PENGEMBANGAN SISTEM TEMU KEMBALI INFORMASI DIGITAL *FULLTEXT* ARTIKEL JURNAL DI PDII – LIPI

Sjaeful Afandi[1*], Firman Ardiansyah[2], Blasius Soedarsono[3]

[1]Magister Profesional Teknologi Informasi untuk Perpustakaan IPB.
[2]Dosen Jurusan Ilmu Komputer, IPB
[3]Pemerhati Perpustakaan PDII - LIPI

*Korespondensi: denolz@yahoo.com

ABSTRACT

One of the tasks in Center for Scientific Documentation and Information – Indonesian Institutes of Sciences (PDII - LIPI) is to disseminate the results of existing research in Indonesia. The research result can be either books or journal articles. Currently, activity of retrieval system on a journal article have still using a traditional retrieve systems regardless of the relevance of data search results. The order of search results based only on the order of data entry that need to be developed information retrieval of digital full text articles as data retrieval alternative. Development of information retrieval system that uses Sphinx Search software. The data used are the result of the conversion from Portable Digital Format (PDF) into XML as much as 1000 file. Data of conversion result, then are processed through "tokenisasi" and indexing techniques using Sphinx Search software. Retrieval system tested with a query that has been determined. Retrieval results calculated using standard recalls eleven that in mind the relevance and accuracy. Data retrieval system produces search results that are relevant and accurate with average presicion (AVP) value is of 79%.

ABSTRAK

Salah satu tugas PDII – LIPI adalah menyebarkan hasil penelitian yang ada di Indonesia. Hasil penelitian dapat berupa buku atau artikel jurnal. Saat ini, sistem temu kembali pada artikel jurnal masih menggunakan sistem pencarian tradisional tanpa menghiraukan relevansi data hasil pencarian. Urutan hasil pencarian hanya didasarkan pada urutan pemasukan data sehingga perlu dikembangkan sistem temu kembali informasi artikel digital *full text* sebagai alternatif pencarian data. Pengembangan sistem temu kembali informasi tersebut menggunakan perangkat lunak *Sphinx Search.* Data yang digunakan adalah hasil dari konversi *file* PDF ke dalam bentuk XML sebanyak 1000 *file*. Data hasil konversi, kemudian diproses melalui teknik tokenisasi dan pengindeksan menggunakan peranti lunak *Sphinx Search*. Sistem temu kembali diuji coba dengan *query* yang telah ditentukan. Hasil temu kembali dihitung dengan menggunakan sebelas *standard recalls* agar diketahui relevansi dan keakuratannya. Sistem temu kembali menghasilkan data hasil pencarian yang relevan dan akurat dengan nilai *average presicion* (AVP) sebesar 79%.

Keywords: Digital article; Retrieval system; PDII-LIPI

1. PENDAHULUAN

Perkembangan ilmu pengetahuan dan teknologi informasi dewasa ini membuat perubahan perilaku dalam pencarian informasi yang berdampak bagi lembaga-lembaga yang bergerak dalam bidang jasa penyedia informasi, seperti perpustakaan. Perpustakaan harus mampu berperan sebagai penyedia sumber daya informasi dan pengetahuan, serta mampu menyediakan akses ke berbagai sumber daya informasi dan pengetahuan yang efisien. Untuk mendukung peran tersebut, perpustakaan perlu melakukan perubahan dalam pemeliharaan dan katalogisasi informasi, dari sistem tercetak menjadi *online* dalam bentuk digital agar dapat diakses dari mana saja. Perubahan sistem tersebut, terlihat pada pengembangan perpustakaan digital. Layanan perpustakaan digital menyediakan akses instan terhadap koleksi/dokumen, baik melalui metada pencarian *keyword*, penulis, maupun judul.

PDII – LIPI memiliki tugas sebagai pusat penyebaran hasil penelitian yang ada di Indonesia, saat ini telah memiliki artikel jurnal sekitar 96.000 *record* dan sekitar 40.000 artikel *full text* digital yang telah diunggah ke database *Indonesian Scientific Database Journal* (ISJD). Hal tersebut tentunya memerlukan sistem pencarian informasi yang cepat, tepat, dan efektif. Sistem temu kembali informasi saat ini masih berdasarkan *record* judul, pengarang, tahun, subjek, tipe koleksi, dan bidang koleksi. Hasil temuan penelusuran informasi tersebut berdasarkan pada *First In First Out (FIFO)*, data pertama yang dientri akan keluar pada urutan pertama. Pencarian temu kembali informasi dilakukan pada basis data relasional, tanpa ada pembobotan dan relevansi yang lebih akurat.

Penelitian ini dilakukan sebagai alternatif dalam temu kembali informasi di PDII – LIPI melalui pengembangan pencarian informasi artikel digital *full text*, yaitu dengan melakukan proses pembobotan dan perangkingan hasil pencarian pada artikel digital *full text* digital agar hasil pencarian relevan dan akurat.

