IDENTIFIKASI NASKAH DOKUMEN TEKS DENGAN METODE *INDEXING*

Heriyanto
Prodi Teknik Informatika UPN "Veteran" Yogyakarta
Jl. Babarsari 2 Tambakbayan 55281 Telp (0274) 485323
email : mr_heriyanto_skom@yahoo.com

Abstract

In everyday life of text data or copy met by many in old stuff and white colars in our life is everyday. Multimedia which consist of text data aggregate, picture and also voice or present image used by many either through computer, internet, even handphone media. Method make an index to text a copy which is through scanning process and parsing can be used to make index a document copy like within reason thick book usually is in it figured in to index for the seeking of a word in book swiftly and index for line, paragraph and word frequency.

Keywords : indexing, text, line, parsing, scanning, document

Dalam kehidupan sehari-hari data teks atau naskah banyak dijumpai di perkantoran dan tidak asing lagi dalam kehidupan kita sehari-hari. Multimedia yang terdiri dari kumpulan data teks, suara maupun gambar atau citra yang sekarang ini banyak digunakan baik melalui komputer, internet, bahkan media handphone. Metode indeks teks suatu naskah yang dilakukan melalui proses *scanning* dan parsing dapat digunakan untuk membuat indeks suatu naskah dokumen seperti layaknya buku-buku yang tebal biasanya di dalamnya disertakan indexing untuk pencarian suatu kata dalam buku dengan cepat dan dilakukan *indexing* untuk baris, paragraph dan frekwensi kata.

Kata kunci : indexing, teks, baris, parsing, scanning, dokumen

1. PENDAHULUAN

Melalui pengolah data naskah atau teks maka bagaimana suatu naskah yang ada di dalam buku atau majalah atau sarana di dalam naskah komputer dapat dibuatkan suatu *indexing* yang dapat memudahkan pencarian data.

a. Bagaimana melakukan dengan proses scanning yang menguraikan satu huruf demi huruf melalui proses pembacaan satu persatu huruf demi huruf.
b. Bagaimana mendeteksi suatu alphabet yang tidak dipakai seperti spasi, koma titik dan lain-lain yang dianggap dibuang.
c. Bagaimana mendeteksi keberadaan suatu paragraph, baris pada suatu naskah
d. Medneteksi berapa banyak kata yang muncul pada suatu paragraph, pada suatu kalimat dan seterusnya.
e. Metode parsing dan scanning untuk menguraikan data huruf dan kata.
f. Menghitung jumlah data kata atau paragraph atau baris pada suatu paragraph.

Pada pembahasan saat ini penulis hanya membatasi pada pengolah data teks melalui proses indek dan pencarian data suatu kata terdapat pada indek dan ditemukan identifikasi kata tersebut ada di berbagai paragraph maupun baris dan frekwensi kemunculannya.

2. TINJAUAN PUSTAKA

Spark Jones 1997 (dalam Rhodes, 2000; hal 45) menyatakan bahwa pencarian dokumen dikelompokkan pada dua aktivitas yang saling berkaitan satu dengan lainnya yaitu : indeks dan pencarian. Indeks mengacu pada dokumen itu sendiri, yaitu informasi yang akan dipanggil, dan pencarian yaitu pernyataan dari pemakai yang membutuhkan informasi dengan tujuan untuk menampilkan informasi yang diinginkan.

Lu (1999;75-76) melihat bahwa tujuan dari suatu sistem IR adalah untuk mendapat kembali materi relevan dari suatu database dokumen sebagai jawaban atas *query* pemakai.

Kebanyakan dari sistem IR yang komersil saat sekarang dapat digolongkan pada sistem IR boolean atau sistem pencarian *text-pattern. Query* pencarian *text-pattern* adalah *string* atau ungkapan *reguler*.

