Design And Implementation of Document Similarity Search System For WEB-Based Medical Journal Management

Mardi Siswo Utomo dan Edi Winarko

Abstract— Document similarity can be used as a reference for other information searches similar. So as to reduce the time-re-appointment for information following a similar document. Document similarity search capability is usually implemented on the features 'related articles'.

Similarity of documents can be measured with a cosine, with preprosesing conducted prior to the document that will be measured. The indexing process and the measurement takes a relatively long excecution time. Problems with a web-based application to conduct the process and measuring the similarity index is a limited execution time, so the processing index and similarity measure in web-based application needs its own programming techniques.

Problems with a web-based application to conduct the process and measuring the similarity index is a limited execution time, so the processing index and similarity measure in web-based application needs its own programming techniques.

The purpose of this research is to design and create a software that give capability for web-based database management system of medical journals in Indonesian language to find other documents similar to the current document in reading at the time.

The results of this research is the mechanism autoreload javascript and session cookies and can break down the process and measurement index similaritas into several small sections, so the process can be performed on web-based applications and the number of relatively large documents.

Results with the cosine similarity measure in the case of Indonesian-language medical journal "Media medika Indonesiana" has a fairly high accuracy of 90%.

Keywords— *document similarity, cosine measure, web-based application.*

M. Siswo Utomo, Universitas Stikubank (UNISBANK), Semarang

E. Winarko, Jurusan Ilmu Komputer dan Elektronika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada Sekip Unit III Bulaksumur Yogyakarta 55281, email : ewinarko@ugm.ac.id

I. PENGANTAR

Perkembangan teknologi internet yang pesat membuat semakin banyaknya pilihan informasi yang tersedia. Dari sekian banyak informasi yang tersedia hanya sedikit informasi yang sesuai dengan keinginan pengguna, selebihnya adalah informasi yang tidak bermanfaat bagi pengguna. Sistem pencarian dan penelusuran informasi yang akurat menjadi hal penting dalam proses pencarian dokumen karena dapat menghemat waktu temu-kembali informasi.

Kemiripan dokumen (Document Similarity) dapat digunakan menjadi acuan pencarian informasi lainya yang sejenis.

Sehingga dapat mengurangi waktu temu-kembali informasi untuk dokumen berikutnya yang sejenis. Kemampuan pencarian kemiripan dokumen biasanya diimplementasikan pada fitur ”artikel terkait” seperti halnya pada kebanyakan situs-situs berita dan jurnal, tidak terkecuali juga artikel Biomedical. Dimana pengguna diberikan beberapa hyperlink yang berisi dokumen yang mirip dengan yang dibaca saat itu.

Banyak pendekatan yang digunakan untuk menyediakan kemampuan tersebut, diantaranya menggunakan data statistik dokumen lainya yang dibuka oleh pengguna-pengguna lain yang juga membuka dokumen tersebut. Atau dapat pula menggunakan data umpan balik relevansi. Penggunaan data umpan balik merupakan cara yang banyak digunakan dan terbukti sukses dalam penggalian informasi.

Pengukuran kemiripan tidak hanya dilakukan pada isi dokumen saja tetapi pada parameter-parameter lain seperti judul kata kunci dan kata yang tercetak tebal. Pada penelitian oleh Syntia Wijaya [14], menyimpulkan bahwa perolehan informasi dengan mengkombinasikan isi, dokumen dan anchor memberikan hasil yang paling baik. Selain itu, penggunaan umpan balik dan perluasan query juga memperbaiki hasil perolehan dokumen pada struktur dokumen web.

Poltak Sihombing [11], menggunakan algoritma genetika untuk implementasi rumus Jaccard dan Dice pada pengukuran kemiripan dokumen sistem temu kembali informasi. Tujuan dari algoritma genetika adalah untuk mencari dokumen yang paling tepat *(best fit)* untuk para pencari. Skoring menggunakan Horng & Yeh's.

Faktor lain yang menunjukan perbaikan perolehan dokumen adalah pembobotan. Pada penelitian yang dilakukan oleh Yi quin [7], perolehan dokumen menjadi lebih baik ketika dilakukan pembobotan terhadap field-field tertentu seperti judul, kata-kata yang ditebalkan dan sebagainya.

