EFEKTIFITAS SELEKSI FITUR DALAM SISTEM TEMU-KEMBALI INFORMASI

Indra Budi dan Rizal Fathoni Aji
Fakultas Ilmu Komputer, Universitas Indonesia
Kampus UI Depok 16424 Indonesia
E-mail: {indra, rizal}@cs.ui.ac.id

ABSTRAKSI

Paper ini mendeskripsikan laporan uji coba penggunaan feature selection (seleksi fitur) pada Information Retrieval System (sistem temu-kembali informasi). Istilah atau index-term yang menjadi indeks merupakan fitur yang diseleksi. Seleksi dilakukan dengan mengurangi istilah pada indeks sebanyak 5%, 10%, 15% dan 20%. Uji coba dilakukan dengan tiga strategi, mengurangi index-term dengan frekuensi besar dan kecil sekaligus, mengurangi index-term dengan frekuensi besar saja dan mengurangi index-term dengan frekuensi kecil saja. Uji coba memperlihatkan bahwa pengurangan index-term dengan frekuensi kecil saja lebih baik dibandingkan dua strategi yang lain.

Kata kunci: *feature selection, information retrieval system, index-term.*

1. PENDAHULUAN

Indeks merupakan representasi dokumen yang dapat menentukan isi dari dokumen. Baik tidaknya sebuah indeks sangat bergantung kepada sejauh mana istilah-istilah yang dipilih menjadi indeks dapat merepresentasikan isi dari dokumen tersebut. Jika dua dokumen berbeda, maka seharusnya tidak terambil bersamaan jika diberikan suatu *query*, sebab istilah-istilah yang dipilih sebagai indeks dapat membedakan kedua dokumen tersebut [4,5].

Pemilihan istilah untuk dijadikan indeks merupakan isu yang penting dalam sistem temu-kembali informasi. Selanjunya proses pemilihan istilah ini disebut dengan seleksi fitur (*feature selection*). Fitur seleksi dapat menyebabkan berkurangnya ukuran indeks sehingga proses *retrieval* suatu dokumen menjadi lebih cepat sebab jumlah indeks yang dicari menjadi lebih sedikit [7].

Tugas utama seleksi fitur adalah menentukan istilah-istilah yang layak dijadikan *term index* atau dengan kata lain membuang (menghilangkan) istilah-istilah yang tidak mungkin dijadikan indeks. Terdapat beberapa cara yang dapat dilakukan untuk mengeliminasi istilah-istilah yang kurang merepresentasikan dokumen tersebut, diantaranya adalah menghilangkan istilah-istilah yang sering muncul pada berbagai dokumen [8]. Istilah-istilah yang sering muncul pada berbagai dokumen biasanya adalah istilah-istilah yang tidak mempunyai arti terhadap dokumen tersebut, jika istilah ini dihilangkan, tidak mengurangi makna dokumennya. Kata sambung seperti *dan, atau* dan *juga* merupakan contoh dari kaat sambung [1].

Disamping istilah yang sering muncul, istilah-istilah yang jarang muncul, atau hanya muncul satu atau dua kali pada satu dokumen juga merupakan kandidat istilah yang dapat dihilangkan. Penelitian ini mencoba untuk meneliti sejauh mana seleksi fitur tersebut berpengaruh terhadap kinerja sistem temu-kembali informasi.

Selanjutnya tulisan ini disusun dengan sistematika sebagai berikut. Landasan teori yang berkaitan dengan pengindeksan dan seleksi fitur didiskusikan pada bagian kedua berikut. Penjelasan mengenai bagaimana karakterisitik dari sistem yang digunakan dijabarkan pada bagian ketiga. Bagian keempat menjelaskan lingkungan dan hasil uji coba sedangkan kesimpulan dijelaskan pada bagian kelima.

2. PENGINDEKSAN DAN SELEKSI FITUR

Sistem temu kembali yang memiliki kinerja baik sangat diperlukan terutama untuk menghadapi perkembangan yang sangat pesat dari dokumen khususnya dokumen berbasis teks seperti laporan penelitian, artikel, skripsi, tesis, dan sebagainya. Sistem temu-kembali informasi adalah suatu sistem yang mengolah data berbasis dokumen atau teks dalam jumlah besar dan memberikan dokumen-dokumen sesuai dengan *query* yang diberikan pemakai.

