Chương II

Các hệ cơ sở dữ liệu phân tán

Nguyễn Kim Anh
anhnk-fit@mail.hut.edu.vn

Bộ môn Hệ thống Thông tin, SoICT

1

Nội dung

- Tổng quan về các hệ CSDLPT
- Phân đoạn dữ liệu
- Biểu diễn các yêu cầu với các mức trong suốt khác nhau
- Thiết kế CSDLPT
- **Xử lý và tối ưu hóa truy vấn phân tán**
- Quản trị giao dịch và điều khiển tương tranh

2

NỘI DUNG

2

MỞ ĐẦU

- Vấn đề tối ưu hoá trên hệ CSDL phân tán là rất quan trọng do tính phân mảnh, nhân bản, tốn kém chi phí cho việc truyền dữ liệu.
- Thuật toán tối ưu truy vấn phân tán cổ điển là vét cạn và leo đồi:
  - Thuật toán vét cạn không phù hợp với sự bùng nổ dữ liệu.
  - Thuật toán leo đồi chỉ tìm kiếm được tối ưu cục bộ.
- Để khắc phục, các giải thuật tìm kiếm ngẫu nhiên và Heuristic được đề xuất có thể tìm ra các giải pháp gần tối ưu chấp nhận được.

3

I. TỔNG QUAN VỀ XỬ LÝ TRUY VẤN PHÂN TÁN

BÀI TOÁN XỬ LÝ TRUY VẤN PHÂN TÁN

Xét một CSDL mẫu mô hình hoá cho một công ty máy tính.
Các thuộc tính của CSDL bao gồm:

- ENO: mã số nhân viên
- ENAME: tên nhân viên
- TITLE: chức vụ trong công ty
- SALE: mức lương
- RESP: nhiệm vụ trong dự án
- DUR: thời gian được phân công trong dự án
- PNO: mã số dự án
- PNAME: tên dự án
- BUDGET: ngân sách dự án

CÁC QUAN HỆ ĐÃ CHUẨN HOÁ

EMP

ENO	ENAME	TITLE
E1	J. Doe	Elect. Eng.
E2	M. Smith	Syst. Anal.
E3	A. Lee	Mech. Eng.
E4	J. Miller	Programmer
E5	B. Casey	Syst. Anal.
E6	L. Chu	Elect. Eng.
E7	R. David	Mech. Eng.
E8	J. Jones	Syst. Anal.

ASG

ENO	PNO	RESP	DUR
E1	P1	Manager	12
E2	P1	Analyst	24
E2	P2	Analyst	6
E3	P3	Consultant	10
E3	P4	Engineer	48
E4	P2	Programmer	18
E5	P2	Manager	24
E6	P4	Manager	48
E7	P3	Engineer	36
E8	P3	Manager	40

PROJ

PNO	PNAME	BUDGET
P1	Instrumentation	150000
P2	Database Develop	135000
P3	CAD/CAM	250000
P4	Maintenance	310000

PAY

TITLE	SAL
Elect. Eng.	40000
Syst. Anal.	34000
Mech. Eng.	27000
Programmer	24000

CHỌN LỰA CHIẾN LƯỢC

SELECT ENAME
FROM EMP, ASG
WHERE EMP.ENO = ASG.ENO
AND DUR > 37

Chiến lược 1

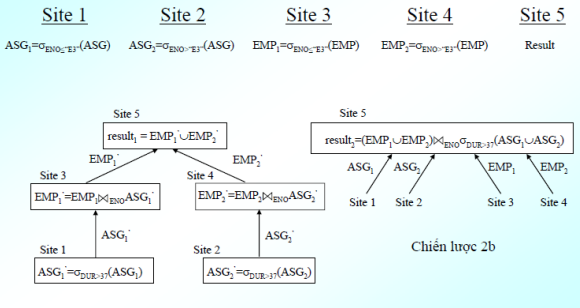
$$\pi_{ENAME}(\sigma_{DUR>37 \wedge EMP.ENO=ASG.ENO}(EMP \times ASG))$$

