Pengindeksan Dan Capaian Dokumen Dalam Pangkalan Data Teks Bebas : Satu Pendekatan

Sarudin Bin Kari
Safie Bin Mat Yatim
Nik Zalbiha bte Nik Mat

Institut Sains Komputer,
Universiti Teknologi Malaysia,
Jalan Semarak, Kuala Lumpur.

Abstrak

Pangkalan data teks bebas adalah sesuatu yang agak unik dan berbeza dari pangkalan data lazim. Dokumen-dokumen di dalamnya merupakan rekod-rekod yang mempunyai panjang yang berbeza-beza dan tanpa struktur yang khusus. Ianya memerlukan bukan sahaja teknik pengolahan data atau dokumen yang khusus, tetapi teknik atau pendekatan pengindeksan yang diperlukannya juga berbeza daripada teknik yang digunakan untuk pangkalan data lazim.

Kertas kerja ini akan membincangkan proses-proses yang perlu dilalui untuk membina satu sistem pengurusan pangkalan data teks bebas yang besar. Di samping itu, teknik pengindeksan dan capaian dokumen juga dibincangkan.

Katakunci : *pangkalan data teks, dokumen, indeks, katakunci, pengindeksan dokumen, capaian dokumen.*

Abstract

Free text database is considered something unusual and different from conventional databases. Records in this database are documents which have no fixed structure or forms. Hence, databases of this kind requires a different kind of data or documents manipulation because of the unconventionality of the techniques used in indexing the database.

This paper discusses all process which are involved in building a large free text database management system. In addition of that, documents indexing and retrieval techniques will also be discussed.

Keywords : *text database, document, index, keyword, document indexing, document retrieval*

1.0. Pengenalan

Kemajuan dan pencapaian di bidang teknologi maklumat dan komputer telah menyediakan satu pilihan yang *bererti* kepada kaedah penyimpanan dan penyampaian maklumat. Implikasi dari kemajuan dan pencapaian ini, sebahagiannya telah mengubah kaedah penyimpanan dan penyampaian maklumat, dari yang berasaskan bahan bercetak kepada kaedah yang berasaskan komputer.

Salah satu bidang yang menjadi tumpuan utama penyelidikan oleh penyelidik-penyelidik sains maklumat sekarang adalah untuk mendapatkan kaedah-kaedah pengkomputeran yang sesuai untuk sistem pengendalian dan capaian dokumen. Ini kerana sebahagian besar daripada maklumat yang terkandung dalam bahan-bahan bercetak (seperti buku dan manual) yang masih lagi menjadi media utama maklumat adalah berbentuk dokumen. Penyelidikan di bidang ini telah menghasilkan beberapa sistem pengendalian dan capaian dokumen, termasuklah sistem-sistem yang dikategorikan di bawah nama "hypertext" yang sangat popular pada masa ini.

Kelebihan utama sistem-sistem pengendalian dan capaian dokumen terutamanya sistem-sistem "hypertext" adalah kebolehan sistem-sistem tersebut menyediakan hubungan dan capaian tak linear antara dokumen-dokumen, di samping kemudahan capaian dokumen secara linear (sama seperti capaian dokumen dalam bahan-bahan bercetak). Walau bagaimanapun, kebanyakan sistem-sistem tersebut memerlukan penyediaan dan pengstrukturan semula dokumen-dokumen mengikut struktur atau format yang sesuai dengan ketetapan sistem-sistem berkenaan.

Sistem pengurusan pangkalan data dokumen teks bebas merupakan satu ***sistem pengindeksan dan capaian dokumen dalam pangkalan data teks bebas*** yang (sedang) dibangunkan untuk mengatasi kekurangan tersebut. Matlamat utama sistem ini adalah bagi membolehkan dokumen-dokumen dalam bentuk teks bebas, yakni dokumen-dokumen yang tidak mempunyai struktur khusus, dikendali dan dicapai tanpa melalui proses penyediaan dan pengstrukturan semula. Sesuatu dokumen di dalam pangkalan data teks bebas ini mungkin dibentuk oleh satu perkataan, satu ayat, satu perenggan, satu mukasurat, ataupun satu artikel. Perkataan merupakan unit terkecil yang boleh membentuk sesuatu dokumen. Ini bermakna, dokumen-dokumen di dalam pangkalan data teks bebas merupakan rekod-rekod yang mempunyai panjang yang berbeza-beza.

