Częstochowa, dn. 29 marca 2024 r.

prof. dr hab. inż. Rafał Scherer
Katedra Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej Anh Nguyen, pt.: Mining Sequence and Inter-Sequence Patterns in Large Databases.

Niniejszą recenzję opracowano na wniosek Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Wrocławskiej. Promotorem jest prof. dr hab. inż. Ngoc Thanh Nguyen, promotorem pomocniczym prof. Bay Vo.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Wykrywanie wzorców sekwencji pozwala na identyfikację trendów i zachowań w danych sekwencyjnych. Analiza taka może być używana do przewidywania przyszłych zdarzeń na podstawie wcześniejszych sekwencji, na przykład w analizie medycznej, w zrozumieniu preferencji klientów, do lepszej personalizacji usług i rekomendacji itd. Ponadto, może ona również służyć do wykrywania anomalii w danych. Odpowiednio zidentyfikowane odstępstwa od typowych wzorców mogą wskazywać na potencjalne problemy lub nieprawidłowości. Oczywiście, w przypadku ogromnych danych z jakimi mamy do czynienia w obecnym świecie, algorytmy takie wymagają dużej ilości pamięci i mocy obliczeniowej. Celem pracy jest zmniejszenie zapotrzebowania na pamięć i moc obliczeniową procesów eksploracji wzorców międzysekwencyjnych. W tym celu Autor proponuje nową strukturę danych oraz model wydobywania wzorców międzysekwencyjnych z ograniczeniami zestawów elementów, aby zmniejszyć liczbę generowanych kandydatów.

2. Zawartość rozprawy

Recenzowana praca składa się ze spisu rysunków, spisu tabel, spisu algorytmów, pięciu rozdziałów oraz bibliografii. Dokument liczy 109 stron.

Pierwszy rozdział jest wprowadzeniem do tematyki. Autor opisał ideę eksploracji danych i jej zastosowania. Podał przykład możliwości odkrywania reguł asocjacyjnych w danych dotyczących sprzedaży. Reguły asocjacyjne są tworzone, aby opisać relacje między różnymi elementami w zbiorze danych. Następnie opisano problem wydobywania sekwencyjnych wzorców polegający na wydobywaniu podsekwencji w zestawie sekwencji, czyli na odkryciu interesujących wzorców lub sekwencji zdarzeń w danych sekwencyjnych lub czasowych. Wydobywanie wzorców międzysekwencyjnych to rozszerzenie wydobywania sekwencyjnych wzorców, które polega na odkrywaniu wspólnych wzorców, powiązań i zależności między sekwencjami w bazie danych sekwencyjnych. Identyfikuje ono wzorce, które są wspólne nie tylko w obrębie tej samej transakcji, ale również między transakcjami. Dalej omówiono motywację do badań, a mianowicie fakt, że dotychczasowe algorytmy eksploracji wzorców sekwencyjnych charakteryzują się zasobożernością. Podano cel pracy, czyli rozwiązanie problemu ograniczeń wydobywania wzorców międzysekwencyjnych pod względem czasu przetwarzania i zajętości pamięci, przez zaproponowanie nowej struktury danych problemu wydobywania wzorców międzysekwencyjnych, mającą na celu minimalizację duplikacji danych w trakcie procesu eksploracji. Zaproponowany będzie również model wydobywania wzorców międzysekwencyjnych z ograniczeniami zbiorów elementów w celu zmniejszenia liczby wygenerowanych kandydatów. Na końcu Rozdziału 1 zestawiono elementy będące wkładem pracy do istniejącego stanu wiedzy oraz omówiono strukturę rozprawy.

Rozdział 2 jest przeglądem literatury. Omówiono metody eksploracji danych sekwencyjnych, począwszy od podstawowych definicji takich jak transakcje, elementy, zbiory elementów, sekwencje, podsekwencje, supersekwencje. Przybliżono algorytmy eksploracji wzorców sekwencyjnych: AprioriAll, FreeSpan, SPADE, PrefixSpan, PRISM, CM-SPADE, oraz eksploracji wzorców sekwencyjnych z ograniczeniami: MSPIC-DBV, MWAPC and EMWAPC, MSRIC-R and MSRIC-P. Następnie omówione są algorytmy do eksploracji danych typu Clickstream, czyli zawierających informacje o interakcjach użytkownika z daną stroną internetową lub aplikacją. Dane takie mogą obejmować informacje takie jak kolejność kliknięć, przeglądanych stron, czas spędzony na poszczególnych stronach, interakcje z formularzami, przewijanie strony itp. I tu omówiono algorytmy CUP, CM-WSPADE, Compact-SPADE oraz SUI (Sequential pattern mining Using Indices). Następnie zebrane są algorytmy eksploracji między-sekwencyjnej: EISP-Miner, DBV-ISP, ISP-IC, iISP-IC, piISP-IC.

