prof. dr hab. Mikołaj Bojańczyk
Instytut Informatyki
Uniwersytet Warszawski

Recenzja rozprawy doktorskiej Tomasza Gogacza,
*Fundamentalne własności teorii będących zbiorami*
*Tuple Generating Dependencies i związki między tymi własnościami.*

Praca dotyczy teorii baz danych, gdzie (relacyjne) bazy danych są rozumiane jako modele skończone w sensie z teorii modeli. Zapytania w bazach danych to są po prostu formuły logiki pierwszego rzędu ze zmiennymi wolnymi, ewaluacja takiej formuły to zbiór krotek w uniwersum bazy danych, które formułę spełniają. Ważnym rodzajem zapytań są *zapytania koniunkcyjne*, które wyglądają na przykład tak

$$\varphi(x,y) = \exists z R(x,x) \wedge R(x,z) \wedge S(x,y,z).$$

W zapytaniu takim można użyć $\exists$ oraz $\wedge$, ale nie można użyć $\forall$, $\vee$ ani $\neg$.

Tematem pracy jest ewaluacja zapytań koniunkcyjnych w modelu skończonym, który nie jest do końca znany. Wiadomo o nim tylko tyle, że zawiera pewien model skończony $\mathbb{D}$, i że jest modelem dla pewnen teorii $\mathcal{T}$. Problem formalizuje się tak: dla modelu skończonego $\mathbb{D}$, dla teorii $\mathcal{T}$ oraz zapytania koniunkcyjnego $\phi$ bez zmiennych wolnych, chcemy wiedzieć czy zachodzi:

(*) $\phi$ jest prawdziwe w każdym modelu skończonym, który rozszerza $\mathbb{D}$ i spełnia wszystkie zdania z $\mathcal{T}$.

Praca rozważa przypadek, gdzie teoria $\mathcal{T}$ składa się ze skończonego zbioru zdań postaci

$$\forall x_1 \cdots \forall x_n (\phi(x_1, \ldots, x_n) \Rightarrow \exists y Q(y_1, \ldots, y_k))$$

gdzie zmienne $y_1, \ldots, y_k$ są wzięte, być może z powtórzeniami i być może w sposób niewyczerpujący, ze zbioru $\{y, x_1, \ldots, x_n\}$. Dla takich teorii używa się w teorii baz danych nazwy *Tuple Generating Dependencies* i jest to najważniejszy przypadek rozważanych teorii w pytaniach typu (*).

Doktorat składa sie z czterech prac (P1)-(P4) oraz wstępu. Prace są moim zdaniem bardzo mocne, o czym napiszę bardziej szczegółowo poniżej. Pierwsza z nich, (P1), jest wersją czasopismową, jeszcze nie opublikowaną. Pozostałe prace są pracami konferencyjnymi. Myślę, że redakcja doktoratu pozostawia nieco do życzenia, część problemu wynika z niejednolitej formy rozprawy. Mam nadzieję, że kiedyś pojawią się pełne czasopismowe wersje omawianych prac, szczególnie (P4), co ułatwi światu zapoznanie się z naprawdę wartościowymi wynikami doktoranta. Lektura prac była dla mnie trudna, być może trudniejsza nawet niż wskazywałby na to materiał; autorzy często chyba widzą trudność nie w tych samych miejscach co ja. Nie chciałbym jednak koncentrować się na krytyce redakcji – w przypadku tej rozprawy dominująca jest jakość wyników, o czym poniżej.

(P1) *Converging to the Chase – Tool for Finite Controllability.* Praca ta dotyczy *Sticky Logic.* Logikę tę wprowadzili Cali, Gottlob i Pieris w 2010. Jest to szczególny przypadek *Tuple Generating Dependencies,* którego ograniczenia mają służyć lepszym własnościom algorytmicznym, w szczególności rozstrzyganiu własności (*). Wprowadzając *Sticky Logic,* autorzy zapytali, czy zachodzi własność *finite controlability* zdefiniowana tak:

- Dla każdego modelu skończonego $\mathbb{D}$ i zapytania koniunkcyjnego $\phi$, warunek (*) jest równoważny swojej nieskończonej wersji, gdzie kwantyfikacja przebiega po dowolnych modelach.

Prace (P1) udziela pozytywnej odpowiedzi na to pytanie. Dopiero dzięki tej odpowiedzi *Sticky Logic* nabiera pełnego sensu, w szczególności wnioskiem z równoważności jest rozstrzygalność warunku (*). Trudność leży rzecz jasna w implikacji z (*) do jego nieskończonej wersji, którą można przeformułować tak: jeśli istnieje pewien model rozerzający $\mathbb{D}$, gdzie zachodzą zarówno $\mathcal{T}$ jak i $\neg\phi$, to istnieje też model skończony o tych własnościach. Dowód tej implikacji jest bardzo trudny technicznie. Konstrukcja wymagania pokolorawania nieskończonego modelu kolorami opisującymi tak dużo informacji, że nieuniknione powtórzenie się kolorów pozwala na zawinięcie modelu w strukturę skończoną. Łatwo powiedzieć, trudno zrobić.

