

Penerapan Methode Query Search Method pada Multidimensional Database sebagai Aplikasi Information Retrieval dalam Datamining

Elfino Sitompul 13508098
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jl. Ganesha 10 Bandung 40132, Indonesia
If18098@students.if.itb.ac.id

Abstract—Nowdays, system just become more interactive and many people make a multidimensional object that needs to have their own database. But A basic kind of database type do not longer need. Because the response times not as good as the performance they need to maximize the system. So the new search method algorithm is needed to empower a new database system.

Index Terms—Database, Multidimensional, Search Method, Algorithrm, Pattern Matching

I. PENDAHULUAN

Alasan utama mengapa kebutuhan basis data saat ini sangat diperlukan adalah data mining sekarang telah menjadi sebuah hal yang menarik untuk dilakukan di dalam industri informasi. Informasi dan pengetahuan yang didapat dari berbagai sistem, seperti manajemen bisnis, kontrol produksi dan analisa pasar dapat disimpan untuk kelak dapat dipakai dan ditelaah lebih lanjut.

Basis data yang dipakai lama kelamaan semakin membesar dan sistem pencarian query yang dipakai akan semakin sulit, untuk itu dibutuhkan suatu sistem basis data yang baru untuk mempermudah berbagai operasi dalam basis data.

Salah satu metode yang ditawarkan adalah basis data dengan template multidimensional database. Dengan ini sistem basis data dapat dipandang dari berbagai sisi sehingga memperingkas operasi yang ada.

Untuk multi dimensional basis data sebenarnya masih menggunakan basis data yang biasa, hanya saja metode observasi basis data tersebut yang ditambah.

Multidimensional data adalah data yang memiliki beberapa dimensi. Hal ini tergantung dari cara kita memandang data tersebut. Penyusunan data itu sendiri dapat disusun berdasarkan keinginan kita. Tergantung dari cara kita untuk menggunakan data tersebut.

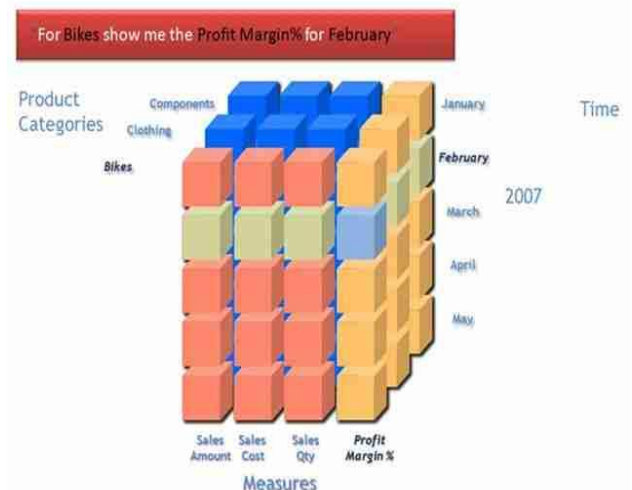
Strukturnya dapat disesuaikan untuk mengoptimisasi analisa dari data data tersebut berdasarkan relational database. Jadi dengan adanya relational ini kita memiliki berbagai sudut pandang untuk suatu data.

Multidimensional data model merupakan variasi dari model relasional yang menggunakan struktur

multidimensional untuk mengorganisasikan data dan menjelaskan relasi antara data.

Data warehouse merupakan repository data dari suatu organisasi, didesain untuk memfasilitasi reporting dan analisis. Walaupun pengertian data warehouse lebih memfokuskan diri pada penyimpanan data, namun pembacaan data, analisa, transformasi data juga dianggap penting sebagai komponen sistem data warehouse.

Datacube merupakan array yang mempunyai 3 atau lebih dimensi yang digunakan untuk menyimpan data. Dari datacube ini metode basis data berbasis relational biasa kita manipulasi sehingga terlihat seperti basis data 3 dimensi. Hal yang paling penting adalah kita mengatur sistem sehingga user biasa yang tidak mempunyai background dalam design sistem informasi dapat mempergunakan datacube secara multidimensional database, bukan basis data relational biasa.



Gambar 1. Contoh data cube

Salah satu contoh penggunaan multidimensional database adalah pada penggunaan sistem enterprise suatu perusahaan yang menggunakan registrasi atau penggunaan user/pelanggan. Banyak dari sistem enterprise ini yang mempergunakan multidimensional database dikarenakan menghemat jalur informasi yang ada.

