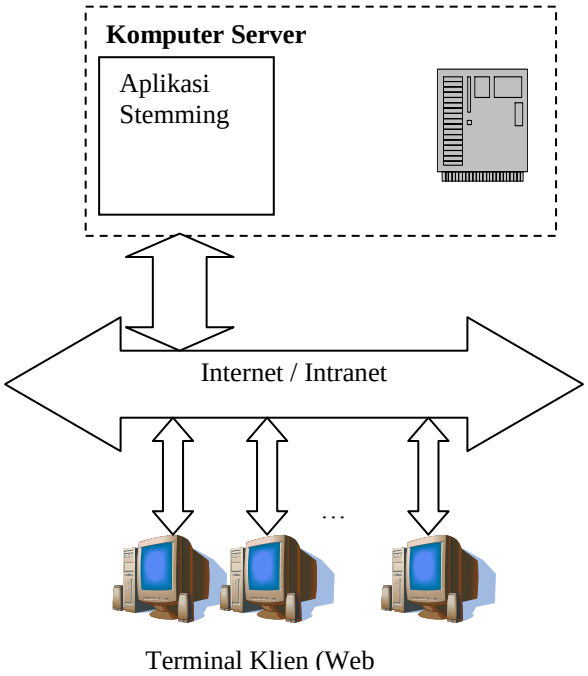
Dalam pengembangan aplikasi ini diharapkan dapat memberikan manfaat sbb :

- Mempermudah pengguna untuk mencari kata dasar pada sebuah kata.
- Mempermudah sistem lain untuk melakukan stemming kata.

Perspektif produk

Aplikasi yang dibangun menggunakan jaringan komputer Client Server. Aplikasi berjalan menggunakan service http dengan format transaksi data html, sehingga dapat dibuka menggunakan terminal yang terkoneksi ke jaringan komputer dan mampu / mempunyai Browser WEB.

Service http dan service basis data menggunakan mesin / komputer yang sama, mengingat aplikasi tidak terlalu membutuhkan resource yang besar. Sedangkan koneksi jaringan menggunakan koneksi internet ataupun intranet dengan protokol TCP/IP. Gambar 5 Menggambarkan perspektif produk aplikasi yang akan dibangun.

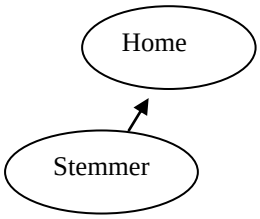


Gambar 5 Perspektif produk

Pada gambar 5 diperlihatkan Komputer Server berfungsi sebagai penyedia layanan aplikasi web dan penyedia layanan RDBMS. Komputer server sebagai server aplikasi dan basis data diakses oleh terminal lainnya melalui jaringan komputer ataupun internet dengan protokol http. Pada terminal klien dibutuhkan aplikasi web browser untuk mengakses aplikasi di server.

Fungsi-fungsi Produk

Produk Aplikasi dibangun dengan antarmuka web, sehingga semua fungsi dapat langsung diakses dari halaman aktif manapun. Dengan demikian fungsi-fungsi yang ada dapat dimanfaatkan oleh pengguna dengan cepat. Gambar 6 merupakan hirarki fungsi dari produk aplikasi



Gambar 6 Fungsi-fungsi produk

Kebutuhan masing - masing fungsi

Pada aplikasi ini terdapat 2 fungsi utama yang dapat digunakan. Administrator sistem dapat menggunakan semua sistem sedang pengguna biasa dapat menggunakan semua fungsi yang ada kecuali fungsi admin dan subfungsinya. Berikut ini penjelasan daru masing-masing fungsi yang tersedia pada aplikasi ini :

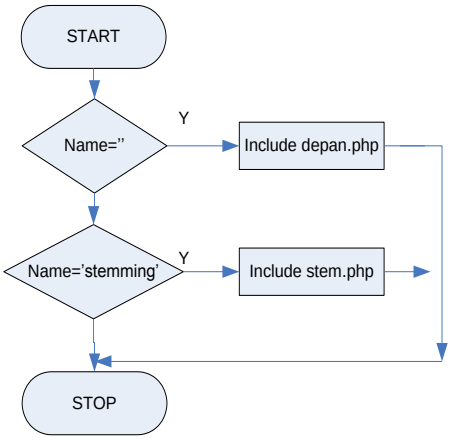
Home : Merupakan tampilan utama / halaman pertama dari aplikasi ini, tidak ada yang ditampilkan selain pejelasan aplikasi ini.

