Rancang Bangun *Information Retrieval System (IRS)* Bahasa Jawa Ngoko pada Palintangan Penjebar Semangad dengan Metode *Vector Space Model (VSM)*

Fatkhul Amin dan Purwatiningtyas
Fakultas Teknologi Informasi, Universitas Stikubank Semarang
Email: fatkhulamin@gmail.com, diba_ian@yahoo.com

Abstrak

Bahasa Jawa adalah bahasa daerah yang paling banyak digunakan di Indonesia yang mulai ditinggalkan. Perlunya pelestarian bahasa jawa dalam bentuk online yang bisa diakses bagi penggunanya sehingga akanmemudahkan dalam pencarian dokumen teks khususnya dokumen bahasa jawa ngoko. *Software IRS* dirancang untuk memberikan hasil pencarian dokumen dalam jumlah yang optimal (*recall* rendah) dan akurat (*precision* tinggi) menggunakan metode *VSM*, sehingga *user* akan mendapatkan hasil pencarian cepat dan akurat. Metode VSM akan melakukan pembobotan tiap dokumen yang ada pada database sehingga antar dokumen memiliki bobot yang berbeda untuk menentukan dokumen mana yang paling mirip (similar) dengan *query*, dokumen dengan bobot tertinggi menempati ranking teratas dalam hasil pencarian. Evaluasi hasil pencarian IRS dilakukan dengan uji *recall* dan *precision*. Studi kasus yang telah dilakukan menggunakan IRS ini didapatkan hasil sistem mampu melakukan proses preprosesing (tokenisasi, *filtering, dan stemming*) dengan waktu komputasi 18 detik. Sistem mampu melakukan pencarian dokumen dan menampilkan hasil pencarian dokumen dalam waktu komputasi rata-rata 2 detik, memiliki rata-rata *recall* 0,04 dan rata-rata *precision* 0,84. Sistem dilengkapi dengan bobot tiap dokumen dan letakknya yang akan memudahkan *user* dalam pencarian dokumen teks bahasa Indonesia.

Kata Kunci: Jawa Ngoko, Vector Space Model

PENDAHULUAN

Bahasa Jawa sebagai bahasa yang paling banyak digunakan di wilayah Indonesia setelah bahasa indonesia, dewasa ini mulai banyak ditinggalkan oleh kebanyakan orang. Media offline dan media online juga kurang mengangkat bahasa jawa sehingga dikhawatirkan bahasa jawa lama-kelamaan akan ditinggalkan oleh bangsa kita. Beberapa media online berbahasa Jawa ada, namun belum menggunakan atau belum menyediakan pencarian informasi menggunakan mesin pencari khusus berbahasa jawa.

Implementasi *Vector Space Model* dapat dirasakan dan dinikmati pada berbagai bidang keilmuan seperti *Computational Linguistics* (Erk dkk, 2010), *Expert Systems* (Kim dkk, 2010), *Medical* (lopez dkk, 2010), *Knowledge-Based Systems* (Yu dkk, 2009), *Data and Knowledge Engineering* (Mao dkk, 2007), dan lain sebagainya. *Vector space model* dapat juga digunakan dalam sistem temu kembali informasi *(information retrieval).* Sistem temu kembali informasi akan memberikan nilai tambah dalam pecarian informasi jika keinginan *user* bisa terpenuhi. Penelitian ini diharapakan dapat membuat sistem temu kembali informasi yang bernilai tambah yaitu menghasilkan pencarian informasi dengan cepat dan akurat.

Pencarian informasi saat ini dilakukan dengan menggunakan mesin pencari atau sistem temu kembali informasi, *user* menuliskan *query* dan mesin pencari akan menampilkan hasil pencarian. Mesin pencari yang sudah ada dan banyak digunakan saat ini memberikan hasil perolehan pencarian yang banyak (banyak dokumen yang terambil), sehingga diperlukan waktu untuk menentukan hasil pencarian yang

relevan. Menentukan hasil yang relevan sesuai dengan keinginan user dengan jumlah hasil pencarian yang banyak akan menyulitkan *user*. Hal ini terjadi karena dokumen yang terambil oleh sistem jumlahnya banyak, maka sistem berkemungkinan menampilkan hasil pencarian yang tidak relevan. Banyaknya dokumen hasil pencarian ini membuat waktu yang dibutuhkan dalam pencarian menjadi lebih banyak dari yang diharapkan.

Perkembangan penelusuran informasi saat ini menghasilkan *recall* yang tinggi dan *precision* yang rendah. *Recall* yang tinggi diartikan bahwa dokumen yang dihasilkan dalam penelusuran dokumen adalah banyak, sedangkan *precision* rendah dapat diartikan bahwa dokumen yang diharapkan dapat ditemukan sedikit.

Solusi untuk mengatasi masalah ini adalah dengan membuat *software Information Retrieval System (IRS)* menggunakan metode *Vector Space Model (VSM).* Metode VSM dipilih karena cara kerja model ini efisien, mudah dalam representasi dan dapat diimplementasikan pada *document-matching*. *Software IRS basa jawa ngoko* diharapkan menghasilkan *recall* rendah dan *precision* tinggi.

