TUG SYNOPSIS UNTUK HIMPUNAN DATABASE RELASIONAL

[1]Hustinawaty
[2]Dini Sundani

[1]Universitas Gunadarma(hustina@staff.gunadarma.ac.id)

[2]Universitas Gunadarma(dinisundani@staff.gunadarma.ac.id)

Abstrak

Tupel Graph (TuG) Synopsis adalah suatu model yang digunakan untuk menghasilkan perkiraan yang akurat dari pemilihan query join yang kompleks pada database relasional. Tupel (record) pada database ralasional digambarkan sebagai sebuah graf, dan query join digambarkan sebagai transversal (perjalanan) graf. Model ini memiliki keunggulan dari teknik sebelumnya dengan mengembangkan suatu algoritma yang efisien untuk membangun TuG synopsis yang akurat pada kapasitas memori yang terbatas.
Kata Kunci : Tupel Graph (TuG) Synopsis; Query Join; Database; Relasional System.

1 PENDAHULUAN

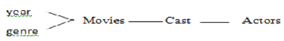
Dalam melakukan pengoptimalan relasi query tugas yang pertama dilakukan adalah mengoptimalkan query join yang kompleks. Untuk memperkirakan efektifitas biaya dari kandidat plan, pengoptimalan ini memerlukan keakuratan estimasi pada ukuran penyimpanan yang dihasilkan oleh operator yang berbeda atau ekivalen, dan memerlukan akurasi estimasi selektivitas untuk hubungan ekspresi query. Estimasi ini biasanya disediakan oleh data synopsis umumnya mengarah kepada data statistik, dengan memperkirakan distribusi data yang tepat dan estimasi yang dihasilkan oleh sebuah query Query join yang kompleks ditunjukkan oleh adanya relasi many-to-many pada suatu database. Sebagai contoh terdapat database sederhana yang terdiri dari empat tabel, movie, actor, casting dan genre. Satu actor dapat membintangi lebih dari satu movie, begitu sebaliknya satu movie dapat dibintangi oleh beberapa actor, sehingga memiliki relasi many-to-many. Dengan adanya relasi many-to-many maka akan memunculkan korelasi yang kompleks. Relational query yang kompleks akan mempengaruhi penggunaan storage yang akan berdampak pada besarnya ukuran penyimpanan[3]. Penelitian-penelitian yang dilakukan sebelumnnya dalam upaya pengoptimalan query dengan menggunakan tabel-level synopsis, seperti, histogram [2] atau wavelet [1], tidak efektif dalam meng-capture korelasi join yang kompleks karena hanya berfokus pada summarization dari satu tabe pada satu waktu.

2 DEFINISI DAN PEMBAHASAN

2.1 Model Data

Fungsi agregasi dari relational database didefinisikan sebagai relasi R = {R1, . . . , Rn}. Diasumsikan bahwa setiap relasi adalah Rj, 1 = j = n, yang mempengaruhi attributes Aj dan kumpulan disjoint dari attributes joint Jj. Attributes joint digunakan untuk mendefinisikan join relation antara relasi. dimana Aj untuk menotasikan kumpulan semua nilai atribut dari skema relasi. Informasi mengenai skema digambarkan dalam bentuk graph tidak berarah GS, yang disebut sebagai skema graph. Himpunan simpul dari GS adalah himpunan relasi dan atribut yang dinotasikan sebagai $R \cup A$. Graph terdiri dari sebuah ruas (R,A) untuk setiap R pada himpunan relasi R dan bergantung dengan nilai A pada himpunan atribut A. Dibawah ini merupakan Informasi database skema :

```
CREATE TABLE Movies (
 mid INTEGER PRIMARY KEY,  genre VARCHAR(40),
year INTEGER );
CREATE TABLE Actors (
 aid INTEGER PRIMARY KEY,  sex CHAR(1)
CHECK( sex in ('M', 'F')));
CREATE TABLE Cast (
 mid INTEGER REFERENCES Movies;  aid
INTEGER REFERENCES Actorss;  PRIMARY KEY (mid,aid);
```



