TEMU KEMBALI INFORMASI BERBASIS KLUSTER UNTUK SISTEM TEMU KEMBALI INFORMASI TEKS BAHASA INDONESIA

Amir Hamzah
Jurusan Teknik Informatika, Fakultas Teknologi Industri
Institut Sains & Teknologi AKPRIND Yogyakarta
Jl. Kalisahak No.28 Komp.Balapan, Yogyakarta 55222
e-mail : amir@akprind.ac.id

ABSTRACT

The exponential growth of textual documents has caused difficulties in the process of informatioan retrieval, mainly in the model of linear retrieval based on word matching that generally ineffective. The word synonimy of a text has triggered to the resulting of non relevan documents in the retrieval, on the other hand polisemy factor has caused many of relevan document remain unretrieved. The application of document clustering can improve the performance of retrieval process according to the hypothesis that the documents relevant to the same query tends to be in the same cluster.

This research studied the application of document clustering to improve the effectiveness of document retrieval by using cluster-based retrieval in the vector space model. In the first step, document collection was clustered using any cluster algorithm and the cluster center was selected to be cluster representative. In the second step, the search process then matched the query to the all cluster representatives and finally the all documents in the cluster that have the highest similarity to the query was selected to present to the user..

The clustering methods used in this study are partitional method (Bisecting K-Mean and Buckshot algorithms) and hierarchical agglomerative method using cluster similarity of UPGMA and Complete Link. The performance of retrieval was measured using F-measure parameter derived from Precision and Recall of retrieva process. The test document collection used are 1000 news text documents with known cluster structure and 3000 news text documents with unknown cluster structure.

The results showed that in the test collection which is evaluated in the retrieval process based on cluster-matching has imporved the performance of 12.3% and 9.5% compare to the process of linear retrieval based on word –matching.

Key words : *information retrieval, clustering, cluster-based retrieval*

INTISARI

Volume informasi teks yang berkembang eksponensial menyebabkan kesulitan dalam proses temu kembali informasi, utamanya pada model perolehan informasi linear berbasis *word matching* yang umumnya tidak efektif. Faktor sinonim dari kata menjadi penyebab munculnya dokumen tidak relevan dalam perolehan, sebaliknya faktor *polisemy* menyebabkan banyak dokumen yang relevan tidak terpanggil. Penerapan *clustering* dokumen dipercaya dapat meningkatkan kinerja berdasar satu *hypothesis* bahwa dokumen yang relevan terhadap suatu *query* cenderung berada dalam kluster yang sama.

Penelitian ini melakukan kajian penerapan *clustering* dokumen untuk meningkatkan perolehan informasi dengan cara melakukan *retrieval* berbasis kluster (*cluster-based retrieval*) dengan model ruang vektor. Koleksi dokumen mula-mula dikluster dan representasi kluster digunakan vektor pusat kluster. Dokumen-dokuman dalam kluster yang pusat klusternya memiliki similaritas tertinggi terhadap *query* dipilih sebagai perolehan.

Metode *clustering* yang dipilih adalah metode *partitional*, yaitu algoritma *Bisecting K-Mean* dan *Buckshot*, dan metode *hierarchical agglomerative* dengan algoritma perhitungan similaritas kluster UPGMA dan *Complete Link*. Kriteria kinerja perolehan informasi diukur dengan parameter F-*measure* yang diturunkan dari *Precision* dan *Recall* dari *retrieval*. Koleksi dokumen yang digunakan adalah 1000 dokumen berita yang telah diketahui struktur klusternya dan 3000 dokumen berita dengan struktur kluster belum diketahui.

Hasil penelitian menunjukkan bahwa pada koleksi yang dicobakan terjadi peningkatan kinerja perolehan informasi berbasis kluster sebesar berturut-turut 12.3% dan 9.5% dibandingkan dengan perolehan linear berbasis *word –matching*.

Kata Kunci : Perolehan informasi, clustering, cluster-based retrieval

PENDAHULUAN

Penerapan teknologi digital dan jaringan komputer telah menyebabkan terjadinya "ledakan" informasi yang berkembang eksponensial. Hal ini menyebabkan Sistem temu kembali informasi (*information retrieval* = *IR*) mengalami kesulitan. Google sebagai mesin pencari terkemuka pada tahun 2006 mengindeks tidak kurang dari 16 milyar dokumen (Google.com, 2006). Sebagian besar (80%) informasi adalah berbentuk teks (Tan,1999).

