Forum Statistika dan Komputasi, Agustus 1996, p:24 - 29
ISSN 0853 - 8115

Vol. 2, No. 1

PENDEKATAN KUANTITATIF UNTUK PENELUSURAN INFORMASI

Julio Adisantoso [1)]

RINGKASAN

Makalah ini menelaah tiga model kuantitatif dalam penelusuran informasi, yaitu model ruang vektor, model peluang, dan model Boolean. Pembahasan dilakukan berdasarkan studi pustaka dari berbagai literatur dan jurnal terkait. Pendekatan ruang vektor merupakan model yang paling sederhana dengan tujuan mengukur kesamaan antar vektor suatu dokumen dan vektor pencarian yang ditentukan, sedangkan model peluang menggunakan prinsip peluang bersyarat untuk memberikan bobot dari suatu dokumen yang ditelusuri. Model Boolean pada mulanya tidak dapat digunakan untuk menentukan urutan derajat kesamaan suatu dokumen berdasarkan pencarian yang diberikan karena ukuran kesamaan antara dokumen dan rekord yang dicari bernilai 0 dan 1. Setelah dikombinasikan dengan model ruang vektor, model ini dapat memberikan ukuran kesamaan penciri dokumen. Model peluang dalam penelusuran informasi tergantung pada dua komponen utama yaitu sekumpulan dokumen-dokumen yang diidentifikasi sebagai rekord-rekord yang relevan dan yang tidak relevan. Relevansi suatu dokumen atau informasi ditentukan melalui keputusan yang diambil oleh pengguna berdasarkan pencarian rekord yang diberikan.

PENDAHULUAN

Otomasi penelusuran informasi (*Information Retrieval* = IR) telah dikembangkan sejak tahun 1940 untuk mempermudah akses buku, jurnal, atau bahan pustaka lainnya dengan alat bantu komputer. Sistem penelusuran informasi pada dasarnya adalah menentukan kesamaan antara informasi yang ada di dalam media penyimpanan dengan permintaan yang diberikan (*queries*), yang diukur dengan membandingkan nilai atribut tertentu dari file informasi yang ada dan yang diminta.

Salton (1979) membagi lingkup penelusuran informasi menjadi tiga topik, yaitu: (1) *database retrieval*, yang memproses berkas data dasar sederhana dengan menggunakan sejumlah atribut yang sudah didefinisikan sebagai ciri dari setiap rekord; (2) *reference retrieval*, dimana rekord data berupa buku, jurnal, majalah, atau bahan pustaka lainnya; dan (3) *fact retrieval*, yang memproses informasi dengan jenis karakteristik rekord yang lebih kompleks. Penelusuran terhadap bahan pustaka yang merupakan informasi berbasis teks adalah serupa dengan penelusuran terhadap rekord data dasar sederhana, yaitu dengan menentukan identitas yang berfungsi sebagai penciri dari setiap

[1)] *Staf pengajar Jurusan Statistika, FMIPA-IPB*

rekord. Karakteristik penciri rekord data berbasis teks dapat berupa kata (*term*), indeks, kata kunci, dan lain-lain. Dasar penelusuran informasi berdasarkan karakteristik seperti ini memungkinkan penggunaan model kuantitatif dalam implementasinya.

Beberapa model kuantitatif untuk penelusuran informasi telah dikembangkan, antara lain model ruang vektor, model peluang, dan model Boolean (Kwok, 1995). Metode penelusuran informasi dari ketiga model ini sangat beragam, yang masing-masing mempunyai keunggulan dan kelemahan, tergantung pada metode itu sendiri dan pola dokumen yang ditelusuri. Pendekatan ruang vektor merupakan model yang paling sederhana dengan tujuan mengukur kesamaan antar vektor suatu dokumen dan vektor pencarian yang ditentukan, sedangkan model peluang menggunakan prinsip peluang bersyarat untuk memberikan bobot dari suatu dokumen yang ditelusuri. Model Boolean konvensional tidak dapat digunakan untuk menentukan urutan derajat kesamaan (*similarity*) suatu dokumen berdasarkan pencarian yang diberikan karena ukuran kesamaan antara dokumen dan rekord yang dicari bernilai 0 dan 1 (Salton, 1989).

