EVALUASI KINERJA SISTEM PENYARINGAN INFORMASI MODEL RUANG VEKTOR

Rila Mandala
Kelompok Keahlian Informatika, Sekolah Teknik Elektro dan Informatika, Institut Teknologi Bandung
Jalan Ganesha 10 Bandung, Indonesia
E-mail: rila@if.itb.ac.id

ABSTRAKSI

Penyaringan Informasi (Information Filtering) adalah sebuah kajian mengenai pemfilteran suatu aliran informasi dinamis dalam volume besar dan menyampaikannya kepada pengguna tertentu untuk memenuhi kebutuhan informasinya. Kajian ini berkerja pada informasi tidak terstruktur. Aliran informasi yang dikaji merupakan sekumpulan dokumen mengenai berbagai topik dan kebutuhan informasi dinyatakan dalam bentuk profil pengguna. Makalah ini membahas sistem Information Filtering yang menggunakan metode Model Ruang Vektor untuk memodelkan dokumen dan profil pengguna. Tiap dokumen dan profil akan direpresentasikan dalam bentuk vektor-vektor. Elemen vektor adalah term-term yang terdapat pada dokumen dan profil dan dilengkapi dengan bobot masing-masing term. Evaluasi performansi dilakukan dengan menggunakan sebuah koleksi dokumen sebagai aliran informasi dan sebuah kumpulan profil sebagai pemfilter. Dengan mengamati hasil pemfilteran, dokumen yang disampaikan sistem kepada pengguna cukup sesuai dengan kebutuhan informasinya. Performansi lebih lanjut diukur dengan kalkulasi nilai Recall dan Precision dengan menggunakan data penilaian manusia akan relevansi dokumen terhadap profil.

Kata kunci: *information filtering, model ruang vektor, Evaluasi Performansi.*

1. PENDAHULUAN

Dalam kehidupan masyarakat saat ini, informasi telah menjadi kebutuhan yang sangat penting. Informasi menjadi salah satu sumber daya yang dapat memberikan nilai tambah bagi pemilik informasi tersebut. Dengan adanya internet, penyebaran informasi semakin cepat dan murah. Siapa saja dapat dengan mudah mengakses informasi yang ditampilkan dalam bentuk halaman-halaman *web* maupun e-mail.

Namun semakin besarnya jumlah informasi yang terdistribusi membuat pengguna sulit menemukan dan memanfaatkan informasi yang betul-betul dibutuhkannya. Hal tersebut juga disebabkan oleh kurangnya perangkat yang dapat dimanfaatkan pengguna untuk mengelola aliran informasi tersebut. Oleh karena itu perlu adanya suatu mekanisme yang membantu pengguna dalam mendapatkan informasi mengenai topik-topik yang sesuai dengan kebutuhannya.

Information Filtering (IF) adalah salah satu metode yang secara cepat berkembang untuk mengelola aliran informasi yang datang kepada pengguna. Tujuan dari *Information Filtering* adalah membawa pengguna kepada hanya informasi yang relevan terhadap kebutuhan mereka. Sistem IF telah dikembangkan beberapa tahun terakhir ini untuk berbagai domain aplikasi. Beberapa contoh dari aplikasi pemfilteran adalah pemfilteran e-mail personal berdasarkan profil personal, pemfilter *browser* yang memblok informasi yang tidak sesuai, filter yang dirancang agar anak-anak hanya dapat mengakses informasi yang sesuai bagi mereka, dan lain-lain. Secara umum tujuan utama dari IF adalah mengarahkan informasi yang paling berharga (relevan) kepada pengguna secara otomatis dan membantu pengguna memanfaatkan waktu membaca dokumen yang terbatas secara lebih optimal.