Jalur informasi yang semakin dihemat tersebut membuat sistem basis data yang ditawarkan semakin interaktif karena response time dari sistem sangatlah bagus. Sehingga semakin mudah untuk digunakan oleh orang yang tidak memiliki background di dunia informasi. Dengan sistem yang semakin interaktif dapat dijamin pula keterlibatan user itu sendiri dalam mempergunakan multidimensional database untuk keperluan yang diperlukan.

Satu hal yang paling penting adalah mendefinisikan transformasi data dari database relational biasa menjadi multidimensional database. Salah satu cara yang sering dipakai adalah dengan membagi baris yang ada menjadi beberapa bagian, (biasanya 3) dan menerapkannya sebagai sisi lain dari database tersebut.

Penggunaan multidimensional database dapat dipakai di dalam data mining. Data mining adalah metode pencarian informasi dari basis data yang ada. Pencarian yang dimaksud merupakan pencarian berbasis analisis. Hal yang perlu diketahui adalah, kita memiliki banyak sekali database tapi terkadang kita menyadari bahwa sedikit sekali informasi yang kita dapatkan dari database ini. Di sinilah data mining bekerja. Data mining sistem akan mencari knowledge atau pengetahuan dari sistem basis data yang ada. Data mining terdiri dari beberapa tahap.

- a. Data cleaning
Membersihkan noise atau informasi yang tidak penting atau tidak berkaitan dengan penelitian.
- b. Data integration
Menggabungkan berbagai sumber data menjadi satu basis data.
- c. Data selection
Mengambil data yang sesuai dengan analisis task dari basis data.
- d. Data transformation
Mengubah data menjadi mudah untuk dilakukan penggalian data, biasanya mengalami operasi summary ataupun aggregation.
- e. Data mining
Proses penting dimana metode berbasis intelijen diterapkan guna mengambil informasi yang ada.
- f. Pattern evaluation
Mencari pola yang benar benar menarik yang merepresentasikan knowledge.
- g. Knowledge presentation
Menggunakan visualisasi dan representasi knowledge kepada user.

Yang akan saya bahas adalah metode bagian e dan f dimana sangat dibutuhkan metode retrieval data yang baik untuk menjamin mutu dari knowledge yang akan didapat.

II. PERANCANGAN SOLUSI

A. Tahap Analisis Permasalahan

Masalah yang paling sering kita temui dalam data mining adalah mencari knowledge yang tepat dengan menggunakan metode yang benar. Tapi dalam makalah

saya ini, saya menganggap atau menjamin ketetapan metode yang dipakai dalam menentukan atau mendapatkan knowledge yang ada. Yang menjadi pokok permasalahan adalah pencarian berdasarkan query.

Pencarian berdasarkan query dalam data yang sangat banyak akan mengakibatkan hasil yang didapat kurang baik. Untuk mendapatkan hasil yang baik maka hasil tersebut harus disadur kembali.

Metode penyaduran yang dipakai, karena kita mengambil tema basis data. Maka hanya dibatasi dalam query.

Metode query adalah metode mencocokkan masukan berupa query dengan database dan meretrieve semua data yang berkaitan dengan query tersebut. Dari data data inilah kelak akan didapatkan hasil yang akan diolah dengan metode pengolahan data mining sehingga menjadi knowledge.

Masalah dalam query, sering sekali kita menemukan hal hal yang tidak berkaitan dengan query yang kita inginkan hal ini bisa terjadi karena beberapa hal. Contoh stopword, dalam kasus stopword yang kita pakai di database sering kita temui dalam bagian keterangan untuk user. Kita banyak melihat kata "atau" yang tidak memiliki arti di dalam pencarian. Sehingga "atau" bisa kita hapus atau kita abaikan.

Contoh lain adalah penggunaan kata yang sudah jelas berhubungan contoh kata jalan di dalam bagian alamat sudah jelas berhubungan sehingga harus kita buang meskipun berkaitan dengan query yang kita inginkan contoh lokasi.

Untuk itu kita akan membuat metode yang mengabaikan hal hal ini sehingga hasil pencarian semakin optimal. Metode ini selanjutnya disebut Inverse Document Frequency atau metode IDF. Dengan ini kita tidak perlu khawatir apakah kata kata tersebut penting atau tidak.