Pada strategi pencarian query berbasis kata (*word-matching*) kesulitan yang dijumpai muncul dari aspek bahasa, yaitu faktor sinonim pada kata telah menyebabkan dokumen yang tidak relevan akan dipanggil hanya semata-mata karena dokumen tersebut mengandung

kata yang ada dalam query. Sebaliknya faktor *polisemy* , yaitu keadaan di mana suatu kata dapat memiliki lebih dari satu makna, menyebabkan ada dokumen relevan dalam koleksi yang tidak dipanggil karena tidak memuat kata yang ada dalam query. Kesulitan ini semakin kompleks manakala pada kenyataannya koleksi dokumen cenderung bertambah besar dan akan menghasilkan hasil (*search result*) yang berpresisi rendah (Zamir, 1999; Tombros,2002).

Menurut Rijbergen (1979), *clustering* dokumen telah lama diterapkan untuk meningkatkan efekifitas temu kembali informasi. Penerapan *clustering* ini bersandar pada suatu hipotesis (*cluster-hypothesis*) bahwa dokumen yang relevan akan cenderung berada pada cluster yang sama jika pada koleksi dokumen dilakukan *clustering*. Beberapa penelitian untuk dokumen berbahasa inggris menerapkan *clustering* dokumen untuk memperbaiki kinerja dalam proses *searching* (Voorhess,1986; Tombros,2002). Sedangkan perbaikan dalam penyajian hasil *search* dilakukan oleh antara lain Cutting et.al.(1992), Zamir(1999), Osinki (2004) dan Widyantoro (2007). Untuk dokumen berbahasa Indonesia penelitian bidang IR adalah oleh Vega (2001) dan Tala (2004) yang meneliti efek *stemming* pada hasil pencarian. Penelitian penerapan *clustering* untuk perbaikan kinerja perolehan informasi untuk dokumen berbahasa Indonesia belum pernah dilakukan. Hal ini mengingat secara umum penelitian tentang komputasi bahasa untuk dokumen Bahasa Indonesia juga masih sangat minim (Nazief, 2000), bahkan *tes-bed* yang dapat digunakan secara standar untuk penelitian IR belum ada (Asian, 2004). Dengan latar belakang tersebut penelitian ini mencoba menyelidiki alternative pencarian berbasis *cluster* untuk dokumen berbahasa Indonesia.

Permasalahan dalam penelitian ini adalah bagaimana merancang sebuah sistem untuk menyimpan dan menemukan informasi teks dengan pendekatan berbasis kluster dan menguji apakah pendekatan ini lebih unggul dibandingkan dengan pencarian berbasis *word-matching*.

Penelitian ini memiliki batasan model yaitu model ruang vektor dengan uji coba sistem berupa dokumen teks berita berbahasa Indonesia.

Dari penelitian ini diharapkan dapat dirancang suatu sistem temu kembali informasi yang memiliki kinerja yang lebih baik dibandingkan dengan pendekatan berbasis kata (*word-matching*) didalam menangani volume data teks yang semakin membesar.

Model Ruang Vektor Untuk Koleksi Dokumen

Model ruang vektor untuk koleksi dokumen mengandaikan dokumen sebagai sebuah vektor dalam ruang kata (*feature*). Klustering dokumen dipandang sebagai pengelompokan vektor berdasarkan suatu fungsi *similarity* antar dua vektor tersebut. Jika koleksi n buah dokumen dapat diindeks oleh t buah *term*/*feature* maka suatu dokumen dapat dipandang sebagai vektor berdimensi t dalam ruang term tersebut. Dengan demikian koleksi dokumen dapat dituliskan sebagai matrik kata-dokumen X, yang dapat ditulis :

$$X = \{x_{ij}\}\ i= 1,2,..t\ ;\ j =1,2,..\ n \qquad (1)$$

x_{ij} adalah bobot term i dalam dokumen ke j

Menurut Luhn (1958), kekuatan pembeda terkait dengan frekuensi term (*term-frequency*, *tf*). *Term* yang memiliki kekuatan diskriminasi adalah *term* dengan frekuensi sedang. Pemotongan term dengan frekuensi tinggi dilakukan dengan membuang *stop-word*, seperti ‘ini’,’itu’,’yang’, ‘yaitu’ dan lain-lain yang dapat mengurangi frekuensi *feature* 30 sampai 40 persen (Steinbach et.al.,2000; Hamzah, 2006).

