PENGINDEKAN DAN PENCARIAN DOKUMEN TEXT

Kusrini, S.Kom
STMIK AMIKOM Yogyakarta

Abstract

We often needs to search a specific or joined word(s) within a document. An application with ability to store and search word document is made in this research. Borland Delphi programming tools is used to develop the application and Interbase as database system.

For efficiency reason, a stop word will be eliminated before data is stored. Also, synonim words need to be keep attended. In document searching it is possible to use operator boolean AND, OR and NOT with word count priority that contained in the document.

Keywords : Index, Searching, Document, Text, Key

1. Pendahuluan

Latar Belakang Masalah

Dalam pengarsipan data text sering diperlukan pancarian data yang mengandung kata-kata tertentu. Kepandaian pengguna dalam menentukan kata kunci sangat menentukan keakuratan hasil pencarian. Selain kata kunci, ada hal lain yang harus diperhatikan yaitu penggunaan operator jika diinginkan pengunaan lebih dari satu kata kunci.

Operator yang sering digunakan dalam pencarian data text diantaranya adalah: AND, OR, NOT. AND jika diinginkan semua kata kunci ada dalam dokumen, OR jika diinginkan salah satu kata kunci ada dalam dokumen dan NOT jika diinginkan suatu kata kunci ada dalam dokumen tetapi kata kunci lainnya tidak terdapat pada dokumen tersebut.

Pada kasus tertentu, untuk mendapatkan relevansi data hasil pencarian dengan data yang dimaksud diperlukan operator selain AND, OR dan NOT. Operator itu

harus mampu mendeteksi keberadaan kata-kata kunci tersebut harus dalam satu padagraph, dalam satu kalimat atau bahkan kata-kata kunci tersebut harus saling berurutan.

Untuk memenuhi kebutuhan pencarian dokumen text tersebut, diperlukan suatu cara untuk menyimpan/mengorganisir data dan cara mengambil kembali data tersebut.

Batasan Masalah

Banyak parameter yang harus diperhatikan dalam penyimpanan data text diantaraya:

1. Menghindari penyimpanan *stop word* dalam data, untuk efisiensi
2. Penyimpanan kata sinonim untuk keperluan pengambilan data yang mungkin relevan meskipun tidak menggunakan kata yang sama

Dalam makalah ini, kedua parameter diatas sudah diakomodir, tetapi hasil penelitian ini belum mampu mengatasi imbuhan yang berupa sisipan (*infix*). Dalam penelitian ini, kata kunci yang digunakan hanya 2 dengan penggunaan operator : AND, OR, NOT, SATU PARAGRAPH, SATU KALIMAT, dan BERURUTAN.

Dalam penelitian ini juga belum dibahas perhitungan bobot berdasarkan relevansi hasil pencarian.

Tujuan dan Manfaat

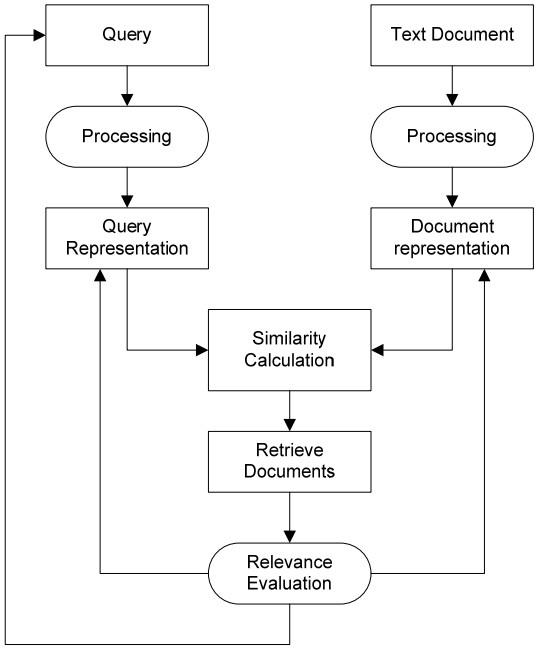
Tujuan dari penelitian ini adalah untuk membuat prototype perangkat lunak yang mempu melakukan penyimpanan dan pengambilan kembali dokumen text.

Sedangkan manfaat yang bisa diperoleh dari penelitian ini yaitu hasil penelitian ini bisa digunakan untuk melakukan penyimpanan dokumen bertipe text dan dapat digunakan untuk pencarian dokumen tersebut berdasarkan kata kunci tertentu.

2. Dasar Teori

Proses Pengambilan Informasi

Proses dasar pengambilan informasi ditunjukkan oleh gambar 1.



