Databázové technologie

Dotazovací systémy v relačním modelu (dokončení)

Petr Krajča

Katedra informatiky
Univerzita Palackého v Olomouci

Nekonečná relace

- uvažujme relaci $usedon$ s relačním schématem $R_{usedon} = \{\texttt{partno}, \texttt{ptype}, \texttt{nused}\}$
- pokud některá z domén atributů je nekonečná, pak např. výraz

$$\{x(R_{usedon}) \mid \neg usedon(x) \vee x(\texttt{ptype}) \neq \text{"707"}\}$$

- představuje relaci, která je nekonečná
- řešení: omezení na bezpečné výrazy n-ticového relačního kalkulu

Rozšířená aktivní doména

- $edom(y, f) \subseteq D_y$
- množina všech hodnot z domény D_y, které se vyskytují v relaci použité v f nebo jako konstanty
- pro relační schéma $R \subseteq Y$ budeme značit

$$edom(R, f) = \Pi_{y \in R}\, edom(y, f)$$

Bezpečné výrazy n-ticového rel. kalkulu

Výraz $\{x(R) \mid f(x)\}$ je bezpečný, pokud platí

1. $I(f(t/x)) = true$, pak $t \in edom(R, f)$.
2. Pro každou podformuli formule f ve tvaru $\exists y(S)g(y, z_1, \ldots, z_k)$, z $I(g(t/y, u_1/z_1, \ldots, u_n/z_n)) = true$ plyne $t \in edom(S, g)$.
3. Pro každou podformuli formule f ve tvaru $\forall y(S)g(y, z_1, \ldots, z_k)$ a $t \notin edom(S, g)$ platí $I(g(t/y, u_1/z_1, \ldots, u_n/z_n)) = true$.

za předpokladu, že $y, z_1, \ldots z_n$ jsou volné proměnné v g.

- bod (1) zajišťuje, že hodnota výrazu bude konečná relace
- body (2) a (3) zajišťují, že podvýrazy s kvantifikátory půjde efektivně spočítat.

Interpretace bezpečných výrazů

- analogické jako pro neomezené výrazy, pro f, které je dáno jako

1. $\exists x(R)g$, x je jediná volná proměnná v g. $I(f) = true$ p.k. existuje alespoň jedno $t \in edom(R, g)$ takové, že $I(g(t/x)) = true$, jinak $I(f) = false$.
2. $\forall x(R)g$, x je jediná volná proměnná v g. $I(f) = true$ p.k. pro všechny $t \in edom(R, g)$ platí, $I(g(t/x)) = true$, jinak $I(f) = false$.

Redukce relační algebry na n-ticový relační kalkulus (1/2)

Konstantní relace

- RA: $\{\langle y, a\rangle\}$
- NRK: $\{t(y) \mid t(y) = a\}$

Relace

- RA: $\mathcal{D}$ nad relačním schématem R
- NRK: $\{t(R) \mid \mathcal{D}(t)\}$

Restrikce

- RA: $\sigma_{y\theta c}(E)$ nebo $\sigma_{y_1\theta y_2}(E)$
- NRK: Pokud $\{t(R) \mid f(t)\}$ odpovídá E:
  $\{t(R) \mid f(t) \wedge t(y)\theta c\}$ nebo $\{t(R) \mid f(t) \wedge t(y_1)\theta t(y_2)\}$

Projekce

- RA: $\pi_S(E)$
- NRK: Pokud $\{t(R) \mid f(t)\}$ odpovídá E:
  $\{t(S) \mid f(t)\}$

Redukce relační algebry na n-ticový relační kalkulus (2/2)

Množinové operace

- RA: $E_1 \cap E_2$
- NRK: Pokud $\{t(R) \mid f(t)\}$ odpovídá E_1 a $\{t(R) \mid g(t)\}$ odpovídá E_2: $\{t(R) \mid f(t) \wedge g(t)\}$
- ostatní operace analogicky

Přirozené spojení

- RA: $E_1 \bowtie E_2$ nad relačními schématy RS a ST
- NRK: Pokud $\{t(RS) \mid f(t)\}$ odpovídá E_1 a $\{u(ST) \mid g(u)\}$ odpovídá E_2:

$$\{z(RST) \mid \exists x(RS)\exists y(ST)(f(x) \wedge g(y) \wedge z(RS) = x(RS) \wedge z(ST) = y(ST)\}$$