Faktor umpan balik juga memiliki peran dalam tingkat relevansi dokumen. Hal ini ditunjukan dalam penelitian yang dilakukan oleh Ryan White [13], bahwa penggunaan umpan balik implisit untuk query expansion dapat menaikan tingkat relevansi dokumen. Dalam peneilitiannya, Ryan White [13], menggunakan dua cara untuk mendapatkan umpan balik yaitu umpan balik ekplisit dan umpan balik implisit.

Penelitian tentang penambangan teks pada artikel biomedical dilakukan oleh [8]. Dalam penelitian ini disebutkan bahwa penambangan teks pada teks pada artikel biomedical dilakukan dengan 5 tahapan proses, yaitu : 1) Text Gathering, adalah proses koleksi dokumen. Proses koleksi dapat melalui penyedia dokumen di internet seperti Pubmed, Citeseer dan sebagainya, dapat juga melalui mesin pecari semacam Google, Yahoo ataupun Bing. Selain itu format dari dokumen juga diperhatikan, disesuaikan dengan kemampuan aplikasi untuk melakukan ekstraksi teks pada dokumen tersebut. 2) Text Preprocesing, proses ini secara umum terdiri dari beberapa tahap yaitu *tokenizing, filtering, stemming, tagging*. Hasil dari proses ini diharapkan memudahkan proses analisa pada tahap selanjutnya. 3) Data Analysis, setelah data melalui tahap preprocessing, akan dilakukan proses analisa pada hasil preprocesing. Data analisis dapat dilakukan dengan 3 metode yang biasa digunakan yaitu a) *Boolean Model*, b) *Vector Model* dan c) *Probabilistic Model*. 4) Visualisation, proses visualisasi merupakan proses yang paling berperan untuk menyampaikan hasil kepada pengguna. Mengekstrak informasi yang tidak dilihat oleh satu orangpun membuat sistem menjadi tidak berguna Visualisasi yang komunikatif mempermudah dan mempercepat pengguna untuk mengerti informasi yang disampaikan. 5) Evaluation, proses evaluasi hasil dimaksudkan untuk mengukur performa dari sistem. Banyak metode yang dapat digunakan untuk mengukur performa dari sistem. Untuk data berkelompok dapat digunakan Precission and Recall, Kluster dapat menggunakan F-Measure dan sebagainya. Kadangkala juga dapat digunakan pendapat pengguna untuk melakukan evaluasi sistem.

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 basis data 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 [1].

II. CARA PENELITIAN

Sistem kemiripan dokumen diharapkan dapat mempermudah pengguna menemukan dokumen sejenis dengan dokumen yang diakses saat itu. Untuk itu diperlukan data kemiripan antar dokumen, pada penelitian ini digunakan *cosine measure* untuk mengukur kemiripan antar dokumen. Data kemiripan disimpan pada tabel basis data untuk mempermudah proses query.

Sistem pencari kemiripan dokumen dengan *cosine measure* yang dibuat diharapkan memiliki kemampuan :

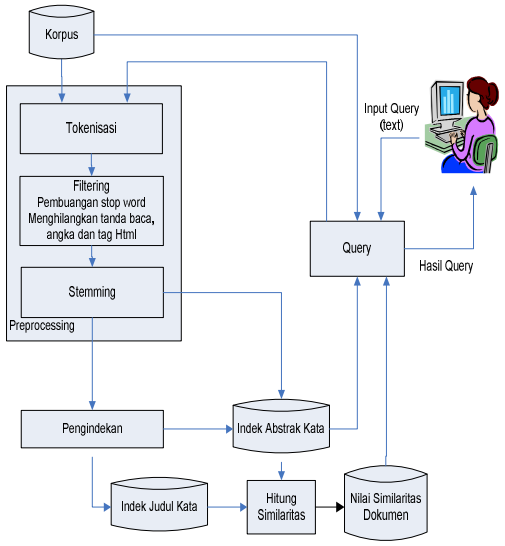
1. Sistem ini harus mampu membuat basis data indexing dalam bentuk tabel indek yang akan digunakan dalam proses query .
2. Sistem harus memiliki fasilitas query yang dapat merangking dengan memperhatikan frekuensi setiap kata pada setiap dokumennya.
3. Sistem harus mampu membuat filter terhadap daftar kata umum (stop word) dan tidak melakukan pengindekan ke dalam basis data.
4. Sistem harus mampu memberikan daftar dokumen yang mirip dengan dokumen yang sedang dibaca saat itu.
5. Program aplikasi ini dapat diakses pengguna melalui antarmuka web baik dengan internet maupun intranet.

