QUERY-SENSITIVE SIMILARITY MEASURE DALAM TEMU KEMBALI DOKUMEN BERBAHASA INDONESIA

Sri Nurdiati[1], Julio Adisantoso[1], Adam Salnor Akbar[2]
[1]Staf Departemen Ilmu Komputer, Fakultas Matematika dan IPA, Institut Pertanian Bogor
[2]Mahasiswa Departemen Ilmu Komputer, Fakultas Matematika dan IPA, Institut Pertanian Bogor

ABSTRAK

Tujuan penelitian ini adalah mengimplementasikan dan menganalisis kinerja dari suatu sistem temu kembali informasi yang menggunakan ukuran kesamaan *Query-Sensitive Similarity Measure* (QSSM) dalam peng-*cluster*-an dokumen, serta membandingkan kinerja ketiga ukuran QSSM (M1, M2, dan M3). Jumlah dokumen yang digunakan dalam pengujian sistem ini sebanyak 700 dokumen dengan 30 kueri beserta gugus jawabannya. Algoritma *Stemming* dan penghilangan *stopwords* untuk pengolahan dokumen digunakan pada implementasi sistem

Hasil penelitian menunjukkan penentuan *threshold* berpengaruh dalam menilai keefektifan suatu ukuran yang digunakan untuk peng-*cluster*-an dokumen. Pada penelitian ini digunakan dua nilai *threshold*, yaitu 0.001 dan 0.025. Pada *threshold* 0.025 ukuran kesamaan M2 lebih unggul dibandingkan kedua ukuran kesamaan lainnya (M1 dan M3) dengan nilai AVP untuk M2=0.497, M1=0.304, dan M3=0.476. Keunggulan M2 dibandingkan M1 dan M3 diperkuat dengan uji *t* berpasangan. Pada *threshold* 0.001 ukuran kesamaan M2 menghasilkan nilai AVP=0.540, M1 menghasilkan nilai AVP=0.491, dan M3 menghasilkan nilai AVP=0.535. Dari nilai AVP yang dihasilkan, ukuran kesamaan M2 lebih unggul dibandingkan M1, M3 juga lebih unggul dibandingkan M1, sedangkan perbandingan antara M2 dan M3 berdasarkan uji *t* berpasangan, keefektifan kedua ukuran ini tidak ada perbedaan, tetapi M2 cenderung lebih unggul pada tiap standar tingkat *recall*.

Kata kunci: Temu Kembali Informasi, *Hierarchic Document Clustering*, *Query Sensitive Similarity Measure.*

1. PENDAHULUAN

Latar Belakang

Dewasa ini informasi banyak dikemas dalam bentuk dijital agar lebih mudah didapatkan oleh pengguna yang membutuhkannya. Untuk lebih memudahkan pengguna dalam memperoleh informasi tersebut, dikembangkan suatu sistem temu kembali informasi. Sistem ini memberikan informasi berdasarkan permintaan atau kueri dari pengguna. Tujuan utamanya adalah memberikan sebanyak mungkin dokumen atau informasi yang relevan dengan kueri dari pengguna dan sesedikit mungkin memberikan informasi yang tidak relevan.

Keefektifan suatu sistem temu kembali informasi dalam memberikan informasi sesuai dengan kueri pengguna dapat ditingkatkan dengan diterapkan metode *Hierarchic Document Clustering*. Metode ini telah diterapkan selama tiga dekade. Acuan dalam pengefektifan ini adalah pada *Clustering Hypothesis*, yang menyatakan dokumen-dokumen yang relevan cenderung memiliki tingkat kesamaan yang tinggi satu sama lain, sehingga dapat dimasukkan ke dalam satu *cluster*.

Salah satu tipe metode peng-*cluster*-an dokumen adalah *query-based*, yaitu suatu informasi dibentuk dari kueri untuk membangkitkan proses peng-*cluster*-an dokumen. Dua pendekatan untuk pembentukan informasi kueri dalam proses peng-*cluster*-an dokumen yaitu berdasarkan tanggapan dari kueri pengguna (*post-retreival clustering*) dan berdasarkan ukuran kesamaan kueri (*query-sensitive similarity measure*). Penelitian ini menekankan pada pendekatan yang kedua, yaitu pembahasan peng-*cluster*-an dokumen berdasarkan ukuran kesamaan kueri dengan *query-sensitive similarity measure* (QSSM) (Tombros 2002).