Gambar 1: Skema Graph

Movies	Mid	Year	Genre
	1	2005	Action
	2	2004	Action
	3	2000	Drama

Cast	Mid	Aid
	1	1
	1	2
	2	3
	3	3
	3	4

Actors	Aid	Sex
	1	Male
	2	Female
	3	Male
	4	Male

Gambar 2: Contoh Sample

2.2 Model Query

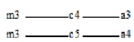
Model query difokuskan pada SQL query yang dapat melakukan perhitungan fungsi agregasi yangk ompleks. Diasumsikan bahwa relasi join adalah R1, . . . , RK . Sebuah query A querydapat dideskripsikan sebagai berikut :

```
SELECT Aggr FROM R1, R2,..RK
```

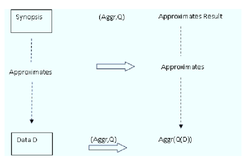
$$\text{WHERE} \bigwedge_{1\le i\le k} C_i \ \text{AND} \bigwedge_{1\le i\ne j\le k} R_i \bowtie R_j$$

Konsep dari graph query dari hasil pemilihan join ditunjukkan oleh SQL Specification berikut

```
SELECT COUNT(*) FROM Movies m,
Cast c, Actors a  WHERE m.year=2000 AND
m.genre='drama' AND a.sex='male'
AND m.mid = c.mid AND c.aid = a.aid
```



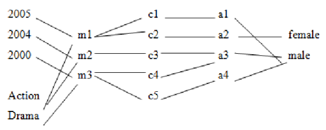
Gambar 5: Hubungan yang mungkin terjadi pada rata graph dari gambar 1



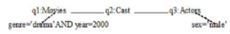
Gambar 6: Perkiraan jawaban query

Tug Sinopsys Sebuah Tug synopsis dari graph GD adalah sebuah graph TG sedemikian rupa sehingga:

- simpul dari TG berkoresponden ke partisi dari partisi TS
- masing-masing simpul r dilabeli dengan relasi R yang sesuai dan berhubungan dengan counter tcount (r) = $|r|$
- untuk setiap atribut A dari R, simpul berasosiasi dengan sebuah nilai vsum (r,A) yang berisi nilai (r, A)



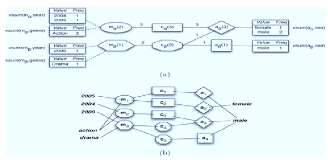
Gambar 3: Hubungan data graph



Gambar 4: Graph query Q

- TG terdiri dari sebuah ruas(r, s) untuk setiap pasangan dari partisi joini r dan s
- setiap ruas beraosiasi dengan sebuah counter jcount (r,s) = $|r, s|$

Gambar 7a memperlihatkan contoh Tug dan simpul untuk data graph daro gambar 7b Simpul $m\alpha$ merepresentasikan dua tupel movie dan simpul ca merepresentasikan tiga tupel Cast. Join dari keduanya dinotasikan oleh mereka dilambangkan oleh ruas yang terhubung dan dinotasikan sebagai jcount $(m\alpha, c\alpha$ = 3 hasil.



Gambar 7: Tug untuk data graph dan simpul

TuG sebagai dasar dalam pendekatan hasil query

- Masalahnya dapat dirumuskan sebagai berikut: Diberi synopsis Tug TG dari database D dan sebuah query (Aggr, Q), hitunglah sebuah pendekatan Aggr (Q (D)) dengan menggunakan informasi yang disimpan di T G.
- Di notasikan Q (T G) sebagai himpunan penyisipan Q pada TG.f Diberikan sebuah penyisipan e dalam Q (TG), fungsi Aggr (e) didefinisikan sebagai perkiraan dari fungsi agregasi dalam himpunan yang sama yang dinotasikan sebagai e. Sehingga diperoleh perkiraan Aggr (Q (D)) dengan mengakumulasi perkiraan setiap penyisipan Aggr (e).
- Sebagai contoh, perkiraan COUNT (Q (D)) dapat diturunkan sebagai

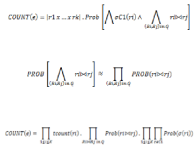
$$\sum_{e\in Q(TG)} COUNT(e)$$

- Dua penyisipan e1 dan e2 yang didefinikan sebagai berikut :

  e1(q1) = mb , e1(q2) = cb , e1(q3) = aa, and e2(q1) = mb , e2(q2) = cb , e2(q3) = ab.