II.I DATA PENELITIAN

Data yang digunakan dalam penelitian ini diambil dari dokumen teks abstrak naskah publikasi Jurnal Kedokteran Media Medika Indonesiana Fakultas Kedokteran Universitas Diponegoro dari tahun 1997 s/d 2008. Dokumen yang diproses adalah dokumen yang memuat teks berbahasa Indonesia.

II.II DIAGRAM ARSITEKTUR

Sistem pencari kemiripan dokumen memiliki beberapa proses yang membangun sistem secara keseluruhan. Proses tersebut terdiri dari : Proses tokenisasi, proses filtering, proses pembuangan stopword, proses stemming dan proses pengindeksan kata. Keempat proses tersebut memproses judul dokumen, abstrak dokumen dan masukan query secara terpisah serta menyimpannya dalam tabel indek yang berbeda. Kemudian proses pengukuran kemiripan dokumen yang menggunakan tabel indek judul kata dan abstrak kata. Secara lengkap arsitektur dari sistem pencari kemiripan dokumen dapat dilihat pada Gambar 1.



Gambar 1 Diagram Arsitektur Sistem Pencari Kemiripan Dokumen

II.II.I PROSES TOKENISASI:

Pada proses tokenisasi kata akan dipisahkan dari dokumennya. Pada umumnya setiap kata terindentifikasi atau terpisahkan dengan kata yang lain oleh karakter spasi, sehingga proses tokenisasi mengandalkan karakter spasi pada dokumen untuk melakukan pemisahan kata. Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan query secara terpisah. Pada pemrograman PHP dapat dilakukan dengan perintah explode yang akan merubahnya menjadi variabel array, sehingga pemrograman dapat dilakukan dengan lebih cepat.

II.II.II PROSES FILTERING:

Proses filtering berfungsi untuk membersihkan teks dokumen dari tanda baca, tag html dan angka. Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan query secara terpisah. Proses pembuangan stop word merupakan bagian dari proses filtering, proses ini lebih mudah dan lebih cepat diproses setelah kata diekstrak dari teks dokumennya. Pembuangan stopword sendiri dimaksudkan untuk menghilangkan kata-kata yang masuk dalam kategori stop word. Stopword merupakan kata yang tidak memiliki arti atau tidak relevan. Kata yang diperoleh dari tahap filtering diperiksa dengan daftar stopword, apabila sebuah kata masuk di dalam daftar stopword maka kata tersebut tidak akan diproses lebih lanjut. Sebaliknya apabila sebuah kata tidak termasuk di dalam daftar stopword maka kata tersebut akan masuk ke proses selanjutnya. Daftar stopword tersimpan dalam suatu tabel, dalam penelitian ini menggunakan daftar stop word yang digunakan oleh Tala [12], sebanyak 765 kata. Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan query secara terpisah.

II.II.III PROSES *STEMMING* :

Proses *stemming* adalah proses pembentukan kata dasar. Kata yang diperoleh dari tahap pembuangan stopword akan dilakukan proses stemming. Algoritma stemming menggunakan algoritma berbasis aturan [12]. Stemming digunakan untuk mereduksi bentuk kata untuk menghindari ketidak cocokan, di mana kata-kata yang berbeda namun memiliki makna dasar yang sama direduksi menjadi satu bentuk. Proses ini dilakukan pada judul dokumen, abstrak dokumen dan masukan query secara terpisah.

II.II.IV PROSES PENGINDEKAN:

Hasil dari kata-kata yang telah diproses sebelumnya disimpan dengan metode tertentu

sehingga mempermudah proses perangkingan dan query. Penelitian ini menggunakan metode *Inverted Index*, dengan struktur terdiri dari : kata, identitas dokumen dan jumlah kemunculan. Kata-kata tersebut adalah himpunan dari kata-kata yang ada pada dokumen, merupakan ekstraksi dari kumpulan dokumen yang ada. Proses ini dilakukan pada judul dokumen, abstrak dokumen secara terpisah.