Chiến lược 2

$$\pi_{ENAME}(EMP \bowtie_{ENO} (\sigma_{DUR>37}(ASG)))$$

Chiến lược 2 tránh được việc sử dụng tích Descartes nên "tốt hơn"

CÁC CHIẾN LƯỢC THỰC THI TRUY VẤN TƯƠNG ĐƯƠNG



Chiến lược 2a

Chiến lược 2b

CHI PHÍ CỦA CÁC CHIẾN LƯỢC

Giả sử

- size(EMP) = 400, size(ASG) = 1000
- Chi phí truy xuất 1 bộ TA = 1; Chi phí truyền 1 bộ TT = 10

Chiến lược 2a

1.	Tạo ASG' bằng cách chọn trên ASG: (10 + 10) * TA	= 20
2.	Truyền ASG' đến các vị trí (site) của EMP: (10 + 10) * TT	= 200
3.	Tạo EMP' bằng cách nối ASG' và EMP': (10 + 10) * TA *.2	= 40
4.	Truyền EMP' đến vị trí (site) nhận kết quả: (10 + 10) * TT	= 200
	Tổng chi phí	460

Chiến lược 2b

1.	Truyền EMP đến vị trí (site) 5: 400 * TT	= 4,000
2.	Truyền ASG đến vị trí (site) 5: 1000 * TT	= 10,000
3.	Tạo ASG' bằng cách chọn trên ASG: 1000 * TA	= 1,000
4.	Nối EMP và ASG': 400 * 20 * TA	= 8,000
	Tổng chi phí	23,000

8

MỤC TIÊU CỦA TỐI ƯU TRUY VẤN

- Cực tiểu hàm chi phí truy vấn

  Chi phí I/O + chi phí CPU + chi phí truyền thông

Có những khác biệt về trọng số trong những môi trường phân tán khác nhau.

- Mạng diện rộng
  - Chi phí truyền thông có ảnh hưởng lớn
    - Dải thông thấp
    - Tốc độ thấp
  - Hầu hết các giải thuật đều bỏ qua các thành phần chi phí khác trong quá trình xử lý cục bộ
- Mạng cục bộ
  - Chi phí truyền thông không có ảnh hưởng lớn
  - Hàm chi phí tổng cộng cần phải được xem xét

9

ĐỘ PHỨC TẠP CỦA CÁC PHÉP TOÁN QUAN HỆ

Giả sử số quan hệ của lực lượng là n

Phép toán	Độ phức tạp
Chọn Chiếu (không loại bỏ trùng lặp)	O(n)
Chiếu (có loại bỏ trùng lặp) Gộp nhóm	O(n*logn)
Nối Nối nửa Chia Các phép toán tập hợp	O(n*logn)
Tích Descartes	$O(n^2)$

10

CÁC KIỂU TỐI ƯU

- Giải thuật tìm kiếm vét cạn
  - Dựa trên chi phí
  - Tối ưu
  - Tổ hợp phức tạp trong một số quan hệ
- Giải thuật Heuristics
  - Không tối ưu
  - Nhóm lại các biểu thức con chung
  - Thực hiện các phép chọn và chiếu trước
  - Thay thế các nối bằng một tổ hợp các nối nửa
  - Sắp xếp lại thứ tự thực hiện các phép toán để giảm các quan hệ trung gian
  - Tối ưu các phép toán riêng lẻ

11

THỜI ĐIỂM TỐI ƯU

- Tĩnh
  - Biên dịch → tối ưu hoá trước khi thực hiện truy vấn
  - Khó khăn trong việc ước lượng kích thước của các kết quả trung gian → lỗi truyền
  - Có thể truyền lại cho nhiều thực thi khác
  - R*
- Động
  - Tối ưu hoá trong khi thực hiện truy vấn
  - Thông tin chính xác về kích thước của các quan hệ trung gian
  - Phải tối ưu lại với các thực thi bội nên tốn nhiều chi phí
  - INGRES
- Lai (hỗn hợp)
  - Biên dịch sử dụng một giải thuật tĩnh
  - Nếu xảy ra lỗi do việc ước lượng kích thước > ngưỡng, phải tối ưu hoá lại lúc chạy chương trình