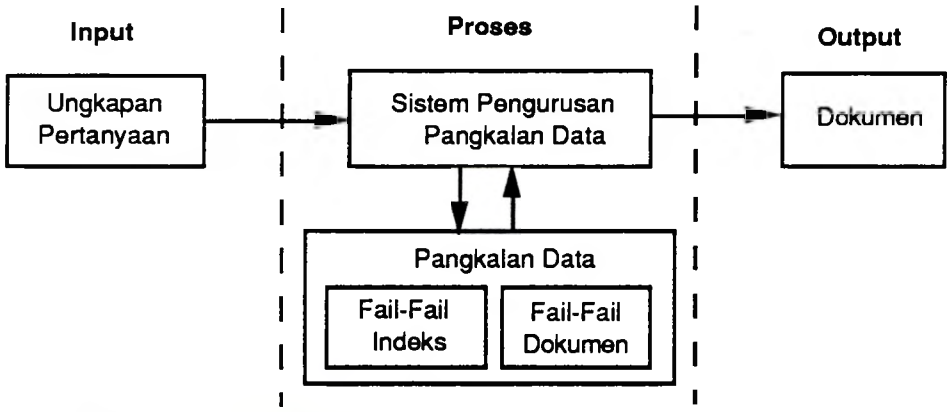
Untuk membolehkan dokumen-dokumen di dalam pangkalan data teks bebas dicapai, dokumen-dokumen tersebut perlu diindeks berdasarkan katakunci-katakunci tertentu. Untuk melakukan pengindeksan ini, setiap perkataan yang wujud dalam fail dokumen akan digunakan sebagai ***katakunci***. Dengan cara ini, proses pengindeksan dokumen boleh dilaksanakan secara automatik, tanpa perlu proses penyediaan dan pengstrukturan semula dokumen-dokumen tersebut. Pendekatan ini juga membolehkan sesuatu dokumen dicapai melalui satu atau lebih katakunci, bergantung kepada bilangan perkataan yang berbeza yang wujud dalam dokumen tersebut. Setiap satu katakunci pula boleh digunakan untuk mencapai sekurang-kurang satu dokumen.

Dalam melaksanakan fungsi-fungsi yang dinyatakan di atas, sistem pengurusan pangkalan data dokumen teks bebas ini dipecahkan kepada 2 subsistem, iaitu **subsistem pengindeksan** dan **subsistem capaian dokumen**. Subsistem pengindeksan merupakan bahagian yang berfungsi membina pangkalan data dokumen dengan melakukan proses-proses seperti penentuan katakunci, pengisihan katakunci, dan pengindeksan dokumen. Subsistem capaian dokumen pula, secara umum dibentuk oleh tiga komponen utama (lihat rajah 1), iaitu:

i. **Input**- menerima ungkapan pertanyaan pangkalan data ;

ii. **Proses**- melakukan penghuraian ungkapan pertanyaan untuk menentukan katakunci-katakunci dan operasi yang diperlukan, serta mencari dan mencapai dokumen-dokumen yang dikehendaki, berdasarkan hasil penghuraian ungkapan pertanyaan; dan

iii. **Output**- memaparkan dokumen-dokumen yang berkaitan sebagai bukti capaian.



Rajah 1 : Rekabentuk Umum Subsistem Capaian Dokumen.

2.0. Ungkapan Pertanyaan Dokumen

Terdapat berbagai bentuk antaramuka untuk pertanyaan pangkalan data yang telah digunakan oleh sistem-sistem pengurusan pangkalan data, tetapi kebanyakannya merupakan pertanyaan berasaskan ungkapan berkatakunci. Dua bentuk antaramuka yang sering disediakan adalah *pilihan menu* dan *ditaip menggunakan papan kekunci*. Pertanyaan pula boleh mengambil bentuk pertanyaan berasaskan satu katakunci atau pertanyaan berasaskan beberapa katakunci yang digabungkan dalam satu ungkapan (seperti ungkapan Boole). Terdapat juga sistem-sistem pengurusan pangkalan data yang menggunakan ungkapan bahasa tabii sebagai antaramuka untuk pertanyaan pangkalan data. Walau bagaimanapun, sistem yang dapat menyediakan kemudahan ini sangat terhad, bukan sahaja dari segi bilangannya, tetapi juga bentuk ungkapan bahasa tabii yang boleh digunakan.

