

Appunti Basi di Dati (Teoria)

Luca Seggiani

13 Marzo 2024

1 Linguaggi per basi di dati

I linguaggi per basi di dati appartengono a 2 categorie distinte:

- Operazioni sullo schema: **Data Definition Language (DDL)**;
- Operazioni sui dati: **Data Manipulation Lanaguage (DML)**
 - Interrogazione (query)
 - Aggiornamento (update).

Interrogazioni di basi di dati

Un'operazione di interrogazione è un'operazione di lettura che richiede l'accesso a una o più tabelle. Per specificare interrogazioni si possono seguire due formalismi:

- **Modo dichiarativo:** si specificano le proprietà del risultato (che cosa)
- **Modo procedurale:** si specificano le modalità di generazione del risultato (come).

L'**algebra relazionale** permette di specificare delle interrogazioni secondo il modello procedurale: cioè elencando i passi "primitivi" necessari alla generazione di una risposta. Il **calcolo relazionale** invece ci permette di definire in modo dichiarativo quello che sarà il risultato dell'interrogazione. Come vedremo, sarà il calcolo relazionale a definire la semantica del linguaggio relazionale, perché permette di fornire un'implementazione slegata dai dettagli procedurali.

2 Algebra relazionale

Un'algebra è una struttura matematica dotata di un'insieme di dati e una serie di operatori che manipolano suddetti dati. Nell'algebra relazionale abbiamo che:

- **Dati:** relazioni
- **Operatori:**
 - su relazioni
 - che producono relazioni
 - che possono essere composti.

essi sono divisi fra:

- **Operatori su insiemi:**
unione, intersezione, differenza
- **Operatori su relazioni:**
ridenominazione, selezione, proiezione,
join:
 - * naturale, prodotto cartesiano, theta

Notazione dell'algebra relazionale

Specifichiamo adesso la notazione usata:

- R, R_1, \dots indicano nomi di relazioni
- A, B, C, A_1, \dots indicano nomi di attributo
- XY è un'abbreviazione di $X \cup Y$
- Una relazione con n-uple t_1, t_2, \dots è indicata con l'insieme $\{t_1, t_2, \dots\}$
- $t_j[A_j]$ indica il valore della n-upla t_j sull'attributo A_j
- $t[X]$ indica l'n-upla ottenuta da considerando solo gli elementi di X .

Operatori su insiemi

Le relazioni sono insiemi di tuple, e non possono avere elementi duplicati (sarebbero altrimenti multiinsiemi). I risultati di operazioni fra relazioni sono a loro volta relazioni, ovvero insiemi di tuple. Gli operatori fra relazioni possono applicarsi solo e soltanto fra relazioni definite sullo stesso insieme

di attributi X , e il risultato sarà a sua volta definito sullo stesso insieme di attributi X .

Unione

L'operatore di unione fra due relazioni sull'insieme di attributi X comporta la formazione di una nuova relazione che ha tutte le n-uple delle relazioni unite. Eventuali n-uple identiche fra le relazioni appariranno nel risultato una volta sola. Il suo simbolo è \cup .

Intersezione

L'operatore di intersezione fra due relazioni sull'insieme di attributi X restituisce una nuova relazione che contiene soltanto gli elementi appartenenti sia alla prima che alla seconda relazione. Il suo simbolo è \cap .

Differenza

L'operatore di differenza fra due relazioni sull'insieme di attributi X restituisce una nuova relazione contenente tutte le n-uple che appartengono alla prima relazione ma non alla seconda. L'operatore di differenza è l'unico fra gli operatori insiemistici a non essere commutativo. Il suo simbolo è il meno ($-$), o più propriamente \setminus .

Notiamo che l'unione è l'unico operatore in grado di creare relazioni con un numero di elementi maggiore di quello degli operandi. Questa caratteristica tornerà poi utile nel calcolo relazionale.

Descriviamo adesso gli operatori non insiemistici (sulle relazioni):

Ridenominazione

L'operatore di ridenominazione è un'operatore monadico che modifica lo schema dell'operando, lasciando inalterata l'istanza. Si scrive, data una relazione R , come:

$$\rho_{B_1 B_2 \dots} \xleftarrow{A_1 A_2 \dots (R)}$$

I pedici a sinistra e a destra della freccia sono insiemi di attributi. Avremo che:

- L'attributo A_1 viene sostituito dall'attributo B_1
- L'attributo A_2 viene sostituito dall'attributo B_2
- ecc...

Selezione

L'operatore di selezione è un operatore monadico che produce un risultato con lo stesso schema dell'operando, e un sottoinsieme di n-uple che rispettano una determinata condizione. Si scrive come:

$$\sigma_F(R)$$

sulla relazione R , dove F è un'espressione Booleana (predicato) ottenuta componendo con gli operatori logici AND, OR e NOT le condizioni atomiche:

- $A \star B$, dove A e B sono attributi di X con domini compatibili e \star un operatore di confronto.
- $A \star k$, dove A è un attributo di X e k una costante con dominio compatibile con A e \star sempre un operatore di confronto.

Notiamo che la condizione atomica è vera solo per valori non nulli in ogni attributo.

Selezione valori NULL

Per riferirsi ai valori NULL occorre usare le apposite condizioni: IS NULL e IS NOT NULL.

