Gdynia, 9 stycznia 2024 r.

Prof. dr hab. inż. Ireneusz Czarnowski
Katedra Systemów Informacyjnych
Uniwersytet Morski w Gdyni
ul. Morska 83, 81-225 Gdynia

RECENZJA

rozprawy doktorskiej mgr inż. Anh Nguyena

pt.: *"Mining sequence and inter-sequence patterns in large databases"*
(Eksploracja wzorców sekwencji i inter-sekwencji w dużych zbiorach danych)

Recenzję przygotowałem zgodnie z pismem Przewodniczącego Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Wrocławskiej z dnia 27 listopada 2023 roku, kierowanym do mojej osoby wskazanej jako recenzenta rozprawy doktorskiej mgr. inż. Anh Nguyena.

Rozprawa doktorska została przygotowana w dyscyplinie informatyka techniczna i telekomunikacja. W tej samej dyscyplinie prowadzone jest postępowanie w sprawie nadania mgr. inż. Anh Nguyenowi stopnia doktora.

1. Problematyka naukowa rozprawy

Tematyka rozprawy dotyczy problemu odkrywania wzorców sekwencji w zbiorach danych, w tym w szczególności odkrywania wzorców międzysekwencyjnych (inter-sekwencji).

W ogólności, odkrywanie wzorców sekwencji w zbiorach danych dotyczy poszukiwania zależności występujących pomiędzy zdarzeniami odzwierciedlonymi w tych danych. Identyfikacja wzorców sekwencji należy do problemów trudnych obliczeniowo, w szczególności, gdy ich poszukiwanie prowadzone jest na dużych zbiorach danych.

Metody odkrywania wzorców sekwencji są szeroko stosowane, w tym w medycynie, handlu i bankowości, analizie koszykowej, telekomunikacji, analizie i identyfikacji zachowań konsumentów, czy identyfikacji zachowań użytkowników aplikacji internetowych, oraz wielu innych. Z drugiej strony, odkrywanie wzorców międzysekwencyjnych jest cennym narzędziem do wydobywania przydatnych informacji z sekwencji danych. Ten aspekt poszukiwań obecnie zyskuje coraz większe znaczenie, pozwala bowiem na szersze poznanie związków i zależności oraz proponowanie nowych rozwiązań technologicznych.

Znane metody odkrywania wzorców sekwencji w danych oraz wzorców międzysekwencyjnych charakteryzują się szeregiem ograniczeń, które przekładają się na wysokie koszty obliczeń, w tym koszt

WPŁYNĘŁO
15-01-2024
RDN-IIT/73/2024

zapotrzebowania na pamięć. Ograniczenia te są szczególnie istotne przy przetwarzaniu dużych i bardzo dużych zbiorów danych. Stąd wraz ze wzrostem ilości gromadzonych i analizowanych danych, rośnie zapotrzebowanie na skalowalne i wydajne algorytmy eksploracji wzorców sekwencji oraz wzorców międzysekwencyjnych. Ponadto istotnym jest docelowe przedstawianie użytkownikowi systemu, opartego na omawianej technice eksploracyjnej, tych odnalezionych wzorców, które mogą potencjalnie być w jego zainteresowaniu. Gdy odkrywanie wzorców międzysekwencyjnych prowadzi do generowania znacznie większej liczby potencjalnych kandydatów (wzorców), niż ma to miejsce w przypadku odkrywania wzorców sekwencji, poszukiwanie efektywnych algorytmów do generowania wzorców międzysekwencyjnych odpowiadających oczekiwaniom użytkowników (lub formułowanym przez nich ograniczeniom) stanowi także aktualne wyzwanie badawcze.

Podsumowując, Doktorant w swojej pracy doktorskiej zajął się poszukiwaniem efektywnych algorytmów odkrywania wzorców międzysekwencyjnych, odpowiednich dla przetwarzania dużych zbiorów danych oraz opartych na przetwarzaniu równoległym. W zainteresowaniu Doktoranta znalazło się także zaproponowanie odpowiednich struktur danych pozwalających na efektywne przeszukiwanie danych. Należy stwierdzić, że podjęta przez Doktoranta problematyka badań (objęta recenzowaną rozprawą doktorską) jest aktualna. Aktualność ta została przez Doktoranta także dobrze uwypuklona, a sam problem badawczy został jasno sformułowany.

