

## **ANALISIS GEROMBOL UNTUK PENYUSUNAN TESAURUS**

*Dwi Prasetyo*

*Staf Pengajar Universitas Nusa Cendana, Kupang*

### **I. PENDAHULUAN**

Sistem temu-kembali informasi umumnya dilakukan terhadap sekumpulan berkas yang berisi record-record dengan struktur yang homogen dimana setiap berkas mempunyai atribut-atribut sebagai karakteristiknya (Adisantoso, 1997). Untuk informasi berbasis teks, kegiatan temu-kembali informasi hampir sama dengan yang dilakukan terhadap record-record dalam basis data. Atribut yang digunakan sebagai karakteristik basis data ini antara lain adalah pengarang, penerbit, tahun terbit, dan lain-lain. Jika record berisi teks seperti judul dokumen, abstrak atau ringkasan dari suatu dokumen, indentifikasi yang digunakan adalah kunci atau kata indeks dari suatu dokumen. Hal ini dilakukan karena temu-kembali terhadap dokumen ini tidak hanya mencari kata-kata yang terdapat di dalam dokumen, melainkan mencari istilah-istilah lain yang relevan dengan yang diinginkan.

Pada saat jumlah kata indeks dalam suatu dokumen semakin besar, proses temu-kembali yang dilakukan akan membutuhkan waktu yang lama dan diperlukan banyak tempat untuk menyimpan seluruh kata indeks di dalam media penyimpanan komputer. Oleh karena itu diperlukan proses reduksi kata-kata indeks sehingga jumlahnya lebih sedikit tetapi tetap mencirikan dokumen-dokumen yang ada. Dengan kata lain, perlu disusun kelompok-kelompok istilah yang homogen

dan mempunyai hubungan yang sangat erat. Kelompok-kelompok yang terbentuk ini disebut dengan *tesaurus* (salton, 1989).

Penyusunan tesaurus untuk temu-kembali informasi dapat dilakukan secara manual, semi otomatis, atau otomatis sepenuhnya (Magdalena, 1996). cara manual melibatkan tenaga yang benar-benar ahli di bidang subyek dokumen sehingga sangat sulit dan lama. Dengan perkembangan teknologi komputer, penyusunan tesaurus mulai dilakukan -dengan menggunakan bantuan komputer untuk sebagian atau bahkan seluruh proses penyusunan proses tesaurus. Karena banyak keraguan terhadap kemungkinan penyusunan tesaurus secara otomatis, maka sampai saat ini banyak dilakukan penelitian mengenai hal tersebut (Srinivasan, 1992 dalam Magdalena, 1996).

Penelitian tentang hal ini banyak dilakukan oleh para ahli untuk teks berbahasa Inggris, sedangkan penelitian untuk teks yang berbahasa Indonesia dilakukan oleh Magdalena (1996) yang meneliti dokumen Lembaga Penelitian Universitas Indonesia dan Fakultas Ilmu Komputer Universitas Indonesia, Adisantoso (1997) meneliti dokumen Hasil Penelitian Lingkungan Hidup Bapedal Jakarta, dan beberapa penelitian terkait lainnya.

Untuk lebih melengkapi penelitian otomatisasi penyusunan

tesaurus, dilakukan penelaahan terhadap metode penyusunan tesaurus untuk teks berbahasa Indonesia dengan domain subyek yang berbeda.

## II. BAHAN DAN METODE

### 2.1. Bahan

Dokumen yang digunakan sebagai masukan proses adalah 119 dokumen yang berisi judul dan abstraksi ringkasan karya ilmiah mahasiswa S 1 FMIPA-IPB, dengan ukuran rata-rata setiap dokumen adalah 2167,06 bytes. Kumpulan dokumen ini tersimpan dalam format ASCII.