Proiezione

La proiezione è un'operatore monadico che produce un sottoinsieme degli attributi dell'operando contenente tutte le sue n-uple ristrette soltanto ad alcuni attributi. Si scrive, data una relazione $R(X)$ e un insieme di attributi $Y \subseteq X$, come:

$$\pi_Y(R)$$

Il risultato è una relazione su Y che contiene l'insieme delle n-uple di R ristrette ai soli attributi di Y . Notare che, di nuovo, non possono esserci righe ripetute. Ogni n-upla ripetuta nell'insieme risultato verrà inserita una volta sola.

Cardinalità di proiezioni

Una proiezione contiene al massimo tante n-uple quante ne contiene l'operando. Inoltre, se l'insieme di attributi X è superchiave della relazione R , allora la proiezione $\pi_X(R)$ avrà tante n-uple quante ne ha l'operando.

Join

Non possiamo, usando gli operatori di selezione e proiezione, estrarre e combinare informazioni da più relazioni diverse fra loro, e non possiamo nemmeno combinare informazioni presenti in n-uple diverse di una stessa relazione. Avremo allora bisogno di un'ulteriore serie di operazioni, le cosiddette operazioni di join. I join ci permettono di unire sulla stessa riga più righe di relazioni diverse. Vediamo nel dettaglio:

Join naturale

Il join naturale è un'operatore con due operandi (generalizzabile), che produce come risultato l'unione degli attributi degli operandi, contenente le n-uple costruite ciascuna a partire da un n-upla di ognuno degli operandi. Formalmente: Date due relazioni R_1X_1 e R_2X_2 , il loro join naturale si scrive

come:

$$R_1 \bowtie R_2$$

Il risultato è una relazione $R(X_1 \cup X_2)$ definita come:

$$R(X_1 \cup X_2) = R_1(X_1) \bowtie R_2(X_2) = \{t | \exists t_1 \in R_1, \quad t_2 \in R_2$$

$$\text{con} \quad t[X_1] = t_1, \quad t[X_2] = t_2\}$$

N.B.: è perfettamente plausibile voler fare il join naturale tra due relazioni senza attributi in comune, e in questo caso si ottiene il prodotto cartesiano fra le due.

Cardinalità del join

Il join di due relazioni R_1 e R_2 contiene un numero di n-uple compreso fra 0 e il prodotto di $|R_1|$ e $|R_2|$. Se il join coinvolge una chiave di R_2 allora il numero di n-uple è compreso fra 0 e $|R_1|$. Se il join coinvolge una chiave di R_2 e un vincolo di integrità referenziale allora il numero di n-uple è uguale a $|R_1|$. Più formalmente:

Il join di $R_1(A, B)$ e $R_2(B, C)$ contiene un un numero di n-uple:

$$0|R_1 \bowtie R_2| \leq |R_1| \times |R_2|$$

Se B è una chiava di R_2 allora il numero di n-uple è

$$0 \leq |R_1 \bowtie R_2| \leq |R_1|$$

Se B è una chiave di R_2 ed esiste un vincolo di integrità referenziale fra B (in R_1) e R_2 allora il numero delle n-uple è:

$$|R_1 \bowtie R_2| = |R_1|$$

Una problematica del join è il fatto che le n-uple che non contribuiscono al risultato ("che non fanno join") restano tagliate fuori. Introduciamo allora altri tipi di operatori di join, i cosiddetti:

Join esterno

Il join esterno estende, con valori NULL, le n-uple che verrebbero altrimenti scartate dal join (interno). Ne esistono tre versioni:

- **Sinistro:** mantiene tutte le n-uple del primo operando (\bowtie_{LEFT});
- **Destro:** mantiene tutte le n-uple del secondo operando (\bowtie_{DESTRO});
- **Completo:** mantiene tutte le n-uple di entrambi gli operandi (\bowtie_{FULL}).

Join e proiezioni

Date due relazioni $R_1(X_1)$ e $R_2(X_2)$:

$$\pi_{X_1}(R_1 \bowtie R_2) \subseteq R_1$$

Date una relazione $R(X)$ con $X = X_1 \cup X_2$

$$(\pi_{X_1}(R) \bowtie \pi_{X_2}(R)) \supseteq R$$

Prodotto cartesiano

Come detto prima, date due relazioni $R_1(X_1)$ e $R_2(X_2)$, senza attributi in comune, cioè con $X_1 \cap X_2 = 0$, la definizione di join naturale e comunque senso e restituisce il prodotto cartesiano delle due relazioni:

$$R = R_1 \bowtie R_2 = R_1 \times R_2$$

la cardinalità del prodotto cartesiano è uguale al prodotto delle cardinalità delle due relazioni.

Theta join

Nella pratica, il prodotto cartesiano ha senso quasi solamente se seguito da una selezione $\sigma_F(R_1 \times R_2)$. Questa combinazione prende il nome di theta join (è un operatore derivato) ed è indicato come:

$$R_1 \bowtie_F R_2$$

dove F è un certo predicato. Spesso F è una congiunzione di atomi di confronto $A_1 \sigma A_2$ dove σ è un operatore di confronto ($\leq, <, =$, ecc...) e A_1 e A_2 attributi di relazioni diverse. Quando l'operatore di confronto è l'uguaglianza parliamo di equi-join.