Hal-hal yang dilakukan oleh suatu sistem temu-kembali informasi diantaranya adalah sebagai berikut [5]:

1. Mengolah *record-record* berupa teks atau dokumen, yaitu mengidentifikasikan sejumlah istilah yang dianggap mewakili isi dokumen.
2. Mengidentifikasikan permintaan informasi (*information request / query*)
3. Menentukan dan mengambil informasi atau dokumen yang sesuai dengan permintaan pemakai.

Tahapan pertama tersebut dikenal dengan pengindeksan. Pengindeksan merupakan cara untuk mendapatkan istilah-istilah yang dianggap mewakili isi dari dokumen.

Pengindeksan dapat dilakukan secara manual atau otomatis. Jika dengan cara manual, dibutuhkan campur tangan seorang manusia yang dikenal dengan *indexer* yang bertugas untuk memlilih istilah-istilah yang terdapat pada dokumen untuk dijadikan *index term* yang merepresentasikan dokumen tersebut. Sedangkan pada pengindeksan yang dilakukan secara otomatis, pemilihan *term index* dilakukan secara otomatis menggunakan program komputer [3].

Ketepatan pemilihan istilah merupakan isu yang menentukan kinerja dari sistem yang dihasilkan. Pada dasarnya setiap kata yang muncul pada dokumen dapat dijadikan *index term*. Namun jika semua kata dijadikan index term, disamping ukuran indeks menjadi besar, belum tentu kata-kata tersebut merepresentasikan isi dokumen [2].

Secara umum, istilah-istilah yang sering muncul pada banyak dokumen tidak layak dijadikan indeks, seperti kata sambung *dan, atau, juga*, dsb. Kata-kata tersebut tidak layak jika dijadikan sebagai *index term* karena [1,8]:

- Mereka muncul sangat sering pada dokumen, bahkan semua dokumen memiliki kata-kata tersebut.
- Kata-kata tersebut tidak menggambarkan isi dari dokumen yang bersangkutan.

Satu lagi jenis kata yang juga kurang baik dijadikan indeks, adalah kata-kata yang jarang muncul, muncul hanya sekali atau dua kali pada dokumen tertentu [1]. Jika kata-kata seperti ini dijadikan indeks maka sangat sedikit dokumen yang akan terambil. Apalagi jika kata-kata tersebut dijadikan dalam satu *query* secara bersamaan maka kemungkinan tidak ada dokumen yang terambil.

Dengan melihat kondisi tersebut diatas maka disarankan menggunakan dua *threshold* (nilai ambang), yaitu untuk menentukan batas atas dimana nilai frekuensi tertinggi dari istilah yang diperbolehkan dan batas bawah untuk menentukan nilai frekuensi terendah. Proses ini disebut dengan seleksi fitur [7, 8].

Diharapkan dengan seleksi fitur dapat mengurangi istilah-istilah yang tidak berpotensi menjadi indeks, sekaligus mengurangi ukuran indeks sehingga mempercepat proses pencarian. Namun diharapkan dengan adanya pengurangan istilah tersebut tidak mengurangi kinerja sistem, atau paling tidak sama dengan kinerja sistem tanpa seleksi fitur. Sehingga dalam penelitian ini dilakukan uji coba untuk melihat sejauh mana efektifitas pengurangan *index-term* pada ketiga strategi diatas terhadap kinerja sistem temu-kembali informasi.

3. KARAKTERISTIK SISTEM

Sistem yang dikembangkan pada penelitian ini menggunakan model *vector space* dimana nilai bobot setiap istilah pada suatu dokumen dihitung dengan menggunakan rumusan sebagai berikut:

$$w_{ik} = tf_{ik} * \log\left[\frac{n}{df_k}\right]$$

dimana:

- w_{ik} adalah bobot istilah k pada dokumen i.
- tf_{ik} Merupakan frekeunsi dari istilah k dalam dokumen i.
- n adalah jumlah dokumen dalam kumpulan dokumen.
- df_k adalah jumlah dokumen yang mengandung istilah k.

Sedangkan seleksi fitur yang digunakan adalah menghilangkan sejumlah istilah dengan frekuensi kemunculan terbesar dan menghilangkan istilah dengan frekuensi kemunculan terkecil. Misalnya untuk seleksi fitur sebesar 5%, artinya menghilangkan 2,5% istilah yang mempunyai frekuensi terbesar dan 2,5% istilah yang mempunyai frekuensi terkecil.