Disamping kumpulan dokumen sebagai masukan, proses juga memerlukan berkas yang berisi kata-kata buangan, yaitu kata tugas, kata kerja, kata punya, kata depan, kata penghubung, dan kata-kata lain yang tidak akan digunakan untuk istilah indeks. Kata-kata buangan selain kata tugas dan kata kerja akan ditambahkan setelah diperoleh daftar kata-kata indeks sesuai dengan frekuensi kemunculannya dan makna dari kata tersebut.

### 2.2. Metode

Metode yang digunakan dalam penelitian ini adalah :

- Pemilihan istilah dari kata-kata yang ada dalam dokumen, dilakukan berdasarkan kriteria tertentu, jumlah kata dalam satu istilah adalah satu (single term),
- Pembentukan Matriks Istilah-Dokumen yang berisi frekuensi istilah pada setiap dokumen,
- Pengelompokan istilah secara non hirarki dengan teknis analisis gerombol menggunakan paket program SPSS.

## III. PERANCANGAN SISTEM

### 3.1. Bahasa Pemrograman

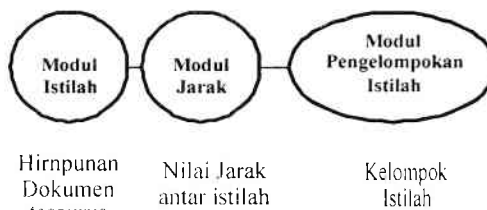
Pemrograman dilakukan dengan menggunakan basis kompilator Clipper karena relatif mudah, mempunyai banyak fungsi baku, dan kecepatan proses yang tidak jauh berbeda dengan bahasa pemrograman tingkat tinggi lainnya seperti Pascal, dan sejenisnya.

### 3.2. Modul

Modul yang dikembangkan dalam sistem penyusunan tesaurus adalah :

1. Modul Istilah
2. Pembentukan Jarak
3. Modul Pengelompokan IsWah (SPSS)

Modul (1) akan menghasilkan istilah dan kumpulan dokumen, setelah itu dengan rumus jarak didapatkan jarak antar istilah dengan menggunakan Model (2), setelah itu dengan cara statistik dikelompokkan dengan menggunakan Model (3). Hubungan antar ketiga Modul tersebut dapat dilihat pada gambar di bawah ini :



Modul Istilah, bertujuan untuk membentuk MDI dan MID yang mempunyai empat proses antara lain :

- Pengambilan kata dari dokumen, pembuangan kata yang tidak digunakan sebagai istilah indeks,

- Pembuangan kata kerja berimbuhan dan kata punya
- Penyusunan MDI dan MID

Setiap dokumen dalam himpunan dokumen masukan yang memiliki istilah indeks dapat dinyatakan dengan Matrik Dokumen-Istilah (MDI), dimana setiap baris dalam matrik tersebut menyimpan satu dokumen dari himpunan dokumen dan istilah-istilah yang dimilikinya, seperti dapat dilihat pada gambar di bawah ini :

	Istilah			
Doc.	I <sub>1</sub>	I <sub>2</sub>	I <sub>3</sub>	I <sub>n</sub>
D <sub>1</sub>	f <sub>11</sub>	f <sub>12</sub>	f <sub>13</sub>	f <sub>1n</sub>
D <sub>2</sub>	f <sub>21</sub>	f <sub>22</sub>	f <sub>23</sub>	f <sub>2n</sub>
D <sub>3</sub>	f <sub>31</sub>	f <sub>32</sub>	f <sub>33</sub>	f <sub>3n</sub>
D <sub>m</sub>	f <sub>m1</sub>	f <sub>m2</sub>	f <sub>m3</sub>	f <sub>mn</sub>

