EVALUASI PENGGUNAAN *SIMILARITY THESAURUS* TERHADAP EKSPANSI KUERI DALAM SISTEM TEMU KEMBALI INFORMASI BERBAHASA INDONESIA

Fridolin Febrianto Paiki
Jurusan Teknik, Universitas Negeri Papua
Jl. Gunung salju Amban, Manokwari
ff.paiki@unipa.ac.id

Abstrak

Istilah merupakan komponen utama dalam sistem temu kembali informasi. Penggunaan istilah sebagai indeks dan kueri sangat mempengaruhi kinerja dari suatu sistem temu kembali informasi. Penelitian ini mengamati pengaruh penggunaan similarity thesaurus dalam proses ekspansi kueri terhadap kinerja sistem temu kembali informasi berbahasa Indonesia. Dengan menggunakan 30 gugus kueri pada 1.000 dokumen di koleksi, dilakukan uji banding antara pemberian bobot yang sama dan yang berbeda terhadap istilah di dalam kueri untuk melihat pengaruhnya terhadap hasil temu kembali sebelum dan setelah melakukan ekspansi kueri. Dengan memanfaatkan cosine sebagai ukuran kesamaan dan bobot istilah dalam kueri, ditentukan istilah-istilah yang akan digunakan dalam ekspansi kueri. Dua perlakuan yang diambil adalah dengan mengambil 5 (TH5) dan 10 (TH10) istilah yang memiliki nilai kesamaan terbesar dengan kueri. Setelah dibandingkan dengan hasil temu kembali tanpa ekspansi kueri (NoTH) diperoleh bahwa penggunaan similarity thesaurus secara keseluruhan meningkatkan kinerja sistem. Namun, hal ini tidak sepenuhnya berlaku sebab sangat dipengaruhi oleh bobot istilah di dalam kueri. Pada tiga percobaan, yaitu tanpa thesaurus (NoTH) dan dengan thesaurus (TH5 atau TH10) untuk bobot istilah dalam kueri yang berbeda (1 dan idf) diperoleh bahwa penggunaan idf sebagai bobot istilah dalam kueri meningkatkan kinerja sistem, baik dalam hal temu kembali biasa maupun ekspansi kueri.

Kata kunci: ekspansi kueri, kesamaan thesaurus, temu kembali informasi

Abstract

Terms or tokens are the main component in information retrieval system. The use of them as index and query affects the performance of the system. This research is conducted to observe how similarity thesaurus improves the performance of the Indonesian information retrieval system through query expansion. By using 30 sets of query and 1.000 documents, series of tests are conducted using different weight of terms in query to measure the performance of the system before and after query expansion. By using cosine as similarity measurement and the weight of the query terms, the terms used in query expansion can be determined. Two treatments that were used are by taking 5 (TH5) and 10 (TH10) terms that has the biggest similarity value with the query. It is found that overall the query expansion improve the performance of the system compared to the one without query expansion (NoTH). However, it also depends on the weight of the terms in the query. On three experiment combined with NoTH, TH5, and TH10, the results show that idf is proved to be better used as weight of the terms in query in order to improve the performance of the system, either using query expansion or without query expansion.

Key words*: query expansion, , similarity thesauru*

1. PENDAHULUAN

1.1 Latar Belakang

Dalam temu kembali informasi, jumlah dokumen relevan yang ditemukembalikan akan dipengaruhi oleh jumlah kata kunci yang digunakan untuk pencarian. Masalah yang dihadapi adalah seringkali pengguna tidak mampu merepresentasikan kebutuhan informasi yang diinginkan ke dalam bentuk kueri. Masalah lain yang sering muncul adalah pilihan kata yang digunakan. Seringkali pilihan kata yang digunakan pengguna di dalam kueri berbeda dengan pilihan kata yang digunakan penulis. Selain itu, kebutuhan informasi dapat direpresentasikan dengan pilihan kata yang berbeda oleh pengguna yang berbeda.

Untuk memecahkan masalah-masalah tersebut salah satu pendekatan yang dapat digunakan adalah ekspansi kueri. Dengan pendekatan ini, kueri yang diberikan pengguna akan diperluas dengan menambahkan kata-kata pada kueri awal kemudian kueri yang baru tersebut akan digunakan untuk pencarian yang berikutnya. Dengan demikian, diharapkan hasil temu kembali menjadi lebih baik.

Salah satu bentuk ekspansi kueri dengan menggunakan analisis global adalah *similarity thesaurus*. Metode ini merupakan salah satu bentuk *automatic query expansion*. Metode ini memanfaatkan kata-kata yang memiliki nilai kesamaan terbesar dengan kueri sebagai kata-kata yang akan digunakan pada ekspansi kueri. *Similarity thesaurus* hanya mengambil istilah-istilah yang paling dekat dengan keseluruhan kueri (Qiu & Frei 1993).

1.2 Tujuan

Penelitian ini bertujuan untuk mengimplementasi dan mengevaluasi penggunaan *similarity thesaurus* dalam proses ekspansi kueri dalam sistem temu kembali informasi untuk koleksi dokumen teks berbahasa Indonesia.

1.3 Ruang Lingkup

Penelitian ini terbatas pada analisis pengaruh *similarity thesaurus* terhadap kinerja dari sistem temu kembali informasi berbahasa Indonesia. Dalam hal ini, pengaruhnya dilihat berdasarkan pada nilai *recall* dan *precision*.

Model sistem temu kembali yang digunakan adalah *vector space model*. Koleksi dokumen yang digunakan merupakan dokumen-dokumen teks berbahasa Indonesia yang berupa kumpulan berita-berita di bidang pertanian secara umum.

2. TINJAUAN PUSTAKA

2.1 Ekspansi Kueri

Ekspansi kueri dengan analisis lokal hanya menggunakan kueri dan dokumen-dokumen yang sudah ditemukembalikan pada pencarian awal. Dalam hal ini, analisis lokal digunakan untuk menentukan istilah-istilah yang tepat untuk ekspansi kueri (Baeza-Yates & Ribeiro-Neto 1999). Ada dua pendekatan yang sering digunakan, yaitu *relevance feedback* dan *local feedback*.