Di dalam suatu *file inverted* untuk masing-masing istilah merupakan suatu indeks terpisah yang dibangun itu menyimpan *record* pengenal untuk semua arsip yang berisi *term*. Suatu masukan *file inverted* pada umumnya berisi suatu kata kunci dan sejumlah dokumen Identitas. Masing-masing kata kunci atau istilah dan *document-Ids* dari dokumen yang berisi kata kunci diorganisasi ke dalam satu baris. Suatu contoh dari suatu *file inverted* ditunjukkan dibawah ini :

Term 1: Record1, Record3
Term 2 : Record 1, Record 2
Term 3 : Record 2, Record 3, Record 4
Term 4 : Record 1, Record 2, Record 3, Record 4

Dimana *term* I (I menjadi 1,2,3 atau 4) adalah nomor ID jumlah indek masukan I, *record* I (menjadi 1,2,3 atau 4) adalah nomor ID jumlah *record* I atau dokumen I (Lu,1999;hal 77).

Teknik yang dipakai untuk dokumen teks banyak digunakan dengan teknik IR (*Indexing* dan *Retrieval*). Teknik IR sangatlah penting di dalam informasi multimedia manajemen sistem untuk dua alasan yaitu :

1. Keberadaannya, sebagian besar dokumen teks banyak digunakan di dalam organisasi seperti perpustakaan.
   Teks adalah sangat penting sumber informasi untuk organisasi. Untuk pengefisienan penggunaan penyimpanan informasi dalam naskah sangat diperlukan sistem IR.
2. Teks dapat digunakan untuk kebutuhan media lain seperti Audio, Gambar dan Video.

(menurut Buku Multimedia Database Management System, Guojun Lu Hal 73)
Ada 4 penggunaan komponen model IR (*Indexing Retrieval*) yaitu :

1. *Exact Match*
2. *Vector Space*
3. *Probabilistic*
4. *Cluster-based*

(menurut Buku Multimedia Database Management System, Guojun Lu Hal 73)
Namun sebagian besar menggunakan teknik *exact Match* dalam model *boolean*.

Text Documents di proses seperti dilakukan indeks yang tersimpan dalam database yang merupakan *document representation* lalu *query /permintaan* yang diinputkan oleh *user* atau dilakukan *entry* data *query* maka diproses dan merupakan *query representation* kemudian dicocokan apa yang ada pada *query* dengan apa yang ada pada *databases* diserver dilakukan *comparison* (*similary calculation*) maka merupakan hasil *retrieved documents* apakah data hasil *retrieved documents* tersebut sudah *relevance* seperti yang diinginkan atau yang dicari jika tidak maka *feed back* dilakukan permintaan /*query* lagi oleh *user* kembali ke awal *query* atau *query representation* yang sudah ada tadi dicari lagi. Contoh misalkan dalam *search engine* pada saat *user* melakukan *query* pada query representation dilakukan pencarian kata haji maka dilakukan pencocokan dengan *text document* yang ada pada proses di indeks database berupa *document representation* lalu dilakukan pencocokan *similary calculation* dan ditampilkan hasilnya beberapa nama yang mendekati kata haji *diretrieved documents* apabila tidak sesuai dengan yang dicari atau tidak *relevance* maka kembali lagi ke *query* atau ke *query representation* yang sudah ada tersebut.

Adapun tujuan yang ingin dicapai penulis antara lain :

1. dapat membuat indek pada suatu naskah atau dokumen teks surat
2. dapat melakukan pencarian kata terdapat pada naskah dengan menamplilkan paragraph, baris, dan frekwensi kemunculan data.

Manfaat
Adapun beberapa manfaat yang didapat yaitu :
mengetahui kata tertentu dengan tepat dan cepat berdasarkan indeks yang sudah dic *create* menemukan posisi data dimanapun berada baik paragraph, baris, dan frekwensi kemunculannya.

3. METODE PENELITIAN

Metode penelitan dengan metode indexing. Dokumen teks atau naskah dapat melakukan indek dengan menganalisa sebagai berikut :

a. Data teks terdiri dari beberapa paragraph

b. Data paragraph terdiri atas beberapa baris
c. Data kata terdiri atas kemunculan / jumlah kemunculan atau frekwensi.

Tahap Pertama

> Makanan bakso munir bangga yang mudah ditemukan dimana-mana, di Indonesia merupakan budaya munir bangga mengandung adi luhur bangsa. Dimana budaya Indonesia sangat rukli nilai-nilai luhur budaya rizki bakso rizki yang bangga bakso pemuda rukli bangsa Indonesia yang luhur budi pekerti dan akhlaknya rukli heri, rizki.

