Bazy danych 2022

na podstawie slajdów Przemysławy Kanarek

14 marca 2022

Przykład 1

Baza danych

$S = (\underline{indeks}, nazwisko, rok)$, $P = (\underline{nazwa}, typ)$, $O = (\underline{indeks}, \underline{przed}, \underline{data}, stop)$

- $x \in S \wedge (\exists y)(y \in O \wedge y.indeks = x.indeks \wedge y.stop = 5 \wedge y.przed =' BD')$
- $\{x \mid x \in S \wedge (\exists y)(y \in O \wedge y.indeks = x.indeks \wedge y.stop = 5 \wedge y.przed =' BD')\}$
- $\{z^{[nazwisko,indeks]} \mid$
  $(\exists x)(x \in S \wedge z.indeks = x.indeks \wedge z.nazwisko = x.nazwisko \wedge$
  $(\exists y)(y \in O \wedge y.indeks = x.indeks \wedge y.stop = 5 \wedge y.przed =' BD'))\}$

Znaczenie zapytań

- x - student, który dostał *5.0* z *BD*.
- Wszyscy studenci, którzy dostali *5.0* z *BD*.
- Nazwiska i indeksy studentów, którzy dostali *5.0* z *BD*.

Przykład 2

Baza danych

$S = (\underline{indeks}, nazwisko, rok)$, $P = (\underline{nazwa}, typ)$, $O = (\underline{indeks}, \underline{przed}, \underline{data}, stop)$

2a. $\{x \mid x \in S \land \neg(\exists y)(y \in O \land y.indeks = x.indeks \land y.stop = 5)\}$

2b. $\{x \mid x \in S \land \neg(\exists y)(y \in O \land y.indeks = x.indeks \land y.stop \neq 5)\}$

2c. $\{x \mid x \in S \land (\exists y)(y \in O \land y.indeks = x.indeks \land y.przed =' BD' \land$
$\neg(\exists y_1)(y_1 \in O \land y_1.indeks \neq x.indeks \land y_1.przed =' BD' \land y_1.stop > y.stop)\}$

Znaczenie zapytań

2a. Studenci, którzy nie dostali żadnej piątki.

2b. Studenci, którzy mają tylko piątki.

2c. Studenci, którzy mają najlepszą ocenę z Baz danych.

Zapytanie relacyjnego rachunku krotek

Formuła rrk — opisuje własności krotek

Formuła atomowa:

- $R(t)$ lub $t \in R$, gdzie R to relacja z bazy danych, a t to zmienna (krotkowa);
- $t.a = c$, gdzie a jest atrybutem t; równość można zastąpić przez: $\neq, <, \leq, >, \geq$, a c jest stałą lub atrybutem innej zmiennej krotkowej;

Formuła:

- formuła atomowa,
- (ϕ), $\neg(\phi)$, $\phi \vee \psi$, $\phi \wedge \psi$, gdzie ϕ i ψ są formułami;
- $(\exists t)(\phi(t))$ lub $(\forall t)(\phi(t))$, gdzie ϕ jest formułą, a t jej zmienną wolną.

Zapytanie rrk — wybiera krotki mające daną własność

$$\{x \mid \phi(x)\} \qquad \{x^{[A_1, A_2, \ldots, A_k]} \mid \phi(x)\},$$

gdzie x jest zmienną krotkową, a ϕ jest formułą relacyjnego rachunku krotek, w której x jest jedyną zmienną wolną;

Przykład 1

Baza danych

$S = (\underline{indeks}, nazwisko, rok)$, $P = (\underline{nazwa}, typ)$, $O = (\underline{indeks}, \underline{przed}, \underline{data}, stop)$

1. $S(x_1, x_2, x_3) \wedge (\exists y_1, y_2, y_3, y_4)((O(y_1, y_2, y_3, y_4) \wedge y_1 = x_1 \wedge y_2 =' BD' \wedge y_4 = 5)$