Doménový relační kalkulus

- velmi podobný n-ticovému rel. kalkulu
- proměnné nabývají hodnot jednotlivých domén (nikoliv ntic)
- formule se skládají z doménových proměnných, relačních proměnných, spojek a pomocných symbolů (závorek)

Příklady

- $\{x\ y \mid parts(x, y, z) \wedge y = 206\}$
- $\{x\ y \mid parts(x, 206, z)\}$

Formule doménového relačního kalkulu

atomy

1. $r(y_1, \ldots, y_n)$, kde r je relační proměnná nad schématem $\{y_1, \ldots, y_n\}$
2. $x\theta y$, $x\theta c$, $c\theta y$, kde x, y jsou doménové proměnné a c konstanta
3. $true$, $false$ (konstanty pravdivostních hodnot)

formule

1. každý atom je formule
2. pokud f je formule, pak $\neg f$ je formule
3. pokud f a g jsou formule, pak $f \vee g$ a $f \wedge g$ jsou formule
4. pokud f je formule, pak $\exists x(y) f$ kde $y \in Y$ a x je doménová proměnná, je formule
5. pokud f je formule, pak $\forall x(y) f$ kde $y \in Y$ a x je doménová proměnná, je formule
6. pokud f je formule, pak (f) je formule

Formule doménového relačního kalkulu (poznámky)

- pro volné a vázané proměnné platí podobné pravidla jako v n-ticovém relačním kalkulu
- typ doménové proměnné je buď konkrétní doména nebo je nedefinován
- pro validní formule je potřeba, aby použití typů bylo konzistentní (nebudeme formalizovat)
- substituce doménové konstanty c za doménovou proměnnou $f(c/x)$ je analogická n-ticovému kalkulu

Výrazy doménového kalkulu

Výraz doménového kalkulu má tvar:

$$\{x_1(y_1)\ x_2(y_2)\ \ldots\ x_n(y_n) \mid f(x_1, x_2, \ldots, x_n)\}$$

- f je validní formule doménového kalkulu, která má právě volné proměnné $x_1, x_2, \ldots, x_n$
- $y_1, y_2, \ldots, y_n$ jsou atributy z Y
- typy proměnných odpovídají typům atributů

hodnota výrazu

Je relace nad schématem $R = \{y_1, y_2, \ldots, y_n\}$, která obsahuje všechny n-tice $\langle c_1, c_2, \ldots, c_n\rangle$ takové, že $c_i \in D_i$ a

$$I(f(c_1/x_1, c_2/x_2, \ldots, c_n/x_n)) = true$$

Redukce n-ticového kalkulu na doménový

- přímočará
- ve výrazu $\{x(R) \mid f(x)\}$ má každá n-ticová proměnná z přiřazené unikátní schéma S. Buď se z vyskytuje v podformuli $\exists z(S)$ nebo $\forall z(S)$, nebo $z = x$ a pak $S = R$.
- n-ticová proměnná t je nahrazena doménovými proměnnými $t_1, \ldots, t_n$
- atom $r(t)$ je nahrazen $r(t_1, \ldots, t_n)$
- atomy ve tvaru $t(y_i)\ \theta\ c$ jsou nahrazeny $t_i\ \theta\ c$
- podformule $\exists t(S) g$ je nahrazena $\exists t_1(y_1) \exists t_2(y_2) \ldots \exists t_n(y_n)\ g$
- analogicky pro $\forall$
- analogicky první část výrazu $x(R)$ je nahrazena $x_1(y_1)\ x_2(y_2)\ \ldots\ x_n(y_n)$

(Odbočka) Relační dělení: Motivace

dotazy ve tvaru „některá A jsou B"

- *Studenti, kteří mají zapsaný nějaký předmět.*
- *Součástky, které patří do některého výrobku.*
- *Dodavatelé, kteří dodávají některé výrobky.*
- ...

dotazy ve tvaru „všechna A jsou B"

- *Studenti, kteří mají splněny všechny vyžadované předměty.*
- *Výrobky, jejichž všechny součástky jsou skladem.*
- *Dodavatelé, kteří dodávají všechny výrobky.*
- ...