Proses pembacaan data dengan scanning teks atau naskah teks misalkana apabila dilakukan indek maka :

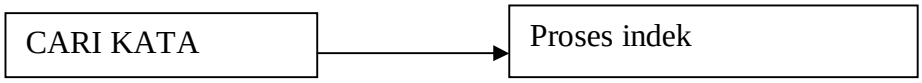
Kata	prgh	baris	frek-
Makanan	1	1	1
bakso	1	1	1
munir	1	1	1
bangga	1	1	1
yang	1	1	1
mudah	1	1	1
ditemukan	1	1	1
dimana-mana	1	1	1
di	1	1	1
Indonesia	1	1	1
merupakan	1	2	1
budaya	1	2	1
munir	1	2	2
bangga	1	2	2
mengandung	1	2	1

dan seterusnya……

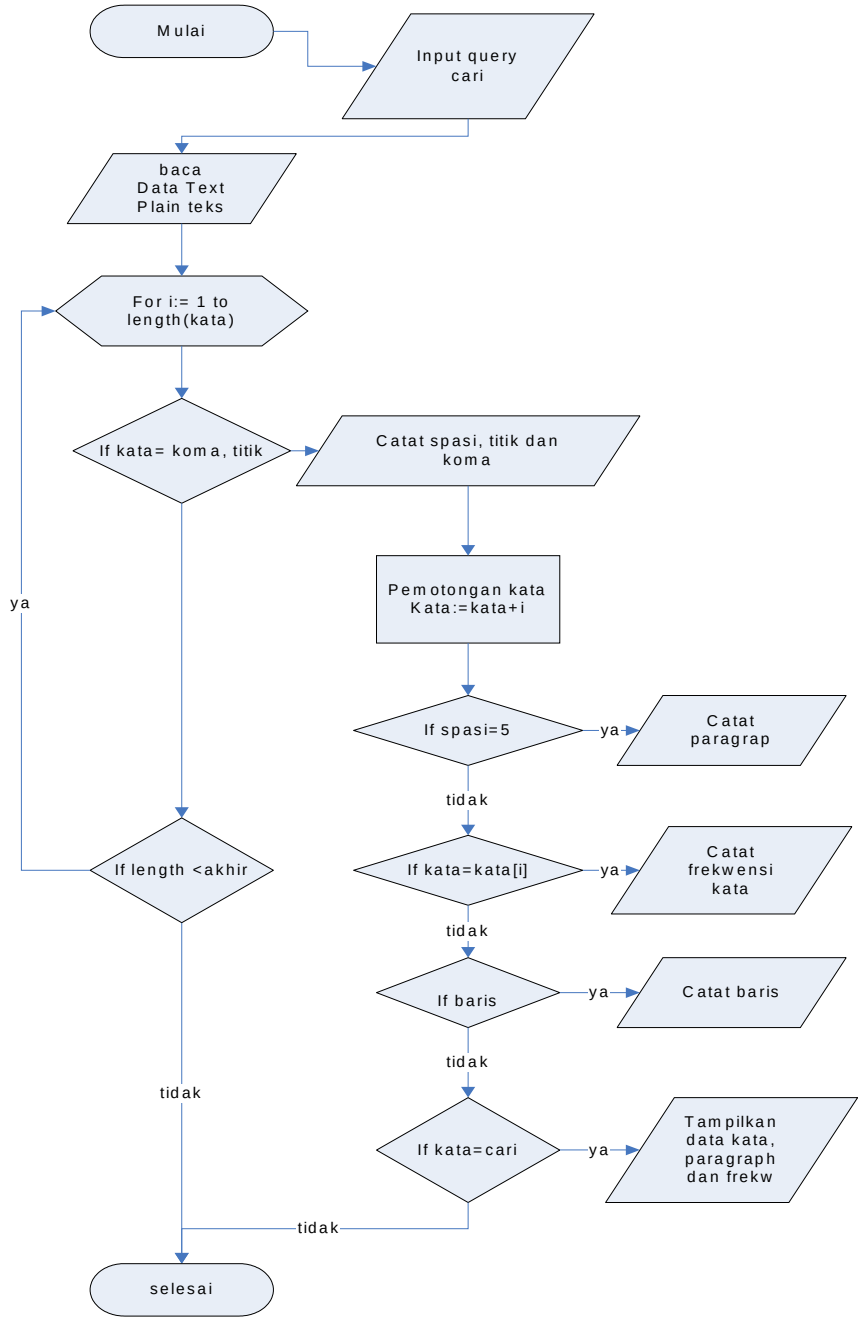
{mengambil dokumen data.txt ke richedit}

```
procedure TForm1.BitBtnBacaClick(Sender: TObject);
var baris:string;
begin
RichEdit1.Clear;
Assignfile(berkastext,'data.txt');
reset(berkastext);
while not (eof(berkastext)) do
  begin
    readln(berkastext,baris);
    RichEdit1.Lines.Add(baris);
  end;
 closefile(berkastext);
 {tutup file}
end;
```