1a. $S(x_1, x_2, x_3) \wedge (\exists y_3)(O(x_1, ' BD', y_3, 5))$

2. $\{x_1, x_2, x_3 \mid S(x_1, x_2, x_3) \wedge (\exists y_3)(O(x_1, ' BD', y_3, 5))\}$

3. $\{x_1, x_2 \mid (\exists x_3)(S(x_1, x_2, x_3) \wedge (\exists y_3)(O(x_1, ' BD', y_3, 5)))\}$

3a. $\{ind, naz \mid (\exists rok)(S(ind, naz, rok) \wedge (\exists dat)(O(ind, ' BD', dat, 5)))\}$

Znaczenie zapytań

1. *x* - student, który dostał *5.0* z *BD*.
2. Wszyscy studenci, którzy dostali *5.0* z *BD*.
3. Nazwiska i indeksy studentów, którzy dostali *5.0* z *BD*.

Przykład 2

Baza danych

$S = (\underline{indeks}, nazwisko, rok)$, $P = (\underline{nazwa}, typ)$, $O = (\underline{indeks}, \underline{przed}, \underline{data}, stop)$

2a. $\{(i, n, r) \mid S(i, n, r) \wedge \neg(\exists p, d)(O(i, p, d, 5))\}$

2b. $\{(i, n, r) \mid S(i, n, r) \wedge \neg(\exists p, d, s)(O(i, p, d, s) \wedge s \neq 5)\}$

2c. $\{(i, n, r) \mid S(i, n, r) \wedge (\exists d, s)(O(i,' BD', d, s) \wedge \neg(\exists i_1, d_1, s_1)(O(i_1,' BD', d_1, s_1) \wedge i \neq i_1 \wedge s_1 > s))\}$

Znaczenie zapytań

2a. Studenci, którzy nie dostali żadnej piątki.

2b. Studenci, którzy mają tylko piątki.

2c. Studenci, którzy mają najlepszą ocenę z Baz danych.

Zapytanie relacyjnego rachunku dziedzin

Formuła rrd — opisuje własności wektorów zmiennych dziedzinowych

Formuła atomowa:

- $R(x_1, x_2, \ldots, x_k)$ lub $(x_1, x_2, \ldots, x_k) \in R$, gdzie R to relacja o arności k, $x_1, x_2, \ldots, x_k$ to zmienne lub stałe (dziedzinowe);
- $x = c$, gdzie x jest zmienną; równość można zastąpić przez: $\neq, <, \leq, >, \geq$, a c jest stałą lub zmienną;

Formuła:

- formuła atomowa,
- (ϕ), $\neg(\phi)$, $\phi \vee \psi$, $\phi \wedge \psi$, gdzie ϕ i ψ są formułami;
- $(\exists t_1, t_2, \ldots, t_\ell)(\phi(t_1, t_2, \ldots, t_\ell)$ lub $(\forall t_1, t_2, \ldots, t_\ell)(\phi(t_1, t_2, \ldots, t_\ell)$, gdzie ϕ jest formułą, a $t_1, t_2, \ldots, t_\ell$ jej zmiennymi wolnymi.

Zapytanie rrd — wybiera wektory wartości zmiennych mające daną własność

$$\{x_1, x_2, \ldots, x_\ell \mid \phi(y_1, y_2, \ldots, y_k)\},$$

gdzie ϕ jest formułą relacyjnego rachunku krotek, a $x_1, x_2, \ldots, x_\ell$ to zmienne dziedzinowe i stałe, wśród których występują wszystkie zmienne wolne ϕ: $y_1, y_2, \ldots, y_k$ i tylko takie zmienne;

Niezależność od dziedziny (domain independence)

Trudny przykład

$E(osoba, temat)$ to baza ekspertów z określonych tematów. Poszukujemy pary ekspertów, którzy razem tworzą zespół znający się na wszystkich tematach.

$\{a, b \mid (\forall d)((\exists c)(E(c, d)) \Rightarrow E(a, d) \vee E(b, d))\}$

$\{x \mid P(x) \wedge (\forall y)L(x, y)\}$

$\{x \mid \neg(\exists y)E(x, y)\}$

Co jest nie tak?

