Prof. dr hab. inż. Stanisław Kozielski
Instytut Informatyki
Politechnika Śląska

Gliwice, 9.05.2014 r.

Recenzja

rozprawy doktorskiej mgr Tomasza Xięskiego nt.: „Wydobywanie wiedzy z danych złożonych"

Promotor rozprawy: prof. dr hab. inż. Alicja Wakulicz-Deja
Promotor pomocniczy: dr Agnieszka Nowak-Brzezińska

1. Aktualność i znaczenie tematyki rozprawy

Problematyka opiniowanej pracy dotyczy eksploracji danych i odkrywania wiedzy – jednej z najbardziej intensywnie rozwijanych gałęzi informatyki. Z kilku przyczyn tego rozwoju warto zwrócić uwagę na lawinowy wzrost ilości danych generowanych przez Internet oraz różnego rodzaju systemy techniczne, wśród nich systemy telekomunikacyjne. Zagadnienie to doczekało się własnej nazwy – *Big Data*, a także nowych metod i narzędzi umożliwiających przetwarzanie i analizę takich ilości danych. Niniejszą rozprawę można również zaliczyć do tej grupy. Analizując ogromne zbiory danych telekomunikacyjnych autor dostrzegł problem w zastosowaniu do analizy tych zbiorów znanych metod i narzędzi. Okazało się m.in., że bardzo duże rozmiary wyników analiz nie pozwalają w sposób czytelny zinterpretować uzyskanych rezultatów, a w wielu przypadkach czasy przetwarzania są nieakceptowalne.

Poszukując sposobu skutecznego zmierzenia się z danymi o ogromnych rozmiarach autor oryginalnie połączył istniejące już metody tworząc nowe jakościowo podejście, a jego implementacja dała skuteczne programowe narzędzie analizy i wydobywania wiedzy ukrytej w bardzo dużych zbiorach danych.

Podjęcie przez autora opiniowanej rozprawy takiej tematyki badań oceniam jako wybór zadań bardzo aktualnych, a przy tym trudnych. Uważam, że rozprawa dotyczy ważnej i aktualnej problematyki badawczej.

2. Zakres pracy

Praca składa się z dziewięciu rozdziałów. Tematyka pracy jest mocno osadzona w realiach konkretnych danych, dotyczących sieci telekomunikacyjnych. Ten akcent został zaznaczony już na początku pracy, bowiem po wprowadzeniu autor przedstawił w rozdziale drugim przegląd podstawowych informacji dotyczących systemów telefonii komórkowej, a także charakterystykę dwóch zbiorów analizowanych danych: zbioru danych o urządzeniach nadawczo-odbiorczych sieci komórkowej rozlokowanych w aglomeracji śląskiej oraz zbioru danych opisujących parametry transmisji i wykorzystanie zasobów przez poszczególne urządzenia sieciowe. Następnie autor krótko scharakteryzował proces eksploracji danych i odkrywania wiedzy, akcentując problem danych złożonych i kluczowe znaczenie wiedzy dziedzinowej w procesie analizy danych.

Trzeci rozdział pracy zawiera przegląd i porównawczą analizę narzędzi programowych umożliwiających prowadzenie eksploracji danych. Autor omówił systemy komercyjne i niekomercyjne, zwracając szczególną uwagę na udostępniane przez te systemy możliwości grupowania danych (analizy skupień) oraz technik graficznego opisu danych, w szczególności

wizualizacji wyników grupowania. W podsumowaniu autor konkluduje, że w tylko niewielkiej liczbie programów dostępne są, najbardziej intersujące w ocenie autora, gęstościowe metody grupowania, a także żaden z ocenianych systemów nie udostępnia interaktywnej metody graficznej reprezentacji skupień. Te stwierdzenia były dla autora inspiracją do rozwinięcia badań, których wyniki zostały przedstawione w dalszych częściach rozprawy.

