REPOSITORI DIGITAL BERBASIS OAI DAN RANTAI KUTIPAN

Adi Wibowo[1], Resmana Lim[2]
Jurusan Teknik Informatika, Fakultas Teknologi Industri, Universitas Kristen Petra
Jl. Siwalankerto 121 - 131, Surabaya, 60236
Telp. (031) 8439040 ext. 3455
E-mail: adiw@petra.ac.id, resmana@petra.ac.id

ABSTRAK
Sebuah organisasi dapat memiliki sebuah repositori digital. Repositori digital adalah sebuah gudang file digital yang dilengkapi dengan metadata untuk mendeskripsikan file-file digital tersebut. File digital dapat berasal dari dalam organisasi tersebut, dan juga berasal dari organisasi eksternal yang mengijinkan koleksi file digitalnya digunakan oleh umum. Untuk mengambil file digital dari repositori eksternal digunakan protokol terbuka yaitu Open Archives Initiative Protocol for Metadata Harvesting. Setelah metadata dan file digital dari repositori eksternal digabungkan dengan koleksi file digital internal, maka perlu dibuat sistem pencarian yang memanfaatkan rantai kutipan yang melibatkan file-file digital tersebut. Nilai dari rantai kutipan dapat membantu menyusun ulang peringkat hasil pencarian agar file-file yang berkualitas tinggi dapat menduduki posisi lebih atas dibandingkan file-file yang berkualitas lebih rendah.

Kata Kunci: repositori digital, oai, citation network

1. PENDAHULUAN

Repositori digital adalah sebuah gudang penyimpanan file-file digital yang kemudian dibagikan (shared) kepada pengguna baik pengguna lokal, maupun pengguna umum melalui internet. Setiap file digital dilengkapi dengan metadata yang mencatat judul, pengarang, deskripsi, tanggal terbit, dan relasi dengan file digital lainnya. Bila file-file digital tersebut adalah artikel dari jurnal, atau makalah penelitian, maka relasi yang dimaksud di atas adalah bahwa file digital tersebut mengutip beberapa file digital lainnya, atau juga dikutip oleh file-file digital lainnya.

Sebuah repositori digital juga dapat berbagi koleksi metadata dan file digitalnya dengan repositori digital lainnya. Biasanya proses berbagi koleksi ini dilakukan melalui kerjasama antar institusi menggunakan *proprietary protocol*. *Proprietary protocol* menyebabkan interoperabilitas antar organisasi repositori digital di seluruh dunia menjadi sulit karena antar protokol tidak selalu memiliki standar metadata yang dapat saling mendukung.

Open Archives Initiative (OAI) adalah sebuah organisasi nirlaba yang memiliki tujuan menyediakan protokol pertukaran metadata yang terbuka (open). *Open* dimaksudkan bahwa protokol tersebut bisa didapatkan secara bebas oleh setiap organisasi yang membutuhkannya. OAI menyediakan protokol untuk memanen (*harvest*) koleksi-koleksi dari beberapa repository digital yang disebut *OAI Protocol for Metadata Harvesting* (OAI-PMH) (OAI, 2002).

Tujuan penelitian ini adalah membangun sebuah repositori digital berbasis OAI yang juga memanfaatkan relasi berupa rantai kutipan (*chain of reference*) antar file-file digital dalam koleksinya.

2. OAI PROTOCOL FOR METADATA HARVESTING (OAI-PMH)

OAI-PMH pada dasarnya adalah sebuah implementasi protokol *web services* berbasis REST. Arsitektur REST terdiri atas server dan client. REST client di OAI-PMH menggunakan operasi GET dan POST untuk mengambil metadata koleksi yang disimpan oleh server. Data yang dikirimkan dari server menuju ke client berbentuk dokumen XML.

Pada OAI-PMH terdapat beberapa *verb*. *Verb* menunjukkan jenis operasi yang diminta oleh client kepada server. *Verb* digunakan baik untuk mengetahui format metadata yang didukung oleh sebuah repositori digital, untuk mengambil satu koleksi dari server, atau mengetahui kategori-kategori yang disediakan oleh server repositori digital. Daftar *verb* lengkap ditunjukkan pada tabel 1.

