Calcolo Relazionale

Angelo Montanari

**Dipartimento di Scienze Matematiche,
Informatiche e Fisiche
Università di Udine**

Introduzione

Calcolo relazionale: famiglia di linguaggi di interrogazione dichiarativi basati sul *calcolo dei predicati del primo ordine*

Prenderemo in considerazione due linguaggi:

- calcolo relazionale su domini
- calcolo relazionale su tuple con dichiarazioni di range (base per il linguaggio SQL)

Alcune assunzioni

- I simboli di predicato corrispondono alle relazioni presenti nella base di dati (più alcuni predicati standard quali uguaglianza e diseguaglianza); *non compaiono simboli di funzione* (data la struttura piatta delle relazioni imposta dalla Prima Forma Normale, non sono necessari)
- Nel calcolo relazionale vengono utilizzate prevalentemente *formule aperte* (formule con variabili libere, il cui valore di verità dipende dai valori assegnati alle variabili libere): il risultato di un'interrogazione (formula aperta) è costituito dalle tuple di valori che, sostituiti alle variabili libere, la rendono vera
- In coerenza con quanto fatto in algebra relazionale (attributi con nome), utilizzeremo una *notazione non posizionale*

Calcolo relazionale su domini - 1

Le *espressioni* del calcolo relazionale su domini hanno la forma:

$$\{A_1 : x_1, \ldots, A_k : x_k \mid f\}$$

dove

- $A_1, \ldots, A_k$ sono *attributi* distinti (possono anche non comparire nello schema della base di dati)
- $x_1, \ldots, x_k$ sono *variabili* (che assumiamo essere distinte, nonostante non sia strettamente necessario)
- f è una *formula* costruita a partire da formule atomiche utilizzando (eventualmente) i connettivi Booleani $\vee$, $\wedge$, e $\neg$ e i quantificatori $\exists x$ e $\forall x$, con x variabile

Calcolo relazionale su domini - 2

Le *formule atomiche* possono essere di due tipi:

- $R(A_1 : x_1, \ldots, A_h : x_h)$, dove $R(A_1, \ldots, A_h)$ è uno schema di relazione e $x_1, \ldots, x_h$ sono variabili
- $x\theta y$ oppure $x\theta c$, dove x, y sono variabili, c è una costante e θ è un operatore di confronto $(=, \neq, \leq, \ldots)$

La lista di coppie $A_1 : x_1, \ldots, A_k : x_k$ è detta *lista degli obiettivi*. Essa definisce la struttura del risultato costituito dalla relazione su $A_1, \ldots, A_k$ che contiene tutte e sole le tuple di valori che, sostituiti a $x_1, \ldots, x_k$, rendono vera la formula f

Calcolo relazionale su domini - 3

Il *valore di verità* di una formula è definito nel modo seguente (assumiamo per semplicità che tutti gli attributi abbiano lo stesso dominio):

- una formula atomica $R(A_1 : x_1, \ldots, A_h : x_h)$ è vera sui valori di $x_1, \ldots, x_h$ che costituiscono una tupla di R
- una formula atomica $x\theta y$ vera sui valori di x, y che rendono vera la condizione θ (lo stesso per $x\theta c$)
- il valore di verità di disgiunzioni $\vee$, congiunzioni $\wedge$ e negazioni $\neg$ è definito nel modo usuale
- una formula della forma $\exists x f$ (rispettivamente $\forall x f$) è vera se esiste almeno un elemento del dominio che (rispettivamente, ogni elemento del dominio), sostituito ad x, rende vera f

Alcuni esempi - 1

Interrogazione 1. Selezionare le tuple della relazione $LAVORA\_A$ contraddistinte da un numero di ore settimanali maggiore di 5.

$\{IMP : i, PROGETTO : p, ORE\_SETTIMANA : o \mid$

$LAVORA\_A(IMP : i, PROGETTO : p, ORE\_SETTIMANA : o) \land o > 5\}$

Il primo congiunto della formula impone che le tuple di valori assegnati alle variabili (i, p, o) appartengano alla relazione $LAVORA\_A$; il secondo impone che il valore della variabile o sia maggiore di 5.

Alcuni esempi - 2

Interrogazione 2. Selezionare le coppie (impiegato,progetto) tali che l'impiegato dedica al progetto un numero di ore settimanali maggiore di 5.

