Nr rejestracyjny: 2016/21/D/ST6/01485; Kierownik projektu: dr Szymon Abram Toruńczyk

Granice baz danych

W tym projekcie będziemy się zajmować ważnym problemem teorii baz danych, mianowicie problemem ewaluacji zapytań koniunkcyjnych. Przykładem takiego zapytania jest zapytanie

$$Q(x, y, z) = E(x, y) \wedge E(y, z) \wedge E(z, x),$$

wybierające wszystkie takie trójki danych (x, y, z) w bazie danych D, że pary $(x, y), (y, z)$ oraz (z, x) znajdują się w tabeli E tej bazy.

Niezwykle istotnym problemem w komercyjnych systemach bazodanowych jest szybkie wyliczanie wyników tego typu zapytanie koniunkcyjne, w sytuacji, gdy baza danych D jest bardzo duża, np. jej tabela E zawiera $N = 10^6 = 1000000$ par danych. Algorytm dokonujący naiwnej ewaluacji powyższego zapytania sprawdziłby wszystkie możliwe kombinacje (x, y) oraz (y', z) dwójek par występujących w tabeli E, i dla każdej takiej dwójki par, jeżeli $y = y'$ oraz para (z, x) też występuje w tabeli E, to trójka (x, y, z) zostałaby wrzucona do wyniku. W tym przypadku, algorytm przeczesałby $N^2 = 10^{12}$ dwójek par w tabeli E. Przypuszczając, że procesor komputerowy potrafi dokonać 10^9 operacji na sekundę (tyle potrafią operacji wykonać procesory taktowane zegarem 1GHz), na wynik czekalibyśmy 1000 sekund.

Okazuje się, że można skonstruować sporo szybszy algorytm. W tym celu należy wpierw zauażyć, że dla dowolnej bazy danych D, zbiór wyników zapytania Q ma conajwyżej $N^{3/2}$ wyników. Jest to nietrywialne twierdzenie, które można który ma dwa inne równoważne sformułowania:

- graf o N krawędziach ma conajwyżej $N^{3/2}$ trójkątów;
- bryła w przestrzeni trójwymiarowej, której rzuty na płaszczyzny xy, yz oraz xz mają pole powierzchni ograniczone przez liczbę N, ma objętość conajwyżej równą $N^{3/2}$.

Liczba $3/2$ w powyższym twierdzeniu jest optymalna, dla rozważanego zapytania Q. Oznacza to, że istnieją bazy danych D mające N par w tabeli E, oraz dla których zapytanie Q zwraca $N^{3/2}$ wynikowych trójek. W związku z tym, w najgorszym wypadku, czas działania optymalnego algorytmu obliczającego wynik zapytania Q na bazie danych D o N parach w tabeli E musi być przynajmniej proporcjonalny do liczby $N^{3/2}$, gdyż algorytm tyle musi wyprodukować odpowiedzi. Okazuje się, że można rzeczywiście taki algorytm skonstruować. Tak więc, dla $N = 10^6$, algorytm działałby w jedną sekundę, co czyni olbrzymią różnicę w porównaniu z naiwną ewaluacją.

W tym projekcie będziemy próbowali skonstruować algorytm który działa jeszcze szybciej, korzystając z dodatkowych informacji na temat bazy danych D, np. statystyk wystąpień każdej danej w każdej kolumnie. Nasze podejście do tego problemu będzie korzystało z metod kombinatoryki i teorii miary. Zamiast rozważać bardzo duże bazy danych, będziemy rozważali nieskończone, graniczne bazy danych, i dla nich będziemy dowodzili twierdzeń matematycznych.