Rozdział 3 omawia problem wydobywania wzorców międzysekwencyjnych (inter-sequence patterns) oraz ograniczenia obecnie stosowanej struktury danych w obecności zduplikowanych danych podczas procesu eksploracji. Autor proponuje wykorzystanie struktury danych pseudo-IDList oraz algorytm o nazwie ISP-PI (Inter-Sequence Pattern mining based on PseudoIndex), specjalnie zaprojektowany dla problemu wydobywania wzorców międzysekwencyjnych. Algorytm wykorzystuje metodę ISP-IC (Inter-Sequence Pattern mining with Index Intersection Checking) w celu zmniejszenia czasu eksploracji. Aby ocenić skuteczność zaproponowanych metod pod względem czasu działania i wykorzystania pamięci w porównaniu z wcześniejszymi algorytmami analizy wzorców międzysekwencyjnych, Doktorant użył sześć zbiorów danych.

Rozdział rozpoczyna się przeglądem istniejących algorytmów od 1996 roku, z dokładnym opisem struktury DBV-PatternList oraz opartej na niej metody DBV-ISPMIC. Doktorant następnie wprowadza strukturę pseudo-IDList. Następnie omawiane są metody generacji

kandydatów, algorytmy ISP-IC i ISP-PI oraz ich złożoność obliczeniowa. Do porównań i sprawdzenia działania algorytmów użyte zostały rzeczywiste zbiory danych MSNBC, Kosarak, FIFA, BMSWebView2 oraz dwa syntetyczne. Procedura pruningu metody ISP-IC usuwała nawet kilkadziesiąt kandydatów, w zależności od zbioru danych. Porównanie szybkości działania i zapotrzebowania na pamięć algorytmu ISP-PI zostało wykonane z algorytmami EISPMiner roku 2009 oraz DBV-ISP z 2012 roku. Zaproponowany algorytm wykazał się bardzo dużym przyspieszeniem oraz zmniejszeniem zapotrzebowania na pamięć porównaniu do dwóch wspomnianych algorytmów z literatury.

Rozdział 4 omawia autorskie metody eksploracji między-sekwencyjnych wzorców z ograniczeniami. Doktorant zaproponował algorytm o nazwie DBV-ISPMIC oraz jego ulepszoną wersję o nazwie DBV-ISPMIC-IMPROVING, wykorzystującą pruning, wraz z jego zrównoleglonąą modyfikacją pDBV-ISPMIC-IMPROVING. Algorytmy te zostały porównane z metodami EISP-Miner z 2009 roku i DBV-ISP z 2012 r. pod kątem szybkości i zapotrzebowania na pamięć na pięciu zbiorach testowych. Na wstępie Doktorant omawia istniejące w literaturze algorytmy MSPIC-DBV oraz ISP-IC oraz strukturę danych DBV-PatternList zmniejszająca zapotrzebowanie na pamięć. Na bazie tej struktury zaproponowany został algorytm DBV-ISPMIC wykorzystujący wykorzystuje strukturę danych DBV-PatternList oraz strukturę drzewa ISP-Tree. Struktura danych DBV-PatternList umożliwia szybkie obliczanie wsparcia kandydatów oraz zmniejsza zużycie pamięci. Doktorant opracował również nową metodę przetwarzania struktur ISP-Tree, przetwarzając je równolegle, co doprowadziło do powstania równoległej wersji powyższego algorytmu – pDBV-ISPMIC. Doktorant samodzielnie zaimplementował algorytmy w języku C# i sprawdzone na zbiorach danych syntetycznych, zbiorach danych typu clickstream oraz opisującym wypożyczenia rowerów.

Rozdział 5 jest podsumowaniem rozprawy z dokładnym zestawieniem elementów nowości. Podano ograniczenia stworzonych metod oraz badań, takie jak ograniczona skalowalność, możliwość stosowania ich to sekwencyjnych zbiorów danych ograniczona wartość parametru maxspan, brak testów na zaawansowanych platformach obliczeniowych, zbadanie metod tylko dla wybranych ograniczeń oraz na statycznych zestawach danych sekwencyjnych.