TUJUAN PENELITIAN

Tujuan yang ingin dicapai dalam penelitian ini adalah;

1. Melestarikan bahasa Jawa agar tidak dilupakan oleh generasi penerus bangsa khususnya orang jawa.
2. Memberikan sumbangsih pemikiran tentang implementasi bahasa jawa di era teknologi informasi
3. Mengembangankan ide kreatif tentang perlunya mesin pencari bebahasa jawa yangbisa digunakan untuk pencarian bahasa jawa
4. Riset Seni Teater Membuat rancang bangun *Information Retrieval System (IRS)* Bahasa *Jawa Ngoko* dengan metode *Vector Space Model*.

METODE

Information Retrieval System dengan *Vector Space Model*

a. *Information Retrieval System (IRS)*

Information Retrieval System menemukan informasi yang biasanya dalam bentuk dokumen dari sebuah data yang tidak terstruktur dalam bentuk teks untuk memenuhi kebutuhan informasi dari koleksi data yang sangat besar umumnya tersimpan dalam *database computer (Manning, 2008).*

information retrieval (IRS) merupakan suatu sistem yang menemukan informasi yang sesuai dengan kebutuhan *user* dari kumpulan informasi secara otomatis. Aplikasi Information Retrieval System sudah digunakan dalam banyak bidang seperti dikedokteran, perusahaan dan lain sebagainya. Salah satu aplikasi dari Information Retrieval System adalah mesin pencari yang dapat diterapkan diberbagai bidang. Pada mesin pencari dengan Information Retrieval System user dapat memasukkan *query* yang bebas dalam arti kata *query* yang sesuai dengan bahasa manusia dan sistem dapat menemukan dokuen yang sesuai dengan *query* yang ditulis oleh *user.*

Prinsip kerja Information Retrieval System jika ada sebuah kumpulan dokumen dan seorang user yang memformulasikan sebuah pertanyaan (*request* atau *query*). Jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan (Salton, 1989).

Information Retrieval System akan mengambil salah satu dari kemungkinan tersebut. Information Retrieval System dibagi dalam dua komponen utama yaitu sistem pengindeksan *(indexing)* menghasilkan basis data sistem dan temu kembali merupakan gabungan dari *user interface* dan *look-up-table*. Information Retrieval System didesain untuk menemukan dokumen atau informasi yang diperlukan oleh *user*.

Information Retrieval System bertujuan untuk menjawab kebutuhan informasi *user* dengan sumber informasi yang tersedia dalam kondisi seperti sebagai berikut (Salton, 1989);

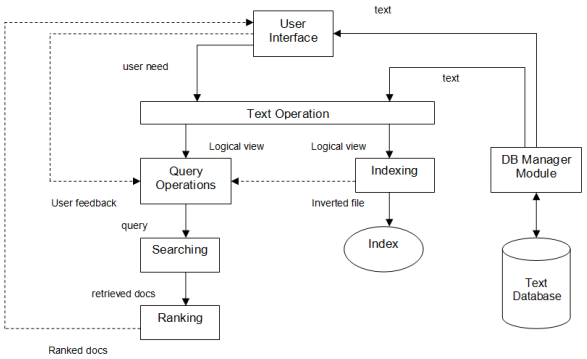
1) Mempresentasikan sekumpulan ide dalam sebuah dokumen menggunakan sekumpulan konsep.
2) Terdapat beberapa pengguna yang memerlukan ide, tapi tidak dapat mengidentifikasikan dan menemukannya dengan baik.
3) Information Retrieval System bertujuan untuk mempertemukan ide yang dikemukakan oleh penulis dalam dokumen dengan kebutuhan informasi pengguna yang dinyatakan dalam bentuk *key word query*/istilah penelusuran.

Fungsi utama Information Retrieval System (Salton, 1989)

1) Mengidentifikasi sumber informasi yang relevan dengan minat masyarakat pengguna yang ditargetkan
2) Menganalisis isi sumber informasi (dokumen)
3) Merepresentasikan isi sumber informasi dengan cara tertentu yang memungkinkan untuk dipertemukan dengan pertanyaan pengguna
4) Merepresentasikan pertanyaan (*query*) *user* dengan cara tertentu yang memungkinkan untuk dipertemukan sumber informasi yang terdapat dalam basis data.
5) Mempertemukan pernyataan pencarian dengan data yang tersimpan dalam basis data
6) Menemu-kembalikan informasi yang relevan
7) Menyempurnakan unjuk kerja sistem berdasarkan umpan balik yang diberikan oleh user.

b. Arsitektur *Information Retrieval System*

Proses *Information Retrieval System* seperti pada gambar 1 menggunakan arsitektur yang sederhana. Sebelum dilakukannya proses temu kembali diperlukan pendefinisian database. Selanjutnya mengikuti tahapan proses; Dokumen-dokumen yang akan digunakan, Operasi yang akan digunakan dalam pencarian, dan model pengolahan teks (Baeza, 1999, h.9).