Tujuan

Penelitian ini bertujuan untuk mengimplementasikan dan menganalisis seberapa efektif suatu sistem temu kembali informasi yang menerapkan QSSM (sebagai

salah satu ukuran kesamaan yang digunakan untuk peng-*cluster*-an dokumen) dalam menemukembalikan dokumen yang relevan dengan kueri pengguna dan membandingkan keefektifan ketiga ukuran kesamaan QSSM (M1, M2, dan M3). Tujuan lainnya adalah menganalisis sejauh mana nilai *threshold* dapat mempengaruhi keefektifan sistem temu kembali berdasarkan *cluster*.

Ruang Lingkup

Penelitian ini difokuskan kepada analisis metode ukuran kesamaan QSSM (M1, M2, dan M3) yang diterapkan pada suatu sistem temu kembali informasi dalam peng-*cluster*-an dokumen. Nilai *threshold* digunakan untuk membatasi jumlah *cluster* yang terbentuk. Dokumen yang akan digunakan dalam penelitian ini dibatasi hanya untuk dokumen berbahasa Indonesia.

2. METODOLOGI

Langkah-langkah Peng-*cluster*-an Dokumen

Langkah-langkah dasar yang mencirikan suatu proses peng-*cluster*-an dokumen (tidak hanya untuk *Hierarchic Document Clustering*) adalah (Rasmussen 1992; Theodoridis & Koutroumbas 1999, diacu dalam Tombros 2002):

1. Representasi dokumen.
2. Pengukuran kesamaan.
3. Penerapan metode peng-*cluster*-an.
4. Representasi hasil peng-*cluster*-an.
5. Validasi dari hasil peng-*cluster*-an.

Dalam *hierarchic document clustering* terdapat dua metode, yaitu metode *agglomerative* (penggabungan) dan *divisive* (pemisahan). Metode yang akan digunakan pada penelitian ini adalah metode *agglomerative*, karena inialisasi awal pada metode *divisive* seluruh dokumen dalam koleksi dimasukkan dalam satu *cluster* besar, lalu dipisah-pisahkan sehingga tiap-tiap *cluster* hanya memiliki satu anggota. Metode *divisive* membutuhkan suatu *term* yang sama yang ada pada seluruh dokumen dalam koleksi, sedangkan pada metode *agglomerative* tidak dibutuhkan. Pada proses awal metode *agglomerative*, jumlah *cluster* sebanyak jumlah dokumen, dengan tiap satu *cluster* berisi satu dokumen, kemudian dibentuk *cluster* baru (penggabungan dari beberapa *cluster*) sampai memenuhi nilai *threshold* (ambang batas). Metode ini selalu mengikuti langkah-langkah berikut (Murtagh 1983, diacu dalam Tombros 2002):

1. Penentuan nilai kesamaan antar dokumen dalam koleksi (pembentukan matriks ukuran kesamaan).
2. Pembentukan sebuah *cluster* dari dua objek atau *cluster* yang terdekat (berdasarkan nilai ukuran kesamaan).
3. Penghitungan kembali nilai kesamaan dari *cluster* terbaru dengan seluruh objek atau *cluster* yang lain dan meninggalkan semua nilai kesamaan yang tidak berubah.
4. Pengulangan langkah 2 dan 3 sampai memenuhi nilai ambang batas.

Representasi Dokumen

Untuk mempermudah dalam pengolahan dokumen nantinya, bentuk standar dari dokumen tersebut harus disamakan. Berdasarkan *corpus* (Adisantoso 2004), bentuk standar dokumen dalam koleksi yaitu berdasarkan *tag-tag* yang mengapit isi dari dokumen:

```
<doc>
<doc no>…</doc no>
<title>…</title>
<author>…</author>
<docsource>…</docsource>
<date>…</date>
<abstract>…</abstract>
<keyword><kw>…</kw></keyword>
<text>
<caption>…</caption>
… (isi dokumen) …
</text>
</doc>.
```

Sebelum masuk ke dalam tahap peng-*cluster*-an, seluruh dokumen yang ada di dalam koleksi dilakukan normalisasi yang terdiri atas beberapa tahap. Pada tahap pertama, tiap dokumen yang ada pada koleksi dilakukan proses *parser* (*tokenizing* dan penghilangan *stopwords*). Tokenizing adalah pengolahan tiap dokumen yang ada pada koleksi menjadi unit-unit yang paling kecil atau disebut dengan token/*term*, kemudian kata-kata yang berupa *stopwords* (Ridha 2002) dihilangkan.

