PENERAPAN BERBAGAI TEKNIK SISTEM TEMU-KEMBALI INFORMASI BERBASIS HIPERTEKS

Zainal A. Hasibuan dan Yofi Andri

Fakultas Ilmu Komputer Universitas Indonesia
Email : zhasibua@caplin.cs.ui.ac.id dan yofi298@puspa.cs.ui.ac.id

ABSTRAK

Pada tulisan ini akan dijelaskan berbagai teknik sistem temu-kembali informasi dan rancangan integrasi sistem ke basis hiperteks. Sistem pengindeksan yang dijelaskan adalah pengindeksan dengan pembobotan berdasarkan frekuensi dan berdasarkan rumus Savoy [1]. Sedangkan teknik temu-kembali informasi yang dijelaskan adalah teknik Boolean biasa, teknik Boolean berperingkat dan teknik Extended Boolean. Kinerja berbagai teknik temu-kembali informasi dan berbagai pengindeksan di ukur dengan menampilkan dokumen yang terambil berikut bobot peringkatnya. Sistem ini dapat digunakan sebagai "benchmarking tool" untuk mengukur kinerja berbagai teknik yang digunakan dalam sistem temu-kembali informasi.

Kata kunci: Sistem pengindeksan, teknik Boolean biasa, teknik Boolean berperingkat, teknik Extended Boolean .

1. LATAR BELAKANG

Ledakan informasi menyebabkan masyarakat akan mengalami kesulitan mendapatkan informasi yang cepat, padat dan relevan dengan kebutuhannya. Untuk mengatasi hal tersebut diperlukan suatu sistem temu-kembali informasi. Menurut Davies & Weeks [2], tahun 1982 pertumbuhan informasi meningkat dua kali lipat setiap 5 tahun. Tahun 1988 diprediksi informasi meningkat dua kali lipat setiap 2.2 tahun dan tahun 1992 berubah lagi menjadi setiap 1,6 tahun. Kecendrungan ini akan selalu berubah dan saat ini terjadi peningkatan informasi dua kali lipat setiap satu tahun.

Kecepatan perubahan dan penambahan informasi menyebabkan dibutuhkannya suatu sistem yang dapat mengakses dan menyediakan berbagai informasi tersebut. Saat ini telah banyak dari berbagai informasi tersebut dapat diakses secara elektronik melalui WWW atau *internet* dengan menggunakan berbagai mesin penelusur (*search engine*). Perbedaan mesin penelusur yang satu dengan yang lain sangat tergantung pada teknik temu-kembali informasi dan teknik pengindeksan yang dipakai.

Pada tulisan ini akan dijelaskan kinerja berbagai teknik sistem temu-kembali informasi berbasis hiperteks. Teknik-teknik yang akan dibahas tersebut adalah: teknik *Boolean* biasa dan teknik *Boolean* berperingkat [3], serta teknik *Extended Boolean* berdasarkan p-norm model [4]. Sedangkan teknik pembobotannya adalah teknik pembobotan berdasarkan frekuensi dan teknik pembobotan berdasarkan rumus Savoy [1].

2. SISTEM TEMU-KEMBALI INFORMASI

Sistem temu-kembali informasi pada prinsipnya adalah suatu sistem yang sederhana. Misalkan ada sebuah kumpulan dokumen dan seorang user yang memformulasikan sebuah pertanyaan (*request* atau *query*). Jawaban dari pertanyaan tersebut adalah sekumpulan dokumen yang relevan dan membuang dokumen yang tidak relevan. Secara matematis hal tersebut dapat dituliskan sebagai berikut :

$Q \xrightarrow{2^n} D$, dimana Q = pertanyaan (queri), D = dokumen, n = jumlah dokumen, 2^n = jumlah kemungkinan himpunan bagian dari dokumen yang ditemukan. Sistem temu-kembali akan mengambil salah satu dari kemungkinan tersebut.

Sistem temu-kembali informasi pada dasarnya dibagi dalam dua komponen utama yaitu sistem pengindeksan (*indexing*) yang menghasilkan basis data sistem dan temu-kembali yang merupakan gabungan dari *user interface* dan look-up-table. Pada bagian selanjutnya akan dijelaskan berbagai macam sistem pengindeksan dan teknik-teknik temu-kembali informasi yang telah dikembangkan.

Makalah diterima [16 September 2001]. Revisi akhir [25 November 2001]

3. INDEXING

Indexing merupakan sebuah proses untuk melakukan pengindeksan terhadap kumpulan dokumen yang akan disediakan sebagai informasi kepada pemakai. Proses pengindeksan bisa secara manual ataupun secara otomatis. Dewasa ini, sistem pengindeksan secara manual mulai digantikan oleh sistem pengindeksan otomatis. Adapun tahapan dari pengindeksan adalah sebagai berikut :

- *Parsing* Dokumen yaitu proses pengambilan kata-kata dari kumpulan dokumen.
- *Stoplist* yaitu proses pembuangan kata buang seperti: tetapi, yaitu, sedangkan, dan sebagainya.
- *Stemming* yaitu proses penghilangan/ pemotongan dari suatu kata menjadi bentuk dasar. Kata "diadaptasikan" atau "beradaptasi" mejadi kata "adaptasi" sebagai istilah.
- *Term Weighting* dan *Inverted File* yaitu proses pemberian bobot pada istilah.

Didalam memberikan bobot pada sebuah istilah, terdapat berbagai macam teknik antara lain yaitu :

1. Teknik pembobotan berdasarkan frekuensi kemunculan istilah pada satu dokumen [3]. Teknik pembobotan ini cukup sederhana dimana bobot suatu istilah pada sebuah dokumen berdasarkan jumlah kemunculannya pada dokumen tersebut.
2. Teknik pembobotan berdasarkan rumus Savoy [1] yaitu:

$W_{ik} = ntf_{ik} * nidf_{k}$

dimana $ntf_{ik} = \frac{tf_{ik}}{Max_j tf_{ij}}$ dan $nidf_k = \frac{\log\left[\frac{n}{df_k}\right]}{\log(n)}$

Dengan :

- W_{ik} adalah bobot istilah k pada dokumen i.
- tf_{ik} merupakan frekuensi dari istilah k dalam dokumen i.
- n adalah jumlah dokumen dalam kumpulan dokumen.
- df_k adalah jumlah dokumen yang mengandung istilah k.
- $Max_j\ tf_{ij}$ adalah frekuensi istilah terbesar pada satu dokumen.