2. Treść rozprawy

Rozprawa doktorska została przygotowane w języku angielskim. Składa się z 109 stron, w tym:

- wprowadzenia,
- przeglądu literatury (w rozdziale 2 rozprawy),
- rozdziałów 3 i 4, w których Doktorant przedstawił proponowane rozwiązania usprawniające proces odkrywania wzorców międzysekwencyjnych,
- rozdziału podsumowującego,
- spisu treści, rysunków, tabel, algorytmów oraz spisu literatury obejmującego 62 pozycje.

Rozdział pierwszy rozprawy (Introduction) stanowi wprowadzenie, gdzie sformułowano problem badawczy oraz cele badania. Rozdział ten uwypukla także wyniki badań, które zostały w rozprawie doktorskiej przedstawione, a które zostały uzyskane w ramach sformułowanych celów.

Rozdział 2 wprowadza podstawowe definicje i oznaczenia związane z podjętym przez Doktoranta problemem badawczym. Obejmuje także przegląd algorytmów i metod odkrywania wzorców sekwencji w danych, w tym algorytmów odkrywania wzorców sekwencji z ograniczeniami oraz algorytmów odkrywania wzorców międzysekwencyjnych. Dla prezentowanych algorytmów, dyskusji poddano ich zalety oraz ograniczenia.

W rozdziale 3 rozprawy Doktorant poczynając od zobrazowania problemu odkrywania wzorców międzysekwencyjnych, wprowadził strukturę opisu danych oznaczoną jako pseudo-IDList, a następnie omówił założenia obliczeniowe dla proponowanej metody ISP-IC. Dla metody tej Doktorant przestawił także dowód słuszności eliminacji kandydujących sekwencji. W następstwie, Doktorant omówił proponowany algorytm ISP-PI, dla którego przedstawił także analizę jego złożoności obliczeniowej. W dalszej części rozdziału zaprezentowano wyniki eksperymentu obliczeniowego, którego celem było porównanie zaproponowanego algorytm ISP-PI z innymi algorytmami, w tym jego wersją podstawową EISP-Miner oraz algorytmem DBV-ISP. Dla różnych zestawów danych, algorytm ISP-PI wykazał przewagę, mając na uwadze czas obliczeń oraz wykorzystanie pamięci RAM.

Rozdział 4 rozprawy został poświęcony odkrywaniu wzorców międzysekwencyjnych z ograniczeniami. W rozdziale tym Doktorant przedstawił algorytm DBV-ISPMIC oraz jego wersję rozszerzoną DBV-ISPMIC-IMPROVING, a także wersję równoległą pDBV-ISPMIC. Algorytm DBV-ISPMIC został również poddany ocenie pod kątem jego złożoności obliczeniowej. W dalszej części rozdziału zaprezentowano wyniki eksperymentu obliczeniowego, którego celem było porównanie opracowanych algorytmów oraz ich konfrontacja z algorytmem Post-EISP-Miner (zmodyfikowaną przez wprowadzenie możliwości weryfikacji ograniczeń wersją algorytmu EISP-Miner). Dla różnych zestawów danych, algorytm DBV-ISPMIC-IMPROVING wykazał przewagę (mają na uwadze czas obliczeń oraz wykorzystanie pamięci RAM). Przewaga algorytmu DBV-ISPMIC-IMPROVING został także potwierdzona dla jego wersji równoległej.

Podsumowanie oraz wnioski z przeprowadzonych badań Doktorant zawarł w rozdziale piątym rozprawy. W rozdziale tym Doktorant odniósł się także do jakości uzyskanych wyników, wskazując ograniczenia, jakie na chwilę prezentacji wyników mogą mieć związek z ich generalizacją. Rozdział ten wskazuje także kierunki przyszłych prac.

