BAB I
PENDAHULUAN

1.1. LATAR BELAKANG

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.

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. Halaman web biasanya dibatasi waktu eksekusi oleh server selama 30 detik, apabila eksekusi lebih dari 30 detik maka proses dianggap overtime dan dimatikan paksa oleh web server.

1.2. PERUMUSAN MASALAH

Berdasarkan latar belakang di atas, maka permasalahan yang dapat dirumuskan adalah bagaimana membuat aplikasi berbasis web yang mampu mengindek kata pada dokumen teks berbahasa Indonesia.

1.3. BATASAN MASALAH

Dalam penelitian ini ada beberapa pembatasan masalah yang dilakukan, yaitu: hanya melakukan implementasi sampai dengan pembuatan fitur pengindek kata pada artikel kesehatan.

1.4. TUJUAN PENELITIAN

Tujuan yang ingin dicapai dalam penelitian ini adalah merancang dan membuat aplikasi berbasis web yang dapat melakukan proses indek pada dokumen berbahasa Indonesia.

1.5 METODOLOGI PENELITIAN

1.5.1. Obyek Penelitian

Obyek penelitian dari penelitian ini adalah artikel kesehatan berb.ahasa indonesia

Data Yang diperlukan

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

Data primer : Data yang diperoleh langsung dari internet dalam hal ini berita dari departemen kesehatan RI.

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

1.5.2 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 :

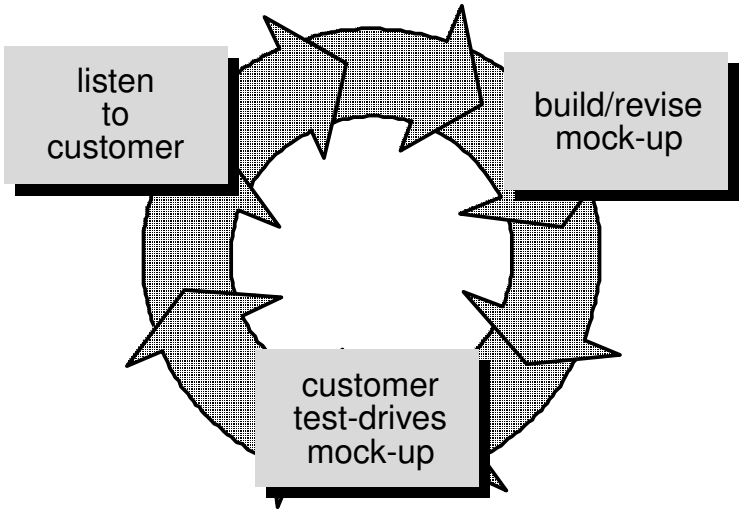
Observasi : Dengan melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan kemampuan pengindekan dokumen.

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

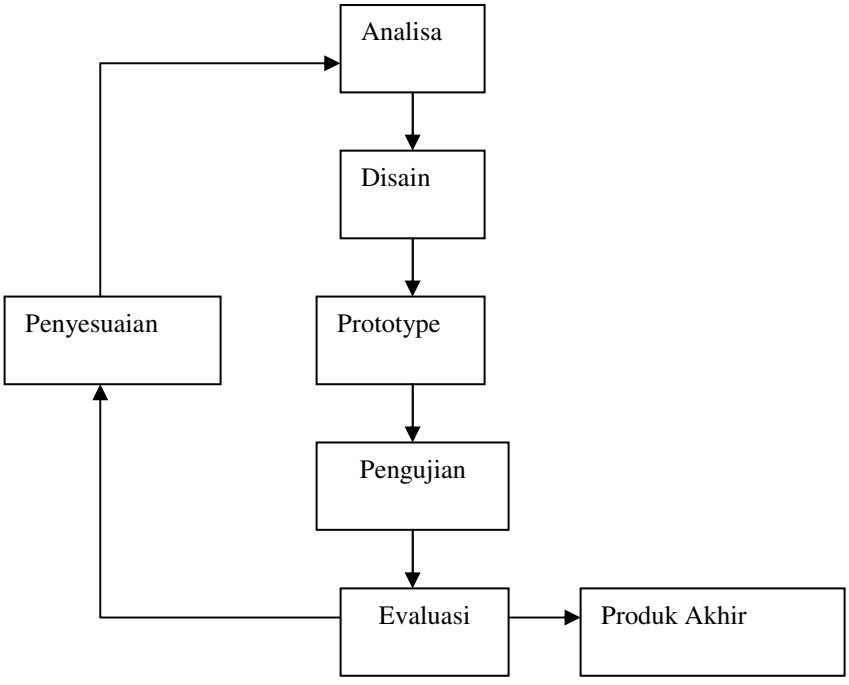
1.5.3. Metode Pengembangan

Penelitian ini menggunakan model *prototyping* . di dalam model ini sistem dirancang dan dibangun secara bertahap dan untuk setiap tahap pengembangan

dilakukan percobaan-percobaan untuk melihat apakah sistem sudah bekerja sesuai dengan yang diinginkan. Sistematika model *prototyping* terdapat pada gambar 1.1, dan pada gambar 1.2 memperlihatkan tahapan pada prototyping



Gambar 1.1. Sistematika Prototyping (Pressman, 1997:40)



Gambar 1.2. Tahapan Protoyping

Berikut adalah tahapan yang dilakukan pada penelitian ini dengan metode pengembangan prototyping.

Analisa : Pada tahap ini dilakukan analisa tentang masalah penelitian dan menentukan pemecahan masalah yang tepat untuk menyelesaikannya. Dalam tahap ini juga dilakukan penyusunan Software Requirement Spesification.

Disain : Pada tahap ini dibangun rancangan sistem dengan beberapa diagram bantu DFD dan ER-D. Perancangan alur program juga dilakukan tahap ini dengan menggunakan diagram Flow Chart.

Prototype : Pada tahap ini dibangun aplikasi berbasis web yang sesuai dengan disain dan kebutuhan sistem.

Pengujian: Pada tahap ini dilakukan pengujian hasil dari proses indek pada aplikasi yang dibangun.

Evaluasi : Pada tahap ini dilakukan evaluasi apakah performa aplikasi sudah sesuai dengan yang diharapkan, apabila belum maka dilakukan penyesuaian-penyesuaian secukupnya.

Penyesuaian : Tahap ini dilakukan apabila pada evaluasi performa aplikasi kurang memadai dan dibutuhkan perbaikan, tahap ini melakukan penyesuaian dan perbaikan pada aplikasi sesuia dengan kebutuhan

1.6 Sistematika Penulisan

Sistematika penulisan terdiri dari empat bab yang masing-masing bab menguraikan hal-hal yang berbeda.

Bab I Pendahuluan

Pada bab ini diuraikan mengenai permasalahan yang dibahas secara umum yang meliputi : latar belakang, perumusan dan pembatasan masalah, tujuan dan manfaat

penelitian, metode pengumpulan data, metodologi penelitian serta sistematika penelitian.

Bab II Landasan Teori

Dalam Bagian ini memuat hal-hal teoritis yang ada hubunganya dengan penyelesaian masalah dalam penelitian ini. Pada bab ini di uraikan antara lain tentang mysql, php dan Pengindekan kata.