Penelitian yang dilakukan oleh Bartell, *et al.* (2002) pada sistem temu kembali informasi dokumen digital, menggunakan teknik pembobotan *Vector Space Model* (VSM) dokumen untuk mengurutkan dokumen dari yang paling relevan sampai dengan yang kurang relevan berdasarkan masukan kata kunci dari pengguna. Salton dan Buckley (1987) menyarankan bahwa pengurutan dokumen dapat dilakukan dengan pembobotan indeks. Penelitian lain dari Faren (2005) yang mengenai sistem temu kembali informasi berbasis Model Boolean pada pencarian informasi dalam bentuk *file* .txt.

Tujuan penelitian ini adalah: a) menganalisis dan mengembangkan kinerja sistem temu kembali informasi pada artikel jurnal digital *full text* di PDII – LIPI; dan b) menganalisa relevansi hasil pencarian dari sistem temu kembali pada artikel digital *full text*. Adapun ruang lingkup bahasan penelitian ini adalah:

1) data yang digunakan diambil dari data koleksi PDII – LIPI, antara tahun 2003 – 2010 dengan jumlah 1000 data;
2) bidang artikel yang digunakan adalah manajemen, pendidikan, kesehatan, ekonomi, rekayasa, yang berdasarkan pada jumlah data artikel terbanyak;
3) format file *full text* digital yang digunakan adalah PDF.

Sementara itu, manfaat penelitian ini adalah: 1) sebagai alternatif sistem temu kembali informasi di PDII – LIPI (selain menggunakan metadata); dan 2) mengetahui relevansi hasil pencarian *fulltext* digital yang ada di PDII - LIPI.

2. METODE

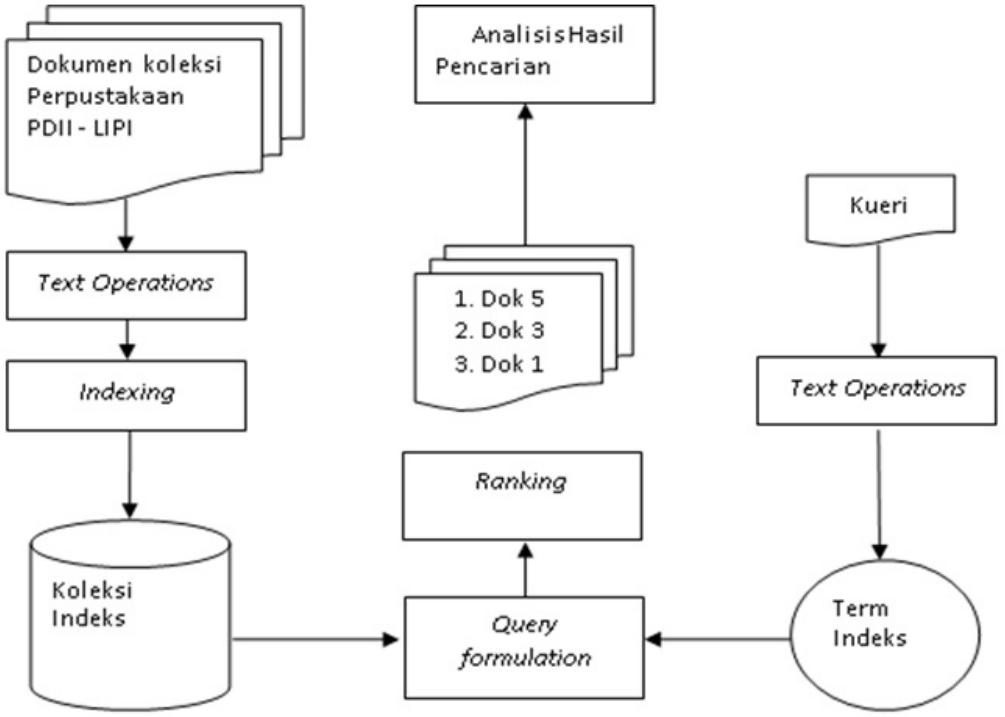
2.1 Jenis Penelitian

Jenis penelitian ini adalah studi kasus (*case study*), yaitu penelitian yang memusatkan perhatian pada suatu kasus tertentu dengan menggunakan individu atau kelompok sebagai bahan studinya. Penelitian ini bersifat deskriptif – eksploratif. Fokus penelitian ini adalah menggali dan mengumpulkan data lebih mendalam terhadap objek yang diteliti agar dapat menjawab permasalahan yang sedang terjadi (Hasibuan, 2007).

2.2 Tahapan Penelitian

Sistem temu kembali informasi menerima *query* dari pengguna, kemudian peneliti melakukan perangkingan terhadap dokumen pada koleksi berdasarkan kesesuaiannya dengan *query*.Hasil

perangkingan merupakan dokumen yang menurut sistem relevan dengan *query* (Mandala dan Setiawan, 2002). Penelitian ini dilaksanakan dalam beberapa tahapan yang diilustrasikan pada Gambar 1. Data yang diproses dalam sistem ini adalah koleksi dokumen artikel jurnal dan *query* yang telah disiapkan sebelumnya.



