Primljen: 25.09.2024.
Prihvaćen: 01.10.2024.

Stručni rad
UDK: 004.921

Grafovske tehnologije u podatkovnoj znanosti: primjena *Neo4j* i *Cypher* jezika

Graph Technologies in Data Science: Application of Neo4j and Cypher Language

[1]Martina Šuman, [2]Sabrina Šuman, [3]Bruno Polonijo
[1]Rinels d.o.o, Grabovac 4, 51000 Rijeka
e-mail: [1]mart.suman@gmail.com
[2]Veleučilište u Rijeci, Vukovarska 58, 51000 Rijeka
e-mail: [2]ssuman@veleri.hr, [3]bpolonijo@veleri.hr

Sažetak: *Grafovi omogućuju intuitivan i vizualno jasan prikaz podataka, lako uočavanje ključnih značajki grupacija entiteta ili anomalija koje u tabličnom formatu nisu očite. Omogućuju bolje razumijevanje veza među entitetima, otkrivanje skrivenih uzoraka i identifikaciju ključnih entiteta unutar mreže podataka. Rad istražuje ulogu i primjenu tehnologija zasnovanih na grafovskom modelu podataka u području podatkovne analitike, daje pregled novih trendova, opisuje praktičnu primjenu, analitiku i vizualizaciju podataka uz korištenje Neo4j okruženja, te primjere upotrebe Cypher jezika za grafovske baze podataka.*

Ključne riječi: *grafovske baze podataka, vizualizacija grafova, GDS, Neo4j, Cypher*

Abstract: *Graphs provide an intuitive and visually clear way of representing data, enabling easy identification of key features, entity groupings, or anomalies that may not be apparent in traditional tabular formats. They enable a more detailed understanding of the relationships between entities, uncover hidden patterns, identify key entities within the data network. This paper explores the role and application of graph-based data modeling technologies in the field of data analytics. It provides an overview of new trends in data analytics, with a particular focus on graph databases as efficient solutions for managing and analyzing networked data. It also describes business applications of graph technologies and presents examples of data analysis and visualization using the Neo4j environment with Cypher, a query language specifical for graph databases.*

Key words: *GDS (Graph Data Science), Graph Databases, Graph Visualisation, Neo4j, Cypher*

1. Uvod

Znanost o podatcima (Data Science) povezuje velike baze podataka (Big Data) s donošenjem odluka kombinirajući računalne znanosti, matematiku, statistiku i domensko znanje (Hazzan i Mike, 2020). Cilj znanosti o podatcima otkrivanje je korisnih uzoraka u velikim skupovima podataka zbog donošenja odluka (Kelleher i Tierney, 2021).

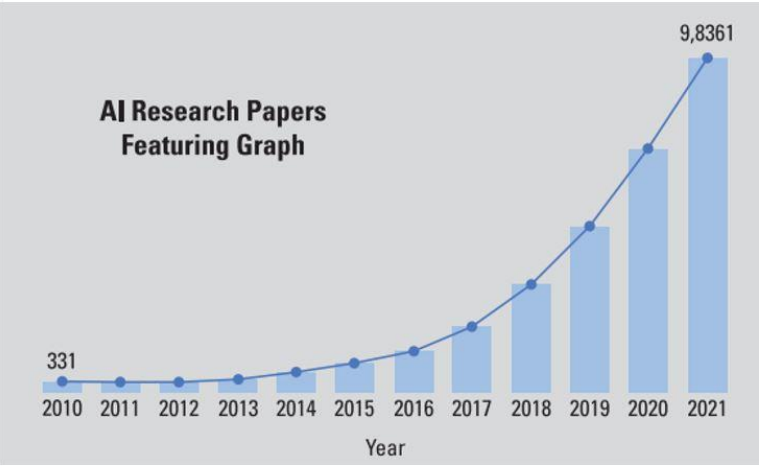
S obzirom na sve veći obujam podataka, analiza visoko povezanih podataka postala je ključna (Park i sur., 2024). Grafovski modeli nude rješenja za analizu odnosa među podatcima (Timon – Reina i sur., 2021), omogućujući intuitivnu analizu veza, otkrivanje uzoraka i identifikaciju ključnih entiteta.