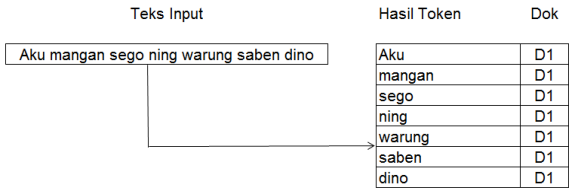
Gambar 1. *The Process of Retrieving Information* (*Baeza*, 1999,h.10)

c. Korpus

Proses IRS dalam aplikasinya membutuhkan database yang didalamnya terdapat satu atau beberapa tabel yang digunakan sebagai tempat penyimpanan data yang akan diolah pada saat proses pencarian. Penelitian dengan menggunakan database pada aplikasinya biasanya memakai korpus untuk proses pembuatan tabel pendukungnya. Penelitian empiris dapat dilakukan dengan menggunakan teks tertulis atau lisan, seperti teks-teks dasar dari berbagai jenis sastra dan analisis linguistik. Tapi gagasan tentang korpus sebagai dasar untuk sebuah bentuk linguistic empiris berbeda dalam beberapa cara mendasar dari teks-teks tertentu.

d. Proses Tokenisasi

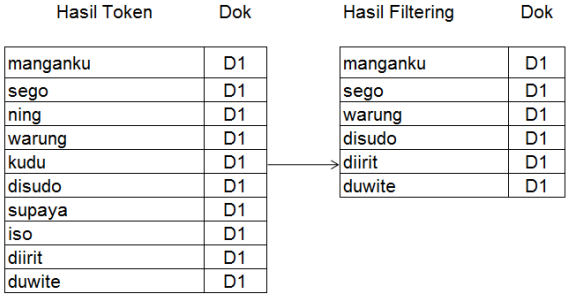
Proses pertama yang dilakukan dalam IRS adalah proses memisahkan kata yang ada pada dokumen berdasarkan spasi kemudian memproses kata yang telah dipisahkan tersebutk kedalam sebuah tabel untuk dilakukan proses berikutnya. Proses Tokenisasi merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi, dan mungkin pada waktu yang bersamaan dilakukan juga proses penghapusan karakter tertentu, seperti tanda baca. Gambar 2 menunjukkan proses tokenisasi.



Gambar 2. Contoh hasil proses tokenisasi

e. Proses Filtering

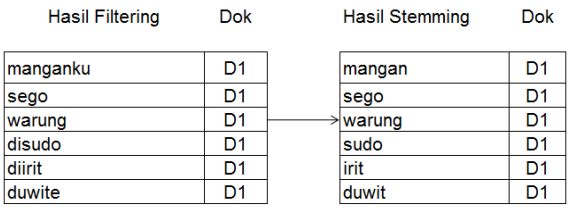
Proses selanjutnya setelah dilakukan pemisahan kata pada dokumen adalah proses *filtering*. *Filtering* akan memproses kata hasil tokenisasi menjadi lebih sedikit dengan cara mengurangi kata tersebut dengan kata yang termasuk dalam *stopwords*. Eliminasi *stopwords* memiliki banyak keuntungan, yaitu akan mengurangi *space* pada tabel *term index* hingga 40% atau lebih (Baeza, 1999, h.167).



Gambar 3. Contoh hasil proses Filtering

f. Proses Stemming

Proses *Stemming* digunakan untuk mengubah *term* yang masih melekat dalam term tersebut awalan, sisipan, dan akhiran. Selanjutnya *term* tersebut diproses untuk dihilangkan awalan, sisipan dan akhiran sehingga menjadi term kata dasar. Proses membuat *term* dasar ini mengacu kepada bahasa jawa ngoko yang benar. Contoh Stemming bisa dilihat pada gambar 4.



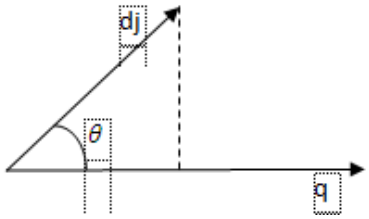
Gambar 4. Contoh hasil proses Stemming

g. Proses Perhitungan VSM

Vector Space Model (VSM) adalah metode untuk melihat tingkat kedekatan atau kesamaan *(similarity)* term dengan cara pembobotan *term*. Dokumen dipandang sebagi sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada *Vector Space Model*, sebuah istilah direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke sebuah *query* didasarkan pada similaritas diantara vektor dokumen dan vektor *query* (Baeza, 1999).

VSM memberikan sebuah kerangka pencocokan parsial adalah mungkin. Hal ini dicapai dengan menetapkan bobot non-biner untuk istilah indeks dalam query dan dokumen. Bobot istilah yang akhirnya digunakan untuk menghitung tingkat kesamaan antara setiap dokumen yang tersimpan dalam sistem dan permintaan user. Dokumen yang terambil disortir dalam urutan yang memiliki kemiripan, model vektor memperhitungkan pertimbangan dokumen yang relevan dengan permintaan user. Hasilnya adalah himpunan dokumen yang terambil jauh lebih akurat (dalam arti sesuai dengan informasi yang dibutuhkan oleh *user*).

Sebuah dokumen dj dan sebuah *query* q direpresentasikan sebagai vektor t-dimensi seperti pada gambar 5.