Setelah proses *parser* selesai, tahap yang kedua adalah proses *stemming*, yaitu penghilangan prefiks dan sufiks dari *term* yang ada pada dokumen (Grossman 2002). Algoritma *stemming* yang akan digunakan mengacu kepada *stemming* Ridha (2002).

Setelah dilakukan *stemming*, tahap selanjutnya adalah penghitungan bobot untuk tiap-tiap *term* yang terdapat pada seluruh dokumen dalam koleksi, yang hasil akhirnya direpresentasikan dalam matriks istilah-dokumen.

Penghitungan Ukuran Kesamaan

Setelah dokumen direpresentasikan, kemudian dilakukan ukuran kesamaan antara dokumen-dokumen yang ada pada koleksi dengan kueri masukan dari pengguna. Ukuran kesamaan yang digunakan adalah dengan QSSM. Prinsip dasar dari ukuran kesamaan ini adalah ukuran kesamaan *Cosine* yang telah dikembangkan (Tombros 2002).

Penjelasan bagaimana langkah melakukan penghitungan ukuran kesamaan adalah sebagai berikut. Setelah kueri dimasukkan oleh pengguna, maka akan dibentuk vektor kueri yang berisi bobot *term* pada kueri. Lalu dilakukan pengukuran kesamaan antara dokumen-dokumen yang ada pada koleksi dengan kueri menggunakan QSSM (M1, M2, dan M3). Nilai parameter yang digunakan untuk ukuran M3 dengan perbandingan ϑ_1 dan ϑ_2 adalah 1:4, karena dinilai lebih efektif dibandingkan perbandingan yang lainnya berdasarkan rata-rata nilai tes *5NN* (Tombros 2002).

Pembentukan *Cluster*

Tahap selanjutnya adalah peng-*cluster*-an dokumen dengan menggunakan metode *group average link*. Hasil dari tahap ini berupa *cluster-cluster* yang berisi satu atau lebih dokumen berdasarkan kedekatan antardokumen yang dilihat dari ukuran kesamaan yang telah dihitung sebelumnya. Penggabungan satu atau lebih dokumen ke dalam satu *cluster* dibatasi nilai ukuran kesamaannya melalui *threshold.*

Evaluasi Sistem Temu Kembali Informasi

Untuk menilai seberapa efektif suatu sistem temu kembali informasi dilakukan, terdapat beberapa teknik dan dalam penelitian ini teknik yang digunakan adalah penghitungan AVP. Sistem temu kembali yang ideal memiliki AVP=1 (berdasarkan dokumen-dokumen yang ditemukembalikan).

Pengujian dilakukan dengan uji *t* berpasangan dengan selang kepercayaan 95% (α=0.05) terhadap pengaruh perbedaan ukuran kesamaan yang digunakan, yaitu:

1. M1 dan M2,
2. M1 dan M3,
3. M2 dan M3.

Koleksi Pengujian

Koleksi dokumen yang digunakan penulis berasal dari *corpus* (Adisantoso 2004) yang berkaitan dengan masalah pertanian dan 30 kueri beserta gugus jawaban (Andika 2005). Jumlah dokumen yang akan digunakan dalam penelitian ini sebanyak 700 dokumen.

3. HASIL DAN PEMBAHASAN

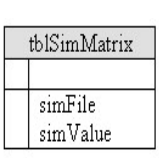
Pengolahan Awal Dokumen

Langkah awal pada peng-*cluster*-an dokumen adalah merepresentasikan dokumen, pada langkah ini terdapat beberapa proses, yaitu pengubahan struktur dokumen ke dalam bentuk *tag-tag*, proses *parser* (meliputi *tokenizer* dan penghilangan *stopwords*) ,dan *stemming.* Dari hasil proses *parser* dan *stemming* didapatkan data yang dapat dilihat pada Tabel 1.

Tabel 1 Hasil *parsing* dan *stemming*.