Vyjadřování dotazů **existenčního** a **všeobecného charakteru**:

- existenční – polospojení (projekce a přirozené spojení)
- všeobecný – dělení (množinový rozdíl, přirozené spojení, projekce)

Relační dělení

Uvažujme následující relace: $\mathcal{D}_1$ (dělenec) na schématu R, $\mathcal{D}_2$ (dělitel) na schématu S a $\mathcal{D}_3$ (prostředník) na schématu T. Pak relace

$$\mathcal{D}_1 \div_{\mathcal{D}_3} \mathcal{D}_2 = \{r(R \cap T) | r \in \mathcal{D}_1 \text{ tak, že pro každou } s \in \mathcal{D}_2 \text{ splňující podmínku } \\ r(R \cap S \cap T) = s(R \cap S \cap T) \text{ platí, že } r(R \cap T)s(S \cap T) \in \pi_{(R \cup S) \cap T}(\mathcal{D}_3)\}$$

se nazývá **podíl** $\mathcal{D}_1$ a $\mathcal{D}_2$ přes $\mathcal{D}_3$ na schématu $R \cap T$.

Interpretace pro $\mathcal{D}_1$ (výrobci), $\mathcal{D}_2$ (výrobek), $\mathcal{D}_3$ a (kdo vyrábí co):

- $\mathcal{D}_1 \div_{\mathcal{D}_3} \mathcal{D}_2$ – výrobci vyrábějící všechny výrobky
- ...ještě jinak: r z dělence (výrobce) je ve výsledném podílu právě tehdy, pokud pro každou s (výrobek) z dělitele platí, že rs je v prostředníkovi (výrobce-výrobek), tj. výrobce vyrábí každý výrobek.
- $\mathcal{D}_2 \div_{\mathcal{D}_3} \mathcal{D}_1$ – výrobky vyráběné všemi výrobci

Příklad dat pro relační dělení

NAME
Abbe
Blangis
Curval
Durcet

COURSE
KMI/DATA1
KMI/DATA2
KMI/PAPR1

NAME	COURSE
Abbe	KMI/DATA1
Abbe	KMI/DATA2
Blangis	KMI/DATA1
Blangis	KMI/DATA2
Blangis	KMI/PAPR1
Curval	KMI/DATA1
Curval	KMI/DATA2
Curval	KMI/PAPR1

NAME	MAJOR
Abbe	CS
Blangis	CS
Curval	EE
Durcet	SS

COURSE	VERSION
KMI/DATA1	2
KMI/DATA2	1
KMI/PAPR1	2

NAME	COURSE	YEAR
Abbe	KMI/DATA1	2011
Abbe	KMI/DATA2	2012
Blangis	KMI/DATA1	2012
Blangis	KMI/DATA2	2012
Blangis	KMI/PAPR1	2007
Curval	KMI/DATA1	2011
Curval	KMI/DATA2	2013
Curval	KMI/PAPR1	2008

Vyjádření $\div$ pomocí $\setminus, \pi$ a $\bowtie$

Pro $\mathcal{D}_1$ na R, $\mathcal{D}_2$ na S, $\mathcal{D}_3$ na T platí:

$$\mathcal{D}_1 \div_{\mathcal{D}_3} \mathcal{D}_2 = \pi_{R \cap T}(\mathcal{D}_1) \setminus \pi_{R \cap T}\left((\pi_{R \cap T}(\mathcal{D}_1) \bowtie \pi_{S \cap T}(\mathcal{D}_2)) \setminus \pi_{(R \cup S) \cap T}(\mathcal{D}_3)\right).$$

Neformální skica důkazu

a $\mathcal{D}_a = \pi_{R \cap T}(\mathcal{D}_1) \bowtie \pi_{S \cap T}(\mathcal{D}_2)$ – všechny kombinace hodnot, které můžeme uvažovat v $\mathcal{D}_3$

b $\mathcal{D}_b = (\mathcal{D}_a \setminus \pi_{(R \cup S) \cap T}(\mathcal{D}_3))$ – všechny n-tice, které mohly být v $\mathcal{D}_3$, ale nejsou.

c $\mathcal{D}_c = \pi_{R \cap T}(\mathcal{D}_b)$ – všechny n-tice z $\mathcal{D}_1$, které nemají v $\mathcal{D}_3$ **nějakou** spojitelnou n-tici.

d $\mathcal{D}_d = \pi_{R \cap T}(\mathcal{D}_1) \setminus \mathcal{D}_c$ – ponechá n-tice $\mathcal{D}_1$, kterým nechybí spojitelná n-tice v $\mathcal{D}_3$.