Bagi sistem pengurusan pangkalan data dokumen yang sedang dibina, ungkapan Boole telah dipilih sebagai perantaraan yang membolehkan pengguna membuat pertanyaan untuk mencapai dokumen. Pilihan ini dibuat kerana penghurai ungkapan Boole lebih mudah untuk dibangunkan berbanding dengan penghurai bagi bahasa tabii. Di samping itu juga, ianya mencukupi untuk membolehkan pertanyaan capaian berbilang dokumen dengan gabungan beberapa katakunci diwakilkan melalui satu ungkapan sahaja. Sistem ini menyediakan dua kaedah pembentukan ungkapan pertanyaan yang boleh digunakan oleh pengguna untuk katakunci dan juga untuk operator Boole iaitu *pilihan menu* dan *menaip melalui papan kekunci*. Operator Boole yang dibenarkan untuk ungkapan pertanyaan bagi sistem ini adalah operator **OR, AND, NOT** dan **kurungan** (penentu atau pemisah keutamaan).

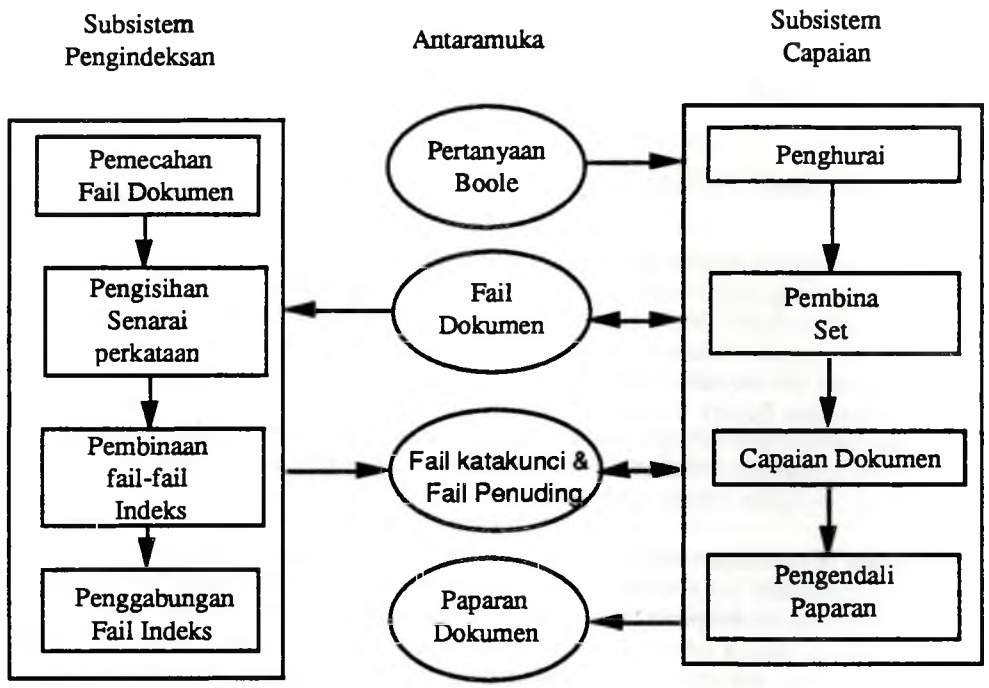
2.1. Keutamaan Operator Boole

Keutamaan operator dan arah imbasan merupakan dua perkara penting yang akan menentukan perjalanan penghurai dan hasil operasi yang akan dilaksanakan. Pada umumnya, terdapat dua pilihan strategi bagi arah imbasan yang sering digunakan iaitu *imbasan dari kiri ke kanan* dan *imbasan dari kanan ke kiri*. Bagi keutamaan operator pula, pada kebiasaannya operator **NOT** diutamakan dahulu, diikuti oleh operator **AND** dan akhirnya operator **OR**. **Kurungan** yang sering digunakan sebagai *penentu atau pemisah keutamaan*, biasanya diberi keutamaan yang paling tinggi dalam melaksanakan sesuatu operasi.

Penghurai bagi sistem yang sedang dibina menggunakan strategi imbasan dari kiri ke kanan. Keutamaan operator pula berdasarkan keutamaan yang digunakan oleh kebanyakan penghurai ungkapan Boole yang terdapat dalam sistem-sistem lain, di mana **kurungan** mempunyai keutamaan yang tertinggi, diikuti oleh operator **NOT**, operator **AND** dan akhir sekali **OR** sebagai operator yang terendah keutamaannya.

