MINISTERUL EDUCAŢIEI ŞI CERCETĂRII AL REPUBLICII MOLDOVA
Universitatea Tehnică a Moldovei
Facultatea Calculatoare Informatică şi Microelectronică
Departamentul Ingineria Software și Automatică

Admis la susţinere
Şef de departament:
Fiodorov I. dr., conf. univ.

„___" _____ 2021

Aplicarea tehnicilor de Graph Mining în algoritmi de detectare a comunităților

Teza de master

Student: _____ Postolachi Valeria, TI-201M

Conducător: _____ Beșliu Victor, conf. univ. dr.

Consultant: _____ Cojocaru Svetlana, lect. univ.

Chişinău, 2022

Rezumat

Lucrarea „Aplicarea tehnicilor de Graph Mining în algoritmi de detectare a comunităților” se bazează pe teoria grafurilor pentru a scoate în evidență specificul acestora, cazurile de utilizare și metodele în care grafurile și funcțiile acestora pot fi implementate pentru rezolvarea problemelor din viața reală cu date deja existente.

Au fost analize diferite tipuri de grafuri, procesele în care acestea sunt utilizate și diverși agloritmi din categoria detecției a comunităților. Sub cuvântul „comunitate”, în cadrul teoriei grafurilor, se înțelege o colecție de noduri din interiorul grafului care au o comunicație mai puternică decât cu celelalte noduri din graf. Detecția comunităților este importantă datorită aplicației sale asupra datelor din viața reală pentru investigarea comportamentelor de grup și a fenomenelor recurente în baza diferitor procese.

În cadrul lucrării a fost utilizat un segment de date de pe transtats.bts.gov și OpenFlights pentru analiza rutelor și zborurilor, și obținerea informației și perspectivelor care nu ar putea fi cunoscute cu ajutorul unor interogări sau funcții simple SQL.

Abstract

The paper „Graph Mining for Community Detection Algorithms” builds on graph theory to highlight their specifics, use cases, and methods in which graphs and their functions can be implemented to solve real-life data problems with already existing data.

Different types of graphs, processes in which they are used and various algorithms for community detection have been analyzed. The term "community" in graph theory means a collection of nodes inside the graph that have stronger communication between each other than with the other nodes in the graph. Community detection is important because of the actual behavior of its application on life data for group investigation and recurring phenomena based on different processes.

The paper used a piece of data from transtats.bts.gov and OpenFlights to analyze routes and flights, and to obtain information and perspectives that could not be known with the help of simple SQL query functions.

CUPRINS

INTRODUCERE **8**

1 ANALIZA DOMENIULUI **10**

1.1 Varietatea grafurilor 11

1.2 Tipuri de algoritmi pe bază de graf 16

1.3 Platformele și procesarea grafurilor 17

1.4 OLTP și OLAP 21

1.5 Cazuri de utilizare pentru grafuri 22

2 ALGORITMI DE DETECTARE A COMUNITĂȚILOR **26**

2.1 Triangle Count și Coeficientul de Clustering 28

2.2 Componente Slab Conectate și Componente Puternic Conectate 30

2.3 Label Propagation 32

2.4 Modularitatea Louvain 34

3 ANALIZA DATELOR **40**

3.1 Analiza exploratorie 42

3.2 Întârzieri de zboruri 44

3.3 Aeroporturi interconectate prin linii aeriene 47

CONCLUZII **52**

BIBLIOGRAFIE **53**

ANEXA A **54**

INTRODUCERE

Lumea este condusă și se bazează pe conexiuni – de la sistemele financiare și de comunicare la procesele sociale și biologice. Posibilitatea de a dezvălui sensul din spatele acestor conexiuni generează progrese în industriile celor mai diverse domenii precum identificarea tentativelor și cazurilor de fraudă și optimizarea recomandărilor la evaluarea puterii unui grup și prezicerea eșecurilor în cascadă (*eng.* cascading failure).

Pe măsură ce conexiunile în lume continuă să accelereze, interesul pentru teoria grafurilor și algoritmii pe bază de grafuri a crescut, deoarece acestea se bazează pe matematică dezvoltată în mod explicit pentru a obține informații din relațiile dintre date. O analiză pe bază de grafuri corectă și calitativă poate descoperi funcționarea sistemelor și rețelelor complexe la scări masive pentru orice organizație.

Pentru a dispune de un punct de referință și de o imagine clară a aplicației grafurilor poate fi analizată similitudinea următoarelor lucruri: analiza atribuțiilor marketing-ului, analiza serviciilor de prevenire a spălării banilor (*eng.* AML – Anti-Money Laundering), modelarea călătoriei clienților, crearea aplicațiilor de hartă, analiza grupului de boli și analiza performanței unui film rulat – toate acestea au în comun utilizarea grafurilor. Se observă că exemplele enumerate mai sus implică entități și relațiile dintre ele, incluzând atât relații directe, cât și indirecte (tranzitive). Entitățile sunt nodurile din graf – acestea pot fi persoane, evenimente, obiecte, concepte sau locuri. Relațiile dintre noduri sunt arcurile din graf. Astfel, esența unui fenomen este în sine reprezentarea activă a entităților (nodurilor) și a relațiilor lor (arcurilor).

