INFORMATION SYSTEM APPLICATION,
Vol. 02, No. 01 ISSN 2503-166X

Perbandingan Pembobotan Kata Dalam Sistem Temu Balik Informasi

[1]**Amril Mutoi Siregar**
[1] Universitas Buana Perjuangan, Jalan HS. Ronggo Waluyo, Telukjambe Timur, Puseurjaya
Telukjambe Timur, Kabupaten Karawang, Jawa Barat 41361
E-mail: amrilmutoi@upbkarawang.ac.id

***Abstrak*— Information retrieval (IR) system adalah sistem yang secara otomatis melakukan pencarian informasi yang relevan terhadap kebutuhan pengguna, yang diekspresikan dalam query, menjadi input bagi IR system dan selanjutnya diproses oleh system kemudian ditampilkan dokumen yang relevan dengan query tersebut.**

Salah satu metode pencarian informasi yang relevan dengan query adalah dengan pembobotan kata baik query maupun dokumen. Metode yang sering digunakan adalah pembobotan kata lokal dan pembobotan global, adapun pembobotan kata lokal yang digunakan tf, logaritmic tf, binary tf dan augmented tf dan pembobotan globalnya adalah idf, idfp, idfb. Dalam penelitian ini selain membandingkan algoritma pembobotan kata. Tingkat keberhasilan algoritma di ukur berdasarkan precision, recall dan niap.

Penelitian ini setelah dilakukan perbandingan dan analisis hasil algoritma pembobotan, hasil algortima yang baru lebih baik daripada yang lama. Kedepan perlu dilakukan dikombinasikan pembobotan dengan algoritma semantic sehingga mendapatkan hasil yang lebih baik.

Kata kunci: ***Sistem temu balik informasi, pembobotan, pembobotan kata lokal, pembobotan global, similarity(kemiripan), vector space model***

I. PENDAHULUAN

Sistem temu kembali informasi *(information retrieval system*) adalah suatu sistem yang melakukan pencarian informasi relevan dari koleksi dokumen terhadap *query*. *Query* tersebut berupa kata/istilah dapat menggambarkan informasi yang sedang dicari dalam bentuk bahasa sehari-hari.

Seiring dengan perkembangan informasi, banyak pihak menyadari bahwa masalah utama telah bergeser dari cara mengakses informasi menjadi memilih informasi yang berguna secara selektif. Usaha untuk memilih informasi ternyata lebih susah dari sekedar mendapatkan akses informasi tersebut. Pemilihan informasi tidak mungkin dilakukan secara manual karena kumpulan informasi yang sangat besar dan terus bertambah besar.

Suatu sistem otomatis diperlukan untuk membantu pengguna dalam menemukan informasi. *Information retrieval system* adalah sistem yang digunakan untuk menemukan informasi yang relevan terhadap kebutuhan penggunanya secara otomatis dari koleksi dokumen.

Sistem melakukan pengindeksan terhadap dokumen untuk mempermudah dan mempercepat proses pencarian. Relevansi ditentukan dengan menghitung nilai kemiripan antara dokumen-dokumen yang ada dengan *query*, dan direpresentasikan ke dalam bentuk tertentu. Dokumen-dokumen yang diperoleh, kemudian sistem mengurutkan berdasarkan tingkat relevansinya terhadap *query*. Banyak algoritma yang telah dilakukan peneliti sebelumnya untuk pembobotan, baik berasal dari dokumen yang disebut sebagai pembobotan kata lokal, dan koleksi dokumen disebut sebagai pembobotan global. Ternyata yang dihasilkan belum cukup baik, apalagi mencari informasi relevan yang berhubungan makna semantik anatara *query* dengan dokumen.

II. KAJIAN TEORI

A. *Sistem Temu Balik Informasi*

Sistem temu balik informasi adalah salah satu penerapan teknologi komputer untuk perolehan, pengorganisasian, penyimpanan, pencarian dan pendistribusian informasi. Tujuan Sistem temu balik informasi (information retrieval) adalah untuk mencari informasi relevan terhadap kebutuhan pengguna, menggunakan search engine dengan memasukkan query untuk mendapatkan informasi yang diinginkan (Baeza-Yates, Ricardo, dan Berthier Ribeiro-Neto,1999).

Query dimasukkan belum tentu memberikan hasil pencarian yang baik. Walaupun begitu, query hanya sebagai kata kunci untuk sebuah search engine dalam melakukan pencarian untuk mendapatkan hasil yang diinginkan oleh pengguna (S.M Jeckson,2002).