Metode analisis lain yang juga sering digunakan dalam ekspansi kueri adalah analisis global. Prinsip dasarnya adalah dengan memanfaatkan konteks suatu kata untuk menentukan kesamaannya dengan kata yang lain (Baeza-Yates & Ribeiro-Neto 1999). Dalam hal ini, prosesnya dilakukan untuk seluruh dokumen di dalam koleksi.

2.2 Similarity Thesaurus

Similarity thesaurus merupakan suatu matriks yang berisi nilai-nilai kesamaan antara suatu istilah dengan istilah yang lain (Schäuble & Knaus 1992). *Similarity thesaurus* dibentuk berdasarkan pada bagaimana istilah-istilah tersebut diindeks menggunakan dokumen-dokumen. *Similarity thesaurus* dibentuk dengan melihat kemampuan dokumen untuk merepresentasikan arti dari istilah (2.1).

Metode pembobotan yang digunakan dalam perancangan *similarity thesaurus* ditunjukkan oleh persamaan berikut.

$$w_{i,j} = \frac{\left(0.5 + 0.5\frac{f_{i,j}}{\max_j(f_{i,j})}\right) \text{x } itf_j}{\sqrt{\sum_{l=1}^{N}\left(0.5 + 0.5\frac{f_{i,l}}{\max_l(f_{i,l})}\right)^2 \text{x } itf_j^2}} \quad ,(2.1)$$

dengan :

$$itf_j = \log\left(\frac{t}{t_j}\right), (2.2)$$

dengan :

- $w_{i,j}$: bobot istilah k_i pada dokumen d_j
- $f_{i,j}$: frekuensi kemunculan istilah k_i pada dokumen d_j
- $max(f_{i,j})$: frekuensi maksimum kemunculan istilah di koleksi
- itf_j : inverse term frequency untuk dokumen d_j
- N : jumlah dokumen di koleksi
- t : jumlah istilah indeks unik di koleksi
- t_j : jumlah istilah indeks pada dokumen d_j

Persamaan (2.5) menunjukkan bahwa dokumen dengan jumlah istilah yang kecil lebih berpengaruh dalam penentuan bobot istilah daripada dokumen dengan jumlah istilah yang besar.

Terdapat beberapa ukuran kesamaan yang dapat digunakan, yaitu *Correlation*, *Cosine*, *Jaccard*, dan *Dice*. Dalam penelitian ini, ukuran kesamaan yang akan digunakan adalah *Cosine*. Persamaannya adalah:

$$c_{x,y} = \frac{\sum_{i=1}^{n} w_{x,i} w_{y,i}}{\sqrt{\sum_{i=1}^{n} w_{x,i}^2}\sqrt{\sum_{i=1}^{n} w_{y,i}^2}} \quad ,(2.3)$$

- dengan :
- $c_{x,y}$: korelasi antara term k_x dan k_y
- $w_{x,i}$: bobot istilah k_x pada dokumen d_i

2.3 Ekspansi Kueri

Selberg (1997) dalam Agusetyawan (2006) menyatakan bahwa ekspansi kueri adalah sekumpulan teknik untuk memodifikasi kueri dengan tujuan untuk memenuhi sebuah kebutuhan

informasi. Ekspansi kueri dapat berarti penambahan maupun pengurangan kata pada kueri.

Proses ekspansi kueri berdasarkan *similarity thesaurus* dilakukan melalui tiga tahap, yaitu :

1. Merepresentasikan kueri dalam ruang vektor.
$$\bar{q} = \sum_{ki \in q} w_{i,q} k_i \text{ ,(2.4)}$$

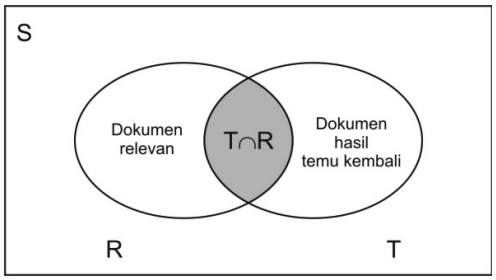
2. Menghitung nilai kesamaan antara istilah-istilah (yang berkorelasi dengan istilah-istilah pada kueri) dan kueri.
$$sim(q, k_v) = \sum_{ku \in q} w_{u,q} \, x \, c_{u,v} \text{ ,} \quad (2.5)$$
dengan :
   - $sim(q,k_v)$:nilai kesamaan antara kueri (q) dengan istilah-istilah yang berkorelasi dengan istilah di dalam kueri (k_v)
   - $w_{u,q}$: bobot istilah dalam kueri
   - $c_{u,v}$: nilai kesamaan antara istilah di kueri (k_u) dengan istilah yang berkorelasi dengannya (k_v)

3. Melakukan ekspansi kueri dengan mengambil r istilah teratas berdasarkan nilai $sim(q,k_v)$. Bobot dari istilah-istilah yang diambil tersebut dihitung kembali dengan persamaan :
$$w_{v,q} = \frac{sim(q,k_v)}{\sum_{ku \in q} w_{u,q}} \text{ ,(2.6)}$$
dengan :
   - $w_{v,q}$: bobot istilah hasil ekspansi yang baru
   - $sim(q,k_v)$:nilai kesamaan antara kueri (q) dengan istilah-istilah yang berkorelasi dengan istilah di dalam kueri (k_v)
   - $w_{u,q}$: bobot istilah dalam kueri

2.4 Evaluasi Sistem Temu Kembali Informasi

Berbagai macam ukuran dapat digunakan untuk mengevaluasi kinerja suatu sistem temu kembali informasi. Cleverdon (1996) menyatakan bahwa terdapat enam jenis ukuran yang dapat digunakan yaitu *coverage*, *time lag*, *presentation*, *effort*, *recall*, dan *precision*. Dari keenam ukuran tersebut, dua ukuran yang paling umum digunakan adalah *recall* dan *precision*.