Bab III Perancangan Sistem

Dalam bagian ini dibahas tentang rancang bangun aplikasi. Rancangan meliputi struktur program, DFD, E-R Diagram dan Flowchart.

Bab IV Implementasi

Dalam bagian ini dibahas tentang langkah-langkah implementasi untuk perangkat lunak aplikasi yang telah selesai di rancang pada bab III. Disini disertakan juga kode sumber dari fungsi-fungsi utama.

Bab V Kesimpulan Dan Saran

Pada bab ini berisi kesimpulan dan saran dari penelitian ini. Selain hasil penelitian berupa perangkat lunak aplikasi, juga dihasilkan juga saran-saran untuk penelitian lebih lanjut pada bidang pengindekan.

BAB II
LANDASAN TEORI

2.1 Internet

2.1.1 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 : 1), internet merupakan sekumpulan jaringan yang saling terhubung dengan jaringan lain menggunakan bahasa yang dikenal dengan TCP/IP.

Sedangkan menurut Ellisworth (1995 : 437), 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.

2.2 Protokol internet

2.2.1 Pengertian protokol internet

Protokol dalam dunia komunikasi data komputer digunakan untuk mengatur bagaimana sebuah komputer berkomunikasi dengan komputer lain. Komputer yang

terhubung ke internet berkomunikasi dengan menggunakan protokol yang sama, karena protokol ini berfungsi mirip dengan bahasa.

2.2.2 Hypertext Transport Protokol (HTTP)

Salah satu protokol yang sering digunakan dalam dunia internet adalah HTTP (Hypertext Transport Protokol). Protokol HTTP ini digunakan untuk berbagai jenis layanan dalam WWW (Worl Wide Web) di jaringan TCP/IP. Protokol HTTP juga dapat digunakan untuk berkomunikasi antara web browser dan web server satu sama lain.

HTTP akan kita gunakan jika pemakai hendak mengakses suatu website tertentu. HTTP memiliki tugas yaitu untuk mentransfer dokumen atau file berupa hypertext yang dalam pelaksanaannya dikenal dengan nama HTML.

Dengan demikian HTTP akan menstransfer HTML ke browser dari server tempat HTML tersebut di simpan. Protokol HTTP berifat request response, yaitu dalam protokol ini client menyampaikan pesan request ke server dan server kemudian akan memberikan respon yang sesuai dengan request tersebut.

Protokol HTTP ini pada dasarnya keseluruhan beroperasi tanpa sepengetahuan pemakai, tidak setiap pemakai diwajibkan perlu tahu TCP/IP bila pemakai hanya sekedar menggunakan internet atau web untuk kebutuhannya.

2.2.3 Transmission Control Protokol / Internet Protokol (TCP/IP)

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.

2.3 Teori Internet Service

2.3.1 World Wide Web (WWW)

World Wide Web (WWW) adalah jaringan komputer yang terdiri dari client dan server dengan menggunakan software khusus membentuk sebuah jaringan yang disebut jaringan client-sever.

WWW juga merupakan jaringan dokumen yang sangat besar yang saling dihubungkan satu sama lain, satu set protokol yang mendefinisikan bagaimana sistem bekerja dan menstransfer data, dan sebuah perangkat lunak yang membuatnya bekerja dengan mulus.

WWW ada 2 hal penting yaitu web server dan web browser. Informasi yang di letakkan di WWW disebut “homepage” dan setiap homepage memiliki alamatnya

masing – masing. WWW menggunakan teknik hypertext dan multimedia yang membuat internet mudah digunakan dan di jelajahi.

2.3.2 Electronic Mail (E-mail)

E-mail merupakan cara pengiriman surat atau pesan secara elektronik. E-mail juga merupakan penggunaan teknologi pasar elektronik yang mengijinkan pengguna komputer untuk berkomunikasi dengan pengguna komputer lainnya dengan berbagai tujuan. E-mail menjadi salah satu alasan mengapa komputer saling terhubung. Transfer E-mail yang lebih cepat adalah server menstransfer E-mail dengan menggunakan STMP (Single Mail Transfer Protokol).

Dengan E-mail dapat mengirim file – file berupa program, gambar, grafik, video dan lain sebagainya. Serta dapat juga mengirim ke lebih dari 1 orang sekaligus pada saat bersamaan tanpa mengenal batas ruang dan waktu.

2.3.3 Feed Back

Merupakan pesan umpan balik dari konsumen yang berisi penilaian terhadap suatu proses layanan yang diberikan oleh perusahaan.

2.3.4 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.

2.3.5 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 2.1:

Tabel 2.1 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

2.3.6 Hypertext Markup Lannguage (HTML)

HTML adalah suatu sistem yang digunakan untuk menandai dokumen dengan pembatas informasional yang mengindikasikan bagaimana teks pada dokuman harus direpresentasikan dan bagaimana dokumen dihubungkan satu sama lain. HTML sendiri termasuk turunan dari SGML (Standard Generalized Markup Language) yang merupakan bahasa standar untuk markup.

Dokumen HTML disebut sebagai markup language karena mengandung tanda tertentu yang digunakan untuk menentukan tampilan suatu teks dan tingkat kepentingan dari teks tersebut pada suatu dokumen. HTML juga mendukung multimedia secar penuh, karena dapat menampilakan seluruh komponen multimedia (text, hypertext, gambar, animasi, audio, video).

2.4 Pemrograman Internet

Agar website yang kita tampilkan dapat bersifat dinamis dan informasi yang akan di tampilkan pada internet dapat di tampilkan dengan baik, maka diperlukan suatu program. Adapun program yang diperlukan untuk pembuatan website dinamis adalah :

2.4.1 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,2003:109)

- 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.
- Dapat memanfaatkan sumber aplikasi yang dimiliki oleh server, misalnya koneksi ke database.
- Script tidak dapat dilihat dengan fasilitas view HTML source.

2.4.2 Web Browser

Web browser merupakan aplikasi yang memungkinkan pengguna untuk menjelajahi world wide web untuk mendapatkan informasi dan berkomunikasi. Pengguna hanya mengetahui alamat halaman web yang dimaksud. Kemudian web browser menunggu informasi yang diminta dikirimkan kembali oleh web server, sehingga pengguna dapat melihat informasi tersebut dari web browser. Contoh web browser : Nescape Communicator, Microsoft Internet Explorer, Opera, dan lain – lain.

Fungsi utama browser adalah :

- Memungkinkan untuk mengambil dan melihat informasi dari komputer server www, gopher, dan FTP di internet, atau media disk yang berisi dokumen HTML.
- Berinteraksi dengan sistem berbasiskan server.
- Merupakan alat untuk melihat dokumen elektronik
- Untuk melakukan download / upload informasi digital.
- Untuk mengirim dan menerima e-mail.

2.4.3 ApacheWeb Server

Web server adalah suatu program yang terletak pada komputer dengan akses internet, yang merespon permintaan browser untuk suatu URL. Web server memenuhi kebutuhan pengguna dengan mensuplai atau melayani permintaan halaman web.

Jadi, halaman web harus diletakkan dalam web server agar dapat dilihat dari internet. Idealnya, web server harus memiliki koneksi internet yang tidak bisa terputus, sehingga halaman – halaman yang ditangani dapat selalu tersedia.

