Fondamenti di Teoria delle Basi di Dati

Riccardo Torlone

Parte 6: Potenza espressiva del calcolo

Calcolo su domini, discussione

- Pregi:
  - dichiaratività
- Difetti:
  - "verbosità": tante variabili!
  - espressioni senza senso:

$$\{ A: x, B: y \mid R(A: x) \}$$

$$\{ A: x, B: y \mid R(A: x) \wedge y=y \}$$

$$\{ A: x \mid \neg R(A: x) \}$$

  queste espressioni sono "dipendenti dal dominio" e vorremmo evitarle;
  - nell'algebra espressioni come queste non sono formulabili: l'algebra è indipendente dal dominio

Indipendenza dal dominio

- Notazione:
  - $E^D(\mathbf{r})$ valore di E su **r** rispetto al dominio D.
- Un dominio è compatibile con un'espressione E e un'istanza **r** se contiene i valori in **r** e le costanti in E.
- Un'espressione E è indipendente dal dominio se, per ogni schema **R**, per ogni istanza **r** di **R** e per ogni coppia di domini D e D' compatibili con E e **r**, risulta $E^D(\mathbf{r}) = E^{D'}(\mathbf{r})$
- Un linguaggio è indipendente dal dominio se tutte le sue espressioni sono indipendenti dal dominio
- Il requisito dell'indipendenza dal dominio è fondamentale per i linguaggi reali.

Indipendenza dal dominio ed espressività

- Il calcolo relazionale su domini (DRC) non è indipendente dal dominio
- **Lemma** Verificare che un’espressione DRC è indipendente dal dominio è un problema indecidibile.
- **Lemma** L’algebra relazionale è indipendente dal dominio
  - **Proof** E’ diretta conseguenza del fatto che l’algebra costruisce i risultati a partire dalle relazioni date
- L’algebra relazionale è quindi meno espressiva del DRC
- La maggiore espressività è poco significativa
- DI-DRC è il linguaggio che contiene solo espressioni DRC indipendenti dal dominio

Calcolo e algebra

- Calcolo e algebra sono equivalenti
  - per ogni espressione del calcolo relazionale che sia indipendente dal dominio esiste un'espressione dell'algebra relazionale equivalente a essa
  - per ogni espressione dell'algebra relazionale esiste un'espressione del calcolo relazionale equivalente a essa (e di conseguenza indipendente dal dominio)

Valutazione limitata

- Fissiamo un'espressione E e un'istanza $\mathbf{r}$.
- Dominio attivo di E e $\mathbf{r}$:

$$D_{E,\mathbf{r}} = D_{\mathbf{r}} \cup C$$

- dove:
  - $D_{\mathbf{r}}$ insieme dei valori del dominio che compaiono in $\mathbf{r}$;
  - C insieme delle costanti in E
- Valore $E_{\mathrm{LE}}(\mathbf{r})$ di un'espressione del DRC secondo la valutazione limitata
  - le sostituzioni sono funzioni da V a $D_{\mathbf{E},\mathbf{r}}$
- Risulta $E_{\mathrm{LE}}(\mathbf{r}) = E^{D_{E,\mathbf{r}}}(\mathbf{r})$

Indipendenza e valutazione limitata

- **Lemma** Per le espressioni indipendenti dal dominio, la valutazione classica e quella limitata producono lo stesso valore: $E_{LE}(\mathbf{r}) = E(\mathbf{r})$.
- **Proof** Se E è indipendente dal dominio, allora $E(\mathbf{r}) = E^{D}(\mathbf{r})$, per ogni possibile dominio D. Il dominio attivo $D_{E,r}$ è un possibile domini e quindi $E(\mathbf{r}) = E^{D_{E,r}}(\mathbf{r})$. Poiché $E_{LE}(\mathbf{r}) = E^{D_{E,r}}(\mathbf{r})$, abbiamo $E_{LE}(\mathbf{r}) = E(\mathbf{r})$

- Proprietà “tecnica”:
  - **Lemma** Per ogni espressione E del DRC, ogni schema di base di dati **R** ed ogni nome di attributo A, esiste un’espressione E_A dell’algebra che calcola, per ogni **r** su **R**, il dominio attivo $D_{E,r}$, producendo una relazione su A.

Equivalenza tra calcolo e algebra (1 parte)

- **Lemma** Per ogni espressione del DI-DRC esiste un'espressione dell'algebra equivalente ad essa.
- **Proof**

$$\text{Sia } E_{\mathrm{DRC}} = \{ A_1 : x_1, \ldots, A_k : x_k \mid f \}$$

- Assunzioni (senza perdita di generalità):
  - le variabili libere di f sono $x_1, \ldots, x_k$
  - ogni variabile quantificata appare libera nella corrispondente sottoformula
  - f non contiene disgiunzioni né quantificatori universali
- Per induzione completa sul numero di connettivi e quantificatori, costruiamo un'espressione dell'algebra equivalente a

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

- Assumendo E_{DRC} indipendente dal dominio, consideriamo la valutazione limitata.