12

VỊ TRÍ QUYẾT ĐỊNH

- Tập trung
  - Chỉ một vị trí xác định chiến lược tốt nhất
  - Đơn giản
  - Cần tri thức về CSDL phân tán toàn vẹn
- Phân tán
  - Nhiều vị trí tham gia vào quá trình chọn ra chiến lược tốt nhất
  - Chỉ cần thông tin về vị trí
- Lai (Hỗn hợp)
  - Một ví trí xác định các quyết định chính
  - Các vị trí khác đưa ra chọn lựa cục bộ
  - System R*

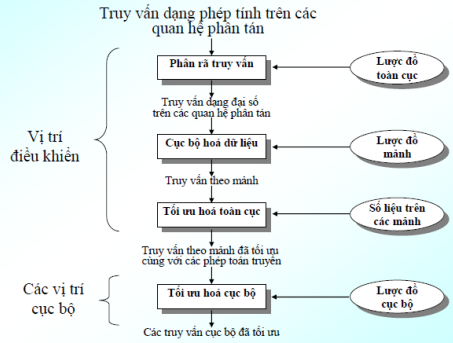
13

CẤU HÌNH MẠNG

- Mạng diện rộng (WAN)
  - Đặc điểm
    - Dải thông thấp
    - Tốc độ thấp
  - Chi phí truyền thông chiếm ưu thế, có thể bỏ qua các nhân tố chi phí khác
  - Sắp xếp toàn thể để tối ưu hoá chi phí truyền thông
  - Sắp xếp cục bộ kéo theo tối ưu truy vấn tập trung
- Mạng cục bộ (LAN)
  - Chi phí truyền thông không đáng kể
  - Hàm tổng chi phí phải được xem xét

14

CÁC TẦNG XỬ LÝ TRUY VẤN PHÂN TÁN



15

II. XỬ LÝ TRUY VẤN PHÂN TÁN

1. PHÂN RÃ TRUY VẤN

- Chuẩn hoá
  - Biến đổi câu truy vấn thành dạng chuẩn để xử lý tiếp
- Phân tích
  - Tìm và loại bỏ các truy vấn không đúng hoặc không cần thiết
- Loại bỏ dư thừa
  - Loại các vị từ thừa
- Viết lại câu truy vấn
  - Truy vấn phép toán quan hệ □ truy vấn đại số quan hệ
  - Cấu trúc lại câu truy vấn đại số
  - Sử dụng các quy tắc biến đổi

16

CHUẨN HOÁ DỮ LIỆU

- Phân tích cú pháp và từ vựng
  - Kiểm tra tính hợp lệ (tương tự bộ biên dịch)
  - Kiểm tra các thuộc tính và quan hệ
- Đưa vào dạng chuẩn
  - Dạng chuẩn hội
    $(p_{11} \vee p_{12} \vee \ldots \vee p_{1n}) \wedge \ldots \wedge (p_{m1} \vee p_{m2} \vee \ldots \vee p_{mn})$
  - Dạng chuẩn tuyển
    $(p_{11} \wedge p_{12} \wedge \ldots \wedge p_{1n}) \vee \ldots \vee (p_{m1} \wedge p_{m2} \wedge \ldots \wedge p_{mn})$
  - AND ($\wedge$) được ánh xạ vào phép nối và chọn

17

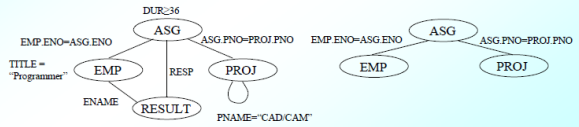
PHÂN TÍCH

- Loại bỏ những câu truy vấn sai
- Sai kiểu
  - Nếu có bất kỳ thuộc tính hoặc tên quan hệ trong câu truy vấn chưa được khai báo trong lược đồ toàn cục.
  - Nếu các phép toán áp dụng cho các thuộc tính có kiểu không thích hợp
- Sai nghĩa
  - Để tìm kiếm câu truy vấn sai, sử dụng:
    - Đồ thị truy vấn
    - Đồ thị kết nối

18

PHÂN TÍCH – VÍ DỤ

```
SELECT   ENAME,RESP
FROM     EMP, ASG, PROJ
WHERE    EMP.ENO = ASG.ENO
AND      ASG.PNO = PROJ.PNO
AND      PNAME = "CAD/CAM"
AND      DUR ≥ 36
AND      TITLE = "Programmer"
```



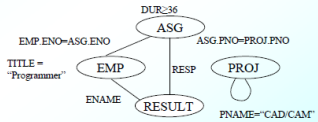
Đồ thị truy vấn **Đồ thị kết nối**

19

PHÂN TÍCH

Nếu đồ thị truy vấn không liên thông, truy vấn bị sai.