Recall dan *precision* mengukur kemampuan sistem dalam menemukembalikan dokumen-dokumen yang relevan dan menahan dokumen-dokumen yang tidak relevan. *Recall* merupakan rasio jumlah dokumen relevan yang ditemukembalikan terhadap jumlah seluruh dokumen relevan di dalam koleksi. *Precision* merupakan rasio jumlah dokumen relevan yang ditemukembalikan terhadap jumlah seluruh dokumen yang ditemukembalikan. Gambaran lebih jelas dapat dilihat pada Gambar 1.



Gambar 1. Ilustrasi *recall* dan *precision* untuk suatu kueri tertentu.

$$Recall = \frac{|T \cap R|}{|R|} \text{, (2.7)}$$

$$Precision = \frac{|T \cap R|}{|T|} \text{, (2.8)}$$

3. METODE PENELITIAN

3.1 Koleksi Pengujian

Dokumen-dokumen yang akan digunakan adalah dokumen berita dalam bidang pertanian secara umum sebanyak seribu dokumen. Koleksi dokumen tersebut merupakan hasil penelitian Adisantoso & Ridha (2004). Koleksi berita ini dikumpulkan dari beberapa sumber di internet dimana gugus kueri dan gugus jawabannya sudah tersedia.

3.2 Pengindeksan

Proses ini akan dibagi ke dalam dua proses yang berurutan, yaitu *tokenizing* dan *weighting*. *Tokenizing* akan menghasilkan *token-token* yang akan digunakan sebagai istilah indeks. Selanjutnya, *weighting* akan menghasilkan bobot terhadap *token-token* yang dihasilkan sebelumnya. Masing-masing proses tersebut dilakukan untuk seluruh dokumen di koleksi.

Pada *tokenizing* akan dilakukan pembacaan karakter per karakter. Tujuannya adalah untuk membedakan karakter-karakter yang bersifat *separator* dan yang bukan. Masing-masing *token* tersebut harus diperiksa keberadaannya di dalam *stoplist*. Jika *token* tersebut termasuk *stoplist*, maka harus dibuang dan sebaliknya jika tidak,

maka *token* tersebut akan digunakan sebagai istilah indeks. Daftar kata-kata yang digunakan sebagai *stoplist* diambil dari penelitian Ridha (2002).

Setelah itu, akan dilakukan pembobotan (*weighting*). Proses ini sangat penting karena hasilnya akan sangat mempengaruhi kinerja sistem. Tujuannya adalah menentukan tingkat kepentingan suatu *token* di dalam dokumen. Metode yang akan digunakan adalah *tf-idf* dan *similarity thesaurus*.

3.3 Matriks Kesamaan

Berdasarkan indeks yang sudah dibuat akan dihasilkan matriks kesamaan secara automatis. Ukuran kesamaan yang digunakan adalah *cosine*. Ukuran kesamaan ini dipilih dengan maksud untuk menyesuaikan dengan model sistem temu kembali, yaitu *vector space model* (VSM).

3.4 Ekspansi Kueri

Metode ekspansi yang akan digunakan dalam penelitian ini adalah Automatic Query Expansion (AQE) dengan menggunakan analisis global, yang akan menghasilkan *similarity thesaurus*. *Similarity thesaurus* tersebut akan digunakan pada saat pemilihan kata-kata yang akan digunakan dalam ekspansi kueri.

Dalam *similarity thesaurus*, ekspansi kueri dilakukan dengan mengambil sejumlah istilah yang paling mirip dengan kueri. Dalam penelitian ini, jumlah istilah yang akan diuji adalah lima istilah dan sepuluh istilah. Tujuannya adalah untuk menyesuaikan dengan ukuran koleksi yang juga kecil (seribu dokumen).

Pada proses pemilihan istilah tersebut terdapat suatu nilai ambang batas (*threshold*) yang membatasi jumlah istilah yang akan dipakai dalam ekspansi kueri. *Threshold* membatasi istilah-istilah dengan nilai kesamaan – terhadap kueri – yang lebih kecil darinya.

Dalam penelitian ini, nilai *threshold* yang digunakan adalah 0,3. Nilai ini dipilih sebab istilah-istilah dengan nilai kesamaan lebih dari 0,1 memiliki rataan sebesar 0,322. Tujuan pemilihan nilai ini (0,3) adalah agar lebih mendekati rataan tersebut.

3.5 Evaluasi Sistem Temu Kembali Informasi

Sistem dasar yang akan digunakan sebagai pembanding adalah sistem temu kembali hasil penelitian Ridha (2002). Sistem ini dikembangkan menggunakan *vector space model* (VSM) dengan pembobotan *tf-idf*.

Metode evaluasi yang umum digunakan untuk mengukur kinerja sistem temu kembali informasi adalah *recall* dan *precision*. Dalam hal ini, koleksi dokumen yang digunakan sudah memiliki gugus kueri dan gugus jawaban. Dalam penelitian ini, akan dibandingkan hasil temu kembali yang diperoleh, baik dengan menggunakan *similarity thesaurus* maupun tanpa menggunakan *similarity thesaurus*.

Dari hasil penemukembalian yang dilakukan sistem, dihitung *precision* pada berbagai tingkat *recall*. Tingkat *recall* yang digunakan adalah sebelas tingkat *recall* standar terinterpolasi. Hasilnya kemudian dirata-ratakan untuk mendapatkan *average precision* (AVP).

Untuk melihat pengaruh *similarity thesaurus* dan pengaruh pemberian bobot kueri dalam ekspansi kueri dilakukan pengukuran *average precision* terhadap enam jenis penemukembalian, yaitu:

- NoTH-1, yaitu penemukembalian tanpa ekspansi kueri dengan bobot istilah pada kueri adalah 1.
- NoTH-*idf*, yaitu penemukembalian tanpa ekspansi kueri dengan bobot istilah pada kueri adalah *idf*.
- TH5-1, yaitu penemukembalian dengan mengambil lima istilah teratas dimana bobot istilah pada kueri adalah 1.
- TH5-*idf*, yaitu penemukembalian dengan mengambil lima istilah teratas dimana bobot istilah pada kueri adalah *idf*.
- TH10-1, yaitu penemukembalian dengan mengambil sepuluh istilah teratas dimana bobot istilah pada kueri adalah 1.
- TH10-*idf*, yaitu penemukembalian dengan mengambil sepuluh istilah teratas dimana bobot istilah pada kueri adalah *idf*.