Ceea ce face ca algoritmii de graf și bazele de date ale acestora să fie atât de interesante și puternice nu este simpla relație dintre două entități, A fiind legat de B. Modelul relațional standard al bazelor de date a inițiat aceste tipuri de relații cu zeci de ani în urmă, în diagrama de relații între entități (ERD). Ceea ce face ca grafurile să fie atât de importante sunt relațiile binare și relațiile tranzitive. În relațiile binare, A poate provoca B, dar nu și contrariul. În relațiile tranzitive, A poate fi legat direct de B și B poate fi legat direct de C, în timp ce A nu este legat direct de C, astfel încât, în consecință, A este legat tranzitiv de C.

Pentru a oferi un exemplu relevant, se va reveni la unul din exemplele enumerate anterior – se ia în considerare un caz de analiză a combaterii spălării banilor (AML): persoanele A și C sunt suspectate de trafic ilicit. Orice interacțiune între cele două (de exemplu, o tranzacție financiară într-o bază de date financiare) ar fi semnalată de autorități și ar fi analizată cu atenție. Cu toate acestea, dacă A și C nu efectuează niciodată tranzacții împreună, ci în schimb desfășoară tranzacțiile financiare prin intermediul

autorității financiare B, sigură, respectată și nesemnalizată, tranzacția în sine nu ar dispune de informații critice cazului dat. În acest caz ar putea fi utilizată o aplicație pe bază de grafuri care ar descoperi relația tranzitivă dintre A și C prin intermediarul nodului B – care inițial poate fi omis.

Cu aceste relații de tranzitivitate, în special atunci când acestea sunt numeroase și diverse, cu multe modele posibile de relații/ rețele și grade de separare între entități, modelul grafurilor descoperă relații între entități care altfel ar putea părea deconectate sau fără legătură și care nu sunt detectate de un model relațional. Prin urmare, modelul grafurilor poate fi aplicat în mod productiv și eficient în multe cazuri de utilizare a analizei rețelelor.

Atunci când se ia în considerare puterea grafurilor, ar trebui să se țină cont de asemenea de faptul că posibil cel mai puternic nod dintr-un model de graf pentru cazurile de utilizare din lumea reală ar putea fi contextul. Contextul poate include timpul, locația, evenimentele conexe, entitățile din apropiere și multe altele. Încorporarea contextului în graf (în formă de noduri și muchii) poate produce astfel capacități impresionante de analiză predictivă și de analiză prescriptivă (ultimele două faze ale analizei de business).

Utilitatea și importanța analizei pe bază de grafuri este un studiu pasionant condus de necesitatea de a descoperi funcționarea interioară a scenariilor complexe. Până de curând, adoptarea analizei grafurilor necesita o expertiză și determinare semnificative, deoarece instrumentele și integrările erau dificile și puțini știau cum să aplice algoritmii pe bază de grafuri la problemele lor. Scopul acestei laturi a matematicii este de a ajuta organizațiile și comunitățile să profite mai bine de analiza pe bază de grafuri, astfel încât să poată face noi descoperiri și să dezvolte mai rapid soluții inteligente.

BIBLIOGRAFIE

[1] Teoria Grafurilor, Curs Introductiv. Disponibil la: https://profs.info.uaic.ro/~vcosmin/pagini/resurse_pregatire/resurse/graf_definitii.pdf

[2] NEEDHAM, Mark, HODLER, Amy E. Graph Algorithms. United States of America: O'Reilly Media, Inc., 2019. 30 p. ISBN 978-1-492-05781-9

[3] Using Graph Theory to Build A Simple Recommendation Engine. Disponibil: https://keithwhor.medium.com/using-graph-theory-to-build-a-simple-recommendation-engine-in-javascript-ec43394b35a3

[4] AL-TAIE, Mohammed Zuhair, KADRY, Seifedine. Python for Graph and Network Analysis. Springer International Publishing AG, 2017. 24 p. ISBN 978-3-319-53003-1

[5] Graph-based Data Mining: A New Approach for Data Analysis. Disponibil: https://www.ijser.org/researchpaper/Graph-based-Data-Mining-A-New-Approach-for-Data-Analysis.pdf

[6] Pregel: A System for Large-Scale Graph Processing. Disponibil: https://www.dcs.bbk.ac.uk/~dell/teaching/cc/paper/sigmod10/p135-malewicz.pdf

[7] DILLON, Tharam S. Mining Of Data With Complex Structures. Springer, 2011. 57 p. ISBN 139783642175565

[8] Graph Mining Algorithms for Memory Leak Diagnosis and Biological Database Clustering. Disponibil: https://vtechworks.lib.vt.edu/bitstream/handle/10919/34008/Maxwell_EK_T_2010.pdf