SELECT ENAME,RESP
FROM EMP, ASG, PROJ
WHERE EMP.ENO = ASG.ENO
AND PNAME = "CAD/CAM"
AND DUR ≥ 36
AND TITLE = "Programmer"



20

LOẠI BỎ DƯ THỪA – VÍ DỤ

SELECT TITLE
FROM EMP
WHERE EMP.ENAME = "J. Doe"
OR (**NOT**(EMP.TITLE = "Programmer")
AND (EMP.TITLE = "Programmer"
OR EMP.TITLE = "Elect. Eng.")
AND **NOT**(EMP.TITLE = "Elect. Eng."))
⇩
SELECT TITLE
FROM EMP
WHERE EMP.ENAME = "J. Doe"

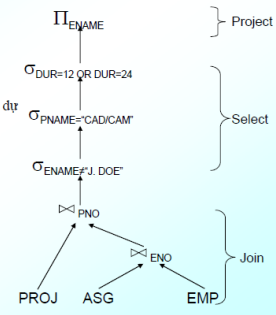
21

VIẾT LẠI CÂU TRUY VẤN

- Biến đổi câu truy vấn từ phép tính quan hệ thành đại số quan hệ
- Tạo cây truy vấn
- Ví dụ

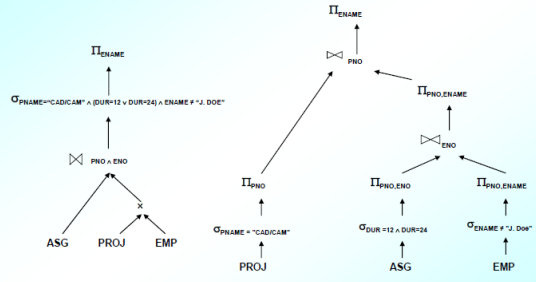
"Tìm tên các nhân viên trừ J. Doe đã làm cho dự án CAD/CAM trong một hoặc hai năm"

SELECT ENAME
FROM EMP, ASG, PROJ
WHERE EMP.ENO = ASG.ENO
AND ASG.PNO = PROJ.PNO
AND ENAME ≠ "J. Doe"
AND PNAME = "CAD/CAM"
AND (DUR = 12 **OR** DUR = 24)



22

VIẾT LẠI CÂU TRUY VẤN



Cây toán tử tương đương

Cây toán tử đã được viết lại

23

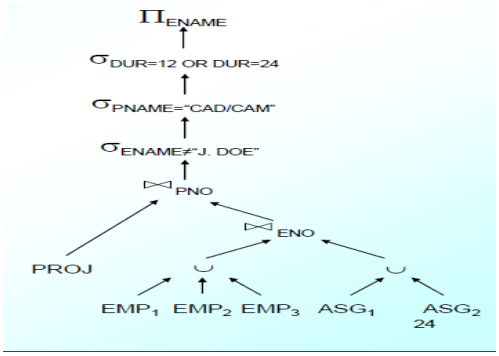
CỤC BỘ HOÁ DỮ LIỆU PHÂN TÁN

Giả sử
–EMP được tách thành ba mảnh ngang EMP_1, EMP_2, EMP_3 như sau:
- $EMP_1 = \sigma_{ENO \leq "E3"}(EMP)$
- $EMP_2 = \sigma_{"E3" < ENO \leq "E6"}(EMP)$
- $EMP_3 = \sigma_{ENO > "E6"}(EMP)$

