Nr rejestracyjny: 2022/45/B/ST6/00457; Kierownik projektu: prof. dr hab. Jerzy Marcinkowski

Zagadki teorii baz danych. Wewnątrz Logiki I Rzędu i poza nią. Opis projektu badawczego (streszczenie popularnonaukowe)

Jerzy Marcinkowski

Czerwiec 2022

Niniejszy projekt mieści się w obszarze teorii baz danych. Teoria ta, inspirowana przez praktykę baz danych, próbuje: • identyfikować, opisywać i badać, w abstrakcyjny matematyczny sposób, podstawowe teoretyczne mechanizmy stojące za praktycznymi aplikacjami w zakresie baz danych; • proponować nowe pojecia i mechanizmy które mogłyby być pożyteczne dla praktyki baz danych; • tłumaczyć – poprzez dowodzenie wyników negatywnych – porażki niektórych próbowanych w praktyce pomysłów.

W pełnym Opisie Projektu Badawczego bardziej szczegółowo omawiam trzy podobszary teorii baz danych i prezentuję kilka otwartych podstawowych technicznych problemów nad którymi planuję pracować. Żaden z tych problemów nie daje się rzetelnie wyjaśnić na jednej stronie popularnego tekstu. Ale spróbujmy trochę opowiedzieć o jednym z nich, ryzykując pewne nieścisłości.

Wyobraźmy sobie, że mamy bazę danych w której przechowujemy informację o tym kto *lubi* kogo, a w tej bazie danych mamy pięć rekordów: *a lubi b*, *b lubi c*, *c lubi a*, *d lubi c* i *a lubi d*.

I wyobraźmy sobie dwa zapytania bazodanowe:

Zapytanie α: Daj mi wszystkie trójki $[x, y, z]$, takie że x *lubi* y i y *lubi* z i z *lubi* x.

Zapytanie β: Daj mi wszystkie trójki $[x, y, z]$, takie że x *lubi* y i x *lubi* z.

Gdy skierujemy do naszej bazy danych zapytanie α, jako odpowiedź otrzymamy sześć trójek[1]: $[a, b, c]$, $[b, c, a]$, $[c, b, a]$, $[a, d, c]$, $[d, c, a]$, $[c, d, a]$.

Gdy skierujemy do naszej bazy danych zapytanie β, w odpowiedzi otrzymamy trójki $[a, b, d]$ i $[a, d, b]$. Ale nie tylko! Zauważmy, że β nie mówi że y i z muszą być różne. W odpowiedzi zatem pojawią się również $[b, c, c]$, $[a, b, b]$, $[d, c, c]$, $[a, d, d]$ i $[c, a, a]$. W sumie zatem β zwróci siedem trójek, o jedną więcej niż α.

Jak się okazuje, nie jest to wcale przypadek. Jakąkolwiek weźmiemy bazę danych, zapytanie β zwróci przynajmniej tyle samo trójek co zapytanie α. Dowód tego faktu nie jest bardzo trudny, ale też wcale nie jest trywialny.

W takiej sytuacji, jak w powyższym przykładzie, mówimy że zapytanie β zawiera zapytanie α.

Byłoby bardzo pożyteczne mieć algorytm[2] który, dla dwóch danych zapytań ϕ i ψ, powiedziałby nam (najlepiej w ciągu ułamka sekundy) czy ϕ zawiera ψ. Taki algorytm byłby ważną cegiełką w budowie silników bazodanowych. Ale kwestia jego istnienia jest znanym otwartym problemem:

> Problem Zawierania Zapytań (Otwarty): Czy istnieje algorytm który, po przeczytaniu dwóch zapytań, powiedziałby czy pierwsze z nich zawiera drugie?

Bardzo zdolni ludzie poświęcili w ostatnich 30 latach sporo czasu starając się rozwiązać ten problem. Osiągnięto przeróżne częściowe wyniki. Odkryto, na przykład, że jeśli pozwolimy w zapytaniach używać nierówności (czyli jeśli będziemy mogli powiedzieć, w treści zapytania, że elementy zwracanej trójki mają być różne) to **takiego algorytmu nie będzie**. Nie dlatego że go nie znamy, tylko dlatego że on po prostu nie istnieje[3].

Jednym z celów naukowych projektu jest dołożenie cegiełki do wysiłku mającego na celu rozwiązać ten problem.

[1] Zauważmy, że porządek elementów w trójkach ma znaczenie.

[2] Inaczej: "program komputerowy".

[3] Czasem możemy *udowodnić* że algorytm nie istnieje. Ma to związek z matematycznym zjawiskiem *nierozstrzygalności*.