Makalah ini menelaah ketiga model kuantitatif dalam penelusuran informasi berdasarkan studi pustaka dari berbagai literatur dan jurnal terkait.

PENCIRI DOKUMEN

Penelusuran informasi secara otomatis pada umumnya dilakukan dengan membandingkan secara langsung antara kata yang diminta dengan kata yang ada di dalam suatu dokumen. Pada kenyataannya, kata yang muncul dalam suatu dokumen sangat beragam sehingga untuk melakukan pembandingan dan perhitungan kata antar dokumen menjadi hal yang sulit dilakukan. Oleh karena itu, perlu ditentukan identitas atau profil yang dapat digunakan sebagai penciri suatu dokumen sehingga dokumen dapat diindeks sesuai dengan penciri dokumen yang bersangkutan.

Proses penentuan indeks dokumen dapat dilakukan dengan dua cara, yaitu manual dan otomatis. Penentuan indeks secara manual melibatkan pakar di bidang ilmu masing-masing yang menjadi isi dari dokumen yang sedang ditelaah. Dengan perkembangan teknologi informasi, penentuan indeks dokumen dapat dilakukan secara otomatis berdasarkan frekuensi kemunculan kata.

Porter (1982) memberikan algoritma penentuan frekuensi kemunculan kata dari suatu dokumen sebagai berikut:

1. ambil setiap kata yang terdapat pada dokumen **d** dimana kata adalah setiap karakter teks yang dipisahkan oleh spasi;
2. dari setiap kata yang diperoleh pada langkah (1), buang semua karakter selain angka dan huruf;
3. buang kata-kata yang hanya terdiri dari satu karakter;
4. buang kata-kata yang tidak perlu, misalnya kata penghubung;
5. ubah setiap karakter menjadi huruf rendah; dan
6. hitung frekuensi kemunculan suatu kata pada dokumen **d**.

MODEL RUANG VEKTOR

Misalkan terdapat n rekord dokumen D_1, D_2, ..., D_n dan t atribut A_1, A_2, ..., A_t yang digunakan sebagai penciri setiap rekord dokumen. Dengan demikian, suatu rekord D_i dapat ditulis sebagai vektor atribut

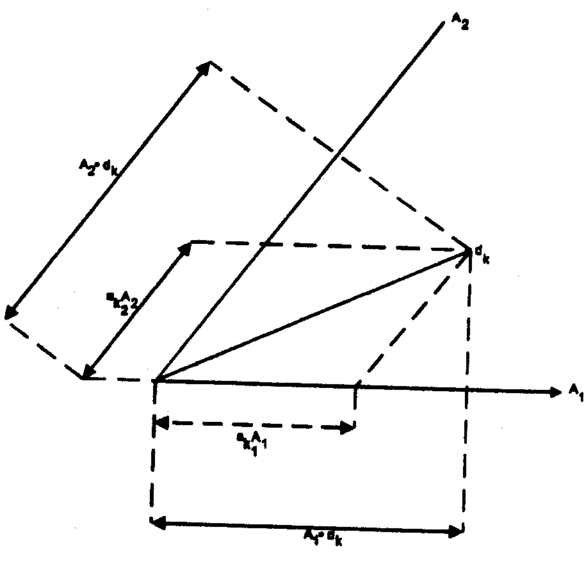
$$\mathbf{d}_i' = (a_{i1}, a_{i2}, \ldots, a_{it}),$$

sedangkan a_{ij} menunjukkan nilai kuantifi-kasi sebagai penciri dari atribut A_j dalam dokumen D_i. Nilai a_{ij} dapat berupa nilai biner yang menunjukkan adanya kata ke-k pada suatu dokumen D_i ($x_{ik} = a_{ik} = 1$, $1 \le k \le t$), dan $x_{ik} = a_{ik} = 0$ untuk selainnya (Salton, 1979; Croft and Harper, 1979). Disamping itu, nilai a_{ij} ini dapat juga berupa frekuensi munculnya kata ke-k dalam dokumen D_i, yaitu f_{ik} (Kwok, 1995).

