Appunti Basi di Dati

Luca Seggiani

14 Marzo 2024

Tutti gli operatori visti finora, ovvero gli operatori insiemistici e quelli relazionali, bastano a realizzare qualsiasi possibile interrogazione. Ogni altro operatore sarà un' operatore derivato dei 5 operatori relazionali e i 3 poeratori insiemistici. Un particolare operatore derivato è:

1 Divisione

Dati due insiemi di attributi disgiunti X_1 e X_2, una relazione r sulla loro unione e una relazione r_2 su X_2, la divisione $r \div r_2$ è una relazione su X_1 che contiene le n-uple ottenute come "proiezione" di n-uple di r che si combinano con tutte le n-uple di r_2 per formare n-uple di r, in una sorta di prodotto cartesiano inverso. In simboli:

$$r \div r_2 = \{t_1[X_1] \quad | \quad \forall t_2 \in r_2 \quad \exists t \in r : t[X_1] = t_1, \quad t[X_2] = t_2\}$$

possiamo dimostrare che è un operatore derivato definendolo come composizione di operatori fondamentali, in questo modo:

$$r \div r_2 = \pi_{X_1}(r) - \pi_{X_1}((\pi_{X_1}(r) \times r_2) - r)$$

Ovvero:

- $\pi_{X_1}(r) \times r_2$ contiene tutte le n-uple di $\pi_{X_1}(r)$ "estese" con tutti i possibili valori di r_2.
- $(\pi_{X_1}(r) \times r_2) - r$ contiene le "estensioni" di $\pi_{X_1}(r)$ che non compaiono in r.
- $\pi_{X_1}((\pi_{X_1}(r) \times r_2) - r)$ contiene le n-uple di $\pi_{X_1}(r)$ per le quali un qualche completamento con r_2 non compare in r.
- Togliendo queste ultime n-uple a $\pi_{X_1}(r)$ otteniamo tutte le n-uple di $\pi_{X_1}(r)$ che si combinano con tutte le n-uple di r_2.

2 Chiusura transitiva

Poniamoci il problema di dover trovare, in un'opportuna tabella supervisione formata da matricole di impiegati e supervisori di tali impiegati, le matricole di tutti i supervisori di un dato impiegato (ammettendo che i supervisori siano impiegati e possano a loro volta avere supervisori). Tale richiesta sarebbe perfettamente valida, ma impossibile da esprimere attraverso gli operatori dell'algebra relazionale.
Nell'algebra relazionale non esiste la possibilità di esprimere interrogazioni che calcolino la chiusura transitiva di una relazione arbitraria. Tale operazione potrebbe infatti richiedere un numero infinito di di join (join illimitato).

3 Espressioni equivalenti

Due espressioni sono equivalenti se producono lo stesso risultato qualunque sia l'istanza fornitagli. In questo caso, sarà opportuno scegliere espressioni di costo minore, dove il loro "costo" è determinato dalle dimensioni delle istanze intermedie che la loro esecuzione genera. Vediamo alcune equivalenze fondamentali:

- **Atomizzazione delle selezioni**: una congiunzione di selezioni può essere sostituita da una sequenza di selezioni atomiche:

$$\sigma_{F_1 \wedge F_2}(E) = \sigma_{F_1}(\sigma F_2(E))$$

- **Idempotenza delle proiezoni**: una proiezione può essere trasformata in una sequenza di proiezioni:

$$\pi_X(E) = \pi_X(\pi_{XY}(E))$$

- **Push selections down**: se una condizione F coinvolge solo attributi dell'espressione E_2:

$$\sigma_F(E_1 \bowtie E_2) = E_1 \bowtie \sigma_F(E_2)$$

- **Push projections down**: se un'espressione E_1 ha attributi X_1, un'espressione E_2 ha attributi X_2, $Y_2 \subseteq X_2$ e gli attributi $X_2 - Y_2$ non sono coinvolti nel join ($X_1 \cap X_2 \subseteq Y_2$):

$$\pi_{X_1 Y_2}(E_1 \bowtie E_2) = E_1 \bowtie \pi_{Y_2}(E_2)$$

4 Ottimizzazione delle interrogazioni

Un modulo presente nel DBMS è il **query processor** (od ottimizzatore). L'ottimizzatore si occupa di scegliere la strategia realizzativa a partire dall'istruzione in linguaggio dichiarativo di alto livello, tenendo conto del costo di implementazioni diverse. Le fasi di eseecuzione di una query saranno:

- Analisi lessicale, sintattica e semantica della query in linguaggio di alto livello (SQL). Al termine di questa analisi, la query verrà tradotta in un'espressione dell'algebra relazionale.
- Ottimizzazione algebraica: a questo punto verranno calcolate una o più nuove espressioni algebriche equivalenti a quella di partenza, sfruttando le equivalenza sopra descritte.
- Ottimizzazoine basata sui costi: fra le alternative calcolate prima, viene selezionata la più efficiente, che viene poi utilizzata per effettuare l'interrogazione effettiva sulla base di dati.

Nella prima e l'ultima di queste fasi, il DMBS interagisce con un componente detto catalogo, che contiene informazioni sugli schemi contenuti nel database, la cardinalità delle loro istanze, ecc..

Profili delle relazioni

Tra le informazioni quantitative memorizzate nel catalogo troviamo:

- Cardinalità di ciascuna relazione;
- Dimensioni delle n-uple;
- Dimensioni dei valori;
- Numero di valori distinti degli attributi;
- Valore minimo e massimo degli attributi.

Queste informazioni vengono usate nella fase di ottimizzazione basata sui costi per stimare le dimensioni dei risultati intermedi di più espressioni algebriche alternative generate dall'ottimizzazione algebrica.

Ottimizzazione algebrica

In verità, il termine ottimizzazione non è completamente accurato: il processo utilizza infatti delle euristiche per trovare risultati migliori. Si basa sul concetto di equivalenza per trovare query che restituiscono lo stesso risultato generando dimensioni d'istanza intermedie minori. Ad esempio, un'ottimizzazione tipica è quella di eseguire selezioni e proiezioni il più presto possibile, ovvero le cosiddette "push selections down" e "push projections down".

5 Grafi

Un grafo $G = (V, E)$ consiste in un insieme V di vertici (o nodi) e un insieme E di coppie di vertici, detti archi. Ogni arco ovviamente connetter fra loro due vertici. Possiamo allora distinguere:

- **Grafi orientati**: detti anche grafi diretti, dove ogni arco è orientato è rappresenta relazioni ordinate fra oggetti;
- **Grafi non orientati**: detti anche grafi indiretti, dove ogni arco non è orientato e rappresenta relazioni simmetriche fra oggetti.

Sui grafi si possono definire **cammini** da un vertice x ad un vertice y. Formalmente, un cammino è:

$$(v_0, ..., v_k) \quad \text{di} \quad V, \quad v_0 = x, \quad v_k = y \quad | \quad 1 \leq i \leq k : (v_{i-1}, v_i) \in E$$

Un cammino che torna da dove parte, ovvero dove $(v_0, ..., v_k) : v_0 = v_k$, è detto ciclico. Si dice che un grafo diretto privo di cicli è **aciclico**.

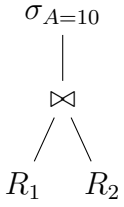
Alberi

Un grafo non orientato si dice connesso se esiste un cammino fra ogni coppia di vertici. Un'albero è un grafo non orientato nel quale due vertici qualsiasi sono connessi da uno e un solo cammino.

Un'interrogazione può essere rappresentata da un'albero, dove le foglie (i nodi finali) sono dati (relazioni, file, ecc...), e i nodi intermedi sono operatori (operatori algebrici, poi effettivi operatori di accesso ai dati). Ad esempio, l'albero corrispondente all'espressione:

$$\sigma_{A=10}(R_1 \bowtie R_2)$$

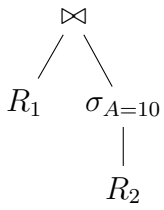
sarà:



La stessa interrogazione dopo il push down della selezione $\sigma_{A=10}$ sarà allora:

$$R_1 \bowtie \sigma A = 10(R_2)$$

con relativo albero:



Procedura euristica di ottimizzazione
La procedura di ottimizzazione consiste quindi nel:

- Decomporre le selezioni congiuntive in successive selezioni atomiche;
- Anticipare il più possibile le selezioni;
- In una sequenza di selezioni, anticipare le più selettive;
- Combinare prodotti cartesiani e selezioni per formare join;
- Anticipare il più possibile le proiezioni.