Cilj rada je dati pregled primjene grafovskih tehnologija i trendova u podatkovnoj znanosti, te prikaz primjera korištenja Neo4j i Cypher.

2. Trendovi u podatkovnoj znanosti- analitika povezanih podataka

Razvoj NoSQL i grafovskih baza podataka potaknut je pojavom velikih podataka (eng. Big data) (Cucen, 2021). Grafovske baze jasno prikazuju povezanost entiteta. Broj istraživačkih radova o primjeni grafova u umjetnoj inteligenciji raste (Frame i Blumenfeld, 2022). Kompanije poput NASA-e, IBM-a i eBaya koriste grafovske tehnologije za analizu odnosa (Neo4j, Customers, 2024). Do 2025. godine očekuje se da će 80 % inovacija u području podataka koristiti.

Slika 1. Frekvencija publikacija s temom grafovske tehnologije



Izvor: Frame, Blumenfeld, 2022

3. Osnovni koncepti teorije grafova i njihova primjena danas

Teorija grafova proučava grafove - matematičke strukture koje ilustriraju odnose među objektima, pružajući temelj za razumijevanje složenih mreža (Divjak, Lovrenčić, 2005). Grafovi modeliraju različite situacije, pohranjujući podatke u obliku vrhova (čvorova) i bridova (veza) s atributima. Primjeri upotrebe grafova uključuju:

- detekcija prijevara: analiza poveznica omogućuje brže i preciznije otkrivanje prijevara;
- nadzor mreže i infrastrukture: monitoring i optimizacija mrežnih resursa;
- preporuka proizvoda u stvarnom vremenu: preporuka temeljena na korisničkim interakcijama i preferencijama;
- *Master Data Management* (MDM): integracija podataka iz različitih izvora;
- društvene mreže: razumijevanje i interpretacija međuljudskih odnosa i dinamike;
- zdravstvo: grafički prikazi napretka bolesti i posjeta pacijenata - dublji uvid u medicinske podatke;
- upravljanje lancem nabave: identifikacija ključnih točaka i optimizacija logističkih ruta;
- formaliziranje baza znanja: stvaranje strukturiranih baza podataka koje olakšavaju povezivanje i pretragu informacija (Knežević, 2022; Neo4j *Use Cases*).

Slika 2. Primjer grafa s atributima, vezama i čvorovima

Izvor: obrada autora

4. Osnovni koncepti teorije grafova i njihova primjena danas

Teorija grafova proučava grafove - matematičke strukture koje ilustriraju odnose među objektima, pružajući temelj za razumijevanje složenih mreža (Divjak, Lovrenčić, 2005). Grafovi modeliraju različite situacije, pohranjujući podatke u obliku vrhova (čvorova) i bridova (veza) s atributima. Primjeri upotrebe grafova uključuju:

- detekcija prijevara: analiza poveznica omogućuje brže i preciznije otkrivanje prijevara;
- nadzor mreže i infrastrukture: monitoring i optimizacija mrežnih resursa;
- preporuka proizvoda u stvarnom vremenu: preporuka temeljena na korisničkim interakcijama i preferencijama;
- *Master Data Management* (MDM): integracija podataka iz različitih izvora;
- društvene mreže: razumijevanje i interpretacija međuljudskih odnosa i dinamike;
- zdravstvo: grafički prikazi napretka bolesti i posjeta pacijenata - dublji uvid u medicinske podatke;
- upravljanje lancem nabave: identifikacija ključnih točaka i optimizacija logističkih ruta;
- formaliziranje baza znanja: stvaranje strukturiranih baza podataka koje olakšavaju povezivanje i pretragu informacija (Knežević, 2022; Neo4j *Use Cases*).

Slika 3. Primjer grafa s atributima, vezama i čvorovima

Izvor: obrada autora

Slika 2. prikazuje mrežu korisnika na društvenoj mreži gdje su korisnici predstavljeni kao čvorovi s atributima (npr. ime: Petar) i povezani su vezama koje označavaju da jedan korisnik prati drugoga. Ovaj pristup omogućuje intuitivno i efikasno mapiranje i analizu odnosa među korisnicima, ilustrirajući snagu i fleksibilnost grafovskih modela u razumijevanju složenih mreža interakcija.