Jika didefinisikan suatu ruang vektor $\mathbb{R}$ dimana setiap vektor dalam $\mathbb{R}$ saling ortogonal, maka dokumen ke-k dapat ditulis dalam bentuk kombinasi linier sebagai berikut:

$$\mathbf{d}_k = \sum_{i=1}^{t} a_{ki} A_i$$

Gambar 1 menunjukkan vektor dokumen dalam ruang vektor berdimensi dua.



Gambar 1. Representasi dokumen dalam ruang vektor

Rekord pencarian (query) yang diinginkan ditulis dalam bentuk:

$$\mathbf{q} = \sum_{i=1}^{t} q_i A_i$$

Dengan demikian, ukuran kesamaan antara $\mathbf{d}_k$ dan $\mathbf{q}$ dalam ruang vektor $\mathbb{R}$ dapat diukur dengan menghitung hasil kali silang kedua vektor seperti berikut:

$$\mathbf{d}_k \cdot \mathbf{q} = \sum_{i,j=1}^{t} a_{ki} q_j A_i \cdot A_j$$

Karena setiap vektor di dalam $\mathbb{R}$ saling ortogonal ($\mathbf{A'A} = 0$) maka persamaan ini menjadi koefisien kesamaan sebagai berikut:

$$\mathrm{sim}(D_k, Q) = \mathbf{d}_k{}'\mathbf{q} = \sum_{i=1}^{t} a_{ki} q_i$$

Salton (1989) melakukan normalisasi ukuran koefisien kesamaan ini menjadi koefisien Dice, Cosine, dan Jaccard. Ketiga koefisien ini berturut-turut adalah:

a. $$\frac{2\,\mathbf{d}_k\,\mathbf{q}}{(\mathbf{d}_k{}'\mathbf{d}_k)\,(\mathbf{q}'\mathbf{q})}$$

b. $$\frac{\mathbf{d}_k\,\mathbf{q}}{\sqrt{(\mathbf{d}_k{}'\mathbf{d}_k)\,(\mathbf{q}'\mathbf{q})}}$$

c. $$\frac{\mathbf{d}_k{}'\mathbf{q}}{(\mathbf{d}_k - \mathbf{q})'\,(\mathbf{d}_k - \mathbf{q})}$$

Ukuran kesamaan setiap vektor dokumen ini selanjutnya digunakan sebagai dasar pemberian peringkat (indeks) setiap dokumen sesuai dengan kunci pencarian rekord yang diinginkan (Salton and Buckley, 1988).

MODEL PELUANG

Model peluang dalam penelusuran informasi tergantung pada dua komponen utama yaitu sekumpulan dokumen-dokumen yang diidentifikasi sebagai rekord-rekord yang relevan dan yang tidak relevan. Relevansi suatu dokumen atau informasi ditentukan melalui keputusan yang diambil oleh pengguna berdasarkan pencarian rekord yang diberikan (Croft and Harper, 1979).

Salton (1979) menunjukkan bahwa penelusuran informasi model peluang dapat diekspresikan sebagai hubungan pertidaksamaan

$$P(rel)\, a_2 \geq [1 - P(rel)]\, a_1$$

sedangkan P(rel) adalah peluang suatu rekord relevan, a_1 adalah parameter kehilangan (*loss*) berkaitan dengan penelusuran suatu rekord yang tidak relevan, dan a_2 adalah parameter yang berkaitan dengan rekord relevan yang tidak ditelusuri. Pertidaksamaan ini dapat dicatat sebagai fungsi diskriminan $g \geq 0$, sedangkan

$$g = \frac{P(rel)}{1 - P(rel)} - \frac{a_1}{a_2}$$

Untuk mengimplementasikan aturan penelusuran informasi dengan menggunakan persamaan di atas, didefinisikan dua buah peluang bersyarat, yaitu:

a. $P(x_i \mid w_1)$:
peluang kata x_i muncul pada rekord setelah diketahui bahwa rekord tersebut relevan bagi pencarian yang diberikan,

b. $P(x_i \mid w_2)$:
peluang kata x_i muncul pada rekord setelah diketahui bahwa rekord tersebut tidak relevan bagi pencarian yang diberikan.