Information retrieval dapat mencari informasi yang berbentuk dokumen yang tidak terstruktur, yang memenuhi kebutuhan informasi dari dalam kumpulan data yang besar. Bidang information retrieval juga mencakup dalam mencari ataupun menyaring kumpulan dokumen atau

pemrosesan lebih lanjut terhadap serangkaian dokumen yang diambil (Manning Christopher D, Prabhakar Raghavan dan Hinrich Schutze,2009).

Sistem temu balik informasi (information retrieval system) adalah suatu sistem yang digunakan untuk mendapatkan informasi-informasi yang relevan terhadap kebutuhan pengguna diambil dari koleksi dokumen yang dapat diakses secara otomatis (Mandala, Rila dan Setiawan, 2002). Aplikasi yang sering temukan di internet adalah berupa search engine atau dikenal dengan mesin pencarian. Pengguna bebas mencari informasi informasi yang relevan baik berupa teks, gambar dan lainya.

Sebuah sistem temu balik informasi tidak memberitahu kepada pengguna, mengenai masalah yang ditanyakannya. Sistem tersebut hanya memberitahukan keberadaan (atau ketidakberadaan) dan keterangan dokumen-dokumen yang berhubungan dengan permintaannya pengguna (Rijsbergen. C.J,1979).

B. Skema Pembobotan kata *(Term Weighting)*

Dalam sistem temu kembali informasi proses pembobotan kata sangat menentukan perangkingan dokumen untuk disajikan kepada pengguna. Proses dimulai dari kata di dalam suatu indeks harus bisa membedakan tujuan dari sebuah dokumen pada sebuah informasi. Caranya yaitu dengan pemberian bobot kepada sebuah kata terhadap suatu dokumen. Semakin tinggi bobot dari sebuah kata maka semakin penting kata tersebut, dibandingkan dengan kata lainnya di dalam sebuah dokumen. Pembobotan kata ini dicantumkan pada *inverted file* untuk digunakan dalam proses temu balik dokumen.

Pada saat pencarian kata tunggal digunakan untuk mengidentifikasi dari isi sekumpulan dokumen, pembedaan harus dilakukan antar kata tunggal berdasarkan perkiraan nilai kata tersebut sebagai pendeskripsi sebuah dokumen. Hal ini menunjukkan penggunaan dari pemboobtan kata yang dicantumkan pada saat proses pengidentifikasian (Salton G and Buckley C ,1988).

Sebagai contoh terhadap sebuah dokumen D dinyatakan seperti:

D={Ti1, 0.2; Ti2, 0.5 ; Ti3, 0,8; Ti, 0,9}

Dari pembobotan diatas dapat disimpulkan bahwa kata ke empat memiliki bobot 0.9 merupakan bobot lebih tinggi sedangkan kata pertama memiliki bobot yang jauh lebih kecil yaitu sebesar 0.2. Penggunaan dari bobot kata selain untuk membedakan kepentingan suatu kata di dalam sebuah dokumen juga dapat digunakan untuk menggunakan pengurutan saat temu balik dengan perangkingan dari bobot yang besar ke kecil sesuai dengan bobot kata kata yang sama antara *query* dan dokumen.

Penelitian ini menggunakan skema pembobotan kata akhir adalah sebagai berikut:

$$T_{i,j} = L_{i,j} \; x \; G_{i,j} \; x \; N_j$$

Keterangan :

$T_{i,j}$ = Pembobotan kata gabungan dalam dokumen *j*

$L_{i,j}$ = Pembobotan kata lokaldalam dokumen *j*,

$G_{i,j}$ = Pembobotan kata global dalam koleksi dokumen

N_j = Faktor normalisasi pada dokumen *j*.

C. Pembobotan Kata Lokal

Penelitian ini menggunakan skema pembobotan kata lokal, diantara skema pembobotan memiliki kekurangan dan kelebihannya yaitu metode Logaritma (LOG) dianggap lebih baik dan stabil apabila dibandingkan dengan metode lain. Karena logaritma digunakan untuk mengatur frekuensi kata dalam dokumen. Misalnya sepuluh kali kata yang muncul dalam dokumen belum tentu sepuluh kali lebih penting daripada kata yang muncul sekali dalam dokumen tersebut. Metode binary (BIN) adalah tidak membedakan antara kata dalam dokumen yang sering muncul dan kata yang jarang muncul. Metode *tf (term frequency*) adalah memberikan terlalu banyak bobot kata untuk kata yang sering muncul dalam dokumen.