Gambar 1. Tahapan penelitian

2.3 Pengumpulan Dokumen

Metadata koleksi artikel jurnal yang digunakan antara tahun 2003 – 2010 yang berjumlah 17.377 *record*. Data tersebut diurutkan berdasarkan bidang artikel jurnal digital *full text* yang paling banyak. Dari hasil penelusuran informasi, dihasilkan 5 bidang artikel dengan urutan terbesar, yaitu manajemen, pendidikan, ekonomi, kesehatan, dan hukum (Tabel 1). Artikel jurnal yang digunakan sebanyak 100 dengan pembagian masing-masing bidang sebanyak 200 dokumen.

Tabel 1. Jumlah Artikel pada Sepuluh Besar Bidang Jurnal

No	Bidang	Jumlah
1	Manajemen	4713
2	Pendidikan	4141
3	Ekonomi	3623
4	Kesehatan	2697
5	Hukum	2170
6	Rekayasa	2107
7	Pertanian	1770
8	Biologi	1306
9	Lingkungan	1209
10	Kimia	1202

Proses pengumpulan artike ljurnal digital dilakukan dengan menggunakan *query* pada subjek metadata di database (Tabel 2). Total *query* yang digunakan dalam pengambilan data sebanyak 30 *query,* terdiri dari:

- satu *query* untuk dua kata subjek;
- dua *query* untuk satu kata subjek yang sama antara dua bidang atau lebih;
- tiga *query* untuk satu kata subjek.

Query yang digunakan diambil secara acak dan tiap bidang harus diperoleh sebanyak 200 *record*. Kata yang didapat dari *query* digunakan untuk menguji sistem temu kembali yang dikembangkan. Jumlah dokumen hasil *query* dapat dilihat di Tabel 2. Pada proses pembuatan *corpus,* penentuan artikel jurnal digital yang relevan ditentukan oleh peneliti. Artikel jurnal digital yang relevan dengan *query* ditandai dengan nomor identitas untuk membantu proses pengolahan data.

2.4 Pra-Proses

Tahapan ini menjalankan praproses pengujian sistem yang terdiri dari.

a. *Text Operations* (operasi terhadap teks), meliputi pemilihan kata-kata dalam *query* atau dokumen (*term selection*) dalam transformasi dokumen menjadi indeks dari kata-kata yang ada. Tahap ini meliputi proses *lowercasing*, tokenisasi, dan pembuangan *stopwords*. *Lowercasing* adalah proses untuk mengubah semua huruf menjadi huruf *non-capital* agar menjadi *case-insensitive* pada saat dilakukan pemrosesan teks dokumen.

Tabel 2. Query Pengumpulan Data

Bidang	Query	Jumlah Dokumen Relevan
Ekonomi	economic development	34
	economic	46
	agroindustry	32
	bankruptcy	25
	transportation	33
	tax	30
Hukum	legal aspects	51
	crimes	19
	contract	48
	agricultural	17
	government	54
	human	11
Kesehatan	avian influenza	35
	dental	33
	nutritional	27
	age	58
	malpractice	17
	medical	30
Pendidikan	teaching material	40
	achievement	35
	accounting	28
	accountants	25
	activity	39
	education	33

Bidang	Query	Jumlah Dokumen Relevan
Manajemen	accounting information	23
	marketing	60
	bureaucracy	21
	entrepreneurs	16
	strategic	11
	auditing	69

Sementara itu, tokenisasi adalah suatu tahap pemrosesan *input* teks yang dibagi menjadi unit-unit kecil yang disebut *token* atau *term*, yang dapat berupa suatu kata atau angka. Proses tokenisasi dilakukan sesuai dengan aturan berikut:

- *whitespace*, berarti karakter ini merupakan karakter pemisah token;
- *alpha*, berarti karakter ini merupakan huruf;
- *numeric*, berarti karakter ini merupakan angka;
- *other*, berarti karakter ini tidak temasuk *whitespace*, *alpha*, dan *numeri;*
- token yang terdiri atas karakter numerik saja tidak diikutsertakan.

Stopwords merupakan daftar kata-kata yang dianggap tidak memiliki makna. Kata yang tercantum dalam daftar ini dibuang dan tidak ikut diproses pada tahap selanjutnya.

b. *Indexing* (pengindeksan) dilakukan untuk membangun indeks dari koleksi dokumen dalam bentuk XML.

c. *Query formulation* (formulasi *query*), yaitu memberi bobot pada indeks kata-kata *query*.

d. *Ranking* (perangkingan), dilakukan dengan memberikan nilai pada dokumen-dokumen yang relevan terhadap *query* dan mengurutkannya berdasarkan relevansi dengan *query.*

3. HASIL DAN PEMBAHASAN

3.1 Koleksi Dokumen Pengujian

Koleksi dokumen yang digunakan untuk menguji sistem berasal dari koleksi *file* artikel jurnal digital PDII-LIPI yang berbentuk PDF. Format ini diubah ke struktur *tag* XML dengan *field* yang digunakan dalam dokumen, terdiri dari:

- <*doc*></*doc*> mewakili keseluruhan isi dokumen, di dalamnya terdapat *tag* lain yang mendeskripsikan isi dokumen secara lebih jelas;
- <*docno*></*docno*> mewakili ID dokumen, ID dokumen yang digunakan adalah ID artikel jurnal di dalam basis data;
- <*title*></*title*> mewakili judul dokumen;
- <*keyword*></*keyword*> mewakili kata kunci dokumen;
- <*abstract*></*abstract*>mewakili abstrak dari dokumen.