B. Tahap Analisis Kebutuhan

Tahap ini akan mencari kebutuhan query apa yang kita inginkan contoh ketika kita dihadapkan untuk pencarian beberapa database kita akan menemukan fakta bahwa jumlah data di tiap database tersebut tidaklah sama. Data yang banyak cenderung akan menghasilkan hasil retrieval yang banyak pula, sedangkan data yang sedikit cenderung akan menghasilkan hasil yang sedikit pula. Hal ini membuktikan bahwa jumlah data yang banyak akan cenderung signifikan mempengaruhi knowledge yang akan kita terima kelak.

Tapi karena kita menganggap bahwa perlu adanya kestabilan perbandingan antara database database ini maka kita akan menerapkan metode yang akan menyelaraskan pencarian informasi pada beberapa database ini.

Solusi yang ditawarkan adalah dengan menggunakan normalisasi dengan ini masukan yang jumlahnya banyak bisa menjadi masukan yang hampir setara dengan masukan dengan jumlah yang sedikit.

Ada dua hal yang menyebabkan kita memerlukan normalisasi.

- Frekuensi dari query yang dicari
Dalam kasus basis data dengan data yang besar kita akan menemukan query yang kita cari akan banyak jumlahnya sehingga mempengaruhi nilai relevansinya. Untuk itu dibutuhkan normalisasi.
- Jumlah query dalam satu basis data
Karena dalam basis data dengan jumlah informasi yang banyak kita. Maka banyak juga jumlah query yang tidak berkaitan dengan query yang kita cari. Sehingga hal ini dapat menurunkan relevansi, untuk itu kita memerlukan normalisasi.

Normalisasi yang kita pakai adalah normalisasi cosine

$$\cos(\vec{Q}, \vec{D}) = \frac{\sum_{i=1}^r w_{qi} \times w_{di}}{\sqrt{w_{q1}^2 + w_{q2}^2 + \dots + w_{qr}^2} \times \sqrt{w_{d1}^2 + w_{d2}^2 + \dots + w_{dr}^2}}$$

Dimana :

wq = bobot $tf \times idf$ dari term i dalam query

wd = bobot $tf \times idf$ dalam dokumen

tf adalah term frekuensi dimana menentukan jumlah query yang berkaitan dengan query yang kita inginkan. Sedangkan idf adalah bobot dari nilai idf untuk mengabaikan kata-kata yang sudah jelas berkaitan dan tidak diperlukan (tahap analisis perancangan)

C. Tahap Perancangan.

Pada tahap sebelumnya kita sudah mendapatkan gambaran solusi apa yang akan dirancang untuk memastikan kebenaran knowledge yang diretrieve dalam data mining.

Sekarang kita menurunkan langkah untuk menggambarkan solusi berdasarkan analisis. Langkah yang pertama kali kita lakukan setelah mendapatkan retrieval data menggunakan pattern matching adalah memastikan nilai tersebut sesuai dengan yang kita inginkan. Kriteria yang kita inginkan adalah tidak mengandung stopword, tidak dipengaruhi jumlah kata yang sudah pasti berkaitan dan tidak dipengaruhi jumlah data yang terlalu banyak.

Solusinya adalah kita akan membuat system information retrieval yang menggunakan stopword removal, normalisasi dan idf . Sehingga hasil dari retrieval ini dapat dipakai untuk metode selanjutnya dalam datamining.

III. PENERAPAN ALGORITMA PATTERN MATCHING

Dari solusi yang kita bangun diatas. System bantuan pada knowledge retrieval dalam datamining kita petakan dalam metode pencarian query. Salah satu penerapan dalam pencarian query adalah dengan menggunakan pattern matching.

Pattern matching dipakai karena tepat guna dan benar benar sesuai dengan kebutuhan. Dan akan sangat mudah

karena pattern matching sebenarnya berupa penerapan baru pada algoritma awal brute force. Pattern matching yang saya terapkan di sini adalah Knuth-Morris-Pratt dan Boyer Moore algorithm. Saya menggunakan dua, karena saya ingin agar kelak knowledge yang diretrieve dapat dibandingkan hasilnya yang menggunakan KMP dan yang menggunakan BM.

KMP adalah pencarian pola menggunakan metode dari kiri ke kanan, yang mana akan mencari pola dalam string yang sesuai dengan pattern yang diinginkan jika tidak ditemukan maka string akan bergerak ke kanan. Yang menjadi perbedaan adalah karena pada KMP akan mencatat string sebelumnya pada suatu stack yang menyebabkannya menjadi berbeda dengan brute force.