Salah satu repositori digital yang mengimplementasikan OAI-PMH adalah CiteSeerX dari Penn State College of Information Sciences and Technology. Contoh operasi GET yang menggunakan verb GetRecord, dan dokumen XML hasil dari operasi tersebut ditunjukkan pada tabel 2.

Tabel 1. Daftar verb dari OAI-PMH

Verb	Fungsi
GetRecord	Mengambil satu record koleksi dari server
Identify	Mendapatkan versi protokol OAI-PMH yang didukung oleh server, email administrator, system penghapusan record, dan tingkat detail dari tanggal.

Verb	Fungsi
ListIdentifiers	Mendapatkan sekumpulan header koleksi.
ListMetadata Formats	Mendapatkan format metadata yang didukung oleh server.
ListRecords	Mendapatkan sekumpulan koleksi sesuai kriteria tanggal atau set tertentu
ListSets	Mendapatkan set (kategori) dari koleksi di server.

Tabel 2. Operasi GET menggunakan verb GetRecord, dan hasil dokumen XML di CiteSeerX

Operasi GET:
http://citeseerx.ist.psu.edu/oai2?verb=GetRecord&identifier=oai:CiteSeerXPSU:10.1.1.40.5588&metadataPrefix=oai_dc

Sebagian dokumen XML yang dihasilkan:

```
<OAI-PMH
xsi:schemaLocation="http://www.openarchives.
org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-
PMH.xsd">
<request
identifier="oai:CiteSeerXPSU:10.1.1.40.5588"
metadataPrefix="oai_dc"
verb="GetRecord">http://citeseerx.ist.psu.ed
u/oai2</request>
<GetRecord>
<record>
<header>
<identifier>oai:CiteSeerXPSU:10.1.1.40.5588<
/identifier>
<datestamp>2009-04-11</datestamp>
</header>
<metadata>
<oai_dc:dc
xsi:schemaLocation="http://www.openarchives.
org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.x
sd">
<dc:title>A Method for Obtaining Digital
Signatures and Public-Key
Cryptosystems</dc:title>
<dc:creator>R.L. Rivest</dc:creator>
<dc:creator>A. Shamir</dc:creator>
<dc:creator> L. Adleman</dc:creator>
<dc:subject>the difficulty of factoring the
published divisor</dc:subject>
<dc:description>An encryption method is
presented ...</dc:description>
<dc:contributor>CiteSeerX</dc:contributor>
<dc:date>2009-04-11</dc:date>
<dc:date>2007-11-22</dc:date>
<dc:date>1978</dc:date>
<dc:format>application/postscript</dc:format
>
<dc:type>text</dc:type>
<dc:identifier>http://citeseerx.ist.psu.edu/
viewdoc/summary?doi=10.1.1.40.5588</dc:ident
ifier>
<dc:source>http://www.matha.mathematik.uni-
dortmund.de/~fv/diplom_i/ars78.ps</dc:source
>
<dc:language>en</dc:language>
<dc:relation>10.1.1.37.9720</dc:relation>
<dc:relation>10.1.1.116.2833</dc:relation>
<dc:relation>10.1.1.115.3569</dc:relation>
<dc:rights>Metadata may be used without
