IMPLEMENTASI SISTEM TEMU KEMBALI INFORMASI
Studi Kasus: Dokumen Teks Berbahasa Indonesia

(IMPLEMENTATION OF INFORMATION RETRIEVAL SYSTEM
Case Study: Text Document in Indonesian Language)

Bernadus Very Christioko
Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang

Abstract

Storage of digital documents is growing rapidly with increasing use of computers. These conditions raise issues to access the desired information quickly and accurately and also the difficulty in finding a document relating to a particular keyword with the precise and accurate results. The purpose of this paper is to compile a collection of documents as part of the testing device information retrieval systems for text documents in Indonesian language and compiled a collection of indexes from the collection of documents and build a search engine to help to search documents using information retrieval methods / Information Retrieval System (IRS). The method used was to build a system with two main parts, namely subsystem Indexing and Searching Subsystem (matching system). The results show that the method the IRS can assist in conducting a search for documents in a collection by the number of one or more keywords or a combined using Boolean functions.

Keyword *: Information Retrieval System, Search Engine*

1. PENDAHULUAN

Penyimpanan dokumen secara digital berkembang dengan pesat seiring meningkatnya penggunaan komputer. Kondisi tersebut memunculkan masalah untuk mengakses informasi yang diinginkan secara akurat dan cepat. Oleh karena itu, walaupun sebagian besar dokumen digital tersimpan dalam bentuk teks dan berbagai algoritma yang efisien untuk pencarian teks telah dikembangkan, kesulitan menemukan suatu dokumen yang berhubungan dengan suatu kata kunci tertentu dengan hasil yang tepat dan akurat masih terjadi. Pencarian terhadap seluruh isi dokumen yang tersimpan bukanlah solusi yang tepat mengingat pertumbuhan ukuran data yang tersimpan umumnya.Temu kembali informasi bertujuan untuk membantu pengguna dalam menemukan informasi yang relevan dengan kebutuhan mereka dalam waktu singkat. Akan tetapi banyak teknik-teknik tersebut yang tergantung pada bahasa yang digunakan dalam dokumen. Untuk mengembangkan teknik-teknik temu kembali informasi bagi dokumen teks berbahasa Indonesia, dibutuhkan perangkat pengujian untuk Bahasa Indonesia. Salah satunya adalah suatu koleksi dokumen dalam Bahasa Indonesia sebagai pendekatan seragam dalam evaluasi sistem temu kembali informasi.

Sistem temu kembali informasi atau *Information Retrieval System* (IRS) mempunyai tujuan memberitahukan keberadaan (atau ketidakberadaan) dan keterangan dokumen-dokumen yang berhubungan dengan permintaan dari pengguna bukan memberitahu mengenai masalah yang ditanyakan. Alasan utamanya adalah karena IRS menangani teks bahasa alami yang tidak selalu terstruktur dengan baik dan bersifat ambigu. IRS bekerja berdasarkan kueri yang diberikan pengguna yang kemudian menghasilkan daftar dokumen yang dianggap relevan dengan 2 bagian utama, yaitu *Indexing subsystem*, dan *Searching subsystem* (*matching system*).

Tujuan dari tulisan ini adalah untuk menyusun koleksi dokumen sebagai bagian

perangkat pengujian sistem temu kembali informasi untuk dokumen teks berbahasa Indonesia dan menyusun kumpulan indeks dari koleksi dokumen dan membangun mesin pencari untuk membantu melakukan pencarian dokumen menggunakan metode Sistem Temu Kembali Informasi.

2. METODE PENELITIAN

Dalam menerapkan metode sistem temu kembali informasi untuk koleksi dokumen berbahasa indonesia dengan membangun 2 bagian utama, yaitu *Indexing subsystem*, dan *Searching subsystem* (*matching system*)dilakukan beberapa tahapan penelitian berikut ini.

2.1. Kata buangan / stop list

Pada tahapan ini, sebelum melakukan upload dokumen dan parsing dokumen, terlebih dahulu menentukan daftar kata-kata buangan yang akan digunakan sebagai penyaring indeks. Setelah kata-kata buangan didefinisikan, parsing dokumen dapat dilakukan dengan melakukan upload dokumen menggunakan upload engine.