II.II.V PROSES HITUNG SIMILARITAS:

Relevansi sebuah dokumen ke sebuah *query* didasarkan pada *similarity* (similaritas) diantara vektor dokumen dan vektor *query*. Koordinat dari bobot istilah secara dasarnya diturunkan dari frekuensi kemunculan dari istilah. Metode yang digunakan untuk menghitung adalah metode *cosine simmilarity* dengan menggunakan rumus seperti diuraikan pada persamaan (1). Masing-masing dokumen akan dihitung kata yang sama antara dokumen yang satu dengan dokumen yang lain. Hasil dari perhitungan akan dihasilkan dokumen dengan nilai similaritas dokumen. Nilai similaritas dokumen yang tertinggi dapat dianggap bahwa dokumen tersebut paling similar, dengan kata lain memiliki banyak kesamaan.

$$Similarity(X,Y) = \frac{X \bigcap Y}{|X|^{\frac{1}{2}} . |Y|^{\frac{1}{2}}} \text{.....(1)}$$

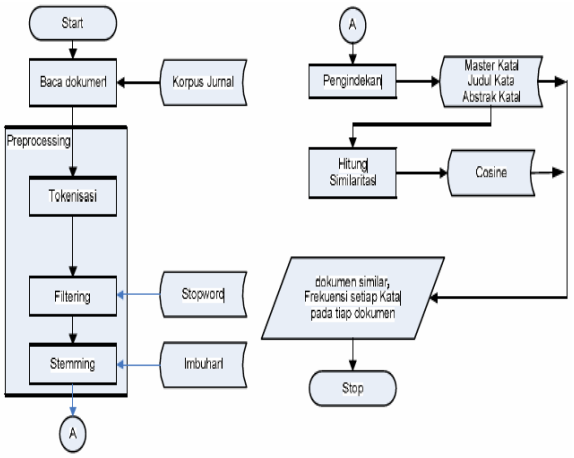
Dimana : $X \bigcap Y$
adalah jumlah terms yang ada di X dan Y
|X| adalah jumlah term yang ada di X
|Y| adalah jumlah term yang ada di Y

II.III DIAGRAM ALIR SISTEM

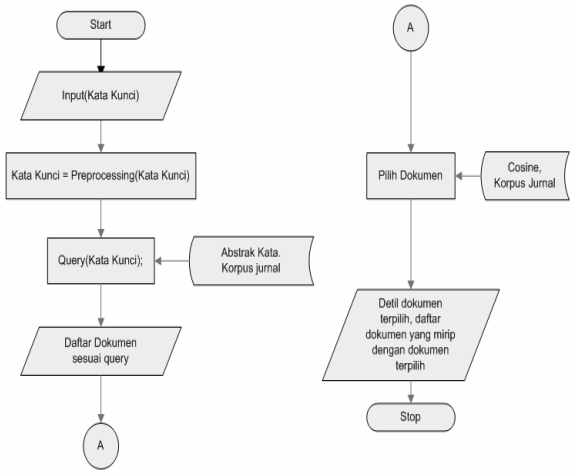
Gambar 2 menjelaskan Diagram Alir sistem untuk basis data indek dan basis data similaritas. Proses untuk pembentukan basis data indek dan basis data similaritas terdiri dari: 1) Proses preprosesing terdiri dari 3 proses : proses tokenisasi, proses filtering, proses pembuangan stopword dan proses *stemming*, 2) proses pengindekan, 3) proses hitung similaritas.

Diagram Alir sistem untuk Sistem Pencari Kemiripan Dokumen dapat dilihat pada Gambar 3. Sistem Pencari Kemiripan Dokumen dimulai dari input berupa *query* (kata kunci). *Query* terlebih dahulu melalui proses preprocessing kemudian akan dilakukan proses koneksi ke basis data indek dan basis data korpus jurnal. Basis data indek merupakan basis data hasil dari diagram alir sistem pembentukan basis data indek dan similaritas. Hasil dari proses koneksi ke basis data indek dan korpus jurnal adalah dokumen dari basis data korpus jurnal yang relevan (sesuai) dengan *query* yang dimasukkan. Apabila dilakukan pemilihan salah satu dokumen maka proses akan menampilkan detil dokumen yang dipilih beserta daftar dokumen yang mirip dengan dokumen terpilih.

Pada Sistem Pencari Kemiripan Dokumen, diagram alir program untuk proses pembentukan basis data indek dan similaritas dokumen yang terdiri dari diagram alir preprocesing, diagram alir tokenisasi, diagram alir filtering, diagram alir pengindekan, diagram alir stemming, diagram alir hitung similaritas dan diagram alir proses query.