Pembobotan dasar dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen karena dipercaya bahwa frekuensi kemunculan *term* merupakan petunjuk sejauh mana *term* tersebut mewakili isi dokumen. Menurut Luhn (1958), kekuatan pembeda terkait dengan frekuensi term (*term-frequency*, *tf*), di mana *term* yang memiliki kekuatan diskriminasi adalah *term* dengan frekuensi sedang. Pembobotan baku yang digunakan adalah *term-frequency invers-document freqeuency* (TF-IDF) (Chisholm and Kolda,1999) sebagai berikut :

$$x_{ij}= tf_i * \log(n/df_i)\ ;\ i=1,2,\ldots,t\ ;\ j=1,2,\ldots,n \qquad (2)$$

dengan t=total term dalam index, n=total dokumen dalam koleksi, df_i=total dokumen yang mengandung term ke-i.

Dalam proses *clustering*, kesamaan antara dokumen D_i dengan dokumen D_j umumnya diukur dengan fungsi similaritas tertentu. Menurut Chisholm and Kolda (1999) untuk tujuan clustering dokumen fungsi yang baik adalah fungsi similaritas Cosine, berikut :

$$\textit{Cosine}\text{-sim}(D_i,D_j)= \frac{\sum_{k=1}^{t} D_{ik} D_{jk}}{\sqrt{\sum_{k=1}^{t} (D_{ik})^2 \sum_{k=1}^{t} (D_{jk})^2}} \qquad (3)$$

Jika vektor Di dan Dj masing-masing ternormalisasi sehingga masing-masing panjangnya satu, maka fungsi *cosine* menjadi :

$$\textit{Cosine}\text{-sim}(D_i,D_j) = \sum_{k=1}^{t} D_{ik} D_{jk} \qquad (4)$$

Dalam Pemrosesan query, similaritas antara query Q dengan dokumen Di juga dapat digunakan formula pada persamaan (4), yaitu :

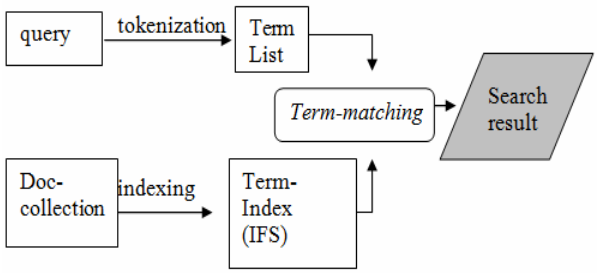
$$\textit{Cosine}\text{-sim}(Q,D_i) = \sum_{k=1}^{t} Q_k D_{ik} \qquad (5)$$

Strategi Pencarian *Query*

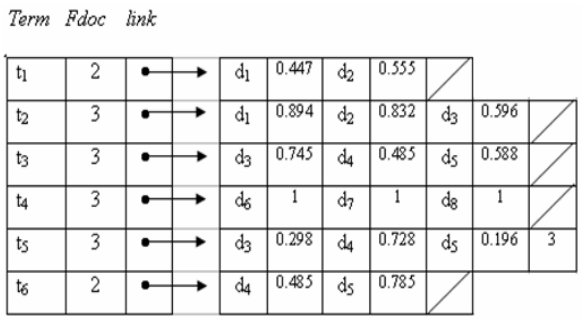
Ada berbagai strategi pencarian (*search strategies*) dalam IR antara lain : *boolean search*, *inverted file search*, *probabilistic search, extended boolean search* (Frakes and Baeza-Yates,1992). Dari model-model *search* tersebut yang banyak digunakan adalah *inverted files search* (IFS) karena alasan efisiensi.