Berbeda dengan algoritma Brute Force, informasi yang digunakan masih dipelihara untuk melakukan jumlah pergeseran. Algoritma menggunakan informasi tersebut untuk membuat pergeseran yang lebih jauh, tidak hanya satu karakter Perbandingan antara algoritma brute force dengan algoritma KMP ditunjukkan dengan perpindahan posisi pattern terhadap posisi teks. Dimana dalam hal pencocokan string, pattern yang dicocokkan berawal dari kiri teks.

Algoritma KMP melakukan proses awal terhadap pattern dengan menghitung fungsi pinggiran Dengan adanya fungsi pinggiran ini, maka dapat dicegah pergeseran yang tidak berguna, seperti yang terjadi pada algoritma brute force. Fungsi pinggiran hanya bergantung pada karakter yang ada di dalam pattern, tidak bergantung kepada karakter di dalam teks.

```

procedure HitungPinggiran(input m :
integer, P : string)
{ Menghitung nilai pinggiran b[1..m]
untuk pattern P[1..m] }

Kamus :
k, g : integer

Algoritma :
b[] = array[1..m] of integer
b[0] = 0
q = 1
k = 0

for q ← 1 to m-1 do
  while ((k > 0) and (P[q] ≠ P[k]))
  do
    k ← b[k]
  endwhile
  if P[q] = P[k] then
    k ← k+1
  endif
  b[q] ← k
endfor

```

Gambar 2. Fungsi Pinggiran dalam KMP

```

function KMPSearch(input m : integer,
n : integer, P : string, T string) →
boolean
( Mengembalikan true jika pattern P
ditemukan dalam teks T )

Kamus :
i, j : integer
ketemu : boolean
procedure HitungPinggiran(input m :
integer, P : string)
( Menghitung nilai pinggiran b[1..m]
untuk pattern P[1..m] )

Algoritma :
HitungPinggiran(m, P)
i ← 0
j ← 0
ketemu = false
while ((j > 0) and (P[j] ≠ T[i]))
do
j ← b[j-1]
endwhile
if P[j] = T[i] then
j ← j+1
endif
if j = m then
ketemu ← true
else
i ← i+1
endif
return ketemu

```

Gambar 3. Fungsi KMP

Boyer Moore adalah pencarian pola dengan menggunakan metode dari kanan ke kiri yang menyocokkan nilai dari query yang sudah di masukkan ke dalam suatu stack ke string yang ingin dicari nilainya.

Algoritma Boyer-Moore adalah salah satu algoritma untuk mencari suatu string di dalam teks, dibuat oleh R.M Boyer dan J.S Moore. Ide utama algoritma ini adalah mencari string dengan melakukan perbandingan karakter mulai dari karakter paling kanan dari string yang dicari. Dengan menggunakan algoritma ini, secara rata-rata proses pencarian akan menjadi lebih cepat jika dibandingkan dengan algoritma lainnya. alasan melakukan pencocokan dari kanan (posisi terakhir string yang dicari) ditunjukkan dalam contoh berikut :

k	a	n	a	n	k	i	r	i	o	k	e	r	a	d	i	o
r	a	d	i	o												

Pada contoh diatas, dengan melakukan perbandingan dari posisi paling akhir string dapat dilihat bahwa karakter “n” pada string “kanan” tidak cocok dengan karakter “o” pada string “radio” yang dicari, dan karakter “n” tidak pernah ada dalam string “radio” yang dicari sehingga string “radio” dapat digeser melewati string “kanan” sehingga posisinya menjadi :

k	a	n	a	n	k	i	r	i	o	k	e	r	a	d	i	o
					r	a	d	i	o							

Dalam contoh terlihat bahwa algoritma Boyer-Moore memiliki loncatan karakter yang besar sehingga mempercepat pencarian string karena dengan hanya memeriksa sedikit karakter, dapat langsung diketahui bahwa string yang dicari tidak ditemukan dan dapat digeser ke posisi berikutnya.

Jadi berdasarkan metode pencarian query seperti ini akan diharapkan hasil hasil yang sesuai ditambah dengan bobot yang sesuai dengan metode information retrieval.