1. Tipovi grafovskih baza podataka

Grafovske baze dijele se na nativne i ne-nativne. Nativne pohranjuju podatke u grafovskom modelu i koriste "susjedstvo bez indeksa" za bržu obradu, dok ne-nativne baze

pohranjuju grafove unutar drugih modela podataka (Matijašević, 2021; Kollegger, 2016).

Najčešći modeli su Labeled Property Graph (LPG) i RDF Triple Stores. LPG model sadrži čvorove i veze s atributima, dok RDF koristi trojke: subjekt, predikat i objekt (Vettrivel, 2022).

Slika 4. Popularnost grafovskih baza podataka

Rank			DBMS	Database Model
May 2024	Apr 2024	May 2023		
1.	1.	1.	Neo4j	Graph
2.	2.	2.	Microsoft Azure Cosmos DB	Multi-model
3.	3.	3.	Aerospike	Multi-model
4.	4.	4.	Virtuoso	Multi-model
5.	5.	5.	ArangoDB	Multi-model
6.	↑ 7.	↑ 11.	GraphDB	Multi-model
7.	↓ 6.	↓ 6.	OrientDB	Multi-model
8.	8.	↑ 9.	Memgraph	Graph
9.	9.	↓ 7.	Amazon Neptune	Multi-model
10.	10.	10.	NebulaGraph	Graph

Izvor: https://db-engines.com/en/ranking/graph+dbms

Prema *DB-Engines*, *Neo4j* je najpopularnija nativna grafovska baza koja koristi LPG model. Od baza podataka koje su dizajnirane isključivo za rad s grafovskim modelom podataka ističe se *Memgraph*, baza podataka u C++ jeziku s *in-memory* infrastrukturom, omogućava brzu analitiku u realnom vremenu (*Memgraph, Graph Database Performance Benchmark*).

5. Neo4j i Cypher

Neo4j nudi napredne funkcionalnosti za upravljanje bazama podataka, uključujući *Neo4j Bloom* i *Graph Data Science* (GDS). *Neo4j* je odabran zbog kvalitetne dokumentacije, jednostavne instalacije i aktivne zajednice.

Za lokalnu analizu korišten je *Neo4j Desktop* 1.5.8 s *Neo4j Browserom* za pisanje upita pomoću *Cypher* jezika. Instalirane su GDS i APOC biblioteke te *Neo4j Bloom* za vizualizaciju.

Cypher je službeni jezik za upite nad *Neo4j* bazom, inspiriran SQL-om i razvijen za standardizaciju jezika, omogućuje intuitivan rad i bolje performanse pri složenim upitima.

6. Primjena grafovskih metoda

Primjenu grafovskih algoritama prikazat će se kroz: *Neo4j Bloom, Neo4j NEuler* i kreiranje projekcije grafa.

6.1. Metode primjene algoritma

- *Neo4j Bloom* omogućuje odabir podskupa grafa i primjenu željenog algoritma. Rezultati su odmah vidljivi i mogu se prilagoditi promjenom boja čvorova, veza i ikona. Vrijednosti su privremeno dostupne tijekom vizualizacije.
- *Neo4j NEuler* nudi sučelje bez kodiranja za izvršavanje algoritama iz GDS biblioteke, konfiguraciju parametara, pohranu rezultata i generiranje *Cypher* koda za kasniju reprodukciju.

Projekcija grafa uključuje kreiranje imenovanoga podskupa grafa i filtriranje relevantnih čvorova, veza i atributa. Algoritmi se izvršavaju kao GDS procedure s opcijama za prikaz statistike, pohranu rezultata u bazu ili projicirani graf.

