prof. dr hab. inż. Marzena Kryszkiewicz
Politechnika Warszawska
Instytut Informatyki
ul. Nowowiejska 15/19
00-665 Warszawa
mkr@ii.pw.edu.pl

Warszawa, 2.09.2015

Recenzja rozprawy doktorskiej
mgr inż. Pawła Boińskiego
zatytułowanej:
Przetwarzanie zbiorów przestrzennych zapytań eksploracyjnych w środowiskach z ograniczonym rozmiarem pamięci operacyjnej

1. Problem badawczy i jego znaczenie

Opiniowana rozprawa doktorska dotyczy zagadnień z dziedziny eksploracji danych. Jest poświęcona metodom i strukturom wspierającym wydajne odkrywanie w przestrzennych zbiorach danych specyficznych grup cech zwanych *powszechnymi kolokacjami przestrzennymi*. Bardzo ważnym wyróżnikiem tej rozprawy jest zaproponowanie rozwiązań uwzględniających rozmiar dostępnej pamięci operacyjnej.

W rozprawie przyjęto typowe założenie, że kolokacje przestrzenne składają się z tych cech, których instancje są obiektami, charakteryzowanymi m.in. przez atrybut położenia (lokalizację). Zakłada się także, że jest zdefiniowana zwrotna i symetryczna relacja sąsiedztwa pomiędzy obiektami z wykorzystaniem atrybutu lokalizacja. Określanie, czy dana kolokacja przestrzenna jest powszechna, wymaga wyznaczenia stosownych *instancji* tejże *kolokacji*. W opiniowanej rozprawie przyjęto założenie, że powszechne kolokacje przestrzenne są zdefiniowane za pomocą instancji kolokacji typu *klika*, którą można scharakteryzować następująco: *Instancją typu klika kolokacji przestrzennej C* składającej się z k cech $c_1, \ldots, c_k$ jest zbiór takich obiektów $o_1, \ldots, o_k$, że obiekt o_i posiada cechę c_i ($i = 1 \ldots k$) oraz każda para obiektów ze zbioru $\{o_1, \ldots, o_k\}$ jest w relacji sąsiedztwa. Tak jak w rozprawie doktorskiej, w dalszej części mojej recenzji, zwrot „instancje kolokacji przestrzennej" należy rozumieć jako „instancje typu klika kolokacji przestrzennej". *Powszechność kolokacji przestrzennej C* jest mierzona za pomocą *indeksu uczestnictwa*, który jest wyznaczany na podstawie zbioru wszystkich instancji kolokacji C i liczby instancji każdej cechy w kolokacji C. *Kolokacja przestrzenna* jest definiowana jako *powszechna*, jeśli jej indeks uczestnictwa jest nie mniejszy niż zadana wartość progowa.

Powszechne kolokacje przestrzenne dostarczają wiedzy, która często znajduje bezpośrednie zastosowanie w wielu dziedzinach, w tym m.in. tak ważnych jak: epidemiologia, ochrona zdrowia, hydrologia, analiza przestępczości. W podrozdziale 3.1.4, Doktorant przedstawia zastosowania wiedzy zawartej w powszechnych kolokacjach przestrzennych w wymienionych powyżej oraz innych dziedzinach.

Zadanie odkrywania powszechnych kolokacji przestrzennych staje się problemem w przypadku zbiorów danych o dużej liczbie cech i licznych instancjach cech. W takim przypadku proces odkrywania może trwać nieakceptowalnie długo przy zastosowaniu standardowych algorytmów ich odkrywania, takich jak np. *Collocation Miner*, *joinless* lub *iCPI-tree*. Taka sytuacja występuje np. wtedy, gdy rozmiar struktur wykorzystywanych w procesie odkrywania kolokacji przestrzennych przekracza rozmiar dostępnej pamięci operacyjnej. Jakkolwiek dostępne w systemach operacyjnych mechanizmy pamięci wirtualnej i stronicowania umożliwiają działanie algorytmów nawet przy niedoborach pamięci operacyjnej, to jednak ich działanie bywa często nieakceptowalnie spowolnione ze względu na liczne, dodatkowe operacje odczytu i zapisu. Zatem zdecydowanie lepiej jest korzystać z algorytmów odkrywania wzorców, a w szczególności, kolokacji przestrzennych, które są dostosowane do takiego korzystania z pamięci operacyjnej i dyskowej, aby liczba operacji we/wy była możliwie mała. To wyzwanie zostało podjęte w opiniowanej rozprawie doktorskiej.