Keenam hasil penemukembalian di atas kemudian akan dibandingkan, baik berdasarkan *average precision* maupun *query-by-query*. Secara spesifik, uji banding yang akan dilakukan adalah:

- NoTH-1 – NoTH-*idf*, TH5-1 – TH5-*idf*, dan TH10-1 – TH10-*idf*: untuk melihat pengaruh pemberian bobot istilah dalam kueri terhadap hasil penemukembalian masing-masing metode.
- NoTH-1 – TH5-1 – TH10-1 dan NoTH-*idf* – TH5-*idf* – TH10-*idf*: untuk melihat pengaruh pemberian bobot istilah dalam kueri terhadap ekspansi kueri.

4. HASIL DAN PEMBAHASAN

4.1 Koleksi Pengujian

Dokumen-dokumen yang terdapat di dalam koleksi merupakan dokumen-dokumen berita yang diperoleh dari beberapa sumber di internet (Tabel 1). Dokumen-dokumen tersebut menyerupai dokumen XML dimana isi dokumennya dikelompokkan ke dalam *tag-tag* tertentu. *Tag-tag* yang digunakan telah menyesuaikan dengan standar TREC, yaitu:

- <doc></doc>
- <docno></docno>
- <title></title>
- <author></author>
- <date></date>
- <text></text>

Tabel 1Jumlah dokumen untuk masing-masing sumber berita.

Sumber	Jumlah
Balai Penelitian	12
Bitra Indonesia	1
Gatra	46
Indobic	4
Indosiar	68
Jurnal	26
Kompas	134
Lablink	1
Lapan	1
Media Indonesia	58
Pembaruan	1
Pikiran Rakyat	16
Poskota	9
Puslitbang	1
Republika	287
Situs Hijau	159
Suara Karya	43
Suara Merdeka	68
Suara Pembaruan	50
Tempo	1
Trubus	4
Warta Penelitian	10

4.2 Pengindeksan

Langkah pertama yang dilakukan adalah dengan melakukan pembobotan setiap istilah unik. Dalam hal ini, persamaan pembobotan yang digunakan adalah persamaan pembobotan *tf-idf*, seperti yang telah disampaikan sebelumnya. Adapun proses pengindeksannya adalah:

- *Tokenizing*, yaitu memecah suatu dokumen ke dalam *token-token*. Dalam proses ini, dilakukan pembuangan istilah-istilah yang termasuk ke dalam *stoplist* dan *stemming*.
- *Weighting*, yaitu menghitung bobot suatu *token* di dalam setiap dokumen di koleksi.

Secara keseluruhan, hasil pengindeksan dapat dilihat pada Tabel 2.

Tabel 2Hasil pengindeksan yang dilakukan.

Data	Koleksi
Jumlah dokumen	1.000
Ukuran dokumen (kb)	4.359
Rataan istilah tiap dokumen	172

Hasil pengindeksan yang dilakukan disimpan dalam tiga file teks, yaitu *token.txt*, *term.txt*, dan *index.txt*. Tujuannya adalah untuk mempercepat pemrosesan, baik saat pengindeksan maupun saat indeks tersebut digunakan. File *token.txt* berfungsi untuk menyimpan hasil *tokenizing* dan frekuensinya di dalam dokumen. File term.txt berfungsi untuk menyimpan nilai *idf* istilah. File *index.txt* berfungsi untuk menyimpan bobot istilah di dalam dokumen.

4.3 Tokenizing

Algoritma *tokenizing* yang digunakan adalah algoritma yang dibuat sendiri oleh penulis. Dalam melakukan *tokenizing*, secara automatis sistem akan membuang istilah-istilah yang termasuk di dalam *stoplist* dan melakukan *stemming*. Daftar istilah yang termasuk ke dalam *stoplist* dan algoritma *stemming* yang digunakan merupakan hasil penelitian Ridha (2002).

Sebelum melakukan *tokenizing*, terlebih dahulu ditetapkan beberapa karakter yang berperan sebagai karakter pemisah. Di antara karakter-karakter pemisah tersebut ada karakter yang berperan sebagai pemisah secara penuh (*full separator*) dan secara sebagian (*partial separator*). Pengertian dari *full separator* adalah sifat karakter tersebut sebagai karakter pemisah tidak dipengaruhi oleh konteks kalimat yang mengandungnya (posisinya terhadap karakter lain). Contoh : *'Space'*, *'Tab'*, *'_'*, dll. Sebaliknya, *partial separator* sangat dipengaruhi

oleh posisinya di dalam kalimat. Contoh : *'.'*, *','*, *'-'*, dll.

Dalam algoritma ini, semua karakter pemisah diasumsikan sebagai *full separator*. Termasuk di dalamnya adalah karakter-karakter numerik. Alasan dibuat seperti ini adalah karena beberapa *token* yang tidak sesuai untuk digunakan sebagai indeks karena menimbulkan ketidakkonsistenan dengan *token-token* lain yang dihasilkan. Misalnya : '00.00', 'apa,lagi'.

Hasil *tokenizing* menunjukkan adanya variasi jumlah *token* yang dihasilkan pada masing-masing dokumen di koleksi. Jumlah *token* yang dihasilkan bervariasi dari 1.161 hingga 24 *token*. Meskipun ada dokumen dengan jumlah *token* yang sangat banyak, namun rata-rata jumlah *token* per dokumen hanya sebesar 172 (Tabel 2). Hal ini menunjukkan bahwa terdapat banyak dokumen berukuran kecil.