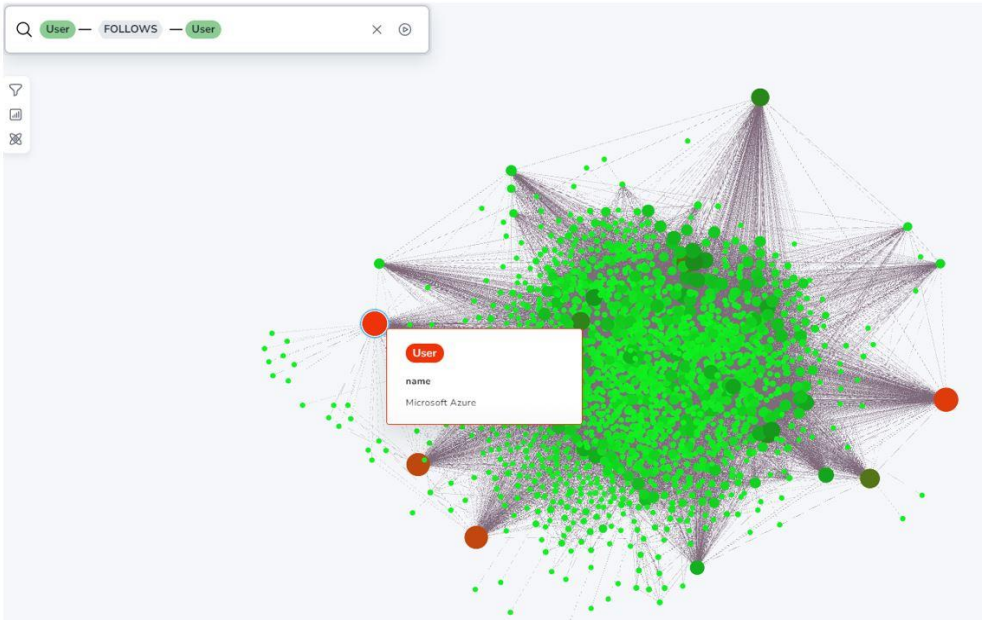
6.1.1. Primjena mjere stupnja vrha (*Degree Centrality*)

Na podskupu grafa s 1807 čvorova i 43738 veza, koji je generiran korištenjem podataka korisnika s društvene mreže Twitter/X, primijenjena je mjera stupnja vrha za identifikaciju najutjecajnijih korisnika:

1. konfiguracija algoritma s obrnutim smjerom veze *FOLLOWS*.
2. izračun ocjene (*score*) za svaki čvor, predstavljajući broj pratitelja
3. vizualizacija rezultata prilagodbom boje i veličine čvorova prema ocjeni

Rezultat ove analize prikazan je na slici 4. Ističe se nekoliko utjecajnih korisnika, s *Microsoft Azure* kao najutjecajnijim. Veći i tamniji čvorovi predstavljaju korisnike s većim brojem pratitelja, dok manji i svjetliji čvorovi predstavljaju manje utjecajne korisnike.

Slika 5. Izgled vizualizacije nakon primjene algoritma



Izvor: obrada autora u Neo4j Bloom alatu

6.1.2. Primjena Louvainove metode za otkrivanje zajednica

Louvainova metoda je hijerarhijski algoritam kreiranja klastera koji optimizira modularnost grafa. Proces uključuje:

1. raspoređivanje čvorova u male zajednice.
2. spajanje čvorova unutar zajednica u jedan čvor.
3. ponavljanje procesa dok zajednice ne postanu stabilne.

Primjena metode uključuje:

1. kreiranje projekcije grafa "Louvain" koristeći vezu *FOLLOWS* između čvorova tipa User (Slika 5).
2. pokretanje algoritma u *"Write"* načinu rada, zapisujući identifikator zajednice (*IDcommunity*) za svaki čvor
3. analiza rezultata, tj. formiranje 890 zajednica (Slika 7).

Slika 6. Kreiranje projekcije grafa naziva Louvain

```
CALL gds.graph.project(
    'Louvain',
    'User',
    {
        FOLLOWS: {
            orientation: 'UNDIRECTED'
        }
    }
)
```

Izvor: obrada autora

Za dodatnu analizu, korišten je *Cypher* upit prikazan na slici 6 za identifikaciju korisnika s najviše veza unutar svake zajednice. Upit omogućuje pronalaženje najpovezanijih korisnika unutar odabranih zajednica, pružajući uvid u ključne aktere.