Najważniejszym problemem, któremu jest poświęcona opiniowana rozprawa doktorska, jest wydajne odkrywanie powszechnych kolokacji przestrzennych w sytuacji, gdy rozmiar struktur wykorzystywanych w algorytmach ich odkrywania przekracza rozmiar dostępnej pamięci operacyjnej. Kolejnym zagadnieniem rozważanym w rozprawie jest zrównoleglenie operacji wyszukiwania powszechnych kolokacji przestrzennych, także w przypadku istotnego ograniczenia rozmiaru pamięci operacyjnej. W rozprawie rozważane jest także zadanie wspólnego wyszukiwania wszystkich kolokacji przestrzennych, z których każda spełnia warunki sformułowane w co najmniej jednym zapytaniu eksploracyjnym z przetwarzanego zbioru zapytań.

W rozprawie Autor postawił ogólną tezę, mówiącą że „Konsolidacja i zrównoleglenie operacji wewnętrznych, zastosowanie technik materializacji, współużytkowania i sekwencyjnego przetwarzania pośrednich struktur danych w algorytmach odkrywania kolokacji przestrzennych umożliwiają poprawę efektywności realizacji zbiorów zapytań eksploracyjnych o kolokacje przestrzenne, w szczególności w środowisku z ograniczonym rozmiarem pamięci operacyjnej." Sformułował także jej następujące tezy szczegółowe:

- „Można efektywnie odkrywać kolokacje przestrzenne w środowisku z ograniczonym rozmiarem pamięci operacyjnej przy wykorzystaniu metody hybrydowego połączenia mieszającego podczas filtrowania zbioru instancji kandydatów na kolokacje przestrzenne.
- W przypadku zbyt dużego rozmiaru struktur pośrednich w stosunku do rozmiaru dostępnej pamięci operacyjnej, można efektywnie zastosować techniki materializacji danych poprzez wykorzystanie pamięci masowej.
- Zastosowanie przetwarzania równoległego przy pomocy procesorów kart graficznych, wraz z uwzględnieniem fizycznych ograniczeń na rozmiar pamięci, może wielokrotnie przyspieszyć działanie algorytmu odkrywania kolokacji przestrzennych.
- Współbieżna, czyli wspólna realizacja zapytań o kolokacje przestrzenne, dla których przetwarzane obszary w przestrzeni nakładają się, może być bardziej wydajna niż oddzielna realizacja tych zapytań.
- Można podzielić eksplorowaną przestrzeń na mniejsze fragmenty w celu efektywnej realizacji jednego lub wielu zapytań w środowisku z ograniczoną pamięcią operacyjną."

Rozprawa ma charakter naukowo-badawczy. Problem badawczy rozważany w rozprawie ma istotne znaczenie praktyczne.