(P2) *On the BDD/FC Conjecture.* Powiemy, że teoria $\mathcal{T}$ ma własność BDD jeśli dla każdego zapytania koniunkcyjnego bez zmiennych wolnych $\phi$ istnieją zapytania koniunkcyjne $\phi_1, \ldots, \phi_n$ bez zmiennych wolnych takie, $\phi_1 \vee \cdots \vee \phi_n$ jest spełnione dokładnie w tych modelach skończonych $\mathbb{D}$, które spełniają warunek (*). W pracy postawiona jest hipoteza, że dla każdej teorii $\mathcal{T}$ ze składni *Tuple Generating Dependencies,* własność BDD implikuje własność finite controlability. Hipoteza ta jest udowodniona dla sygnatur, gdzie wszystkie predykaty mają arność najwyżej dwa. Dowód korzysta z podobnych technik jak praca (P1). Myślę, że przełomem były pomysły z (P1), a praca (P2) jest twórczym i nietrywialnym ich zaadaptowaniem.

(P3) *All-instances termination of chase is undecidable.* Rozważając problem (*), w naturalny sposób pojawia się procedura *Chase,* który zaczyna od modelu skończonego $\mathbb{D}$ i dodaje do niego elementy i krotki w relacjach zgodnie z teorią $\mathcal{T}$, aż do osiągnięcia punktu stałego. Procedura ta jest jednym z fundamentalnych narzędzi w teorii baz danych. Trudność polega na tym, że być może punkt stały jest modelem nieskończonym, a teoria baz danych interesuje się modelami skończonymi. Stąd naturalne pytanie:

- Dane: Teoria $\mathcal{T}$ w składni *Tuple Generating Dependencies.*
- Pytanie: Czy dla każdego modelu skończonego $\mathbb{D}$, procedura *chase* kończy się w skończonej liczbie kroków dla $(\mathcal{T}, \mathbb{D})$?

W różnych wariantach problem ten był badany od lat i poświęcono mu wiele prac. Dotychczasowe prace głównie podawały warunki na $\mathcal{T}$, które gwarantują, że odpowiedź na pytanie jest twierdząca. W pracy (P3) pytanie jest potraktowane jako problem decyzyjny i udowodnione jest, że problem jest nierozstrzygalny. Dowód nierozstrzygalności jest całkiem naturalny i nieskomplikowany i aż trudno uwierzyć, że tyle trzeba było na niego czekać, biorąc pod uwagę zaintersowanie problemem. Ale, jak piszą sami autorzy, wynik jest łatwy – jak się go już zna.

(P4) *The Hunt for a Red Spider: Conjuntive Query Determinacy Is Undecidable.* Praca ta jest moją ulubioną w rozprawie. Praca pokazuje, że nierozstrzygalny jest następujący problem:

- Dane: Zapytania koniunkcyjne $Q_1, \ldots, Q_n, Q$. Zapytania te mają zmienne wolne, a więc dla każdego modelu produkują zbiór krotek.
- Pytanie: Czy dla każdego, być może też nieskończonego, modelu, wynik $Q_1, \ldots, Q_n$ jednoznacznie determinuje wynik $Q$?

Pytanie o rozstrzygalność było otwarte od lat osiemdziesiątych i ma bogatą bibliografię. Problem jest o tyle subtelny, że łatwo się pomylić rozwiązując go, czego przykładem prace podające błędne rozwiązania, o których można przeczytać we wstępie pracy. Dowód nierozstrzygalności w pracy (P4) jest trudny i urodziwy. Jest to jeden z najlepszych dowodów nierozstrzygalności, jakie widziałem. Jak to zwykle bywa z dowodami nierozstrzygalności, dowód pokazuje, że problem można traktować jako język programowania, ale ten akurat język programowania jest wyjątkowo nieprzyjazny dla użytkownika.

Wszystkie prace są napisane wspólnie z promotorem, ale na podstawie rozmów z promotorem i doktorantem, wyciągam wniosek, że wkład doktoranta w powstanie wyników był kluczowy.

**Podsumowanie.** Myślę, że nie trzeba tutaj dużo pisać – doktorat jest świetny. Rozwiązuje otwarte problemy, korzystając z technik trudnych i, co ważne, autentycznie pomysłowych. Wnoszę o dopuszczenie Tomasza Gogacza do dalszych etapów przewodu doktorskiego. Wnioskuję o wyróżnienie rozprawy.

Warszawa, 21 stycznia 2015
Mikołaj Bojańczyk