Slika 7. Primjena Louvain-ove metode za otkrivanje zajednica

```
MATCH (u:User)-[r]-()
WHERE u.IDcommunity IN [2632, 3285, 860, 2295, 678]
WITH u.IDcommunity AS idCommunity, u, count(r) AS brojVeza
ORDER BY idCommunity, brojVeza DESC
WITH idCommunity, COLLECT({user: u, brojVeza: brojVeza})[0] AS najviseVeza
RETURN najviseVeza
```

Izvor: obrada autora

Rezultati primjene Louvain algoritma prikazani su na slici 7, gdje je formirano 890 zajednica. Grafikon prikazuje top 5 zajednica po broju korisnika, što omogućava uvid u najaktivnije i najpovezanije grupe unutar mreže.

Slika 8. Rezultati Louvain algoritma

```
CALL gds.louvain.write('Louvain', { writeProperty: 'IDcommunity' })
YIELD communityCount, modularity, modularities
```

	communityCount	modularity	modularities
1	890	0.42666343398113166	[0.401522611763479, 0.42653080933496607, 0.42666343398113166]

```
MATCH (u:User)
RETURN u.IDcommunity AS idCommunity, count(u) AS brojKorisnika
ORDER BY brojKorisnika DESC
LIMIT 5
```

	idCommunity	brojKorisnika
1	2632	1491
2	3285	1416
3	860	805
4	2295	536
5	678	487

Izvor: obrada autora u Neo4j Browser alatu

Vizualni prikaz zajednica otkrivenih Louvainovom metodom prikazan je na slici 8. Različite boje predstavljaju različite zajednice, dok veličina čvorova može predstavljati broj veza ili važnost unutar zajednice. Ovaj prikaz pomaže u razumijevanju strukture mreže i identifikaciji ključnih grupa.

6.1.3. Analiza rezultata

Analiza je otkrila sljedeće:

1. identifikacija utjecajnih korisnika: *Microsoft Azure* je najutjecajniji korisnik u mreži, vidljivo iz veličine i boje čvora na Slici 4.
2. struktura zajednica: Formirano je 890 zajednica, što ukazuje na kompleksnu strukturu mreže s brojnim podgrupama.
3. predstavnici zajednica: Među profilima koji predstavljaju zajednice nalaze se *Neo4j, TigerGraphDB* i *Amazon Web Services*, tj. zajednice se često formiraju oko poznatih brendova ili interesnih područja.
4. modularnost: njeno povećanje uzastopnim izvođenjem algoritma ukazuje na jasno definiranu strukturu zajednica.

Ovo pokazuje sposobnost otkrivanja skrivenih obrazaca i struktura unutar kompleksnih mreža, dublje razumijevanje dinamike interakcija, identifikaciju ključnih aktera i mapiranje interesnih zajednica.

Korištenje *Neo4j alata, Bloom*-a za vizualizaciju i *NEuler*-a za jednostavnu primjenu algoritama olakšava proces analize i interpretacije rezultata što u kombinaciji s *Cypher* jezikom pruža moćan okvir za istraživanje i analizu povezanih podataka u različitim domenama.

7. Zaključak

Analitika podataka putem grafova danas se primjenjuje u različitim područjima. Rad ističe trendove u podatkovnoj znanosti, analitici povezanih podataka, osnovne koncepte teorije grafova i njihovoj primjeni. Predstavljena je klasifikacija grafovskih baza podataka i rang - lista najpopularnijih u 2024. godini.

Prikazani su primjeri izrade analitike podataka s fokusom na Neo4j koji omogućuje pohranu i čitanje, ali i analitiku i vizualizaciju podataka. Demonstrirana je vizualizacija analize društvene mreže Twitter/X korištenjem alata *Neo4j Bloom* i *Cypher* upita.