3.0. Rekabentuk Sistem Dan Struktur Pangkalan Data

Sebagaimana yang telah dijelaskan sebelum ini, subsistem pengindeksan merupakan bahagian yang membina pangkalan data. Subsistem capaian dokumen pula adalah subsistem yang mencapai pangkalan data yang telah dihasilkan oleh subsistem pengindeksan. Dalam konteks sebenar, kedua-dua subsistem ini merupakan sistem-sistem yang lengkap, di mana sistem-sistem *dilarikan* secara berasingan. Kaitan secara tidak langsung atau secara logikal di antara kedua-dua subsistem ini hanyalah melalui fail dokumen, dan dapat digambarkan seperti rajah 2.



Rajah 2 : Proses Dan Antaramuka Antara Subsistem

3.1. Subsistem Pengindeksan

Subsistem pengindeksan melibatkan empat modul atau proses utama iaitu :

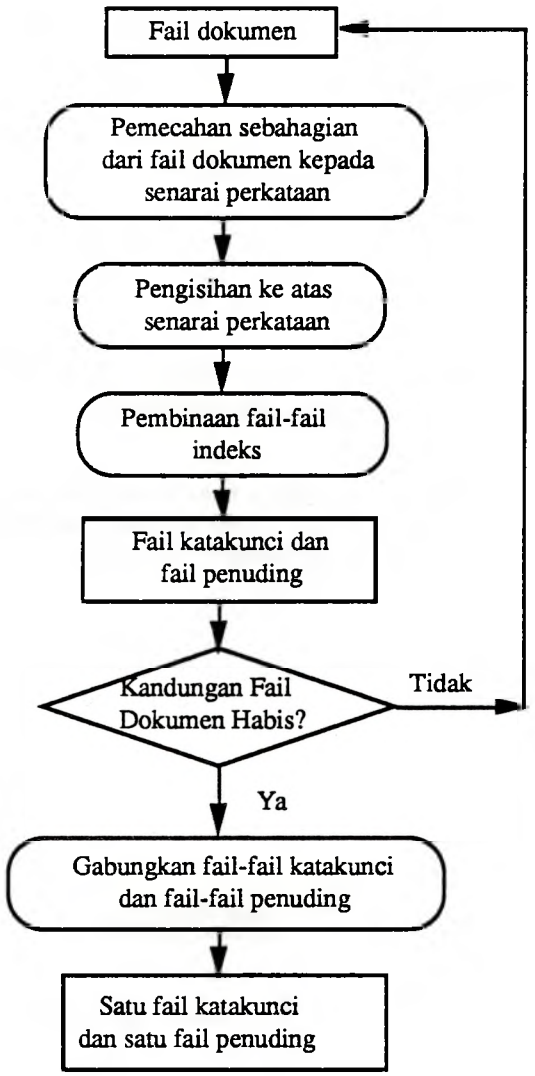
i. **Pemecahan Fail Dokumen** - Untuk memecahkan bahagian-bahagian fail dokumen kepada perkataan-perkataan bagi mendapatkan katakunci-katakunci.

ii. **Pengisihan Senarai Perkataan** - Untuk melakukan isihan ke atas senarai perkataan bagi menghasilkan senarai katakunci terisih.

iii. **Pembinaan Fail-Fail Indeks** - Untuk membina fail-fail indeks (fail katakunci dan fail penuding) bagi setiap bahagian fail dokumen.

iv. **Penggabungan Fail-Fail Indeks** - Untuk menggabungkan fail-fail katakunci, dan fail-fail penuding bagi setiap bahagian fail dokumen menjadi satu fail katakunci dan satu fail penuding sebagai hasil akhir proses pengindeksan sesuatu fail dokumen.

Aliran dan hubungan antara proses-proses bagi subsistem pengindeksan ini boleh digambarkan seperti dalam rajah 3 berikut :



Rajah 3: Subsistem Pengindeksan

Sinopsis pendekatan yang digunakan untuk proses-proses yang dilaksanakan oleh subsistem pengindeksan adalah :

i. Program akan membaca kandungan fail dokumen (input) dan semasa proses pembacaan ini, operasi berikut akan dilakukan :

a) Pengesanan dan pemecahan kandungan fail dokumen kepada perkataan dilakukan berdasarkan ruang kosong (blank) dan aksara-aksara khas tertentu sebagai tanda pemisah perkataan. (Di samping ruang kosong, semua simbol-simbol khas seperti tanda seruan (!), tanda soal (?) dan lain-lain tidak dikira sebagai satu aksara dalam perkataan dan sebahagiannya digunakan sebagai penentu atau pembeza antara dua perkataan yang bersebelahan).

b) Penentuan lokasi permulaan dokumen bagi perkataan.

c) Pembentukan rekod yang terdiri daripada katakunci dan lokasi permulaan dokumen untuk katakunci (perkataan) berkaitan.