Stemmer : Fungsi ini digunakan untuk menampilkan masukan kata yang akan di stem dan proses semming itu sendiri.

DIAGRAM ALIR APLIKASI

Diagram alir menu utama

Aliran Proses Menu Utama aplikasi diperlihatkan pada gambar 7 Variabel name adalah parameter yang diberi nilai melalui hyperlink Contoh : http://localhost/modules.php?name=stemming.

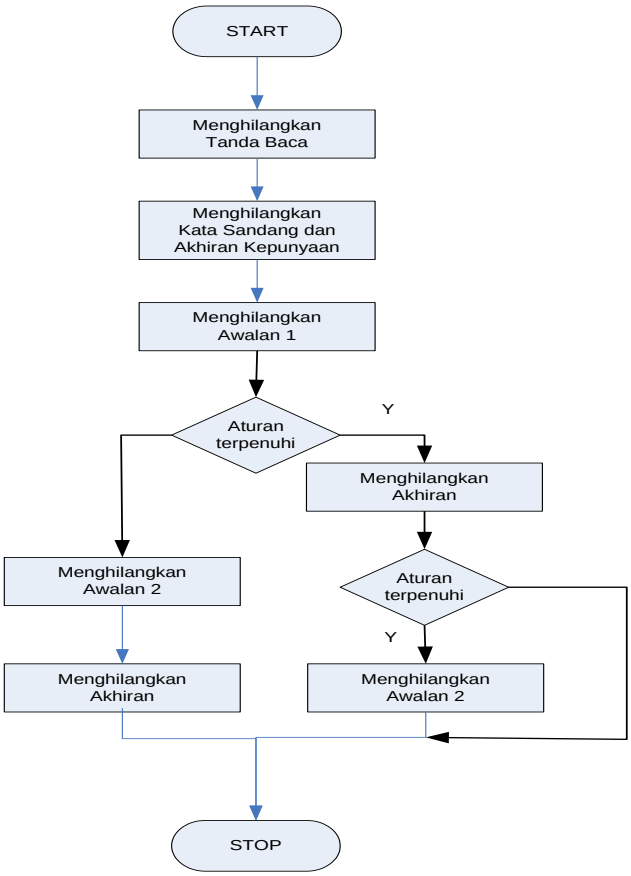


Gambar 7 Diagram aliran proses menu utama aplikasi

Pada gambar 7 diperlihatkan pada saat variabel *name* tidak terdefinisi atau kosong maka modul halaman depan akan dimuat. Sedang apabila variabel *name* berisi stemming maka modul halaman stemming akan dimuat di web.

DIAGRAM ALIR PROSES STEMMING

Aliran Proses Fungsi stemming diperlihatkan pada gambar 8.



Gambar 8 Diagram aliran proses stemming

IMPLEMENTASI TEXT REPROCESSING

Dalam text preprocessing ada beberapa langkah yang perlu dilakukan untuk mendapatkan teks yang bebas derau (noise) atau bebas kata-kata yang tidak bermakna. Selain membebaskan dari derau, text preprocessing juga mengembalikan kata menjadi kata dasar atau root word (Baesa, 1998).

Langkah-langkah dalam Text preprocessing dalam bahasa Indonesia adalah :

a) Proses Filtering.

b) Proses Tokenizing

c) Proses Stemming.

Proses Filtering penghilangan tanda baca dan angka dilakukan sebelum dilakukan tokenizing. Hal ini dilakukan untuk menghemat waktu ekseskusi setiap dokumennya.