Doktorant podaje również możliwości przyszłych badań podzielone tematycznie. Są to np. sprawdzenie skalowalności metod poprzez testy na większych zbiorach danych, sprawdzenie aplikowalności metod dla innych typów danych takich jak szeregi czasowe, dane multimedialne czy dane rosnące w czasie, sprawdzenie możliwości zastosowania innych platform obliczeniowych oraz przetwarzania równoległego i rozproszonego, sprawdzenie różnych typów ograniczeń. Ponadto podano pomysły na rozwój badań w kierunku innych wyzwań związanych z eksploracją danych. Na końcu występuje spis siedmiu publikacji doktoranta związanych z pracą doktorską.

Dalej następuje podsumowanie pracy oraz bibliografia.

3. Ocena rozprawy

W ramach rozprawy doktorskiej Doktorant zaproponował zestaw oryginalnych metod związanych z eksploracją wzorców międzysekwencyjnych w obszernych bazach danych sekwencyjnych. Cele pracy są trafnie i jasno sformułowane, a tematyka pracy jest aktualna. Oryginalny dorobek autora polega na

- stworzeniu algorytmu DBV-ISPMIC do eksploracji wzorców międzysekwencyjnych z ograniczeniami zestawów elementów,

- stworzeniu równoległej wersji powyższego algorytmu pDBV-ISPMIC,
- zaproponowaniu algorytmu ISP-PI, który stosuje strukturę danych pseudo-IDList do wydobywania wzorców międzysekwencyjnych, wraz z metodą pruningu kandydatów ISP-IC i jej udoskonaleniem,
- zweryfikowaniu poprawności i skuteczności zaproponowanych algorytmów przez eksperymenty w użyciem kilku zbiorów danych.

Zaprezentowany materiał pokazuje, że Doktorant zrealizował cel pracy. Rozprawa doktorska uwidacznia ogólną wiedzę teoretyczną i praktyczną mgra inż. Anh Nguyena. Doktorant opublikował siedem prac naukowych – sześć w czasopismach z listy JCR oraz jeden w materiałach konferencji. Rozprawa doktorska wykazuje umiejętność samodzielnego prowadzenia pracy naukowej Doktoranta. Rozprawa doktorska stanowi oryginalne rozwiązanie problemu naukowego. Zaproponowane metody mają duże znaczenie aplikacyjne dla nauk technicznych i przemysłu.

Niezależnie od mojej bardzo wysokiej oceny pracy, wykraczającej poza poziom przeciętny, poniżej wymieniam drobne uwagi, które nasunęły się w czasie czytania:

Sekcja 3.2 nazwana jest „Data Structure", ale omawia więcej niż jedną strukturę. Brak wprowadzenia na początku tej sekcji mówiącego czego ona dotyczy. Podobnie jest w przypadku Sekcji 4.3.

Skrót IPM ma stronie 48 nie jest zdefiniowany, ale można domyślić się jego znaczenia.

Dlaczego jedne algorytmy implementowane były w języku Java, a inne w C#?

Czy równolegle wersje zaproponowanych algorytmów wykorzystywały jakieś programistyczne i softwarowe mechanizmy tworzenia aplikacji równoległych, czy jedynie wbudowane w języki programowania? Czy wydajność zależała od użytego sprzętu (procesora/procesorów, szybkości dostępu do pamięci)?

Od kilku lat pojawiają się metody sequential pattern mining oparte na głębokich sieciach neuronowych. Czy mogą one być obecnie konkurencją dla tradycyjnych algorytmów?

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Anh Nguyen w rozprawie doktorskiej „Mining Sequence and Inter-Sequence Patterns in Large Databases" zrealizował cel rozprawy. Zaprezentowane rezultaty stanowią oryginalny wkład Autora w rozwój dyscypliny Informatyka Techniczna i Telekomunikacja. Pan Anh Nguyen wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w zakresie eksploracji danych, szczególnie sekwencyjnych. Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej Informatyka Techniczna i Telekomunikacja. Wnoszę o jej przyjęcie i dopuszczenie do dalszych etapów postępowania doktorskiego. Ponadto ze względu na ponadprzeciętny poziom rozprawy oraz fakt opublikowania sześciu prac związanych bezpośrednio z tematyką rozprawy w czasopismach znajdujących się na liście JCR, wnioskuję o wyróżnienie pracy.

Rafał Scherer