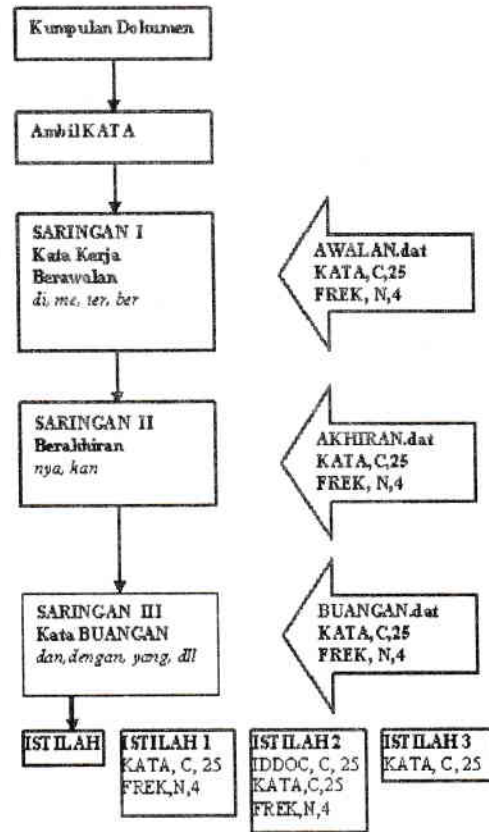
Karena yang dilakukan dalam proses tersebut adalah pengelompokan istilah, bukan dokumen. Dalam setiap baris dari matrik untuk pengelompokan istilah seharusnya disimpan satu istilah dengan dokumen-dokumen yang memiliki istilah tersebut. Matrik -tersebut disebut Matrik Istilah.

Dokumen (MID) seperti terlihat pada gambar di bawah ini :

	Doc.			
Istilah	D <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	D <sub>m</sub>
I <sub>1</sub>	f <sub>11</sub>	f <sub>12</sub>	f <sub>13</sub>	f <sub>1m</sub>
I <sub>2</sub>	f <sub>21</sub>	f <sub>22</sub>	f <sub>23</sub>	f <sub>2m</sub>
I <sub>3</sub>	f <sub>31</sub>	f <sub>32</sub>	f <sub>33</sub>	f <sub>3m</sub>
I <sub>n</sub>	f <sub>n1</sub>	f <sub>n2</sub>	f <sub>n3</sub>	f <sub>nm</sub>

Matrik Istilah-Dokumen

Sedangkan metodologi dari tahap-tahap pemrosesan teks digambarkan sebagai berikut :



Arsip yang berhubungan dengan modul ini terdiri-dari lima arsip yaitu :

1. Berkas AWALAN.dat yang berisi kata kerja berawalan yang tidak digunakan sebagai istilah indeks, seperti : "di", "me", "ter", "ber" dan sejenisnya. Format data untuk setiap kata buangan ini adalah :

KATA, karakter (25) (kata buangan)  
 FREK. Numerik(4) (Frekwensi kata yang dihitung secara otomatis oleh Program)

2. Berkas AKHIRAN.dat yang berisi kata kerja berakhiran yang tidak digunakan sebagai istilah indeks, seperti: "nya", "kan", "lah", "kah", "kah", "pun", "ku", "mu" dan sejenisnya. Format data untuk setiap kata buangan ini adalah:

KA TA. karakter (25) *{kata buangan}*  
 FREK. Numerik(4) *{Frekwensi kata yang dihitung secara otomatis oleh Program}*

3. Berkas BUANGAN.dat yang berisi kata buangan yang tidak digunakan sebagai istilah indeks, seperti: "dan", "dengan", "yang" dan sejenisnya. Format data untuk setiap kata buangan ini adalah :

KATA. karakter (25) *{kata buangan}*  
 FREK. numerik(4) *{Frekwensi kata yang dihitung secara otomatis oleh Program}*

4. Berkas ISTILAH1.dat yang berisi istilahistilah dari kumpulan dokumen yang digunakan sebagai istilah indeks. Format data untuk setiap kata buangan ini adalah :

KATA. karakter (25) *{kata buangan}*  
 FREK. numerik(4) *{Frekwensi kata yang dihitung secara otomatis oleh Program}*

Berkas ISTILAH2.dat yang berisi istilah istilah dari masing-masing dokumen yang digunakan sebagai istilah indeks. Format data untuk setiap kata buangan ini adalah :