Similarity (kesamaan) antara *query* dengan dokumen adalah dengan menggunakan rumusan *dot product* sebagai berikut [5]:

$$Q.D = q1*d2 + q1*d2 + \ldots + qn*dn$$

dimana:
Q = vektor *query*
D = vektor dokumen
n = jumlah istilah

4. UJI COBA

Pada bagian uji coba ini dijelaskan mengenai karakteristik dokumen uji coba, evaluasi kinerja sistem dan hasil uji coba serta analisisnya.

4.1 Dokumen Uji Coba

Koleksi dokumen yang dipergunakan untuk uji coba meliputi 524 dokumen yang diperoleh dari surat kabar Suara Merdeka Online (www.suaramerdeka.com) pada tanggal 2 Januari 2002 sampai dengan 11 Januari 2002. Dokumen-dokumen tersebut merupakan dokumen berbahasa Indonesia yang dikelompokkan dalam delapan kategori yaitu: budaya, daerah ekonomi, internasional, nasional, olahraga, semarang dan solo. Jumlah term yang dihasilkan dari proses *parsing* terhadap koleksi dokumen sebanyak 17.473 *term* tanpa melakukan proses *stemming* dan pembuangan terhadap *stop words* standar. Sedangkan kueri berjumlah 125 dokumen yang didapatkan dari 524 koleksi dokumen tersebut. Rincian koleksi dokumen yang digunakan dapat dilihat pada Tabel 1.

Tabel 1. Jumlah dokumen per kategori

Kategori	***Jumlah***
budaya	28
daerah	118
ekonomi	25
internasional	24
nasional	107
olahraga	75
semarang	96
solo	51
Total	524

4.2 Evaluasi Kinerja

Evaluasi standar yang dilakukan untuk mengetahui kinerja sistem temu-kembali informasi mengacu pada tiga nilai parameter berikut, yaitu:

Recall = A/B
Precision = A/C
*F-Measure = (2 * Recall * Precision) / (Recall + Precision)*

dimana:

A = jumlah dokumen relevan yang terambil oleh sistem
B = jumlah dokumen relevan yang terdapat pada koleksi dokumen
C = jumlah dokumen yang terambil

Penilaian relevansi dari dokumen yang terambil dilakukan dengan cara membandingkan kategori dari dokumen kueri dengan kategori dari dokumen yang terambil. Dokumen yang terambil dikatakan **relevan** apabila memiliki kategori yang sama dengan dokumen kueri.

Sedangkan untuk klasifikasi teks (dokumen), secara umum digunakan pengukuran *microaveraging*, yaitu dengan menghitung nilai *recall*, *precision* dan *F-Measure* untuk semua kueri sekaligus, tidak untuk setiap kueri [6].

Penghitungan *recall* dan *precision* menggunakan *microaveraging* dijelaskan sebagai berikut. Misalnya untuk setiap kueri didapatkan nilai A, B dan C-nya, kemudian nilai-nilai tersebut dijumlahkan seperti terlihat pada Tabel 2.

Tabel 2. Penghitungan *microaveraging*

Kueri	***A***	***B***	***C***
q_1	a_1	b_1	c_1
q_2	a_2	b_2	c_2
: :	: :	: :	: :
q_{125}	a_{125}	b_{125}	c_{125}
Total	**TA**	**TB**	**TC**

Sehingga didapatkan nilai *recall*, *precision* dan *F-Measure* menggunakan *microaveraging* adalah:

Recall = TA/TB
Precision = TA/TC
*F-Measure = (2 * Recall * Precision) / (Recall + Precision)*

dimana:

$TA = \sum_{i=1}^{125} a_i$, $TB = \sum_{i=1}^{125} b_i$ dan $TC = \sum_{i=1}^{125} c_i$

Nilai 125 merupakan jumlah kueri yang diujicobakan. Sedangkan nilai a_i, b_i, dan c_i adalah jumlah dokumen relevan terambil, jumlah dokumen relevan pada koleksi dan jumlah dokumen terambil pada kueri ke-i.

4.3 Hasil Uji Coba dan Pembahasan

Uji coba dilakukan dengan tiga strategi yaitu:

1. Membuang *term-term* dengan frekuensi besar dan kecil sekaligus
2. Membuang *term-term* yang memiliki frekuensi besar
3. Membuang *term-term* yang memiliki frekuensi kecil.