2. Wkład autora

Autor przeprowadził analizę wydajności algorytmów odkrywania kolokacji przestrzennych ze względu na rozmiar dostępnej pamięci operacyjnej. Na przykładzie algorytmów *joinless* i *iCPI-tree*, przeprowadził badania eksperymentalne, których zadaniem było zidentyfikowanie typowych etapów w algorytmach odkrywania o największej czasochłonności, która wiązała się z wykonywaniem operacji na strukturach o znaczących rozmiarach, w tym przekraczających rozmiar dostępnej pamięci operacyjnej. Za takie uznał etapy dotyczące operacji na kandydatach na instancje kandydatów na powszechne kolokacje przestrzenne, w tym operację weryfikacji tychże kandydatów na instancje z użyciem instancji powszechnych kolokacji przestrzennych, wyznaczonych w poprzedniej iteracji w przypadku algorytmu *joinless* oraz operacji tworzenia kandydatów na instancje z użyciem struktury przechowującej informacje o sąsiedztwie instancji w przypadku algorytmu *iCPI-tree*. Autor zauważa, że potencjalnie bardzo duże rozmiary tych struktur mogą prowadzić do znaczącego spadku wydajności algorytmów odkrywania kolokacji przestrzennych w związku z dodatkowymi operacjami we/wy wynikającymi ze stronicowania. W rezultacie, Doktorant w swojej rozprawie koncentruje się na zaproponowaniu struktur w pamięci operacyjnej i/lub na dysku oraz operacji na nich w taki sposób, aby liczba operacji we/wy była jak najmniejsza. W szczególności, Doktorant proponuje grupowanie kandydatów na powszechne kolokacje przestrzenne, aby weryfikację ich kandydatów na instancje przeprowadzać z użyciem tych samych danych, a tym samym, potencjalnie uniknąć wielokrotnego pobierania tych danych z dysku do pamięci operacyjnej. Doktorant przedstawia zaproponowane przez siebie rozwiązania dotyczące krytycznych etapów algorytmów odkrywania kolokacji przestrzennych na przykładzie algorytmów *joinless* i *iCPI-tree*. Doktorant także umiejętnie adaptuje rozwiązania znane z dziedziny baz danych do przypadku wydajnego odkrywania kolokacji przestrzennych, w tym metodę *Hybrid Hash Join* do weryfikacji kandydatów na instancje oraz strukturę plików reprezentujących sąsiedztwa cech.

W ramach rozprawy Doktorant zaproponował:

- metodę *BCM* dostosowującą algorytm *joinless* do odkrywania powszechnych kolokacji przestrzennych z uwzględnieniem rozmiaru dostępnej pamięci operacyjnej (podrozdz. 5.1);
- metody *MiCPI-tree i PMiCPI-trees* dostosowujące algorytm *iCPI-tree* do odkrywania powszechnych kolokacji przestrzennych z uwzględnieniem rozmiaru dostępnej pamięci operacyjnej (podrozdz. 5.2 i 5.3);
- algorytm *Common iCPI-tree* dostosowujący algorytm *iCPI-tree* do wykonywania zbioru pewnego typu zapytań o powszechne kolokacje przestrzenne (rozdz. 7);
- algorytm *PCommon iCPI-trees* stanowiący adaptację algorytmów *Common iCPI-tree* i *iCPI-tree* do wykonywania zbioru pewnego typu zapytań o powszechne kolokacje przestrzenne z uwzględnieniem rozmiaru dostępnej pamięci operacyjnej (rozdz. 7);
- algorytm odkrywania kolokacji przestrzennych z wykorzystaniem procesorów kart graficznych (rozdz. 6).

Kluczowymi cechami *BCM* metody są:

- Zaobserwowanie, że zbiór instancji typu klika kandydatów na powszechne kolokacje przestrzenne w algorytmie joinless można uzyskać poprzez wykonanie operacji złączenia na zbiorze instancji typu gwiazda kandydatów na powszechne kolokacje przestrzenne i zbiorze instancji powszechnych kolokacji przestrzennych znalezionych w poprzedniej iteracji oraz zaproponowanie wykonania takiego złączenia z użyciem metody *Hybrid Hash Join* autorstwa DeWitta i innych, która uwzględnia ograniczenie na rozmiar dostępnej pamięci operacyjnej, w tym ewentualną konieczność

materializacji pojemników z przydzielonymi w wyniku zastosowania funkcji mieszającej instancjami typu gwiazda.