–ASG được tách thành hai mảnh ngang ASG_1 and ASG_2 như sau:
- $ASG_1 = \sigma_{ENO \leq "E3"}(ASG)$
- $ASG_2 = \sigma_{ENO > "E3"}(ASG)$

Thay thế
EMP bằng $(EMP_1 \cup EMP_2 \cup EMP_3)$ và
ASG bằng $(ASG_1 \cup ASG_2)$ trong câu truy vấn

CỤC BỘ HOÁ DỮ LIỆU PHÂN TÁN



RÚT GỌN CHO PHÂN MẢNH NGANG

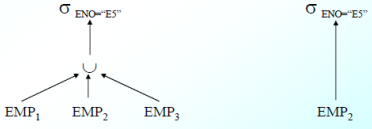
Rút gọn với phép chọn

Cho quan hệ R và $F_R = \{R_1, R_2, \ldots, R_w\}$ trong đó $R_j = \sigma_{pj}(R)$

$\sigma_{pi}(R_j) = \phi$ nếu $\forall x$ thuộc R: $\neg(pi(x) \wedge pj(x))$

Ví dụ:

```
SELECT   *
FROM     EMP
WHERE    ENO="E5"
```



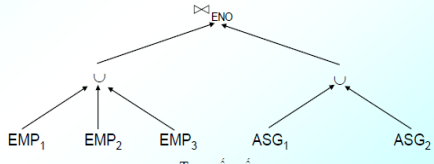
25

RÚT GỌN CHO PHÂN MẢNH NGANG

Rút gọn với nối

Xem câu truy vấn

```
SELECT   *
FROM     EMP, ASG
WHERE    EMP.ENO=ASG.ENO
```

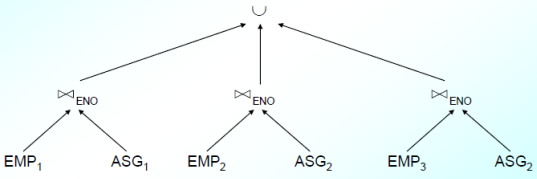


Truy vấn gốc

26

RÚT GỌN CHO PHÂN MẢNH NGANG

- Phân phối nối trên hợp
- Áp dụng quy tắc rút gọn



Truy vấn đã rút gọn

27

RÚT GỌN CHO PHÂN MẢNH DỌC

Tìm các quan hệ trung gian vô dụng (không rỗng)

Quan hệ R được định nghĩa trên các thuộc tính A = {A_1, ..., A_n} và được phân mảnh dọc thành $R_i = \Pi_{A'}(R)$ trong đó A' ⊆ A:

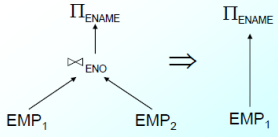
$\Pi_{D,K}(R_i)$ là vô dụng nếu tập các thuộc tính chiếu D không trong A'

Ví dụ:

$EMP_1 = \Pi_{ENO,ENAME}(EMP)$

$EMP_2 = \Pi_{ENO,TITLE}(EMP)$

SELECT ENAME
FROM EMP



28

RÚT GỌN CHO PHÂN MẢNH NGANG SUY DIỄN

Quy tắc

- Phân phối các nối trên hợp
- Áp dụng việc loại bỏ các nối trên phân mảnh ngang

Ví dụ

$ASG_1 = ASG \ltimes ENO\ EMP_1$

$ASG_2 = ASG \ltimes ENO\ EMP_2$

$EMP_1 = \sigma\ TITLE="Programmer"\ (EMP)$

$EMP_2 = \sigma\ TITLE\neq"Programmer"\ (EMP)$

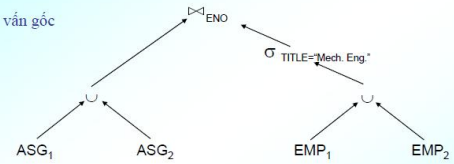
Truy vấn

SELECT *
FROM EMP, ASG
WHERE ASG.ENO = EMP.ENO
AND EMP.TITLE = "Mech. Eng."