restrictions as long as the oai identifier
remains attached to it.</dc:rights>
</oai_dc:dc>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

3. FORMAT METADATA

Pada table 2 terlihat bahwa CiteSeerX menggunakan format metadata (metadata prefix) "oai_dc". OAI_DC adalah format Dublin Core dengan spesifikasi khusus dari OAI.

Fornat Dublin Core (DCMI, 2010) adalah format metadata untuk dokumen elektronik. Format Dublin core menggunakan 15 elemen untuk menyimpan data tentang sebuah file elektronik. Elemen-elemen tersebut adalah *contributor*, *coverage*, *creator*, *date*, *description*, *format*, *identifier*, *language*, *publisher*, *relation*, *rights*, *source*, *subject*, *title*, dan *type*. Setiap institusi yang mengadopsi Dublin Core boleh menambah elemen baru yang dibutuhkan, atau menggunakan beberapa elemen yang dibutuhkannya saja.

Dari *xml schema definition* yang diberikan oleh OAI format Dublin Core yang digunakan oleh OAI menggunakan 15 elemen tersebut tanpa tambahan elemen baru. Setiap elemen dapat tidak memiliki isi apapun, tetapi juga dapat berisi jumlah data tak terbatas. Jadi sebuah file digital dapat tidak memiliki judul, dan file digital yang lain dapat memiliki judul dalam jumlah yang tidak dibatasi. Elemen *relation* dalam oai_dc menyimpan identifier dari file-file digital lainnya yang mengutip file tersebut, dan juga identifier dari file-file digital lainnya yang dikutip oleh file digital tersebut.

4. IMPLEMENTASI

Sistem pada implementasi ini disebut sebagai Web Content Management. File digital yang disimpan oleh sistem adalah artikel dari jurnal, atau paper dari sebuah konferensi ilmiah.

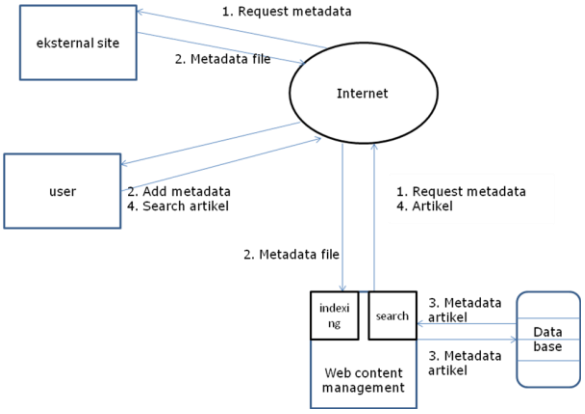
4.1 Sistem Keseluruhan

Desain sistem keseluruhan ditunjukkan pada gambar 1. Sistem melakukan *request metadata* pada situs eksternal seperti CiteSeerX menggunakan protokol OAI-PMH dan menyimpan metadata dan file digital yang diperoleh ke dalam sistem. Selain itu *user* (anggota sebuah organisasi) juga dapat memasukkan metadata dan file digital internal organisasi ke dalam sistem. Metadata akan disimpan ke sebuah database.

Setelah metadata dan file digital tersimpan, sistem secara berkala akan melakukan proses *indexing* pada metadata dan file digital tersebut. Metode *indexing* yang digunakan tidak ditentukan secara spesifik oleh penelitian ini. Organisasi dapat

 ISSN: 1907-5022