Information filtering, seperti juga *Information Retrieval* (sistem temu balik informasi), menangani ruang informasi tidak terstruktur dan kebutuhan pengguna akan informasi spesifik[WON97]. Sistem temu balik informasi menangani ruang informasi yang stabil dan kebutuhan pengguna akan informasi yang bervariasi / dinamis, sedangkan *Information Filtering* menangani ruang informasi yang dinamis dan kebutuhan pengguna akan informasi yang relatif stabil. Permasalahan dalam *Information Filtering* dengan demikian dapat dinyatakan sebagai berikut: terhadap sejumlah objek informasi dinamis, sistem *Information Filtering* mencocokkan karakteristik dari objek informasi tersebut dengan profil pengguna, yaitu deskripsi dari kebutuhan informasi pengguna, untuk mendapatkan perkiraan relevansi antara objek informasi tersebut terhadap kebutuhan informasi. Sistem profil akan menunjukkan ketertarikan dan pilihan pengguna, dan penggunaannya membantu pengguna untuk melakukan akses terkendali terhadap bagian yang relevan dari informasi[WON97]. Sistem profil tersebut akan bertindak sebagai intermediator antara pengguna dan objek informasi. Dalam penentuan relevansi tersebut, pendekatan yang digunakan adalah pendekatan *binary classification system* dimana dokumen hanya akan diklasifikasikan sebagai dokumen relevan atau tidak, dan tidak dilakukan pengurutan peringkat (*ranking*)[LEW96].

2. KONSEP *INFORMATION FILTERING*

Bahasa adalah bentuk yang memungkinkan manusia mengekspresikan ide, dan teks adalah bentuk dimana bahasa paling mudah ditangani oleh sistem komputer saat ini. Informasi dalam jumlah besar disimpan dalam bentuk teks.

Information Filtering adalah sebuah kajian mengenai pemfilteran suatu aliran informasi dinamis dalam volume besar dan menyampaikannya kepada pengguna tertentu untuk memenuhi kebutuhan informasinya[OAR96].

Kajian mengenai perolehan informasi yang telah banyak dikenal adalah Temu Balik Informasi (*Information Retrieval*). Kemudian kajian mengenai *Information Filtering*, yang bersama kajian temu balik informasi disebut sebagai "keping mata uang dengan dua sisi", mulai berkembang.

Information filtering, seperti juga *Information Retrieval* (Temu Balik Informasi), menangani ruang informasi tidak terstruktur dan kebutuhan pengguna akan informasi spesifik[WON97]. Hal ini yang membedakan keduanya dengan konsep basis data yang menangani *query* dan kumpulan data berstruktur tertentu. Kajian Temu Balik Informasi menangani ruang informasi yang stabil dan kebutuhan pengguna akan informasi yang bervariasi/dinamis, sedangkan *Information Filtering* menangani ruang informasi yang dinamis dan kebutuhan pengguna akan informasi yang relatif stabil.

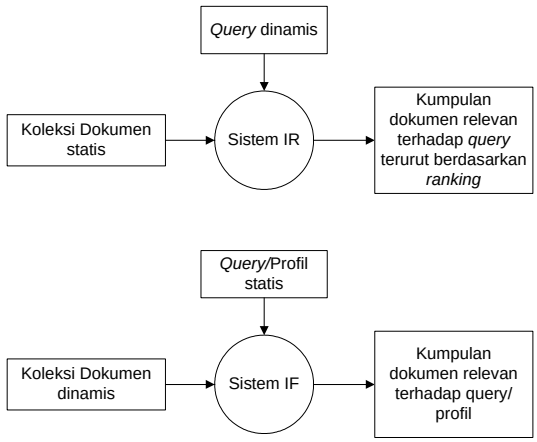
Permasalahan dalam *Information Filtering* dengan demikian dapat dinyatakan sebagai berikut: terhadap sejumlah objek informasi dinamis, sistem *Information Filtering* mencocokkan karakteristik dari objek informasi tersebut dengan profil pengguna, yaitu deskripsi dari kebutuhan informasi pengguna, untuk mendapatkan perkiraan relevansi antara objek informasi tersebut terhadap kebutuhan informasi.

Sistem profil akan menunjukkan ketertarikan dan pilihan pengguna, dan penggunaannya membantu pengguna untuk melakukan akses terkendali terhadap bagian yang relevan dari informasi[WON97]. Sistem profil tersebut akan bertindak sebagai intermediator antara pengguna dan objek informasi. Dalam penentuan relevansi tersebut, pendekatan yang digunakan adalah pendekatan *binary classification system* dimana dokumen hanya akan diklasifikasikan sebagai dokumen relevan atau tidak, dan tidak dilakukan pengurutan peringkat (ranking)[LEW96].