Pencarian Linear model IFS

Sekema IR model IFS dapat dilihat seperti pada Gambar 1. Dalam *indexing* model IFS term terindex akan menunjuk pada *list* yang memuat daftar dokumen yang mengandung *term* tersebut (Gambar 2), sehingga jika suatu *query* diberikan maka dengan cepat akan diberikan jawaban daftar dokumen yang memuat *term* tersebut.



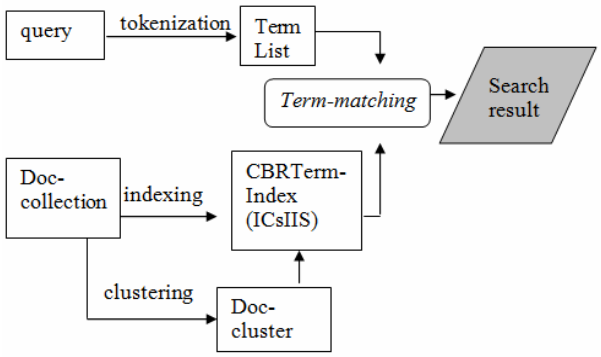
Gambar 1. Pencarian Query berbasis kata dengan IFS



Gambar 2. Struktur Data Pada Pencarian Query model IFS

Pencarian berbasis kluster

Pada pencarian berbasis kluster dokumen yang telah dikluster diindeks berdasarkan term IFS dan indeks kluster (Gambar 3). Jika suatu query diberikan maka similaritas query dengan pusat kluster dihitung, selanjutnya kluster yang pusat klusternya paling dekat dengan query ditampilkan sebagai jawaban.

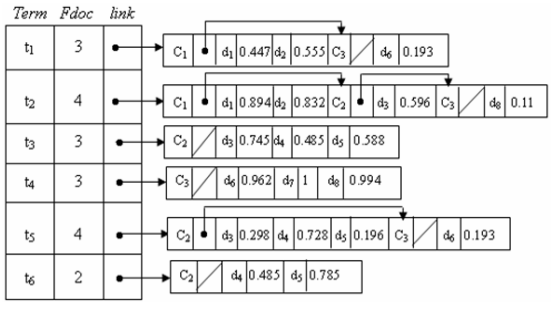


Gambar 3. Pencarian Query berbasis Kluster

Sebagai ilustrasi jika dimiliki koleksi 8 dokumen yang terkluster menjadi 3 kluster (Gambar 4). Struktur data yang dirancang untuk implementasi disajikan seperti pada Gambar 5. Struktur ini terdiri dari *inverted-index* untuk *centroid vector* (IC) dan *CBR implementation using skips* (ICsIIS). Dengan struktur ini pencarian query dengan model kluster akan dapat dilakukan dengan cepat (Can et.al.,2004).

Term	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
t_1	0.447	0.555	0	0	0	0.193	0	0
t_2	0.894	0.832	0.596	0	0	0	0	0.11
t_3	0	0	0.745	0.485	0.588	0	0	0
t_4	0	0	0	0	0	0.962	1	0.994
t_5	0	0	0.298	0.728	0.196	0.193	0	0
t_6	0	0	0	0.485	0.785	0	0	0
Cluster	C_1	C_1	C_2	C_2	C_2	C_3	C_3	C_3

Gambar 4. Contoh Indexing term dokumen terkluster



Gambar 5. Contoh Struktur Data dokumen terkluster

***Clustering* Dokumen**

Clustering didefinisikan sebagai upaya pengelompokan data ke dalam kluster sehingga data-data didalam kluster yang sama memiliki lebih kesamaan dibandingkan dengan data-data pada kluster yang berbeda (Jain,1988). Dikenal dua pendekatan, yaitu *herarchical* dan

partitional dengan masing-masing memiliki banyak variasi.

Metode *Hierarchi Agglomerative* untuk *Clustering* dokumen

Metode klustering secara *aglomerative* berawal dari n= cacah dokumen sebagai cluster. Dengan menggunakan fungsi similaritas antar kluster kemudian proses penggabungan kluster terdekat dilakukan. Ukuran similaritas antar kluster antara lain, misalnya: *UPGMA,*CST, *Single Link, Complete Link* (Jain,1988). Berikut ini ringkasan masing-masing teknik tersebut:

- *Unweighted Pair Group Method Average similarity* (UPGMA): Similaritas dua kluster diukur dengan rata-rata hitung similaritas antar seluruh pasangan titik antara kedua kluster.
- *Centorid- Similarity Technique(CST) :* Jarak antar kluster ditentukan dengan jarak antar pusat kluster.
- *Single Link* (SL) : jarak terbaik dua kluster diwakili oleh jarak terdekat (similaritas tertinggi) dari dua titik dari dua kluster.
- *Complete Link(CL)* : jarak terbaik dua kluster diwakili oleh jarak terjauh (similaritas terendah) dari dua titik dari dua kluster.