2.2. Koleksi Dokumen

Untuk melakukan implementasi system temu kembali diperlukan sebuah koleksi dokumen yang terdiri dari kumpulan-kumpulan jurnal atau paper yang mempunyai topic yang sama. Oleh karena itu, dalam tahap ini, pengumpulan dokumen dilakukan sebanyak 10 (sepuluh) dokumen dengan topic E-Learning, E-Government, Teknologi Informasi dengan format dokumen berekstensi .doc (Microsoft Word).

File-file dokumen yang telah dikumpulkan akan diindeks berdasarkan judul dan abstrak dengan cara mengupload dokumen ke server database menggunakan upload engine menggunakan bahasa pemrograman PHP dan database MySQL.

2.3. Parsing Dokumen

Dalam proses upload dokumen ke server menggunakan upload engine, judul dan abstrak akan diparsing dengan membuang kata-kata buangan (stop list) yang tidak perlu untuk dijadikan sebagai indeks, sehingga setiap kata dari judul dan abstrak yang menjadi indeks akan tersimpan dalam database di MySQL.

2.4. Data / Indeks

Hasil dari tahap parsing dokumen adalah daftar indeks dari setiap dokumen yang terdapat dalam koleksi dokumen. Daftar indeks sudah terbebas dari kata-kata buangan (stop list), daftar indeks inilah yang akan menjadi acuan dalam melakukan pencarian informasi dengan menggunakan kata kunci pencarian yang akan dicocokkan dengan daftar indeks yang ada.

2.5. Matriks Dokumen Indeks

Matriks dokumen indeks diperoleh dari hasil indeks dan dokumen dengan mengukur frekuensi kemunculan suatu indeks dalam setiap dokumen.

2.6. Query Dokumen

Query terhadap dokumen dilakukan dengan menggunakan kata kunci pencarian untuk melakukan pengujian dari koleksi indeks yang telah terbentuk pada proses parsing dokumen. Untuk melakukan query dibangun sebuah mesin pencari dengan menerima input berupa kata kunci tunggal maupun ganda (gabungan dengan fungsi Boolean AND dan OR).

3. HASIL DAN PEMBAHASAN

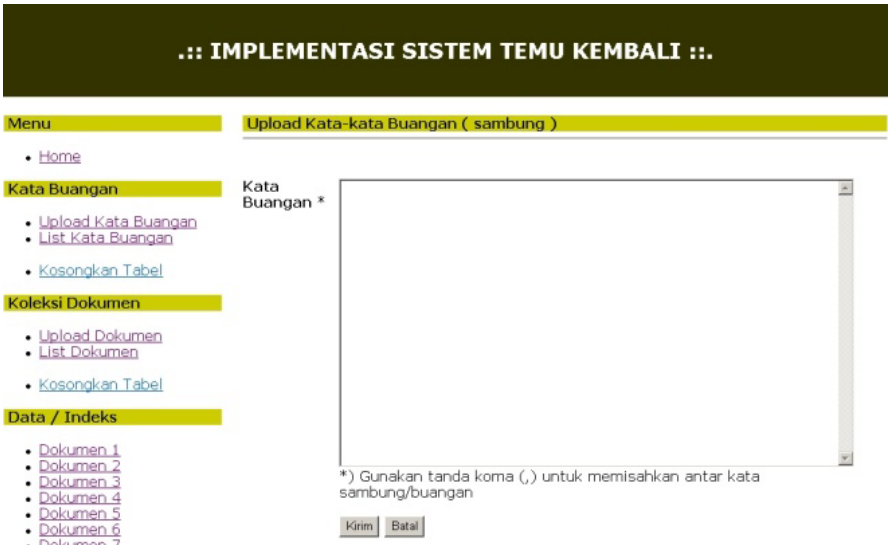
3.1. Kata buangan / stop list

Kata-kata buangan yang telah ditentukan diupload ke database server MySQL untuk disimpan ke dalam table ‘Common’ menggunakan modul ‘Upload Kata Buangan’ . Table ‘common’ mempunyai struktur sebagai berikut :

Tabel 1.Tabel Common

Field	Type
word	varchar(255)
no	int(1)

Tampilan dari modul ‘Upload Kata Buangan’ seperti gambar di bawah ini :