IDDOC. karakter (25) *{ id dokumen yang membedakan dengan dokumen yang lain}*

KATA karakter (25) *{ istilah }*  
 FREK. Numerik(4) *{ frekwensi istilah yang dihitung secara otomatis oleh program}*

**algoritma Submodul ISTILAH 1 :**

Untuk setiap kata yang terdapat dalam suatu **dokumen**, lakukan,

Ubah semua karakter dalam kata menjadi huruf kecil  
 Periksa apakah kata terdapat dalam p A WALAN.dat  
 Jika ya, lakukan :  
 Tambahkan 1 ke dalam FREK sesuai dalam arsip AWALAN.dat

Periksa apakah kata terdapat dalam arsip AKHIRAN.dat  
 Jika ya, lakukan :  
 Tambahkan 1 ke dalam FREK sesuai dalam arsip AKHIRAN.dbf

Periksa apakah kata terdapat dalam arsip BUANGAN.dat  
 Jika ya, lakukan :  
 Tambahkan 1 ke dalam FREK sesuai dalam arsip BUANGAN.dat  
 Jika tidak, kirim data dan id dokumen ke modul MDI dan, kemudian ke modul MID

**algoritma Submodul ISTILAH2 :**

Periksa apakah kata dan id dokumen sudah ada dalam arsip ISTILAH2.dat  
 Jika tidak,  
 tambahkan record baru yang berisi kata dan id dokumen ke dalam arsip ISTILAH2.dat dan isikan FREK=1  
 Jika sudah ada,  
 tambahkan 1 ke dalam FREK dari kata dan id dokumen yang sesuai

Periksa apakah kata sudah ada dalam arsip ISTILAH1.dat  
 Jib tidak,  
 tambahkan record baru yang berisi kata ke dalam arsip ISTILAH1.dbf dan isikan FREK=1  
 Jib sudah ada,  
 tambahkan 1 ke dalam FREK.

**Pengelompokan Istilah,**

Bertujuan untuk membentuk kelompok kelompok istilah dengan metode Analisis

Gerombol Non Hirarki, pada Proses pengelompokan istilah ini menggunakan paket program SPSS.

#### IV. HASIL DAN PEMBAHASAN

Bab ini berisi uraian tentang hasil dan pembahasan sistem temu-kembali informasi yang diimplementasikan dengan menggunakan bahasa pemrograman berbasis kompilator Clipper. Pembahasan dimulai dengan menelaah karakteristik dokumen masukan.

##### 4.1. Karakteristik Dokumen

Dokumen yang diproses terdiri dari 119 dokumen yang seluruhnya mempunyai domain subyek yang sama, yaitu pertanian. Deskripsi dokumen tercantum di bawah ini :

Tabel 1

No	Uraian.	Jumlah
1	Banyaknya dokumen	119
2	Jumlah kata	4708
3	Rataan banyaknya kata	395.64
4	Rataan ukuran dokumen	2167.06

##### 4.2. Pemilihan Istilah

Setelah diuji coba untuk memilih suatu istilah, ternyata program yang ada banyak memiliki keterbatasan sebagai berikut:

1. Beberapa kata benda dianggap sebagai kata kerja karena diawali karakter "me", "memper", "di", "ber", "ter"

Contoh: "media", "diploid", "tersier"

2. Penggunaan beberapa istilah yang berbeda pada dokumen masukan untuk menyatakan konsep yang sama belum diatasi.

Contoh: "log-linear" dan "log-Linier" "metoda" dan "metode" "analisis" dan "analysis" "matematik" dan "malematis" "komoditas" dan "komoditi"

3. Kesalahan penulisan istilah dalam dokumen masukan yang terdeteksi dan dibetulkan secara manual
4. Dalam teks yang berdomain kimia sering dijumpai istilah untuk menyatakan unsur-unsur kimia, yang terdiri dari satu atau dua karakter, sehingga batasan pemilihan istilah terhadap kata minimal mempunyai tiga karakter sulit dilakukan secara otomatis.