Hasil uji coba dapat dilihat pada Tabel 3, Tabel 4 dan Tabel 5.

Tabel 3. Ujicoba dengan menghilangkan term-term dengan frekuensi besar dan kecil

Treshold	***Recall***	***Precision***	***F-Measure***
0%	0.952	0.173	0.293
5%	0.827	0.182	0.298
10%	0.599	0.196	0.295
15%	0.449	0.216	0.292
20%	0.330	0.229	0.270

Tabel 4. Ujicoba dengan menghilangkan term-term dengan frekuensi besar

Treshold	***Recall***	***Precision***	***F-Measure***
0%	0.952	0.173	0.293
5%	0.599	0.195	0.294
10%	0.330	0.229	0.271
15%	0.202	0.267	0.230
20%	0.125	0.314	0.179

Tabel 5. Ujicoba dengan menghilangkan term-term dengan frekuensi kecil

Treshold	***Recall***	***Precision***	***F-Measure***
0%	0.952	0.173	0.293
5%	0.952	0.173	0.293
10%	0.952	0.173	0.293
15%	0.952	0.173	0.293
20%	0.952	0.174	0.294

Berdasarkan Tabel 3 dapat terlihat bahwa nilai *recall* cenderung menurun dan nilai *precision* tidak dapat dipastikan. Namun secara umum nilai *F-Measure* cenderung menurun jika dilakukan pengurangan *index term* yang memiliki frekuensi besar dan kecil sekaligus.

Hal yang sama juga terjadi apabila dilakukan pengurangan term-term dengan frekuensi besar saja, seperti terlihat pada Tabel 4. Sedangkan Tabel 5, memperlihatkan bahwa nilai *recall*, *precision* dan *F-Measure* cenderung konstan (stabil).

Secara umum, pemotongan *index-term* yang mempunyai frekuensi kecil lebih baik dari pada strategi lainnya. Hal ini mungkin terjadi karena *index-term* dengan frekuensi kecil sangat banyak

jumlahnya dan tidak relevan dengan keseluruhan isi dokumen.

5. KESIMPULAN

Feature selection dengan mengurangi *index-term* sangat mungkin untuk digunakan dalam sistem temu-kembali informasi sebagai cara untuk optimisasi indeks. Berdasarkan uji coba yang dilakukan, efektifitas sistem sistem temu-kembali informasi tidak banyak berubah walaupun dilakukan pemotongan sampai 20%.

Dari hasil uji coba dapat disimpulkan bahwa pemotongan *index-term* dengan frekuensi kecil lebih baik dibandingkan yang lain. Namun, hal ini akan tergantung dari kumpulan koleksi yang ada dalam indeks. Dalam uji coba ini, koleksi dokumen mempunyai indeks yang mengandung *index-term* dengan frekuensi kecil lebih banyak, sehingga lebih efektif jika pemotongan dilakukan terhadap *index-term* ini. Jika suatu indeks mengandung lebih banyak *index-term* dengan frekuensi besar, maka pemotongan akan lebih efektif jika dilakukan terhadap *index-term* tersebut.

REFERENSI

[1] Korfhage, Robert R., *Information Storage and Retrieval*, Jhon Wiley & Sons, Inc, New York, 1997.

[2] Luk, Robert, WP, et al.2002. A survey in Indexing and Searching XML Documents, *Journal of the American Society for Information Science and technology*, Vol. 53 No. 6 April 2002.

[3] Meulen, W. A. Van Der and P.J.F.C. Janssen. 1977. Automatic versus Manual Indexing. *Information Processing & Management*, Vol. 13 pp. 13-21.

[4] Ricardo Baeza-Yates, Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, London, 1999.

[5] Salton, Gerard., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley Publishing Company, New York, 1989.

[6] Sebastiani, F., Machine learning in automated text categorization. *ACM Computing Surveys, 34*(1), 1-47, 2002.

[7] Wibowo W., Williams HE, Simple and Accurate Feature Selection for Hierarchical Categorization, *DocEng*, November 2002.

[8] Yang Y, Pedersen JO. A Comparative Study on Feature Selection in Text Categorization, *Proceedings of {ICML}-97, 14th International Conference on Machine Learning*, 1997.