W rozdziale czwartym zostały przedstawione wybrane metody opisu danych, w szczególności techniki statystycznego opisu danych. Autor dość szczegółowo omawia na wstępie miary tendencji centralnej oraz miary rozproszenia. Wprowadzenie tych pojęć było niezbędne do przedstawionej dalej dyskusji form graficznej prezentacji rozkładu danych. Omówiono dwa rodzaje histogramów oraz szczegóły tzw. wykresów pudełkowych. Końcowa część rozdziału została poświęcona krótkiej dyskusji nad sposobami reprezentowania wyników grupowania danych. Autor przedstawił koncepcję czterech wariantów tworzenia reprezentantów grup i zapowiedział wizualizację struktury grup za pomocą techniki tzw. map prostokątów. Treść tej części rozdziału wyprzedza rozdział dotyczący grupowania danych, co nie ułatwia zrozumienia intencji autora i nie jest najszczęśliwszym rozwiązaniem.

Rozdział piąty jest poświęcony omówieniu cech algorytmów grupowania danych bazujących na pojęciu gęstości danych. Autor skupił się na dwóch algorytmach gęstościowych: *DBSCAN* oraz *OPTICS*. W oparciu o źródłowe prace, w których po raz pierwszy przedstawiono te algorytmy, autor bardzo szczegółowo omówił podstawy koncepcji grupowania gęstościowego oraz własności wymienionych algorytmów.

Szósty rozdział pracy jest poświęcony graficznym metodom reprezentacji skupień. Po ogólnych rozważaniach dotyczących motywacji wykorzystania technik wizualizacji danych autor definiuje i szerzej omawia proces graficznej analizy eksploracyjnej. Autor deklaruje wykorzystanie tego procesu do wykrywania odchyleń, a przede wszystkim graficznej reperzentacji skupień (wyników grupowania danych). Następnie autor omawia różne metody takiej reprezentacji, poczynając od najczęściej spotykanych, którymi są dendrogramy oraz diagramy Woronoja. Kolejno przedstawiane są inne metody graficznej prezentacji danych, mające również zastosowanie do prezentacji skupień, takie jak mapy ciepła, wykresy osiągalności, wykresy kołowe, techniki grafowe, drzewa stożkowe, drzewa hiperboliczne, techniki typu *sunburst* oraz *icicle*, a także mapy prostokątów. Po dyskusji zalet i wad tych metod autor wybrał tę ostatnią metodę i przeanalizował znane techniki generowania map prostokątów, m.in.: *Slice and Dice, Squarified, Split, Strip, Pivot* i wybrał spośród nich algorytm *Squarified.*

W rozdziale siódmym przedstawiono kluczowy dla całej rozprawy projekt systemu *DensGroup*, umożliwiającego wydobywanie wiedzy z danych z wykorzystaniem technik analizy skupień. W założeniach projektowych autor przyjął, że grupowaniu mogą podlegać bardzo duże zbiory (liczące dziesiątki tysięcy lub więcej) obiektów, co może prowadzić, przy zastosowaniu np. rozważanych w pracy gęstościowych algorytmów grupowania, do bardzo licznych (np. rzędu tysięcy) zbiorów grup. Takie założenie postawiło duże wyzwanie przed zadaniem wizualizacji wyników grupowania, która dla autora stanowi kwintesencję procesu odkrywania wiedzy z wykorzystaniem metod grupowania. Aby umożliwić interaktywną wizualizację, zapewniającą czytelną interpretację wyników grupowania przy tej skali problemu, autor zaproponował dwuetapowy proces grupowania. W pierwszym etapie jest realizowane grupowanie gęstościowe (z użyciem algorytmów *DBSCAN* lub *OPTICS*). W drugim etapie odbywa się grupowanie reprezentantów skupień (grup) uzyskanych w ramach pierwszego etapu, przy wykorzystaniu aglomeracyjnego algorytmu grupowania AHC. Algorytm ten tworzy pełną strukturę hierarchiczną skupień, którą przed procesem wizualizacji można przyciąć albo do ustalonej liczby grup, albo też na podstawie zadanego progu podobieństwa. Tak przycięte drzewo hierarchii skupień poddawane jest wizualizacji

metodą map prostokątów. Dla zwiększenia czytelności uzyskanych wyników rozmiary i kolory prostokątów można powiązać z cechami skupień, np. wielkość i tonacja kolorów prostokątów mogą określać liczbę obiektów w skupieniach.