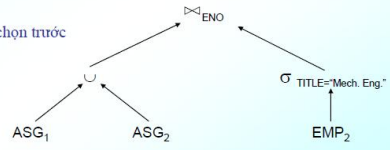
29

RÚT GỌN CHO PHÂN MẢNH NGANG SUY DIỄN

Truy vấn gốc

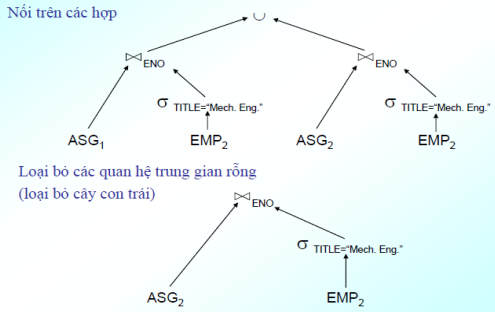


Phép chọn trước



30

RÚT GỌN CHO PHÂN MẢNH NGANG SUY DIỄN



31

RÚT GỌN CHO PHÂN MẢNH HỖN HỢP

Kết hợp các quy tắc đã có:

- Loại bỏ các quan hệ rỗng được tạo ra bởi các phép chọn mâu thuẫn trên các mảnh ngang;
- Loại bỏ các quan hệ vô dụng được tạo ra từ các phép chiếu trên các mảnh dọc;
- Phân phối các nối cho các hợp nhằm cô lập và loại bỏ các nối vô dụng.

32

RÚT GỌN CHO PHÂN MẢNH HỖN HỢP

Ví dụ

– Giả sử có phân mảnh hỗn hợp sau:

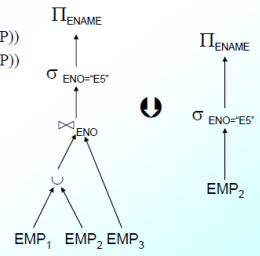
EMP1= $\sigma_{ENO\leq"E4"}(\Pi_{ENO,ENAME}(EMP))$

EMP2= $\sigma_{ENO>"E4"}(\Pi_{ENO,ENAME}(EMP))$

EMP3= $\Pi_{ENO,TITLE}(EMP)$

– Và câu truy vấn

SELECT ENAME
FROM EMP
WHERE ENO="E5"



33

III. TỐI ƯU TRUY VẤN PHÂN TÁN

Input: Truy vấn phân mảnh

Tìm kế hoạch tổng quát tốt nhất (không nhất thiết phải tối ưu)

- Cực tiểu hoá hàm chi phí
- Xử lý các nối phân tán, sử dụng nối nửa.
- Phương pháp nối: Lặp lồng và thứ tự nối (nối trộn và nối băm)

Không gian tìm kiếm

- Tập của các biểu thức đại số tương đương (các cây truy vấn).

Hàm chi phí (trong quan hệ thời gian)

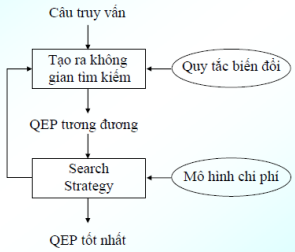
- Chi phí I/O + chi phí CPU + chi phí truyền

Thuật toán tìm kiếm

- Di chuyển bên trong không gian tìm kiếm
- Các thuật giải heuristic (lặp cải tiến, mô phỏng luyện thép, di truyền,…)

34

TỐI ƯU HOÁ TRUY VẤN



35

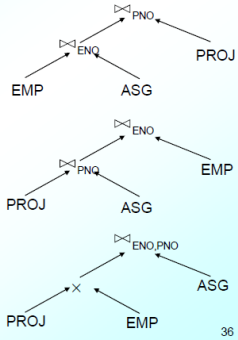
KHÔNG GIAN TÌM KIẾM

Không gian tìm kiếm đặc trưng bởi việc chọn lựa hoạch định thực thi

Trọng tâm trên cây nối

Với N quan hệ, có O(N!) cây nối tương đương có thể thu được bằng cách áp dụng các quy tắc giao hoán và kết hợp