$\{IMP : i, PROGETTO : p \mid \exists o(LAVORA\_A(IMP : i,$

$PROGETTO : p, ORE\_SETTIMANA : o) \wedge o > 5)\}$

o, equivalentemente (la quantificazione esistenziale può essere omessa):

$\{IMP : i, PROGETTO : p \mid LAVORA\_A(IMP : i,$

$PROGETTO : p, ORE\_SETTIMANA : o) \wedge o > 5\}$

Alcuni esempi - 3

Interrogazione 3. Selezionare nome e cognome dei manager.

$$\{NOME : n, COGNOME : c \mid IMPIEGATO(NOME : n, INIZIALE : i, COGNOME : c, CF : f, \ldots) \land DIPARTIMENTO(\ldots, MANAGER : f, \ldots)\}$$

Si noti come l'operazione di join dell'algebra relazionale venga espressa attraverso la congiunzione di formule atomiche, una per ogni relazione coinvolta, e come variabili ripetute si possano usare per esprimere condizioni di uguaglianza.

Alcuni esempi - 4

Interrogazione 4. Selezionare codice fiscale e stipendio dei supervisori degli impiegati che guadagnano più di 40000 euro.

$\{CF : fs, STIPENDIO : ss \mid IMPIEGATO(\ldots,$

$STIPENDIO : s, SUPERVISORE : fs, \ldots) \wedge$

$IMPIEGATO(\ldots, CF : fs, \ldots, STIPENDIO : ss, \ldots)$

$\wedge s > 40000\}$

Si noti che se per confrontare tuple diverse di una stessa tabella è sufficiente includere nella formula più condizioni sullo stesso predicato, utilizzando variabili diverse.

Alcuni esempi - 5

Interrogazione 5. Selezionare il codice fiscale degli impiegati che guadagnano più del loro supervisore.

$\{CF : f \mid IMPIEGATO(\ldots, CF : f, \ldots, STIPENDIO : s,$

$SUPERVISORE : fs, \ldots) \wedge IMPIEGATO(\ldots, CF : fs,$

$\ldots, STIPENDIO : ss, \ldots) \wedge s > ss\}$

Si noti come l'uso di variabili diverse consenta di confrontare valori dello stesso attributo provenienti da tuple diverse.

Alcuni esempi - 6

Interrogazione 6. Trovare i dipartimenti i cui impiegati guadagnano tutti più di 40000 euro.

$\{DNUMERO : d \mid DIPARTIMENTO(DNUMERO : d, \ldots) \wedge$

$\neg(\exists n(\exists i(\ldots(\exists st(\exists su(IMPIEGATO(NOME : n, INIZIALE : i,$

$\ldots, STIPENDIO : st, SUPERVISORE : su, DIP : d) \wedge$

$st \leq 40000)))\ldots)))\}$

In alternativa, si può utilizzare la quantificazione universale (consentita nel calcolo relazionale).

Alcuni esempi - 7

Interrogazione 6. (soluzione alternativa) Trovare i dipartimenti i cui impiegati guadagnano tutti più di 40000 euro.

$$\{DNUMERO : d \mid DIPARTIMENTO(DNUMERO : d, \ldots) \wedge$$
$$\forall n(\forall i(\ldots(\forall st(\forall su(\neg IMPIEGATO(NOME : n, INIZIALE : i,$$
$$\ldots, STIPENDIO : st, SUPERVISORE : su, DIP : d) \vee$$
$$st > 40000)))\ldots))\}$$

Potenza e limiti del calcolo su domini

Il calcolo relazionale su domini ammette *espressioni prive di significato* (dal punto di vista pratico). Ad esempio, l'espressione

$\{A_1 : x_1, A_2 : x_2 \mid R(A_1 : x_1) \wedge x_2 = x_2\}$

produce quale risultato una relazione su A_1, A_2 costituita da tuple il cui valore su A_1 compare nella relazione R, mentre il valore A_2 è un qualunque valore del dominio. In particolare:

- se il dominio è infinito, il risultato è infinito
- se il dominio cambia, cambia il risultato (*espressione dipendente dal dominio*)

Altro esempio ("operatore di complementazione"):

$\{A_1 : x_1 \mid \neg R(A_1 : x_1)\}$

Condizione di indipendenza dal dominio

Un'*espressione* di un linguaggio di interrogazione si dice *indipendente dal dominio* se il suo risultato, su ciascuna istanza della base di dati, non varia al variare del dominio rispetto al quale l'espressione viene valutata (purché ogni dominio contenga almeno i valori presenti nell'istanza e nell'espressione).

Un *linguaggio* si dice *indipendente dal dominio* se tali sono tutte le sue espressioni.