Literatura, na którą Doktorant powołał się w rozprawie jest aktualna oraz właściwa dla podjętego problemu badawczego.

3. Najważniejsze wyniki uzyskane w pracy

Do najważniejszych osiągnięć rozprawy zaliczam:

- sformułowanie problemu odkrywania wzorców międzysekwencyjnych z ograniczeniami,
- zaproponowanie algorytmu DBV-ISPMIC dedykowanego do odkrywania wzorców międzysekwencyjnych z ograniczeniami na zbiory elementów (zdarzeń) w sekwencji oraz opartego na dedykowanej strukturze danych (DBV-PatternList) listy kandydatów dla zbioru sekwencji oraz strukturze drzewa (ISP-Tree) do przechowywania wzorców częstych,
- zaprezentowanie wersji rozszerzonej algorytmu DBV-ISPMIC (DBV-ISPMIC-IMPROVING) oraz jego wersji równoległej (pDBV-ISPMIC oraz pDBV-ISPMIC-IMPROVING),
- zaproponowanie algorytmu ISP-PI zoptymalizowanego pod kątem odkrywania wzorców międzysekwencyjnych z dużych zbiorów danych i opartego na opisie danych z wykorzystaniem pseudo-IDList, oraz metody ISP-IC służącej do ograniczania zbiorów z tzw. kandydującymi sekwencji.

Wybrane wyniki badań przedstawione w rozprawie zostały zawarte w artykułach naukowych, których Doktorant jest współautorem i które zostały opublikowane w takich czasopismach jak:

- Applied Intelligence (2 artykuły),
- Journal of Intelligence and Fuzzy Systems (2 artykuły),
- Expert Systems with Applications (2 artykuły)

oraz w serii wydawniczej Lecture Notes in Computer Science (1 artykuł). Publikacje te powstały w latach 2018-2024.

4. Uwagi i pytania do recenzowanej rozprawy

Recenzowana rozprawa doktorska nasuwa również kilka uwag i pytań szczególnie do części eksperymentalnej. Kiedy jedną z przesłanek do podjętej pracy badawczej był aspekt skalowalności, to eksperymenty zostały przeprowadzone na mało reprezentatywnych zbiorach danych. Oczekiwaniem byłoby użycie znaczenie większych zbiorów danych, w tym o znaczenie większej liczbie zdarzeń i

transakcji, a w tym większej liczbie unikalnych zbiór elementów w zbiorniach zdarzeń. Ponadto, interesującym byłaby ocena realnych korzyści z efektu zrównoleglenia przy przetwarzaniu znaczenie większych zbiorów danych, gdy te przetwarzane byłyby w środowisku obliczeń równoległych. Uwaga ta silnie koreluje także z tytułem oraz celem badań, gdzie wskazano, iż przetwarzanie dużych zbiorów danych leży u podstaw motywacji do podjętych badań.

Ważnym dla oceny proponowanych algorytmów byłoby także szersze uwypuklenie aspektu ograniczeń na zbiory elementów (zdarzeń) w sekwencji, skoro głównym aspektem badań były algorytmy odkrywania wzorców międzysekwencyjnych z ograniczeniami. Należy także dodać, że wzmocnienia wniosków i konkluzji dostarczyłoby także porównanie wyników z udziałem szerszej grupy konkurencyjnych algorytmów (wymienionych i omówionych chociażby w rozdziale 2 rozprawy).

W przypadku wersji równoległej algorytmu pDBV-ISPMIC, w tym przetwarzania danych w oparciu o strukturę ISP-Tree, zachodzi pytanie o jego implementację, zastosowaną technikę zrównoleglenia oraz technologię tej implementacji, tu szczególnie dla potrzeb prowadzenia eksperymentów. Czy implementacja ta oraz zastosowana technologia pozwoliła na realne zrównoleglenie obliczeń w wykorzystanym środowisku obliczeniowym? Ponadto zachodzi pytanie o koszt operacji sekwencyjnych, czy w ogóle mają one znaczenie dla implementacji i oczekiwanych korzyści ze zrównoleglenia.