TAHAPAN INDEKS



Gambar 1. Indek yang dicari



Gambar 2. Flow Chart pembacaaan kata

Algoritma pembacaan perkata

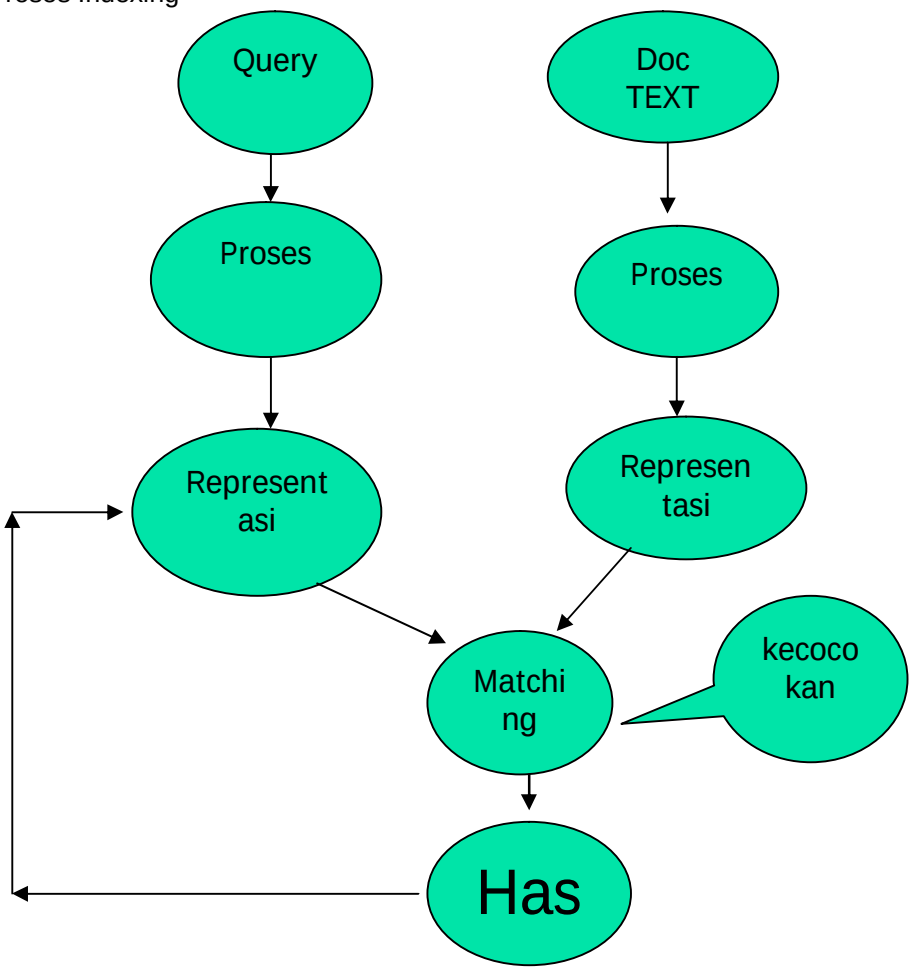
0. Mulai
1. Masukkan query
2. Baca teks
3. For I 1 to length kata
4. Jika If kata=spasi koma titik
5. Catatat koma titik spasi
6. Spasi maka Kata=kata+kata[I]