Pendekatan hierarchical memiliki kompleksitas waktu dan ruang $O(N^2)$.

K-Means Clustering

Algoritma K-*means clustering* merupakan algortima iteratif dengan meminimalkan jumlah kuadrat *error* antara vektor objek dengan pusat kluster terdekatnya (Jain,1988), yaitu :

$$\sum_{j=1}^{k} \sum_{x \in \pi_j} \left\| x - m_j \right\|^2 \qquad (6)$$

di mana $\mathbf{m}_j$ adalah pusat kluster (*mean vector*) dalam kluster ke j. Proses dimulai dengan mula-mula memilih secara random k buah dokumen sebagai pusat kluster awal.

Bisecting K-Means Clustering

Metode *Bisecting K-means* (Steinbach, et.al.,2000) mencoba menggabungkan pendekatan *partitional* dengan *divisive hierarchi*, yaitu mula-mula seluruh dokumen dibagi dua dengan cara *K-means* (*bisecting-step*). Selanjutnya cara itu dikenakan pada tiap-tiap kluster sampai diperoleh K buah kluster.

Buckshot Clustering

Algoritma *Buckshot* menggunakan pendekatan *hierarchie agglomerative* untuk mendapatkan k buah vektor sebagai pusat kluster awal. Langkah *Buckshot* mula-mula mengambil sampel acak sebesar $\sqrt{kn}$ dokumen, dikluster dengan prosedur *hierarchie agglomerative* untuk mendapatkan k buah kluster. Selanjutnya dari partisi awal *Buckshot* proses *refinement* dilakukan sebagaimana dalam K-*means clustering*

Evaluasi *Retrieval*

Evaluasi suatu model *retrieval* oleh suatu sistem IR yang paling umum adalah ukuran *Recall* dan *Precision* (Rijsbergen,1979). *Recall* didefinisikan sebagai rasio cacah dokumen relevan terpanggil dengan cacah total dokumen terpanggil, sedangkan *Recall* didefinisikan sebagai rasio antara cacah dokumen relevan terpanggil dengan total cacah dokumen relevan dalam koleksi. Parameter tunggal ukuran keberhasilan retrieval yang menggabungkan *Recall* dan *Precision* adalah parameter F-*measure* (Rijsbergen,1979) :

$$\text{F-}measure = \frac{(\beta^2 + 1)PR}{\beta^2 P + R} \qquad (7)$$

dengan β parameter kepentingan relative aspek Precision dan Recall. Jika *Recall* (R) dan *Precision* (P) memiliki bobot yang sama penting, β =1, maka parameter F-measure menjadi :

$$\text{F-}measure = \frac{2PR}{P + R} \qquad (8)$$

METODOLOGI

Bahan penelitian ini berupa koleksi dokumen teks berbahasa Indonesia, yang terdiri dari dua buah koleksi berita dn koleksi abstract , yaitu seperti tersaji dalam Tabel 1 berikut :

Tabel 1. Koleksi-koleksi dokumen untuk Tes Retrieval

Koleksi	∑doc	∑term	∑term index	∑cluster	∑Query
News500	500	11.637	3.994	13	5
News1009	1009	18.255	5.233	21	10
Abstract	302	5.110	1.119	17	10

Adapun daftar Query untuk masing-masing koleksi adalah seperti pada Tabel 2, Tabel 3 dan Tabel 4.