SELECT ENAME,RESP
FROM EMP, ASG, PROJ
WHERE EMP.ENO=ASG.ENO
AND ASG.PNO=PROJ.PNO



36

CHIẾN LƯỢC TÌM KIẾM

Cách di chuyển trong không gian tìm kiếm

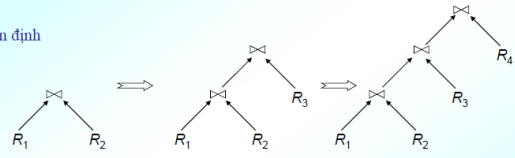
Đơn định
- Bắt đầu từ các quan hệ cơ sở và xây dựng các hoạch định bằng cách thêm vào một quan hệ ở mỗi bước
- Quy hoạch động: theo chiều ngang
- Thuật toán thiển cận: theo chiều sâu

Ngẫu nhiên
- Tìm kiếm lời giải tối ưu xung quanh một số điểm đặc biệt
- Đánh đổi thời gian tối ưu với thời gian thực thi
- Tốt nhất khi số quan hệ lớn hơn 5-6
- Mô phỏng luyện thép (Simulated Annealing)
- Lặp cải tiến (Iterative improvement)

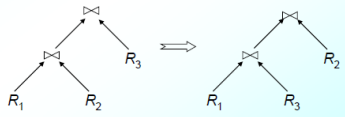
37

CHIẾN LƯỢC TÌM KIẾM

Đơn định



Ngẫu nhiên



38

MÔ HÌNH CHI PHÍ PHÂN TÁN

- Tổng thời gian (hay Tổng chi phí): Tổng các thành phần chi phí
- Thời gian đáp ứng: Thời gian tính từ khi khởi hoạt cho đến khi hoàn thành câu truy vấn
- Mạng diện rộng WAN
  - Khởi tạo và truyền thông điệp với chi phí cao
  - Xử lý cục bộ có chi phí thấp (với mainframes hoặc minicomputers)
  - Tỷ lệ truyền và thời gian xuất nhập có chi phí = 20:1
- Mạng cục bộ LAN
  - Phải xét cả chi phí cục bộ lẫn chi phí truyền
  - Tỷ lệ = 1:1.6

39

CÁC THUẬT TOÁN TỐI ƯU TRUY VẤN PHÂN TÁN

Ba thuật toán cơ bản đại diện cho nhiều lớp thuật toán khác nhau là:

- Ingres phân tán
- System R*
- SDD-1

THUẬT TOÁN INGRES PHÂN TÁN

Nguyên liệu: MRQ (Multi Relation Query, truy vấn đa quan hệ)

Thành phẩm: kết quả cuối cùng của truy vấn đa quan hệ

Begin

{thực hiện mọi truy vấn một quan hệ}

For mỗi ORQ_i khả tách trong MRQ **do**

Run(ORQi)

End For

{thay MRQ bằng một danh sách có n truy vấn đã tối giản (bất khả tách)}

MRQ'_list ← REDUCE(MRQ)

41

THUẬT TOÁN INGRES PHÂN TÁN (tt)

While n <> 0 **do**

Begin

{chọn một truy vấn bất khả tách có chứa các mảnh nhỏ nhất}

MRQ' ← SELECT_QUERY(MRQ'_list)

{xác định các mảnh cần truyền và vị trí xử lý cho MRQ'}

Fragment_site_list ← SELECT_STRATEGY(MRQ')

For mỗi cặp (F, S) trong Fragment_site_list **do**

di_chuyển_mảnh_F_đến_vị_trí_S

End For

run(MRQ')

dec(n)

End While {MRQ' cuối cùng là thành phẩm}

End.

42

THUẬT TOÁN SYSTEM R*

Nguyên liệu: QT (Query Tree, cây truy vấn)

Thành phẩm: strat (chiến lược có chi phí nhỏ nhất)

Begin

For mỗi quan hệ $R_i \in$ QT **do**

Begin

For mỗi đường truy xuất Ap_{ij} đến R_i **do**

xác định cost(Ap_{ij});

End For;

best_$Ap_i \leftarrow Ap_{ij}$ có chi phí nhỏ nhất

End;

43

THUẬT TOÁN SYSTEM R* (tt)