- Zaproponowanie grupowania kandydatów na powszechne kolokacje przestrzenne w taki sposób, aby weryfikacja poprawności instancji typu gwiazda wszystkich kandydatów na powszechne kolokacje przestrzenne w każdej z utworzonych grup mogła być wykonana z użyciem tego samego podzbioru pojemników z przydzielonymi instancjami kolokacji przestrzennych, znalezionych w poprzedniej iteracji.
- Zaproponowanie dodatkowej, *wczesnej eliminacji instancji typu gwiazda kandydatów na powszechne kolokacje przestrzenne*, poprzedzającej zgrubną weryfikację instancji. Zaproponowana eliminacja polega na odrzuceniu wytworzonej instancji typu gwiazda jednego z kandydatów na powszechne kolokacje przestrzenne przetwarzanej grupy, jeśli nie istnieje pojemnik z instancjami kolokacji przestrzennych z poprzedniej iteracji, identyfikowany przez wartość mieszającą sprawdzanej instancji typu gwiazda.

Kluczowymi cechami metody *MiCPI-tree* są;

- Zaproponowanie sposobu materializacji struktury drzewa sąsiedztw, który umożliwia szybkie wyszukiwanie węzłów drzewa z dysku.
- Zaproponowanie nowej metody wytwarzania instancji typu klika dla kandydata na powszechną kolokację przestrzenną z wykorzystaniem instancji powszechnej kolokacji przestrzennej, zawierającej o jedną cechę mniej niż rozważany kandydat. Charakterystyczną własnością nowej metody jest jednokrotne wyszukiwanie obiektów o danej cesze będących sąsiadami danego obiektu bez względu na jego liczbę wystąpień w instancjach kolokacji powszechnej. Zaproponowana metoda wymaga sortowania zbioru instancji kolokacji powszechnej k-1 razy, gdzie k jest liczbą cech tej kolokacji.

Najistotniejszymi cechami metody *PMiCPI-trees* są;

- Zaproponowanie metody rozszczepienia drzewa *iCPI-tree* na zbiór drzew, w taki sposób, aby każda instancja typu klika dowolnej kolokacji była wyprowadzalna z jakiegoś drzewa *PMiCPI*. Parametrami tej metody jest maksymalna odległość pomiędzy sąsiadami i rozmiar dostępnej pamięci operacyjnej. Cechą uboczną zaproponowanej metody jest to, że drzewa mogą częściowo współdzielić gałęzie i, w rezultacie, możliwe jest tworzenie zduplikowanych instancji kandydatów na kolokacje.
- Zaproponowanie sposobu wykrywania duplikatów instancji kandydata na powszechną kolokację przestrzenną z użyciem złączenia mieszającego.

Kluczowymi cechami algorytmu *Common iCPI-tree* są;

- Założenie, że w zapytaniach o powszechne kolokacje przestrzenne stosowana jest ta sama relacja sąsiedztwa, natomiast progi powszechności, cechy i przestrzenie, do których mogą się odnosić zapytania, mogą być zróżnicowane.
- Założenie, że przestrzeń, do których odnoszą się zapytania, będzie podzielona na rozłączne obszary.
- Założenie, że generowanie kandydatów na powszechne kolokacje przestrzenne, wyznaczanie ich instancji oraz struktura przechowująca relacje sąsiedztwa będą wspólne dla wszystkich zapytań z danego zbioru.
- Przechowywanie (np. z użyciem mapy bitowej) dla każdego zapytania informacji o wynikowych obszarach, do których się odnosi.
- Przechowywanie (np. z użyciem mapy bitowej) dla każdej instancji kolokacji informacji o wynikowych obszarach, w których występuje co najmniej jeden element tej instancji.
- Użycie struktury *Common iCPI-tree* do przechowywania informacji o sąsiedztwach, które rożni się tym od *iCPI-tree*, że dodatkowo w każdym węźle reprezentującym

instancję cechy, jest dodatkowo przechowywana informacja (np. z użyciem mapy bitowej) o wynikowym obszarze, na którym ta instancja się znajduje.