Konversi *file* PDF ke XML untuk 100 *file* dilakukan secara manual dengan menggunakan perangkat lunak Adobe Acrobat Pro Extended 9. Hal ini dilakukan karena hasil OCR dengan menggunakan script PHP tidak memberikan hasil teks yang baik dan mengalami kerusakan seperti terlihat pada Gambar 2.

Gambar 2. Hasil konversi PDF ke XML yang mengalami kerusakan teks

Dengan mempertimbangkan masalah hasil konversi, data yang diambil dari sebuah dokumen dibatasi pada judul, *keyword*, dan abstrak sebagai sumber informasi (Gambar 3). Format hasil akhir XML yang diperoleh ditunjukkan pada Gambar 4.

Gambar 3. Hasil konversi PDF ke XML berisi judul, abstrak, dan *keyword*

Gambar 4. Hasil akhir konversi PDF ke XML

3.2 Pra-Proses

Tahapan praproses dalam temu kembali merupakan tahapan pembangunan indeks dari koleksi dokumen yang akan digunakan. Kualitas indeks mempengaruhi efektivitas dan efisiensi sistem temu kembali informasi (Chu, *et al*., 2002). Indeks dokumen adalah himpunan *term* yang menunjukkan isi atau topik yang dikandung oleh dokumen. Ukuran indeks yang kecil dapat memberikan hasil buruk dan mungkin beberapa *item* yang relevan terabaikan. Indeks yang besar memungkinkan ditemukan banyak dokumen yang relevan, sekaligus dapat menaikkan jumlah dokumen yang tidak relevan dan menurunkan kecepatan pencarian (Hyusein dan Patel, 2003).

Tahapan proses *text operation*, *indexing*, *query formulation* dan *ranking* dilakukan oleh Sphinx Search. Pada tahap awal dilakukan proses konfigurasi dokumen, sebagai berikut.

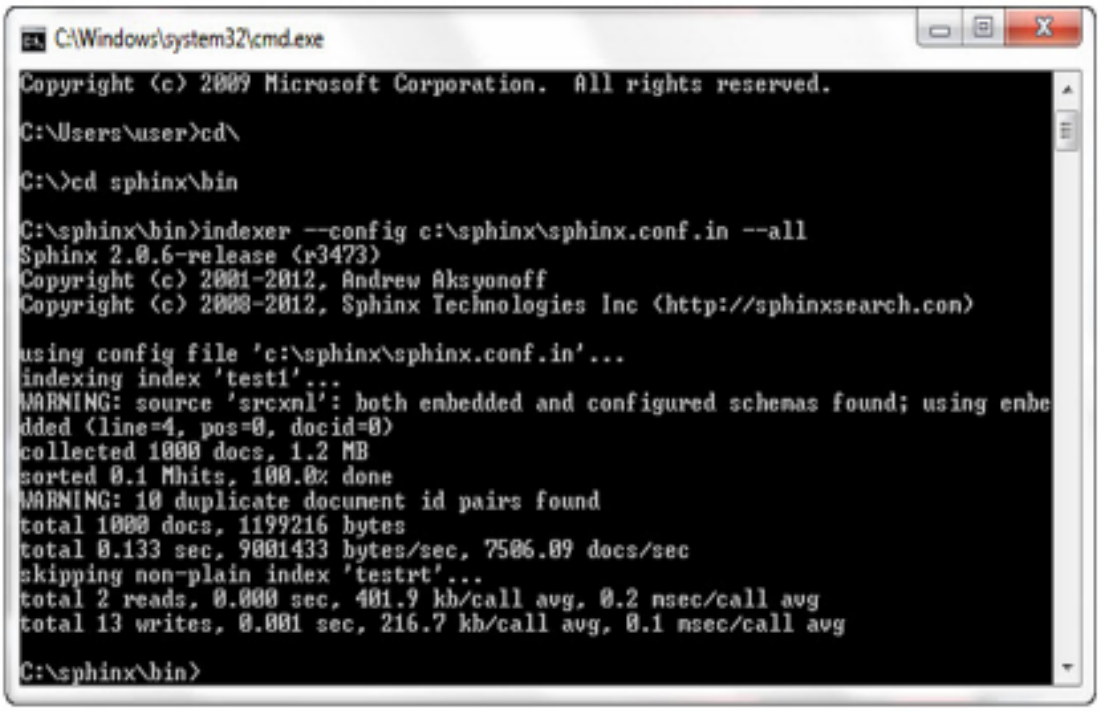
```
index test2
{
  source = srcxml
  path = c:/sphinx/data/test2
  docinfo = extern
  min_word_len = 3
  charset_type = utf-8
  enable_star = 0
  html_strip = 0
  stopwords = c:/sphinx/data/stopwords.txt
}
```

Penjelasan untuk konfigurasi di atas, sebagai berikut.