Contoh: "h", "fe", "nm", "zn"

Jenis dan jumlah kata yang diperoleh selama proses Pemilihan Istilah dari 119 dokumen masukan dapat dilihat pada tabel 1, jumlah kata keseluruhan dari dokumen masukan adalah 47081 kata.

Tabel 2.

No	Jenis Kata	Jumlah
1	Kata yang terpilih menjadi Istilah	1286
2	Kata Buangan ; kata kerja berimbuhan	625
3	Kata Buangan; kata dengan partikel atau kata punya	165

Rendahnya persentase kata yang terpilih terjadi karena banyak kata yang ditambahkan ke dalam daftar kata buangan karena dianggap tidak mampu digunakan sebagai penciri dokumen, dan juga karena kata dengan partikel/kata

punya dan kata kerja berimbuan tidak digunakan sebagai istilah. Lampiran 6 menyajikan daftar kata dengan partikel kata punya, sedangkan lampiran 4 menyajikan daftar kata kerja berimbuan, dan lampiran 7 menyatakan jenis kata buangan dengan frekuensi sangat tinggi di dominasi oleh kata-kata yang tidak penting, seperti kata penghubung, kata keterangan dan lain-lain.

Dan juga didapati bahwa hasil pengelompokkan tidak proposional antara satu dengan yang lainnya.

**4.3. Hasil Pengelompokkan**

Hasil Analisis Gerombol tak berhirarki, dalam penelitian ini terlihat pada pengelompokkan dari sebagian hasil uji coba yang dilakukan pada dokumen masukkan tesis S1 mahasiswa FMIPA Institut Pertanian Bogor.

Tabel.3

Kel	Jum	Istilah
1	1	Erysi
2	2	daerah, periode
3	2	Ekstrak, etanol
4	2	Aku,upt
5	3	Mmat, ilmu, persepsi
6	2	Dekade,wonogiri
7	1	Lahun
8	1	Tamah
9	1	Pola
10	1	Fraksi
11	1124	Terlampir
12	1	Isolat
13	2	Tingkat,status
14	1	Data
15	1	nilai
16	1	Siswa
17	6	Antigen,antiserum,serotipe, Kelmel, babi, daging
18	2	Galur, seleksi
19	7	Iklan, jadwal, televisi, efektivitas,biaya,efisiensi, evaluasi
20	1	Hasil

21	1	Inflasi
22	3	Lentinus, spp, lignosus
23	13	Kecambah, Spesies,kotiledon, hayati,apikal, tomentella, kultur, akar, tunas, induksi, kedelai, biji, kombinasi
24	5	Kabupaten, sungai, indonesia, persen,total
25	3	Hari, waktu, deret
26	4	Abung, bandar,redoks, potensial
27	1	Air
28	1	Kelompok
29	2	Tahun, rata
30	7	Mahasiswa, standar,program, matematika,studi,statistika, kimia
31	4	I,sl,oksidasi,aterosklerosis,genstein
32	12	Sutera,instar,murbei, kanva, alba, eath, ayana,kokon,ulat,bobot,sasag,daun,je nis
33	10	Sentral,ser,call,nomor,proposisi, telepon,sirkit,turangga, panggil,pesawat
34	6	Timpasan,curah,bisnis,sawah, maksimum,industri
35	5	Ppni,limbah,lateks,pekat,platensis
36	2	Spiramisisn, residu
37	1	Hujan
38	1	Usaba
39	1	Wilayah
40	4	Kasting, npk, kebun, pupuk
41	15	Ramah, tangga, dana, investasi, survei, elastisitas, harga, sifat, nonmakanan, konsumsi, kuning, jemih, warna, sektor, bunga
42	2	Hutan, tajak
43	1	Plastik
44	1	Analisis
45	1	Model
46	2	Fosforus, erapan
47	2	Kering, musim
48	3	Varietas, jagung, komponen
49	4	Peubah
50	6	Konstasepsi, anak, jarak, responden, asosiasi, ibu