Gambar 5. The Cosines of θ is adopted as sim d_j, q (Baeza, 1999)

Dalam VSM koleksi dokumen direpresentasikan sebagai sebuah matrik *term* document (atau matrik *term frequency*). Setiap sel dalam matrik bersesuaian dengan bobot yang diberikan dari suatu *term* dalam dokumen yang ditentukan. Nilai nol berarti bahwa term tersebut tidak ada dalam dokumen. Gambar 6

menunjukkan matrik *term document* dengan n dokumen dan t *term*.

	T_1	T_2	T_3	$T_{...}$	T_t
D_1	W_{11}	W_{21}	W_{31}	$\square_{...}$	T_{t1}
D_2	W_{12}	W_{22}	W_{32}	$\square_{...}$	T_{t2}
D_3	W_{13}	W_{23}	W_{33}	$\square_{...}$	T_{t3}
$D_{...}$	$\square_{...}$	$\square_{...}$	$\square_{...}$	$\square_{...}$	$\square_{\square}$
D_n	W_{1n}	W_{2n}	W_{3n}	$\square_{...}$	T_{tn}

Gambar 6. Matrik *term-document (Baeza, 1999)*

Proses perhitungan VSM melalui tahapan perhitungan *term frequency (tf), Inverse Document Frequency (idf), term frequency Inverse Document Frequency (tfidf),* Jarak query dan dokumen, pengukuran Similaritas *query document (inner product),* dan pengukuran *Cosine Similarity* (menghitung nilai kosinus sudut antara dua vector).

Melalui VSM dan *tfidf weighting* akan didapatkan representasi nilai numerik dokumen sehingga dapat dihitung kedekatan antar dokumen. Semakin dekat dua vektor didalam suatu VSM, maka semakin mirip dua dokumen yang diwakili oleh dua vektor tersebut. Kemiripan antar dokumen dapat dihitung menggunakan suatu fungsi ukuran kemiripan *(similarity measure).* Ukuran ini memungkinkan perankingan dokumen sesuai dengan kemiripannya atau relevansinya terhadap *query*. *Cosine Similarity* atau *Sim(q,dj)* digunakan untuk mengevaluasi tingkat similaritas atau kemiripan dari dokumen (d_j) berkaitan dengan query (q) sebagai korelasi antara vektor dj dan q. Korelasi ini bisa diukur, dengan persamaan (1)

$$Sim(q,d_j) = \frac{q.d_j}{|q| * |d_j|} = \frac{\sum_{i=1}^{t} W_{iq} . W_{ij}}{\sqrt{\sum_{j=1}^{t} (W_{iq})^2} * \sqrt{\sum_{i=1}^{t} (W_{ij})^2}} \quad (1)$$

METODOLOGI PENELITIAN

a. Obyek Penelitian

Obyek penelitian dari penelitian ini adalah Bahasa Jawa Ngoko.

b. 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:

1) Observasi

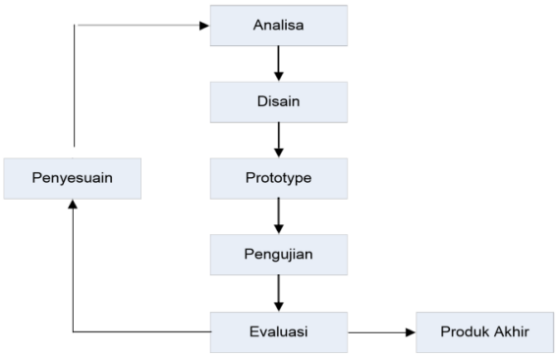
Melakukan pengamatan dan pencatatan secara sistematis tentang hal-hal yang berhubungan dengan basis data dokumen teks dan kemampuan pencarian kemiripan dokumen.

2) Studi Pustaka

Pengumpulan data dari bahan-bahan referensi, arsip, dan dokumen yang berhubungan dengan permasalan dalam penelitian ini.

c. 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 7 memperlihatkan tahapan pada *prototyping*.



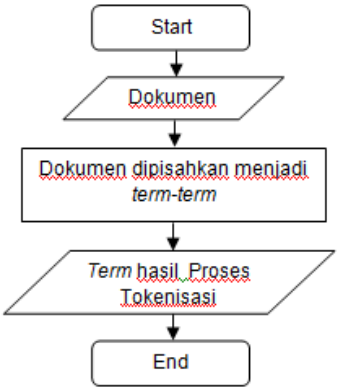
Gambar 7. Tahapan *Prototyping (Pressman, 2001)*

HASIL

a. *Flowchart* Tokenisasi

Proses Tokenisasi dirancang untuk dapat memisahkan dokumen menjadi *term-term* yang