Pada teknik pembobotan ini, bobot istilah telah dinormalisasi. Dalam menentukan bobot suatu istilah tidak hanya berdasarkan frekuensi kemunculan istilah di satu dokumen, tetapi juga memperhatikan frekuensi terbesar pada suatu istilah yang dimiliki oleh dokumen bersangkutan. Hal ini untuk menentukan posisi relatif bobot dari istilah dibanding dengan istilah-istilah lain di dokumen yang sama. Selain itu teknik ini juga memperhitungkan jumlah dokumen yang mengandung istilah yang bersangkutan dan jumlah keseluruhan dokumen. Hal ini berguna untuk mengetahui posisi relatif bobot istilah bersangkutan pada suatu dokumen dibandingkan dengan dokumen-dokumen lain yang memiliki istilah yang sama. Sehingga jika sebuah istilah mempunyai frekuensi kemunculan yang sama pada dua dokumen belum tentu mempunyai bobot yang sama.

4. TEKNIK-TEKNIK TEMU-KEMBALI INFORMASI

Ada beberapa teknik temu-kembali informasi yang telah dikembangkan yaitu teknik *Boolean* sederhana dan teknik *Boolean* berperingkat [3], serta teknik *Extended Boolean* berdasarkan p-norm model [4]. Untuk lebih jelasnya mengenai perbedaan dan keunggulan masing-masing teknik ini dapat dilihat pada penjelasan berikut.

1. Teknik *Boolean*

Teknik *Boolean* merupakan suatu cara dalam mengekspresikan keinginan pemakai ke sebuah kueri dengan memakai operator-operator *Boolean* [5] yaitu: "and", "or", dan "not". Adapun maksud dari operator "and" adalah untuk menggabungkan istilah-istilah kedalam sebuah ungkapan, dan operator "or" adalah untuk memperlakukan istilah-istilah sebagai sinonim, sedangkan operator "not" merupakan sebuah pembatasan. Pada teknik *Boolean* sederhana, kueri diproses sesuai dengan operator yang digunakan dan menampilkan dokumen berdasarkan urutan dokumen ditemukan. Sedangkan pada teknik *Boolean* berperingkat, dokumen diperingkat berdasarkan bobot dari dokumen. Adapun pembobotan dari masing-masing dokumen berdasarkan aturan sebagai berikut :

A and B → $D_{1A\cap B}, D_{2A\cap B}, \ldots \rightarrow d_{1A\cap B} > d_{2A\cap B} >$

... dengan $d_{A\cap B} = \min(d_A, d_B)$

A or B → $D_{1A\cup B}, D_{2A\cup B}, \ldots \rightarrow d_{1A\cup B} > d_{2A\cup B} >$

... dengan $d_{A\cup B} = \max(d_A, d_B)$

Not A → $U - d_A$

Dimana d_A menyatakan bobot istilah A pada dokumen D. Bobot istilah ini didapat dari hasil proses *Indexing*. $\text{Min}(d_A, d_B)$ berarti bahwa sebuah dokumen di *retrieve* dengan bobot

sebesar nilai terkecil dari bobot-bobot istilah yang dipunyainya. Max(d_A,d_B) berarti bahwa sebuah dokumen di *retrieve* dengan bobot sebesar nilai terbesar dari bobot-bobot istilah yang dipunyainya.

2. Teknik *Extendei Boolean*
   Teknik *Extended Boolean* berdasarkan p-norm model merupakan pengembangan lebih lanjut dari model *Boolean*. Teknik ini memakai operator yang dikomputasi berdasarkan rumus Savoy [1], sebagai berikut :

Query	*Retrieval Status Value* (RSV)
A OR <p> B	$\sqrt{\frac{W_{ia}^{\ p} + W_{ib}^{\ p}}{2}}$
A AND <p> B	$1 - \sqrt[p]{\frac{(1-W_{ia})^p + (1-W_{ib})^p}{2}}$
NOT A	$1 - W_{ia}$

Dimana :

- p adalah nilai p-norm yang dimasukkan pada kueri.
- W_{ia} adalah bobot istilah A dalam indeks pada dokumen D_i.
- W_{ib} adalah bobot istilah B dalam indeks pada dokumen D_i.

Pemeringkatan yang dipakai bisa dua cara:

- Langsung mengurutkan dokumen (dari besar ke kecil) berdasarkan bobot dokumen yang didapat dengan rumus RSV (*retrieval status value*) di atas.
- Memakai rumus *Learning Scheme*.

$$RSV(D_i) = RSV_{init}(D_i) + \sum_{k=1}^{r} \alpha_{ik}\ norm * RSV_{init}(D_k)$$

untuk i= 1, 2,...., n,

Dimana :

- $RSV_{init}(D_i)$ merupakan *retrieval status value* dari dokumen *i* yang dikomputasi berdasarkan rumus teknik *retrieval* P-norm model.
- α_{ik} merupakan bobot keterhubungan antara dokumen *i* dan *k*. Bobot keterhubungan ini didapat dari nilai *relevance link* yang merupakan hasil dari proses pembelajaran.

5. SISTEM TEMU-KEMBALI INFORMASI BERBASIS HIPERTEKS

Pada awalnya, hiperteks dan temu-kembali informasi merupakan bidang penelitian yang berbeda satu dengan yang lain. Hiperteks berkisar pada masalah *user-disorientation*, strategi pengembangan dokumen hiperteks dan mekanisme konversi dokumen tekstual menjadi bentuk hiperteks [6]. Sedangkan temu-kembali informasi bergerak pada topik manipulasi kueri, konsep basis data tekstual dan relevansi dokumen terhadap kueri [7]. Penggabungan kedua bidang ini dapat memecahkan masalah-masalah dalam bidang temu-kembali informasi. Misalnya, sistem temu-kembali informasi yang didasarkan pada penggunaan operator *Boolean*, mengandalkan kemampuan pemakai dalam memformulasikan kueri. Hal ini sering mempersulit pengguna. Dengan adanya sistem hiperteks, hal ini dapat di permudah dengan penyediaan antar muka yang memakai pencarian dengan metode *browsing*.