- Przechowywanie (np. z użyciem mapy bitowej) dla każdego kandydata na kolokację powszechną informacji o zapytaniach, dla których może stanowić poszukiwaną odpowiedź.

Wyróżnikiem algorytmu *PCommon iCPI-trees* jest wykorzystanie możliwości wspólnego wyszukiwania kolokacji dla zapytań z danego zbioru, zgodnie z ideą przedstawiona w algorytmie *Common iCPI-tree*, jednakże z użyciem drzew *PCommon iCPI-trees* stanowiących poddrzewa *Common iCPI-tree* w celu wydajniejszego zarządzania pamięcią operacyjną. Idea tworzenia tychże drzew oraz uzyskiwania odpowiedzi z ich użyciem jest oparta na algorytmie *PMiCPI-trees*.

Doktorant zaproponował także algorytm równoległego wyznaczania powszechnych kolokacji przestrzennych z uwzględnieniem rozmiaru dostępnej pamięci na kartach graficznych. Istotnymi operacjami zaproponowanymi przez Doktoranta w przypadku odkrywania powszechnych kolokacji przestrzennych z wykorzystaniem procesorów kart graficznych są:

- Wydajne wyszukiwanie wspólnych sąsiadów dla pojedynczej instancji kolokacji.
- Wydajne przetwarzanie wielu instancji wielu kandydatów na powszechne kolokacje przestrzenne z uwzględnieniem ograniczeń technologii CUDA.

Większość przedstawionych w rozprawie rozwiązań w zakresie odkrywania powszechnych kolokacji przestrzennych została opublikowana w 2 czasopismach międzynarodowych i 5 materiałach z konferencji międzynarodowych oraz zaprezentowana w 3 raportach technicznych.

3. Poprawność

Prezentacja znanych i proponowanych przez Autora rozwiązań oraz wywody na ich temat są prowadzone w sposób logiczny, przekonujący, zrozumiały oraz są bardzo dobrze zilustrowane licznymi przykładami. Przeprowadzone eksperymenty i dyskusja otrzymanych wyników przekonują o dużej użyteczności rozwiązań zaproponowanych przez Doktoranta. Wydajność zaproponowanych rozwiązań jest zdecydowanie większa od wydajności ich oryginalnych odpowiedników w przypadku, gdy rozmiar wykorzystywanych struktur przekracza rozmiar dostępnej pamięci operacyjnej oraz niewiele ustępuje wydajności oryginalnych rozwiązań w przeciwnym przypadku.

Pewne zastrzeżenia może budzić sposób przedstawienia niektórych definicji i fragmentów pseudokodów (co przedstawiam poniżej w punkcie poświęconym zauważonym słabościom/nieścisłościom rozprawy). Na ogół zastrzeżenia dotyczą kwestii, które, w moim przekonaniu, zostały wadliwie przedstawione już w literaturze poświęconej odkrywaniu powszechnych kolokacji przestrzennych. Niemniej ilustracje definicji i algorytmów, zamieszczone w rozprawie, pozwalają w pełni zrozumieć intencje Doktoranta jak i intencje autorów standardowych algorytmów. Jakość prezentacji standardowych algorytmów w ujęciu Doktoranta nie jest gorsza niż oryginalna, a bywa nawet znacząco lepsza. Doktorant wszystkie algorytmy, zarówno nieautorskie, jak i autorskie, przedstawia w jednolity sposób, a mianowicie: prezentuje algorytm w postaci pseudokodu, opisuje poszczególne kroki algorytmu, ilustruje sposób wykonania. Dzieje się tak nawet w przypadku nieautorskich algorytmów, dla których oryginalnie nie dostarczono pseudokodu. W rezultacie prezentacja algorytmu w rozprawie bywa bardziej precyzyjna niż oryginalny zapis.