Proof (continua)

Indichiamo con E_x l'espressione che produce una relazione, sull'attributo x che presenta, per ogni istanza, tutti i valori del dominio attivo.

Base: 0 connettivi (formule atomiche).

- $f = R(A_1 : x_1, \ldots, A_k : x_k)$.
  L'espressione $\rho_{x_1,\ldots,x_k \leftarrow A_1,\ldots,A_k}(R)$ è equivalente a
  $$E_{\mathsf{DRC}} = \{x_1 : x_1, \ldots, x_k : x_k \mid R(A_1 : x_1, \ldots, A_k : x_k)\}$$

- $f = x_1 \theta x_2$.
  $\sigma_{x_1 \theta x_2}(E_{x_1} \bowtie E_{x_2})$ è equivalente a
  $$E_{\mathsf{DRC}} = \{x_1 : x_1, x_2 : x_2 \mid x_1 \theta x_2\}$$

- $f = x_1 \theta c$.
  $\sigma_{x_1 \theta c}(E_{x_1})$ è equivalente a
  $$E_{\mathsf{DRC}} = \{x_1 : x_1 \mid x_1 \theta c\}$$

Proof (continua)

Induzione

Tre casi, sulla base del connettivo o quantificatore più esterno.

- $f = f_1 \wedge f_2$.
  E_{f_1} e E_{f_2} corrispondono a f_1 e f_2.
  Intuitivamente, E_f è l'intersezione di E_{f_1} e E_{f_2}.
  Però, f_1 e f_2 possono non avere le stesse variabili libere e quindi E_{f_1} e E_{f_2} non sono definite sullo stesso schema.
  Se V_1 e V_2 sono le variabili in f_1 e f_2 risp.,
  e $V_2 - V_1 = \{y_1, \ldots, y_h\}$ e $V_1 - V_2 = \{z_1, \ldots, z_k\}$,
  "estendiamo" E_{f_1} e E_{f_2} allo stesso schema:
  $E_1 = E_{f_1} \bowtie E_{y_1} \bowtie \ldots \bowtie E_{y_h}$
  $E_2 = E_{f_2} \bowtie E_{z_1} \bowtie \ldots \bowtie E_{z_k}$ e quindi:
  $E_f = E_1 \cap E_2$, che è equivalente a $E_{f_1} \bowtie E_{f_2}$.

Proof (continua)

- $f = \neg(f_1)$,
  con $x_1, \ldots, x_n$ variabili libere in f_1.
  E_{f_1} espressione algebrica equivalente a
  $\{x_1 : x_1, \ldots, x_n : x_n \mid f_1\}$ (ipotesi induttiva).
  Allora $(E_{x_1} \bowtie \ldots E_{x_n}) - E_{f_1}$ è equivalente a
  $\{x_1 : x_1, \ldots, x_n : x_n \mid \neg(f_1)\}$.
- $f = \exists x(f_1)$,
  con x e $x_1, \ldots, x_n$ variabili libere in f_1;
  E_{f_1} corrispondente a f_1.
  Allora, $E_f = \pi_{x_1 \ldots x_n}(E_{f_1})$.

L'induzione è completata.

Infine, se E_f è equivalente a $\{x_1 : x_1, \ldots, x_k : x_k \mid f\}$ allora $\rho_{A_1,\ldots,A_k \leftarrow x_1,\ldots,x_k}(E_f)$ è equivalente a $\{A_1 : x_1, \ldots, A_k : x_k \mid f\}$ $\diamond$

Equivalenza tra calcolo e algebra (2 parte)

- **Lemma** Per ogni espressione dell'algebra esiste un'espressione del DI-DRC equivalente ad essa.
- **Proof**
- Sia E_{RA} una espressione dell'algebra.
- Assunzioni (senza perdita di generalità):
  - intersezioni non compaiono
  - i join naturali sono prodotti cartesiani
  - le condizioni di selezione sono atomiche
  - ogni proiezione "elimina" un solo attributo
- Per induzione completa sul numero di operatori, costruiamo un'espressione del DRC equivalente a E_{RA}. Essa è DI perché equivalente a E_{DRC}.

Proof (continua)

Base: nessun operatore in E_{RA}.