4.4 Pembobotan (*Weighting*)

Telah disebutkan bahwa metode pembobotan yang digunakan adalah *tf-idf* dan *similarity thesaurus*. Pembobotan dilakukan setelah *tokenizing* dilakukan.

Tabel 3Jumlah istilah pada setiap rentang nilai bobot pada pembobotan *tf-idf*.

Bobot (w)	Jumlah Istilah
$0,0 \leq w < 0,3$	13.293
$0,3 \leq w < 0,6$	2.296
$0,6 \leq w < 0,9$	503
$0,9 \leq w < 1,2$	160
$1,2 \leq w < 1,5$	88
$1,5 \leq w < 1,8$	51
$1,8 \leq w < 2,1$	32
$2,1 \leq w < 2,4$	14
$2,4 \leq w < 2,7$	13
$2,7 \leq w \leq 3,0$	44

Dalam pembobotan *tf-idf*, nilai bobot minimum adalah 0,0184 ('tani') dan nilai bobot maksimum adalah 3. Pada Tabel 3 terlihat bahwa dengan meningkatnya bobot jumlah istilah cenderung semakin menurun. Dapat dilihat bahwa jumlah istilah yang berbobot rendah (<0,9) sebanyak 16.092 istilah (97,56%), yang berbobot sedang sebanyak 331 istilah (2,01%), dan yang berbobot tinggi (>2,1) sebanyak 71 istilah (0,43%).

Dalam pembobotan *similarity thesaurus*, nilai bobotnya berkisar dari 0 sampai 1 dimana nilai minimumnya adalah 0,0237 pada kata 'tani'. Dapat dilihat pada Tabel 4 bahwa jumlah istilah yang berbobot tinggi lebih besar daripada istilah yang berbobot rendah. Hal ini berbeda dengan hasil pembobotan *tf-idf* yang menunjukkan hasil yang sebaliknya.

Tabel 4Jumlah istilah pada setiap rentang nilai bobot pada pembobotan *similarity thesaurus*.

Bobot (w)	Jumlah Istilah
$0,0 \leq w < 0,35$	2.907
$0,35 \leq w < 0,7$	3.677
$0,7 \leq w \leq 1,0$	9.910

4.5 Matriks Kesamaan

Matriks kesamaan dibentuk dengan menghitung nilai kesamaan antar istilah dengan menggunakan ukuran *cosine*. Dalam hal ini, diperlukan bobot istilah yang diperoleh dari pengindeksan.

Dari hasil pengindeksan terhadap dokumen-dokumen di koleksi diperoleh jumlah istilah sebanyak 16.494. Tujuannya adalah untuk menyamakannya dengan gugus kueri yang tersedia yang mana semuanya merupakan karakter non-numerik.

Total jumlah elemen matriks yang terpakai adalah sebanyak 3.941.320 elemen (1,45% dari jumlah elemen yang seharusnya tersedia, yaitu 272.052.036 elemen). Penyebabnya adalah banyaknya elemen matriks yang dibuang karena nilainya nol. Hal ini bertujuan mengurangi ukuran penyimpanan matriks.

4.6 Ekspansi Kueri

Pada dasarnya, ekspansi kueri menggunakan *similarity thesaurus* dilakukan secara automatis saat pengguna memasukkan kueri ke dalam sistem. Untuk setiap istilah pada kueri yang dimasukkan, sistem akan mengambil bobot istilahnya dan nilai kesamaannya dengan istilah yang lain (2.5).

Pada proses pemilihan istilah, terlihat perbedaan yang mencolok antara pemberian bobot 1 dan *idf* terhadap istilah dalam kueri. Rata-rata bobot istilah hasil ekspansi meningkat untuk setiap gugus kueri saat *idf* digunakan sebagai bobot istilah dalam kueri. Dalam hal ini, bobot istilah pada kueri yang diekspansi tidak berpengaruh. Maksudnya adalah meskipun suatu istilah pada kueri awal berbobot rendah, istilah-

istilah hasil ekspansi pasti berbobot tinggi. Hal inilah yang mendorong peningkatan kinerja sistem temu kembali.

Telah dijelaskan sebelumnya bahwa suatu kueri akan berkorelasi dengan banyak istilah jika istilah-istilah pada kueri tersebut memiliki nilai *df* yang besar. Semakin besar df suatu istilah, maka semakin banyak calon istilah yang dapat digunakan dalam ekspansi kueri. Dapat dilihat bahwa istilah-istilah yang digunakan pada gugus kueri adalah istilah-istilah dengan *df* yang tinggi. Hal ini mengakibatkan pemilihan istilah sebanyak lima dan sepuluh istilah dari banyaknya istilah yang berkorelasi dengan kueri menjadi lebih rumit (waktu komputasinya lebih lama).

Pemberian bobot yang sama terhadap istilah dalam kueri mengakibatkan pemilihan istilah lebih dipengaruhi oleh nilai kesamaan antar istilah yang mana secara tidak langsung dipengaruhi juga oleh bobot istilah di dalam dokumen. Jika rata-rata bobot suatu istilah di dalam koleksi tinggi maka nilai kesamaannya dengan istilah lain juga tinggi. Sayangnya hal ini tidak sepenuhnya berlaku sebab sebagian besar istilah di kueri berbobot rendah.

Pemberian *idf* sebagai bobot istilah dalam kueri memberikan hasil yang lebih baik sebab setiap istilah yang digunakan dalam kueri diperlakukan secara berbeda. Dalam hal ini, istilah yang berbobot tinggi lebih berpengaruh dalam pemilihan istilah. Pengaruhnya adalah jika bobot istilah tersebut tinggi maka kemungkinan besar nilai kesamaannya dan *idf*-nya juga tinggi. Dengan demikian, istilah-istilah yang berkorelasi dengannya lebih diutamakan.