Berikut beberapa pembobotan kata lokal yang umum digunakan dalam sistem temu balik informasi.

Tabel 2.1 Pembobotan kata lokal

Formula	Name	Singkatan
$Tf_{i,j} \frac{Tf_{i,j}}{\max(Tf_{i,j})}$	*Term frequency*	FREK
$1 + \log(Tf_{i,j})$, if $Tf_{i,j}$> 0 0 , if $Tf_{i,j}$=0	*Logaritmic*	LOG
1, $if\ Tf_{i,j} > 0$ 0, $if\ Tf_{i,j} = 0$	*Binary*	BIN
$0.5 + 0.5 * \frac{Tf_{i,j}}{\max(Tf_{i,j})}$, if $Tf_{i,j}$>0 0, $if\ Tf_{i,j} = 0$	*Augmented term frequency*	ATP

D. Pembobotan Kata Global

Pembobotan kata global yang umum digunakan adalah kebalikan frekuensi dokumen (*IDF*), dan menggunakan dua variasi yaitu *IDF* dan *IDFP*, didalam rumusnya, mana *N* adalah jumlah dokumen dalam koleksi dan *ni* adalah jumlah kata di dalam dokumen yang mengandung kata *i (*Frakes William and Baeza-Yates,Ricardo,1992). *IDF* adalah logaritma dari kebalikan kata yang muncul dalam dokumen secara acak. *IDFP* adalah logaritma dari kebalikan dari probabilitas bahwa kata yang muncul dalam dokumen secara acak. *IDF* dan *IDFP* adalah sama bahwa kedua algoritma mempunyai bobot tinggi untuk kata yang muncul didalam dokumen, dan bobot rendah untuk kata yang jarang muncul didalam dokumen. Namun perbedaannya *IDFP* sebenarnya memberikan bobot negatif untuk bobot kata yang muncul lebih dari setengah dari dokumen dalam koleksi.

Tabel 2.2 Pembobotan global lama

Formula	Name	Singkatan
$\log(\frac{N}{ni})$	*Inverse Document Frequency*	IDF
$\log(\frac{N-ni}{ni})$	*Probabilistic Inverse Document Frequency*	IDFP
$\log\ (\frac{N-ni+0.5}{ni+0.5})$	*BM25 Inverse Document Frequency*	IDFB

E. Faktor Normalisasi

Faktor normalisasi berfungsi untuk menormalkan frekuensi kata yang terlalu tinggi dalam dokumen. Karena dokumen memiliki karakteristik yang panjang dan beragam. Ketimpangan terjadi dalam dokumen yang panjang, cenderung mempunyai frekuensi kemunculan kata yang besar. Sehingga untuk mengurangi ketimpangan tersebut diperlukan faktor normalisasi dalam pembobotan kata. Untuk mengurangi pengaruh perbedaan panjang dalam dokumen, pada pembobotan kata digunakan satu faktor untuk men-normalisasikan panjang dokumen, adapun normalisasi yang digunakan adalah:

Tabel 2.3. Faktor Normalisasi

Formula	*Name*	Singkatan
$\frac{1}{\sqrt{\sum_{i=0}^{n}(G_i.L_{i,j})2}}$	*Cosinus* Normalisasi	*NORM*
1	*None*	*NONE*

F. *Recall, Precision, Niap*

Evaluasi dari sistem temu kembali informasi dipengaruhi oleh dua parameter utama yaitu *recall* dan *precision*. *Recall* adalah rasio antara dokumen relevan yang berhasil ditemu kembalikan, dari seluruh dokumen relevan yang ada di dalam sistem, sedangkan *precision* adalah rasio dokumen relevan yang berhasil ditemu kembalikan dari seluruh dokumen yang berhasil ditemu kembalikan (Grossman,D,1992).

NIAP (non interpolated average precision) digunakan untuk mengecek keberhasilan pencarian dari perangkat lunak yang dibangun (Mizzaro,S,1998). Dan ukuran ideal dari keefektifan suatu sistem temu kembali informasi adalah apabila *rasio Recall* dan *Precision* sama besarnya (1:1). Dengan demikian, dapat disimpulkan bahwa bagian terpenting dalam proses temu balik informasi adalah ketika kebutuhan informasi pengguna tercapai dikarenakan *precision* (ketepatan) yang dihasilkan.

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

NIAP adalah rata rata ukuran yang menggambarkan performansi semua dokumen yang relevan.