Dengan menggunakan formula Bayes dapat ditentukan $P(w_i \mid \mathbf{x})$, yaitu:

$$P(w_i|\mathbf{x}) = \frac{P(\mathbf{x}|w_i)P(w_i)}{P(\mathbf{x})}, i = 1,2$$

sedangkan w_1 dan w_2 menunjukkan rekord yang relevan dan tidak relevan.

Jika $a_1 = a_2 = 1$ maka fungsi diskriminan $f \geq 1$ adalah

$$f(\mathbf{x}) = \frac{P(w_1|\mathbf{x})}{P(w_2|\mathbf{x})} = \frac{P(\mathbf{x}|w_1)P(w_1)}{P(\mathbf{x}|w_2)P(w_2)}$$

Persamaan ini dapat dilinierkan menjadi

$$g(\mathbf{x}) = \ln f(\mathbf{x}) = \ln\frac{P(\mathbf{x}|w_1)}{P(\mathbf{x}|w_2)} + \ln\frac{P(w_1)}{P(w_2)}$$

Dengan asumsi bahwa kemunculan suatu kata dalam setiap dokumen adalah saling bebas dan $x_i = [0,1]$, $i=1, 2, \ldots, t$, maka $P(\mathbf{x}|w_i)$ dapat ditulis dalam peluang binom sebagai berikut:

$$P(\mathbf{x}|w_1) = \prod_{i=1}^{t} p_i^{x_i}(1-p_i)^{1-x_i}$$

dan

$$P(\mathbf{x}|w_2) = \prod_{i=1}^{t} q_i^{x_i}(1-q_i)^{1-x_i}$$

sedangkan $p_i = P(x_i=1|w_1)$ dan $q_i = P(x_i=1|w_2)$. Dengan demikian, persamaan $g(\mathbf{x})$ di atas dapat ditulis sebagai:

$$g(\mathbf{x}) = \sum_{i=1}^{t} x_i b_i + \sum_{i=1}^{t} \ln\frac{1-p_i}{1-q_i} + \ln\frac{P(w_1)}{P(w_2)}$$

sedangkan $b_i = \ln\frac{p_i}{1-p_i} + \ln\frac{1-q_i}{q_i}$. Nilai b_i ini selanjutnya digunakan sebagai pembobot bagi penciri ke-i pada rekord dokumen dan rekord kunci yang diminta (Roberston and Sparck Jones, 1976 dalam Croft and Harper, 1979).

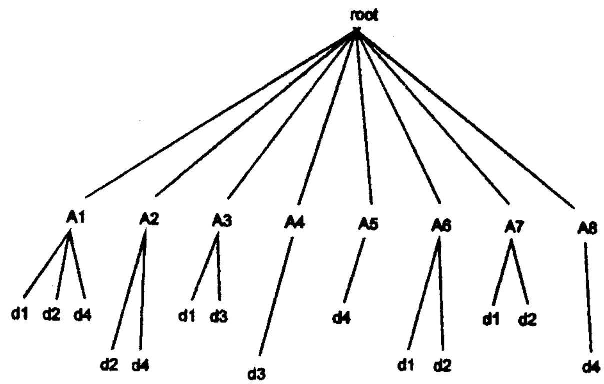
Peluang p_i dan q_i dapat diduga berdasarkan pada sekumpulan dokumen contoh yang relevan dan yang tidak relevan dengan vektor permintaan $\mathbf{q}' = (q_1, q_2, \ldots, q_t)$, sedangkan q_i adalah kemunculan kata ke-i dalam kunci permintaan yang diberikan. Kwok (1990) menambahkan vektor **q** ke dalam sekumpulan dokumen sebagai dokumen ekstra sehingga vektor **q** ini sebagai dokumen yang relevan dan dokumen yang sedang ditelusuri sebagai dokumen yang tidak relevan. Dengan demikian, matrik data kemunculan suatu kata yang ditangani berbentuk sebagai berikut:

$$D = \begin{bmatrix} x_{11} & x_{12} & \cdots & \cdots & \cdots & x_{1t} \\ x_{21} & x_{22} & \cdots & \cdots & \cdots & x_{2t} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & \cdots & \cdots & x_{nt} \\ -- & -- & -- & -- & -- & -- \\ q_1 & q_2 & \cdots & \cdots & \cdots & q_t \end{bmatrix}$$