Untuk mengatasi hal tersebut maka proses text preprocessing dilakukan tiap satu persatu dokumen, maksudnya adalah hanya 1 dokumen yang akan diproses setiap waktunya oleh web server setiap kali url program di muat / dipanggil. Kemudian menggunakan mekanisme variabel session untuk menyimpan data pointer posisi dokumen terakhir diproses. Sehingga setiap kali url dimuat maka pointer akan bergeser ke dokumen selanjutnya sampai pointer menunjuk pada dokumen terakhir.

Mekanisme pemanggilan / pemuatan ulang url program secara otomatis dapat menggunakan bantuan javascript autoreload. Setiap kali script autoreload dipanggil maka browser secara otomatis memanggil / memuat ulang halaman tersebut, demikian seterusnya sampai semua dokumen selesai diproses. Pada algoritma 1 diperlihatkan pseudo code proses implementasi mekanisme autoreload text preprocessing pada aplikasi berbasis web.

Algoritma 1 *Pseudo code proses implementasi mekanisme autoreload.*

```
<iframe autoreload >
        $pos=$_SESSION[pos];
        $max=$_SESSION[max];
        $id=$_SESSION[id];
  while (($pos<$max) and ($pos<$pos+10)) {
    TextPreprocessing($id[$pos]);
    $pos++;
        } else {
          Halt("Proses Selesai");

        }
  $_SESSION[pos]:=$pos;
        echo "<script type='text/javascript'>
        window.onload=
        setTimeout('window.location.reload()',1)
        ;</script>";
</iframe>
```

Pada pseudo code diperlihatkan setiap pemanggilan program hanya akan diproses dokumen sebanyak 10 buah saja, pembatasan ini untuk memotong proses menjadi lebih kecil. Setelah program selesai dijalankan, program akan dipanggil ulang oleh javascript autoreload pada bagian bawah.

IMPLEMENTASI TEXT

Sebelum kata dipisahkan dari kalimatnya, terlebih dahulu dibersihkan dari tanda baca, tag html dan angka. Untuk membersihkan dapat digunakan perintah ekspresi regular yang ada pada bahasa pemrograman PHP. Pembersihan dilakukan sebelum proses tokenizing dimaksudkan untuk memperkecil hasil dari tokenizing. Dengan demikian diharapkan keluaran dari tokenizing berupa kata-kata yang bersih dari tanda baca, tag html dan angka.

Proses pembersihan tanda baca dan angka diperlihatkan pada pseudo code pada algoritma 2

Algoritma 2 *Pseudo code proses pembersihan tanda baca dan angka*

```
while
(ereg("<(/?[[:alpha:]]*)[[:space:]]*([^>]*)>",$st
r,$reg)) {
   $i = strpos($str,$reg[0]);
   $l = strlen($reg[0]);
   $tag = "";
   $tmp .= substr($str,0,$i) . $tag;
   $str = substr($str,$i+$l);
}
$str = $tmp . $str;
$str=eregi_replace (chr(13), " ", $str);
$str=eregi_replace (chr(10), " ", $str);
$str=eregi_replace("([(,),%,=,.,?,<,>,-
,:,;,`,~,!,@,#,$,%,^,&,*,+,\,|,{,},',/])"," ",$str);
$str=str_replace('[',' ',$str);
$str=str_replace(']',' ',$str);
$str=str_replace('-',' ',$str);
$str=str_replace('"',' ',$str);
$str=eregi_replace("([0-9])"," ",$str);
$str=eregi_replace(" ([a-z]) "," ",$str);
$str=eregi_replace(" ([a-z])([a-z]) "," ",$str);
return $str;
```

Proses filtering selanjutnya dilakukan setelah kata di stem dan tersimpan dalam tabel master kata, transaksi judul kata dan transaksi abstrak kata. Proses filter tersebut menghilangkan kata-kata yang masuk didalam daftar stopword.