Rekod-rekod yang terhasil daripada operasi di atas ditempatkan atau disenaraikan dalam satu penimbal (di dalam ingatan utama) dan proses pembacaan fail dokumen akan digantung sementara setelah penimbal tersebut penuh atau mencapai satu had yang ditetapkan.

ii. Proses pengisihan rekod-rekod berdasarkan katakunci akan dilaksanakan dalam penimbal tersebut untuk menghasilkan satu senarai rekod yang terisih.

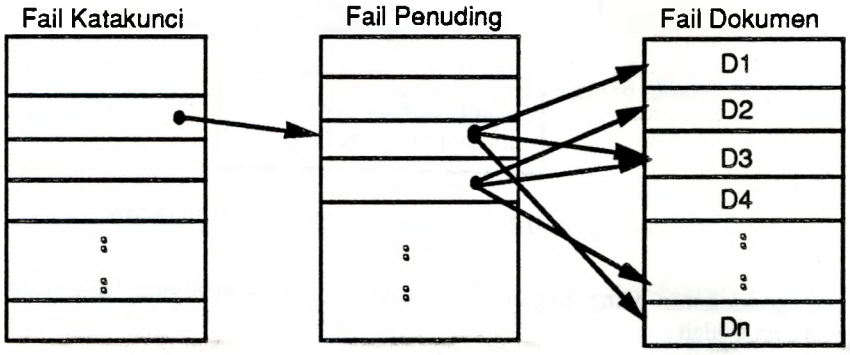
iii. Proses pemecahan (katakunci dan penuding) dan penyalinan rekod-rekod yang terisih dari penimbal ke dalam fail katakunci dan fail penuding yang baru.

iv. Proses (i) hingga (iii) akan diulang sehingga keseluruhan kandungan fail dokumen diindekskan.

v. Setelah semua kandungan fail dokumen diindekskan, fail-fail katakunci dan fail-fail penuding untuk setiap bahagian fail dokumen tadi digabungkan menjadi satu fail katakunci dan satu fail penuding.

Pendekatan yang dijelaskan di atas diambil bagi membolehkan fail dokumen yang bersaiz besar diindekskan. Pendekatan pengindeksan secara berperingkat-peringkat ini adalah penting untuk membolehkan sistem yang dibangunkan ini digunakan pada komputer yang mempunyai ingatan utama yang terhad. Di samping itu, proses penggabungan fail-fail katakunci dan penggabungan fail-fail penuding dilaksanakan dengan menggunakan satu kaedah pengstrukturan yang akan ditentukan. Penentuan kaedah ini akan dibuat berasaskan kepada keberkesanan dan kecekapan dari segi penggunaan ruang dan masa yang boleh disediakan oleh sesuatu kaedah dalam melakukan capaian ke atas dokumen. Untuk tujuan ini, beberapa kaedah telah dikaji dan sedang diuji, di antaranya adalah kaedah pepohon-B, pepohon dedua, dan "trie".