Setiap pendekatan yang dilakukan untuk *Information Filtering* maupun *Information Retrieval* memiliki empat komponen dasar:

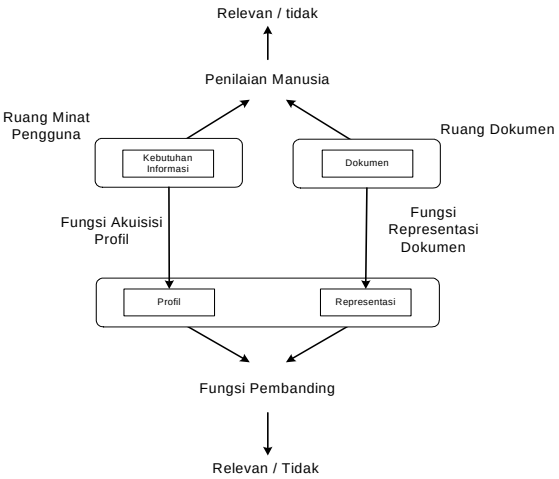
1. Teknik untuk merepresentasikan dokumen
2. Teknik untuk merepresentasikan kebutuhan informasi (misalnya: konstruksi profil)
3. Cara untuk membandingkan profil dengan representasi dokumen
4. Cara untuk menggunakan hasil dari pembandingan tersebut

Perbedaan sistem *Information Filtering* dan *Information Retrieval* pada umumnya dapat dilihat pada **Gambar 1**.



Gambar 1. Perbandingan Sistem *Information Filtering* dengan *Information Retrieval*

Objektif dari konsep yang dikaji adalah untuk mengotomasi proses pemeriksaan dokumen dengan melakukan komputasi perbandingan antara representasi kebutuhan (profil) dengan representasi dokumen. Proses yang diotomasi ini disebut berhasil apabila memberikan hasil yang mirip dengan hasil perbandingan secara manual oleh manusia. Model sistem *Information Filtering* dapat digambarkan seperti pada **Gambar 2**.



Gambar 2. Model Sistem *Information Filtering*

3. MODEL RUANG VEKTOR

Model Ruang Vektor merupakan suatu metode yang cukup banyak digunakan dalam sistem *Information Retrieval*. Namun karena konsep *Information Filtering* memiliki prinsip yang mirip dengan *Information Retrieval*, maka Model Ruang Vektor ini juga dapat diimplementasikan pada aplikasi *Information Filtering*. Kakas-kakas

 ISSN: 1907-5022

pendukung Model Ruang Vektor seperti untuk *stemming* juga tetap dapat dimanfaatkan.

Dengan menggunakan Model Ruang Vektor, dokumen-dokumen yang ada dan profil pengguna akan dipetakan menjadi n dimensi vektor[GRO98]. Profil pengguna akan menjadi *query* yang relatif tetap bagi dokumen-dokumen yang berubah secara dinamis.

Banyaknya dimensi dari ruang vektor akan ditentukan oleh jumlah kata signifikan yang terdapat dalam dokumen ataupun profil/*query*. Diasumsikan terdapat *t* buah term yang berbeda, yang disebut *vocabulary* atau *term index,* pada koleksi dokumen dan profil. Kumpulan term tersebut akan dibentuk menjadi sebuah ruang vektor dengan dimensi = *t* = |*vocabulary*|.

Tiap dokumen atau profil akan direpresentasikan sebagai sebuah vektor. Panjang dari tiap vektor akan bervariasi sesuai dengan bobot dari masing-masing term yang menjadi elemen vektor. Setiap term *i*, dalam dokumen atau profil, *j,* diberi bobot bertipe real, w_{ij}. Representasi dokumen dan profil dalam bentuk vektor berdimensi *t* digambarkan seperti pada **Gambar 3** dan dengan representasi matematis:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj}) \quad (1)$$

contoh:
$D1 = 2T1 + 3T2 + 5T3$
$D2 = 3T1 + 7T2 + T3$
$Q = 0T1 + 0T2 + 2T3$

Koleksi dari *n* buah dokumen dapat direpresentasikan dalam Model Ruang Vektor dengan sebuah matriks term-dokumen sebagai berikut:

$$\begin{pmatrix} & T_1 & T_2 & .... & T_t \\ D_1 & w_{11} & w_{21} & \ldots & w_{t1} \\ D_2 & w_{12} & w_{22} & \ldots & w_{t2} \\ : & : & : & & : \\ : & : & : & & : \\ D_n & w_{1n} & w_{2n} & \ldots & w_{tn} \end{pmatrix}$$

Gambar 3. Representasi Vektor dokumen dan query

Sebuah entri dalam matriks term-dokumen berkorespondensi dengan bobot sebuah term dalam dokumen tertentu. Nilai 0 berarti keberadaan term tersebut tidak signifikan dalam dokumen atau term tersebut tidak terdapat dalam dokumen.