III. METODE PENELITIAN

3.1. Model Rancangan Penelitian

Dalam penelitian ini menggunakan tahapan dalam melakukan eksperimen, tujuannya adalah untuk memaksimalkan hasil sistem temu balik informasi. Perlu diperhatikan bagaimana proses mengolah dokumen tersebut, agar lebih mudah dalam pencarian dokumen yang relevan. Adapun susunan metode penelitian yang dilakukan seperti pada Gambar 3.1.

Metode yang diusulkan untuk penelitian ini menggunakan dua alur dalam sistem temu balik informasi yang terdiri dari *query* dan dokumen yang masing masing mempunyai tahapan. Dan dapat analiasa bahwa terdapat dua buah alur operasi pada sistem temu kembali informasi tersebut. Alur pertama dimulai dari koleksi dokumen dan alur kedua dimulai dari *query* pengguna. Alur pertama yaitu pemrosesan terhadap koleksi dokumen menjadi *inverted file* dan tidak tergantung pada alur kedua. Sedangkan alur kedua tergantung dari keberadaan *inverted file* yang dihasilkan pada alur pertama.

3.2. Query Dan Koleksi Dokumen

Pengguna melakukan pencarian dokumen dengan membuat suatu *query,* penelitian ini menggunakan 35 *query* dalam data set *ADI*. Dalam test koleksi dokumen terdiri dari: Koleksi dokumen, *query*, *relevan judgment*

Jumlah *query* dalam koleksi dokumen pengujian terdiri dari beberapa *query* yang berbeda. Hal ini dilakukan untuk mendekati keadaan dunia nyata. Koleksi pengujian yang digunakan untuk evaluasi pengaruh umpan balik pada performansi sistem temu kembali informasi, terdiri dari 3 buah koleksi pengujian yaitu *dataset, query*, relevansinya (*relevan judgment*).

3.3. Preprocessing

Tahap sebelum pengolahan dalam temu balik informasi adalah penerapan beberapa teknik untuk *query* dan dokumen dalam mengkonversikan ke format yang lebih ringkas sebelum ketahap *similarity calculation*. Teknik yang banyak digunakan dalam sistem temu balik informasi adalah penghapusan *stopword, stemming*. *Stopword* adalah penghapusan kata-kata yang terlalu sering dalam dokumen.

3.4. Query Formulation and *Indexing*

*Query formulation (*formulasi terhadap *query*) adalah tahap pemberian bobot terhadap indeks kata-kata dalam *query* yang dimasukan oleh pengguna. *Indexing* adalah proses sistem yang merepresentasikan koleksi dokumen kedalam bentuk tertentu untuk memudahkan dan mempercepat proses pencarian dan penemuan kembali dokumen yang relevan. Pembangunan *index* dari koleksi dokumen merupakan tugas pokok pada tahapan *preprocessing* di dalam sistem temu balik informasi. Kualitas *index* mempengaruhi efektifitas dan efisiensi sistem.

3.5. Comparison and Similarity Calculation

Comparison adalah salah satu model sistem temu balik informasi berupa menghitung perbedaan, kesamaan, probabilitas. Pada tahap ini *query* dan dokumen telah berubah menjadi terstruktur atau objeknya sebelum dilakukan perbandingan. *Comparison* menggunakan fungsi kecocokan untuk memberikan ukuran numerik dari yang terkait (diperkirakan relevansi) antara setiap dokumen dan *query*. Ukuran numerik yang dihasilkan dan dibatasi hanya oleh model itu sendiri. Dokumen-dokumen yang

mempunyai kemiripan dengan *query* akan dikembalikan pengguna.

3.6. *Retrieved Document*

Retreieved document adalah proses mengambilan dokumen-dokumen relevan terhadap *query yang dimasukkan oleh pengguna* dan mengurutkan dokumen tersebut berdasarkan kemiripan dengan *query* dengan dokumen yang telah dihitung tingkat kemiripannya, kemudian berikan kepada pengguna dalam bentuk perangkingan dokumen.

3.7. *Tampilan Dokumen*

Tampilan dokumen dalam penelitian ini adalah menampilkan hasil dokumen yang diambil oleh sistem temu balik informasi. Didalam tampilan akan terlihat hasil *recall, precision, dan non interpolated avarege precision (NIAP)*