Tabel 2. Daftar Query untuk Koleksi News500

No	Query	∑doc Rel
1	Pemberangkatan jamaah haji	38
2	Pertandingan Piala dunia	183
3	Pasar uang dolar	67
4	Penumpasan Gam aceh	61
5	Kerusuhan ambon maluku	51

Tabel 3. Daftar Query untuk Koleksi News1009

No	Query	∑doc Rel
1	Pemberangkatan haji	38
2	Pemberangkatan jamaah haji	38
3	Piala dunia	183
4	Pertandingan Piala dunia	183
5	Pasar uang dolar	67
6	Perkembangan Pasar uang dolar	67
7	Penumpasan Gam aceh	61
8	Kerusuhan ambon maluku	51
9	Kunjungan megawati ke laur negeri	36
10	Penyelesaian kasus tommy suharto	67

Tabel 4. Daftar Query untuk Koleksi Abstract

No	Query	∑doc Rel
1	Aplikasi logika fuzzy	16
2	Sistem informasi	40
3	Jaringan syaraf tiruan	14
4	Pengolahan citra	9
5	Algoritma genetika	17
6	Database	14
7	Sistem pendukung keputusan	11
8	GPS GPRS komunikasi data	25
9	Rekayasa perangkat lunak	23
10	Keamanan system informasi	10

Proses *pre-processing* berupa ekstrak kata, penyusunan indeks dan struktur IFS maupun struktur ICsIIS dilakukan dengan kode program JAVA (jdk1.4.2).

Hasil pengujian kinerja *feature* kata dan frasa diukur melalui nilai F-*measure* yang membandingkan *feature* kata saja, frasa saja dan *feature* campuran. Uji statistik hasil dengan uji t *wilcoxon sign-rank* untuk pengamatan berpasangan.

PEMBAHASAN

Hasil pengujian untuk koleksi pertama News500 ketika diberikan query seperti yang ada dalam daftar memberikan hasil bahwa pemanggilan berbasis kluster (CBR) menghasilkan nilai F-*measure* yang lebih tinggi dibandingkan dengan pemanggilan linear (IFS). Hal ini berlaku untuk semua query. Tabel berikut adalah Rata-rata F-*measure* untuk pemanggilan query berbasis kluster dan pemanggilan linear untuk koleksi News500. Rata-rata diambil untuk retrieval berbasis kluster pada setiap model clustering, baik hierarchical maupun partitional. Hasil uji statistic menunjukkan bahwa perbedaan rata-rata adalah signifikan.

Tabel 5. Rata-rata F-*measure* untuk koleksi News500

Query	F-measure IFS	F-measure CBR
1	0.5685	**0.6574**
2	0.4534	**0.5643**
3	0.5643	**0.7654**
4	0.6343	**0.8875**
5	0.7543	**0.7845**

Ket : cetak bold lebih tinggi

Pengaruh algoritma clustering pada hasil pemanggilan berbasis kluster dapat diberi contoh seperti Tabel 6. berikut, untuk suatu query :"pertandingan piala dunia".

Tabel 6. Pengaruh algoritma *Clustering* pada *Retrieval*

Metode Clustering		IFS		CBR Search	
	Doc Retriev	Rel Doc Retriev	F-measure	RelDoc Retriev	F-measure
UPGMA	75	62	0,7848	75	0,9494
ClusCtr	79	64	0,7901	79	0,9753
CompLink	104	73	0,7807	83	0,8877
K-Mean	85	66	0,7857	66	0,7857
Bsc-KMean	60	52	0,7273	56	0,7832
Buckshot	83	66	0,7952	83	1,0000

Ket : cetak tebal nilainya lebih tinggi

Dari Tabel 6 terlihat bahwa pada algoritma *hierarchical* kinerja *clustering* lebih baik dalam memberikan nilai F-*measure* daripada algoritma partitional K-means dan Bisecting K-mean, tetapi kinerja masih dibawah algoritma *buckshot*.

Pada koleksi News1009 dan koleksi Abstract pengujian query diambil untuk metode kluster yang relatif cepat dan dengan kompleksitas komputasi lineer, yaitu buckshot. Tabel 7 menyajikan hasil pengujian untuk seluruh Query dari koleksi News1009 untuk jumlah retrieval pada IFS tidak dibatasi.