Jumlah Dokumen	Jumlah Token Keseluruhan	Jumlah Token Unik
700	131362	15273

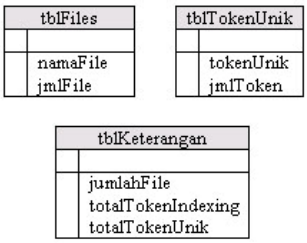
Hasil dari proses *parsing* dan *stemming* selanjutnya dilakukan penghitungan bobot token, yang akan membentuk sebuah matriks istilah-dokumen. Matriks ini akan disimpan di dalam basisdata. Alasan penulis melakukan penyimpanan di dalam basisdata adalah agar token dan nilai bobotnya bisa digunakan dalam proses berikutnya tanpa harus melakukan *parsing*, *stemming*, dan penghitungan bobot setiap saat, kecuali ada penambahan dokumen baru ke dalam koleksi. Tabel dalam basisdata yang digunakan untuk menampung matriks istilah-dokumen hanya berupa satu tabel, yaitu *tblWeight*, dengan jumlah *field* sebanyak lima buah. Representasi *tblWeight* dapat dilihat pada Gambar 2.



Gambar 2 Representasi tabel *tblWeight* dalam basisdata.

Selain menggunakan tabel *tblWeight* untuk menampung matriks istilah-dokumen, disediakan juga beberapa tabel tambahan

sebanyak tiga tabel yang digunakan untuk menampung keterangan tambahan dari hasil pembentukan matriks istilah-dokumen, yaitu *tblFiles, tblTokenUnik,* dan *tblKeterangan*. Representasi dari ketiga tabel dapat dilihat pada Gambar 3.



Gambar 3 Representasi ketiga tabel keterangan tambahan.

Algoritma dari pengolahan dokumen di atas yang telah diterapkan dalam sistem adalah sebagai berikut:

1. Dilakukan proses *tokenizer* untuk seluruh dokumen dalam koleksi, sehingga dihasilkan *term-term* atau token-token, kemudian diberi id masing-masing token sesuai keberadaan token dalam dokumen. Selanjutnya dihitung jumlah frekuensi kemunculan *term* tersebut dalam suatu dokumen.
2. Sebelum dimasukkan ke basisdata, token-token hasil *tokenizer* dicek apakah termasuk *stopwords*, apabila tidak termasuk, token dapat dimasukkan ke dalam basisdata.
3. Langkah selanjutnya dilakukan proses *stem* sesuai dengan algoritma *stemming* Ridha (2002).
4. Setelah ketiga proses di atas, dilakukan penghitungan bobot masing-masing token, kemudian dimasukkan ke basisdata sehingga terbentuk matriks istilah-dokumen.

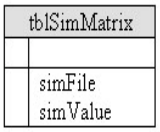
Nilai *Threshold*

Penentuan nilai *threshold* atau ambang batas mempengaruhi jumlah *cluster* dan anggota *cluster* yang terbentuk. Nilai ambang batas berkisar antara 0 sampai 1. Semakin mendekati 1, jumlah *cluster* yang terbentuk semakin banyak, dengan jumlah anggota *cluster* yang semakin sedikit. Semakin mendekati 0, jumlah *cluster* yang terbentuk semakin sedikit, dengan jumlah anggota *cluster* yang semakin banyak.

Matriks Kesamaan Pasangan Dokumen

Sebelum dilakukan penghitungan ukuran kesamaan antara dokumen dengan kueri menggunakan QSSM, terlebih dahulu dibentuk matriks ukuran kesamaan antara dua dokumen. Matriks ini berisi ukuran kesamaan pasangan dokumen, dengan ukuran n x n (n menyatakan banyak dokumen dalam koleksi). Pembentukan matriks ini bertujuan untuk mempermudah dalam penghitungan ukuran kesamaan berikutnya dengan menggunakan QSSM.

Matriks yang terbentuk akan dimasukkan ke dalam basisdata (Gambar 4), agar tidak perlu dilakukan penghitungan ulang dalam menghitung nilai QSSM, karena proses pembentukan matriks ini cukup memakan waktu lama. Misalnya n adalah banyak dokumen dalam koleksi, t adalah waktu penghitungan tiap proses, maka waktu pembentukan matriks ini adalah (nC2) x t. Oleh karena itu, waktu pembentukan matriks ini bertambah secara eksponensial terhadap pertambahan jumlah dokumen ke dalam koleksi.



Gambar 4 Representasi tabel *tblSimMatrix* dalam basisdata.