memilih antara pendekatan berbasis fuzzy, vector, probabilistik atau pendekatan lainnya. Selain sistem indexing yang metodenya ditentukan sendiri oleh organisasi tersebut, sistem juga akan menghitung sebuah nilai file digital berdasarkan rantai kutipannya (*chain of reference*).



Gambar 1. Sistem Keseluruhan

4.2 Struktur Database File Digital

Untuk menyimpan metadata dari file digital yang disimpan di repositori digunakan struktur database yang ditunjukkan pada gambar 2.

4.3 Sistem *Indexing* dan *Searching* File Digital

Sistem pencarian yang digunakan pada penelitian ini adalah OKAPI BM25. Pendekatan ini dipilih karena proses *indexing* dan *searching* yang cepat dibandingkan pendekatan *information retrieval* lainnya.

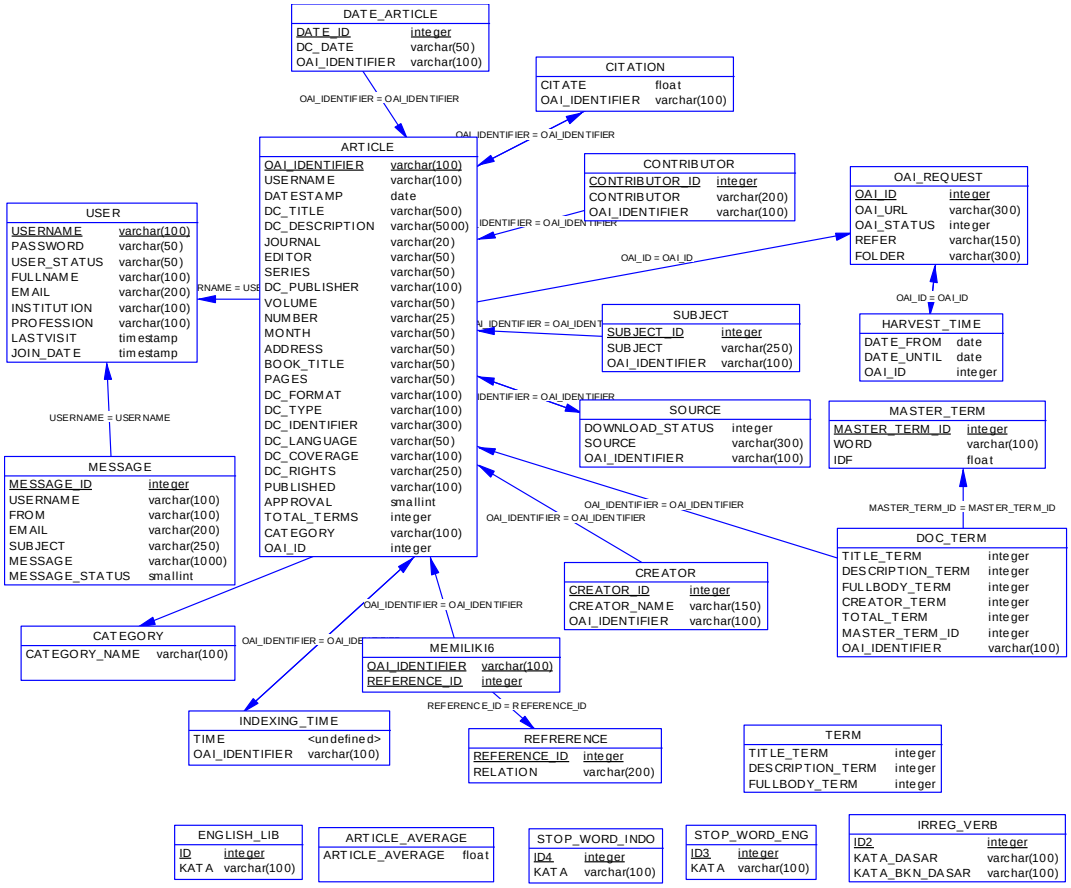
BM25 (Robertson, 1977) adalah model probabilistik yang mengasumsikan bahwa kemungkinan sebuah file digital relevan dengan *query* dari pengguna didapatkan dari jumlah kemungkinan kata-kata dalam file tersebut relevan terhadap query dari pengguna. Persamaan untuk mencari tingkat kemungkinan relevansi sebuah file digital ditunjukkan pada persamaan (1).

$$RSV_d = \sum_{t \in q} \log\left[\frac{N}{\mathrm{df}_t}\right] \cdot \frac{(k_1+1)\mathrm{tf}_{td}}{k_1((1-b)+b \times (L_d/L_{ave}))+\mathrm{tf}_{td}} \quad (1)$$

tf_{td} adalah jumlah term t pada file d. df_t adalah jumlah file yang mengandung term t. L_d adalah panjang dari file d dalam satuan term. L_{ave} adalah rata-rata dari panjang seluruh file yang tersimpan di database. *k1* adalah parameter untuk menentukan besar pengaruh tf_{td}. *b* adalah parameter untuk menentukan besar normalisasi dari panjang file.

4.4 Perhitungan Nilai Kutipan dari Sebuah File Digital

Dalam rangka proses *searching* tersebut selain menemukan nilai probabilitas relevansi file terhadap *query* pengguna menggunakan OKAPI BM25,



Gambar 2. Struktur database implementasi OAI-PMH

 ISSN: 1907-5022

penelitian ini juga mengusulkan adanya faktor tambahan, yaitu nilai kelayakan sebuah file. Karena setiap file dalam database adalah artikel jurnal atau konferensi ilmiah, maka nilai kelayakan didapatkan dari bobot kutipan dalam artikel tersebut, dan berapa banyak artikel tersebut telah dikutip artikel lainnya.

Asumsi yang digunakan dalam menghitung nilai sebuah file digital berdasarkan rantai kutipan adalah:

1. Semakin banyak sebuah file digital dikutip maka nilai file digital tersebut semakin besar.
2. Semakin banyak file-file digital bernilai besar yang dikutip oleh sebuah file digital, maka diasumsikan nilai file tersebut juga semakin besar.

Untuk menghasilkan nilai kutipan dari sebuah file digital digunakan persamaan (2).