Pembobotan term untuk Model Ruang Vektor dapat dilakukan dengan konsep *Term Frequency* (*Tf*) dan *Inverse Document Frequency* (*Idf*).

Term frequency suatu term adalah jumlah term tersebut dapat ditemukan dalam dokumen atau profil. Semakin banyak suatu term ditemukan dalam dokumen berarti term tersebut semakin penting dan semakin indikatif terhadap suatu topik.

Document Frequency suatu term adalah banyaknya dokumen dalam koleksi, yang mengandung kemunculan term tersebut. *Inverse Document Frequency* suatu term adalah invers dari *Df*, yaitu 1/*Df*. Nilai *Idf* suatu term yang tinggi menunjukkan bahwa term tersebut penting/signifikan karena tidak banyak dipakai di dokumen-dokumen dalam koleksi. Sebaliknya nilai *Idf* suatu term yang rendah menunjukkan bahwa term tersebut tidak penting karena banyak dipakai di dokumen-dokumen dalam koleksi.

Sebuah term yang sering muncul dalam sebuah dokumen, namun jarang muncul di sisa dokumen dalam koleksi, akan memperoleh bobot yang tinggi.

Profil/*query* diperlakukan sama seperti dokumen dalam pembentukan vektor dan diberi bobot dengan pembobotan *Tf-Idf* yang sama.

Relevansi antara *query* dengan suatu dokumen (dikenal juga dengan istilah similaritas) adalah suatu ukuran kemiripan antara representasi dokumen dengan representasi kebutuhan pengguna (profil). Ukuran kemiripan (*Similarity Measure*) adalah sebuah fungsi untuk menentukan derajat kemiripan (*degree of similarity*) antara dua buah vektor. Penggunaan ukuran kemiripan antara *query* dan dokumen-dokumen memungkinkan:

1. Perankingan dokumen-dokumen yang diperoleh sesuai urutan relevansi
2. Pembuatan nilai ambang (*threshold*) tertentu agar ukuran himpunan dokumen yang diperoleh dapat dikendalikan.

Ukuran kemiripan yang sering digunakan adalah *dot product* dan *Cosine Similarity Measure*. Ukuran kemiripan dengan *dot product* adalah perhitungan yang cukup sederhana dalam menentukan similaritas suatu dokumen dengan *query*. *Dot product* memberikan tingkat kemiripan antara vektor dokumen d_i dan *query q* dengan menghitung *dot product* antara dua vektor:

$$\text{sim}(d_j, q) = d_j \bullet q = \sum_{i=1}^{t} w_{ij} . w_{iq} \quad (2)$$

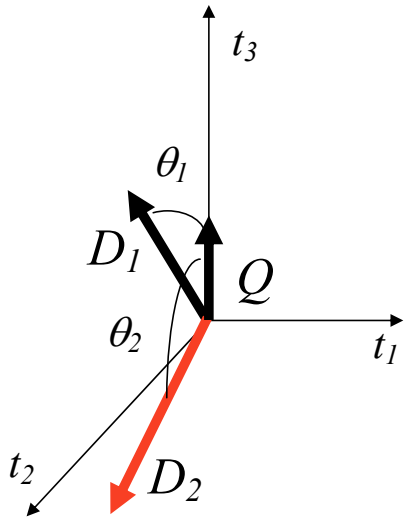
dimana d_j adalah vektor dokumen *j*, *q* adalah vektor *query*, w_{ij} adalah bobot term *i* dalam dokumen *j*, dan w_{iq} adalah bobot term *i* dalam *query*.