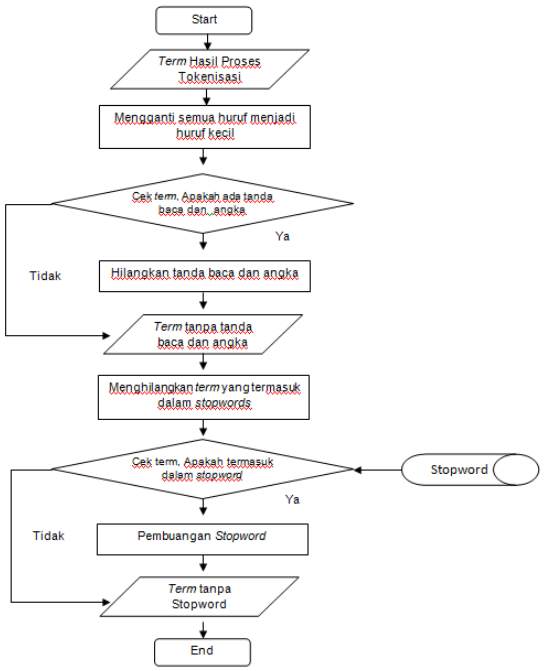
akan diproses pada tahap *filtering*. Proses tokenisasi diawali dengan *scanner* dokumen yang ada pada korpus kemudian diproses menjadi *term*. *Flowchart* tokenisasi bisa dilihat pada gambar 8.



Gambar 8. *Flowchart* Proses Tokenisasi

b. Flowchart Filtering

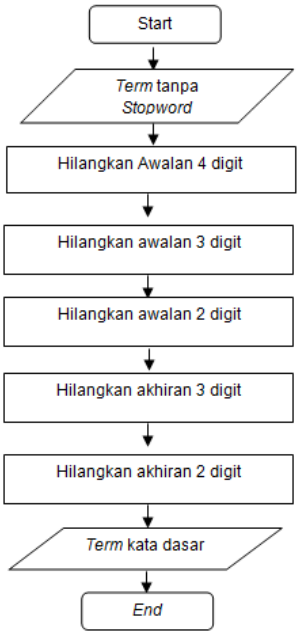
Proses *Filtering* dirancang untuk menghasilkan *term* tanpa *stopwords*. *Flowchart filtering* dimulai dengan mengganti huruf kapital menjadi huruf kecil, menghilangkan tanda baca dan angka, dan menghilangkan *term* yang termasuk dalam *stopwords*. Gambar 9. menunjukkan *flowchart* proses *filtering.*



Gambar 9. *Flowchart* Proses *Filtering*

c. Flowchart Stemming

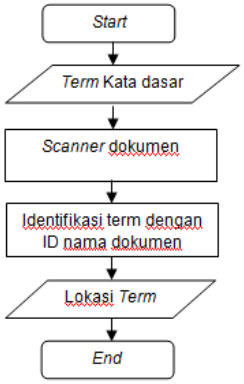
Proses *stemming* dirancang agar *term* hasil *filtering* diubah menjadi *term* kata dasar. Proses *stemming* dimulai dengan menghilangkan awalan dan akhiran. Proses ini juga dirancang dapat melakukan *replace* ketika awalan dihilangkan dan menggantinya dengan huruf yang sesuai. Proses menghilangkan awalan, akhiran, dan *replace* sisipan dilakukan dalam satu tahap proses. Gambar 10 menunjukkan *flowchart stemming*.



Gambar 10. *Flowchart* Proses *Stemming*

d. Flowchar Indexing

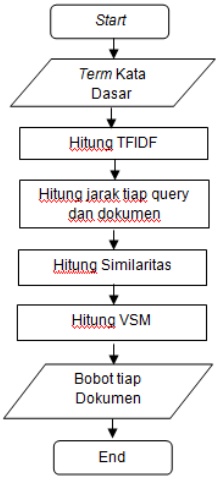
Term kata dasar hasil proses *stemming* selanjutnya dimasukkan dalam tabel untuk diproses pada perhitungan *Vector Space Model*. Proses *indexing* menggunakan metode *inverted indexing*, yaitu dengan membedakan letak tiap term dalam dokumen. Gambar 11 menunjukkan *flowchart indexing*.



Gambar 11. *Flowchart* Proses *Indexing*

e. Flowchart Hitung VSM

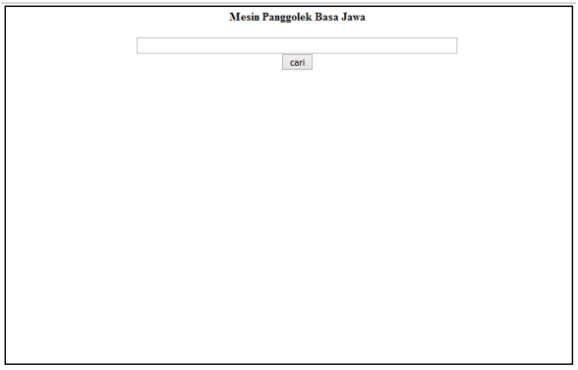
Proses selanjutnya adalah proses perhitungan pembobotan menggunakan metode VSM. Proses ini dimulai dengan perhitungan tf, idf, tfidf, jarak dokumen dan query, similaritas dan *Cosine Similarity* (gambar 12). Proses hitung VSM dirancang menghasilkan dokumen hasil pencarian disertai dengan letak dokumen dan bobot dokumen.