Przedstawiony system *DensGroup* tworzy bardzo interesujące, oryginalne narzędzie eksploracji danych. Nieco dyskusyjny jest przy tym sposób jego prezentacji, bowiem omawiany rozdział siódmy pracy ma postać dokumentacji programu *DensGroup,* przedstawiono więc kolejno: instalację i wymagania sprzętowe programu, jego interfejs i funkcjonalność, strukturę plików wejściowych i przykład użycia. W takim ujęciu ciekawa koncepcja połączenia dwuetapowego procesu grupowania z możliwościami wizualizacji jego wyników jest przysłonięta szczegółami technicznymi programu. Bardziej czytelne mogłoby być przedstawienie najpierw samej koncepcji zaproponowanego podejścia, a dopiero potem (może nawet w dodatku do pracy) całej dokumentacji programu. Przy takim rozwiązaniu do opisu dwuetapowego procesu grupowania można by też włączyć omówienie sposobów tworzenia reprezentantów skupień, obecnie zamieszczone, niezbyt moim zdaniem szczęśliwie, w rozdziale czwartym (punkt 4.3).

Rozdział ósmy pracy został poświęcony omówieniu eksperymentów obliczeniowych przeprowadzonych przez autora w celu odkrycia różnych elementów wiedzy zawartej w dwóch zbiorach złożonych danych telekomunikacyjnych, przedstawionych wcześniej w rozdziale drugim.

Pierwsze eksperymenty dotyczyły tworzenia i analizy histogramów oraz analizy możliwości wykorzystania klauzul grupujących języka SQL do wykrycia związków między atrybutami, bądź obiektami badanych zbiorów danych. Autor umiejętnie wykorzystał tu najnowsze rozszerzenia języka SQL dostępne m.in. w systemie MS SQL Server (klauzula *over* z argumentem *partition by*, mechanizm ramek) do wydobywania z danych ciekawych informacji na temat sprawności urządzeń telekomunikacyjnych. Zauważyć przy tym można, że wyciąganie części wniosków wymagało konsultacji z ekspertem dysponującym głęboką wiedzę dziedzinową. Kolejny eksperyment posłużył autorowi do doboru parametrów startowych do gęstościowych algorytmów grupowania. Przeprowadzone badania pozwoliły na sformułowanie heurystyki do wyznaczania takich parametrów. Autor nazywa je parametrami „optymalnymi", ale z uwagi na bardzo szacunkowe oceny ten przymiotnik należy raczej traktować jako eufemizm. Wykorzystując wyznaczone wartości parametrów w następnym eksperymencie autor z powodzeniem sprawdził przydatność opracowanego programu *DensGroup* do dwuetapowego procesu grupowania (po pierwszym etapie prawie osiem tysięcy grup), a potem wizualizacji i analizy wyników grupowania. Otrzymane wyniki potwierdziły i poszerzyły efekty analiz z wykorzystaniem języka SQL. Podobne badania, z analogicznymi efektami autor przeprowadził dla drugiego zbioru rzeczywistych danych telekomunikacyjnych.

Przeprowadzone eksperymenty i analizy stanowią ważną część pracy. Potwierdziły one przydatność opracowanego autorskiego systemu *DensGroup* do wymagającej analizy bardzo dużych zbiorów złożonych danych rzeczywistych. Pokazały też głęboką wiedzę i doświadczenie autora w prowadzeniu eksploracji zbiorów, których rozmiary kwalifikują je do kategorii *Big Data.*

W dziewiątym rozdziale pracy autor podsumował uzyskane wyniki i przedstawił końcowe wnioski.

3. Ocena rozprawy

Autor rozprawy wykazał się głęboką wiedzą w dziedzinie eksploracji danych, w szczególności w zakresie metod grupowania danych oraz technik graficznej prezentacji wyników procesu eksploracji danych. Do najważniejszych rezultatów badawczych uzyskanych przez autora zaliczam:

1) Opracowanie koncepcji dwuetapowego grupowania, obejmującej połączenie metod grupowania gęstościowego w pierwszym etapie, z wyborem reprezentantów tak utworzonych grup i aglomeracyjnym grupowaniem hierarchicznym w drugim etapie.
2) Opracowanie interaktywnej metody wizualizacji wyników grupowania powiązanej z dwuetapowym grupowaniem, dostosowanej do grupowania hierarchicznego i umożliwiającej, dzięki wykorzystaniu techniki map prostokątów, interesującą prezentację własności wyznaczonych grup.
3) Opracowanie systemu *DensGroup,* oryginalnego narzędzia eksploracji danych, w którym zaimplementowano przestawioną koncepcję dwuetapowego grupowania i metodę wizualizacji jego wyników.