Implementasi text tokenizing

Pada kalimat, pemisah antar kata adalah karakter spasi. Sehingga proses deteksi token dapat dilakukan dengan melihat keberadaan karakter spasi. Pada pemrograman PHP terdapat perintah untuk mengubah string menjadi array dengan pemisah karakter tertentu. Perintah *explode([separator],[teks])* dapat digunakan dengan mengisi [teks] dengan variabel string dan [separator] diisi dengan karakter spasi. Setelah perintah dieksekusi, semua kata akan terpisah dari string dan tersusun dalam suatu array.

Setelah token dideteksi maka array hasil dari deteksi tersebut diolah oleh proses berikutnya. Pemrosesan pada proses berikutnya dilakukan kata-perkata untuk meringankan proses.

Implementasi Proses Indek

Setelah kata telah dikembalikan dalam bentuk asal (kata dasar), kata-kata tersebut disimpan dalam master kata, kemudian untuk setiap kata yang tampil di judul disimpan pada tabel transaksi judul kata, demikian pula setiap kata yang ada pada abstraksi disimpan pada tabel transaksi abstraksi kata. Sebelum dilakukan pengindekan terlebih dahulu tabel master kata, abskata dan judul kata dibersihkan dari stopword..

Tabel artikel berelasi dengan tabel master kata menghasilkan tabel transaksi judulkata. Berikut ilustrasi tabel master artikel pada tabel 2 yang berisi IDaritkel sebagai key dan judul yang berisi string dari judul artikel. Setelah melalui proses preprocessing maka akan dihasilkan tabel 3 yang berisi kata-kata yang pernah digunakan di judul artikel dengan key idkata. Setelah proses preprocessing selain menghasilkan tabel master kata, akan dihasilkan juga tabel transaksi judulkata pada tabel 4. Pada tabel 4 pada kolom pertama diperlihatkan bahwa Idartikel 1 mempunyai kata dengan.id 1 sebanyak 1 buah, demikian seterusnya.

Tabel 2 Tabel master artikel.

IDArtikel	Judul
1	Tanaman Obat untuk Sakit Kepala
2	Obat Sakit Kepala Untuk Anak Balita
3	Kelainan Kepala Pada Balita

Tabel 3 Tabel master kata

IDKata	Kata
1	Tanam
2	Obat
3	Sakit
4	Kepala
5	Anak
6	Balita
7	Lain

Tabel 4 Tabel transaksi judulkata

IDArtikel	**IDKata**	Jumlah
1	1	1
1	2	1
1	3	1
1	4	1
2	2	1
2	3	1
2	4	1
2	5	1
2	6	1
3	7	1
3	4	1
3	6	1

KESIMPULAN

Pembatasan waktu eksekusi pada sistem informasi berbasis web dapat dihindari dengan mekanisme autoreload, membagi pemrosesan dokumen dan melakukan proses per-dokumen sehingga meningkatkan jumlah dokumen yang mampu diproses dan terhidar dari terminasi proses oleh server.

Penggunaan basisdata untuk menyimpan data indek dapat mempercepat proses pencarian kata untuk temu kembali informasi.

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. Dapat juga diteliti pada corpus yang sama tetapi dengan jumlah dokumen yag lebih banyak 5000> sehingga dapat diukur perfoma dan kemampuan sistem.

DAFTAR PUSTAKA

Iain Shigeoka, 2002, *Instant Messaging in Java The Jabber Protocols*, Manning Publications Co.,

Miller J. P. Saint-Andre, 2003, "*XMPP Core Draft-IETF-XMPP-Core-12*" www page, May 2003, Expire on November 2,

Robin Cover, 2002, "*IETF Charters Extensible Messaging and Presence Protocol(XMPP) Working Group*.," WWW page, http://xml.coverpages.org/.

Stephen Lee and Terence Smelser, 2002, *Jabber Programming*, Hungry Minds, Inc.,

http://www.jabber.org "*What Is jabber*," www page, 2003,.