Apache merupakan pengembangan dari server yang dikeluarkan oleh NSCA yaitu NSCA HTTP pada tahun 1995 dan saat ini merupakan tulang punggung dari

World Wide Web (WWW). Apache berfungsi memenuhi permintaan dari client dengan browaer seperti Internet Ekxplorer, Mozilla.

2.4.4 Macromedia Dreamweaver MX

Macromedia Dreamweaver adalah sebuah HTML editor professional untuk mendesign secara visual dan mengelola situs web maupun halaman web. Dreamweaver membuat menjadi lebih mudah dengan menyediakan tool – tool yang sangat berguna dalam meningkatkan kemampuan dan pengalaman dalam membuat web.

Dreamweaver MX juga terdapat banyak tool untuk kode – kode dalam hal web beserta fasilitas – fasilitasnya, antara lain : referensi HTML, CSS, Javascript, Javascript debugger, dan editor code yang mengijinkan pengeditan kode javascript, XML, dan dokumen teks lain secara langsung dalam dreamweaver.

2.4.5 MySQL (My Structured Query Language)

MySQL adalah sebuah program pembuat database yang bersifat open source, artinya siapa saja boleh menggunakan dan tidak dicekal (Nugroho, 2004:29).

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.

2.5 Diagram Arsitektur Informasi

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

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

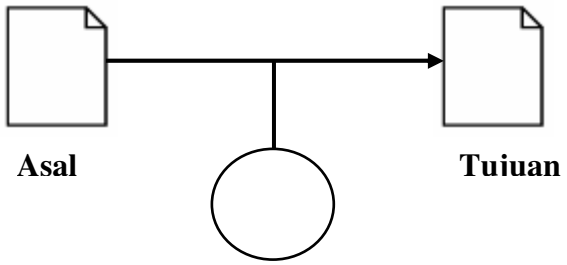
Meskipun model visual yang diusulkan oleh Garret sudah dapat digunakan dalam menggambarkan arsitektur informasi, akan tetapi model tersebut mempunyai kelemahan dimana diagram yang digunakan tidak dapat menunjukkan relasi antara kelompok informasi dengan proses yang dibutuhkan untuk menghasilkan informasi 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 :8)

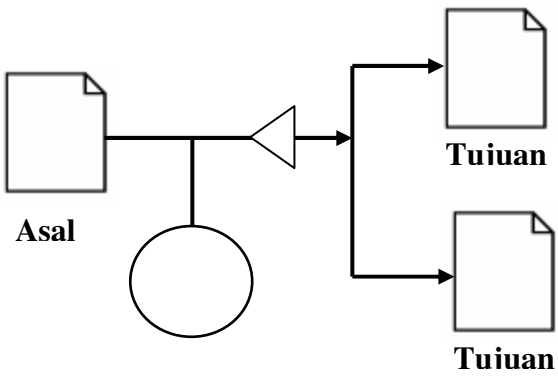
Informasi yang berisi informasi statis digambarkan sebagai sebuah halaman seperti diperlihatkan pada Gambar 2.1.a Apabila Informasi mempunyai informasi yang lebih rinci maka kelompok informasi tersebut dapat digambarkan dengan menggunakan komponen pada Gambar 2.1.b

Gambar 2.1. (a) *kiri*: **Simbol Kelompok Informasi Tunggal; (b)** *kanan*: **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 2.2.a Apabila sebuah proses menghasilkan dua atau lebih kemungkinan hasil maka dapat digunakan tanda segitiga untuk menunjukkan kemungkinan yang muncul. (Gambar 2.2.b)



Gambar 2.2. (a) Informasi dinamis yang dihasilkan melalui sebuah proses

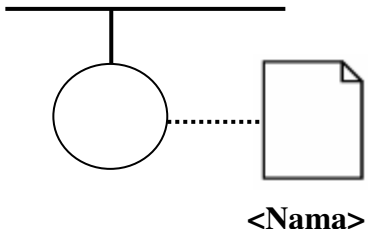


Gambar 2.2.(b) Proses yang menghasilkan kemungkinan dua informasi

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

- Menyediakan antar muka yang baku.
- Mempersingkat waktu pengembangan
- 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 2.3



Gambar 2.3. Pemakaian Template

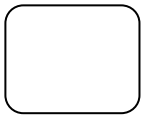
Jalur informasi digambarkan sebagai sebuah panah dengan arah panah menunjukkan arah informasi berikutnya yang dapat diakses oleh pemakai (Gambar 2.4).

Gambar 2.4. Komponen Arah informasi

Keterangan mengenai kondisi yang menyebabkan jalur tersebut dipilih dapat diletakkan di atas atau di bawah tanda panah.

Situs eksternal adalah situs yang berada di luar situs yang sedang diakses oleh pemakai. Halaman yang berada di situs eksternal digambarkan dengan menggunakan *rounded rectangle* seperti dicontohkan pada Gambar 3.6. Sebuah halaman dianggap

berada di situs eksternal apabila alamat URL dari situs tersebut tidak sama, sebagian atau seluruhnya, dari alamat situs yang sedang diakses. Sebagai contoh alamat URL :http://www.unisbank.ac.id/info akan dianggap sebagai situs eksternal apabila pemakai sedang mengakses halaman web yang berada di alamat : http://www.unisbank.ac.id/info_dosen.

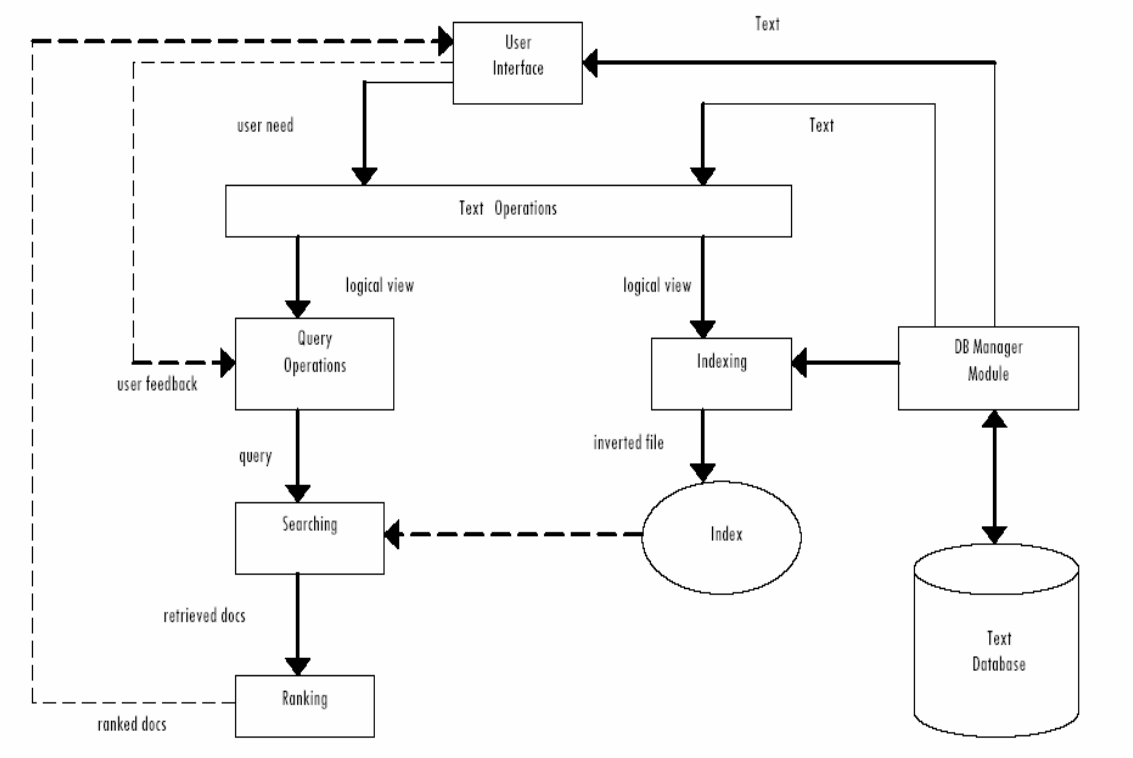


Gambar 2.5. Komponen Situs Eksternal.