Smeaton di dalam [6] juga menyatakan bahwa hiperteks dan temu-kembali informasi itu saling berkomplemen satu sama lain. Hiperteks membutuhkan lebih banyak *searching* sedangkan temu-kembali informasi membutuhkan lebih banyak *browsing*. Hal yang dimaksud adalah hiperteks akan semakin baik jika disertai dengan fasilitas *search* dan temu-kembali informasi membutuhkan *browsing* dalam melakukan pencarian yang efisien. Adapun maksud dari *searching* adalah berusaha mendapatkan atau mencapai tujuan spesifik sedangkan *browsing* adalah mengikuti suatu path sampai mencapai suatu tujuan. Menurut Brown [8], *browsing* itu bisa diibaratkan dengan **From Where to What**. Maksudnya adalah kita tahu dimana posisi kita dalam *database* dan kita ingin tahu apa yang ada disana (*database*). Sedangkan *Searching* bisa diibaratkan dengan **From What to Where**. Maksudnya adalah kita tahu apa yang kita inginkan dan kita ingin menemukan dimana dia didalam *database*.

Penggabungan sistem temu-kembali kedalam basis hiperteks lebih dikenal dengan nama search engine, dimana sistem ini dapat dibagi kedalam dua kategori berdasarkan sumber informasinya yaitu:

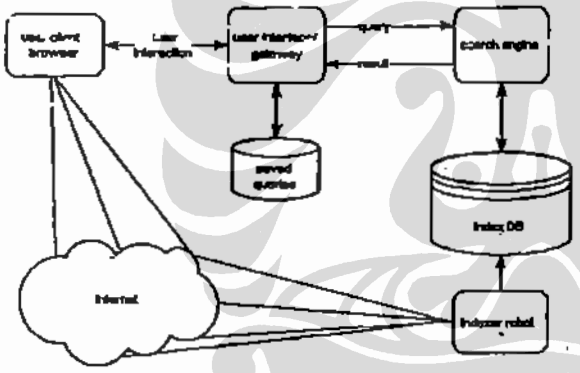
1. *Worldwide Search Engine*
   Worldwide Search Engine adalah suatu sistem temu-kembali informasi yang mengambil data-data dari berbagai server di seluruh penjuru dunia. Data-data tersebut diambil melalui program yang disebut dengan "robot" atau "bot". Program inilah yang melakukan pencarian data pada setiap server, yang kemudian dikirim ke server pusat pada selang waktu tertentu.

2. *Local Search Engine*

   Local search engine adalah suatu sistem temu-kembali informasi yang mengambil data-data dari server tertentu saja. Kata "local", yang berarti lokal atau setempat, memberi penekanan akan lokasi sumber data yang akan digunakan. *Local search engine* tidak dirancang untuk mengarungi belantara internet seperti *worlwide search engine.* Tujuan implementasi local search engine dimaksudkan untuk pencarian pada objek spesifik dan lebih kecil lingkupnya dibandingkan internet sendiri.

Mengenai pemilihan penerapan sistem temu-kembali berbentuk *local search engine* atau *worlwide search engine* tergantung kepada masalah atau jenis informasi yang akan kita sediakan. Penerapan kedua kategori ini hanya akan mempengaruhi cara sistem pengindeksan dari temu-kembali. Sedangkan teknik *retrieval* dan rancangan penerapan teknik pada hiperteks akan sama saja, baik pengindeksannya secara *local search engine* ataupun *worldwide search engine.*

Penelitian mengenai penerapan sistem temu-kembali berbasis hiperteks telah mulai dilakukan seiring dengan perkembangan internet akhir-akhir ini. Penelitian yang dilakukan [9] menggunakan rancangan/arsitektur seperti terlihat pada gambar 1.

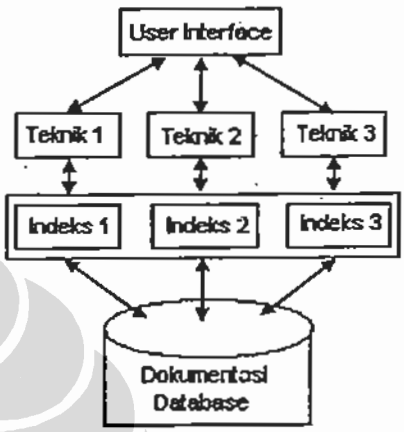


Gambar 1. Arsitektur sistem temu-kembali [9]

Arsitektur yang dirancang ini terdiri dari dua komponen utama yaitu: *Index Builder* dan *Search Engine. Index builder* merupakan sebuah sistem pengindeksan yang memanfaatkan "robot" yang berkomunikasi dengan menggunakan HTTP (*Hypertext Transfer Protocol*) untuk mencari informasi yang akan di indeks. Sedangkan *Search engine* merupakan teknik dari temu-kembali dalam menemukan dokumen dan sekaligus mengeksekusi algoritma peringkat dalam menampilkan dokumen. Sedangkan komunikasi antara pemakai dan *search engine* dalam memformulasikan kueri dilakukan melalui *User Interface.* Setelah pemakai menemukan dokumen yang relevan dengan kueri, dapat langsung melakukan *browsing* ke sumber informasi dalam hal ini adalah alamat tempat *www.*

6. IMPLEMENTASI

Adapun arsitektur dari sistem temu-kembali yang dikembangkan dalam penelitian ini dapat dilihat pada gambar 2 di bawah ini.



Gambar 2. Rancangan Sistem Temu-Kembali Berbasis Hiperteks

Sistem temu-kembali yang dibangun terdiri dari berbagai macam teknik *retrieval* seperti teknik *Boolean* biasa dan *Boolean* berperingkat serta teknik *Extended Boolean* berdasarkan p-norm model. Sedangkan teknik pengindeksannya juga terdiri dari beberapa macam antara lain teknik berdasarkan frekuensi kemunculan istilah dan teknik pengindeksan yang dinormalisasi berdasarkan aturan Savoy [1]. Pada sistem ini, teknik *retrieval*, basis data indeks dan kumpulan dokumen berada dalam sebuah komputer *server* yang sama (*local*). Sedangkan antar muka dari sistem yang dikembangkan adalah berbasis hiperteks.

