

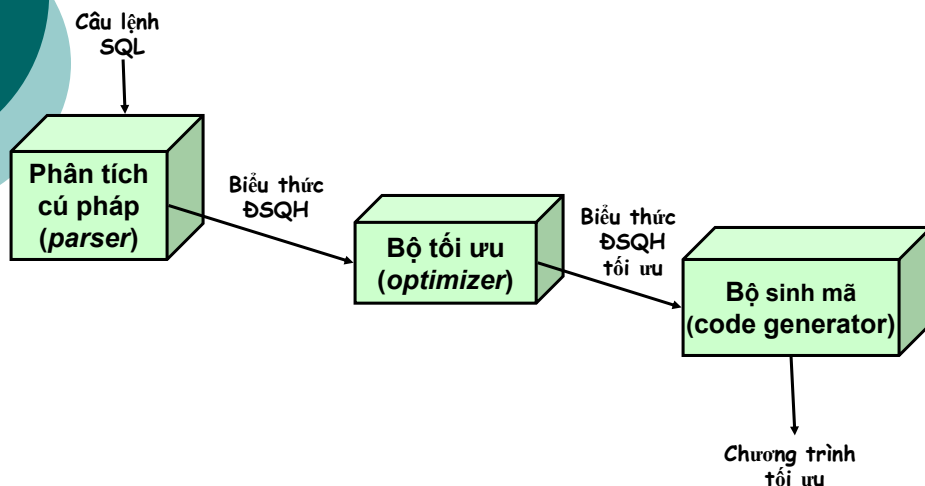
Tối ưu hoá câu hỏi

Vũ Tuyết Trinh

trinhvt@it-hut.edu.vn

Bộ môn Các hệ thống thông tin, Khoa Công nghệ thông tin
Đại học Bách Khoa Hà Nội

Xử lý câu hỏi truy vấn

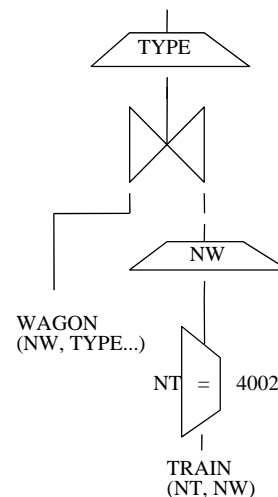


Tối ưu hoá

- Biến đổi biểu thức ĐSQH để tìm 1 biểu thức hiệu quả
- Tối ưu dựa trên cấu trúc và nội dung của dữ liệu
- Nâng cao hiệu quả thực hiện câu hỏi trên 1 hay nhiều tiêu chí: thời gian, sử dụng bộ nhớ, ...
- Lưu ý:
 - Không nhất thiết phải tìm biểu thức tối ưu nhất
 - Chú ý tới tài nguyên sử dụng cho tối ưu

Kỹ thuật tối ưu hoá

- 2 kỹ thuật chính
 - Tối ưu logic (rewriting)
 - Tối ưu vật lý (access methods)
- Mục đích của các kỹ thuật tối ưu
 - Giảm số bản ghi
 - Giảm kích thước bản ghi
- Ví dụ
WAGON (NW, TYPE, COND, STATION, CAPACITY, WEIGHT)
TRAIN (NT, NW)