2.6 Sistem Temu Kembali Informasi

Sistem temu-kembali teks (*teks retrieval)* adalah sistem penemuan kembali informasi dalam bentuk dokumen dengan mengukur kemiripan (*similarity*) antara informasi yang tersimpan dalam basis data dengan query yang dimasukkan oleh pengguna (Baesa dkk, 1998:19).

Teknik pencarian informasi pada sistem Information Retrieval berbeda dengan sistem pencarian pada sistem manajemen basisdata (DBMS) . Dalam sistem temu kembali terdapat dua bagian utama yaitu bagian pengindeksan (*indexing*) dan pencarian (*searching*). Kedua bagian tersebut memiliki peran penting dalam proses temu kembali informasi. Gambar 2.6 menggambarkan proses temu kembali informasi.



Gambar 2.6 Proses temu kembali informasi (Baesa dkk , 1998:10)

DB Manager Module **:** Bersama *Text Database* berfungsi melakukan layanan basis data penyimpanan teks untuk dapat digunakan oleh text operation dan oleh user. Modul ini dapat berupa layanan RDBMS ataupun sekedar perantara untuk mengakses data teks.

User Interface **:** Digunakan untuk berkomunikasi dengan pengguna, sehingga pengguna dapat menggunakan sistem dengan mudah dan cepat. Pada user interface yang baik, pengguna menemukan dokumen yang dimaksud dan tidak merasa menjawab pertanyaan dari sistem untuk menghasilkan pencarian dokumen yang dikehendakinya.

Text Operations **:** mengolah data teks dari DB Manager Module sesuai dengan kebutuhan pengguna. Hasilnya digunakan untuk proses indek berupa teks yang telah diproses dan fitur-fitur yang ada yang nantinya digunakan untuk operasi query.

Query Operation **:** digunakan untuk memanajemen query yang akan dieksekusi. Query dihasilkan dari serangkaian nilai dari jawaban pengguna (*user feedback*) dan dari hasil Text Operation berupa fitur-fitur yang akan dicari. Query tersebut akan dieksekusi ke tabel indek oleh bagian *Searching* dan kemudian menghasilkan daftar dokumen terurut berdasarkan rangkingnya (*Ranking*) yang akan di tampilkan pada User interface.

2.7 Text Mining

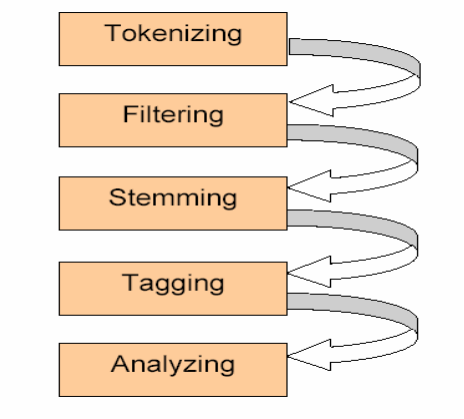
Data mining sendiri digunakan pada proses pengindeksan pada Sistem temu kembali informasi. Pada proses pengindeksan informasi-informasi penting di ekstrak dari dokumen-dokumen yang ada.

Text Mining merupakan bagian dari data mining dimana data mining sendiri mempunyai banyak arti diantaranya adalah : Proses pencarian informasi yang berharga dari data dengan ukuran besar. Data mining juga di definisikan sebagai Ekplorasi dan analisa data ukuran besar untuk menemukan pola-pola dan aturan-aturan yang bermanfaat. Tetapi datamining dapat didefinisikan dengan sederhana yaitu : mengekstrak atau menambang pengetahuan yang bermanfaat dari data berukuran besar (Kamber dan Han 2000:6)

Menurut Salton tipe informasi dapat dikategorikan menjadi 3 macam yaitu informasi berformat teks, informasi berformat suara dan informasi berformat grafik ataupun gambar (Salton 1989:4). Text mining atau sering disebut text data mining dalam bahasa Indonesia disebut dengan penambangan data teks merupakan proses penambangan data berformat teks dari suatu dokumen. Dengan penambangan teks, dapat dicari kata-kata yang dapat mewakili isi dari suatu dokumen. Suatu artikel berita dapat dianalisis apakah artikel berita tersebut tersebut termasuk ke dalam

kategori olah raga, kesehatan, selebriti, kriminal, ekonomi, politik atau yang lain, dicocokkan dengan database kata kunci yang sebelumnya telah dibuat. Sehingga diharapkan dapat membantu sistem redaksi elektronik untuk dapat memilah atau mengetahui kategori dari sebuah artikel berita tanpa memerlukan seorang editor. Hal ini akan menghemat waktu dan biaya dalam menjalankan bisnis pada model kantor berita elektronik on-line berbasis internet (Adrifina dkk 2008).

Pada gambar 2.7 diperlihatkan tahapan-tahapan yang umum dilakukan pada saat melakukan penambangan teks. Proses penambangan melibatkan 5 proses yaitu : a) Tokenizing; b) Filtering; c) Stemming; d) Tagging; e) Analyzing.

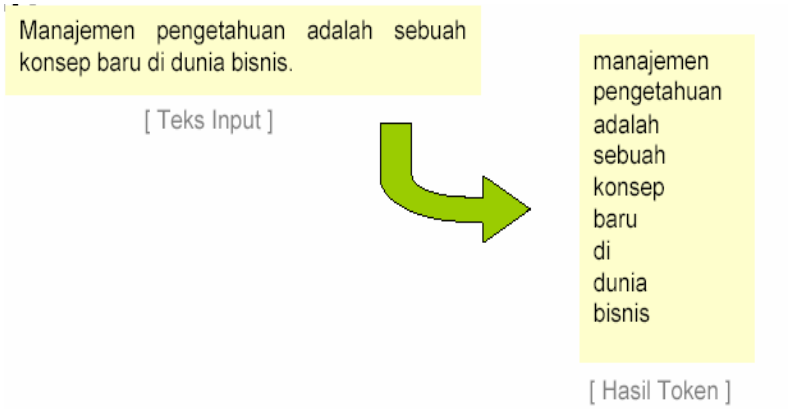


Gambar 2.7 Tahapan penambangan teks

2.7.1 Tokenizing

Proses tokenizing adalah proses pemotongan string masukan berdasarkan tiap kata yang menyusunnya. Pada prinsipnya proses ini adalah memisahkan setiap kata yang menyusun suatu dokumen. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenizing mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata. Gambar 2.8 merupakan gambaran dari hasil proses tokenizing. Pada gambar tersebut diperlihatkan serangkaian kalimat utuh, yang dipisahkan oleh spasi setiap katanya, setelah melalui proses tokenizing maka kalimat tersebut menjadi sekumpulan array yang setiap selnya berisi kata-kata yang ada pada kalimat tersebut. Pada proses