Dari tabel 3. Dapat dilihat pengelompokkan istilah, jumlah Istilah yang dikelompokkan berjumlah 1286 istilah, tersebar dalam 50 kelompok. Pada kelompok 11 terdapat pengelompokkan istilah yang banyak yaitu sebanyak 1124 istilah, maka untuk kelompok 11 ini

Diproses ebih lanjut lagi, supaya mendapatkan kelompok-kelompok yang lebih kecil. Setelah dilakukan proses maka didapat pengelompokkan seperti terlihat pada Tabel 4. Dimana ke 1124 istilah tersebut dikelompokkan menjadi 20 kelompok, dan masih di lihat ada pengelompokkan pada kelompok 16, yaitu 1006 istilah :

Kel	Jum	Istilah
1	3	Waduk, konservasi, kerus
2	13	Titik, unsur, arah, target, matriks, silang, teknik, kategori, lajur, koresponden, pasur, konsumen, simultan
3	2	Embrio, kalus
4	1	Alternatif
5	5	Dominan, nasional, jangka, tukar, domestik
6	1	Defisit
7	188	Terlampir
8	2	Parameter, gerombol
9	3	Volume, pulau, komoditi
10	1	Ganda
11	4	Sekunder, cenderung, interkasi, buruh
12	14	Fisik, biologis, ekor, telaah, isolasi, Dasar, aspergillus, nisbah, derajat, viskositas, Spektroskopi, zona, hewan, visual
13	5	Vegetasi, lokasi, dampak, massa, Radiasi
14	4	Xanthomonas, campestris, glycines, Profil
15	4	Mutu, macam, spektrum, inisiasi
16	1	Aktivitas
17	5	Kayu, batang, baku, klorofom, pelarut
18	2	Nitrogen, protein
19	3	Cekama, ph, tumbuh
20	3	Reaksi, serum, garis

Pada tahap ini selain dilakukan proses seperti proses sebelumnya ditambahkan pula proses ke-khas-an suatu istilah terhadap dokumen-dokumen yang digunakan, maksudnya adalah istilah yang diproses adalah jika suatu istilah terdapat dalam beberapa naskah, dan didapat hasil pada tabel 5 yaitu kelompok 16,

Tabel 5, kelompok 16

Kel	Jum	Istilah
1	3	Citokidia, inang, lama
2	12	Sistem, manajemen, kredit, tipologi, tindak, prototipe, pakur, bank, informasi, proses, rantai, markov
3	3	Utama, karakteristik, regresi
4	1	Jumlah
5	10	Bakteri, populasi, rumen, jerami, protozoa, domba, ruminan, gas, asam, produksi
6	1	Tinggi
7	6	Kapur, fosfat, ortogonal, busah, taral, Bentuk
8	10	Sedimentasi, penampungan, skenario, tata, sub, intensitas, das, hidrologi, ton, efektif
9	5	Padat, tenaga, modal, ciri, kerja
10	8	Buyi, lintas, ekonomi, desa, unit, kota, sosial
11	5	Optimal, kriteria, prinsip, kontrol, masalah
12	10	Tineke, mawar, segar, atmosfer, modifikasi, kamar, ruang, hybrida, rosa, konsentrasi
13	8	Kalium, kg, faktorial, acuk, dosis, identik, rendah, eacing
14	7	Besar, kecil, induk, kelas, laju, koefisien, hulu
15	5	Uji, tempo, kolesterol, aorta, panjang
16	1006	Terlampir
17	12	Peluang, dpl, iklim, peta, palawija, udara, Andosol, gamma, pudr, tipe, sebaran, utara
18	5	Pektin, labu, stam, kadar, buah
19	5	Lai, surplus, energi, partikel, neraca
20	2	Suhu, lengas