Gambar 3. Upload Kata Buangan

Setelah kata-kata buangan terupload, dapat dilihat melalui modul ‘List Kata Buangan’, seperti gambar berikut :

Gambar 4. List Kata Buangan

3.2. Koleksi Dokumen

Sepuluh dokumen yang telah disiapkan kemudian diupload ke server menggunakan modul ‘Upload Dokumen’, berupa form isian judul dan abstrak dari setiap dokumen yang ada. Dokumen secara otomatis akan tersaring dari kata-kata buangan yang sudah ditentukan sebelumnya dan disimpan ke dalam table ‘content’, ‘keytable’, dan ‘link’. Struktur dari ketiga table sebagai berikut:

Tabel 2. Tabel Content

Field	Type
contid	mediumint(9)
title	text
abstract	longtext

Tabel 3. Tabel Keytable

Field	Type
keyid	mediumint(9)
keyword	varchar(100)

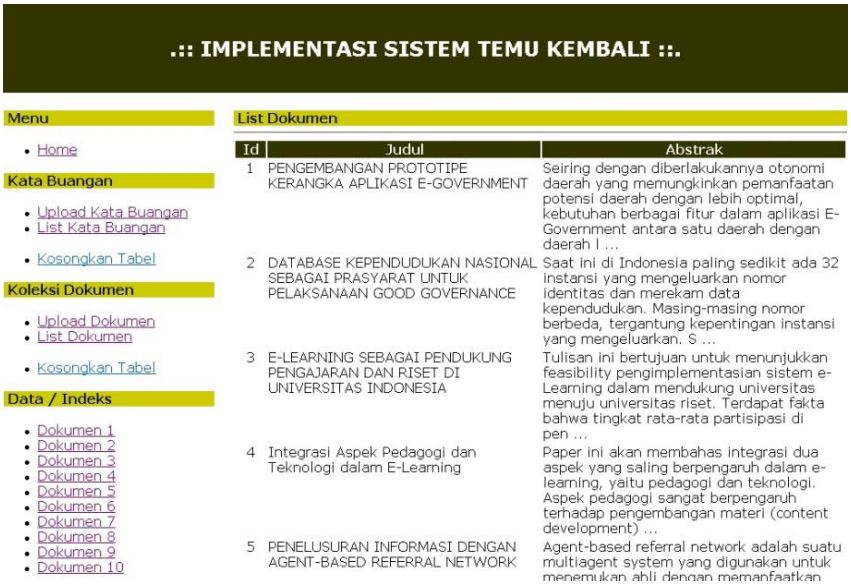
Tabel 4. Table Link

Field	Type
keyid	mediumint(9)
contid	mediumint(9)

Tampilan dari modul 'Upload Dokumen' seperti gambar di bawah ini :

Gambar 5. Upload Dokumen

Setelah dokumen terupload, dapat dilihat melalui modul 'List dokumen', seperti gambar berikut :



Gambar 6. List Dokumen

3.3. Hasil Indexing

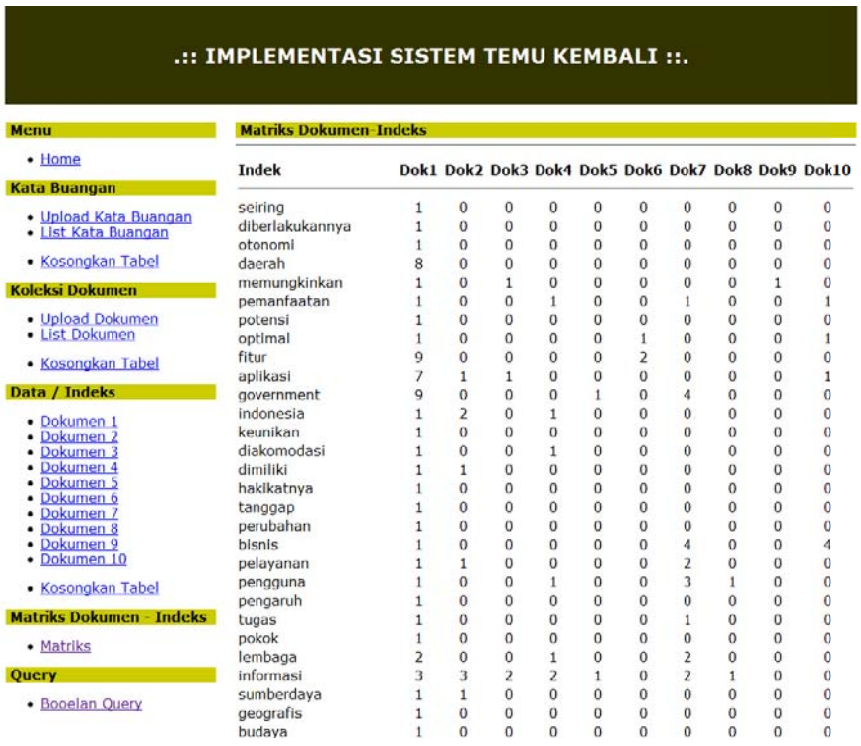
Setelah seluruh koleksi dokumen terupload, indeks dokumen sudah terbentuk. Indeks dapat dilihat pada modul 'indeks-dokumen', seperti gambar di bawah ini :