Penghitungan QSSM dan *Threshold*

Hasil penelitian menunjukkan nilai *threshold* (yang digunakan untuk membatasi sampai sejauh mana kedekatan ukuran kesamaan antar dokumen dapat digabungkan menjadi satu *cluster*) mempengaruhi keefektifan sistem temu kembali. Ukuran kesamaan yang digunakan untuk pembentukan *cluster* dalam penelitian ini adalah ukuran QSSM (terdiri dari M1, M2, dan M3). Langkah awal penghitungan ukuran kesamaan ini adalah setelah kueri dari pengguna dimasukkan dan sebelum dilakukan peng-*cluster*-an dokumen, dilakukan penghitungan fungsi yang kedua dalam QSSM (M2) atau proses pembentukan matriks ukuran kesamaan antara pasangan dokumen dengan kueri. Hal ini dilakukan, karena kedua ukuran QSSM lainnya melibatkan fungsi kedua QSSM dalam penghitungan ukuran kesamaan. Setelah matriks pasangan dokumen-kueri terbentuk, kemudian dilakukan penghitungan ukuran kesamaan dengan ketiga ukuran QSSM (M1, M2, dan M3). Lalu dilakukan proses *cluster* dengan metode *agglomerative* dan penghitungan *group average link.*

Penentuan nilai *threshold*, cukup signifikan mempengaruhi jumlah *cluster* yang terbentuk dan nilai AVP yang dihasilkan berdasarkan jumlah dokumen relevan yang

ditemukembalikan. Berdasarkan pengamatan yang telah dilakukan, dalam penelitian ini penulis menggunakan dua nilai *threshold*, yaitu 0.025 dan 0.001. Sebelumnya telah dilakukan beberapa kali percobaan dalam menentukan nilai *threshold* dan diusahakan agar semua dokumen relevan untuk tiap kueri (yang telah ditentukan sebelumnya) dapat ditemukembalikan. Dari hasil percobaan pada Tabel 2 dapat dilihat rata-rata jumlah dokumen relevan yang ditemukembalikan per kueri mendekati nilai rata-rata jumlah dokumen relevan per kueri yang sebenarnya, sehingga ditetapkan nilai *threshold* yang digunakan adalah 0.025 dan 0.001.

Tabel 2 Deskripsi koleksi pengujian.

Keterangan	Ukuran
Jumlah kueri	30
Jumlah dokumen	700
Rataan kata per kueri	2,6
Rataan dokumen relevan per kueri	31,6

Bila *threshold* yang digunakan lebih dari 0.025 maka jumlah dokumen relevan yang ditemukembalikan terlalu sedikit, sehingga menurunkan nilai AVP. Bila *threshold* yang digunakan kurang dari 0.001, jumlah dokumen yang ditemukembalikan terlalu banyak, bahkan seluruh koleksi ditemukembalikan. Jadi kedua nilai *threshold* inilah yang paling efisien digunakan dalam penelitian ini. Untuk melihat lebih jelas data dari rataan dokumen relevan yang ditemukembalikan pada masing-masing *threshold* dapat dilihat pada Tabel 3.

Tabel 3 Rataan dokumen relevan yang ditemukembalikan.

Threshold	Rataan Dokumen Relevan		
	M1	M2	M3
0.001	27,2	30,2	31
0.025	11,2	29,2	28,2

Penentuan *Cluster* Relevan

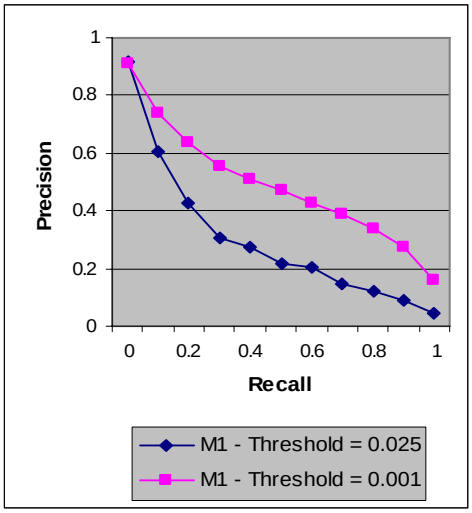
Hasil yang diperoleh dari penghitungan QSSM dan proses penggabungan beberapa dokumen menjadi satu *cluster* dengan metode *agglomerative* dan penghitungan *group average link* adalah terbentuknya beberapa *cluster*. Hasil keluaran dari kueri yang dimasukkan hanya mengembalikan satu *cluster* yang paling relevan dengan kueri. Penentuan *cluster* yang paling relevan ini adalah dari rata-rata tertinggi ukuran kesamaan tiap dokumen yang ada dalam *cluster* dengan kueri.