Gambar 2. Diagram Alir Pembentukan Basis Data Indek dan Similaritas



Gambar 3. Diagram Alir Sistem Pencari Kemiripan Dokumen

Pada proses query metode perangkingan dokumen hasil query menggunakan pembobotan *Term Frequency* dan jika terdapat lebih dari 1 kata kunci digunakan operator boolean "OR". Rumus perangkingan bobot dokumen terhadap kata kunci dapat dilihat pada rumus (2).

$$Bobot\ (DxTK\ ) = \sum_{i=1}^{t} D_x TK_i \quad \text{.....(2)}$$

Dimana : D_xTK_i adalah jumlah kata kunci ke i yang muncul pada dokumen x
t = Jumlah term dalam query

Selain dicari dokumen-dokumen yang mengadung array kata juga dicari jumlah keseluruhan kata-kata dalam array kata yang muncul di setiap dokumen tersebut. Jumlah setiap kata yang ada di array kata di jumlahkan disetiap dokumennya sehingga dihasilkan jumlah total kata-kata yang ada di array kata pada setiap dokumen. Kemudian hasil daftar dokumen ditampilkan terurut mulai dari yang memiliki total jumlah terbanyak kata-katanya yang muncul di array kata.

Apabila pengguna memilih salah satu dokumen dari daftar dokumen yang ditampilkan, maka proses akan dilanjutkan dengan mengambil dan menampilkan detil dokumen terpilih dari tabel korpus jurnal. Setelah itu daftar dokumen yang mempunyai nilai kemiripan >=20% dengan dokumen terpilih juga ditampilkan. Daftar dokumen yang mirip dengan dokumen terpilih diambil dari tabel cosine.

Pada sistem juga ditambahkan kemampuan untuk menambahkan daftar stopword pada tabel stopword dan antarmuka penambahan imbuhan untuk proses stemming.

III. IMPLEMENTASI SISTEM

Perangkat lunak pada penelitian ini, ditulis dengan menggunakan bahasa pemrograman PHP pada platform PHP 5.0. Pemrograman PHP digunakan untuk implementasi proses-proses dalam Sistem Pencari Kemiripan Dokumen. Basis data yang digunakan untuk menyimpan data adalah MySQL 4.0.

Untuk mempermudah implementasi pada penelitian, keseluruhan fungsi Sistem Pencari Kemiripan Dokumen dibagi menjadi 4 modul tambahan yaitu: 1) Modul Indek jurnal, digunakan untuk membangun basis data indek dan basis data similaritas. 2) Modul Tambah imbuhan, digunakan untuk mempermudah pengguna menambahkan imbuhan yang akan diproses oleh proses stemming ke dalam tabel imbuhan. 3) Modul Tambah stopword, digunakan untuk mempermudah pengguna menambahkan kata dalam daftar stopword yang akan disimpan di dalam tabel imbuhan. 4) Modul Tampilan Jurnal, digunakan untuk menampilkan daftar dokumen hasil query, menampilkan detil dokumen dan daftar dokumen yang mirip.

Modul indek jurnal berfungsi untuk membangun basis data indek *term frequency* dan melakukan penghitungan nilai similaritas antar dokumen. Pada Algoritma 1 diperlihatkan urutan pross Keseluruhan proses pembangunan basis data indek dan similaritas. Pada modul ini terdapat proses preprocessing, proses pengindekan dan penghitungan similaritas.

Algoritma 1 *Keseluruhan Proses Pembangunan Basis Data Indek Dan Similaritas*

```
Untuk Setiap Dokumen {
        Baca(ID,Judul,Abstrak);
        Astr=Preprocesing(Judul);
        Indexing(Astr,ID,Judul);
        Astr=Preprocesing(Abstrak);
        Indexing(Astr,ID,Abstrak);
}
Proses pengukuran
Untuk Setiap Pasangan Dokumen {
        Ukur jarak cosine
(dokumen1,dokumen2);
        Simpan dalam tabel cosine;
}
```

Proses 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 preprocessing dilakukan tiap satu persatu dokumen, dengan demikian hanya 1 dokumen yang akan diproses setiap waktunya oleh web server setiap kali url program di muat / dipanggil. Kemudian menggunakan