Gambar 7. indeks Dokumen 1

Gambar 8. indeks Dokumen 2

3.4. Matriks Dokumen Indeks

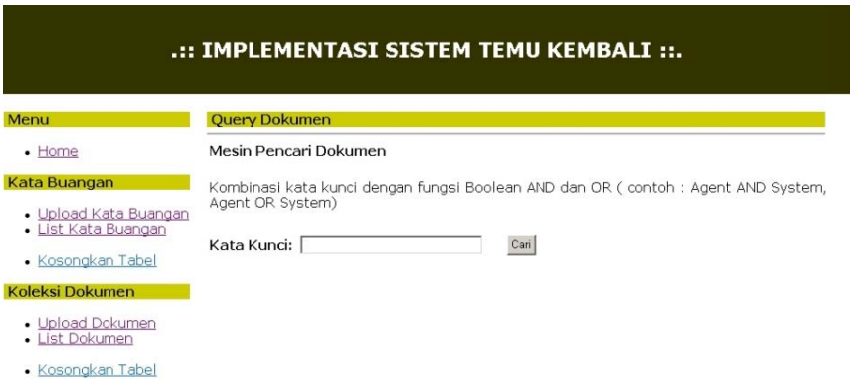
Matriks dokumen antara dokumen dan indeks dapat dilihat pada modul 'matriks'. Frekuensi kemunculan suatu indeks dalam koleksi dokumen ditunjukkan dengan banyaknya jumlah indeks dalam suatu dokumen tertentu, seperti gambar di bawah ini :



Gambar 9. Matriks Dokumen Indeks

3.5. Hasil Query

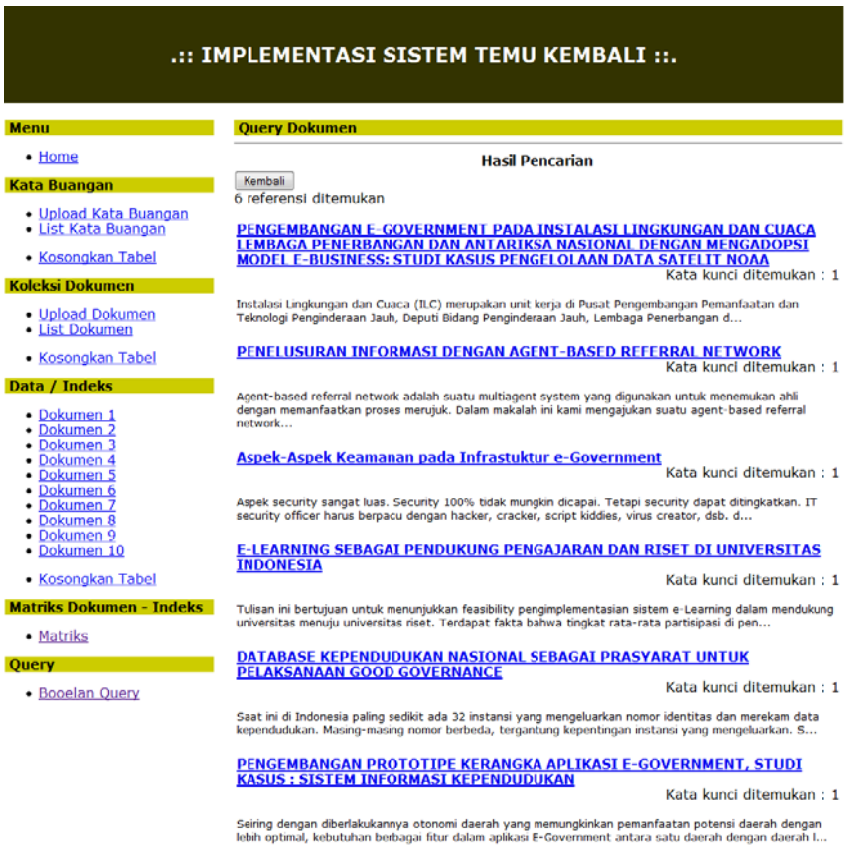
Untuk melakukan uji query dibangun sebuah mesin pencari yaitu modul ‘query boolean’, seperti gambar di bawah ini :