Nội dung

- ✓ Giới thiệu chung
- Tối ưu logic
- Tối ưu vật lý
- Mô hình giá

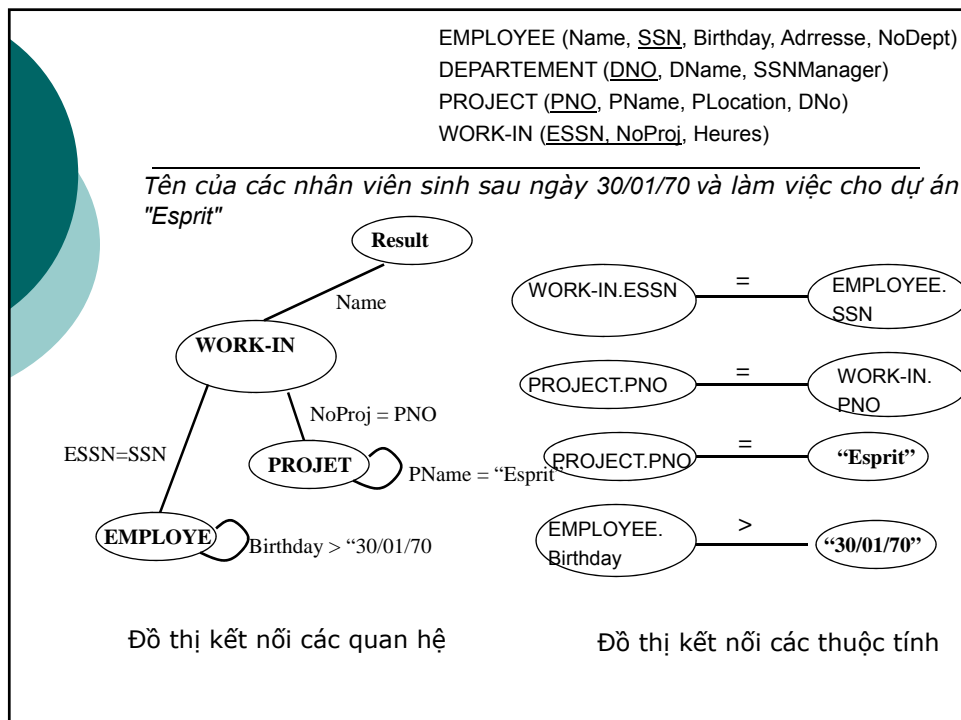


Tối ưu hoá logic

- Sử dụng các phép biến đổi tương đương để tìm ra biểu thức ĐSQH tốt
- Gồm 2 giai đoạn
 - Biến đổi dựa trên ngữ nghĩa
 - Biến đổi dựa trên tính chất của các phép toán ĐSQH

Tối ưu dựa trên ngữ nghĩa

- Mục đích:
 - Dựa trên các ràng buộc dữ liệu để xác định các biểu thức tương đương
 - Viết lại câu hỏi trên khung nhìn dựa trên các định nghĩa của khung nhìn
- Ví dụ
EMPLOYEE (FirstName, LastName, SSN, Birthday, Adresse, NoDept)
DEPARTEMENT (DNO, DName, SSNManager)
PROJECT (PNO, PName, PLocation, DNo)
WORK-IN (ESSN, PNO, Heures)



Tối ưu dựa trên ngữ nghĩa (2)

- Loại bỏ các đồ thị con không liên kết trong đồ thị kết nối các quan hệ
- Kiểm tra mâu thuẫn trong đồ thị kết nối các thuộc tính
- Biến đổi câu hỏi tương đương

Tính chất của phép toán ĐSQH

$A \sim$ tập các thuộc tính, $C \sim$ biểu thức điều kiện

1. Phép chiếu và phép chọn

$$\Pi_A(R) \Rightarrow \Pi_A(\Pi_{A_1}(R)) \text{ nếu } A \subseteq A_1$$

$$\sigma_C(R) \Rightarrow \sigma_{C_1}(\sigma_{C_2}(R)) \text{ nếu } C = C_1 \wedge C_2$$

2. Tính giao hoán đối với phép chọn và chiếu

$$\sigma_{C_1}(\sigma_{C_2}(R)) \Rightarrow \sigma_{C_2}(\sigma_{C_1}(R))$$

$$\sigma_{C_1}(\Pi_{A_2}(R)) \Rightarrow \Pi_{A_2}(\sigma_{C_1}(R))$$

nếu các thuộc tính của C_2 thuộc A_1

$$\Pi_{A_1}(\sigma_{C_2}(R)) \Rightarrow \sigma_{C_2}(\Pi_{A_1}(R))$$

$$\Pi_{A_1}(\Pi_{A_2}(R)) \Rightarrow \Pi_{A_1}(R)$$

nếu $A_1 \subseteq A_2$

Tính chất của phép toán ĐSQH (2)

3. Tính giao hoán và kết hợp của các phép toán

$*, \cap, \cup, -, \times$

$$R \times S \Rightarrow S \times R$$

$$R * S \Rightarrow S * R$$

$$R \cap S \Rightarrow S \cap R$$

$$R \cup S \Rightarrow S \cup R$$

$$(R \times S) \times T \Rightarrow R \times (S \times T)$$

$$(R \cap S) \cap T \Rightarrow R \cap (S \cap T)$$

$$(R \cup S) \cup T \Rightarrow R \cup (S \cup T)$$

$$(R *_{C1} S) *_{C2} T \Rightarrow R *_{C1} (S *_{C2} T) \quad \text{chỉ nếu} \\ \text{Attr}(C2) \subseteq \text{Attr}(S) \cup \text{Attr}(T)$$