Gambar 12. *Flowchart* Proses Hitung *Vector Space Model*

f. IRS Jawa Ngoko

Interface ini akan ditampilkan kolom *query* yang bisa digunakan untuk memasukkan *query* oleh pengguna. Kotak *button* dengan label cari digunakan untuk memproses setelah *query* di *input*. Tombol *button* cari jika sudah diklik akan menampilkan abstraksi hasil pencarian. Gambar rancangan *Interface* dapat dilihat pada gambar 13.



Gambar 13. *Interface IRS Jawa Ngoko*

PEMBAHASAN

a. IRS Bahasa Jawa Ngoko

IRS bekerja melalui beberapa tahapan proses sebagai berikut; Proses Tokenisasi dilakukan dengan mekanisme jika dokumen pada korpus ditemukan spasi, maka *term* yang ada diantara spasi akan di *retrieved* (akan diambil oleh sistem) kemudian *term* ditempatkan dalam tabel tabelawal. Hasil proses berupa *term* asli (term yang masih memiliki imbuhan, tanda baca yang melekat, dan angka). Keunggulan proses ini waktu komputasi cepat.

Proses *Filtering* dilakukan dengan mekanisme jika *term* pada tabel tabelawal ditemukan tanda baca, huruf kapital, dan angka. Maka program akan menghilangkan (tanda baca dan angka) dan mengganti (huruf kapital menjadi huruf kecil), kemudian memeriksa *term* dengan *stopwords*. Hasil proses berupa *term* pilihan (tanpa tanda baca, tanpa huruf kapital, dan bukan termasuk *stopwords*). Keunggulan proses ini sistem mampu mereduksi tanda baca, angka, merubah term menjadi huruf kecil, dan memeriksa term stopwords dengan waktu komputasi yang cepat.

Proses Tokenisasi dan *Filtering* diproses dalam 1 (satu) tahap dengan waktu komputasi 5 detik (jumlah dokumen 252).

Proses *Stemming* dilakukan program dengan cara menghilangkan imbuhan yang terdapat pada term hasil *filtering*. Proses menghilangkan dilakukan dengan

menghilangkan awalan, sisipan, dan akhiran. Hasil proses ini dimasukkan dalam tabel tabelfreq. Hasil proses berupa *term* kata dasar. Keunggulan proses ini waktu komputasi 13 detik (jumlah dokumen 252)

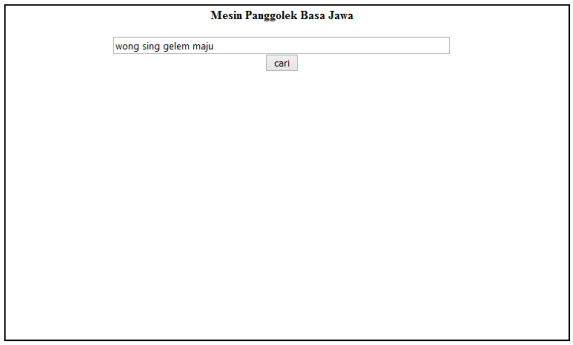
Proses Pembobotan dokumen dengan metode VSM dilakukan dalam proses pencarian dokumen. Program akan bekerja ketika *user* melakukan *query*, selanjutnya program akan memproses *query* tersebut dengan perhitungan-perhitungan *tf, idf, tfidf,* jarak *query* dan dokumen, similaritas dan *cosine similarity*. Hasil proses pembobotan dengan metode VSM berupa dokumen hasil pencarian disertai dengan bobot dokumen, letak dokumen dan disusun *descending* (dokumen dengan bobot terbesar diletakkkan di atas). Keunggulan Proses waktu komputasi rata-rata 1,5 detik dengan hasil pencarian yang akurat (*precision* tinggi).

b. Proses VSM

Information Retrieval System akan melakukan proses perhitungan dimulai dari menghitung tfidf, menghitung jarak *query* dan jarak dokumen, menghitung similaritas produk, dan menghitung bobot dokumen. *Query* yang di *input* oleh user selanjutnya akan dilakukan pencarian pada tabel freq kemudian dilakukan perhitungan pembobotan menggunakan metode *Vector Space Model*. Perhitungan dilakukan dalam sistem pencarian, sistem pencarian akan melakukan perhitungan kemudian akan menampilkan hasilnya. Hasil pencarian akan menampilkan nama dokumen di korpus, kemudian bobot similaritas dan disusun berdasarkan perankingan. Bobot terbesar akan menempati ranking teratas pada hasi pencarian.