IV. HASIL DAN PEMBAHASAN

4.1 Hasil Penelitian

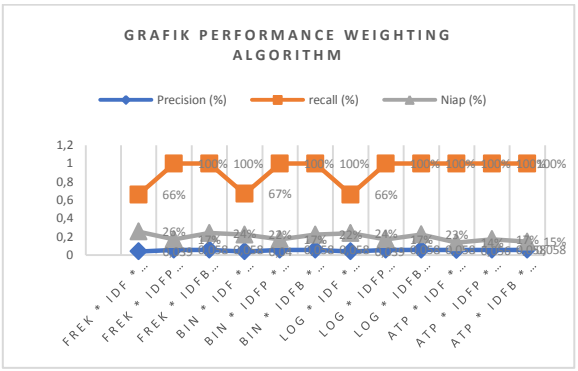
Eksperimen yang dilakukan mendapatkan hasil yang berbeda antara algoritma yang satu dengan yang lain. Dan ukuran hasil dapat di evaluasi menggunakan *recall, precision* dan *non-interpolated average precision (niap)*. Dalam eksperimen ada beberapa dievaluasi yaitu :

- Membandingkan hasil algoritma yang satu dengan lainnya.
- Mencari hasil algoritma yang paling optimal diantara algoritma yang ada.

Mencari kelebihan dan kelemahan berdasarkan hasil performasi algoritma

Tabel 4.1 Hasil penelitian

Weighting Algorithm	*Precision (%)*	*recall* (%)	*Niap* (%)
Frek * idf * norm	0.039	66%	25.61%
Frek * idfp * norm	0.058	100%	17.20%
Frek * idfb * norm	0.058	100%	24.00%
Bin * idf * norm	0.040	67%	22.40%
Bin * idfp * norm	0.058	100%	17.20%
Bin * idfb * norm	0.058	100%	22.30%
Log * idf * norm	0.039	66%	23.90%
Log * idfp * norm	0.058	100%	17.20%
Log * idfb * norm	0.058	100%	22.50%
Atp * idf * norm	0.056	100%	13.70%
Atp * idfp * norm	0.058	100%	17.20%
Atp * idfb * norm	0.058	100%	14.60%



Grafik 4.1 Hasil penelitian

V. KESIMPULAN

Kesimpulan hasil penelitian yang dilakukan terhadap beberapa algoritma pembobotan kata diukur dengan presicion, recall dan niap adalah sebagai berikut:

- Hasil niap dalam performansi algoritma pembobotan kata yang paling tinggi adalah FREK * IDFB * NORM sebesar 0.24.
- Semakin tinggi frekuensi kata dalam dokumen akan semakin kecil nilai idf, dan sebaliknya, dengan penambahan bobot global dapat memperbaiki hasil niap.
- Untuk meningkatkan hasil presisi, recall dan niap tidak tergantung sepenuhnya pada frekuensi kata dalam dokumen.

REFERENSI

[1] Baeza-Yates, Ricardo, Berthier Ribeiro-Neto,1999. Modern Information Retrieval. Harlow. Addison.Wesley.

[2] Frakes William and Baeza-Yates,Ricardo,1992. Information retrieval data structures & algorithms. Prentice Hall.

[3] Grossman,D,1992. IR Book. tp://www.ir.iit.edu/ ~dagr/ cs529/ files/ ir_ book / 7 Maret 2002.

[4] Ingwersen, P, 1992. Information Retrieval Interaction. Taylor Graham Publishing. http://www.db.dk/pi/iri [29 Agustus 2005].London.

[5] Mandala Rila dan Setiawan, 2002. Peningkatan Performansi Sistem Temu Kembali Informasi dengan Perluasan Quer Secara Otomatis. Departemen Teknik Informatika Institut Teknologi Bandung. Bandung.

[6] Manning Christopher D, Prabhakar Raghavan dan Hinrich Schutze,2009. An Introduction To Information Retreival. England. Cambridge University Press.

[7] Mizzaro,S,1998. How many relevances in information retrieval interacting with computers," vol 10(3):305-322.

[8] Rijsbergen C.J,1979. Information Retrieval. Second Edition. Butterworths. London.

[9] Salton G,1969. Automatic Text Analysis. Technical Report No. 69-36. Department of Computer Science. Cornell University, Ithaca.New York.

[10] Salton G, Wong A and Yang CS,1975. A vector space model for automatic indexing. Communication. ACM 18(11):613–620.

[11] Salton G and Buckley C,1988. Term weighting approaches in automatic text retrieval. Information Processing &Management 24(5):513–523.

[12] Sparck Jones K,1972. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21.

[13] Robertson, S,2005. Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, Vol. 60, pp. 502–520. England.S.M Jeckson,2002. A Scientist Practitioner Approach. John Wiley dan Sons.

[14] Witten et all,1999. Compressing and Indexing document dan Images Second Edition. Morgan Kaufmann Publishers. SanFransisco