Implementasi program temu-kembali informasi berbasis hiperteks mempunyai kelebihan dibandingkan dengan sistem konvensional, terutama dalam hal navigasi sistem. Sebagai gambaran lebih jelas dapat di lihat pada gambar 3. Untuk membaca isi dokumen, cukup dengan memilih *hiperlink* judul atau nama pengarang. Demikian pula untuk melihat relasi istilah guna membentuk kueri baru, cukup dengan membuka *selection list* yang di dalamnya terdapat istilah lain yang berelasi. Relasi istilah ini

telah di proses sewaktu pemakai melakukan pencarian berdasarkan istilah-istilah yang terdapat pada kueri pemakai.

Pada sistem temu-kembali berbasis hiperteks ini, selain menemukan dokumen yang relevan juga menampilkan relasi istilah pada kueri dalam satu proses. Semua modul yang ada pada sistem ini menjadi satu kesatuan mode, dimana dokumen yang ditemukan secara langsung merupakan input bagi modul lain seperti modul relasi istilah.

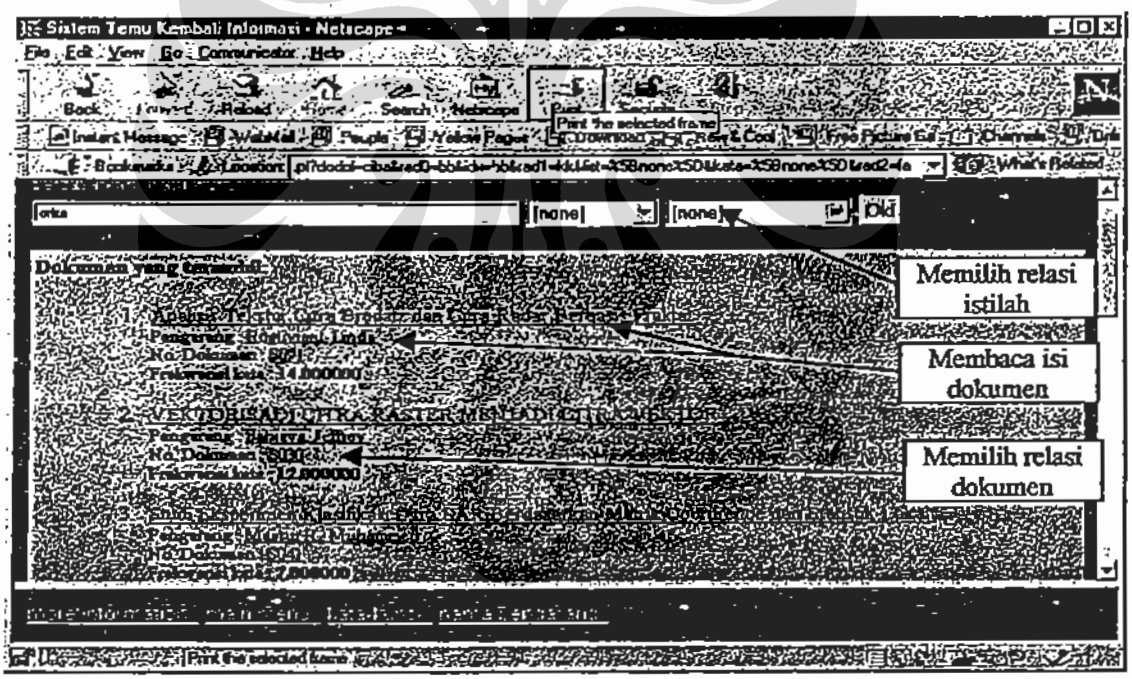
7. HASIL DAN INTERPRETASI BERBAGAI SISTEM PENGINDEKSAN

Pada bagian ini akan dijelaskan mengenai pemakaian berbagai sistem pengindeksan oleh masing-masing teknik. Interpretasi hasil dijelaskan dengan dua pendekatan yaitu, 1) menjelaskan mengenai perbedaan dari dua sistem *indexing* terhadap satu teknik, dan 2) mengenai perbedaan dari tiga teknik terhadap suatu sistem *indexing*.

7.1. Hasil Sistem *Indexing*

Tabel 1 menampilkan hasil kueri "citra and komputer" dengan Teknik *Boolean*. Indeks 1 merupakan sistem pengindeksan berdasarkan frekuensi, dan indeks 2 merupakan hasil pengindeksan berdasarkan rumus Savoy [1]. Hasil penemuan dokumen dan bobotnya pada tabel 1, dapat dijelaskan sebagai berikut :

- Perhatikan dokumen S048 pada hasil kueri dengan indeks 1 dan indeks 2. Pada indeks 1, dokumen S048 mempunyai bobot 2 atau mempunyai nilai RSV yang tertinggi. Pada hasil indeks 2, dokumen S048 mempunyai bobot 0.039120 atau mempunyai RSV dengan urutan peringkat ke-dua yang lebih kecil dari S005. Walaupun dokumen S048 dihasilkan oleh istilah yang mempunyai frekuensi yang tinggi pada sebuah dokumen dengan memakai indeks 1, tetapi pada sistem pengindeksan Savoy [1] (indeks 2), istilah yang mempunyai frekuensi tinggi yang menghasilkan dokumen S048 belum menjamin bahwa dokumen S048 itulah yang lebih relevan dengan kueri yang di masukkan. Hal ini disebabkan karena sistem pengindeksan



Gambar 3. Hasil dokumen yang terambil dengan menggunakan Teknik *Boolean* Berbasis Hiperteks

Tabel 1. Hasil kueri "citra and komputer" teknik *Boolean*

	Indeks 1		Indeks 2
1. S048	2.000000	1. S005	0.099570
2. S005	1.000000	2. S048	0.039120
3. S006	1.000000	3. T044	0.031300
4. S030	1.000000	4. S006	0.026080
5. S067	1.000000	5. T005	0.022350
6. T005	1.000000	6. S030	0.013040
7. T044	1.000000	7. S067	0.013040

Tabel 2. Hasil Kueri "citra or komputer" Teknik *Boolean*

	Indeks 1		Indeks 2
1. S091	14.000000	1. S006	0.497870
2. S030	12.000000	2. S030	0.497870
3. T039	9.000000	3. S041	0.497870
4. T042	8.000000	4. S091	0.497870
5. S041	7.000000	5. T044	0.398300
6. S006	6.000000	6. S048	0.311170
7. S040	6.000000	7. T005	0.284500
8. T032	6.000000	8. S040	0.248940
9. S048	5.000000	9. T026	0.199150
10. S055	5.000000	10. S093	0.156480
11. S093	5.000000	11. T032	0.156480
12. T002	5.000000	12. T034	0.149360
13. S005	4.000000	13. S055	0.130400
14. S085	4.000000	14. S005	0.125180
15. S086	4.000000	15. S085	0.089420