Tabel 7. Hasil retrieval untuk koleksi News1009 dengan Retrieval IFS tidak dibatasi

No	Query	IFS	F-CBR
1	Pemberangkatan haji	0,7037	**0,9189**
2	Pemberangkatan jamaah haji	0,6667	**0,8095**
3	Piala dunia	0,7154	**0,9777**
4	Pertandingan Piala dunia	0,7059	**0,9862**
5	Pasar uang dolar	0,5038	**0,9778**
6	Perkembangan Pasar uang dolar	0,4258	**0,9635**
7	Penumpasan Gam aceh	0,8414	**0,9677**
8	Kerusuhan ambon maluku	0,7500	**0,8224**
9	Kunjungan megawati ke laur negeri	0,1967	**0,5410**
10	Penyelesaian kasus tommy suharto	0,5654	**0,8049**

Jika retrieval IFS dibatasi sejumlah dokumen sesuai dengan jumlah dokumen yang dikembalikan oleh CBR maka hasil retrieval adalah seperti table 8 berikut . Terlihat beberapa query CBR bernilai sama dengan IFS, dan ada satu query yang IFSnya lebih tinggi dari CBR.

Tabel 8. Hasil retrieval untuk koleksi News1009 dengan Retrieval IFS dibatasi sebanyak CBR

No	Query	IFS	F-CBR
1	Pemberangkatan haji	0,7568	0,9189
2	Pemberangkatan jamaah haji	0,7619	0,8095
3	Piala dunia	0,8603	0,9777
4	Pertandingan Piala dunia	0,8595	0,9862
5	Pasar uang dolar	0,9778	0,9778
6	Perkembangan Pasar uang dolar	0,9635	0,9635
7	Penumpasan Gam aceh	0,9677	0,9677
8	Kerusuhan ambon maluku	0,8411	0,8224
9	Kunjungan megawati ke laur negeri	0,4754	0,5410
10	Penyelesaian kasus tommy suharto	0,6951	0,8049

Untuk koleksi Abstract pemanggilan IFS yang tidak dibatasi dan IFS yang dibatasi sebanyak dokumen dari CBR hasilnya berturut-turut adalah tersaji Tabel9 dan Tabel 10.

Tabel 9. Hasil F-measure untuk IFS dan CBR untuk koleksi Abstract dengan jumlah dokumen IFS tidak dibatasi

No	**Query**	**IFS**	**F-CBR**
1	Aplikasi logika fuzy	0,2388	**0,5143**
2	Sistem informasi	0,2989	**0,4516**
3	Jaringan syaraf tiruan	0,4286	**0,7200**
4	Pengolahan citra	0,5294	**0,6957**
5	Algoritma genetika	0,4063	**0,4667**
6	Database	0,2029	**0,6207**
7	Sistem pendukung keputusan	0,1106	**0,2778**
8	GPS GPRS komunikasi data	0,1795	**0,2162**
9	Rekayasa perangkat lunak	0,3297	**0,4324**
10	Keamanan system informasi	0,0858	**0,3077**

Tabel 10. Hasil F-measure untuk IFS dan CBR untuk koleksi Abstract dengan jumlah dokumen IFS dibatasi dengan jumlah dokumen CBR

No	**Query**	**IFS**	**F-CBR**
1	Aplikasi logika fuzy	**0,6857**	0,5143
2	Sistem informasi	0,3871	**0,4516**
3	Jaringan syaraf tiruan	0,7200	0,7200
4	Pengolahan citra	0,6957	0,6957
5	Algoritma genetika	0,4000	**0,4667**
6	Database	0,6207	**0,6207**
7	Sistem pendukung keputusan	**0,3889**	0,2778
8	GPS GPRS komunikasi data	0,2162	0,2162
9	Rekayasa perangkat lunak	**0,5405**	0,4324
10	Keamanan system informasi	**0,3590**	0,3077

Dari Tabel 10 terlihat bahwa untuk koleksi abstract jika jumlah dokumen yang di*retrieve* oleh IFS dibatasi sama dengan jumlah dokumen yang diretreiev oleh CBR, maka kinerja retrieval CBR akan menurun dan beberapa queryu IFS mengahsilkan retrieval yang lebih baik daripada CBR.