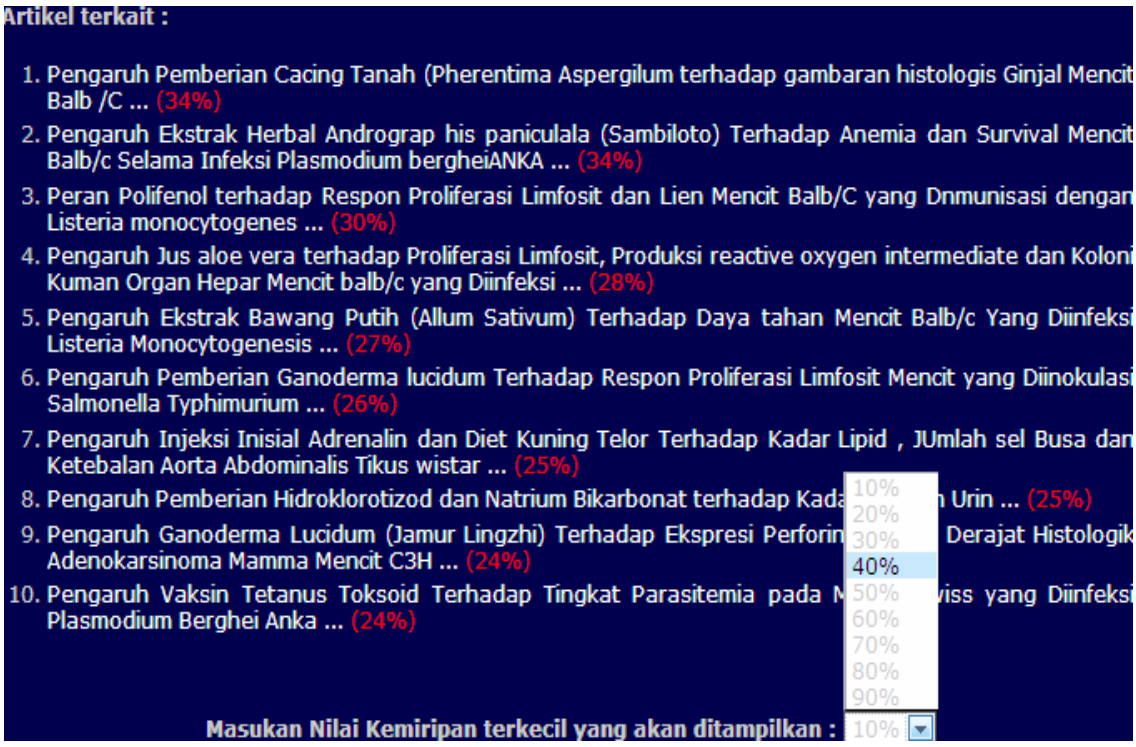
mekanisme variabel session untuk menyimpan data pointer posisi dokumen terakhir diproses. Sehingga setiap kali url dimuat maka pointer akan bergeser ke dokumen selanjutnya sampai pointer menunjuk pada dokumen terakhir.

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 pemrograman PHP proses filtering pembuangan tanda baca, angka dan tag html lebih mudah dilakukan pertama kali sebelum proses tokenisasi, tetapi proses stopword removal tetap dilakukan setelah proses tokenisasi. Pada penelitian ini proses preprocessing untuk teks berbahasa Indonesia menggunakan urutan sbb: a)Proses *Filtering.* b) Proses *Tokenizing* c) Proses *Stopword Removal* d) Proses *Stemming*

III.I VISUALISASI HASIL

Visualisasi similaritas dilakukan saat sebuah jurnal di baca, pada bagian bawah jurnal ditampilkan daftar judul, hyperlink dan nilai kemiripan jurnal lain yang mempunyai nilai kemiripan tertinggi dengan jurnal yang sedang dibaca. Daftar kemiripan jurnal ditampilkan terurut berdasarkan nilai similaritas tertinggi, dengan nilai similaritas minimal 20% atau 0,2. Hanya dokumen dengan nilai similaritas <= 0,2 yang akan ditampilkan dan dengan jumlah maksimal 15 Jurnal. Nilai minimal similaritas dapat diatur oleh pengguna mulai dari 10% s/d 90% dengan kelipatan 10% sesuai dengan kebutuhan pengguna, tetapi dengan jumlah hasil tetap dibatasi yaitu 10 dokumen termirip saja. Antarmuka masukan nilai similaritas oleh pengguna diperlihatkan pada gambar 4.