Zauważone słabości/nieścisłości:

- Na str. 31 rozprawy pojęcie *cechy przestrzennej* określono jako własność przestrzeni w danym jej punkcie, a jako przykłady cech przestrzennych podano miasto, lotnisko, bankomat, rzekę. Przykłady te sugerują, że to, co Doktorant (wzorem m.in. autorów artykułów [177, 212]) rozumie pod pojęciem *cechy przestrzennej* obiektu, odpowiada *wartości atrybutu opisowego* obiektu (a nie *atrybutu przestrzennego*) w terminologii systemów informacji przestrzennej.
- Mimo, że Definicja 3.7 jest poświęcona pojęciom: *wzorca kolokacji przestrzennej i instancji kolokacji przestrzennej*, tak naprawdę, nie definiuje ona pojęcia *instancji kolokacji przestrzennej*, a podana tam definicja *wzorca kolokacji przestrzennej*, odpowiadająca ściśle definicji z artykułu autorstwa J.S. Yoo i S. Shekhara, opublikowanego w *IEEE Transactions on Knowledge and Data Engineering* [212], w moim przekonaniu jest niepoprawna/nieprecyzyjna w tym samym stopniu co pierwowzór. Należy odnotować, że zwrot *instancje cech* nie jest równoważny zwrotowi *instancje zbioru cech*.
- Doktorant badał czas wykonania poszczególnych etapów algorytmów, co pozwoliło mu upewnić się, dla których etapów algorytmów należy opracować wydajniejsze rozwiązania. Niemniej w przypadku uwzględniania ograniczenia na rozmiar dostępnej pamięci operacyjnej, ciekawą i istotną informacją byłoby również zbadanie, ilu jest tworzonych kandydatów na powszechne kolokacje przestrzenne i ile jest powszechnych kolokacji przestrzennych oraz ilu jest tworzonych kandydatów na instancje dla tych zbiorów cech. Niestety, w rozprawie nie przedstawiono tego typu wyników.
- Na stronie 40 w przykładzie 3.5 przedstawiony sposób obliczania prawdopodobieństwa warunkowego reguły $\{A\} \rightarrow \{C, D\}$ jako $P(\{A\} | \{C, D\})$ jest nieprawidłowy.
- W stworzonych przez siebie pseudokodach standardowych algorytmów: *Co-location Miner* (Algorytm 3.2, str. 42), *joinless* (Algorytm 3.3, str. 49), *iCPI-tree* (Algorytm 3.4, str. 54), Doktorant po etapie wyznaczenia powszechnych kolokacji przestrzennych o k cechach, zamieszcza wywołanie pewnej funkcji wyznaczania reguł kolokacyjnych. Wśród parametrów tej funkcji nie ma liczności zbiorów instancji powszechnych kolokacji przestrzennych znalezionych w poprzednich iteracjach, które są potrzebne do wyznaczenia prawdopodobieństw warunkowych reguł kolokacyjnych. Tę sama wadę posiada pseudokod algorytmu *joinless* w [212]. Doktorant nie rozważa, jak przechowywać te dane, aby szybko wyznaczyć reguły kolokacyjne, ani nie przedstawia, ile czasu zajmuje ich wyznaczenie (np. w porównaniu z czasem wyszukiwania powszechnych kolokacji przestrzennych). Przypuszczam, że etap wyznaczania reguł kolokacyjnych nie został zaimplementowany przez Doktoranta i że eksperymentalne porównanie wydajności standardowych i nowych rozwiązań odnosi się wyłącznie do odkrywania powszechnych kolokacji przestrzennych. Takie ograniczenie byłoby uzasadnione, zwłaszcza, że rozprawa i tak jest bardzo obszerna, niemniej Autor powinien w rozprawie poinformować o takiej decyzji, jeżeli ją podjął.
- Zapis $c \cup f_{k+1}$ na str. 43. gdzie c jest zbiorem cech, a f_{k+1} jedną z cech, jest niepoprawny.
- Rys. 3.7 (str. 44) zawiera pseudokod generacji instancji (k+1)-elementowego kandydata na powszechną kolokację przestrzenną na podstawie instancji CI^x i CI^y k-elementowych kolokacji w ramach algorytmu *Co-location Miner*. Znaczenie CI^x, CI^y jest przedstawione wyłącznie w komentarzu zamieszczonym w pseudokodzie. Nie przedstawiono, ani nie omówiono w tekście rozprawy, w jaki sposób CI^x i CI^y są