7. Jika spasi-5 maka paragrh
8. Jika kata=kata[I] maka catat frek
9. Jika akhir baris maka cata baris
10. Cek jika cari=kata maka tampilkan
11. Jika length < akhir maka lanjut 3
12. Jika tidak langkah 13
13. selesai

Algoritma pencarian kata

1. Mulai Baca Data
2. Scanning
3. Tokken=token spasi titik dan koma di hilangkan
4. Tahapan parshing pada kata-kata dengan mendeteksi dengan representasi diantaranya:
   Paragraph dengan spasi 5 maka paragraph
   Jika ketemu spasi, titik dan koma maka itu kata
   Kata sampai dengan length kalimat maka satu baris
   Catat frekwensi kata yang sama kata=kata[I];
   Catat paragprah dan baris dengan counter
5. Tampilkan hasil

Bagan Proses Indexing



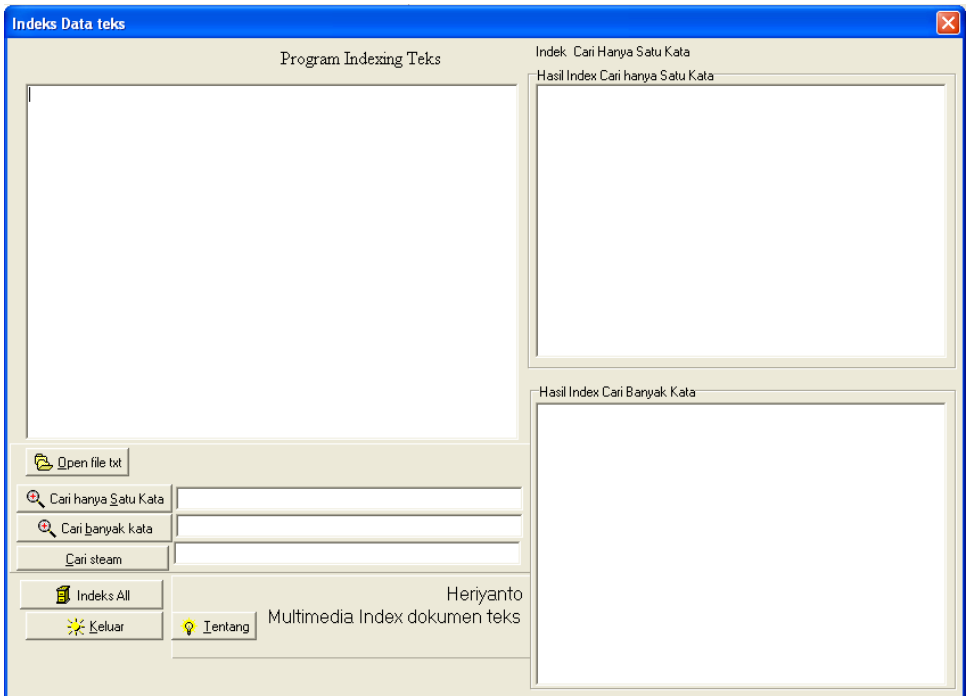
Gambar 3. Proses Indexing

Data *query* yang diminta akan dicek dengan dokumen teks dilakukan proses indek dengan representasi antara data proses indeks dengan dengan dokumen dilakukan kecocokan dengan

matching dilakukan *counter* baik paragraph, baris dan jumlah kata dalam frekwensi kemunculan.