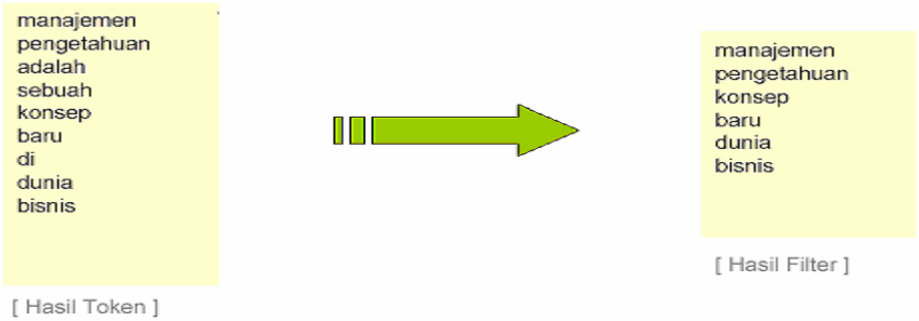
tokenizing biasanya juga ditambahkan informasi jumlah kemunculan setiap kata pada kalimat tersebut.



Gambar 2.8 Contoh tokenizing

2.7.2 Filtering

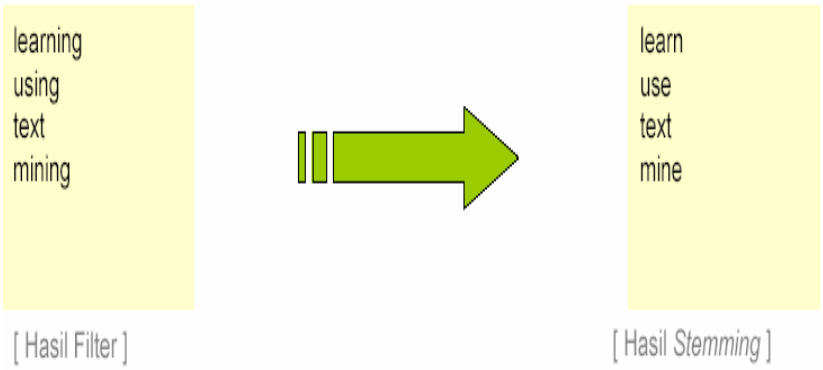
Proses Filtering adalah proses pengambilan kata-kata yang dianggap penting atau mempunyai makna saja. Pada proses ini kata-kata yang dianggap tidak mempunyai makna seperti kata sambung akan dihilangkan. Pada proses ini biasanya digunakan Stop Word List yang tersimpan dalam suatu tabel basis data, yang nantinya digunakan sebagai acuan penghilangan kata. Stop word list berbeda untuk setiap bahasanya. Gambar 2.9 merupakan gambaran dari hasil proses filtering, pada gambar tersebut diperlihatkan kata seperti 'di', 'adalah' dan 'sebuah' melalui proses penghilangan, karena kata-kata tersebut tidak mempunyai makna dan hanya berfungsi sebagai kata sambung saja.



Gambar 2.9 Contoh filtering dengan stop word

2.7.3 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 2.10 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.

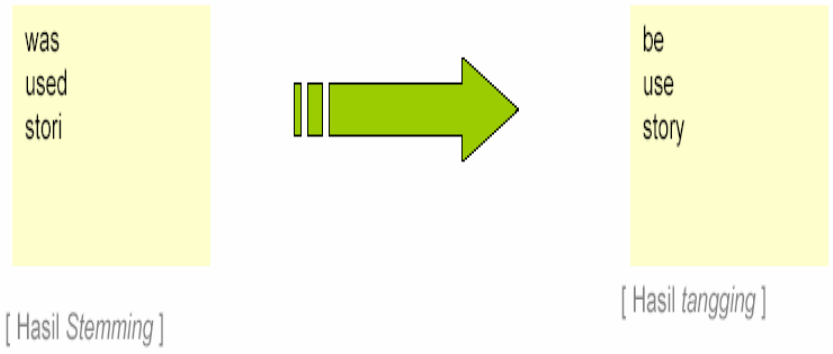


Gambar 2.10 Contoh proses stemming bahasa inggris

2.7.4 Tagging

Proses tagging adalah mencari bentuk utama/root dari suatu kata lampau. Proses tagging tidak digunakan pada dokumen berbahasa Indonesia dikarenakan bahasa Indonesia tidak mengenal kata bentuk lampau. Gambar 2.11 merupakan gambaran

dari proses tagging, pada gambar tersebut diperlihatkan kata *was* berubah ke bentuk pertama *be*, kata *used* berubah ke bentuk pertama *use*.



Gambar 2.11 Contoh proses tagging bahasa inggris

2.7.5 Analyzing

Proses analyzing adalah proses analisa dari hasil proses tagging sehingga diketahui seberapa jauh tingkat keterhubungan antar kata-kata dan antar dokumen yang ada. Ada 3 pendekatan untuk melakukan pembobotan hubungan antar dokumen yaitu (Baesa dkk, 1998:25) :

***Model Boolean*:** Model ini merepresentasikan dokumen-dokumen dengan himpunan dari isitilah-isitilah dokumen, dan sebuah query dengan ekspresi boolean dari istilah-isitilah query. Banyak mesin pencari informasi didasarkan pada model ini. Suatu kecocokan diantara sebuah dokumen dan sebuah query biasanya diturunkan dengan menggunakan operasi teori himpunan boolean. Pada himpunan istilah-istilah dokumen dan istilah-istilah query.

***Model Vektor* :** Vector Space Model merepresentasikan dokumen dan query dengan vektor-vektor bobot istilah dalam sebuah ruang multidimensi. Dalam VSM, sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Jadi, relevansi sebuah dokumen ke sebuah query didasarkan pada similaritas diantara vektor dokumen dan vektor query. Similaritas dari dua vektor biasanya dihitung dengan

sudut, yakni, **cosine measure**, diantara dua vektor. Pendekatan tree distance (Praven dkk, 2007) merupakan contoh pendekatan dengan model vektor space.

***Model Probabilistic* :** Ide dasar dari model probabilistic adalah bahwa jika diketahui beberapa dokumen relevan ke sebuah query, maka bobot yang lebih tinggi diberikan ke isitilah-isitilah yang mana muncul dalam dokumen-dokumen tersebut untuk mencari dokumen-dokumen lain yang relevan. Dalam model probabilistik, probabilitas munculnya setiap istilah dilatih dengan sebuah himpunan dokumen, himpunan query dan himpunan penentuan similaritas diantara tiap dokumen dan tiap query. Teorema Bayes sering digunakan dalam model ini untuk memberitahu bagaimana memperbaharui atau merevisi kepercayaan berkaitan dengan query yang baru dan dokumen baru. Dokumen yang relevan ke query yang baru dapat diperoleh didasarkan pada probabilitas kemunculan istilah query dalam himpunan dokumen training.