identyfikowane dla poszczególnych kandydatów na kolokacje, mimo, że oryginalny opis *Co-location Miner* zawiera wyjaśnienie, w jaki sposób dla każdego (k+1)-elementowego kandydata na kolokację zidentyfikować parę instancji k-elementowych kolokacji, aby wyznaczyć instancje rozważanego kandydata.

- Na podstawie przeprowadzonych eksperymentów z użyciem własnej implementacji algorytmu *joinless* Doktorant wnioskuje, że generacja zbiorów kandydatów na instancje kandydujących kolokacji trwa ok. dwukrotnie krócej niż weryfikacja końcowa ich podzbiorów pozostałych po weryfikacji wstępnej. Niestety nie są podane informacje o liczbach kandydatów na instancje po ich wygenerowaniu i po weryfikacji wstępnej, ani o liczbie instancji po weryfikacji końcowej. Nie wiadomo, jaka jest wydajność wykonania weryfikacji końcowej. Doktorant nie przedstawił zastosowanego w swojej implementacji algorytmu *joinless* sposobu weryfikacji końcowej kandydatów na instancje. Nie jest jasne, czy w ramach weryfikacji usuwano niepoprawne instancje, czy też kopiowano instancje poprawne. Nie wiadomo zatem, co jest przyczyną, że wydajność weryfikacji końcowej kandydatów na instancje jest mniejsza niż wydajność generacji kandydatów na instancje.
- Na str. 80 Doktorant podaje warunki podziału zbioru kandydatów C_k na grupy. Zapis słowny warunku trzeciego jest sformułowany poprawnie, natomiast zapis formalny tego warunku aktualnie nie jest poprawny, jednak stanie się poprawny po dodaniu, że grupa kandydatów g_1 jest różna od grupy kandydatów g_2.
- W algorytmie *BCM*, przedstawionym na str. 84, *SN* jest zdefiniowane jako $\{SN_1, \ldots, SN_{fn}\}$, gdzie $\{f_1, \ldots, f_n\}$ jest zbiorem n cech. Jednocześnie *SN* jest skomentowane jako zbiór sąsiedztw typu gwiazda. Ponieważ jest n cech, to wynika z tego, że *SN* składa się z n sąsiedztw typu gwiazda - po jednym na każdą cechę. Jednak w ogólnym przypadku, sąsiedztw typu gwiazda dla ustalonej cechy może wiele. Zgaduję zatem, że pisząc SN_{fi}, Doktorant miał na myśli wszystkie sąsiedztwa typu gwiazda dla cechy f_i. Pojawia się jednak pytanie, czy *SN* miało być rodziną zbiorów sąsiedztw typu gwiazda, czy też zbiorem sąsiedztw typu gwiazda. W tym drugim przypadku potrzebna byłaby korekta równania definiującego *SN*. Dodatkowo w linii 8 i 16 algorytmu *BCM*, zmienna n jest użyta do oznaczania kolejnych elementów *SN*. Jest to dość niezręczne, ponieważ n zostało użyte do oznaczenia liczby cech. Nie jest też jasne, dlaczego dla każdej grupy kandydatów wywoływana jest funkcja *filterStarInstances* z każdym elementem *SN*.
- W rozprawie nie przedstawiono wyników eksperymentalnych, które ilustrowałyby, jak bardzo zaproponowana operacja wczesnej eliminacji instancji typu gwiazda kandydatów na kolokacje redukuje ich liczbę.
- Na rys. 7.3 (str. 159), węzły dla cechy *C* i *D* powinny mieć mapę bitową [111], a nie [110].