Pada proses pemilihan istilah ini ditemukan juga bahwa terdapat beberapa istilah yang memiliki nilai kesamaan yang sama dengan kueri. Hal ini sering terjadi pada istilah-istilah yang digunakan dalam ekspansi kueri. Pada kondisi yang demikian, istilah-istilah yang diambil akan disesuaikan dengan urutan abjad. Pada dasarnya, kondisi di atas tidak memberikan jaminan bahwa istilah dengan urutan abjad yang lebih kecil lebih baik daripada istilah dengan urutan abjad yang lebih besar. Hal ini dilakukan hanya bertujuan memudahkan pemilihan kata.

Nilai *threshold* yang sudah ditentukan sebelumnya ternyata tidak banyak berpengaruh. Penyebabnya adalah semua istilah-istilah yang digunakan pada ekspansi kueri, baik pada TH5 maupun TH10, memiliki nilai kesamaan yang melebihi nilai *threshold*. Nilai ini hanya bermanfaat untuk mengurangi ukuran matriks kesamaan.

4.7 Evaluasi Sistem Temu Kembali Informasi

Untuk menentukan baik atau tidaknya kinerja suatu sistem temu kembali informasi dilakukan uji banding dengan menggunakan 30 jenis kueri yang sudah ditentukan gugus jawabannya.

Dalam metode penelitian telah disebutkan bahwa evaluasi sistem temu kembali informasi akan dilakukan dalam dua uji banding. Uji banding yang pertama bertujuan melihat pengaruh pemberian bobot istilah dalam kueri terhadap hasil penemukembalian masing-masing metode. Uji banding yang kedua bertujuan melihat pengaruh pemberian bobot istilah pada kueri terhadap ekspansi kueri. Hasil kedua uji banding tersebut dapat dilihat pada Tabel 5 dan 6.

Tabel 5Nilai *average precision* berdasarkan pengaruh pemberian bobot istilah dalam kueri terhadap masing-masing metode.

Metode	AVP	% Perubahan
NoTH-1	0,3352	0 %
NoTH-*idf*	0,3856	15,06 %
TH5-1	0,3059	0 %
TH5-*idf*	0,3963	29,57 %
TH10-1	0,2703	0 %
TH10-*idf*	0,3983	47,33 %

Tabel 6 Nilai *average precision* berdasarkan pengaruh pemberian bobot istilah dalam kueri terhadap ekspansi kueri.

Metode	AVP	% Perubahan
NoTH-1	0,3352	0 %
TH5-1	0,3059	8,73 %
TH10-1	0,2703	19,35 %
NoTH-*idf*	0,3856	0 %
TH5-*idf*	0,3963	2,78 %
TH10-*idf*	0,3983	3,27 %

Tabel 5 menunjukkan bahwa pemberian *idf* istilah sebagai bobot istilah dalam kueri memberikan pengaruh yang lebih baik dibandingkan dengan memberikan bobot yang sama pada masing-masing istilah tersebut. Diketahui bahwa *idf* merepresentasikan *inverse* dari frekuensi munculnya istilah di koleksi. Hal ini berarti bahwa semakin banyak dokumen yang mengandung istilah tersebut semakin kecil nilai *idf*-nya. Misalnya, kata ‘tani’ (stem = ‘nani’) yang muncul di 921 dokumen memiliki nilai *idf*

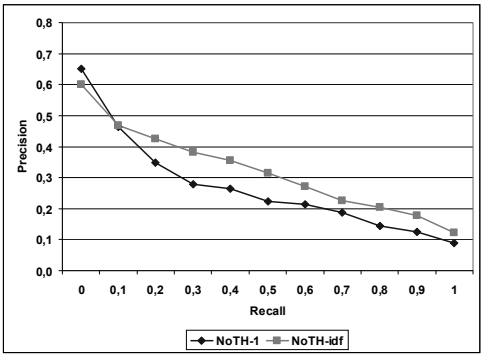
sebesar 0,036, sedangkan kata ‘panen’ (stem = ‘manen’) yang muncul di 254 dokumen memiliki nilai *idf* sebesar 0,595. Jelas bahwa untuk suatu kueri yang mengandung kata ‘tani’ dan ‘panen’ sistem akan lebih mengutamakan dokumen-dokumen yang mengandung kedua kata tersebut daripada dokumen yang hanya mengandung salah satunya saja.

Contoh di atas menunjukkan bahwa *idf* membatasi sistem dalam menemukembalikan dokumen-dokumen yang tidak relevan di awal penemukembalian. Jumlah dokumen yang ditemukembalikan untuk masing-masing kueri antara NoTH-1 dan NoTH-*idf* adalah sama. Namun, *average precision* yang dihasilkan tidak sama dimana NoTH-*idf* memberikan hasil yang lebih baik (Tabel 5). Dari hasil temu kembali untuk NoTH-1 – NoTH-*idf* hanya ada enam kueri yang mengakibatkan turunnya AVP.

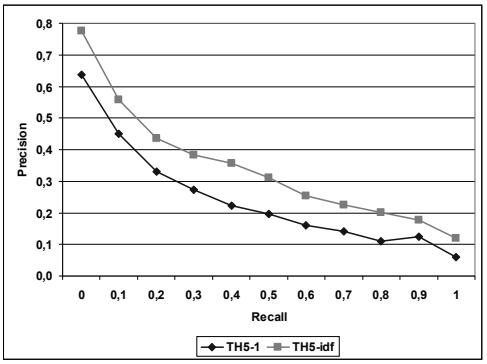
Kondisi yang muncul pada TH5-1 – TH5-*idf* dan TH10-1 – TH10-*idf* berbeda dengan sebelumnya dimana meskipun jumlah dokumen yang ditemukembalikan pada TH5-1 dan TH10-1 bertambah pada setiap kueri, AVP-nya justru menurun (Tabel 5). Contohnya kueri ‘tadah hujan’ pada TH5-1 meningkatkan jumlah dokumen hingga 190,52% dan pada TH10-1 kueri tersebut meningkatkan jumlah dokumen hingga 324,14% dari jumlah dokumen yang ditemukembalikan untuk kueri yang sama pada NoTH-1. Sebaliknya, untuk TH5-*idf* dan TH10-*idf*, meskipun jumlah dokumen yang ditemukembalikan pada setiap kueri rata-rata tidak bertambah, AVP-nya meningkat cukup besar (29,57% dan 47,33%) dari TH5-1 dan TH10-1 (Tabel 5). Hal ini membuktikan bahwa penggunaan *idf* sebagai bobot istilah dalam kueri lebih baik daripada suatu konstanta tertentu (bobot yang sama) dalam meningkatkan kinerja sistem dan tidak dipengaruhi oleh panjang kueri.