Należy podkreślić, że opracowany system umożliwia eksplorację bardzo dużych zbiorów danych, które można klasyfikować w kategorii *Big Data.*

Wysoko oceniam eksperymentalną część pracy, która potwierdziła przydatność opracowanego systemu *DensGroup* do analizy bardzo dużych zbiorów złożonych danych rzeczywistych. Prowadzone eksperymenty potwierdziły również dużą biegłość i doświadczenie autora w prowadzeniu złożonej analizy danych.

Warto w tym miejscu podkreślić, że rezultaty badań przedstawione w pracy były wcześniej opublikowane w jedenastu publikacjach, w tym w dwóch pracach indeksowanych przez *Web of Science.* Prace te były też prezentowane na kilku konferencjach, w tym również międzynarodowych. Wyniki te zostały więc zweryfikowane w środowisku naukowym.

4. Uwagi do pracy

Lektura pracy nasuwa następujące uwagi i pytania:

1) Określenie „dane złożone", które pojawia się w rozprawie już w jej tytule, może mieć w dziedzinach eksploracji danych, baz danych i innych, bardzo szerokie znaczenie, np. do danych złożonych zaliczane są dane o złożonej strukturze (np. dokumenty XML, dane przestrzenne itp.), dane multimedialne: obrazy, dane audio i video, teksty i szereg innych. Autor odnosi to określenie do danych o dużej liczbie atrybutów i typów danych (str. 6). Warto byłoby zwrócić uwagę, że w literaturze spotykane jest szersze znaczenie tego terminu.
2) W kilku przypadkach autor wybrał arbitralnie wartości pewnych parametrów, np. procent próbkowanych danych, parametry początkowe gęstościowych algorytmów grupowania, próg podobieństwa w algorytmie AHC i inne. Intersujące byłoby sprawdzenie, nawet wyrywkowe, jak wrażliwe są wyniki badań na wartości wybieranych parametrów.
3) Dyskusyjna wydaje się organizacja treści rozdziału siódmego rozprawy. Bardziej czytelnym rozwiązaniem mogłoby być przedstawienie najpierw samej koncepcji zaproponowanego podejścia dwuetapowego grupowania i wizualizacji, a dopiero potem prezentacja całego interfejsu programu *DensGroup.*

4) Interesujące byłoby przedstawienie przez autora w ostatnim rozdziale pracy planów rozwoju dalszych badań w podejmowanej dziedzinie, a w szczególności planów wykorzystania systemu *DensGroup*.

5) Lekturę pracy ułatwiłaby dostępność wykazu stosowanych oznaczeń wraz z odniesieniem do miejsc ich definicji.

Kilka drobnych uwag redakcyjnych naniosłem w treści pracy i przekazałem autorowi.

Przedstawione uwagi nie zmieniają mojej pozytywnej i wysokiej oceny rozprawy.

5. Wniosek końcowy

Podsumowując stwierdzam, że w recenzowanej rozprawie został sformułowany, a następnie poprawnie rozwiązany oryginalny problem naukowy obejmujący opracowanie nowych metod i narzędzi programowych dla eksploracji danych.

Autor wykazał się głęboką wiedzą w zakresie eksploracji danych, a w szczególności grupowania danych i wizualizacji jego wyników, a także umiejętnościami i doświadczeniem praktycznym, udokumentowanymi opracowaniem systemu eksploracji danych i udaną weryfikacją opracowanej koncepcji poprzez eksperymentalną analizę dużych zbiorów danych rzeczywistych

Stwierdzam, że oceniana praca spełnia całkowicie wszystkie wymagania stawiane rozprawom doktorskim przez obowiązujące przepisy. Wnoszę o dopuszczenie Pana mgr Tomasza Xięskiego do dalszych etapów przewodu doktorskiego.

Ponadto, biorąc pod uwagę wysoki poziom merytoryczny pracy, opracowanie oryginalnego systemu eksploracji danych o nowych możliwościach i dużej praktycznej użyteczności, a także znaczący dorobek publikacyjny autora **wnioskuję o wyróżnienie rozprawy**.