Gambar 4. Antarmuka Masukan Pilihan Minimal

IV. HASIL DAN PEMBAHASAN

Total Jurnal yang digunakan dalam penelitian ini adalah 314 tetapi hanya 269 dokumen saja yang memenuhi syarat (mempunyai judul dan abstrak berbahasa Indonesia), sehingga didapatkan proximity matrik sebesar 269 x 269 = 72361 elemen, dengan distribusi nilai diperlihatkan pada tabel 1

Tabel 1. Hasil Pengukuran Proses Similaritas

Similaritas	Jumlah sel
100%	269
60% - 70%	4
50% - 60%	6
40% - 50%	32
30% - 40%	92
20% - 30%	522
10% - 20%	5712
1% - 10%	62370
0 %	3354
Total :	72361

IV.I VALIDASI HASIL SIMILARITAS

Pada validasi hasil similaritas digunakan sistem kuis ke 2 orang pakar yang berkompeten dibidang ini. Masing-masing pakar diberikan 5 jurnal acak dengan tingkat kemiripan yang berbeda 20%. Pakar akan memberikan penilaian kemiripan antar dokumen jurnal tersebut.

Tabel 2 Hasil Pengukuran Pakar 1

No	IdJurnal	Urutan Pakar	Urutan Sistem	%
1.	18	244, 82, 243, 117	244, 82, 243, 117	100
2.	244	18, 82, 243, 117	18, 82, 243, 117	100
3.	243	117, 18,244, 82	117, 18,244, 82	100
5.	117	243, 82, 244, 18	243, 82, 244, 18	100
5.	82	18, 244, 117, 243	18, 244, 117, 243	100
			Rata - rata :	100

Tabel 2 diperlihatkan hasil dari pendapat pakar pertama, semua urutan dari sistem sama persis dengan urutan kemiripannya menurut pakar pertama. Sehingga mempunyai score masing –masing 100 %.

Tabel 3 diperlihatkan hasil dari pendapat pakar kedua, terdapat 1 jurnal yang sama urutan kemiripannya menurut pakar kedua, yaitu jurnal dengan id 203. sedangkan 4 jurnal lainya terdapat 1 kesalahan urutan kemiripan sehingga mempunyai score masing –masing 75 %. Rata-rata score dari pakar kedua adalah 80%.

Tabel 3. Hasil Pengukuran Pakar 2

No	IdJurnal	Urutan Pakar	Urutan Sistem	%
1.	17	204, 160, 66, 203	**160**, 204, 66, 203	75
2.	160	204, 17, 66, 203	**17**, 204, 66, 203	75
3.	204	160, 17, 66, 203	**17**, 160 , 66, 203	75
5.	203	204, 160, 66, 17	204, 160, 66, 17	100
5.	66	17, 204,160, 203	**160**, 17, 204, 203	75
			Rata – rata	80

Hasil rata-rata pengukuran perbandingan sistem dengan pakar adalah :

= (100+80) / 2

= 90%

Dari ke 2 hasil diatas dapat disimpulkan bahwa, rangking dari sistem 80% mirip dengan rangking yang dihasilkan oleh para pakar. Sehingga system menghasilkan keputusan yang cukup bagus.

V. KESIMPULAN

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

1. Pembatasan waktu eksekusi pada sistem informasi berbasis web dapat dihindari dengan mekanisme autoreload, membagi pemrosesan dokumen menjadi 2 bagian (Indexing dan Pengukuran Similaritas) dan melakukan proses per-dokumen sehingga meningkatkan jumlah dokumen yang mampu diproses dan terhindar dari terminasi proses oleh server.
2. 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 apabila jumlah dokumen yang diproses berkembang / bertambah setiap waktunya.
3. Penggunaan basisdata untuk menyimpan data indek dapat mempercepat proses pengukuran kemiripan.
4. Hasil pengukuran kemiripan dengan cosine coefficient mempunyai akurasi 90% .
5. Tampilan dan fleksibilitas filter hasil kemiripan memudahkan pengguna untuk

mengkases dokumen yang mirip, serta variasi jumlah hasil dengan mengatur batas minimal kemiripan.

V.I SARAN

Berdasarkan hasil penelitian yang diperoleh dapat disarankan beberapa hal sbb:

1. Penelitian ini menggunakan corpus yang relatif kecil (abstrak), dapat diteliti lebih lanjut pada corpus yang lebih besar lagi misalnya isi jurnal, skripsi, tesis atau desertasi, untuk melihat kualitas hasil pengukuran.
2. Dapat juga diteliti pada corpus yang sama tetapi dengan jumlah dokumen yag lebih banyak 5000> sehingga dapat diukur perfoma dan kemampuan sistem.
3. Penggunaan algoritma stemming bahasa Indonesia berbasis kamus dan aturan dapat meningkatkan kualitas indek dan mungkin juga akan meningkatkan akurasi hasil pengukuran kemiripan.
4. Visualisai hasil dalam bentuk grafik ataupun SOM (Self Organizing Map) dapat lebih mempermudah untuk mencari dokumen yang mirip.