- source = srcxml, konfigurasi untuk menandakan bahwa sumber yang digunakan atau *file* yang akan diindeks berupa *file* dengan format XML.
- path=c:/DTF/data/test2, konfigurasi untuk mengatur lokasi *file* hasil *indexing* disimpan.
- docinfo = extern, konfigurasi untuk penyimpanan dokumen hasil *indexing*. Dalam hal ini, nilai *extern* menunjukkan bahwa hasil *indexing* akan disimpan dalam *file* terpisah dengan nama *file* yang sama.
- min_word_len = 3, konfigurasi ini menjelaskan panjang minimal kata yang diindeks, yaitu minimal 3 karakter.

- charset_type = utf-8, konfigurasi ini menunjukkan tipe karakter yang digunakan, yaitu utf-8.
- enable_star = 0, konfigurasi untuk pengindeksan prefiks. Digunakan nilai 0 yang menunjukkan bahwa tidak dilakukan pengindeksan untuk prefiks.
- html_strip = 0, konfigurasi untuk menghilangkan *tag*. Digunakan nilai 0 yang berarti tidak menghilangkan *tag*.
- stopwords=c:/sphinx/data/stopwords.txt, merupakan konfigurasi untuk eliminasi kata buangan.

Hasil *indexin*g terlihat pada Gambar 5, besar *file* hasil indeks adalah 1.2 MB untuk 1000 dokumen yang digunakan di sistem. Waktu yang dibutuhkan untuk *indexing* adalah 0.133 detik. Pembobotan dokumen hasil temu kembali informasi menggunakan ukuran kesamaan BM25 yeng merupakan fungsi *default Spinx Search* (SPH_RANK_PROXIMITY_BM25). Urutan dokumen yang ditampilkan sesuai dengan kemiripan antara suatu dokumen dengan *query* yang diberikan menggunakan SPH_SORT_RELEVANCE.



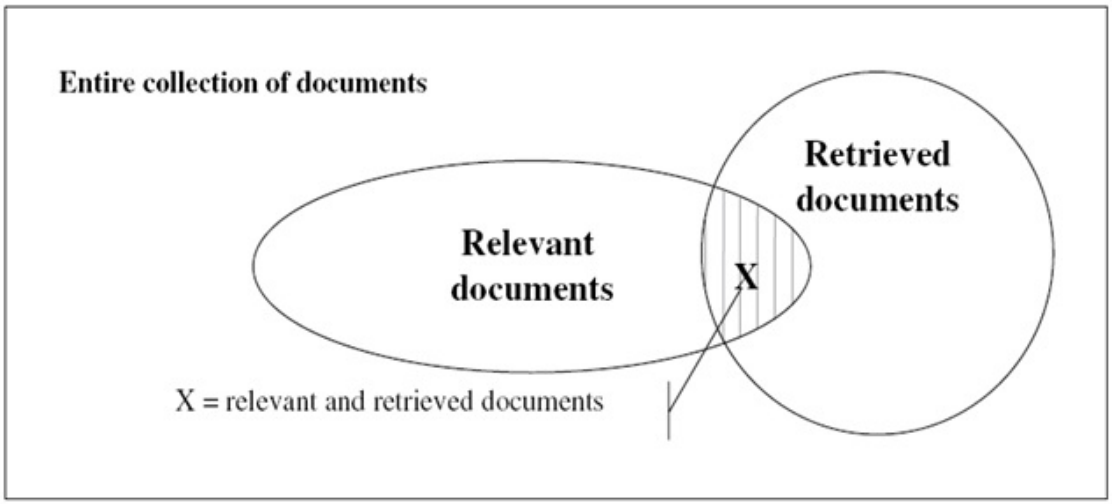
Gambar 5. Hasil *indexing* dokumen

3.3 Pengujian Sistem Temu Kembali

Proses pengujian sistem temu kembali dokumen dilakukan oleh lima orang, terdiri dari 2 orang peneliti, 2 orang pustakawan, dan 1 orang fungsional umum. Alasan pemilihan fungsional peneliti, pustakawan, dan umum adalah agar diperoleh keragaman hasil temu kembali informasi serta dapat diketahui perbedaan persepsi hasil pencarian (benar atau salah). Fungsional peneliti dipilih karena pengunjung perpustakaan PDII-LIPI didominasi oleh peneliti, baik yang datang langsung maupun yang melakukan akses dari katalog *online*. Sementara itu, fungsional pustakawan dibutuhkan dalam pengecekan bibliografi koleksi perpustakaan, dan fungsional umum dipilih karena posisi ini relatif tidak memperdulikan teknik penelusuran informasi tetapi hanya fokus pada hasil pencarian sesuai dengan keinginan.