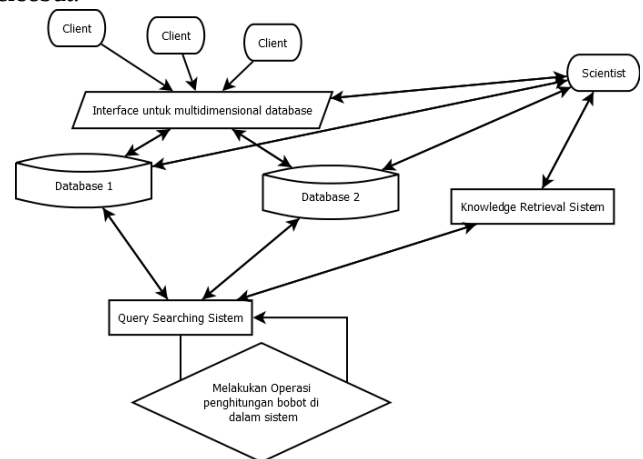
IV. IMPLEMENTASI PADA SISTEM SEBAGAI SOLUSI

Dari langkah langkah yang dilakukan sebelumnya perancangan implementasi harusnya memakai stopword , idf , tf untuk mencegah data yang diretrieve tidak tepat.

Karena system basis data ingin dapat diakses maka system basis data menggunakan multidimensional database yang mana dalam arsitektur system akan disertakan suatu antar muka yang dapat menjamin bahwa system akan dapat dilihat dari berbagai sisi sesuai dengan kriteria multidimensional database.

Karena system menggunakan data mining dan ingin memiliki respons yang cepat maka untuk metode algoritma KMP dan BM akan diimplementasikan pada tools lain di luar system. Sehingga system hanya akan terbebani oleh pencarian knowledge itu sendiri.

Sistem yang diinginkan juga ingin dapat handle user yang banyak maka untuk itu dalam pengambilan informasi kelak akan digunakan normalisasi yang akan memperbaiki hasil untuk hasil yang tidak sebanding jumlahnya dan berpotensi merusak bobot dari query tersebut.



Gambar 4. Sistem yang dirancang

Dari gambar 4 dapat dilihat bahwa ada interface untuk multidimensional database sehingga respon dan tipe data dapat semakin interaktif, hanya saja client hanya berguna untuk memberikan nilai pada database.

Scientist dapat melihat database dari 2 sisi, multidimensional database ataupun database biasa (relational). Scientist juga dapat memasukkan nilai berdasarkan kedua cara ini juga. Ketika scientist mengaktifkan pada knowledge retrieval system, maka knowledge retrieval system langsung meminta data kepada query searching system. Query searching system akan mengambil semua nilai yang sesuai dan akan mengesampingkan nilai yang tidak sesuai dengan normalisasi dan system pembobotan yang sudah diatur sebelumnya. Data yang ada lalu dikirim ke knowledge retrieval system yang kelak akan diproses dengan proses datamining untuk menghasilkan knowledge yang akan dikembalikan dalam tahap “Knowledge Presentation” yang akan dikembalikan pada user dalam bentuk grafik ataupun dalam bentuk knowledge itu sendiri.

V. KESIMPULAN

Dari makalah ini dapat ditarik kesimpulan.

- Penggunaan pattern matching dalam metode pengembalian data sangatlah penting karena merupakan tahap awal dalam metode itu sendiri.
- Sistem multidimensional database dapat menjadi lebih responsive dari system database biasa jika dapat ditemukan interface yang benar benar sesuai.
- Ketepatan pencarian data berdasarkan pattern matching sangatlah bergantung pada fungsi filtering data (pembobotan query).

VII. DAFTAR GAMBAR

Gambar 1. Contoh data cube

(sumber : www.technet.com waktu akses 20 Desember 2012 pukul 20.00 WIB)

Gambar 2. Fungsi Pinggiran dalam KMP

Gambar 3. Fungsi KMP

(sumber:

www.informatika.org/~rinaldi/Stmik/Makalah/MakalahStmik10.pdf waktu akses 20 Desember 2012 pukul 20.00 WIB)

Gambar 4. Sistem yang dirancang

REFRENSI

- [1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.
- [2] V. Ganti, J. Gehrke, R. Ramakrishnan. Mining very large databases. *COMPUTER*, 32(8):38-45, 1999.
- [3] Rila Mandala, *Improving Information Retrieval System Performance by Combining Text-Mining Techniques*, **International Journal of Intelligent Data Analysis**, Volume 4, Number 6, pp. 489-511, IOS Press, ISSN : 1088-467X,
- [4] Eko Gunocipto Hartoyo, Yus Gias Vembrina, and Anggia Ferdina Meilana, "Analisis Algoritma Pencarian String (String Matching)
- [5] E. Ide, New Experimen in Relevance Feedback, in G. Salton, editor, *The SMART Retrieval System*, pp. 313-323, Prentice Hall, Engelwood Cliffs, NJ, 1971.

PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Bandung, 20 Desember 2012



Elfino Sitompul 13508098