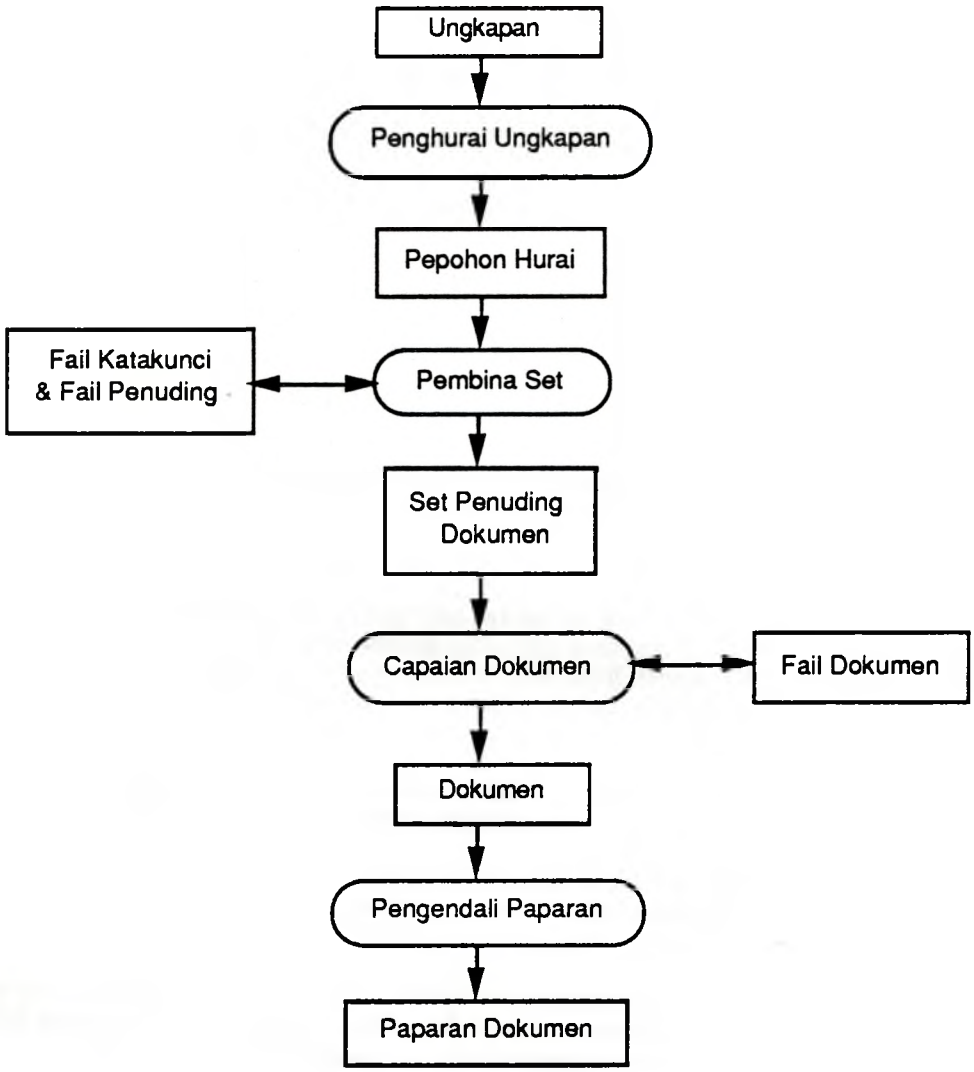
Pangkalan data yang dihasilkan oleh subsistem ini dibentuk oleh tiga fail iaitu *fail katakunci*, *fail penuding* dan *fail dokumen*. Hubungan secara logikal di antara ketiga-tiga fail tersebut dapat digambarkan dengan rajah 4.



Rajah 4 : Struktur Pangkalan Data

3.2. Subsistem Capaian Dokumen

Subsistem capaian dokumen dibentuk oleh empat modul atau proses utama iaitu penghurai,



Rajah 5 : Subsistem Capaian Dokumen.

pembina set, capaian dokumen, dan pengendali paparan. Aliran dan hubungan antara proses-proses tersebut boleh digambarkan seperti dalam rajah 5.

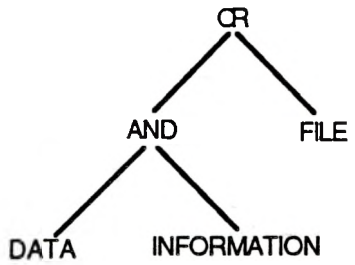
3.2.1. Penghurai

Ungkapan pertanyaan untuk mencapai dokumen yang diajukan oleh pengguna akan melalui proses pengimbasan dan pemecahan. Proses ini dilaksanakan oleh penghurai untuk menentukan katakunci-katakunci dan operasi-operasi yang perlu dilaksanakan. Seterusnya penghurai akan melaksanakan proses pembinaan pepohon hurai yang sesuai berdasarkan ketetapan keutamaan operator Boole (lihat 2.1) bagi ungkapan pertanyaan tersebut.

Sebagai contoh, satu bentuk ungkapan pertanyaan Boole yang mungkin diajukan oleh pengguna untuk mencapai dokumen-dokumen ialah :

(DATA AND INFORMATION) OR FILE

Terdapat tiga katakunci dalam ungkapan pertanyaan di atas iaitu DATA, INFORMATION dan FILE, serta tiga operator iaitu *kurungan*, AND dan OR. Berdasarkan hasil proses imbasan dan pecahan tersebut, proses pembinaan pepohon hurai untuk ungkapan di atas akan menghasil pepohon seperti dalam rajah 6.