E' immediato vedere che il *calcolo relazionale su domini non è* indipendente dal dominio, mentre l'*algebra relazionale* lo *è* (essa costruisce i risultati a partire dalle relazioni presenti nella base di dati, senza far mai riferimento ai domini degli attributi: i valori che compaiono nei risultati sono tutti presenti nell'istanza cui l'espressione viene applicata)

Calcolo su domini e algebra relazionale - 1

Definiamo due linguaggi *equivalenti* se per ogni espressione dell'uno esiste un'espressione dell'altro ad essa equivalente (ossia che produce lo stesso risultato) e viceversa.

Se limitiamo la nostra attenzione al sottoinsieme del calcolo su domini costituito dalle sole espressioni indipendenti dal dominio, possiamo provare la seguente equivalenza:

- per ogni espressione del calcolo relazionale su domini indipendente dal dominio esiste un'espressione dell'algebra relazionale ad essa equivalente e viceversa

In entrambi i versi la dimostrazione è costruttiva e si basa su una induzione sulla struttura dell'espressione

Calcolo su domini e algebra relazionale - 2

Elementi fondamentali della **corrispondenza** tra espressioni dell'algebra relazionale e del calcolo su domini:

esiste una corrispondenza tra condizioni elementari di selezione e formule atomiche delle forme $x\theta y$ e $x\theta c$, proiezioni e quantificazioni esistenziali, join e congiunzioni, unioni e disgiunzioni, negazioni e differenze insiemistiche. I quantificatori universali possono essere ignorati in quanto possono essere ricondotti a quantificazioni esistenziali.

La *completezza espressiva* dell'insieme degli operatori di base dell'algebra relazionale *segue* dalla corrispondenza tra algebra relazionale e calcolo relazionale su domini (ciò che non può essere espresso nell'algebra relazionale, ad esempio le chiusure ricorsive, è ciò che non può essere espresso nel calcolo relazionale su domini)

Variabili: singoli vs. tuple di valori

Il calcolo relazionale su domini presenta anche lo svantaggio di richiedere *numerose variabili*, spesso una per ciascun attributo di ciascuna relazione coinvolta (lo stesso con i quantificatori).

Dal calcolo relazionale su domini al *calcolo relazionale su tuple*: le variabili denotano tuple, non singoli valori.

Vantaggio: una variabile per ciascuna relazione coinvolta. Occorre, ovviamente, associare una *struttura* (insieme degli attributi della relazione) a ciascuna variabile che consenta di individuare le singole componenti delle tuple.

Calcolo su tuple con dichiarazioni di range - 1

Le *espressioni* del calcolo relazionale su tuple con dichiarazioni di range hanno la forma:

$$\{\mathcal{T} \mid \mathcal{L} \mid f\}$$

dove

- $\mathcal{T}$ è la *lista degli obiettivi*, con elementi del tipo $Y : x.Z$ ($x.Z$ se $Y = Z$), con x variabile e Y, Z liste di attributi di pari lunghezza; gli attributi in Z devono comparire nello schema di relazione che costituisce il range di x (se in $x.Z$ Z contiene tutti gli attributi della relazione, si può scrivere $x.*$)
- $\mathcal{L}$ è una lista di elementi (detta *range list*) della forma $x(R)$, con x variabile e R nome di relazione

Calcolo su tuple con dichiarazioni di range - 2

- f è una *formula* costruita a partire da formule atomiche utilizzando (eventualmente) i connettivi Booleani $\vee, \wedge$, and $\neg$ e i quantificatori $\exists x(R)$ e $\forall x(R)$, che associano i range (schemi di relazioni) alle variabili ($\exists x(R) f$ significa "esiste nella relazione R una tupla x che soddisfa la formula f").

Le *formule atomiche* sono del tipo:

- $x_1.A_1 \theta x_2.A_2$ oppure $x.A\theta c$, dove x, y sono variabili (che denotano tuple), $x_1.A_1$ (rispettivamente $x_2.A_2, x.A$) denota il valore che x_1 assume sull'attributo A_1 (rispettivamente x_2 su A_2, x su A), c è una costante e θ è un operatore di confronto $(=, \neq, \leq, \ldots)$

Si noti che le dichiarazioni di range (nella range list e nelle quantificazioni) specificano l'insieme dei valori che possono essere assegnati alle variabili. Non occorrono più condizioni atomiche che vincolano una tupla ad appartenere ad una relazione.

Alcuni esempi - 1

Interrogazione 1. Selezionare le tuple della relazione $LAVORA\_A$ contraddistinte da un numero di ore settimanali maggiore di 5.

$$\{i.* \mid i(LAVORA\_A) \mid i.ORE\_SETTIMANA > 5\}$$

La dichiarazione di range impone che le tuple di valori assegnate alla variabile i appartengano alla relazione $LAVORA\_A$; la condizione $i.ORE\_SETTIMANA > 5$ impone che il numero di ore settimanali sia maggiore di 5.