Total count dapat dihitung sebagai COUNT(Q(T G)) = COUNT(e1) + COUNT(e2).

- Dengan demikian, masalah dari perkiraan jawaban TuG dapat dibagi menjadi dua sub-masalah, yaitu : (a) pendekatan hasil daril Aggr (e) dan (b) secara efisien menggabungkan perkiraan COUNT (e) pada semua penyisipan di Q (T G).
  - Perkiraan jawaban untuk penyisipan tunggal



Atau dinotasikan dengan :

$count(\varrho 1 = tcount(m\beta)tcount(a\alpha)Prob(\sigma sex = male(a\alpha))$

$Prob(m\beta c\beta)Prob(c\beta a\alpha Prob(\sigma year = 2000(m\beta))Prog(\sigma genre = drama(m\beta)))$

Contoh : Didefinisikan $\varrho 1(q1) = m\beta, \varrho 1(q2) = c\beta, \varrho 1(q3) = a\alpha$. Estimasi dari Count dapat dihitung sebagai berikut :

$Prob(m\beta \bowtie c\beta) = \frac{jcount(m\beta,c\beta)}{tcount(m\beta)tcount(c\beta)} = 1$

$Prob\alpha year = 2000(m\beta)) \bowtie \sigma) = \frac{jcount(c\beta,\sigma\beta)}{tcount(c\beta)tcount(\sigma\beta)} = \frac{1}{6}$

$Prob\alpha year = 2000(m\beta)) = \frac{1}{1}$

$Prob\alpha genre = Drama(m\beta)) = \frac{1}{1}$

$Prob\alpha sex = male(a\alpha)) = \frac{2}{3}$

Estimasi akhir $COUNT(\varrho 1)$ adalah $= \frac{2}{3}$

  - Kombinasi Estimasi untuk Semua Penyisipan

$COUNT(\varrho 1) + COUNT(\varrho 2)$

$= tcount(m\beta)Prob(\sigma year = 2000(m\beta))Prob(\sigma genre = Drama(m\beta))Prob(m\beta \bowtie c\beta)$

$(tcount(c\beta)tcount(\sigma\alpha)Prob(c\beta \bowtie \sigma\alpha)Prob(\sigma sex = male(\sigma\alpha))$

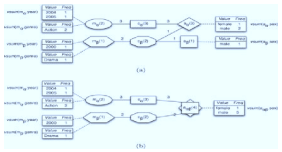
$+tcount(c\beta)tcount(\sigma\beta)Prob(c\beta \bowtie \sigma\alpha)Prob(\sigma sex = male(\sigma\beta)))$

$= tcount(m\beta)Prob(\sigma year = 2000(m\beta))Prob(\sigma genre = Drama(m\beta))Prob(m\beta \bowtie c\beta)COUNT(\varrho') + COUNT(\varrho''))$

$count(r, i) = \Pi_{cinCkj}Prob(\sigma c(f)).tcount(r).(\Pi_{J(i)(r,s)}\Sigma_{intg}Prob(r \bowtie s).count(s, j))$

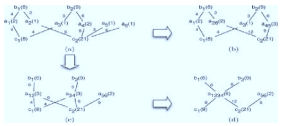
Operasi Dasar Kompresi TuG Synopsis terdiri dari operasi node merge dan lossless merge

- Operasi Node-merge, digunakan untuk mengurangi penyimpanan TuG dari kelebihan beberapa simpul menjadi satu simbul baru