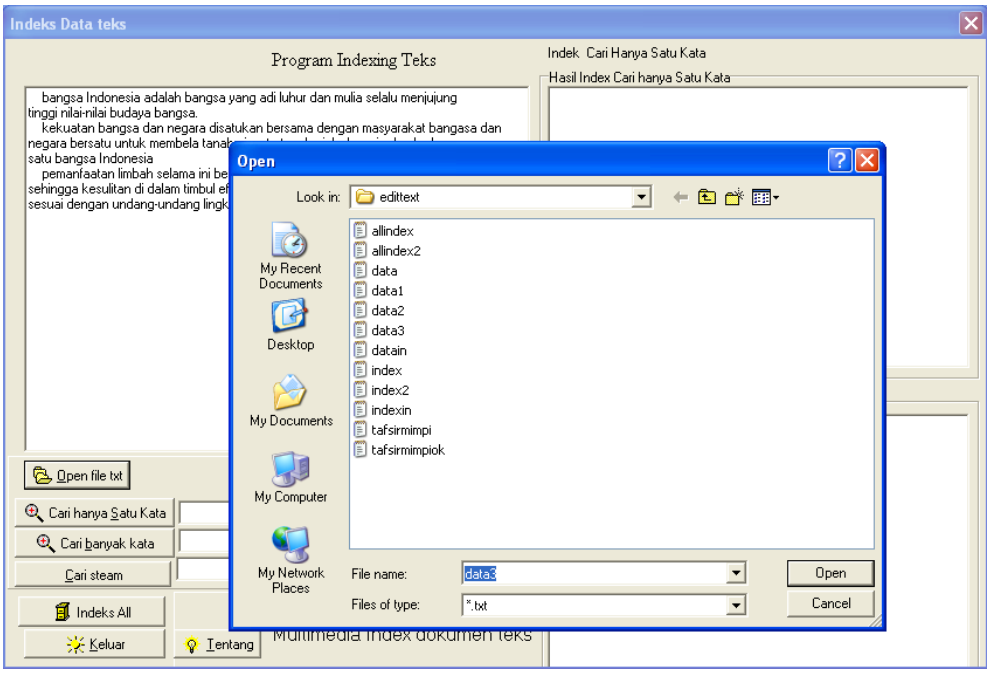
4. DESAIN DAN RANCANGAN

Dalam program tampilan rancangan sebagai berikut :

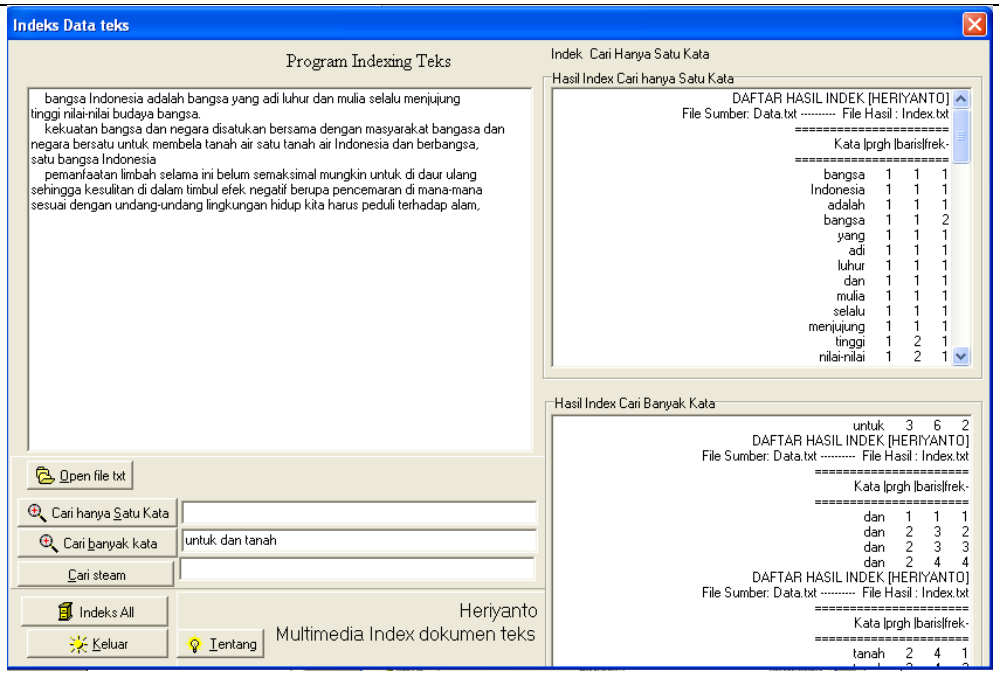


Gambar 4. menu utama *indexing*

Tampilan menu utama *indexing* menu sebelah kiri data sumber naskah dokumen teks dapat diambil dari file atau open folder.