Pengurutan Hasil Temu Kembali

Ketika sudah didapatkan *cluster* yang paling relevan dengan kueri, anggota *cluster* berupa dokumen-dokumen akan diurutkan ketika akan ditampilkan ke pengguna. Hal ini bertujuan untuk meningkatkan *precision*, sehingga dapat meningkatkan nilai AVP. Pengurutan anggota *cluster* ini berdasarkan nilai ukuran kesamaan antara dokumen dengan kueri dari yang terbesar ke yang terkecil.

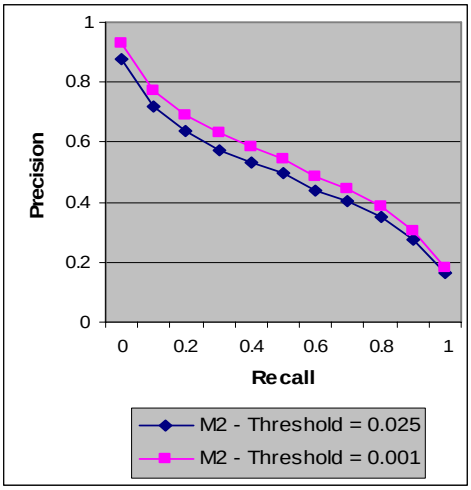
Evaluasi QSSM

Hasil penelitian ukuran kesamaan M1 pada *threshold* 0.001, nilai rata-rata *precision* pada 11 standar tingkat *recall* mengalami kenaikan sebesar 61.71% dibandingkan ukuran kesamaan M1 pada *threshold* 0.025. Hal ini diakibatkan jumlah dokumen relevan yang ditemukembalikan pada *threshold* 0.001 mengalami peningkatan. Kurva perbandingan nilai *precision* pada 11 standar tingkat *recall* dapat dilihat pada Gambar 5.



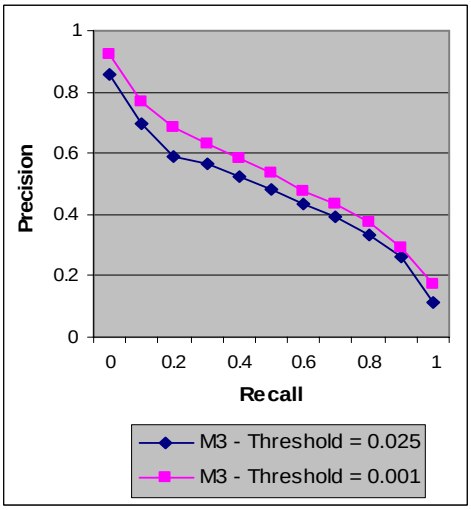
Gambar 5 Kurva *recall-precision* pada *threshold* 0.025 dan 0.001 untuk M1.

Pada ukuran kesamaan M2, hasil penelitian menunjukkan kenaikan nilai rata-rata *precision* pada *threshold* 0.001 dibandingkan dengan *threshold* 0.025 tidak terlalu besar dibandingkan dengan ukuran kesamaan M1, yaitu sebesar 8.72%. Kurva perbandingan nilai *precision* pada 11 standar tingkat *recall* dapat dilihat pada Gambar 6.



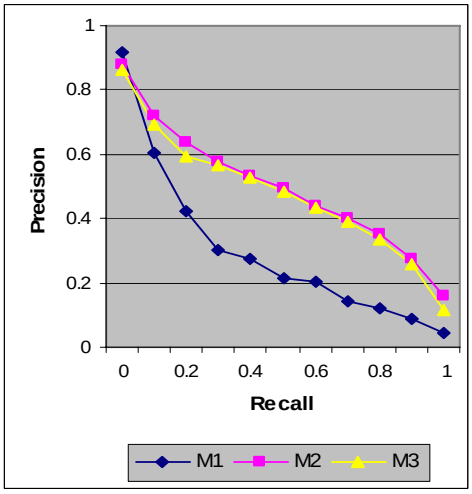
Gambar 6 Kurva *recall-precision* pada *threshold* 0.025 dan 0.001 untuk M2.

Pada ukuran kesamaan M3, hasil penelitian menunjukkan pada *threshold* 0.001 nilai rata-rata *precision* mengalami kenaikan sebesar 11.88% dibandingkan ukuran kesamaan M3 pada *threshold* 0.025. Kurva perbandingan nilai *precision* pada 11 standar tingkat *recall* dapat dilihat pada Gambar 7.