3.4 Evaluasi

Sistem temu kembali mengembalikan satu set dokumen sebagai jawaban atas *query* pengguna. Terdapat dua kategori dokumen yang dihasilkan oleh sistem temu kembali terkait pemrosesan *query*, yaitu *relevant document* (dokumen yang relevan dengan *query*) dan *retrieved document* (dokumen yang diterima pengguna). Gambar 6 menunjukkan hubungan antara kedua kategori ini digambarkan menggunakan diagram Venn (Cios, *et al.* 2007).



Gambar 6. Relasi antara *relevant* dan *retrieved* dokumen

Ukuran umum yang digunakan untuk mengukur kualitas dari *text retrieval* adalah kombinasi *precision* dan *recall*. *Precision* mengevaluasi kemampuan sistem temu kembali untuk menemukan kembali *top-ranked* yang paling relevan, dan didefinisikan sebagai presentase dokumen yang ditemukembalikan relevan terhadap *query* pengguna.

$$Precision = \frac{X}{retrieved\_documents}$$

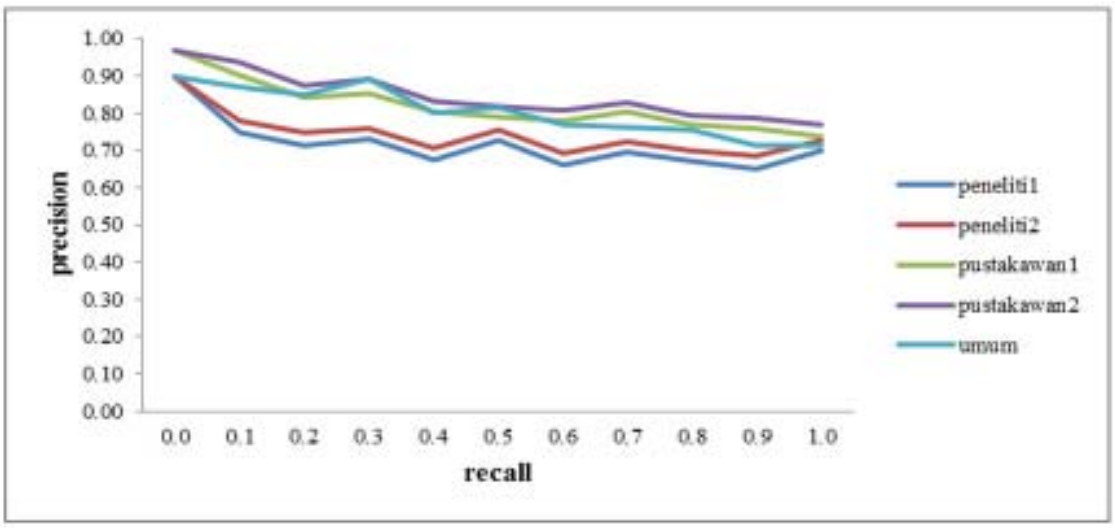
Sementara itu, *recall* mengevaluasi kemampuan sistem temu kembali untuk menemukan semua *item* yang relevan dari dalam koleksi dokumen dan didefinisikan sebagai presentase dokumen yang relevan terhadap query pengguna dan yang ditemukembalikan.

$$Recall = \frac{X}{relevant\_douments}$$

Pada tahap evaluasi dilakukan penilaian tingkat keefektivan proses temu kembali terhadap sejumlah koleksi pengujian dengan menghitung nilai *recall* dan *precision* dari proses temu kembali dokumen berdasarkan penilaian relevansinya. Nilai *recall* yang digunakan adalah 11 *standard recalls* (0,0.1,0.2,...,1). Nilai ini menunjukkan jumlah bagian dokumen dari seluruh dokumen terambil

untuk perhitungan nilai *precision*. Misalkan, untuk nilai *recall* 0.1 berarti jumlah dokumen yang digunakan untuk perhitungan nilai *precision* adalah 10% dari seluruh dokumen relevan yang ada. Nilai *precision* untuk nilai *recall* 0.1 adalah perbandingan banyaknya dokumen relevan yang terambil dari seluruh dokumen dengan jumlah tersebut. Dengan menggunakan nilai dari *recall* dan *precision* akan dicari nilai dari *average precision* untuk menghitung keefektifian dan keakuratan dari sistem temu kembali. *Average precision* adalah suatu ukuran evaluasi sistem temukembali informasi yang diperoleh dari cara menghitung rata-rata *precision* pada seluruh tingkat *recall* (Grossman, 1992).