- Szukamy takich par (a, b), że dla każdego tematu d (czyli zmiennej występującej w drugiej kolumnie E) a lub b jest ekspertem w tym temacie (występuje w parze z d w relacji E).
- Jeśli nie ma tematów (E jest pusta), to każda para wartości $(?, ?)$ spełnia formułę zapytania.
- Jeśli jest jeden ekspert *Wszystkowiedzący* znający się na wszystkim, to każda para $(\text{'Wszystkowiedzący'}, ?)$ spełnia formułę zapytania.
- W obu przypadkach wynik zależy od przyjętego uniwersum!
- AAAAAAAAAA! Przecież to może być zbiór nieskończony!

Niezależność od dziedziny (domain independence)

Niech $\phi_B(\mathcal{A})$ oznacza zbiór krotek będący wynikiem zapytania ϕ obliczonego na bazie $\mathcal{A}$ przy założeniu, że dziedzina (zbiór możliwych wartości zmiennych) to B.

Dziedzina formuły

Zbiór D nazwiemy **dziedziną (aktywną)** formuły ϕ, gdy jest to zbiór wszystkich wartości występujących we wszystkich kolumnach wszystkich relacji wspomnianych w ϕ oraz wszystkich stałych występujących jawnie w ϕ.

Zapytanie ϕ jest *dziedzinowo niezależne* wtedy i tylko wtedy gdy nie istnieje:

- taka baza danych $\mathcal{A}$ oraz
- dwie dziedziny D_1 i D_2 zawierające dziedzinę formuły ϕ takie, że $\phi_{D_1}(\mathcal{A}) \neq \phi_{D_2}(\mathcal{A})$.

Niezależność od dziedziny (domain independence)

Formuły, które zależą od dziedziny są niebezpieczne!

Sprawdzenie czy formuła jest dziedzinowo niezależna jest nierozstrzygalne!

Czy zapytania algebry relacji są dziedzinowo niezależne?

Formuły bezpieczne

Niech D będzie dziedziną formuły ϕ.

Kwantyfikatory ograniczone

Formuła ϕ jest bezpieczna, jeśli poniższe modyfikacje nie wpływają na jej wartość:

- $(\exists x)\phi(x)$ można zamienić na $(\exists x \in D)\phi(x)$, czyli $(\exists x)D(x) \wedge \phi(x)$;
- $(\forall x)\phi(x)$ można zamienić na $(\forall x \in D)\phi(x)$, czyli $(\forall x)D(x) \Rightarrow \phi(x)$;
- $\{x \mid \phi(x)\}$ można zamienić na $\{x \in D^k \mid \phi(x)\}$, gdzie k jest arnością x, a D^k to iloczyn kartezjański k kopii D.

Co to ma wspólnego z tym, że formuła jest niezależna od dziedziny? Istnieją takie ograniczenia składni, które dają gwarancję bezpieczeństwa.

Twierdzenie

Języki zapytań dla modelu relacyjnego:

- *algebra relacji,*
- *relacyjny rachunek krotek i relacyjny rachunek dziedzin*

nie są sobie równoważne.

Każde wyrażenie algebry relacji zwraca zbiór skończony!

Twierdzenie (Twierdzenie)

Języki zapytań dla modelu relacyjnego:

- *algebra relacji,*
- *relacyjny rachunek krotek ograniczony do formuł bezpiecznych i*
- *relacyjny rachunek dziedzin ograniczony do formuł bezpiecznych*

są sobie równoważne.

Proste ćwiczenie 1: Dla każdego wyrażenia algebry relacji istnieje równoważna mu bezpieczna formuła w relacyjnym rachunku krotek.

Bardzo proste ćwiczenie 2: Dla każdej bezpiecznej formuły w relacyjnym rachunku krotek istnieje równoważna mu bezpieczna formuła w relacyjnym rachunku dziedzin.

Twierdzenie 1: Dla każdej bezpiecznej formuły w relacyjnym rachunku dziedzin istnieje równoważne mu wyrażenie algebry relacji.