$$Citation_k = R + \sum_{i=1}^{n} \frac{citation_i}{R_i} \qquad (2)$$

$Citation_k$ adalah nilai kutipan dari sebuah file digital k. R adalah jumlah file digital yang mengutip file k. n adalah jumlah file digital yang dikutip oleh file k. $citation_i$ adalah nilai kutipan dari file i. R_i adalah jumlah file digital yang dikutip ole file i.

4.5 Nilai Total File Digital

Untuk menghasilkan peringkat sebuah file dalam sebuah pencarian, maka nilai yang didapatkan dari OKAPI BM25 dijumlahkan dengan nilai kutipan dari file tersebut. Persamaan untuk menghasilkan nilai total ditunjukkan pada persamaan (3).

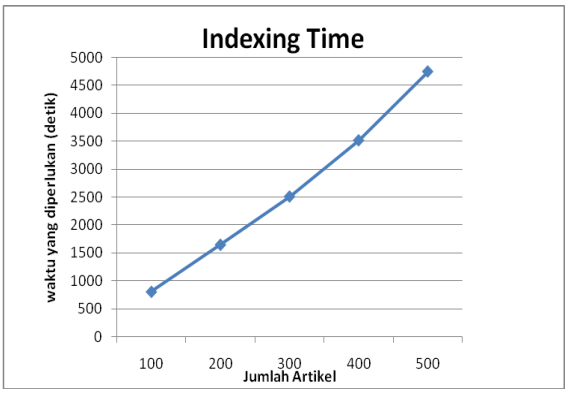
$$N(d) = a * RSV_d + b * Citation_d \qquad (3)$$

a dan b adalah parameter yang menentukan besar pengaruh tiap nilai terhadap nilai total, dimana a + b = 1.

5. PENGUJIAN

5.1 Indexing Time

Pengujian pertama berusaha untuk mendapatkan lama waktu *indexing* dari sekumpulan file.

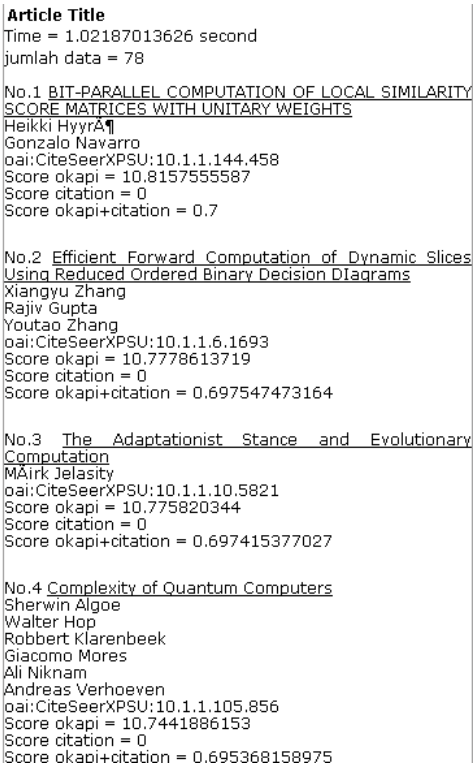


Gambar 3. Pengujian waktu *indexing*

Terlihat pada gambar 3 bahwa waktu *indexing* adalah linear terhadap jumlah file yang diindex.

5.2 Pengujian *Searching*

Hasil pengujian *searching* terhadap 500 file dengan menggunakan term "computer" ditunjukkan pada gambar 4. Pengujian *searching* ini hanya menggunakan OKAPI BM25.


Article Title
Time = 1.02187013626 second
jumlah data = 78

No.1 BIT-PARALLEL COMPUTATION OF LOCAL SIMILARITY SCORE MATRICES WITH UNITARY WEIGHTS
Heikki HyyrÄ¶
Gonzalo Navarro
oai:CiteSeerXPSU:10.1.1.144.458
Score okapi = 10.8157555587
Score citation = 0
Score okapi+citation = 0.7

No.2 Efficient Forward Computation of Dynamic Slices Using Reduced Ordered Binary Decision DIagrams
Xiangyu Zhang
Rajiv Gupta
Youtao Zhang
oai:CiteSeerXPSU:10.1.1.6.1693
Score okapi = 10.7778613719
Score citation = 0
Score okapi+citation = 0.697547473164

No.3 The Adaptationist Stance and Evolutionary Computation
MÃirk Jelasity
oai:CiteSeerXPSU:10.1.1.10.5821
Score okapi = 10.775820344
Score citation = 0
Score okapi+citation = 0.697415377027

No.4 Complexity of Quantum Computers
Sherwin Algoe
Walter Hop
Robbert Klarenbeek
Giacomo Mores
Ali Niknam
Andreas Verhoeven
oai:CiteSeerXPSU:10.1.1.105.856
Score okapi = 10.7441886153
Score citation = 0
Score okapi+citation = 0.695368158975

Gambar 4. Pengujian *searching* menggunakan term *"computer"*

5.3 Pengujian Nilai Kutipan