Gambar 10. Mesin Pencari

Pengujian 1:

Pengujian pertama, dilakukan pencarian dokumen dengan kata kunci ‘sistem’. Dari pencarian dengan kata kunci ‘sistem’ dihasilkan 6 dokumen yang relevan dengan kata kunci ‘sistem’. Gambar seperti di bawah ini :



Gambar 11. Hasil Pencarian Pengujian 1

Pengujian 2:
Pengujian kedua, dilakukan pencarian dokumen dengan kata kunci ‘informasi’.Dari pencarian dengan kata kunci ‘informasi’ dihasilkan 7 dokumen yang relevan dengan kata kunci ‘informasi’. Gambar seperti di bawah ini:

Gambar 12. Hasil Pencarian Pengujian 2

Pengujian 3:
Pengujian ketiga, dilakukan pencarian dokumen dengan menggunakan fungsi Boolean ‘AND’ → ‘sistem AND informasi’.Dari pencarian dengan kata kunci ‘sistem AND informasi’ dihasilkan 6 dokumen yang relevan dengan kata kunci, dimana kedua kata kunci baik system dan informasi terkandung di setiap dokumen yang ditemukan. Gambar seperti di bawah ini :

Gambar 13. Hasil Pencarian Pengujian 3

Pengujian 4:

Pengujian keempat, dilakukan pencarian dokumen dengan menggunakan fungsi Boolean 'OR' → 'sistem OR informasi'.Dari pencarian dengan kata kunci 'sistem OR informasi' dihasilkan 7 dokumen yang relevan dengan kata kunci, dimana kedua kata kunci baik system atau informasi terkandung di setiap dokumen yang ditemukan. Gambar seperti di bawah ini :

Gambar 14. Hasil Pencarian Pengujian 4

4. KESIMPULAN

Dari hasil pengujian dapat dilihat bahwa dengan menerapkan konsep-konsep dalam sistem temu kembali informasi dan membangun suatu mesin pencari dapat membantu dalam melakukan pencarian dokumen dalam suatu koleksi dokumen dimana pencarian dapat dilakukan dengan memberikan kata kunci dengan jumlah kunci satu atau lebih atau dapat juga menggabungkan lebih dari satu kata kunci menggunakan fungsi *Boolean*yaitu fungsi AND dan OR.

5. DAFTAR PUSTAKA

Keraf, Goris. 1984, *Tata Bahasa Indonesia*, Nusa Indah.

Wibisono, Yudi. 2008. *Stop words untuk Bahasa Indonesia*. (Online). (http://yudiwbs.wordpress.com/2008/07/23/stop-words-untuk-bahasa-indonesia/, diakses tanggal 9 Agustus 2011)

Dharan, M.Murali. 2003. *Dynamic Document Search Engine.* (Online).*(*http://www.phpbuilder.com, diakses tanggal 9 Agustus 2011)

Baeza-Yates, R. & Ribeiro-Neto, B. 1999. *Modern Information Retrieval*. Addison-Wesley.

Lancaster, F. & Warner, A. 1993. *Information Retrieval Today*. Information Resources Press, Arlington.

Liddy, E. 2001. *How a Search Engine Works*. Searcher 9(5). Information Today, Inc.

Adisantoso, Julio.2004. *Corpus Dokumen Teks Bahasa Indonesia untuk Pengujian Efektivitas Temu Kembali Informasi*. Institut Pertanian Bogor.