BAB III
PERANCANGAN SISTEM

3.1 Spesifikasi Kebutuhan Perangkat Lunak Aplikasi.

3.1.1 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 pengindekan kata, sehingga diketahui hasil akhir term yang muncul disetiap dokumen dan jumlah setiap term yang muncul.
- Sistem lain dapat menggunakan fungsi dan prosedur yang digunakan untuk melakukan pengindekan kata.
- Aplikasi ini dapat berjalan pada server yang terhubung ke internet ataupun hanya terhubung lokal intranet.

Dalam pengembangan aplikasi ini diharapkan dapat memberikan manfaat sbb :

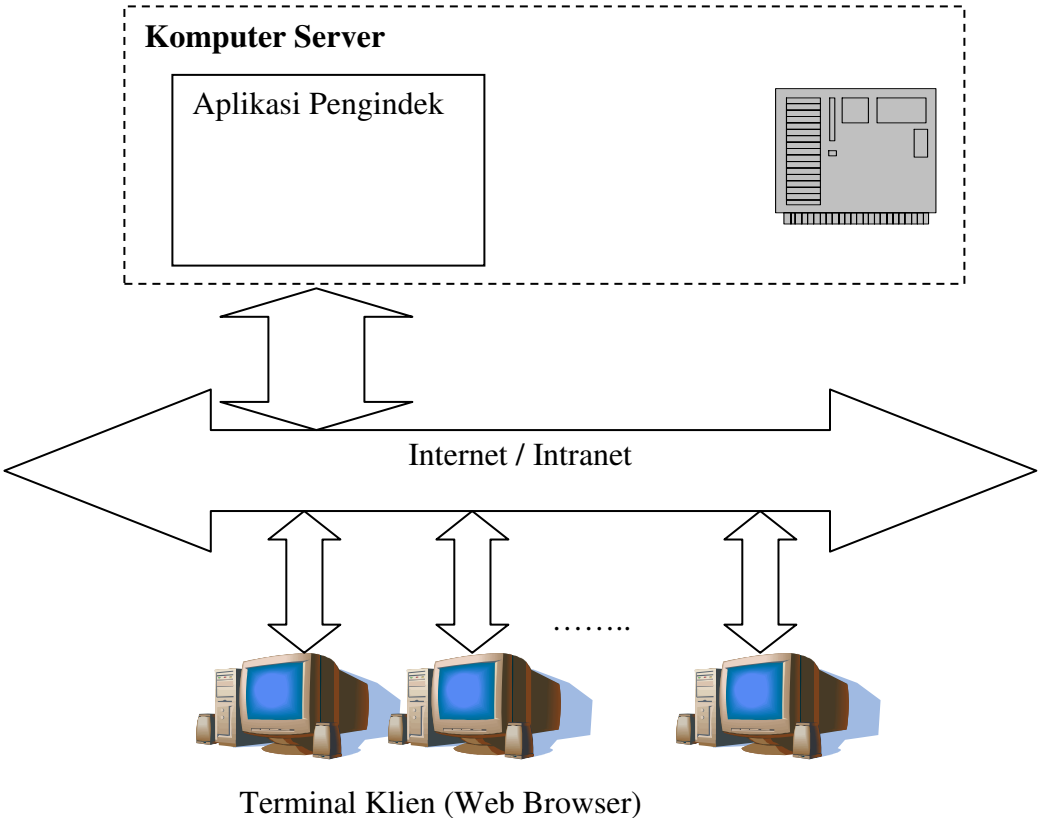
- Mempermudah pengguna untuk melakukan proses indek kata pada dokumen.
- Mempermudah sistem lain untuk pengindekan kata.

3.1.2 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 3.1 Menggambarkan perspektif produk aplikasi yang akan dibangun.



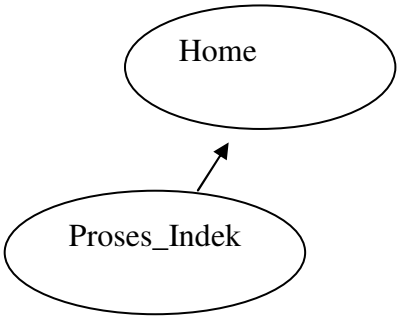
Gambar 3.1 Perspektif produk

Pada gambar 3.1 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.

3.1.3 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 3.2 merupakan hirarki fungsi dari produk aplikasi



Gambar 3.2 Fungsi-fungsi produk

3.1.4 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.

Proses_Indek : Fungsi ini digunakan untuk melakukan proses pengindekan.

3.2 Aturan Bisnis Aplikasi

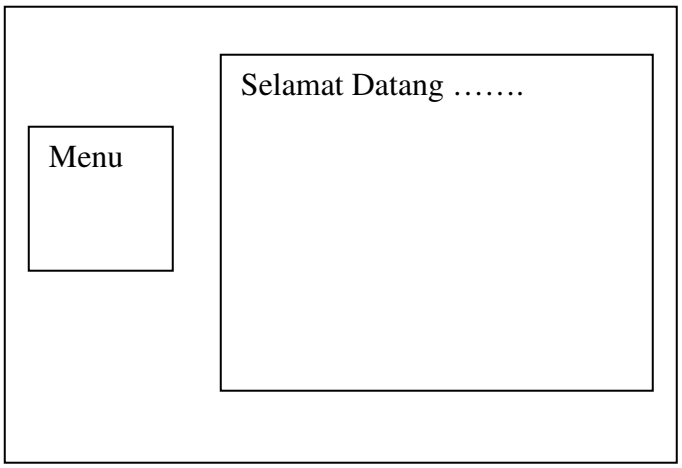
Aturan bisnis digunakan sebagai acuan kemampuan dari aplikasi yang akan dibuat. aturan bisnis untuk pengindek kata adalah sbb:

- Aplikasi berbasis web digunakan untuk melakukan proses indek kata, yang terdapat pada setiap dokumen.