Secara komputasional, tidak ada perbedaan yang signifikan antara pemberian bobot yang sama (1) dengan bobot yang berbeda (*idf*), untuk setiap istilah di kueri. Namun, hal ini sangat mempengaruhi hasil temu kembali. Dapat diperhatikan pembobotan istilah di kueri dengan nilai *idf* lebih menunjukkan peningkatan khususnya untuk kueri dengan istilah-istilah yang berbobot rendah dengan rata-rata peningkatan sebesar 82% untuk NoTH-1 – NoTH-*idf*, 377,50% untuk TH5-1 – TH5-*idf*, dan 282,34% untuk TH10-1 – TH10-*idf*. Di lain pihak, untuk kueri dengan istilah-istilah berbobot tinggi rata-rata peningkatannya adalah sebesar 25% untuk NoTH-1 – NoTH-*idf*, 36,82% untuk TH5-1 – TH5-*idf*, dan 32,61% untuk TH10-1 – TH10-*idf*.

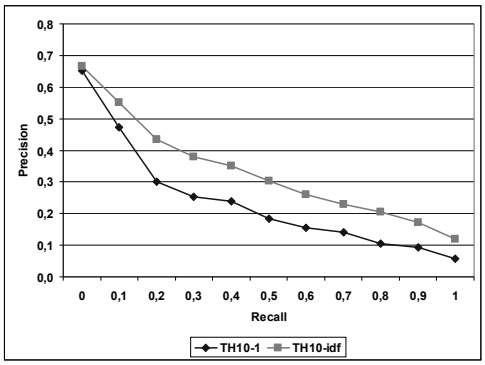
Visualisasi perbandingan antara masing-masing percobaan dengan bobot istilah dalam kueri yang berbeda dapat dilihat pada Gambar 2, 3, dan 4. Dari ketiga gambar tersebut, terlihat perbedaan yang cukup besar antara penggunaan 1 dan *idf* terhadap hasil temu kembali. Berdasarkan ketiga gambar tersebut, penggunaan *idf* sebagai bobot istilah di kueri benar-benar meningkatkan *precision* pada semua tingkat *recall*. Berdasarkan data yang diperoleh, rata-rata selisih *precision* antara setiap pasangan percobaan adalah 0,1 (khususnya untuk TH5-1 – TH5-*idf* dan TH10-1 – TH10-*idf*).



Gambar 2Kurva *recall-precision* untuk NoTH-1 dan NoTH-*idf*.

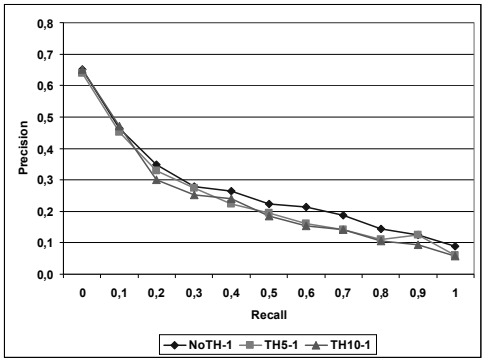


Gambar 3Kurva *recall-precision* untuk TH5-1 dan TH5-*idf*.

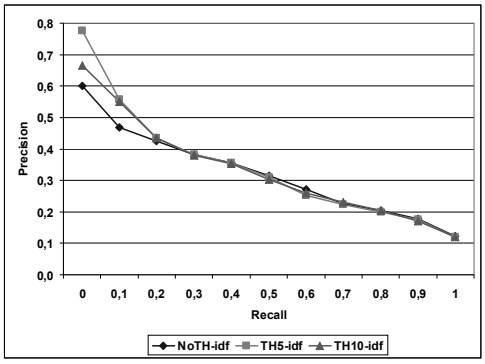


Gambar 4Kurva *recall-precision* untuk TH10-1 dan TH10-*idf*.

Dalam ekspansi kueri, metode pemberian bobot yang digunakan sangat mempengaruhi proses pemilihan istilah yang akan dipakai dalam melakukan ekspansi. Pada hasil uji banding yang kedua (Tabel 6), ekspansi kueri menghasilkan kinerja yang buruk saat nilai 1 dipakai sebagai bobot istilah dalam kueri. Sebaliknya, saat *idf* digunakan sebagai bobot istilah dalam kueri, ekspansi kueri meningkatkan kinerja sistem. Lebih jelasnya dapat dilihat pada Gambar 5 dan 6.



Gambar 5Kurva *recall-precision* untuk NoTH-1, TH5-1, dan TH10-1.



Gambar 6Kurva *recall-precision* untuk NoTH-*idf*, TH5-*idf*, dan TH10-*idf*.

Tabel 6 menunjukkan adanya penurunan AVP pada TH5-1 dan TH10-1 seiring dengan bertambahnya jumlah istilah yang dipakai saat ekspansi. Pada Gambar 5 terlihat jelas bahwa penurunan AVP tersebut terjadi pada semua tingkat *recall*. Jika dilihat AVP untuk masing-masing kueri, ekspansi kueri tersebut menurunkan AVP pada sekitar 19 kueri. Seperti yang sudah dijelaskan sebelumnya, penyebab menurunnya AVP adalah pemberian bobot yang sama pada istilah di dalam kueri dan banyaknya istilah di kueri yang berbobot rendah.