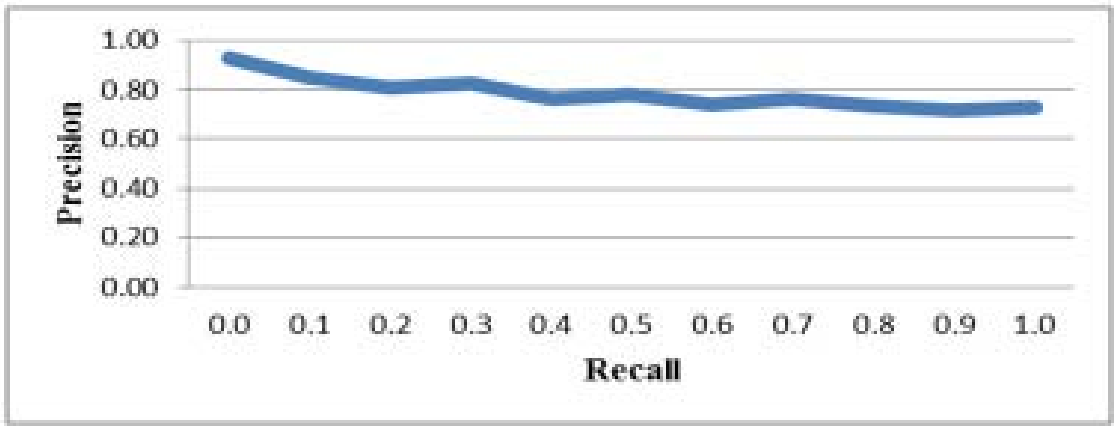
Grafik yang diperoleh dari hasil perhitungan *recall* dan *precision* (Gambar 7) menunjukan bahwa fungsional peneliti mengambil semua dokumen yang relevan saja sesuai dengan query yang dimasukan ke dalam sistem temu kembali. Kinerja sistem temu kembali terhadap hasil pencarian cukup bagus, dengan nilai *average precision* antara 0.74 sampai dengan 0.85, artinya antara 74% sampai dengan 85% dokumen yang ditemukan oleh sistem bersifat relevan bagi pengguna.



Gambar 7. Hasil perhitungan *precision* dan *recall*

Pustakawan mengambil semua hasil temu kembali, baik yang relevan maupun yang mendekati relevan. Hal ini didasarkan bahwa pustakawan membutuhkan banyak data yang diambil dalam pengecekan metadata yang ada, juga untuk alternatif masukan temu kembali dokumen untuk pengguna. Kinerja sistem temu kembali cukup baik, yaitu antara 72% - 80% dokumen berhasil ditemukan oleh sistem. Kinerja sistem temu kembali pada fungsional umum sebesar 82% dokumen berhasil ditemukan. Fungsional Umum mengambil semua atau beberapa saja yang relevan, karena yang terpenting dokumen yang dicari ditemukan.

Gambar 8 menggambarkan bahwa nilai *average precision* (AVP) dari hasil pengujian keseluruhan sebesar 0.79. Hasil uji kinerja sistem menunjukan bahwa kinerja sistem sudah cukup baik disebabkan sistem telah menemukembalikan sebesar 79% dokumen relevan yang dicari oleh pengguna.