Gambar 7 Kurva *recall-precision* pada *threshold* 0.025 dan 0.001 untuk M3.

Hasil penelitian ketiga ukuran kesamaan QSSM, untuk nilai rata-rata *precision* pada 11 standar tingkat *recall* dengan *threshold* 0.025 disajikan dalam bentuk kurva (Gambar 8). Dari kurva terlihat bahwa ukuran M2 dan M3 lebih unggul dibandingkan ukuran M1, tetapi bila dibandingkan antara ukuran M2 dengan ukuran M3, perbedaan kedua ukuran ini tidak terlalu signifikan.



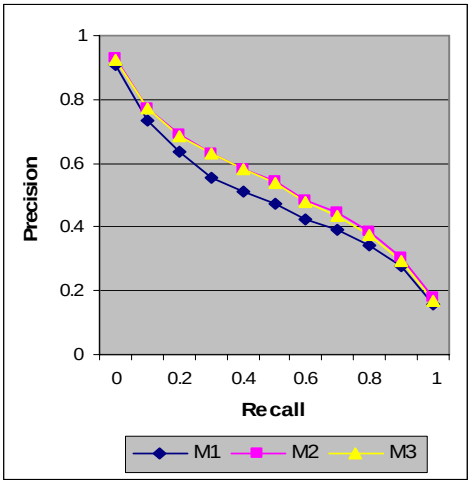
Gambar 8 Kurva perbandingan *recall–precision* M1, M2, dan M3 pada *threshold* 0.025.

Berdasarkan uji *t* berpasangan dengan selang kepercayaan 95% (α=0.05) dapat disimpulkan ukuran M2 lebih unggul dibandingkan dengan M1, ukuran M3 lebih unggul dibandingkan ukuran M1, dan ukuran M2 lebih unggul dibandingkan M3. Evaluasi sistem temu kembali dengan penghitungan nilai AVP (30 kueri masukan) untuk *threshold* 0.025 dan uji *t* berpasangan dapat disimpulkan ukuran kesamaan M2 memiliki nilai keefektifan tertinggi dibandingkan kedua ukuran kesamaan lainnya. Data hasil penghitungan nilai AVP dapat dilihat pada Tabel 4.

Tabel 4 Nilai AVP untuk M1, M2, dan M3 pada *threshold* 0.025.

No.	Ukuran Kesamaan	AVP
1.	M1	0.304
2.	M2	0.497
3.	M3	0.476

Hasil penelitian ketiga ukuran kesamaan QSSM untuk nilai rata-rata *precision* pada 11 standar tingkat *recall* dengan *threshold* 0.001 disajikan dalam bentuk kurva (Gambar 9). Dari kurva terlihat bahwa ukuran M2 dan M3 lebih unggul dibandingkan ukuran M1, tetapi bila dibandingkan antara ukuran M2 dengan ukuran M3, perbedaan kedua ukuran ini tidak terlalu signifikan.



Gambar 9 Kurva perbandingan *recall–precision* M1, M2, dan M3 pada *threshold* 0.001.

Berdasarkan uji *t* berpasangan dengan selang kepercayaan 95% (α=0.05) dapat disimpulkan ukuran M2 dan M3 lebih unggul dibandingkan ukuran M1, tetapi untuk perbandingan antara ukuran M2 dan M3, keefektifan dari kedua ukuran ini tidak ada perbedaan. Evaluasi sistem temu kembali dengan penghitungan nilai AVP (30 kueri masukan) untuk *threshold* 0.001 menghasilkan ukuran kesamaan M2 dan M3 memiliki nilai keefektifan lebih tinggi dibandingkan ukuran kesamaan M1, tetapi perbandingan antara ukuran kesamaan M2 dan M3 berdasarkan nilai AVP, M2 cenderung lebih tinggi pada tiap standar tingkat *recall*. Data hasil penghitungan nilai AVP dapat dilihat pada Tabel 5.

Tabel 5 Nilai AVP untuk M1, M2, dan M3 pada *threshold* 0.001.