VI. DAFTAR PUSTAKA

[1] Adrifina, A; Putri, J dan Simri, IW, 2008, *Pemilihan Artikel Berita Dengan text Mining*, KOMMIT 2008

[2] Arifin, A.Z. dan Setiono, A.N. (2002), *Classification of Event News Documents in Indonesian Language Using Single Pass Clustering Algorithm, in 'Proceedings of the Seminar on Intelligent Technology and its Applications (SITIA)*, Teknik Elektro, Sepuluh Nopember Institute of Technology, Surabaya, Indonesia.

[3] Baesa, R dan Ribeiro, B, 1998, *Modern Information Retrieval*, ACM Press New York USA

[4] Crasswell, N dan Hawking, D, "*Effective Site Finding Using Link Anchor Information*".

[5] Klose, A; N¨urnberger, A; Krusel, R; Hartmann, G dan Richards, M, 2000, *Interactive Text Retrieval Based on Document Similarities*, Institute for Knowledge and Language Processing, University of Magdeburg, Jerman

[6] Lee, M; Pincombe, B dan Welsh, M, 2005, *An Empirical Evaluation of Models of Text Document Similarity*, University of Adelaide, Australia

[7] 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)

[8] Mathiak, B dan Eckstein, S, 2005, *Five Steps to Text Mining in Biomedical Literature*, Technische Universität Braunschweig, Jerman.

[9] Jurnal Kedokteran FK UNDIP Media Medika Indonesiana, http://www.mediamedika.net, (diakses tanggal 25 Oktober 2009)

[10] PHPNuke Official Site, http: // www.phpnuke.org, (diakses tanggal 25 Oktober 2009)

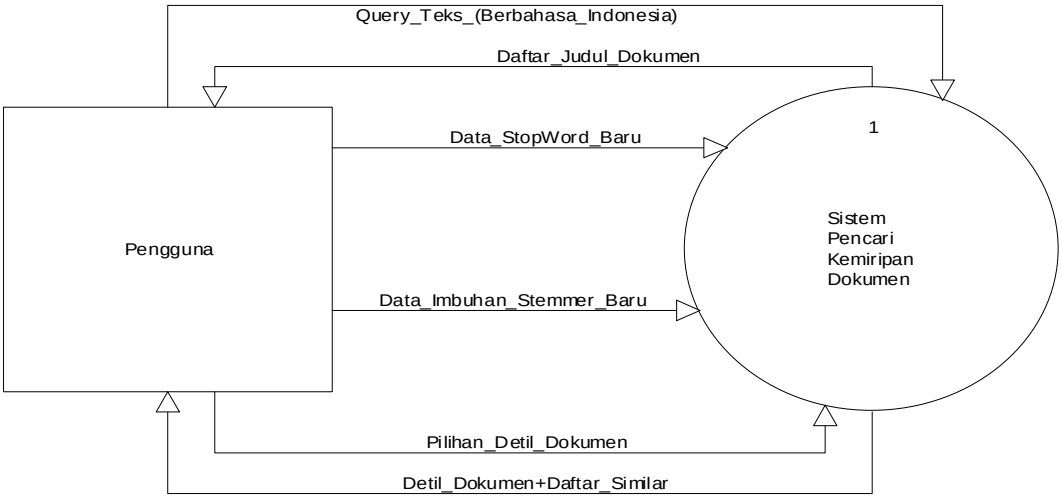
[11] Sihombing, P; Embong, A dan Sumari, P, 2006, *Comparison of Document Similarity in Information Retrieval System by Different Formulation*, Universiti Sains Malaysia, Penang, Malaysia.

[12] Tala, Z, 2003, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands.

[13] White, RW; Ruthven, I; Jose, JM, "*The Use of Implicit Evidence for Relevance Feedback in Web Retrieva*l". http://www.dcs.gla.ac.uk/~whiter/white.pdf (diakses tanggal 25 Oktober 2009)

[14] 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

Lampiran 1 : Diagram Kontek Sistem Pencari Kemiripan Dokumen



Lampiran 2 : Diagram Alir Data Level 1 Sistem Pencari Kemiripan Dokumen