Tính chất của phép toán ĐSQH (3)

4. Tính phân phối σ và Π trên các phép toán $*, \cap, \cup, -, \times$

Nếu $C = (CR \wedge CS)$ và nếu $\text{Attr}(CR) \subseteq R$ và $\text{Attr}(CS) \subseteq S$ thì :

$$\sigma_C(R *_{JC} S) \Rightarrow \sigma_{CR}(R) *_{JC} \sigma_{CS}(S)$$

$$\sigma_C(R \times S) \Rightarrow \sigma_{CR}(R) \times \sigma_{CS}(S)$$

Biến đổi biểu thức ĐSQH

| | | |
|-----------|---|---|
| T1 | $R: F1 \wedge F2 \wedge \dots \wedge Fn$ | $((\dots(R:F1) : F2) : \dots) : Fn$ |
| T2 | $(R[Y]) [Z]$ | $R[Z] \text{ n u } Z \subseteq Y$ |
| T3 | $(R[Y]) : F(X)$ | $(R : F(X)) [Y] \text{ n u } X \subseteq Y$ |
| | $(R: F(X)) [Y]$ | $(R[X \cup Y]) : F(X)) [Y] \text{ n u } X \not\subseteq Y$ |
| T4 | $(R(X) \times S(Y)) : F(Z)$ | $(R(X):F) \times S(Y) \text{ n u } Z \subseteq X$ |
| | $(R(X) \times S(Y)) : F(Z1) \wedge F(Z2)$ | $(R(X):F(Z1)) \times (S(Y): F(Z2))$ $\text{ n u } Z1 \subseteq X \text{ và } Z2 \subseteq Y$ |
| T5 | $(R \cup S): F$ | $(R:F) \cup (S:F)$ |
| T6 | $(R - S): F$ | $(R:F) - S$ |
| T7 | $(R(X) \times S(Y)) [Z]$ | $R[X \cap Z] \times S[Y \cap Z]$ |
| T8 | $(R \cup S) [Z]$ | $(R[Z]) \cup (S[Z])$ |

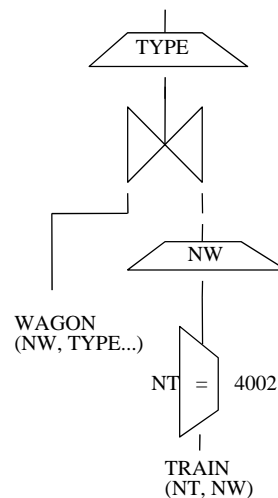
Trình tự áp dụng

- Khai triển phép lựa chọn dựa trên nhiều điều kiện: T1
- Hoán vị phép chọn với tích đề-các, hợp, trừ: T3, T4, T5, T6
- Hoán vị phép chiếu với tích đề-các, hợp : T2, T7, T8
- Nhóm các điều kiện chọn bởi T1 và áp dụng T2 để loại các phép chiếu dư thừa

Bài tập

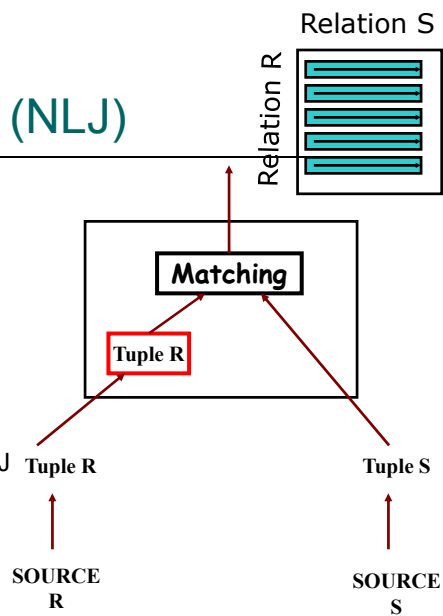
Lựa chọn cách truy nhập dữ liệu

- Giả thiết
 - TRAIN : có chỉ số trên NT
 - WAGON : có chỉ số trên NW
- Thực hiện phép kết nối
 - Lựa chọn 1 giải thuật.
 - Lựa chọn cách truy nhập các quan hệ