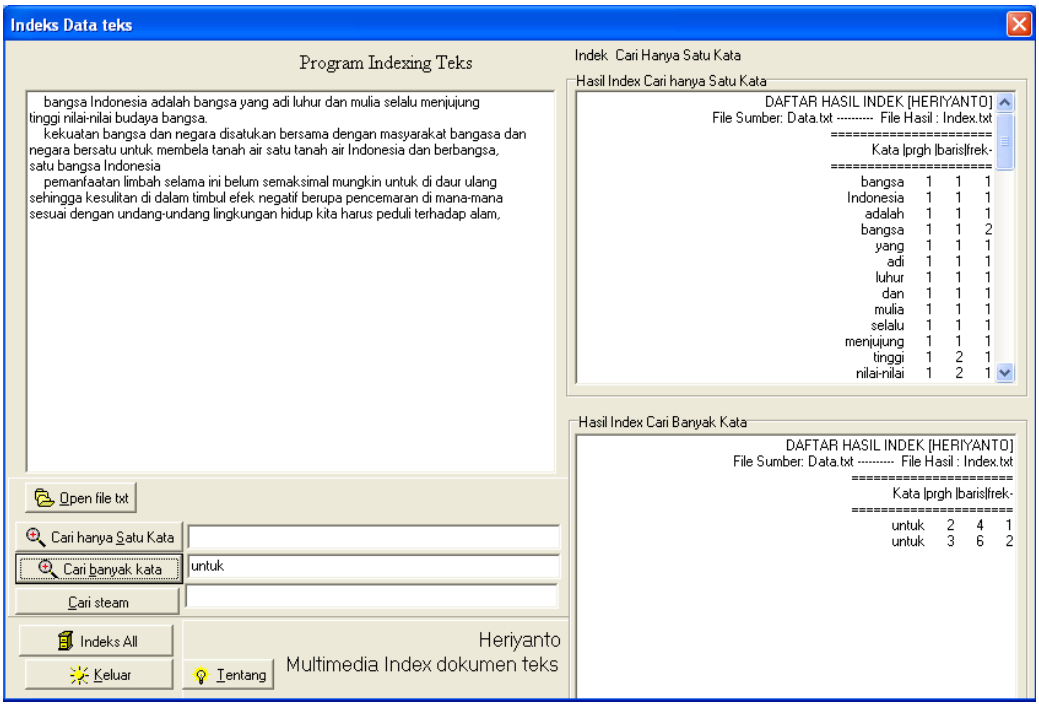


Gambar 5. membuka dokumen teks pada data file

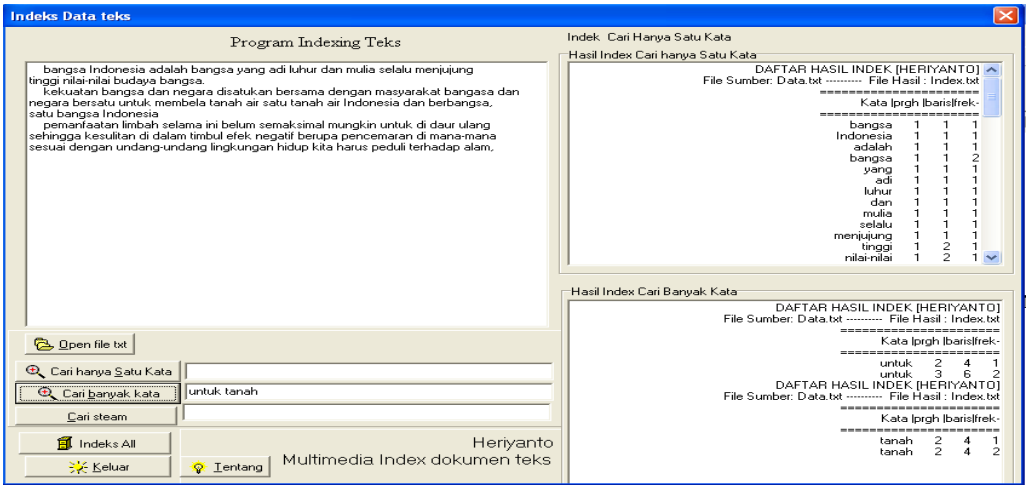


Gambar 6. melakuan indek data kata untuk dan tanah

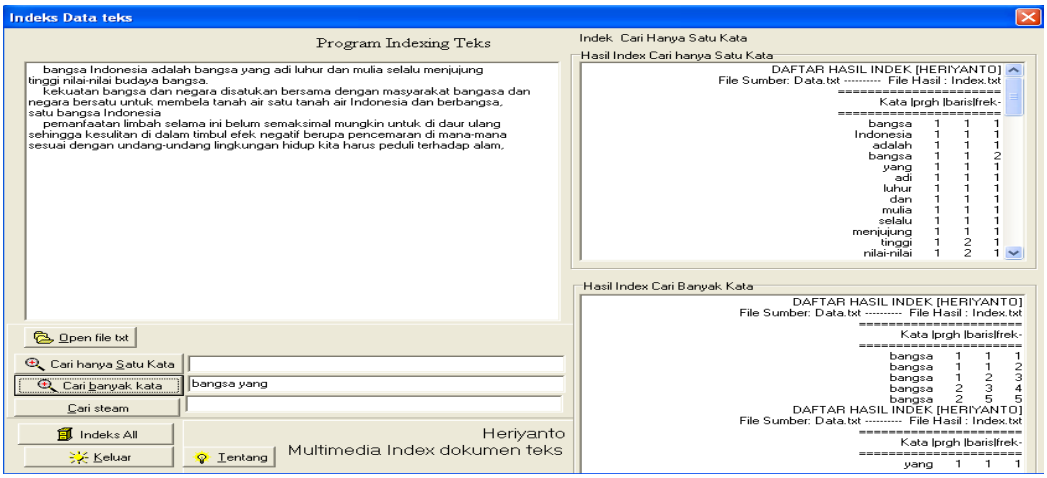
Program indexing data diambil dari file dilakukan pencarian kata dengan banyak kata ditampilkan di sebelah kanan.



Gambar 7. melakuan pencarian data kata “untuk



Gambar 8. melakuan pencarian data kata “untuk” dan “tanah”



Gambar 9. melakuan pencarian data kata

Hasil pencarian banyak kata ditampilan letak paragraph, letak pada baris dan jumlah kata.

5. KESIMPULAN

Secara garis besar sistem yang dibangun untuk mengetahui naskah sumber dan naskah target dilakukan untuk identifikasi dengan pembuatan indeks mencari letak kata, paragraph dan jumlah frekwensi kata tersebut dalam suatu paragraph. Melakukan scanning awal pertama untuk memisahkan kata, koma, titik dan dilakukan identifikasi pasda masing-masing kata terdapat pada baris, paragraph dan frekwensi kata tersebut. Pencarian selanjutnya dengan mengecek keberadaan kata tersebut dengan pencarian satu kata, dua kata atau banyak kata maka dilakukan indeks data kata tersebut terdapat pada baris, paragraph dan kemunculan kata tersebut dengan menghitung *counter* frekwensi kata tersebut.

DAFTAR PUSTAKA

Candra, Ian , *Utility Komputer Multimedia*, 1999, Elex Media Komputindo, Jakarta

Lu, Guajun, *Multimedia Database Manajemen Systems,* 1999 Artech House, Inc

Martina, Inge, *36 Jam Belajar Komputer Delphi 5.0 Database Client/Server Menggunakan Delphi*, 2000, PT. Elex Media Komputindo, Jakarta

Sanjaya Dwi*, Bertualang dengan Struktur Data di Planet Pascal*, edisi Pertama 2001, J& J Learning Yogyakarta

Silberschatz, Korth, Sudarsan, *Databases Systems Concept, 4th ed*, 2002, McGrawHill