KESIMPULAN

Beberapa kesimpulan yang dapat diambil dari penelitian ini adalah :

- Pemrosesan query dengan pendekatan berbasis kluster (*cluster-based retrieval*) terbukti mampu secara signifikan meningkatkan kinerja sistem IR jika dibandingkan dengan pemrosesan linear model IFS.
- Kinerja pemrosesan query berbasis kluster dipengaruhi oleh model koleksi dokumen. Pada koleksi dokumen berbahasa Indonesia untuk jenis dokumen ilmiah seperti kumpulan abstrak dari makalah ilmiah yang umumnya banyak mengandung kosa kata bahasa inggris kinerja *retrieval* berbasis kluster cenderung menurun. Tetapi pada dokumen berita kinerja retrieval berbasis kluster terlihat sangat baik.
- Hasil *clustering* dengan *hierarchical* menunjukkan kinerja yang lebih baik dari *partitional*. Meskipun demikian algoritma partitional tetap memberikan kinerja retrieval berbasis kluster yag lebih baik daripada retrieval dengan model IFS.
- Masih diperlukan pengujian dengan berbagai jenis koleksi dokumen berbahasa Indonesia, seperti makalah penuh, atau jenis tulisan yang lain.

Pustaka

Asian, J., H. E. Williams, and S. M. M. Tahaghoghi, T*esbed for Indonesian Text Retrieval*, 9th Australian Document Computing Symposiom, Melbourne December, 13, 2004

Can,F., I.S. Altingode, E.Damir,2004, Efficiency and Effectiveness of Query Processing in Cluster-Based Retrieval, Information System, 29(2004),697-719.

Chisholm, E. and T. G. Kolda, *New Term Weighting Formula for the Vector Space Method in Information Retrieval*, Research Report, Computer Science and Mathematics Division, Oak Ridge National Library, Oak Ridge, TN 3781-6367, March 1999.

Cutting, D. R., D. R. Karger, J. O. Pederson, and J. W. Tukey,1992, *Scatter/Gather:A Cluster-based Approach to Browsing Large Document Collection*, Procedding 15th Annual Int 7ACM SIGIR Conference on R&D in IR, 1992.

Frakes,W.B. and Baeza-Yates,R.,1992, *Information Retrieval,Data Structure and Algorithm*, Prentice Hall, Englewood Clifs, New Jersey.

Jain, A.K. and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, 1988

Hamzah, A., F. Soesianto, A.Susanto, J.E.,Istyanto : *Seleksi Feature Kata Berdasarkan Variansi Kemunculan Kata Dalam Peningkatan Unjuk Kerja Document Clustering Untuk Dokumen Berbahasa Indonesia,* Pakar, Jurnal Teknologi Informasi dan Bisnis , Vol.7,No.3. , pp. 181-190, 2006.

Luhn, H.P., *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development, 2:159-165 , 1958

Nazief, B., *Development of Computational Linguistic Research: a Challenge for Indonesia*", Computer Science Center, University of Indonesia ,2000

Osinki, S. , 2004, *Dimensionality Reduction Techniques for Search Engine Results Clustering*, Master Thesis, University of Sheffield, UK.

Rijsbergen, C. J., *Information Retrieval*, Information Retrieval Group, University of Glasgow , UK ,1979

Steinbach, M., Karypis, G., Kumar, V., *A Comparison of Document Clustering Techniques,* University of Minnesota, Technical Report #00-034, at http://www.cs.umn. edu/tech_reports, 2000

Tala, F. Z., 2004, *A Study of Stemming Effect on Information Retrieval in Bahasa Indonesia*, Master Thesis, Universiteit van Amsterdam, The Netherlands

Tombros, A., 2002, *The Effectiveness of Query-Based Hierarchic Clustering of Documents for Information Retrieval*, PhD Thesis, Univerity of Glasgow

Vega, V. B. , 2001, *Information Retrieval for the Indonesian Language*, Master's thesis, National University of Singapore.

Voorhees,E.M., 1986, *Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval*. Information Processing & Management, 22:465-76.

Widyantoro,D.H.,2007,*Toward the Development of The Next Generation Search Engine*, Proceeding of The International Conference on Electrical Engineering and Informatics, ICEEI2007, Bandung 17-19 Juni 2007.

www.google.com

Zamir, O.E., *Clustering Web Document : A Phrase-Based Method for Grouping Search Engine Result*, PhD. Dissertation, University of Washington,1999