Alcuni esempi - 2

Interrogazione 2. Selezionare le coppie (impiegato,progetto) tali che l'impiegato dedica al progetto un numero di ore settimanali maggiore di 5.

$\{i.(IMP, PROGETTO) \mid i(LAVORA\_A) \mid$

$i.ORE\_SETTIMANA > 5\}$

Interrogazione 3. Selezionare nome e cognome dei manager.

$\{i.(NOME, COGNOME) \mid i(IMPIEGATO),$

$d(DIPARTIMENTO) \mid i.CF = d.MANAGER\}$

Alcuni esempi - 3

Interrogazione 4. Selezionare codice fiscale e stipendio dei supervisori degli impiegati che guadagnano più di 40000 euro.

$$\{CF\_SUP, STIPENDIO\_SUP : i'.(CF, STIPENDIO) \mid$$

$$i(IMPIEGATO), i'(IMPIEGATO) \mid$$

$$i.SUPERVISORE = i'.CF \land i.STIPENDIO > 40000\}$$

Alcuni esempi - 4

Interrogazione 5. Selezionare il codice fiscale degli impiegati che guadagnano più del loro supervisore.

$\{i.CF \mid i(IMPIEGATO), i'(IMPIEGATO) \mid$

$i.SUPERVISORE = i'.CF \wedge i.STIPENDIO >$
$i'.STIPENDIO\}$

Alcuni esempi - 5

Interrogazione 6. Trovare i dipartimenti i cui impiegati guadagnano tutti più di 40000 euro.

$\{d.DNUMERO \mid d(DIPARTIMENTO) \mid$

$\neg(\exists i(IMPIEGATO)(i.DIP = d.DNUMERO \wedge$

$i.STIPENDIO \leq 40000))\}$

In alternativa, si può utilizzare la quantificazione universale (consentita nel calcolo relazionale).

Alcuni esempi - 6

Interrogazione 6. (soluzione alternativa) Trovare i dipartimenti i cui impiegati guadagnano tutti più di 40000 euro.

$\{d.DNUMERO \mid d(DIPARTIMENTO) \mid$

$\forall i(IMPIEGATO)(\neg(i.DIP = d.DNUMERO) \lor$

$i.STIPENDIO > 40000))\}$

Limiti del calcolo

Il calcolo relazionale su tuple con dichiarazioni di range non permette di esprimere tutte le interrogazioni esprimibili in algebra relazionale (e nel calcolo relazionale su domini).

Non possono essere espresse interrogazioni i cui risultati possono provenire indifferentemente da due e più relazioni (interrogazioni che in algebra relazionale utilizzano l'*operatore di unione*): i risultati sono costruiti a partire dalle variabili libere e ogni variabile ha come range una sola relazione.

Soluzione (adottata anche in SQL): introdurre un costrutto esplicito di unione.

Problemi con l'unione

Se consentissimo di associare ad una variabile un range costituito da più relazioni, risolveremmo il problema per la *semplice unione* di due o più relazioni. Ad esempio, date due relazioni aventi lo stesso schema $R_1(A, B), R_2(A, B)$, l'interrogazione (formulata in algebra relazionale come) $R_1(A, B) \cup R_2(A, B)$ si potrebbe esprimere come:

$$\{i.* \mid i(R_1, R_2) \mid true\}$$

Non riusciremmo, comunque, a formulare *unioni complesse* i cui operandi siano sottoespressioni che non corrispondono direttamente ad alcuno schema di relazione. E' questo il caso, ad esempio, dell'interrogazione (espressa in algebra relazionale) $\Pi_{BC}(R_1) \cup \Pi_{BC}(R_2)$, che coinvolge le relazioni $R_1(A, B, C)$, $R_2(B, C, D)$

Nessun problema con intersezione e differenza

A differenza dell'operatore di unione, gli operatori di intersezione e differenza insiemistica possono essere agevolmente espressi nel calcolo relazionale delle tuple con dichiarazioni di range.

Date le relazioni $R_1(A, B, C)$, $R_2(B, C, D)$,

l'interrogazione $\Pi_{BC}(R_1) \cap \Pi_{BC}(R_2)$ si può esprimere come

$\{i.(B, C) \mid i(R_1) \mid \exists j(R_2)(i.B = j.B \wedge i.C = j.C)\}$

mentre l'interrogazione $\Pi_{BC}(R_1) \setminus \Pi_{BC}(R_2)$ si può esprimere come

$\{i.(B, C) \mid i(R_1) \mid \neg\exists j(R_2)(i.B = j.B \wedge i.C = j.C)\}$