Pencarian berbasis OKAPI BM25 di atas dikombinasikan dengan nilai kutipan setiap file dengan referensi seperti ditunjukkan pada tabel 3. Agar lebih singkat maka setiap file hanya diberi nomor filenya saja. Oai1 dan Oai2 hanya memiliki metadata yang tersimpan dalam database, tetapi tidak memiliki file digitalnya sehingga tidak dapat ditentukan file-file yang direferensi oleh kedua file tersebut.

Tabel 3. File yang digunakan dalam pengujian dan referensinya.

File	Referensi
Oai1	-
Oai2	-
Oai3	Oai2
Oai4	Oai1
Oai5	Oai2
Oai6	Oai1 Oai3
Oai7	Oai1 Oai3

File	Referensi
	Oai5
Oai8	Oai2 Oai5
Oai9	Oai1 Oai5 Oai6 Oai8
Oai10	Oai2 Oai4 Oai6 Oai7

Dengan menggunakan *query* term "detail" dilakukan *searching* file baik menggunakan OKAPI BM25 saja, dan dibandingkan dengan bila ditambahkan nilai kutipannya. Hasil pengujian ditunjukkan pada tabel 4.

Tabel 4. Hasil pengujian *searching* menggunakan OKAPI BM25 dan nilai kutipan menggunakan term "detail"

Artikel	*Rank* artikel dengan Okapi+Citation	*Rank* artikel dengan Okapi
Oai2	1	4
Oai3	2	2
Oai7	3	1
Oai6	4	5
Oai8	5	3
Oai4	6	7
Oai10	7	6

Dari tabel 4 tersebut terlihat bahwa file Oai2 yang dikutip oleh banyak file lainnya pada urutan hasil OKAPI saja menempati posisi 4, sedangkan bila nilai kutipannya diperhitungkan menempati posisi 1. Sedangkan Oai7 yang dikutip oleh lebih sedikit file lainnya turun dari posisi 1 menjadi posisi 3 ketika nilai kutipan diperhitungkan.

6. KESIMPULAN

Penelitian ini mengusulkan implementasi repositori file digital menggunakan OAI-PMH sebagai protokol untuk mengumpulkan file-file digital dan metadatanya yang berguna bagi organsisasi tersebut. Untuk melakukan temu kembali terhadap file-file tersebut dapat digunakan OKAPI BM25 dan didukung oleh nilai kutipan. Nilai kutipan dapat membantu menyusun ulang peringkat file digital sesuai asumsi kualitas file (artikel) tersebut. Kualitas file (artikel) diasumsikan tinggi bila banyak dikutip oleh file lainnya, dan juga banyak mengutip file-file berkualitas lainnya.

PUSTAKA

DCMI - Dublin Core Metadata Initiative (2010). *Dublin Core Metadata Element Set, Version 1.1*. Diakses pada 20 Maret 2011 dari http://dublincore.org/documents/dces/

OAI - Open Archives Initiative (2002). *The Open Archives Initiative Protocol for Metadata Harvesting*. Diakses pada 20 Maret 2011 dari http://www.openarchives.org/OAI/openarchivesprotocol.html

Manning, C. D., Raghavan, P., dan Schütze, H. (2008). *Introduction to Information Retrieval*, Cambridge University Press.

Robertson, S. E. dan Jones K. S. (1977) "Relevance weighting of search terms," *Journal of the American Society for Information Science*.