Gambar 8: Node merge

- Operasi Lossless Merge, yaitu operasi-operasi yang tidak mempengaruhi keakuratan hasil synopsis. Operasi ini penting untuk efisiensi algoritma konstruksi Tug, karena memungkinkan kompresi yang cepat dari keakuratan sebuah sinopsi yang besar menjadi lebih kecil



Gambar 9: Lossless Merge

2.3 Konstruksi TuG

Digunakan untuk membangun TuG synopsis yang akurat dalam media penyimpanan yang spesifik, beroperasi di empat tahap berikut.

1. mencari sebuah atribut numerik dengan kisaran yang tepat sebelum proses kompressing untuk mengurangi kompleksitas nilai domain dan distribusi nilai.
2. menginisialisasi sebuah synopsis dengan sebuah partisi (satu tupel per partisi) dan merangkum nilai-nilai yang sama dan melanjutkan untuk operasi merge. Hasil akhir adalah membangkitkan partisi rendah yang menjaga keakuratan synopsis awal
3. mengkompres partisi yang lebih besar dengan menggunakan operasi merge, sehinggan ukuran dari kompresi tidak mengorbarbankan keakuratan

4. mensubstitusi detil nilai ringkasan dengan dilakukan kompresi dimensi histogram single. Selama proses ini TuG bergantung pada sebuah struktur penyimpanan yang efisien dan skala algoritma ms untuk menangani dataset yang besar dengan memori yang terbatas.

3 HASIL EKSPERIMEN

Tabel 1: hasil estimasi error dari relasi many to one dan many to many

	Percentiles				
Methode	**0%**	**25%**	**50%**	**75%**	**100%**
TuG	0	0.1	8.1	104.9	238.766
Histogram	2.0	149.8	14.412	120.491	5.866.480
Join Synop	0	0	0	0	0

Tabel 2: Menampilkan estimasi error dari tiga teknik untuk worksload query many to one

	Percentiles - Full IMDB				
Methode	**0%**	**25%**	**50%**	**75%**	**100%**
TuG	0	0.1	0.8	5.8	382.2
Histogram	3	258.2	1081.4	4583.1	20218.2

Tabel 3: error dari tiga teknik untuk worksload query many to many

	Percentiles - Scale Down IMDB				
Methode	**0%**	**25%**	**50%**	**75%**	**100%**
TuG	0	0.1	0.3	1.1	12.7
Wareless	0	0.1	0.4	2.0	23.7

4 Kesimpulan

1. TuG menggunakan ringkasan dalam bentuk graph untuk menggambarkan relasi many to many

2. TuG mampu membangun algoritma yang dapat menangani hubungan yang kompleks yang akan melibatkan aplikasi untuk database

3. informasi yang dihasilkan dalam synopsis TuG adalah kompatibel, yang mengarah pada perkiraan hasil yang akurat dibandingkan dengan teknik yang sudah ada sebelumnya.

4. Untuk selanjutnya TuG dapat dibangun untuk model clustering dengan perhitungan yang kompleks.

Pustaka

[1] Yannis E.Ioanidis and Viswanath Poosala. Histogram-based approximate of set-valued query answer. In *Proceeding of the 26th VLDB Conference, Edinburgh, Scotland*, pages 679–698. IEEE Trans Pattern Analysis and Machine Intelligence, November 1999.

[2] Rajeev Rastogi Kaushik Chakrabarti, Minos Garafalakis and Kyuseok Shim. Approximate query processing using wavelet. In *Proceeding of the 26th VLDB Conference, Cairo, Egypt*. IEEE Transactions on Systems, Man, and Cybernetics, 2000.

[3] Joshua Spiegel and Neoklis Polyzotis. Tuple graph synopsises for relational data sets. In *Proceedings of ACM Sigmod International Conference on Management of Data*, volume 7, pages 195–206. IEEE Transactions, September 2006. Visualization and Computer Graphics.