Uszczegółowienia wymagałby też sposób pomiaru czasu oraz zużycia pamięci operacyjnej. Czy zostały zachowane odpowiednie warunki dla przeprowadzonych eksperymentów oraz czy zastosowana technika obliczeń, w tym zrównoleglenia, pozwoliła na obiektywne zebranie wyników oraz ich porównanie. Można by w tym miejscu zwrócić także uwagę na kwestię oceny statystycznej uzyskanych wyników, co wzmocniłoby sformułowane wnioski i pozwoliło na ewentualne odrzucenie wyników nieistotnych statystycznie.

Choć większość powyższych uwag została uchwycona przez Doktoranta, do czego odniósł się w zakończeniu rozprawy (Rozdział 5.2), podczas obrony interesującym byłoby przedstawienie przez Doktoranta chociażby kwestii implementacji proponowanej wersji równoległej algorytmu. Dobrze byłoby również przedstawić wersję równoległą algorytmu w postaci pseudokodu i uwypuklenie poszczególnych jego części wykonujących się sekwencyjnie i równolegle, wraz z analizą skalowalności i sprawności algorytmu. Być może udałoby się także przedstawić szersze spektrum wyników eksperymentów obliczeniowych na bardziej reprezentatywnych zbiorach danych - dla uchwycenia skalowalności oraz radzenia sobie przez proponowane algorytmy ze wspomnianymi ograniczeniami na zbiory elementów (zdarzeń) w sekwencji.

Ponadto należy dodać, że wprowadzenie w pracy listy zastosowanych oznaczeń i skrótów pozwoliłoby na sprawniejszą ocenę sposobu działania omawianych algorytmów. Pewnym ograniczeniem jest także brak ujednolicenia oznaczeń dla omawianych algorytmaów (szczególnie w rozdziale 2), a także ujednolicenia sposobu prezentacji algorytmów (uwaga dotyczy zarówno algorytmów przywoływanych jak i zaproponowanych).

5. Ocena redakcji i przygotowania rozprawy

Praca została napisane w sposób przejrzysty. Ma właściwą dla rozpraw doktorskich strukturę. Nie budzi wątpliwości strona językowa rozprawy. Drobne błędy stylistyczne i językowe nie ujmują jakości pracy.

6. Konkluzja

Uzyskane przez Doktorant wyniki badań oceniam jako wartościowe. Doktorant podjął się ważkiego problemu badawczego związanego z poszukiwaniem efektywnych algorytmów odkrywania wzorców międzysekwencyjnych przy ograniczeniach.

Rozprawa prezentuje i potwierdza ogólną wiedzę teoretyczną odpowiednią dla osoby ubiegającej się o nadanie stopnia doktora. Pomimo sformułowanych w recenzji pytań i uwag, uważam, że wyniki badań zostały zaprezentowane w sposób wystarczający dla oceny ich oryginalności i ważności dla dyscypliny informatyka techniczna i telekomunikacja, oraz ich wykorzystania w praktycznych implementacjach. Wyniki te mogą stanowić także podstawę do dalszych badań nad problemami odkrywania wzorców międzysekwencyjnych.

Ponadto stwierdzam, że Doktorant wykazał się umiejętnością samodzielnego rozwiązywania problemów badawczych, w tym doborem odpowiednich metod, potwierdził też, że posiada umiejętności związane z metodyką i metodologią prowadzenia badań naukowych.

Podsumowując, uważam, że rozprawa doktorska mgr. inż. Anh Nguyena pt. "*Mining sequence and inter-sequence patterns in large databases*" spełnia wymogi stawiane przy ubieganiu się o nadanie stopnia doktora w dyscyplinie informatyka techniczna i telekomunikacja. Wnioskuję też o dopuszczenie rozprawy doktorskiej mgr. inż. Anh Nguyena do obrony.

Ireneusz Czarnowski

Signed by / Podpisano przez:

Ireneusz Czarnowski

Date / Data: 2024-01-09 15:47