Rajah 6 : Contoh Pepohon Hurai

3.2.2. Pembina Set

Berdasarkan pepohon hurai yang dihasilkan oleh penghurai, pembina set akan membina satu **set** penuding kepada dokumen-dokumen yang perlu dicapai. Semasa proses pembinaan set ini, pembina set akan mencapai **fail katakunci** untuk menentukan kewujudan katakunci dan dokumen yang berkaitan dan mencapai **fail penuding** untuk menentukan set penuding kepada dokumen-dokumen bagi katakunci yang berkaitan.

Sebagai contoh, berdasarkan pepohon hurai dalam rajah 6, pembina set akan melakukan proses yang boleh diringkaskan seperti berikut (dengan anggapan semua katakunci wujud) :-

i. Menyemak katakunci DATA dalam fail katakunci dan membentuk set penuding kepada dokumen-dokumen yang mempunyai katakunci DATA dengan mencapai rekod berkaitan dalam fail penunding - **SET 1.**

ii. Menyemak katakunci INFORMATION dalam fail katakunci dan membentuk set penuding kepada dokumen-dokumen yang mempunyai katakunci INFORMATION dengan mencapai rekod berkaitan dalam fail penuding - **SET 2.**

iii. Melakukan operasi AND bagi **SET 1** dan **SET2** untuk menghasilkan satu set penuding kepada dokumen-dokumen yang mempunyai katakunci DATA dan INFORMATION - **SET 3.**

iv. Menyemak katakunci FILE dalam fail katakunci dan membentuk set penuding kepada dokumen-dokumen yang mempunyai katakunci FILE dengan mencapai rekod berkaitan dalam fail penuding - **SET 4.**

v. Melaksanakan operasi OR bagi **SET 3** dan **SET 4** untuk menghasilkan set penuding kepada dokumen-dokumen yang mempunyai katakunci DATA dan INFORMATION, atau/dan FILE - **SET 5** (set penuding kepada dokumen-dokumen yang akan dicapai dan dipamerkan).

3.2.3. Capaian Dokumen

Modul capaian dokumen akan mencari dan mencapai dokumen-dokumen yang diperlukan oleh pengguna dari fail dokumen berdasarkan set penuding yang telah dihasilkan oleh pembina set.