Contoh perhitungan:
D1 = 2T1 + 3T2 + 5T3
D2 = 3T1 + 7T2 + 1T3
Q = 0T1 + 0T2 + 2T3

sim(D1 , Q) = 2*0 + 3*0 + 5*2 = 10
sim(D2 , Q) = 3*0 + 7*0 + 1*2 = 2

 ISSN: 1907-5022

Pada Model Ruang Vektor, penentuan kemiripan yang lebih tepat didasarkan pada perhitungan besarnya sudut yang dibentuk oleh vektor *query* dan vektor dokumen tersebut. Sudut yang semakin kecil menunjukkan derajat relevansi yang semakin besar. Untuk mengatasi kesulitan menentukan besar sudut antar vektor, dapat digunakan konsep kosinus dimana sudut yang kecil memiliki nilai kosinus besar dan sudut yang besar memiliki nilai kosinus yang kecil. *Cosine Similarity Measure* memberikan tingkat kemiripan antara vektor dokumen d_i dan *query* q dengan melakukan perhitungan besar kosinus dari sudut yang dibentuk oleh dua vektor. Contoh dua vektor dokumen dan sebuah vektor *query* ditampilkan pada **Gambar 4** berikut ini:



Gambar 4. Contoh tiga vektor dan sudut yang terbentuk

Q dan D adalah vektor *query* dan dokumen. Perkalian dalam (*Inner Product*) dari kedua vektor tersebut adalah:

$$Q \bullet D = |Q||D|\cos\theta \tag{3}$$

sedangkan

$$|D| = \sqrt{\sum_{i=1}^{n} D_i^2} \quad \text{dan} \quad |Q| = \sqrt{\sum_{i=1}^{n} Q_i^2} \tag{4}$$

merupakan panjang vektor atau jarak Euclidean suatu vektor dengan titik nol. Dengan demikian, ukuran kosinus sudut antara kedua vektor dapat dinyatakan sebagai:

$$CosSim(d_j,q) = \frac{\sum_{i=1}^{t}\left(w_{ij}.w_{iq}\right)}{\sqrt{\sum_{i=1}^{t} w_{ij}^{\ 2}.\sum_{i=1}^{t} w_{iq}^{\ 2}}} \tag{5}$$

Perhitungan *Cosine Similarity Measure* tersebut memperhitungkan normalisasi dari panjang dokumen dengan adanya penyebut $|d_j|$ dan $|q|$, bila dibandingkan dengan konsep *Dot Product*. Normalisasi ini dilakukan karena dokumen yang panjang cenderung mengulang-ulang penggunaan term-term sehingga nilai *tf* term-term tersebut akan sangat tinggi. Akibatnya pembobotan dengan *dot product* akan menjadi kurang adil dimana hasil dari pencocokan(*matching*) akan cenderung lebih mengutamakan dokumen-dokumen yang panjang.

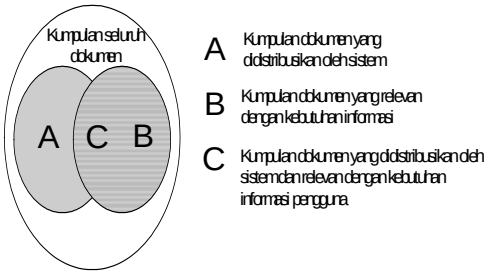
Model Ruang Vektor memiliki kelemahan dimana term-term yang memiliki makna yang sama (misalnya: *walk*, *walked*, *walking*) bisa dianggap sebagai elemen vektor yang berbeda. Untuk itu diperlukan kakas pendukung yaitu kakas *stemming* yang akan merepresentasikan term-term demikian ke dalam satu bentuk (misalnya: *walk*, *walked*, *walking* semuanya dibentuk menjadi *walk*).

4. EKSPERIMEN

Pengukuran performansi dari sistem *Information Filtering* dapat dilakukan dengan besaran-besaran pengukuran tertentu.

Kriteria yang digunakan untuk mengevaluasi keakuratan sistem *Information Filtering* adalah:

1. Semua dokumen yang relevan dengan kebutuhan informasi pengguna didistribusikan pada pengguna yang tepat.
2. Tidak ada satu pun dokumen tidak relevan yang didistribusikan pada pengguna yang tidak menginginkannya.



Gambar 5. Model Kumpulan Dokumen

Dua besaran pengukuran yang sering digunakan dalam sistem temu balik informasi adalah *Recall* dan *Precision*. Dengan mengamati **Gambar 5**, perhitungan nilai *Recall* dan *Precision* adalah:

$$Recall = \frac{C}{B}, \quad 0 \le Recall \le 1 \tag{6}$$

$$Precision = \frac{C}{A}, \quad 0 \le Precision \le 1 \tag{7}$$

Sistem ideal adalah sistem dengan *Recall*=1 dan *Precision*=1. *Recall* dan *Precision* cenderung memiliki nilai yang saling berbanding terbalik.