Gambar1. Proses pengambilan informasi

Struktur File

Struktur file yang digunakan dalam system ini adalah inverted file yang diimplementasikan dalam database. Data disimpan dalam susunan : Kata, NoParagraph, NoKalimat, NoKata.

Operasi Kata dan Pengindekan Otomatis

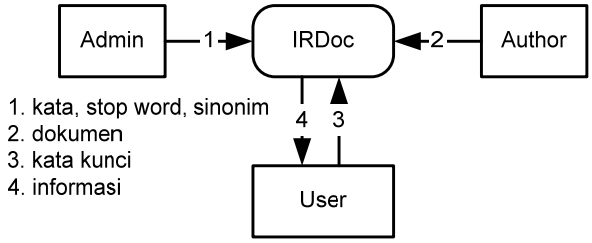
Berikut ini proses-proses yang harus dilakukan dalam operasi kata kunci

1. Identifikasi kata dalam dokumen
2. Buang kata tidak penting dari daftar kata setelah dicocokkan dengan kamus khusus atau daftar kata tidak penting
3. Identifikasi sinonim dengan mencocokkan setiap kata pada daftar kata sinonim
4. Ambil akar kata dengan menggunakan suatu algorithm yang menghilangkan kata depan dan belakang
5. Hitung frequency kata dalam tiap dokumen
6. Hitung bobot kata
7. Buat basis data berbasis invert untuk term-term tersebut beserta bobotnya

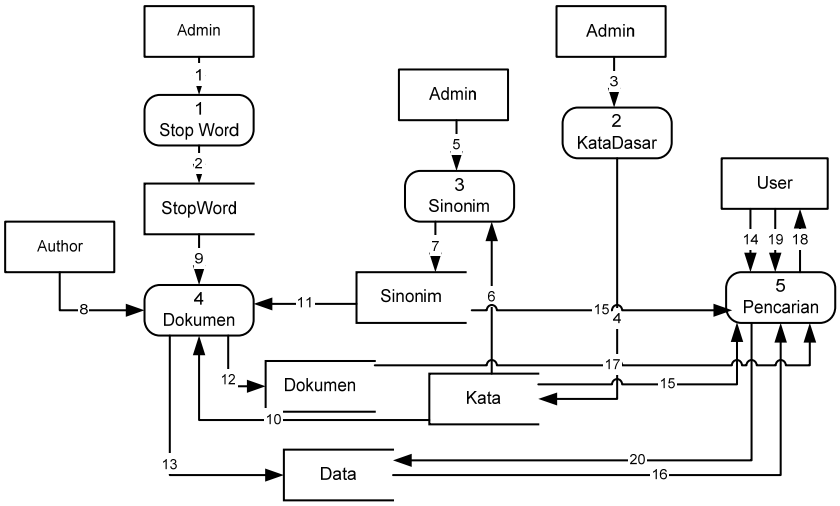
3. Perancangan dan Implementasi Sistem

3.1 Data Flow Diagram

Pada gambar 2 ditunjukkan context diagram dari system ini, sedangkan DFD level 1 ditunjukkan pada gambar 3.



Gambar 2. Context Diagram



Gambar 3. Data Flow Diagram Level 1

3.2 Pembobotan

Dalam penelitian ini bobot suatu dokumen dihitung berdasarkan pada jumlah kata kunci yang ditemukan dalam dokumen.

3.3 Perancangan Database

Dalam implementasi penelitian ini, digunakan bahasa pemrograman Delphi, dan Software database Interbase. Berikut ini adalah rancangan database yang digunakan dalam implementasi system ini:

Kata
Kata

Stop
Kata

Sinonim
Kata

Dokumen
Kddok
Judul
NamaFile
Pengarang

Data
Kata
KdDok
NoPrg
NoKal
NoKata

Hasil
KdDok
Judul
Pengarang
NamaFile
Jml

4. Hasil dan Implementasi

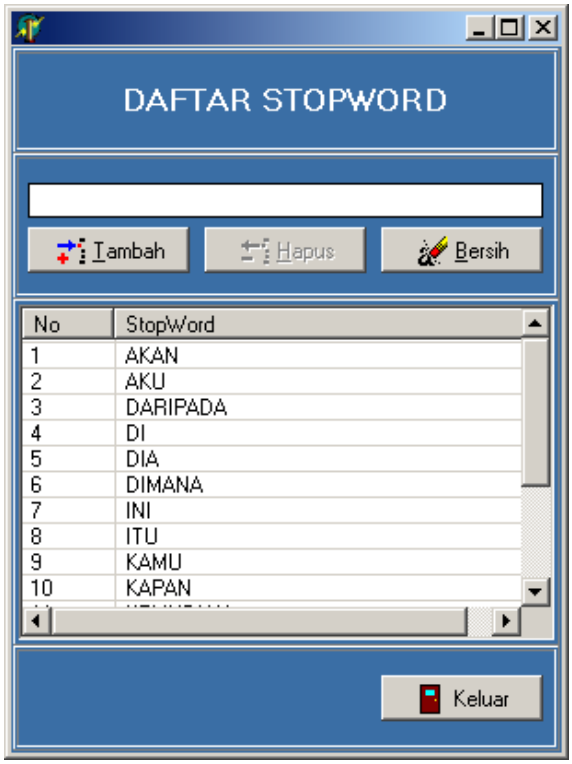
1. Form utama

Gambar 4. Form Utama

2. Form kata dasar
   Kata dasar digunakan untuk mengelola data kata dasar

Gambar 5. Form Kata Dasar

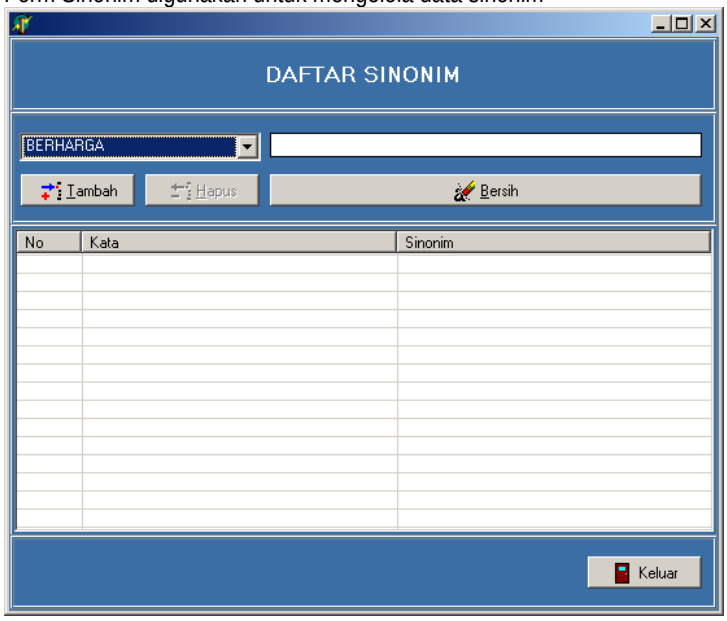
3. Form StopWord
   Form ini digunakan untuk mengelola data StopWord



Gambar 6. Form StopWord

4. Form Sinonim

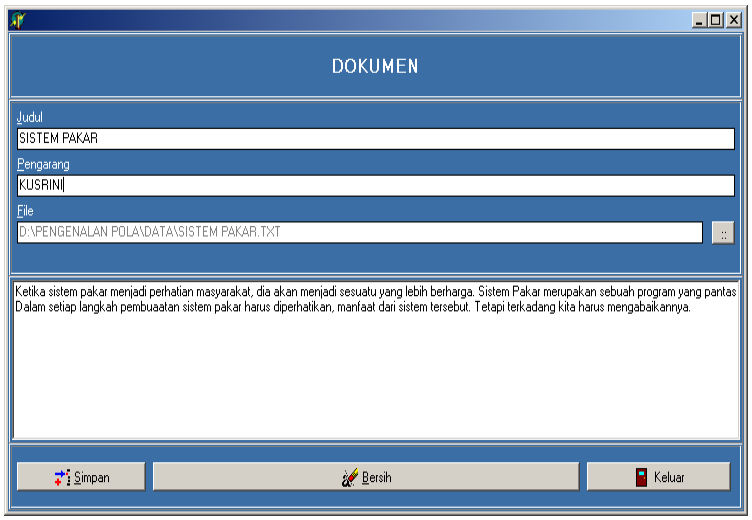
Form Sinonim digunakan untuk mengelola data sinonim