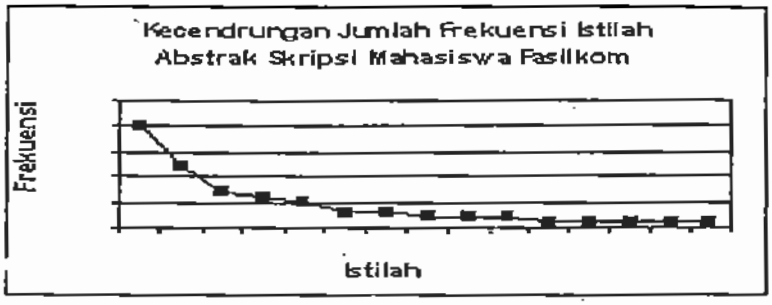
Savoy [1] tidak hanya mempertimbangkan nilai frekuensi saja, tetapi mempertimbangkan posisi relatif dari istilah tersebut. Dengan kata lain rumus Savoy [1] mempertimbangkan seberapa besarnya frekuensi istilah jika dibandingkan dengan frekuensi istilah terbesar pada dokumen, dan juga mempertimbangkan posisi dokumen yang mengandung istilah dimaksud terhadap jumlah dokumen keseluruhannya.

- Kalau dilihat dari hasil indeks 1, dokumen urutan 2 sampai 7 tidak dapat diperingkat dan dianggap sama, karena masing-masing dokumen mempunyai frekuensi kemunculan istilah yang sama. Kalau dilihat pada hasil indeks 2, dokumen-dokumen yang tidak dapat diurutkan pada indeks 1, dapat diperingkat berdasarkan ketinggian bobotnya. Hal ini dapat dilakukan karena rumus Savoy [1] tidak hanya berdasarkan frekuensi kemunculan istilah saja, tetapi juga mempertimbangkan jumlah istilah terbesar dalam dokumen dan jumlah dokumen dalam kumpulan dokumen yang ada.

Hal yang sama berlaku pula untuk operator "OR" seperti terlihat pada tabel 2 dan "NOT".

Dari hasil kueri berdasarkan tabel 2, terlihat perbedaan bobot dokumen yang ditemukan antara indeks 1 dan indeks 2. Hasil dari indek berdasarkan frekuensi (indeks 1), nilai *Retrieval Status Value* ditemukan mempunyai banyak persamaan. Hal ini disebabkan karena banyaknya persamaan frekuensi dari istilah-istilah pada indeks. Hal ini sesuai juga dengan penelitian yang dilakukan, bahwa kecendrungan dari frekuensi istilah pada abstrak skripsi mahasiswa Fasilkom adalah, sedikit dari istilah indeks yang mempunyai frekuensi istilah yang tinggi, dan kebanyakan dari frekuensi istilah bernilai kecil dan bahkan banyak dari istilah itu hanya berfrekuensi satu atau dua pada abstrak tersebut. Kebanyakan dari istilah yang mempunyai frekuensi tinggi tersebut adalah istilah-istilah yang berhubungan dekat dengan domain dari kumpulan dokumen, dalam hal ini adalah domain komputer. Untuk lebih jelas dapat dilihat pada gambar 4.

Pada gambar 4 ini terlihat bahwa hanya sedikit istilah pada dokumen yang mempunyai frekuensi tinggi, dan banyak dari istilah tersebut yang hanya mempunyai frekuensi rendah berkisar antar satu dan dua kali kemunculan. Sebenarnya sebuah indek yang baik itu tidak mempunyai frekuensi yang terlalu tinggi dan juga tidak mempunyai frekuensi yang terlalu rendah. Akan tetapi sejauh mana menentukan batas atas dan batas bawah dari frekuensi istilah yang baik dijadikan indeks, sampai sekarang masih menjadi sebuah isu yang menarik untuk diteliti lebih lanjut.



Gambar 4. Kecendrungan frekuensi istilah pada abstrak skripsi Fasilkom

Tabel 3. Perbandingan Teknik *Boolean* biasa, *Boolean* peringkat dan P-norm dengan memakai sistem *indexing* Savoy [1]

Citra and Komputer -Bool. Biasa-		Citra and Komputer -Bool. Peringkat-		Citra and Komputer P-norm, p=10		Citra and Komputer P-norm, p=5000	
1.S005		1.S005	0.099570	1.T032	0.177624	1.S005	0.099695
2.S006		2.S048	0.039120	2.S093	0.177624	2.S048	0.039253
3.S030		3.T044	0.031300	3.S086	0.160106	3.T044	0.031434
4.S048		4.S006	0.026080	4. S013	0.145057	4.S006	0.027710
5.S067		5.T005	0.022350	5.S006	0.114821	5.T005	0.022486
6.T005		6.S030	0.013040	6.S005	0.111546	6.S030	0.016120
7.T044		7.S067	0.013040	7.S017	0.107115	7.S067	0.013177
				8.S021	0.106112	8.S091	0.004531
				9.S043	0.104643	9.S041	0.004531
				10.S030	0.104018	10.T032	0.000755
				11.S048	0.100304	11.S093	0.000755
				12.T044	0.095401	12.S086	0.000755
				13.S091	0.093214	13.S013	0.000709
				14.S041	0.093214	14.S021	0.000601
				15.T005	0.083877	15.S017	0.000601

Pemakaian sistem pengindeksan berdasarkan frekuensi (indeks 1), lebih sesuai bagi pemakai yang menginginkan dokumen yang ditemukan itu adalah dokumen yang mengandung paling banyak istilah yang terdapat pada kueri. Selain itu pemakai sudah mengetahui persis bahwa informasi yang diinginkannya itu biasanya mengandung suatu istilah yang pasti dan sering terdapat pada informasi yang diinginkannya tersebut. Sistem pembobotan berdasarkan frekuensi ini, kurang cocok diterapkan pada dokumen-dokumen yang mempunyai kecendrungan jumlah frekuensi istilah seperti pada gambar 4 di atas, apalagi dengan memakai teknik *Boolean*. Sistem pembobotan ini kurang mampu untuk memberikan bobot suatu istilah yang lebih unik terhadap sebuah dokumen, sehingga bobot istilah yang dihasilkan mempunyai banyak persamaan. Hal ini juga akan menyulitkan pemakai untuk menentukan mana dokumen yang dijadikan prioritas dalam melakukan *browsing*, apalagi jika dokumen yang ditemukan sangat banyak.