Pokażemy, że dla każdego $\phi(x_1, \ldots, x_k)$ — bezpiecznej formuły relacyjnego rachunku dziedzin o zmiennych wolnych $x_1, x_2, \ldots, x_k$ istnieje wyrażenie algebry relacji W_ϕ, którego wartością jest $\{x_1, x_2, \ldots, x_k \mid \phi(x_1, x_2, \ldots, x_k)\}$.

1. Zdefiniujmy dziedzinę ϕ:

$$D_\phi \;=\; \{c_1, c_2, \ldots, c_\ell\} \cup \bigcup \pi_{A \in attr(R_i)}(R_i),$$

   gdzie $c_1, c_2, \ldots, c_\ell$ to wszystkie stałe występujące w ϕ, a $R_1, R_2, \ldots, R_m$ to symbole wszystkich relacji występujących w ϕ.

2. Przekształćmy formuły atomowe występujące w ϕ tak, by nie zawierały stałych i powtarzających się zmiennych:

$$R(x, y, x, 13) \;\;\rightarrow\;\; (\exists z, u) R(x, y, z, u) \wedge x = z \wedge u = 13$$

3. Przekształćmy ϕ w ten sposób, by nie zawierała spójników $\wedge$ i kwantyfikatorów $\forall$.

Podstawa indukcji

- dla $\phi(x, y, z) \equiv R(x, y, z)$ definiujemy

$$W_\phi = R, \textit{ ewentualnie } W_\phi = \rho_{x,y,z}(R)$$

- dla $\phi(x) \equiv x > const$ definiujemy

$$W_\phi = \sigma_{x>const}(\rho_x(D))$$

- dla $\phi(x) \equiv x = const$ definiujemy

$$W_\phi = \sigma_{x=const}(\rho_x(D)) \textit{ lub } \rho_x(\{const\})$$

- dla $\phi(x, y) \equiv x \neq y$ definiujemy

$$W_\phi = \sigma_{x \neq y}(\rho_x(D) \times \rho_y(D))$$

Krok indukcyjny

- dla $\phi(x) \equiv \neg\psi(x)$ i wyrażenia W_ψ z atrybutem x definiujemy

$$W_\phi = \rho_x(D) \setminus W_\psi$$

- dla $\phi(x, y, z) \equiv \psi(x, y) \vee \eta(y, z)$ i wyrażeń W_ψ i W_η z atrybutami odpowiednio: x, y oraz y, z definiujemy

$$W_\phi = (W_\psi \times \rho_z(D)) \cup (\rho_x(D) \times W_\eta)$$

- dla $\phi(x) \equiv (\exists y)\psi(x, y)$ i wyrażenia W_ψ z atrybutami x, y definiujemy

$$W_\phi = \pi_x(W_\psi)$$

Wnioski

1. Mamy trzy równoważne języki zapytań dla modelu relacyjnego:
   - Dwa z nich, rachunki relacyjne, są deklaratywne — formułując zapytanie podajemy jego znaczenie, a nie sposób obliczania.
   - Jeden z nich, algebra relacji, jest imperatywny — pisząc wyrażenie podajemy sposób wyliczania, ale znaczenie wyrażenia nie musi być jasne.
2. Moc, czyli możliwości ekspresji rachunków relacyjnych, jest bardzo dobrze znana:
   - to logika pierwszego rzędu, w której można opisać wiele własności,
   - nie można jednak wyrazić np. domknięcia tranzytywnego, do czego potrzebne jest kwantyfikowanie po relacjach (powiedzenie, że "istnieje relacja, taka że. . .),
   - łatwiej zastanawiać się czy zapytania są równoważne lub zawierają się w sobie (optymalizacje!).

Zapytania koniunkcyjne

- Wiele naturalnych własności można wyrazić wyrażeniami algebry relacji używającymi wyłącznie selekcji, projekcji, złączeń i przemianowań.
- Fragmentowi temu odpowiadają zapytania koniunkcyjne, są to formuły rrd/rrk postaci

$$\varphi(\vec{x}) = \exists \vec{y} \bigwedge_i R_i(\vec{x}, \vec{y})$$