Lista zadań nr 4
Bazy Danych 2022

Z16. (3 pkt.)

c) (1pkt.) Przekrój baz danych zdefiniowany jest dla baz o tym samym schemacie i oznacza bazę o tym samym schemacie, w której stan każdej relacji jest przekrojem stanów tej relacji z wszystkich wejściowych baz.

Niech $\mathcal{I}$ oraz $\mathcal{J}$ będą zbiorami baz danych bez nulli (np. $\mathcal{I}$ może być równe $\text{rep}(D)$ dla pewnej bazy z nullami D). Powiemy, że $\mathcal{I}$ oraz $\mathcal{J}$ są *nierozróżnialne w języku zapytań* $\mathcal{L}$ gdy dla każdego zapytania $Q \in \mathcal{L}$

$$\bigcap\{Q(I) | I \in \mathcal{I}\} = \bigcap\{Q(I) | I \in \mathcal{J}\}$$

Niech $\mathcal{L}$ będzie algebrą relacji z operatorami π, σ, ρ.

Rozważmy następującą definicję: *odpowiedzią na zapytanie* $Q \in \mathcal{L}$ *na* D nazywamy takie Q_D, że $\text{rep}(Q_D)$ i $Q(\text{rep}(D))$ są nierozróżnialne w $\mathcal{L}$. Zauważ, że pozwala to mieć w Q_D jakieś dodatkowe krotki z nullami, byle zbiory pewnych odpowiedzi na zapytania w $\mathcal{L}$ się zgadzały.

Pokaż, że dla dowolnej bazy z nullami D istnieje odpowiedź na dowolne $Q \in \mathcal{L}$. Wskazówka: po prostu pokaż jak poprawnie obliczać selekcję i projekcję na relacjach z nullami.

d) (1pkt) Niestety definicja odpowiedzi na zapytanie z poprzedniego podpunktu wciąż nie jest satysfakcjonująca jeśli $\mathcal{L}$ będzie algebrą relacji z π, σ, ρ oraz $\times$ (a zatem również z $\bowtie$).

Przyczyna problemu z ewaluacją złączeń jest dość intuicyjna - ze względu na to, że każdy null w bazie danych jest unikalny (tj. nie można użyć tej samej zmiennej dwukrotnie) nie potrafimy przechowywać informacji, że pewne dwa nulle muszą być w każdych okolicznościach zwartościowane tak samo. Na przykład pomyśl o wyniku zapytania $P \bowtie R$ dla bazy

$$\{P(x, 1), R(1, 1), R(1, 2)\}$$

o schemacie $P(A, B), R(B, C)$. Od razu widać, że wyliczanie zapytań ze złączeniami na takim wyniku będzie problematyczne.

Sformalizuj powyższą intuicję i zademonstruj istnienie bazy z nullami D i zapytania Q takie, że nie istnieje Q_D będąca odpowiedzią Q na D.

Pomysł na dowód (nie dałoby się wymyśleć czegoś prostszego?): Rozważmy bazę danych D z jednym symbolem relacyjnym $T(A, B, C)$ taką, że $D = \{T(a, x, c), T(a', x', c')\}$. Niech $Q = \pi_{AC}(T) \bowtie \pi_B(T)$. Załóżmy, że istnieje takie Q_D, że $\text{rep}(Q_D)$ i $Q(\text{rep}(D))$ są nierozróżnialne w $\mathcal{L}$ $(\star)$.

Niech $Q' = \pi_{AC}(\pi_{AB}(T) \bowtie \pi_{BC}(T))$. Spróbuj zapisać zapytanie $Q \circ Q'$ jak najprościej nie korzystając z symbolu złożenia. Zauważ, że $(a', c) \in \text{certain}(Q \circ Q', D)$. Z $(\star)$ wynika, że $\text{certain}(Q', Q_D) = \text{certain}(Q \circ Q', D))$ (dlaczego?). Poszukaj wartościowania w, dla którego $(a', c) \notin Q'(w(Q_D)))$.

e) (1pkt) Okazuje się, że wystarczy dopuścić aby w krotkach bazy danych ta sama zmienna mogła się dowolnie powtarzać. Pokaż, że następujący algorytm obliczania odpowiedzi Q na dowolnej bazie D dla algebry relacji z $\pi, \sigma, \rho, \bowtie$ jest poprawny: zapominamy, że zmienne są nullami, traktujemy je jak parami różne, nowe stałe i przeprowadzamy wszystkie operacje tak jak gdyby w bazie nie było nulli.

Z17 (1 pkt.) Rozważmy następujące zapytanie w Datalogu.

```
T(X, Y) :- E(X,Y).
T(X, Y) :- T(X,Z), T(Z,Y).
```

Przypomnij definicję semantyki dla Datalogu, a następnie pokaż, że dla każdego $i \in \mathbb{N}_+$ zachodzi $T^i = \{(a,b) |$ istnieje ścieżka z a do b o długości $\leq 2^{i-1}\}$

Z18 (2 pkt., po 0.5 za podpunkt) Napisz następujące zapytania datalogowe. Użyj stałych n i m tam gdzie jest to potrzebne. W tym zadaniu interesują nas ścieżki niezerowej długości.

a) Zwróć wierzchołki, do których można dojść ścieżką z n lub ścieżką z m.

b) Zwróć wierzchołki, do których można dojść ścieżką z n i ścieżką z m.

c) Zwróć pary wierzchołków, do których można dojść z wierzchołka n ścieżkami o tej samej długości.

d) Zwróć pary wierzchołków x, y, takie, że do x oraz do y można dojść z wierzchołka n ścieżkami, które mają różną długość.

Z19 (1 pkt.) Graf jest k-kolorowalny jeśli każdemu wierzchołkowi tego grafu możemy przyporządkować jeden z k kolorów w taki sposób aby każde dwa wierzchołki połączone krawędzią miały różne kolory.

Wiadomo, że graf jest 2-kolorowalny wtedy i tylko wtedy gdy nie zawiera cyklu o nieparzystej długości. Napisz zapytanie datalogowe $Q()$ spełnione w grafach, które nie są 2-kolorowalne.

Z20. (1 pkt.) Odpowiedz konstruując odpowiednie zapytanie lub dowodząc, że takie nie istnieje.

1. Czy można napisać zapytanie datalogowe spełnione wtw gdy w grafie nie ma ścieżki pomiędzy wyróżnionymi wierzchołkami n i m?
2. Czy można napisać zapytanie datalogowe spełnione wtw gdy graf zawiera parzystą liczbę wierzchołków?