Gambar 6. Form Sinonim

5. Form Dokumen

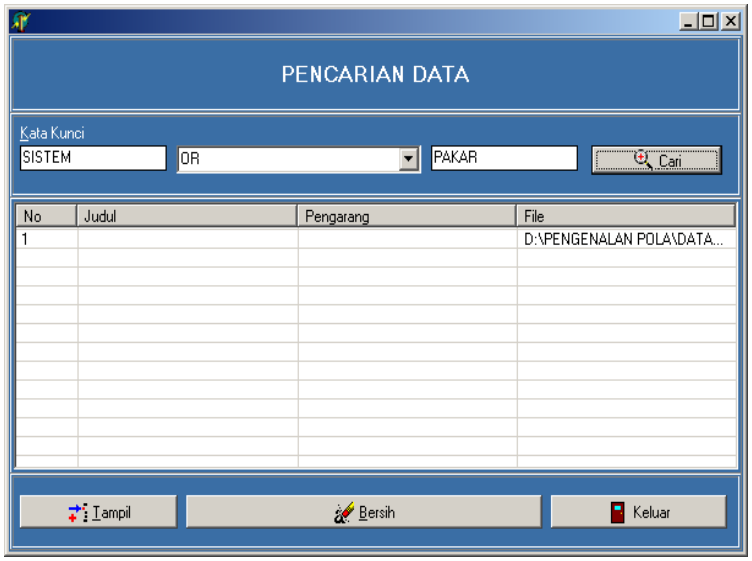
Form dokumen digunakan untuk mengelola data dokumen text yang akan disimpan. Pada saat penyimpanan akan dilakukan penyimpanan otomatis kata yang belum ada dalam daftar kata atau sinonim.



Gambar 7. Form Dokumen

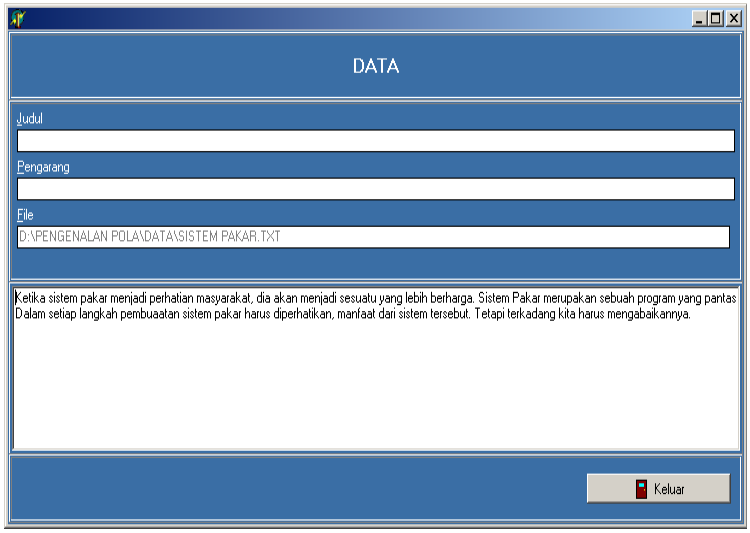
6. Form Pencarian

Form pencarian digunakan untuk mencari dokumen berdasarkan kata kunci yang dimasukkan. Dalam pencarian digunakan operator AND, OR, NOT, SATU PARAGRAPH, SATU KALIMAT dan BERURUTAN. Operator AND digunakan untuk mencari dokumen yang mengandung 2 kata kunci sekaligus. Operator OR digunakan untuk mencari dokumen yang mengandung salah satu kata kunci. Operator NOT digunakan untuk mencari dokumen yang mengandung kata kunci pertama tetapi tidak terdapat kata kunci kedua. Operator SATU PARAGRAPH digunakan untuk mencari dokumen yang terdapat 2 kata kunci dalam satu paragraph. Operator SATU KALIMAT digunakan untuk mencari dokumen yang terdapat 2 kata kunci dalam satu kalimat. Operator BERURUTAN digunakan untuk mencari dokumen yang terdapat 2 kata kunci secara berurutan.



Gambar 8. Form Pencarian Data

Untuk menampilkan data hasil pencarian yang dianggap relevan oleh pengguna, dilakukan dengan penekanan tombol tampil, dan akan menampilkan data seperti tampak pada gambar 9.



Gambar 9. Form untuk Menampilkan Dokumen Hasil Pencarian

5. Kesimpulan

Penelitan ini telah berhasil mengimplementasikan penyimpanan dan pengambilan data dokumen text. Penyimpanan dan pengambilan data dengan memperhatikan adanya stop word dan sinonim.

Dalam penelitian ini, kata kunci yang digunakan hanya 2 degan operator AND, OR, NOT, SATU PARAGRAPH, SATU KALIMAT, dan BERURUTAN.

6. Daftar Pustaka

Wright, L.w., Nardini, H.K.G, Aronson, A.R., RindFlesch, T.C, Hierarchical Concept Indexing of Full-text Documents in the UMLS® Information Sources Map, skr.nlm.nih.gov/papers/references/jasis98.pdf, Tanggal akses 5 Juli 2005

Yang, Y, 1997, An Evaluation of Statistical Aproaches to text Categorization, http://citeseer.ist.psu.edu/yang97evaluation.html, Tanggal akses 1 Agustus 2005