Powyższe uchybienia nie wpływają na moją bardzo wysoką ocenę zaproponowanych przez Doktoranta rozwiązań.

4. Wiedza kandydata

Autor omawia istniejący stan wiedzy, mający związek z problemem badawczym rozpatrywanym w rozprawie, głównie w rozdziałach 2-4 oraz w dodatkach. Rozdział drugi wprowadza w zagadnienia eksploracji klasycznych zbiorów danych oraz przestrzennych zbiorów danych. Uzupełnieniem tego rozdziału jest Dodatek A, w którym Autor omawia reprezentację rastrową oraz wektorową danych przestrzennych. W rozdziale trzecim Autor przedstawia problem odkrywania powszechnych kolokacji przestrzennych, któremu

poświęcona jest rozprawa doktorska. W rozdziale tym Doktorant szczegółowo prezentuje cztery ważne algorytmy odkrywania takich kolokacji: *Co-location Miner*, *Multiresolution Co-location Miner*, *joinless* i *iCPI-tree*. Ilustracje wykonania algorytmów *Co-location Miner*, *joinless* i *iCPI-tree* są zawarte w Dodatku B. Dodatkowo, w rozdziale trzecim, Autor wzmiankuje o alternatywnych definicjach kolokacji przestrzennych i algorytmach ich odkrywania. Rozdział czwarty stanowi wprowadzenie do problematyki przetwarzania zbiorów zapytań eksploracyjnych o kolokacje przestrzenne i problematyki przetwarzania danych w przypadku ograniczonego rozmiaru dostępnej pamięci operacyjnej oraz w przypadku stosowania *GPGPU*. Podrozdział 6.1 zawiera dodatkowe informacje dotyczące przetwarzania równoległego na kartach graficznych.

Przeprowadzona w pracy prezentacja i analiza literatury światowej potwierdza głęboką wiedzę autora w dziedzinie informatyki w zakresie odkrywania wzorców częstych, w tym kolokacji przestrzennych. Potwierdza także głęboką wiedzę w zakresie struktur danych i organizacji fizycznej plików, systemów baz danych, systemów operacyjnych, architektury komputerów. Bibliografia liczy 220 prac i jest właściwie skorelowana z tematyką rozprawy.

5. Inne uwagi

Rozprawa jest zredagowana bardzo starannie, choć nie udało się uniknąć pewnych literówek.

Zaproponowane rozwiązania dotyczące wyszukiwania powszechnych kolokacji przestrzennych można wykorzystać w systemach informacji przestrzennej do zwiększenia wydajności przetwarzania danych. Można je także z pożytkiem zaadaptować do wykonywania także innych zadań eksploracji danych, takich jak klasyfikacja, grupowanie, wyszukiwanie innych rodzajów wzorców.

6. Podsumowanie

Biorąc pod uwagę opinie zaprezentowane w poprzednich punktach i wymagania zdefiniowane przez artykuł 13 Ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym (z późniejszymi zmianami) moja ocena rozprawy pod względem trzech podstawowych kryteriów jest następująca:

A. Czy rozprawa zawiera oryginalne rozwiązanie problemu naukowego?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

B. Czy po przeczytaniu rozprawy zgadzasz się, że kandydat posiada ogólną wiedzę teoretyczną w dyscyplinie Informatyka?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

C. Czy kandydat posiada umiejętność samodzielnego prowadzenia pracy naukowej?

[X]	[ ]	[ ]	[ ]	[ ]
Zdecydowanie TAK	*Raczej TAK*	*Trudno powiedzieć*	*Raczej NIE*	*Zdecydowanie NIE*

Ponadto, biorąc pod uwagę dużą przydatność zaproponowanych rozwiązań dla odkrywania powszechnych kolokacji przestrzennych oraz możliwość ich adaptacji w celu zwiększenia wydajności wykonywania innych zadań eksploracji danych, wnioskuję o wyróżnienie rozprawy doktorskiej.