MODEL BOOLEAN

Penelusuran informasi model Boolean menggunakan prinsip kesesuaian antara kata yang dicari dengan kata-kata yang terdapat di dalam dokumen dan dikombinasikan dengan operator logika AND, OR, atau NOT. Misalkan terdapat empat dokumen (d_1, d_2, d_3, dan d_4) yang secara keseluruhan mempunyai delapan atribut (A_1, A_2, ..., A_8) dan dapat digambarkan dalam bentuk diagram pohon seperti tercantum pada Gambar 2.

Gambar 2. Struktur Data Contoh



Jika diberikan penelusuran q = ((q_1 AND q_2) OR (q_5 AND q_8) maka diperoleh jawaban d_2 dan d_4.

Pendekatan kuantitatif untuk model penelusuran informasi Boolean dapat dilakukan dengan mengkombinasikan operator Boolean dengan bobot kata seperti yang digunakan dalam model ruang vektor. Bobot ini ditentukan berdasarkan frekuensi kemunculan kata dalam masing-masing dokumen. Ukuran kesamaan antara dokumen yang ada dengan yang diinginkan (*query*) dikoreksi oleh parameter khusus yang disebut nilai p, dimana $1 \leq p \leq \infty$. Jika $p=\infty$ maka pendekatan model ini dapat diinterpretasikan sebagai model Boolean biasa, sedangkan jika p=1 maka pengaruh operator Boolean sama dengan model ruang vektor (Watters, 1989; Salton, 1989).

PENUTUP

Model-model penelusuran informasi yang berkembang umumnya menggunakan pendekatan kuantitatif agar dapat diimplementasikan secara otomatis dengan menggunakan alat bantu komputer.

Dari ketiga model penelusuran yang telah dibahas dalam makalah ini, seluruhnya diarahkan untuk menentukan ukuran kesamaan antara dokumen yang ada dengan yang dicari berdasarkan keberadaan kata. Metode ini mempunyai kelemahan karena langsung akan menghilangkan dokumen yang tidak mengandung kata yang dicari sama sekali meskipun dokumen tersebut dari segi isi cukup relevan. Oleh karena itu perlu dilakukan analisis lanjutan untuk menentukan ukuran kesamaan yang tidak hanya tergantung pada keberadaan suatu kata dalam dokumen, misalnya dengan memasukkan faktor korelasi antar vektor **d** kedalam ukuran kesamaan.

Disamping itu, perlu ditelaah kemungkinan penggunaan metode indeks dokumen berdasarkan frekuensi kemunculan kata untuk dokumen berbahasa Indonesia karena sudah dapat dipastikan bahwa konsistensi kata dalam bahasa Indonesia sangat rendah dan sering dijumpai ketidakbakuan struktur kata dalam kalimat dokumen berbahasa Indonesia.

DAFTAR PUSTAKA

Croft, W.B. and D.J. Harper. 1979. Using Probabilistic Models of Document Retrieval Without Relevance Information. Journal of Documentation. 35(4).

Frakes, W.B and R. Baeza-Yates. 1992. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, New Jersey.

Kwok, K.L. 1995. A Network Approach to Probabilistic Information Retrieval. ACM Transaction on Information System. 13(3):324 - 353.

Porter, M.F. 1982. Implementing a Probabilistic Information Retrieval System. Information Technology: Research and Development. 1:131-156.

Salton, G. 1979. Mathematics and Information Retrieval. Journal of Documentation. 35(1): 1-29.

Salton, G. 1989. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley Publishing Company, Inc.

Salton, G. and C. Buckley. 1988. Term-Weighting Approach in Automatic Text Retrieval. Information Processing and Management. 24(5): 513 - 523.

Watters, C.R. 1989. Logic Framework for Information Retrieval. Journal of The American Society for Information Science. 40(5):311 - 324.

⊗