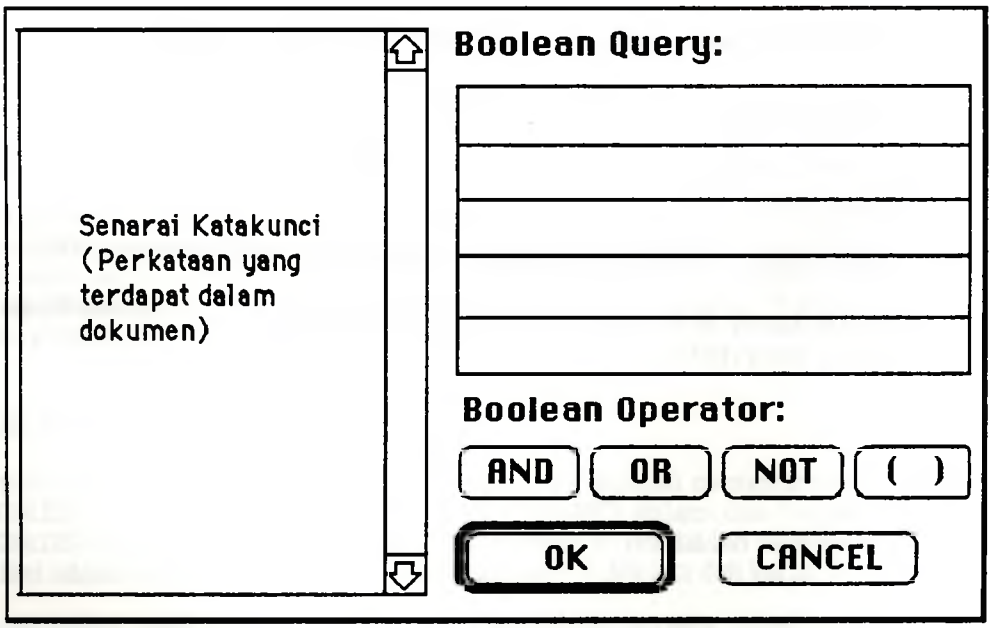
3.2.4. Pengendali Paparan

Dokumen yang dicapai oleh modul capaian dokumen akan dihantar kepada pengendali paparan untuk dipamerkan di tetingkap paparan dokumen. Pengendali paparan menyediakan kemudahan yang membolehkan pengguna melihat kandungan sesuatu dokumen seperti *menggulung* (scrolling) kandungan dokumen ke atas dan ke bawah. Di samping itu, modul ini juga membenarkan pengguna memilih dokumen (ahli set semasa dokumen-dokumen yang dicapai) yang hendak dipaparkan seterusnya, sama ada dokumen sebelum atau selepas dokumen semasa.

4.0. Antaramuka Sistem dan Pengguna

Rekabentuk antaramuka pengguna dan sistem merupakan satu perkara yang penting kerana ianya menentukan sama ada sesuatu sistem itu mudah digunakan atau sebaliknya. Bagi sistem pengurusan pangkalan data dokumen teks bebas yang sedang dibina, dua tetingkap antaramuka pengguna dan sistem disediakan iaitu *tetingkap paparan dokumen* dan *tetingkap pertanyaan capaian dokumen*. Kedua-dua tetingkap ini sentiasa tersedia untuk dicapai oleh pengguna.

Tetingkap paparan dokumen merupakan tetingkap output yang digunakan untuk mempamerkan dokumen-dokumen yang diminta oleh pengguna melalui ungkapan pertanyaan Boole. Tetingkap pertanyaan capaian dokumen pula adalah tetingkap input yang membolehkan pengguna menyampaikan ungkapan pertanyaan Boole (untuk mencapai dokumen). Seperti yang telah dijelas sebelum ini (dalam 2.0), melalui tetingkap ini (lihat rajah 7) pengguna boleh membentuk ungkapan pertanyaan Boole di ruang yang disediakan sama ada melalui *pilihan menu* atau *menaip dengan menggunakan papan kekunci*.



Rajah 7 : Tetingkap pertanyaan capaian dokumen - input.

5.0 Penutup

Sistem pengurusan pangkalan data dokumen teks bebas merupakan satu sistem perisian yang dibangunkan khas untuk membolehkan pangkalan data dokumen dibina dari fail teks bebas secara

automatik dan capaian semula dokumen-dokumen di dalam pangkalan data tersebut dibuat. Dengan sistem ini, dokumen-dokumen teks bebas tidak perlu melalui proses penyediaan dan pengstrukturan semula mengikut format tertentu. Kelebihan inilah yang membezakan sistem ini daripada sistem-sistem pengendalian dan capaian dokumen yang lain walaupun mungkin terdapat perbezaan tertentu dari segi penggunaannya.

Perkara yang diberi tumpuan utama untuk sistem yang sedang dibangunkan ini adalah kebolehan mengendali pangkalan data teks bebas yang besar secara cekap dan berkesan. Dengan kemampuan tersebut, sistem ini juga boleh dijadikan asas untuk menyediakan sistem-sistem yang memerlukan atau melibatkan amaun teks bebas yang besar dan yang berasaskan bahasa tabii seperti sistem pangkalan data bahasa atau korpus dan sistem bantuan pembinaan kamus.

Bibliografi

[**Salton 1983**] Salton, G. and M.J. McGill, Introduction to Modern Information Retrieval. McGraw Hill International Book Company (1983).

[**Harbron 1988**] Harbron, T. R , File Systems: Structures and Algorithms. Prentice-Hall International Edition (1988).

[**Wiggins 1986**] Wiggins, R. and P.Wolberg, Searching for Strings with Boyer-Moore, Computer Language (Nov 1986).

[**Menico 1989**] Menico, C., Faster String Searches, Dr.Dobb's Journal. (July 1989).

[**Smith 1987**] Smith, E. J., Multiple Word Searches witch, Computer Language (Nov. 1987).

[**Purdum 1987**] Purdum, J., Pattern Matching Alternatives: Theory vs. Practice, Computer Language. Nov. (1987).

[**Holub 1987**] Holub , A.I, The C Companion. Prentice-Hall software series (1987).

[**Knuth 1973**] Knuth ,D.E., The Art Of Computer Programming. Addison-Wesley Publishing Company, Vol 3 (1973).