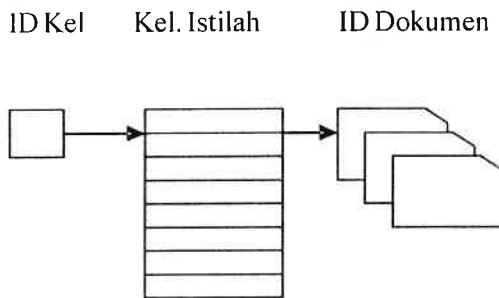
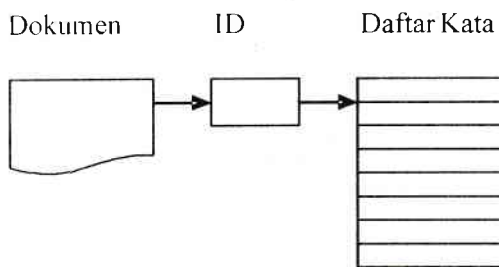
Kelompok istilah yang terbentuk dari proses di atas dapat memiliki istilah-istilah yang mempunyai hubungan yang jelas dan tidak jelas. Hubungan yang jelas, terlihat pada kelompok-kelompok yang mengandung istilan-istilah dalam satu dokumen dan juga istilah-istilah dalam suatu subyek domain misal : subyek kimia saja, biologi saja, dan seterusnya. Contoh kelompok yang mempunyai hubungan yang jelas terlihat pada kelompok 17, yaitu sebagai berikut : "kayu", "batang", "baku", "klorofom", dan "pelarut",

Sedangkan hubungan yang kurang jelas dalam suatu istilah terlihat

pada kelompok 2, yaitu sebagai berikut: "titik", "unsur", "arah", "target", "matriks", "silang", "teknik", "kategori", "lajur", "korespondensi", "pasar", "kosumen" dan "simultan".

#### 4.4. Implementasi Query

Untuk pencarian informasi tentang dokumen dari istilah-istilah yang kita ketahui dapat digambarkan sebagai berikut:



Kalau ingin mencari dokumen untuk topik yang diinginkan, maka Cukup mengetahui kata kunci atau istilah dari dokumen yang ingin dicari. Setelah melalui proses maka akan didapat informasi dokumen dokumen apa yang relevan dengan kata kunci atau istilah tersebut

## V. KESIMPULAN DAN SARAN

### 5.1. Kesimpulan

Karya ilmiah ini bertujuan untuk membuat suatu sistem penyusunan tesaurus dengan menggunakan analisis gerombol non hirarki dari suatu kumpulan dokumen yang dapat dimanfaatkan oleh sistem temu-kembali informasi. Karya ilmiah ini mengkaji kata kata yang berpotensi untuk dipergunakan dalam penyusunan tesaurus dari sekumpulan dokumen berbahasa Indonesia. Dari hasil uji coba dapat disimpulkan :

1. Kata-kata dalam dokumen berbahasa Indonesia yang menjadi masukan memiliki potensi untuk dipergunakan sebagai istilah indeks.
2. Pembuangan kata kerja berimbuhan memiliki potensi untuk meningkatkan efisiensi sistem penyusunan tesaurus dari dokumen masukan dalam bahasa Indonesia.
3. Keterbatasan menggunakan analisis gerombol pada penyusunan tesaurus adalah banyak istilah-istilah yang masuk dalam kelompok tertentu, walaupun tidak memiliki kemiripan dengan istilah lain dalam kelompok tertentu.