Sistem pengindeksan berdasarkan rumus Savoy [1], pada dasarnya merupakan pengembangan lebih lanjut dari sistem pengindeksan berdasarkan frekuensi. Sistem pengindeksan ini dapat memberikan bobot istilah yang baik terhadap sebuah dokumen. Walaupun suatu istilah mempunyai frekuensi yang sama, tetapi sistem pengindeksan ini dapat memberikan bobot yang berbeda, dengan cara menambah perhitungan dengan faktor lain seperti jumlah dokumen yang mengandung istilah tersebut, atau jumlah frekuensi istilah terbesar. Bobot dokumen yang dihasilkan lebih variatif, dan juga tidak menutup kemungkinan bahwa bobot sebuah istilah pada beberapa dokumen sama. Sistem pembobotan Savoy [1] lebih sesuai diterapkan pada dokumen-dokumen yang mempunyai kecendrungan jumlah frekuensi istilah seperti pada gambar 4 di atas. Sistem pembobotan ini mampu untuk memberikan bobot yang lebih spesifik pada dua dokumen yang punya frekuensi istilah yang sama, sehingga mudah diperingkat.

7.2. Hasil Berbagai Teknik Temu-Kembali Informasi

Pada tabel 3, dapat dilihat hasil dokumen *retrieval* dengan menggunakan teknik *Boolean* biasa,

Boolean berperingkat, dan teknik P-norm dengan menggunakan sistem pengindeksan Savoy [1] memakai operator "and". Berdasarkan tabel 3, hal-hal yang dapat diamati adalah sebagai berikut:

- Perbedaan antara *Boolean* biasa dengan *Boolean* peringkat terlihat dari bobotnya. Pada *Boolean* biasa tidak mempunyai bobot dokumen karena teknik ini hanya menemukan dan menampilkan dokumen berdasarkan urutan kata yang ditemukan pada dokumen. Dari hasil dokumen yang ditemukan jika dibandingkan dengan teknik *Boolean* berperingkat terdapat perbedaan yang mendasar dari segi urutan dokumen yang ditampilkan. Pada teknik *Boolean* biasa dokumen yang ditampilkan paling atas belum tentu mempunyai tingkat relevan yang lebih baik dari dokumen dibawahnya karena teknik ini hanya mempertimbangkan ada atau tidaknya kata-kata kueri pada koleksi dokumen dan tidak mengukur urutan tingkat kerelevanan dokumen tersebut dengan kueri yang dimasukkan.
- Pada teknik *Boolean* berperingkat, telah ada perbaikan dari hasil temu-kembali dimana dokumen yang ditemukan telah diberi bobot dan diperingkat sesuai dengan bobotnya. Ini berarti bahwa pemakai telah diberi kemudahan untuk memilih dokumen yang benar-benar relevan dari dokumen-dokumen hasil yang ditampilkan.
- Perhatikan hasil kueri operasi teknik p-norm, dengan nilai p=10 dan p=5000. Pada saat nilai p=5000, maka terdapat penurunan bobot yang cukup tajam seperti dokumen S093 dan T032. Dokumen-dokumen yang nilai bobotnya tidak terlalu jauh perbedaan mempengaruhi peringkatnya pada saat nilai p=10 adalah dokumen S006, S030 dan S048. Ketiga dokumen ini juga terdapat pada dokumen-dokumen yang dihasilkan oleh operasi *Boolean* dengan operator *"and"*, di mana artinya bahwa ketiga dokumen ini mengandung semua istilah yang ada pada kueri. Sedangkan dokumen-dokumen yang peringkatnya turun adalah dokumen yang mengandung salah satu istilah yang ada pada kueri. Ada juga dokumen yang peringkatnya naik seperti dokumen S048, T044 dan T005. Naiknya peringkat dokumen ini karena dokumen ini juga mengandung semua istilah pada kueri dan peringkatnya naik seiring dengan makin besarnya nilai p. Kalau melihat kembali ke salah satu teori yang mengatakan bahwa jika (T1 AND <p> T2) di mana nilai p mendekati ∞, maka sebuah dokumen akan ditemukan jika kedua istilah T1 dan T2 ada pada dokumen tersebut. Maksudnya adalah jika semakin besar nilai p-nya maka dokumen-dokumen yang dihasilkan mempunyai bobot yang semakin kecil (mendekati 0), di mana penurunan bobot bagi dokumen yang mempunyai semua istilah yang ada pada kueri akan sedikit dan sebaliknya dokumen yang tidak mengandung semua istilah pada kueri maka penurunan bobotnya akan tinggi, sehingga dokumen yang diperoleh nantinya akan terpisah antara dokumen-dokumen yang mengandung semua istilah dengan dokumen-dokumen yang tidak mengandung semua istilah. Dokumen-dokumen yang mengandung semua istilah pada kueri akan diurutkan sama dengan dokumen-dokumen yang ditemukan pada teknik *Boolean* (lihat tabel 3). Hal ini disebabkan karena bobot istilah mempunyai nilai dalam rentang [0,1], sehingga jika dilakukan pemangkatan dengan suatu bilangan yang semakin besar (nilai p) maka akan menghasilkan suatu bilangan yang semakin kecil (mendekati 0), dan hal ini menyebabkan bobot istilah yang paling kecil dalam sebuah kueri terlebih dahulu akan mencapai nilai nol, dan hasil pemangkatan dari rumus RSV dari teknik p-norm akan dipengaruhi oleh bobot istilah yang terbesar. Pada kasus dengan operator *"and"*, *inverse* dari hasil pemangkatan yang dipengaruhi oleh maksimal dari bobot-bobot istilah adalah minimal dari bobot-bobot istilah, sehingga hal ini sama dengan perhitungan peringkat dari teknik *Boolean* yaitu bobot dokumen yang didapat berdasarkan minimal dari bobot istilah pada kueri untuk operator *"and"*.

Selanjutnya kita coba lihat karakteristik dari masing-masing teknik dengan operator "or", berdasarkan tabel 4.