Dapat dilihat pada persamaan (2.5), pemilihan istilah untuk digunakan dalam ekspansi ditentukan oleh bobot istilah dalam kueri dan nilai kesamaan antara istilah tersebut dengan istilah yang lain. Jika bobot setiap istilah dalam kueri adalah sama, maka pemilihan istilah-istilah yang akan dipakai dalam ekspansi hanya ditentukan oleh nilai kesamaan antar masing-masing istilah dengan istilah di kueri. Dalam hal ini, hanya perlu dicari pasangan istilah dengan nilai kesamaan terbesar.

Pada persamaan (2.3), nilai kesamaan antar istilah ditentukan oleh bobotnya pada setiap dokumen di koleksi. Semakin besar bobot kedua istilah tersebut, semakin besar nilai kesamaannya. Dengan memakai *idf* sebagai bobot istilah dalam kueri, pemilihan istilah lebih diprioritaskan pada istilah-istilah dengan bobot yang lebih tinggi. Dengan kata lain, semakin rendah nilai *idf* atau bobot suatu istilah, semakin kecil kemungkinan istilah-istilah yang berkorelasi dengannya digunakan dalam proses ekspansi dan begitu juga sebaliknya.

Telah dijelaskan sebelumnya bahwa pemberian bobot yang sama untuk setiap istilah di kueri menghasilkan kinerja yang lebih rendah (Gambar 5) dibandingkan dengan pemberian *idf* sebagai bobot istilah di kueri (Gambar 6). Hal ini dapat dilihat juga pada Gambar 2, 3, dan 4 yang menunjukkan meskipun jumlah istilah di dalam kueri bertambah melalui ekspansi, sistem tetap memberikan hasil yang lebih baik untuk penggunaan *idf* sebagai bobot istilah dalam kueri (Tabel 6).

Tabel 6 juga menunjukkan bahwa adanya peningkatan AVP pada TH5-*idf* dan TH10-*idf*. Pada Gambar 6, terlihat bahwa peningkatan hanya terjadi pada R<20%. Hal ini menunjukkan bahwa semakin banyak dokumen relevan yang ditemukembalikan pada posisi yang lebih tinggi daripada hasil penemukembalian untuk NoTH-*idf*. Selain itu, rata-rata jumlah dokumen yang

ditemukembalikan untuk masing kueri pada NoTH-*idf*, TH5-*idf*, dan TH10-*idf* tidak berbeda jauh. Hal ini membuktikan bahwa pada pemakaian *idf* secara keseluruhan tidak menambah jumlah dokumen yang ditemukembalikan. Pemakaian *idf* hanya membatasi penemukembalian dokumen-dokumen yang tidak relevan.

Secara keseluruhan, meskipun ekspansi kueri seharusnya meningkatkan kinerja sistem, ternyata kondisi itu bergantung pada pemberian bobot istilah dalam kueri. Ternyata pemakaian *idf* sebagai bobot istilah dalam kueri menunjukkan hasil yang lebih baik, baik dari sisi penemukembalian maupun dari sisi ekspansi kueri.

5. KESIMPULAN

Berdasarkan hasil penelitian yang diperoleh dapat disimpulkan bahwa:

1.Ekspansi kueri berdasarkan *similarity thesaurus* dapat meningkatkan kinerja sistem temu kembali dengan memudahkan dan mempercepat pengguna dalam menemukan dokumen yang relevan. Peningkatan yang diperoleh mencapai 3,27%.

2.Pemberian bobot yang sama untuk setiap istilah pada kueri tidak meningkatkan kinerja sistem, khususnya pada saat ekspansi kueri. Penggunaan idf sebagai bobot istilah di kueri mampu meningkat kinerja sistem sampai dengan 47,33% saat ekspansi kueri.

3.Ekspansi kueri menghasilkan istilah-istilah yang berbobot tinggi saat *idf* dipakai sebagai bobot istilah di kueri. Sebaliknya, pemberian bobot yang sama menghasilkan istilah-istilah yang berbobot rendah saat ekspansi kueri.

4.Penentuan nilai *threshold* tidak berpengaruh secara signifikan dalam proses pemilihan kata.

DAFTAR PUSTAKA

[1] Adisantoso J, A. Ridha. 2004. *Corpus* Dokumen Teks Bahasa Indonesia untuk Pengujian Efektivitas Temu Kembali Informasi. Laporan Akhir Hibah Penelitian SP4, Departemen Ilmu Komputer FMIPA IPB, Bogor.

[2] Agusetyawan AW. 2006. Relevance Feedback pada Temu Kembali Teks Berbahasa Indonesia dengan Metode Ide-Dec-Hi dan Ide-Regular. Skripsi. Jurusan Ilmu Komputer IPB, Bogor.

[3] Baeza-Yates R, B. Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley, New York.

[4] Cui H, Ji-Rong W, Jian-Yu N, dan Wei-Ying M. 2003. Query Expansion by Mining User Logs. IEEE Trans. Knowledge Data Eng. 15(4):3-6.

[5] Prasetyo D. 1998. Penyusunan Thesaurus Menggunakan Analisis Gerombol (Cluster Analysis). Skripsi. Jurusan Ilmu Komputer IPB, Bogor.

[6] Qiu Y, H.P. Frei. 1993. *Applying a Similarity Thesaurus to a Large Collection for Information Retrieval*. Department of Computer Science Swiss Federal Institute of Technology (ETH), Zurich.

[7] Ridha A. 2002. Pengindeksan Otomatis dengan Istilah Tunggal untuk Dokumen Berbahasa Indonesia. Skripsi. Jurusan Ilmu Komputer IPB, Bogor.

[8] Tombros A. 2002. *The Effectiveness Of Query-Based Hierarchic Clustering of Documents for Information Retrieval*. Tesis. Department of Computing Science Faculty of Computing Science, Mathematics and Statistics University of Glasgow.

[9] Van Rijsbergen CJ. 1979. *Information Retrieval*. http://www.dcs.gla.ac.uk/Keith/pdf/index.html [15 Juli 2005]

[10] Wandari FA. 2005. Evaluasi Stemmer Berbasis Bahasa Indonesia dengan dan tanpa Menggunakan Kamus Kata Dasar. Skripsi. Jurusan Ilmu Komputer IPB, Bogor.

.