Prednost korištenja grafovskih baza podataka je efikasnost analize visokopovezanih ili duboko hijerarhijskih podataka, zahvaljujući svojstvu susjedstva bez indeksa koje eliminira potrebu za pisanjem dugih i resursno zahtjevnih složenih upita. Tradicionalna analitika relacijskih podataka ne može biti potpuno zamijenjena analitikom grafova, ali važno je iskoristiti prednosti oba pristupa kad je to primjenjivo. Buduća istraživanja mogla bi se fokusirati na integraciju procesa strojnoga učenja s grafovskim metodama.

Literatura

[1] Cucen, E. (2021): Connecting the Dots: Harness the Power of Graphs & ML dostupno na: https://opencredo.com/connect-the-dots-harness-the-power-of-graphs-ml-ebook/ (5.6.2023.)

[2] Divjak, B., Lovrenčić, A. (2005): Diskretna matematika s teorijom grafova. Varaždin: Fakultet organizacije i informatike

[3] Frame, A, Blumenfeld, Z. (2022): Graph Data Science 2nd Neo4j Special Edition

[4] Gartner (2021): Gartner Identifies Top 10 Data and Analytics Technology Trends for 2021, dostupno na: https://www.gartner.com/en/newsroom/press-releases/2021-03-16-gartner-identifies-top-10-data-and-analytics-technologies-trends-for-2021 (15.05.2023.)

[5] Hazzan, O., Mike, K. (2020): Ten Challenges of Data Science Education, dostupno na: https://cacm.acm.org/blogs/blog-cacm/246219-ten-challenges-of-data-science-education/fulltext (30.5.2023)

[6] Kelleher, J., Tierney, B. (2021): Znanost o podacima, Zagreb: Mate d.o.o..

[7] Knežević, T, (2022): Poslovne primjene grafova - Vizualizacija povezanih podataka i detaljan uvid u njihove međuodnose, dostupno na: https://www.bug.hr/biznis/poslovne-primjene-grafova-vizualizacija-povezanih-podataka-i-detaljan-uvid-u-28271 (15.5.2023)
[8] Kollegger, A. (2016): Graph Databases for Beginners: Native vs. Non-Native Graph Technology, dostupno na: https://dzone.com/articles/graph-databases-for-beginners-native-vs-non-native (18.2.2024.)
[9] Matijašević, M. (2021): Grafovska NoSQL baza podataka kao izvještajni sustav, Diplomski rad, Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet, dostupno na: https://urn.nsk.hr/urn:nbn:hr:217:733445 (6.6.2023.)
[10] Memgraph (2023): Graph Database Performance Benchmark, https://memgraph.com/white-paper/performance-benchmark-graph-databases (26.6.2023.)
[11] Neo4j, Customers (2024.): dostupno na: https://neo4j.com/customers/ (1.4.2024.)
[12] Neo4j (2023): Graph Databases for Beginners: Why Graph Technology Is the Future, dostupno na: https://neo4j.com/blog/why-graph-databases-are-the-future/ (15.6.2023)
[13] Neo4j (2023.): Louvain algorithm, dostupno na: https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/, (09.07.2023)
[14] Neo4j (2023): Use cases, dostupno na: https://neo4j.com/use-cases/ (1.6.2023.)
[15] Park, S., Lee, Y., & Yu, K. (2024): Integrated knowledge graph construction framework for places-of-interest retrieval using a property graph database. GIScience and Remote Sensing, 61(1). https://doi.org/10.1080/15481603.2024.2331861
[16] Sasaki, B. (2018): Graph Databases for Beginners: Other Graph Technologies, dostupno na: https://neo4j.com/blog/other-graph-database-technologies/(24.5.2023.)
[17] Timón-Reina, S., Rincón, M., & Martínez-Tomás, R. (2021):. An overview of graph databases and their applications in the biomedical domain. In Database (Vol. 2021). Oxford University Press. https://doi.org/10.1093/database/baab026
[18] Vettrivel, V. (2022): Knowledge Graphs: RDF or Property Graphs, Which One Should You Pick?, dostupno na: https://www.wisecube.ai/blog/knowledge-graphs-rdf-or-property-graphs-which- one-should-you-pick