Dari hasil kueri dengan operator *"or"* di atas, jika nilai p-nya semakin besar maka bobot dari masing-masing dokumen akan semakin tinggi (mendekati 1), dan dokumen-dokumen teratas yang ditemukan merupakan dokumen yang mempunyai maksimal bobot dari bobot istilah-istilah yang ada pada kueri, dan dokumen-dokumen yang ditemukan akan sama dengan dokumen-dokumen yang ditemukan pada teknik *Boolean*. Kasus operator *"or"* ini sama dengan operator *"and"* yaitu karena adanya pemangkatan dengan suatu bilangan yang semakin besar (nilai p), dan bilangan yang dipangkatkan mempunyai nilai dalam rentang [0,1], sehingga hasil pemangkatan dari rumus RSV berdasarkan rumus teknik p-norm untuk operator *"or"* akan dipengaruhi oleh bobot istilah yang terbesar, dan hal ini akan sama dengan cara pembobotan dokumen dengan teknik *Boolean* peringkat yaitu berdasarkan maksimal dari bobot istilah yang ada pada kueri untuk operator *"or"*.

Tabel 4. Perbandingan Teknik *Boolean* peringkat dan p-norm dengan memakai sistem *indexing* Savoy [1]

Citra or komputer (indeks 2) –B.Per–		Citra or <1> komputer (indeks 2)		Citra or <100> komputer (indeks 2)		Citra or <405> komputer (indeks 2)	
1. S006	0.497870	1.S006	0.345678	1.S091	0.659241	1. S091	0.662691
2. S030	0.497870	2.S030	0.339882	2.S041	0.659241	2. S041	0.662691
3. S041	0.497870	3.S091	0.334087	3.S030	0.659241	3. S030	0.662691
4. S091	0.497870	4.S041	0.334087	4.S006	0.659241	4. S006	0.662691
5. T044	0.398300	5.T032	0.244873	5.T032	0.486363	5. T044	0.397619
6. S048	0.311170	6.S093	0.244873	6.S093	0.486363	6. S048	0.310638
7. T005	0.284500	7.S086	0.216117	7.S086	0.429249	7. T005	0.284014
8. S040	0.248940	8.T044	0.214800	8.T044	0.395549	8. S040	0.248514
9. T026	0.199150	9.S013	0.193461	9.S013	0.384250	9. T026	0.198809
10. S093	0.156480	10.S048	0.175145	10.S048	0.309021	10.T047	0.000000
11. T032	0.156480	11.T005	0.153425	11.T005	0.282535	11.T046	0.000000
12. T034	0.149360	12.S017	0.140196	12.S017	0.278455	12.T042	0.000000
13. S055	0.130400	13.S021	0.139094	13.S021	0.276266	13.T039	0.000000
14. S005	0.125180	14.S043	0.137278	14.S043	0.272660	14.T038	0.000000
15. S085	0.089420	15.S040	0.124470	15.S040	0.247220	15.T036	0.000000

Tabel 5. Teknik p-norm dengan memakai sistem pengindeksan frekuensi

Citra and <5> komputer (indeks 1)		Citra and <10> komputer (indeks 1)		Citra and <100> komputer (indeks 1)	
1. T047	1.000000	1. T046	0.066967	1. T046	0.006908
2. T036	1.000000	2. T038	0.066967	2. T038	0.006908
3. S045	1.000000	3. T025	0.066967	3. T025	0.006908
4. S032	1.000000	4. T009	0.066967	4. T009	0.006908
5. T046	0.129449	5. S089	0.066967	5. S089	0.006908
6. T044	NaN	6. S088	0.066967	6. S088	0.006908
7. T042	NaN	7. S083	0.066967	7. S083	0.006908
8. T039	NaN	8. S078	0.066967	8. S078	0.006908
9. T038	0.129449	9. S076	0.066967	9. S076	0.006908
10. T034	NaN	10. S067	0.066967	10. S067	0.006908
11. T032	NaN	11. S066	0.066967	11. S066	0.006908
12. T031	1.000000	12. S065	0.066967	12. S065	0.006908
13. T026	NaN	13. S060	0.066967	13. S060	0.006908
14. T025	0.129449	14. S053	0.066967	14. S053	0.006908
15. T023	NaN	15. S046	0.066967	15. S046	0.006908

Untuk lebih jelasnya berdasarkan tabel 4 di atas, dapat kita amati bahwa teknik p-norm mulai dari nilai p=1 sampai nilai p=405, dokumen yang ditemukan berangsur-angsur seiring dengan penambahan nilai p-nya akan sama diperingkat dengan dokumen yang ditemukan pada teknik *Boolean* dengan operator "or".

Untuk kasus kueri citra or komputer dengan nilai p=405, terdapat dokumen yang mempunyai bobot 0.0 adalah disebabkan karena nilai p-nya yang semakin besar, sedangkan nilai bobotnya dalam rentang [0,1], maka sebelum nilai di akar-kan, nilai bobot yang dipangkatkan dengan nilai p telah menjadi nol terlebih dahulu (lihat rumus teknik p-norm dengan operator "or"), sehingga hasil RSV dari dokumen adalah 0.

Pada bagian sebelumnya telah disinggung bahwa teknik p-norm dengan memakai sistem pengindeksan berdasarkan frekuensi tidak menghasilkan dokumen ter-*retrieve* lebih baik dibandingkan dengan memakai teknik Savoy [1] yang dimilikinya sendiri. Sedangkan teknik *Boolean* peringkat yang sebelumnya mempunyai sistem pengindeksan berdasarkan frekuensi, setelah menggunakan sistem pengindeksan Savoy [1], dapat menghasilkan dokumen ter-*retrieve* yang baik, ditandai dengan dapatnya mengurutkan dokumen-dokumen yang ditemukan, dimana sebelumnya tidak dapat dilakukan (lihat tabel 1).