### 5.2. Saran

Berdasarkan kesimpulan, kelemahan yang ditemukan, ada beberapa saran untuk pengembangan sistem lebih lanjut :



1. Pengadaan dokumen masukan yang memiliki domain subjek yang spesifik, dan telah disimpan dalam media berkas komputer sehingga dapat langsung digunakan dengan format file yang sama, sehingga dapat langsung digunakan.
2. Pencarian karakteristik yang lebih ketat dan jelas untuk memilih istilah indeks dalam bahasa Indonesia secara otomatis. Hal ini seharusnya dilakukan dengan memperhatikan karakteristik kumpulan dokumen yang digunakan dan istilah-istilah yang biasa digunakan pemakai dalam kueri.
3. Penggunaan daftar kata yang berisi kata-kata yang lebih lengkap dan dapat digunakan untuk menentukan jenis kata, seperti kata dasar, kata benda, kata kerja, atau kata sifat, sehingga istilah yang dihasilkan dapat lebih dikontrol.
4. Penggunaan pendeteksian ejaan dan pembetulan kesalahan ketik pada dokumen masukan secara otomatis.
5. Perlu dilakukan evaluasi subjektif, terhadap pengelompokan istilah dalam tesaurus yang disusun secara otomatis.
6. Pengembangan modul kueri yang menggunakan kelompok-kelompok istilah dalam tesaurus yang disusun secara otomatis. Modul tersebut bertujuan mencari istilah yang sekelompok dengan istilah kueri. Istilah-istilah dalam kelompok tersebut kemudian digunakan sebagai istilah indeks untuk meningkatkan efektivitas temu-kembali informasi.
7. Perlu disediakan daftar istilah-istilah

baku berbahasa Indonesia yang tersimpan dalam komputer agar dapat digunakan sebagai acuan pemrosesan teks dalam sistem temu-kembali informasi.

#### DAFTAR PUSTAKA

- Adisantoso, J.** 1997. *Temu-Kembali Informasi Menggunakan Peluang Benyarat*. Fakultas Ilmu Komputer, Universitas Indonesia, Tesis S2.
- Anderberg, M.R.** 1973. *Cluster Analysis for Application*. Academic Press. New York.
- B. Frakes, William and Baeza-Yates, Ricardo.** 1992. *Information Retrieval: data structures and algorithms*. Prentice Hall, New Jersey.
- Couch, C.J. and B. Yang.** 1992. *Experiments in Automatic Statistical Thesaurus Construction*. Proceeding of the 15<sup>th</sup> Annual International ACM SIGR, Denmark.
- Departemen Pendidikan dan Kebudayaan RI.** 1988. *Tata Bahasa Buku Bahasa Indonesia*. Perum Balai Pustaka.
- Duran, B.S. and P.L. Odell.** 1974. *Cluster Analysis*. Springer-Verlag Berlin, New York.
- Johnson, R.A. and D.W. Wichern.** 1988. *Applied Multivariate Stastical Analysis*. 2<sup>nd</sup> ed. Prentice-Hall, Inc.

- Kwok, K.L. 1995.** *A Network Approach to Probabilistic Information Retrieval.* ACM Transactions on Information Systems. 13(3):324-353.
- Magdalena, R. 1996.** *Pembrosesan Teks Penyusunan Tesaursus secara otomatis dalam Bahasa Indonesia dengan Analisis Kluster Hubungan Lengkap (Complete Link).* Fakultas Ilmu Komputer. Universitas Indonesia. Skripsi S1.
- Paice, C.D. 1991.** A thesaurus of Information Retrieval. Information Processing dan Management, Vol.27 No.5, 433-477.
- Pollard, R. 1993.** *Hypertext-Based Thesaurus as a Subject Browsing Aid for Bibliographic Databases.* Information Processing dan Management, V 01.29 No.3, 245-257.
- Salton, G. 1989.** *Automatic Text Processing.* Addison- Wesley Publishing Company. United State of America.
- Salton, G., R. Dartola, and D.M. Murray\_ 1967.** *An Experiment in Automatic Thesaurus Construction.* Information Storage and Retrieval. Scientific Report No.IRS-13 to the National Science Foundation. Ithaca, New York.
- Srinivasan, P. 1992.** *Thesaurus Construction Information Retrieval: Data Structure and Algorithms.* Prentice Hall, New Jersey.