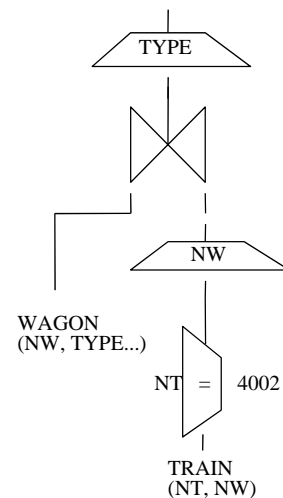


Nested-loop-join (NLJ)

- Nguyên tắc
 - Duyệt 1 lần trên quan hệ ngoài R & lặp trên quan hệ trong S
- Các mở rộng của thuật toán
 - Tuple-based NLJ, block-based NLJ, index-based NLJ



Thực hiện như thế nào?



Thông tin về các quan hệ

- Kích thước của các quan hệ và bản ghi

| Relation | Cardinality | Record size |
|----------|-------------|-------------|
| WAGON | 200000 | 60 |
| TRAIN | 60000 | 30 |
| TRAFFIC | 80000 | 20 |

- Thông tin về các thuộc tính

| Attribute | Cardinality | Size | min - max |
|-----------|-------------|------|-----------|
| NW | 200000 | 20 | |
| TYPE | 200 | 5 | |
| COND | 5 | 15 | |
| CAPACITY | 400 | 15 | 5-45 |
| NT | 2000 | 10 | |
| DATE | 800 | 6 | |

- Thông tin về các chỉ số

| Relation | Attributes | Unique | Type | Num of pages |
|----------|------------|--------|-----------|--------------|
| WAGON | NW | Yes | Principal | 45 |
| WAGON | TYPE | No | Secondary | 25 |
| WAGON | COND | No | Secondary | 30 |
| WAGON | CAPACITY | No | Secondary | 25 |
| TRAIN | NT | No | Principal | 18 |
| TRAFFIC | NT | No | Principal | 20 |
| TRAFFIC | DATE | no | Principal | 40 |

| Relation | Cardinality | Record size (num of rec/page) | Num. of pages (NP) |
|----------|-------------|----------------------------------|-----------------------|
| WAGON | 200000 | 60(100) | 1500(375) |
| TRAIN | 60000 | 30 (200) | 225(60) |
| TRAFFIC | 80000 | 20 (300) | 200(60) |

Mô hình giả

- Chỉ phí thực hiện câu hỏi phụ thuộc:

- đọc/ghi bộ nhớ ngoài (số trang nhớ)
- Kích thước dữ liệu phải xử lý

- Chi phí truy nhập dữ liệu

- Đọc ghi dữ liệu
- xử lý
- Truyền thông giữa các trạm làm việc

$$CTA = \sigma * NBPAGES + \tau * NBNUPLETS (+ \mu * NBMESSAGES)$$

- Trọng số

- σ = trọng số đọc/ghi dữ liệu (ví dụ = 1)
- τ = trọng số xử lý của CPU (ví dụ = 1/3)
- μ = trọng số truyền dữ liệu



Tối ưu hoá dựa trên mô hình giá

- Mục đích: Chọn phương án thực hiện câu hỏi với chi phí thấp nhất
- Nhận xét:
 - Chi phí cho liệt kê các phương án trả lời câu hỏi
 - Chi phí cho lượng hoá các phương án theo mô hình giá
 - Có thể sử dụng các « mẹo » (heuristics) để giảm không gian tìm kiếm của câu hỏi



Kết luận

- Tối ưu hoá nhằm tìm phương án tốt nhất để thực hiện một câu hỏi
 - Cần lưu ý: chi phí thực hiện tối ưu hoá và chi phí thực hiện câu hỏi
- Các kỹ thuật tối ưu
 - Logic : kiểm tra điều kiện ràng buộc của các thuộc tính/quan hệ và điều kiện lựa chọn trong câu hỏi, biến đổi tương đương các biểu thức ĐSQH
 - Vật lý : tổ chức vật lý của dữ liệu trên đĩa, mô hình giá
 - Không nhất thiết phải áp dụng tất cả các kỹ thuật trên khi thực hiện tối ưu hoá 1 câu hỏi