For mỗi thứ tự ($R_{i1}, R_{i2}, \ldots, R_{in}$) với i = 1, …, n! **do**

Begin

xây dựng chiến lược $(\ldots((best\_Ap_{i1} \bowtie R_{i2}) \bowtie R_{i3}) \bowtie \ldots \bowtie R_{in})$

tính chi phí của chiến lược

End For

strat ← chiến lược có chi phí nhỏ nhất

For mỗi vị trí k có lưu quan hệ có mặt trong QT **do**

Begin

$LS_k \leftarrow$ chiến lược cục bộ (chiến lược, k)

{mỗi chiến lược cục bộ được tối ưu hoá tại vị trí k}

send(LS_k, vị trí k)

End For;

End.

44

THUẬT TOÁN SDD-1

Nguyên liệu: QG: đồ thị truy vấn có n quan hệ;

số liệu thống kê cho mỗi quan hệ;

Thành phẩm: ES: chiến lược thực thi truy vấn

Begin

ES ← thao tác cục bộ (QG)

sửa lại số liệu thống kê để phản ánh tác dụng của xử lý cục bộ

BS ← ∅ {tập các nối nửa lợi ích}

For mỗi nối nửa SJ trong QG **do**

If cost(SJ) < benefit(SJ) **then**

BS ← BS ∪ SJ

End If

End For

While BS <> ∅ do {chọn các nối nửa lợi ích}

45

THUẬT TOÁN SDD-1 (tt)

Begin

SJ ← most_benefit(BS)

BS ← BS – SJ

ES ← ES + SJ

sửa lại số liệu thống kê để phản ánh tác dụng của việc gắn SJ

BS ← BS – các nối nửa không lợi ích

BS ← BS ∪ các nối nửa lợi ích mới

End While;

AS(ES) ← chọn vị trí i sao cho i có chứa dữ liệu lớn nhất

sau khi thực hiện tất cả mọi thao tác cục bộ

ES ← ES ∪ truyền các quan hệ trung gian đến AS(ES)

For mỗi quan hệ R_j tại AS(ES) **do**

For mỗi nối nửa SJ của R_i với R_j **do**

If cost(ES) > cost(ES – SJ) **then** ES ← ES – SJ

End If

End For

End For

End.

46

SO SÁNH CÁC THUẬT TOÁN TỐI ƯU HOÁ

Thuật toán	Thời điểm tối ưu	Hàm mục tiêu	Hệ số tối ưu hoá	Topo mạng	Nối nửa	Số liệu thống kê	Phân mảnh
Dist. INGRES	Động	Thời gian đáp ứng hoặc tổng chi phí	Kích thước TB, chi phí xử lý	Tổng quát hoặc phát tán	Không	1	Ngang
R*	Tĩnh	Tổng chi phí	Lượng TB, kích thước TB, IO, CPU	Tổng quát hoặc cục bộ	Không	1, 2	Không
SDD-1	Tĩnh	Tổng chi phí	Kích thước TB	Tổng quát	Có	1, 3, 4, 5	Không

1. Lực lượng của quan hệ; 2. Số giá trị duy nhất của mỗi thuộc tính; 3. Hệ số tuyển chọn nối;
4. Kích thước của nối trong mỗi thuộc tính nối; 5. Kích thước thuộc tính và kích thước bộ

47

KẾT LUẬN

- Mục tiêu của việc xử lý truy vấn phân tán là hạ thấp tối đa hàm chi phí
- Nguyên liệu quan trọng của bài toán tối ưu truy vấn là số liệu thống kê CSDL và các công thức dùng để đánh giá kích thước các kết quả trung gian.
- Phép toán quan trọng nhất trong xử lý truy vấn phân tán là phép toán nối.
- Việc sử dụng thuật toán nào là còn tuỳ theo từng điều kiện cụ thể:
  - Với mạng diện rộng WAN, nên sử dụng thuật toán SDD-1.
  - Với mạng cục bộ LAN, có thể dùng thuật toán D-Ingres hoặc R* do không sử dụng các nối nửa, trong đó
    - D-Ingres tối ưu động thích hợp với phân mảnh ngang
    - R* tối ưu tĩnh thích hợp với các truy vấn được dùng thường xuyên.

48