Kedua besaran pengukuran tersebut juga dapat diterapkan pada sistem *Information Filtering*. *Recall* merupakan jawaban untuk kriteria keakuratan sistem yang pertama, sedangkan *Precision*

merupakan jawaban untuk kriteria kedua. *Recall* dan *Precision* dihitung dengan membandingkan hasil pemfilteran/pe-*retrieve*-an oleh sistem dengan hasil pemfilteran/pe-*retrieve*-an oleh manusia. Hasil pemfilteran/pe-*retrieve*-an oleh manusia tersebut tersimpan sebagai penilaian relevansi (*relevance judgement*) antara dokumen dan *query*/profil. Perhitungan *Recall* dan *Precision* berlaku untuk *set measure* atau perhitungan dalam model himpunan seperti pada pendekatan *binary classification system* ini.

Hasil pengujian tingkat performansi IFVector menunjukkan bahwa nilai *Recall* IFVector cukup baik pada kisaran nilai ambang 0,05-0,1 yaitu sekitar 0,75-0,97. Nilai *Recall* tersebut semakin menurun dengan semakin tingginya nilai ambang yang digunakan. Nilai *Precision* juga menunjukkan hasil yang cukup baik yaitu berkisar pada nilai 0,69-0,93.

Peningkatan nilai ambang menyatakan semakin tingginya tingkat similaritas antara dokumen dan profil yang diinginkan. Hal tersebut berakibat pada semakin sedikitnya jumlah dokumen yang lolos dari pemfilteran daripada seharusnya (*Recall mengecil*) dan makin sedikit pula dokumen tidak relevan yang diterima pengguna (*Precision* membesar). Demikian juga sebaliknya dimana penurunan nilai ambang menyatakan semakin rendahnya tuntutan tingkat similaritas antara dokumen dan profil yang diinginkan. Hal tersebut berakibat pada semakin banyaknya jumlah dokumen relevan yang lolos dari pemfilteran (*Recall* membesar) dan makin besar pula kemungkinan dokumen tidak relevan ikut diterima pengguna (*Precision* mengecil).

5. KESIMPULAN

Makalah ini membahas tentang sistem penyaringan informasi dengan menggunakan model ruang vektor. Performansi dari sistem ini juga diukur dengan menggunakan besaran recall dan precision. Hasil evaluasi menunjukkan bahwa model ruang vektor dapat memberikan performansi yang baik.

DAFTAR PUSTAKA

[GRO98] Grossman, David A., Ophir Frieder (1998). Information Retrieval: Algorithms and Heuristics. Kluwer Academic Publisher.

[LEW96] Lewis, David D. (1996). The TREC-4 Filtering Track. http://trec.nist.gov/pubs/trec4/t4_proceedings.html Tanggal akses: 27 Februari 2003

[WON97] Wondergem, B.C.M., P. van Bommel, T.W.C. Huibers and T.P. van der Weide (1997). Towards an Agent-Based Retrieval Engine (Profile Information Filtering Project). British Computer Society. http://ewic.bcs.org/conferences/1997/irsg/papers/paper13.pdf Tanggal akses: 5 Maret 2004

[FRA92] Frakes, William B., Riardo Baeza-Yates (1992). Information Retrieval Data Structures & Algorithms. Prentice Hall.

[GRO98] Grossman, David A., Ophir Frieder (1998). Information Retrieval: Algorithms and Heuristics. Kluwer Academic Publisher.

[RIJ79] Rijsbergen, C.J. van (1979). Information retrieval. http://www.dcs.gla.ac.uk/Keith/Preface.html. Tanggal akses: 4 Maret 2004

[WON97] Wondergem, B.C.M., P. van Bommel, T.W.C. Huibers and T.P. van der Weide (1997). Towards an Agent-Based Retrieval Engine (Profile Information Filtering Project). British Computer Society. http://ewic.bcs.org/conferences/1997/irsg/papers/paper13.pdf Tanggal akses: 5 Maret 2004

[OAR96] Oard, Douglas W., Gary Marchionini (1996). A Conceptual Framework for Text Filtering. http://www.ee.umd.edu/medlab/filter/filter.html Tanggal akses: 5 Maret 2004

ISSN: 1907-5022