Redukce doménového kalkulu na relační algebru (1/4)

Pro každou podformuli g budeme rekurzivně definovat výraz relační algebry F_g, který bude ekvivalentní výrazu

$$\{x_1\ x_2\ \ldots\ x_n \mid g(x_1, x_2, \ldots, x_n)\}$$

atomy

Pokud je g atom ve tvaru $u\ \theta\ v$, $u\ \theta\ u$, $u\ \theta\ c$ nebo $c\ \theta\ u$, pak odpovídající algebraické výrazy jsou

$$\sigma_{A\theta B}([A] \times [B])$$
$$\sigma_{A\theta A}([A])$$
$$\sigma_{A\theta c}([A])$$
$$\sigma_{c\theta A}([A])$$

Kde $[A]$ označuje relaci se stupněm jedna obsahující všechny hodnoty z domény D_A, kde A označuje atribut příslušný u nebo v.

Redukce doménového kalkulu na relační algebru (2/4)

Pokud je g atom ve tvaru $r(a_1, \ldots, a_n)$, kde a_i je konstanta nebo doménová proměnná a $R = \{y_1, \ldots, y_n\}$ je relační schéma r, pak odpovídající algebraický výraz F_g je tvaru

$$\pi_X(\rho_N(\sigma_C(r)))$$

- C je konjunkce porovnání na rovnost pro každé a_i, které je konstantou
- N je přejmenování, které každé doménové proměnné přiřadí odpovídající atribut y_i'
- X je množina atributů $\{y_i' \mid \text{pro kterou je } a_i \text{ proměnná}\}$

Negace
Pokud g je podformule $\neg h$ a F_h je alg. výraz odpovídající h, pak

$$F_g = \bar{F_h}.$$

Redukce doménového kalkulu na relační algebru (3/4)

binární operace

Podformule g je tvaru $h_1 \wedge h_2$, kde

- h_1 obsahuje volné proměnné $z_1, \ldots z_k$ a $v_1, \ldots, v_p$
- h_2 obsahuje volné proměnné $z_1, \ldots z_k$ a $w_1, \ldots, w_q$
- F_{h_1} a F_{h_2} jsou algebraické výrazy odpovídající h_1 a h_2, pak

$$
\begin{aligned}
F_1 &= F_{h_1} \bowtie [C_1] \bowtie \cdots \bowtie [C_q] \\
F_2 &= F_{h_2} \bowtie [B_1] \bowtie \cdots \bowtie [B_p] \\
F_g &= F_1 \cap F_2
\end{aligned}
$$

- $[C_1], \ldots, [C_q]$ odpovídají relačním proměnným $w_1, \ldots, w_q$
- $[B_1], \ldots, [B_p]$ odpovídají relačním proměnným $v_1, \ldots, v_p$
- smyslem je doplnit relační schémata, aby byla kompatibilní
- analogicky pro $\vee$

Redukce doménového kalkulu na relační algebru (4/4)

existenční kvantifikátor
Podformule g je tvaru $\exists x(y)h$ a F_h je alg. výraz na relačním schématu R pro h, pak odpovídající výraz má tvar

$$F_g = \pi_{R \setminus \{y\}}(F_h)$$

univerzální kvantifikátor
Podformule g je tvaru $\forall x(y)h$ a F_h je alg. výraz na relačním schématu R pro h, pak odpovídající výraz má tvar

$$F_g = F_h \div [y]$$

Shrnutí vzájemných vztahů

- ukázali jsme převoditelnost jednotlivých dotazovacích systémů
- ukazuje vztah mezi deklarativním a operativním zpracováním dotazů
- důležité pro studium vyjadřovací síly dotazovacích jazyků
- některé dotazy ale i tak neuchopitelné (např. hierarchická data)
- možné další aplikace

Query-by-Example (QBE)

Restrikce

$instock$:

partno	location	amount
	NYC	> 20

Spojení

$instock$:

partno	location	amount
_PN	NYC	

$parts$:

partno	supbartof	partname
_PN		bolt M3