3.3 Tampilan Layar Aplikasi

3.3.1 Tampilan layar home / utama

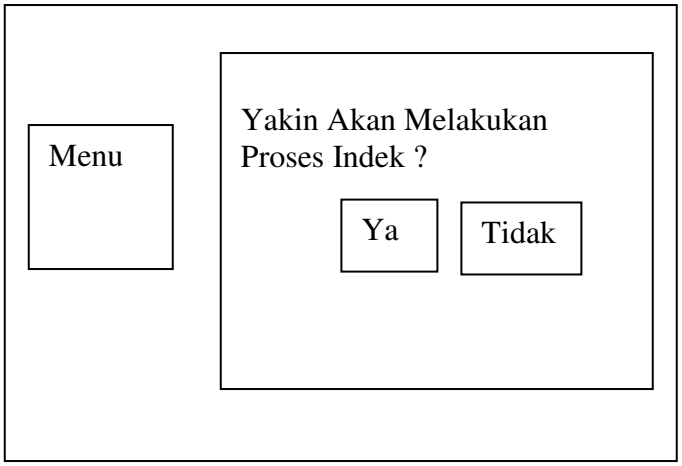
Tampilan layar home / Utama aplikasi diperlihatkan pada gambar 3.3



Gambar 3.3 tampilan layar utama

3.3.2 Tampilan layar pengindek kata

Tampilan layar fungsi pengindek kata diperlihatkan pada gambar 3.4

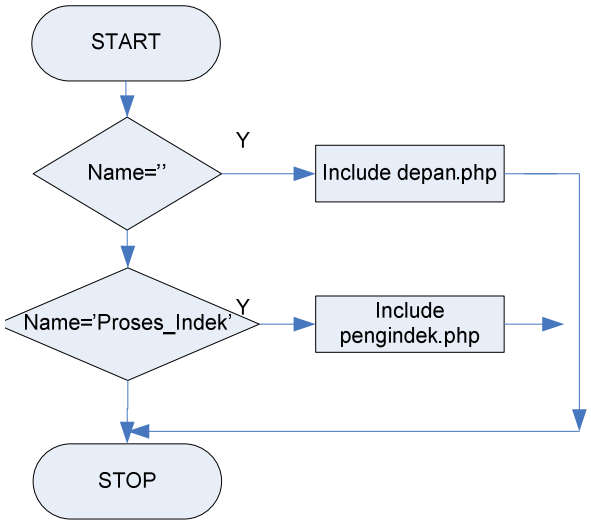


Gambar 3.4 tampilan layar fungsi Indek

3.4. Diagram Alir Aplikasi

3.4.1 Diagram alir menu utama

Aliran Proses Menu Utama aplikasi diperlihatkan pada gambar 3.5. Variabel name adalah parameter yang diberi nilai melalui hyperlink Contoh : http://localhost/modules.php?name=Proses_Indek.

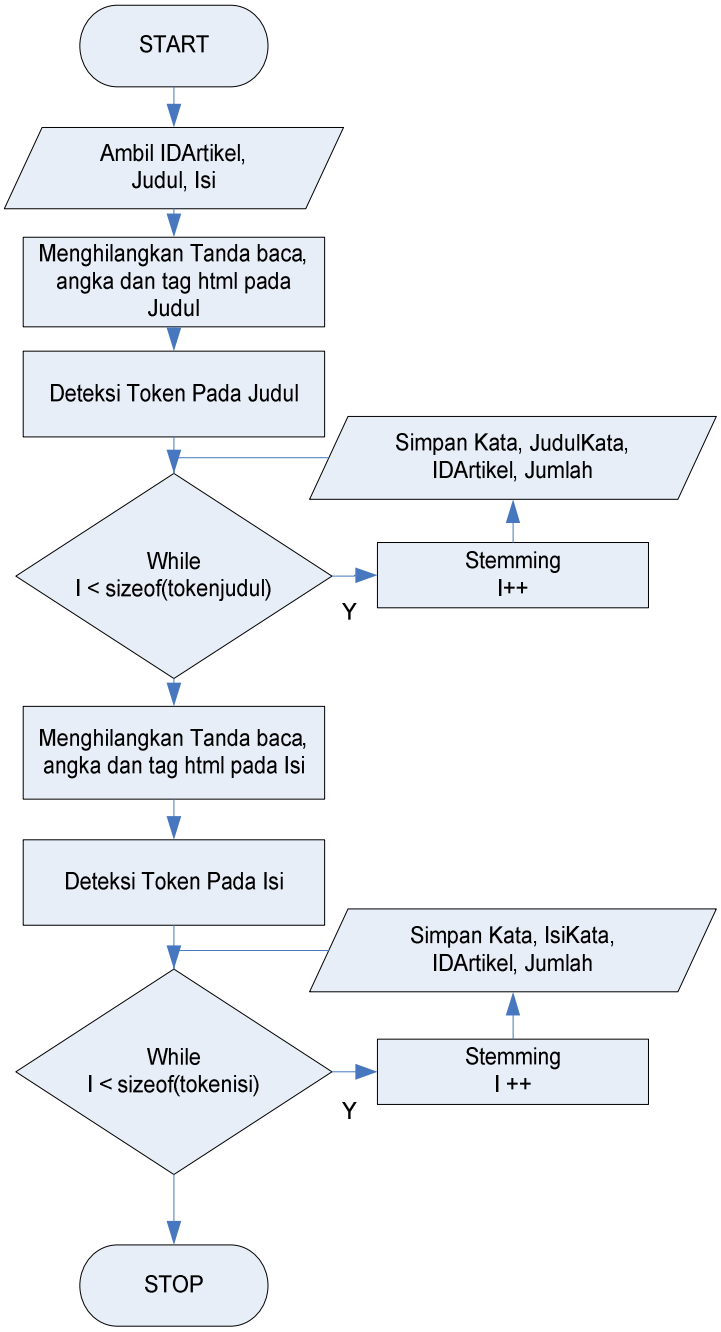


Gambar 3.5 Diagram aliran proses menu utama aplikasi

Pada gambar 3.5 diperlihatkan pada saat variabel *name* tidak terdefinisi atau kosong maka modul halaman depan akan dimuat. Sedang apabila variabel *name* berisi Proses_Indek maka modul halaman pengindek akan dimuat di web.

3.4.2 Diagram alir proses pengindek kata

Aliran Proses Fungsi Pengindek diperlihatkan pada gambar 3.6.



Gambar 3.6 Diagram aliran proses indek

BAB IV
IMPLEMENTASI SISTEM

4.1 Implementasi Modul Tambahan CMS PHPNuke

Pada penelitian ini implementasi sistem menggunakan CMS PHPNuke sebagai manajemen kontennya. Penggunaan CMS diharapkan dapat mempercepat proses implementasi. Fungsi-fungsi umum pada manajemen konten sudah tersedia, tinggal menambah beberapa fungsi utama yang tidak tersedia oleh CMS.

Implementasi sistem manajemen konten PHPNuke pada jaringan internet dapat menggunakan script fantastico yang tersedia oleh layanan web hosting. Instalasi dapat juga mandiri pada komputer lokal baik terhubung ke jaringan atau tidak. Instalasi pada komputer lokal terlebih dahulu aplikasi web server dan RDBMS mysql diinstall terlebih dahulu. Aplikasi layanan web dan RDBMS dapat digunakan paket aplikasi seperti phptriad, appserve, wampserver ata xammpserver. Sedang kode sumber CMS PHPNuke dan cara instalasinya dapat di download dari http://www.phpnuke.org.

Modul tambahan untuk Pengindekan kata disalin pada Direktori "./modules" diikuti dengan Direktori sesuai dengan nama modul, nama modul tidak diperkenankan mengandung spasi atau tanda baca lainnya.

Pada penelitian ini digunakan modul tambahan yaitu modul Indek_Berita sehingga diperoleh direktori tambahan yaitu Direktori "./modules/indek Berita". Kemudian program untuk modul tambahan terletak dalam Direktori tersebut dengan nama file pada masing direktori adalah index.php.