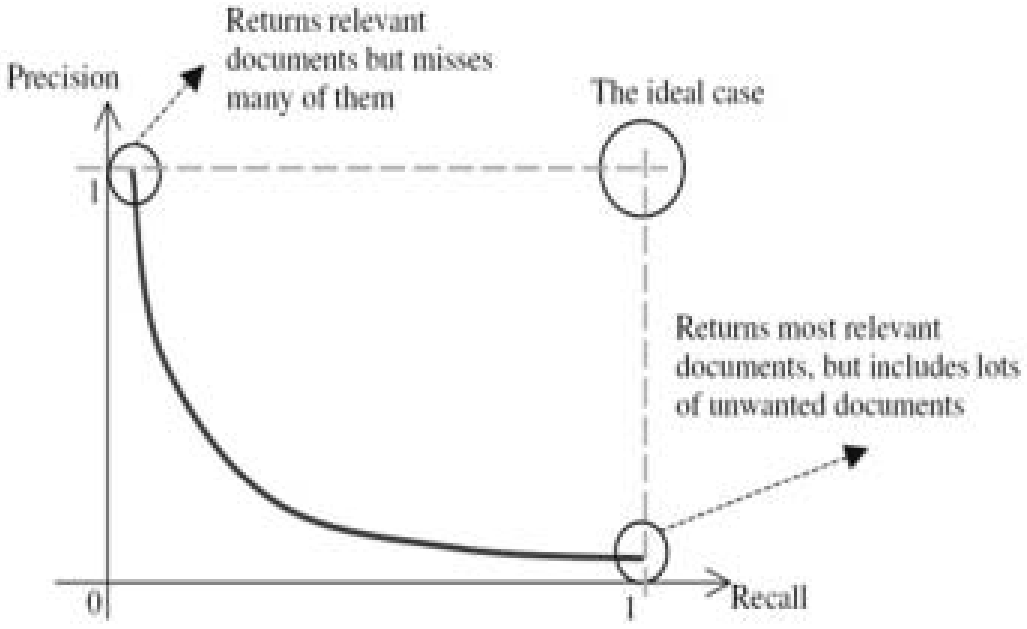


Gambar 8. Hasil perhitungan *average precision* dari lima pengujian sistem

Sistem temu kembali yang baik diharapkan untuk dapat memberikan nilai *precision* dan *recall* mendekati , seperti pada Gambar 9. *Precision* dan *recall* adalah faktor penting dalam mengevaluasi sistem temu kembali, tetapi ada keadaan *trade-off* (tarik-ulur) antara *recall* dan *precision* (Zhang, 2008). Kondisi *trade-off* antara *precision* dengan *recall* mengakibatkan terjadi 2 situasi ekstrim berikut (*Cios, et al.* 2007).

a. *Recall* sangat tinggi dan *precision* relatif rendah. Sistem mengembalikan sejumlah besar dokumen yang mengikut sertakan hampir semua dokumen relevan, tetapi juga mencakup sebagian besar dokumen yang tidak diharapkan.

b. *Precision* terlalu tinggi dan *recall* rendah. Sistem mengembalikan beberapa dokumen dan hampir semuanya relevan, tetapi sejumlah besar dokumen relevan lain terabaikan.



Gambar 9. Trade-off antara *precision* dan *recall* (Zhang 2008)

Gambar 9 menunjukkan bahwa kondisi ideal dari sistem temu kembali pada nilai 1 untuk posisi *precision* dan *recall*. Kinerja dari pengembangan sistem temu kembali artikel jurnal digital PDII – LIPI mendekati nilai 1 untuk *precision* (Gambar 8). Hal ini berarti sistem mengembalikan

sejumlah besar dokumen yang mengikutsertakan hampir semua dokumen relevan, tetapi masih menemukembalikan sebagian dokumen yang tidak diharapkan.

4. KESIMPULAN

Berikut ini merupakan kesimpulan dari implementasi dan ujicoba terhadap kinerja sistem temu kembali informasi pada artikel digital *full text* di PDII – LIPI: a) *software Sphinx Search* cukup membantu pengguna dalam temu kembali informasi, dilihat dari analisis hasil temu kembali informasi pada sistem yang telah dibuat dan diujicobakan; b) tidak konsistennya pengguna dalam mencari dan menghitung hasil pencarian; c) erlu penyeragaman bahasa, agar mempermudah proses *index, stopword* dan *stemming* data; d) evaluasi yang dilakukan terhadap seluruh hasil pengujian menunjukkan bahwa sistem menemukan sejumlah besar dokumen yang dicari dengan nilai AVP sebesar 0.79, di mana sistem mendekati kesesuaian hasil pencarian yang diinginkan; e) penggunaan koleksi dokumen yang lebih banyak dengan topik yang bervariasi akan memberikan perbedaan pada saat memilih kata yang akan digunakan sebagai *query*.

Untuk pengembangan lebih lanjut mengenai sistem temu kembali informasi pada artikel digital *full text* di PDII – LIPI, peneliti menyarankan sebagai berikut. a) pengembangan selanjutnya aplikasi dilengkapi dengan fasilitas yang dapat menangani kesalahan inputan dari pengguna serta fasilitas *auto complete* yang dapat memudahkan pengguna memilih kata kunci; b) proses OCR tidak menggunakan *software* Adobe Acrobat, agar mempermudah proses konversi PDF dalam bentuk teks; c) perlu penyeragaman bahasa, agar mempermudah proses *index, stopword* dan *stemming* data.

DAFTAR PUSTAKA

Bartell, et al. 2002. *Optimizing Parameters in a Ranked Retrieval System Using Multi-Query Relevance Feedback*. California: Institute of Neural Computation and Departement of Science & Engineering University of California, San Diego.

Cios, et al. 2007. *Data Mining a Knowledge Discovery Approach*. Springer.

Faren. 2005. "Implementasi Operasi Boolean Sistem Retrieval Informasi untuk Dokumen Digital". *Skripsi.* Yogyakarta: Program Studi Ekstensi Ilmu Komputer Jurusan Matematika Universitas Gadjah Mada.

Hasibuan, Z.A. 2007. *Metodologi Penelitian Pada Bidang Ilmu Komputer dan Teknologi Informasi* Jakarta: Fakultas Ilmu Komputer Universitas Indonesia.

Mandala R. , Setiawan H. 2002. *Peningkatan Performansi Sistem Temu Kembali Informasi dengan Perluasan Query secara Otomatis.* Bandung: Institut Teknologi Bandung.

Salton G. and Buckley C. 1987. "Term Weighting Approaches in Automatic Text Retrieval". *Technical Report No. 87-881*. New York: Department of Computer Science Cornell University Ithaca.

Zhang, H. 2008. "Formulating Complex Queries Templates". *A Thesis*. Canada: University of Waterloo.