- Relazione della base di dati:

$$E_{\mathsf{RA}} = R$$

con $R(A_1 \ldots A_k) \in \mathbf{R}$

E_{RA} è equivalente a

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_k : x_k \mid R(A_1 : x_1 \ldots A_k : x_k)\}$$

Proof (continua)

Induzione

Sei casi, sulla base dell'operatore più esterno.

Per gli operatori monadici, sia E'_{RA} equivalente a

$$E'_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi\}$$

- Proiezione: $E_{\mathsf{RA}} = \pi_{A_1 \ldots A_{n-1}}(E'_{\mathsf{RA}})$.

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_{n-1} : x_{n-1} \mid \exists x_n(\phi)\}$$

- Ridenominazione: $E_{\mathsf{RA}} = \rho_f(E'_{\mathsf{RA}})$.

$$E_{\mathsf{DRC}} = \{f(A_1) \ldots f(A_n) : x_n \mid \phi\}$$

- Selezione: $E_{\mathsf{RA}} = \sigma_{A_j \theta A_k}(E'_{\mathsf{RA}})$ o $E_{\mathsf{RA}} = \sigma_{A_j \theta c}(E'_{\mathsf{RA}})$

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi \wedge x_j \theta x_k\}$$

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi \wedge x_j \theta c\}$$

Proof (continua)

Operatori binari.

- Unione: $E_{\mathsf{RA}} = E'_{\mathsf{RA}} \cup E''_{\mathsf{RA}}$.
  Le due espressioni DRC equivalenti a E'_{RA} e E''_{RA} possono avere diversi insiemi di variabili libere:

$$E'_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi'\}$$

$$E''_{\mathsf{DRC}} = \{A_1 : y_1 \ldots A_n : y_n \mid \phi''\}$$

  È però possibile uniformare tali insiemi:

$$E'''_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi'''\}$$

  e quindi avere

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi' \vee \phi'''\}$$

Proof (continua)

- Differenza: $E_{\mathsf{RA}} = E'_{\mathsf{RA}} - E''_{\mathsf{RA}}$. Simile all'unione

$$E_{\mathsf{DRC}} = \{A_1 : x_1 \ldots A_n : x_n \mid \phi' \wedge \neg(\phi''')\}$$

- Prodotto cartesiano: $E_{\mathsf{RA}} = E'_{\mathsf{RA}} \bowtie E''_{\mathsf{RA}}$.
  E'_{RA} e E''_{RA} sono definite su insiemi di attributi disgiunti. Si può assumere che lo stesso valga per le variabili libere in

$$E'_{\mathsf{DRC}} = \{A_1 : x_1, \ldots, A_n : x_n \mid \phi'\}$$

$$E''_{\mathsf{DRC}} = \{B_1 : y_1, \ldots, B_n : y_n \mid \phi''\}$$

  Quindi:

$$E_{\mathsf{DRC}} = \{A_1 : x_1, \ldots, A_n : x_n, B_1 : y_1, \ldots, B_n : y_n \mid \phi' \wedge \phi''\}$$

$\diamond$

Calcolo su tuple con dich. di range

- Per superare le limitazioni del calcolo su domini:
  - Verbosità: bisogna "ridurre" le variabili
    - idea: una variabile per ciascuna tupla
  - Indipendenza dal dominio: bisogna far riferimento alla valutazione limitata
    - idea: associare le variabili direttamente a (sottoinsiemi del) dominio attivo
- Introduciamo allora il calcolo su tuple con dichiarazioni di range

Calcolo su tuple con dich. di range

- In TRC-RD le espressioni hanno la forma:

  { TargetList | RangeList | Formula }

- Le variabili denotano tuple
- RangeList elenca le variabili libere della Formula ognuna con il relativo campo di variabilità (una relazione)
- TargetList ha elementi del tipo Y: x.Z (oppure x.Z o anche x.*)
- Formula ha:
  - atomi di confronto x.A ϴ c, x.A ϴ y.B
  - connettivi
  - quantificatori che associano un range alle variabili
    ∃x(R)(...) ∀x(R)(...)

Base di dati per gli esempi

Impiegati(<u>Matricola</u>,Nome, Età, Stipendio)

Supervisione(Capo, <u>Impiegato</u>)

Impiegati

Matricola	Nome	Età	Stipendio
7309	Rossi	26	55
5998	Neri	34	64
9553	Milano	47	44
5698	Neri	52	74

Supervisione

Capo	Impiegato
5998	7309
5698	9553

Esempio 1

- Trovare matricola, nome ed età degli impiegati che guadagnano più di 40 mila

$$\pi_{\text{Matricola, Nome, Età}}(\sigma_{\text{Stipendio>40}}(\text{Impiegati}))$$