Tabel 6. Persamaan dan perbedaan dari masing-masing teknik

No.	Keterangan	T1	T2	T3
1	Melakukan pembobotan dokumen	-	√	√
2	Melakukan peringkat dokumen	-	√	√
3	Mudah dalam pemakaian operator	√	√	-
4	Membutuhkan pengetahuan tambahan dalam penggunaan operator	-	-	√
5	Adanya fasilitas dalam mengontrol dokumen hasil	-	-	√
6	Pemakaian sistem indexing mempengaruhi hasil dokumen yang ditemukan	-	√	√
7	Mempunyai hasil yang lebih baik jika memakai indeks berdasarkan frekuensi	-	√	-
8	Mempunyai hasil yang lebih baik jika memakai indeks berdasarkan rumus savoy	-	√	√
9	Jumlah dokumen yang dihasilkan dengan memakai operator 'and' dan 'or berbeda	√	√	-
10	Dokumen Teratas merupakan dokumen paling relevan dengan kueri pemakai	-	√	√√
11	Adanya komputasi tambahan pada bagian teknik-nya	-	-	√

Catatan:
T1 = Teknik *Boolean* Biasa
T2 = Teknik *Boolean* Peringkat
T3 = Teknik *Extended Boolean* dengan p-norm model

Untuk lebih jelasnya tentang teknik p-norm dengan sistem pengindeksan berdasarkan frekuensi menggunakan operator "*and*" dapat di lihat pada hasil kueri tabel 5.

Berdasarkan tabel 5 di atas, dapat diamati bahwa teknik p-norm dengan memakai sistem pengindeksan berdasarkan frekuensi kurang baik menghasilkan dokumen ter-*retrieve*. Hal ini dapat dilihat dari bobot dokumen (RSV) yang didapat. Hal ini disebabkan karena teknik p-norm itu mengharuskan bahwa bobot dari indeks istilah tersebut harus dalam rentang [0,1]. Sedangkan bobot dari indeks berdasarkan frekuensi adalah besar dari satu, sehingga RSV dokumen yang di-*retrieve* menyalahi kaedah dari teknik p-norm itu sendiri, dimana RSV/bobot dokumen yang didapat tidak bermakna.

Secara umum persamaan dan perbedaan dari ketiga teknik sistem temu-kembali di atas dapat dilihat pada tabel 6 berikut ini.

8. KESIMPULAN

Penelitian ini bertujuan untuk melakukan integrasi sistem temu-kembali, yaitu teknik *Boolean* biasa, teknik *Boolean* berperingkat dan teknik p-norm ke basis hiperteks. Masing-masing teknik diberi kemampuan untuk dapat saling mengakses sistem pengindeksan yang ada, yaitu sistem pengindeksan berdasarkan frekuensi istilah dan berdasarkan rumus Savoy [1]. Pengintegrasian ke basis hiperteks, membuat sistem lebih mudah diakses oleh pemakai, dan antar muka yang disediakan juga lebih *user friendly* dibandingkan dengan sistem temu-kembali informasi sebelumnya. Penerapan konsep *browsing* pada sistem temu-kembali informasi juga mempermudah proses pencarian informasi, sementara dengan menggunakan teknik-teknik *retrieval* akan mempersingkat proses pencarian informasi.

Dengan diberi kemampuan suatu teknik untuk mengakses lebih dari satu sistem pengindeksan akan memberi banyak pilihan bagi pemakai dalam menemukan informasi yang relevan. Selain itu yang harus diperhatikan adalah tidak setiap teknik temu-kembali informasi cocok dengan semua jenis sistem pengindeksan. Untuk itu pemakai harus tahu benar dengan teknik dan sistem pengindeksan yang dipakai. Hal ini dapat dilihat pada kasus teknik p-norm yang memakai sistem pengindeksan berdasarkan frekuensi. Bobot dari sistem pengindeksan berdasarkan frekuensi ini lebih besar dari satu (tidak dalam range [0,1]), maka bobot dari dokumen-dokumen yang ditemukan juga tidak bermakna.

Bobot-bobot dokumen yang dihasilkan oleh setiap teknik pembobotan juga dipengaruhi oleh data yang di indeks. Pada abstrak skripsi mahasiswa Fasilkom, sedikit dari istilah indeks yang mempunyai frekuensi istilah yang tinggi dan kebanyakan frekuensi istilah bernilai kecil.

REFERENSI

[1] Savoy, J. "A Learning Scheme for Information Retrieval in Hypertext". *Information Processing & Management*, 30(4), 515-533. 1993.

[2] Stolt,Hakan. *Agents, Filter and Search Engines : An evaluating survey on technologies for effective search for information from internet resources.* Department of Computing Science. Umea University. Graduation Thesis. 1997.

[3] Fitriyanti, Masayu. *Sistem Temu-kembali Informasi dengan Mengimplementasikan Operasi Boolean, Sistem Peringkat, Perbaikan Query, dan Pemanfaatan Tesaurus.* Fakultas Ilmu Komputer, Universitas Indonesia. Skripsi. 1997.

[4] Andri, Yofi. *Teknik Learning Scheme Berdasarkan Model P-Norm pada Sistem Temu-kembali Informasi*. Fakultas Ilmu Komputer, Universitas Indonesia, Skripsi.1997.

[5] Salton, G. *Automatic Text Processing: The ransformation, Analysis , and Retrieval of Information by Computer.* Addison-Wesley Publishing Company, Inc. United Sates of America.1989.

[6] Ellist, David. *Progress and Problems in Information Retrieval.* Departement of Information Studies, University of Sheffield.1996

[7] Bodhitama, Ananta D. *Implementasi Local Search Engine Pada Sistem Temu-kembali Informasi*. Fakultas Ilmu Komputer, Universitas Indonesia, Skripsi.1997.

[8] Agosti, Maristella. "Hypertext and Information Retrieval". *Information Processing & Management*, 29(3), 283-285. 1993.

[9] Yuwono,Budi.*Search and Ranking Algorithms for Locating Resources on the World Wide Web.* Information Science,The Ohio State University. 1995.

[10] Dunlop, M.D. et.al. "Hypermedia and Free Text Retrieval". *Information Processing & Management*, 29(3), 287-298. 1993.

[11] Lucarella, D. "Information Retrieval From Hypertext: An Approach Using Plausible Inference". *Information Processing & Management*, 29(3), 299-312. 1993.

[12] Rada, Roy. Et al. "Retrieval Hierarchies in Hypertext", *Information Processing & Management*, 29(3), 359-371. 1993.

[13] Weston, Andrew. *State of the Art in Searching and Indexing the World Wide Web.* Information Science,Queens University. Canada. 1996

[14] Wiersma, Williamm.*Research Methods in Education: An introduction*. The University of Toledo, 1995.