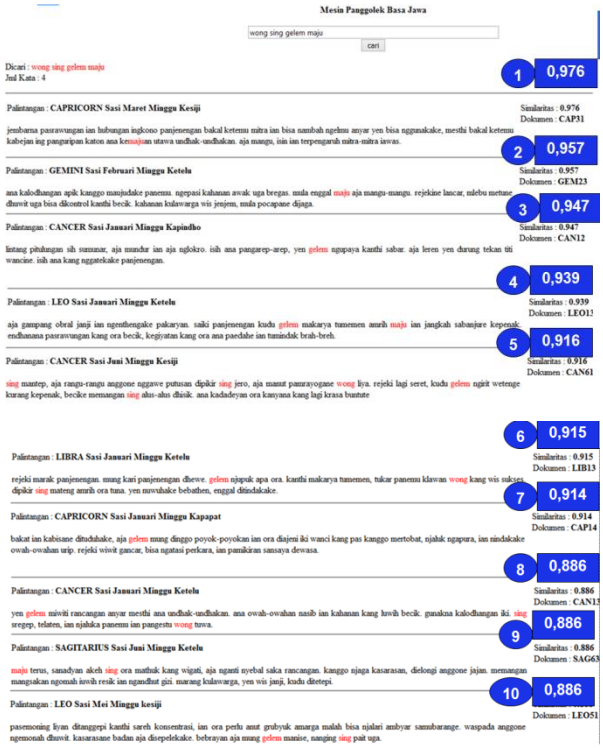
c. Aplikasi *keyword* Bahasa *Jawa ngoko*

Studi kasus pada aplikasi Information Retrieval System ini menggunakan dokumen-dokumen Palintangan Basa Jawa pada Majalah Online Penjebar Semangad yang terdapat pada 12 Palintangan (zodiak) yaitu; Capricorn, Aquarius, Pisces, Aries, Taurus, Gemini, Cancer, Leo, Virgo, Libra, Scorpio, dan Sagitarius. *Query* yang dimasukkan pada Information Retrieval System adalah *keyword* dengan 2 *term* yaitu "*seneng ngalah*", 3 *term* "*wong angel sukses*". 4 *term* "*wong sing gelem maju*" (gambar 14). 5 *term* "*wong sing sabar lan waspada*". 6 *term* "*wong sing angel sukses lan mutungan*". Dan 7 *term* "*wong sing angel sukses lan angel maju*".



Gambar 14. Aplikasi Mesin Panggolek Basa Jawa

Hasil pencarian dengan term " Wong Sing Gelem Maju" bisa dilihat pada gambar 15.



Gambar 15. Hasil Pencarian *keyword*

Hasil pencarian dokumen dengan keyword "*Wong sing gelem maju*", menunjukkan dokumen dengan bobot tertinggi adalah dokumen letak dokumen CAP31 (bobot 0,976). Dokumen CAP31 memiliki frekuensi kemunculan *term* tertinggi dan tidak banyak

muncul didokumen lain untuk *term* "*Wong sing gelem maju*". Dokumen CAP31 (dokumen Palintangan Capricorn Sasi Maret Minggu Kesiji nomer 31) memiliki bobot tertinggi atau memiliki tingkat kemiripan tertinggi dibandingkan dengan dokumen lain yang ada pada korpus.

Dari Hasil pencarian *Term* "*Wong sing gelem maju*" dengan hasil didapatkan 10 peringkat teratas paling banyak hasil didapatkan oleh palintangan CANCER (3 kali muncul).

d. Pengujian *Recall* dan *Precision*

Pengujian *recall (P)* dan *precision (R)* dilakukan dengan cara *input query* ke dalam Information Retrieval System *input* 1 *term*, 2 *term* dan 3 *term, 4 term, 5 term, 6 term dan 7 term*. Perhitungan *recall* dan *precision* menggunakan persamaan (8) dan persamaan (9). Hasil pengujian *recall* dan *precision* dengan menguji 1 *term*, 2 *term* dan 3 *term* sampai dengan 7 term menunjukkan bahwa jika *recall* rendah maka *precision* akan tinggi, selengkapnya terlihat pada tabel 5.18.

Tabel 1. Hasil Pengujian beberapa keyword

No	Query	Recall	Precision
1	Sabar	0.05	1.00
2	Sukses	0.02	1.00
3	seneng ngalah	0.04	0.82
4	wong angel sukses	0.03	0.80
5	wong sing susah asile	0.07	0.77
6	wong sing gelem maju	0.06	0.89
7	wong sing ora gelem maju	0.06	0.78
8	wong sing sabar lan waspada	0.04	0.77
9	wong sing angel sukses lan mutungan	0.04	0.90
10	wong sing angel sukses lan angel maju	0.01	0.67

e. Pengujian Waktu Program *IRS Jawa ngoko*

Preprosesing yang dididalamnya terdapat proses Tokenisasi, *filtering* dan *Stemming*. Proses preprosesing pada penelitian ini dibagi menjadi dua tahap, yaitu pada tahap awal adalah memproses data yang ada di korpus kemudian memasukkannya pada tabel tabelawal dan tabel tabelkedua. Selanjutnya dengan menggunakan tabelkedua diolah dan data berupa *term* yang sudah menjadi kata dasar dimasukkan pada tabel tabelfrek.

Waktu yang dibutuhkan dalam proses komputasi proses tokenisasi dengan jumlah dokumen 252 adalah 3 detik. Sedangkan waktu komputasi untuk proses *filtering* adalah 13 detik. Jadi total waktu yang dibutuhkan dalam proses preprosesing dengan jumlah dokumen 252 adalah 18 detik (perhitungan waktu komputasi dilakukan menggunakan *stopwatch*).