{ Matricola: m, Nome: n, Età: e |
Impiegati(Matricola: m, Nome: n, Età: e, Stipendio: s) $\wedge$
s > 40 }

{ i.(Matricola,Nome,Età) | i(Impiegati) | i.Stipendio > 40 }

Esempio 2

- Trovare le matricole dei capi degli impiegati che guadagnano più di 40 mila

{ Capo: c | Supervisione(Capo:c,Impiegato:m) $\wedge$ Impiegati(Matricola: m, Nome: n, Età: e, Stipendio: s) $\wedge$ s > 40 }

{ s.Capo | i(Impiegati) , s(Supervisione) | i.Matricola=s.Impiegato $\wedge$ i.Stipendio > 40 }

Verso un linguaggio reale

- Prendiamo l'espressione:

  { s.Capo | i(Impiegati) , s(Supervisione) |
  i.Matricola=s.Impiegato $\wedge$ i.Stipendio > 40 }

- Vi ricorda qualcosa?

  SELECT s.Capo
  FROM Impiegati AS i , Supervisione AS s
  WHERE i.Matricola=s.Impiegato
  AND i.Stipendio > 40

- SQL base è “sostanzialmente” equivalente al calcolo su tuple con dichiarazione di range

Esempio 3

- Trovare nome e stipendio dei capi degli impiegati che guadagnano più di 40 mila

$$\{\ \text{NomeC: nc, StipC: sc}\ |$$
$$\text{Impiegati(Matricola: m, Nome: n, Età: e, Stipendio: s)} \wedge$$
$$s > 40\ \wedge \text{Supervisione(Capo:c,Impiegato:m)} \wedge$$
$$\text{Impiegati(Matricola:c, Nome:nc, Età:ec, Stipendio:sc)}\ \}$$

$$\{\ \text{NomeC,StipC: i'.(Nome,Stip)}\ |$$
$$\text{i'(Impiegati), s(Supervisione), i(Impiegati)}\ |$$
$$\text{i'.Matricola=s.Capo} \wedge \text{i.Matricola=s.Impiegato} \wedge$$
$$\text{i.Stipendio} > 40\ \}$$

Esempio 4

- Trovare gli impiegati che guadagnano più del rispettivo capo, mostrando matricola, nome e stipendio di ciascuno di essi e del capo

{ Matr: m, Nome: n, Stip: s, NomeC: nc, StipC: sc |
Impiegati(Matricola: m, Nome: n, Età: e, Stipendio: s) $\wedge$
Supervisione(Capo:c,Impiegato:m) $\wedge$
Impiegati(Matricola: c, Nome: nc, Età: ec, Stipendio: sc) $\wedge$
s > sc}

{ i.(Nome,Matr,Stip), NomeC,MatrC,StipC: i'.(Nome,Matr,Stip) |
i'(Impiegati), s(Supervisione), i(Impiegati) |
i'.Matricola=s.Capo $\wedge$ i.Matricola=s.Impiegato $\wedge$
i.Stipendio > i'.Stipendio }

Esempio 5

- Trovare matricola e nome dei capi i cui impiegati guadagnano tutti più di 40 mila.

$$\{\text{Matricola: } c, \text{ Nome: } n \mid$$
$$\text{Impiegati(Matricola: } c, \text{ Nome: } n, \text{ Età: } e, \text{ Stipendio: } s) \wedge$$
$$\text{Supervisione(Capo:}c, \text{ Impiegato:}m) \wedge$$
$$\neg\, \exists m'(\exists n'(\exists e'(\exists s'(\text{Impiegati(Matr: } m', \text{ Nome: } n', \text{ Età: } e', \text{ Stip: } s') \wedge$$
$$\text{Supervisione(Capo:}c, \text{ Impiegato:}m') \wedge s' \leq 40\}$$

$$\{\ i.(\text{Matricola, Nome}) \mid s(\text{Supervisione}), i(\text{Impiegati}) \mid$$
$$i.\text{Matricola}=s.\text{Capo} \wedge \neg(\exists i'(\text{Impiegati})(\exists s'(\text{Supervisione})$$
$$(s.\text{Capo}=s'.\text{Capo} \wedge s'.\text{Impiegato}=i'.\text{Matricola} \wedge$$
$$i'.\text{Stipendio} \leq 40)))\}$$

Attenzione!

- Il calcolo su tuple con dichiarazioni di range non permette di esprimere alcune interrogazioni importanti, in particolare le unioni:

$$R_1(AB) \cup R_2(AB)$$

- Quale potrebbe essere il range per una variabile? Oppure due variabili?
- Per questa ragione SQL prevede un operatore esplicito di unione, ma non tutte le versioni prevedono intersezione e differenza

Relazione tra linguaggi