Sebelum modul-modul tambaahan tersebut dapat digunakan, modul-modul tersebut harus diaktifkan terlebih dahulu. Untuk mengaktifkan digunakan login setingkat admin untuk dapat mengakses menu manajemen modul. Setelah modul diaktifkan maka menu dari modul-modul baru dapat terlihat dan dapat digunakan oleh pengguna.

PHPNuke mensyaratkan file index.php pada modul-modul tambahan harus diawali dan dan diakhiri oleh beberapa baris program agar dapat berjalan dengan baik

dengan PHPNuke. Pola file index.php diperlihatkan pada Gambar 4.1, blok program untuk modul tambahan diletakan setelah perintah OpenTable(); dan pada akhir blok program disertakan perintah CloseTable(); dan diikuti dengan pemanggilan file footer.php dengan perintah include.

```
if (!eregi("modules.php", $_SERVER['PHP_SELF'])) {
    die ("You can't access this file directly...");
}
$module_name = basename(dirname(__FILE__));
include("header.php");
OpenTable();
   BLOK PROGRAM MODUL TAMBAHAN
CloseTable();
 include("footer.php");
```

Gambar 4.1 Pola file index.php pada modul tambahan PHPNuke

4.2 Implementasi Layanan Basis Data

Pada penelitian ini digunakan RDBMS Mysql sebagai penyedia layanan basis data. Penggunaan basis data diharapkan dapat mempercepat proses pengambilan informasi dan proses manajemen indek. Selain itu pemilihan layanan basis data menggunakan MySql dengan pertimbangan penggunaan CMS PHPNuke sebagai manajemen kontennya. Ketersediaan layanan basis data MySql juga banyak dijumpai pada penyedia layanan jasa web hosting sehingga dapat lebih leluasa memilih penyedia layanan web hosting sesuai yang dikendaki.

Layanan basis data Mysql biasanya langsung tersedia ketika berlangganan web hosting yang menyediakan layanan basis data Mysql. Pada saat instalasi PHPNuke dengan script fantastico, database untuk keperluan PHPNuke juga otomatis dibuat oleh script fantastico. Manajemen database Mysql biasanya digunakan aplikasi berbasis web PhpMyadmin yang juga tersedia oleh server.

4.3 Implementasi algoritma text preprocessing

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 :

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.

Proses text preprocessing dilakukan pada semua data yang ada, untuk data yang besar dibutuhkan waktu yang lama juga. Pada aplikasi berbasis web masalah akan timbul jika waktu eksekusi lebih dari 30 detik, karena sebagian besar web server membatasi waktu eksekusi permintaan layanan maksimal 30 detik. Walaupun waktu eksekusi dapat diperlama tetapi capaian proses tidak dapat diinformasikan secara cepat dan akurat oleh sistem, karena biasanya server baru akan memberikan informasi setelah proses 100% selesai. Selain itu memperbesar waktu eksekusi akan menggangu kompaktibiltas dengan web hosting yang ada.

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 4.1 diperlihatkan pseudo code proses implementasi mekanisme autoreload text preprocessing pada aplikasi berbasis web, dimana variabel $_SESSION['id'] digunakan untuk menyimpan array id artikel, variabel $_SESSION['dok-POS'] menyimpan pointer posisi dokumen dan variable $_SESSION['dok-MAX'] untuk menyimpan jumlah dari artikel yang diproses.

Algoritma 4.1 *Pseudo code proses implementasi mekanisme autoreload.*

```
session_start();
$_SESSION[dok-pos]:=0;
$_SESSION[id];= array of idartikel;
$_SESSION[dok-max];=sizeof($id);
number_of_record(data)
<iframe autoreload >
        $pos=$_SESSION[pos];
        $max=$_SESSION[max];
        $id=$_SESSION[id];
        while (($pos<$max) and ($pos<$pos+5)) {
                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 5 buah saja, pembatasan ini untuk memotong proses menjadi lebih kecil. Setelah program selesai dijalankan, program akan dipanggil ulang oleh javascript autoreload pada bagian bawah.

4.3.1 Implementasi text filtering.

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 4.2

Algoritma 4.2 *Pseudo code proses pembersihan tanda baca dan angka*

```
$tmp = "";
$str=trim($str);
while (ereg("<(/?[[:alpha:]]*)[[:space:]]*([^>]*)>",$str,$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);
$str=eregi_replace("rsquo"," ",$str);
$str=eregi_replace("nbsp"," ",$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);
$str=str_replace('  ',' ',$str);
$str=str_replace('  ',' ',$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.

Query 4.1 digunakan untuk menghapus stopword dari daftar transaksi kata dan master kata, baik pada transaksi judul maupun transaksi abstrak. Proses ini lebih cepat dibanding melakukan query satu persatu untuk setiap kata setiap kali akan di stem.

Query 4.1 *Query untuk menghilangkan stopword*

```
$r=_query("delete from kata
            where teks in (select kata from stopword)");
$r=_query("delete from judulkata
            where idkata not in (select id from kata)");
$r=_query("delete from abskata
            where idkata not in (select id from kata)");
```

4.3.2 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.

4.4 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, seperti yang dijelaskan pada bagian 4.3.1.

Proses selanjutnya adalah pengindekan kata-kata dasar tersebut. Tujuan akhir dari proses indek ini adalah proximity matrik. Pada proximity matrik diketahui jarak antar dokumen berdasarkan dari jumlah kata yang berpotongan dihitung dengan rumus cosine coeficient.

Tabel artikel berelasi dengan tabel master kata menghasilkan tabel transaksi judulkata. Berikut ilustrasi tabel master artikel pada tabel 4.1 yang berisi IDaritkel sebagai key dan judul yang berisi string dari judul artikel. Setelah melalui proses preprocessing maka akan dihasilkan tabel 4.2 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 4.3. Pada tabel 4.3 pada kolom pertama diperlihatkan bahwa Idartikel 1 mempunyai kata dengan.id 1 sebanyak 1 buah, demikian seterusnya.

Tabel 4.1 Tabel master artikel.

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

Tabel 4.2 Tabel master kata

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

Tabel 4.3 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

BAB V
KESIMPULAN DAN SARAN

5.1. Kesimpulan

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

- 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.

5.2 Saran

Berdasarkan hasil penelitian yang diperoleh dapat disarankan beberapa hal sbb:

- 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.
- Visualisasi indek dalam bentuk grafik ataupun SOM (Self Organizing Map) dapat lebih mempermudah untuk mencari dokumen yang mirip.

DAFTAR PUSTAKA

Baesa R; Ribeiro B, 1998, *Modern Information Retrieval*, ACM Press New York USA

JISC Briefing Paper, 2006, *Text mining*, JISC, Inggris

Liu Y; Hui C; Hang M dan Ma S, 2004 *Finding abstract field of web pages or query specific retrieval*, Text Retrieval Conference. http://trec.nist.gov (diakses tanggal 24 Maret 2009)

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

Nugroho, Bunafit, 2004, *PHP & MySQL Dengan Editor Dreamweaver MX,* Andi, Yogyakarta

Pressman R, 1997, *Software Engineering,* Mc Graw Hill, USA

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

Wijaya S; Nugroho B; Khoerniawan T dan Mirna A, 2007, *Analisis struktur dokumen pada perolehan informasi dokumen web*, Faculty of computer science University of Indonesia, Indonesia