No.	Ukuran Kesamaan	AVP
1.	M1	0.491
2.	M2	0.540
3.	M3	0.535

4. KESIMPULAN DAN SARAN

Kesimpulan

Melalui hasil penelitian ini dapat ditarik beberapa kesimpulan sebagai berikut:

1. Penentuan nilai *threshold* ternyata mempengaruhi keefektifan sistem temu kembali yang menggunakan sistem peng-*cluster*-an dokumen. Apabila nilai *threshold* yang ditentukan tinggi (mendekati 1), maka jumlah dokumen yang ditemukembalikan sedikit, sehingga nilai AVP menjadi turun. Bila nilai *threshold* yang ditentukan rendah (mendekati 0), maka jumlah dokumen yang ditemukembalikan banyak dan ada kemungkinan seluruh dokumen dalam koleksi ditemukembalikan, sehingga sistem menjadi tidak efektif.
2. Ukuran kesamaan M2 lebih unggul dibandingkan kedua ukuran kesamaan lainnya pada *threshold* 0.025. Pada *threshold* 0.001, ukuran M2 dan M3 lebih unggul dibandingkan ukuran M1, tetapi tidak ada perbedaan keefektifan pada kedua ukuran ini secara keseluruhan, walaupun perbandingan antara ukuran kesamaan M2 dan M3 berdasarkan nilai AVP, M2 cenderung lebih tinggi pada tiap standar tingkat *recall*.
3. Sistem temu kembali berdasarkan *cluster* dengan menggunakan ukuran kesamaan QSSM dapat digunakan sebagai acuan dalam pengimplementasian suatu sistem temu kembali informasi, karena hasil evaluasi dari sistem ini secara rata-rata menghasilkan nilai keefektifan yang cukup baik.

Saran

Untuk pengembangan penelitian ini, disarankan hal-hal sebagai berikut:

1. Koleksi pengujian dalam penelitian diharapkan memiliki ukuran yang lebih besar, agar dapat diketahui apakah keefektifan ukuran kesamaan QSSM masih bertahan.
2. Keefektifan penggunaan ukuran kesamaan QSSM dalam peng-*cluster*-an dibandingkan dengan penggunaan ukuran kesamaan lainnya.
3. Dalam Tombros (2002), panjang kueri mempengaruhi keefektifan ukuran kesamaan QSSM dalam peng-*cluster*-an dokumen, penulis berharap dalam pengembangan penelitian ini digunakan panjang kueri lebih dari 10 kata.

5. REFERENSI

Adisantoso J. 2004. *Corpus* Dokumen Teks Bahasa Indonesia untuk Pengujian Efektifitas Temu Kembali Informasi. Laporan Akhir Hibah Penelitian SP4, Departemen Ilmu Komputer FMIPA IPB, Bogor.

Anderberg MR. 1973. *Cluster Analysis for Applications*. New York: Academic Press.

Baeza-Yates R, Ribeiro-Neto B. 1999. *Modern Information Retrieval*. England: Addison-Wesley Publishing Company.

Belew RK. 2000. *Finding out about: A cognitive perspective on search engine technology and the WWW*. Cambridge: Cambridge University Press.

Grossman D. *IR Book*. http://www.ir.iit.edu/~dagr/cs529/files/ir_book/ [7 Maret 2002].

Murtagh F. 1983. *A survey of recent advances in hierarchical clustering algorithms*. Computer Journal, 26:354-359.

Rasmussen E. 1992. *Clustering Algorithms*. In Frakes, W.B. and Baeza-Yates, R. (editors) Information Retrieval: Data Structures and Algorithms. New Jersey: Prentice Hall.

Ridha A. 2002. Pengindeksan Otomatis dengan Istilah Tunggal untuk Dokumen Berbahasa Indonesia. Skripsi. Institut Pertanian Bogor. [20 November 2005].

Rijsbergen CJ van. 1979. *Information Retrieval, Second Edition***.** Butterworths, London.

Salton G, McGill JM. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill Book.

Theodoridis S, Koutroumbas K. 1999. *Pattern Recognition*. San Diego: Academic Press.

Tombros A. 2002. *The Effectiveness of Query-Based Hierarchic Clustering of Documents for Information Retrieval*. Tesis. University of Glasgow. [15 September 2005]

Voorhees EM. 1985. *The effectiveness and efficiency of agglomerative hierarchic clustering in document retrieval*. Tesis. Technical Report TR 85-705 of the Department of Computing Science, Cornell University.

Voorhees EM. 1986. *Implementing agglomerative hierarchic clustering algorithms for use in document retrieval*. Information Processing & Management, 22(6):465-476.

Witten IH, Moffat A, Bell TC. 1999. *Managing Gigabytes: compressing and indexing documents and images*. USA: Academic Press.