Waktu komputasi preprosesing bergantung pada jumlah data yang akan diolahnya. Semakin besar data yang akan diproses membutuhkan waktu yang lebih lama. Waktu preposesing untuk *scanner file* dari tabelkedua menjadi tabelfrek dipilih dengan jumlah dokumen 252. Waktu komputasi proses pembobotan bergantung pada jumlah *term* yang terambil untuk diolah. Semakin besar jumlah *term* yang harus diolah akan membutuhkan waktu yang lebih lama. Waktu pencarian rata-rata untuk pencarian 1, 2, 3, 4, 5, 6 dan 7 *keyword* adalah 5 detik

KESIMPULAN DAN SARAN

Kesimpulan

a. IRS mampu melakukan pencarian dokumen teks dan menampilkan hasil pencarian dokumen teks berbahasa Jawa Ngoko dengan disertai bobot tiap dokumen beserta letak dokumen

b. Hasil Uji *recall* dan *precision* STKI menunjukkan hasil pencarian dokumen teks memiliki rata-rata *recall* = 0,04 dan rata-rata *precision* = 0,84.

c. IRS yang dibangun memiliki keunggulan mampu melakukan pencarian dokumen teks bahasa jawa ngoko dan hasil pencarian yang akurat (*precision* = 0,84), serta dilengkapi dengan bobot dan letak dokumen pada database.

Saran

a. Penulisan bahasa jawa yang benar perlu dilakukan karena berkemungkinan tidak akan terambil oleh sistem jika penulisannya salah.

b. Proses *stemming* yang ada masih belum bisa sepenuhnya membuat semua *term* kedalam bentuk *term* kata dasar dengan benar. Proses ini akan mempengaruhi hasil untuk proses indexing, sehingga akan mempengaruhi hasil akhir perhitungan.

DAFTAR PUSTAKA

Budi, I., Aji, R.F., 2006. Efektifitas Seleksi Fitur dalam Sistem Temu Kembali Informasi. Seminar Nasional Aplikasi Teknologi Informasi (SNATI), ISSN: 1907-5022.

Bum, K.Y., 2010. *An autonomous assessment system based on combined latent semantic kernels. Expert Systems with Applications: An International Journal, Volume 37 Issue 4.*

Bunyamin., 2008. Aplikasi IR CATA dengan Metode *Generalized Vector Space Model*, Jurnal Informatika, Vol. 4. No. 1

Bo, Y., 2009. Combining neural networks and semantic feature space for email classification. Knowledge-Based *Systems*, Volume 22 Issue 5.

Erk, K., 2008. *A Structured Vector Space Model for Word Meaning in Context. Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*. 897-906.

Erk, K., 2010. *A flexible, corpus-driven model of regular and inverse selectional preferences . Computational Linguistics*, Volume 36 Issue 4.

Haryono, M.E.A., 2005. Pembentukan Intisari Topik secara Otomatis dalam suatu Paragraf dengan *Vector Space Model.* Seminar Nasional Aplikasi Teknologi Informasi, SNATI 2005.

Haryono, M.EA., Wahyudi., 2005. *Customer Information Gathering* menggunakan Metode Temu Kembali Informasi dengan Model Ruang Vektor. Seminar Nasional Aplikasi Teknologi Informasi, SNATI 2005.

Hasugian, J., 2006. Penggunaan Bahasa Alamiah dan Kosa Kata Terkendali dalam Sistem Temu Balik Informasi berbasis Teks. Jurnal Studi Perpustakaan dan Informasi, Vol 2, No 2 Desember.

Kadir, A., 2001. Dasar Pemrograman Web Dinamis menggunakan PHP. Penerbit Andi. Yogyakarta.

Lopez, C., Ribeiro, C., 2010. *Using Local Precision to Compare Search Engines in Consumer Health Information Retrieval*. **SIGIR '10:** *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*.

Manning, C., Raghavan, P., 2007. An Introduction to Information Retrieval, Stanford. USA.

Mao, W., 2007. *The phrase-based vector space model for automatic retrieval of free-text medical documents*. *Data & Knowledge Engineering*, Volume 61 Issue 1.

Meadow, C.T., 1997. *Text Information Retrieval Systems*. Academic Press.New York.

Mcenery,1998.**W3Corporaproject**.*1998*http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction2.html diakses tanggal 14 juli 2012.

Mondal, D., Gangopadhyay, A., 2010. *Medical Decision Making Using Vector Space Model.* IHI '10**:** *Proceedings of the 1st ACM International Health Informatics Symposium.*

Peranginangin, K., 2006. Aplikasi Web dengan PHP dan MySQL. Penerbit Andi, Yogyakarta.

Tala, F.Z., 2003, *A Study of Stemming Effects on Information Retrieval in bahasa Indonesia*. Institut for logic, Language and

Computation Universiteit van Amsterdam The Netherlands.

Rijsbergen, C.J.V., 1979. Information *Retrieval, Second Edition. Butterworths*. London.

Salton, G., 1989, *Automatic Text Processing, The Transformation, Analysis, and Retrieval of information by computer*. Addison – Wesly Publishing Company, Inc. USA.

Yates, R.B, 1999. *Modern Information Retrieval, Addison Wesley*-Pearson international edition, Boston. USA.