■ *Študijsko gradivo*

Andrej Kastrin

Omrežja znanja in njihova uporaba v biomedicini

Povzetek. Učno gradivo celovito predstavi področje proučevanja omrežij znanja. Omrežje znanja je formalno definirano kot heterogeno omrežje, sestavljeno iz vozlišč in povezav različnih semantičnih tipov. Na kratko je predstavljena zgodovina raziskovanja omrežij znanja, formalna definicija in temeljne lastnosti. Predstavljeni so primeri uporabe omrežij znanja na področju biomedicine, vključno s pregledom virov podatkov, metodami konstrukcije omrežja (luščenje entitet/relacij, normalizacija in integracija omrežij) ter reprezentacijskim učenjem nad njimi.

Ključne besede: reprezentacija znanja; semantična mreža; biomedicina; viri podatkov; besedilno rudarjenje; algoritmi.

Knowledge Networks and Their Use in Biomedicine

Abstract. The tutorial provides a comprehensive introduction to the field of knowledge networks. Knowledge network is defined as a heterogeneous network consisting of nodes and relations of different semantic types. The history of knowledge networks research, formal definition, and basic properties are briefly presented. Examples of the application of knowledge networks in the biomedical domain are provided, including possible data sources, construction methods (entity/relationship extraction, normalisation, and network integration), and representation learning.

Key words: knowledge representation; semantic network; biomedicine; data sources; text mining; algorithms.

■ **Infor Med Slov** 2022; 27(1-2): 44-50

Institucije avtorjev / Authors' institutions: Medicinska fakulteta, Univerza v Ljubljani.

Kontaktna oseba / Contact person: doc. dr. Andrej Kastrin, Univerza v Ljubljani, Medicinska fakulteta, Vrazov trg 2, 1000 Ljubljana, Slovenija. E-pošta / E-mail: andrej.kastrin@mf.uni-lj.si.

Prispelo / Received: 21. 1. 2023. Sprejeto / Accepted: 28. 3. 2023.

Uvod

Odgovor na vprašanje, kaj je znanje, ni preprost. Že bežen pregled literature razkriva paleto različnih opredelitev pojma. V nadaljevanju bomo privzeli preprosto definicijo, da je znanje skupek urejenih informacij, ki nam omogočajo razumevanje obravnavanega pojava.

Najbrž se bo vsakdo strinjal, da je ustrezen način organizacije znanja ključen, da uspešno opravimo dano nalogo. Študent, ki se na zahteven izpit pripravlja sistematično in za ponavljanje uporablja različne miselne sheme, bo predvidoma dosegel boljši uspeh kot njegov kolega, ki je študijsko gradivo le bežno pregledal. Bolj učeno lahko rečemo, da je študentov uspeh povezan z uspešnostjo njegove reprezentacije oz. predstavitve znanja. Boljša kot bo predstavitev znanja v študentovem spominu, bolje ga bo ta razumel, laže ga bo dopolnjeval in o njem bolj poglobljeno razmišljal. Tako osvojeno znanje bo tudi bolj modularno, saj bo lahko posamezne dele uporabil kot gradnike, na osnovi katerih bo razširjal svoje védenje v druge problemske domene.

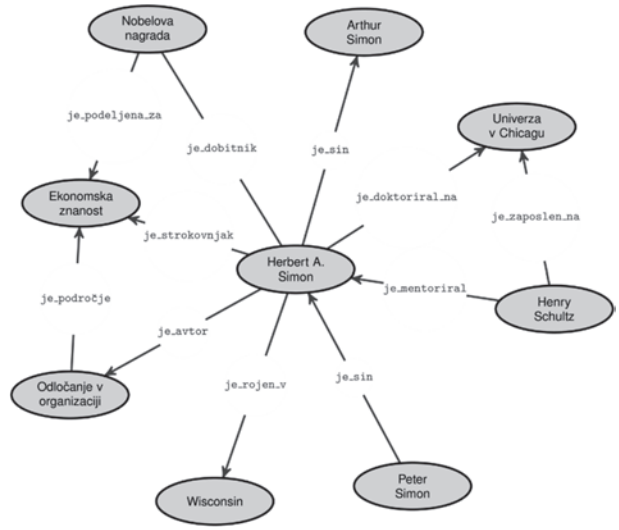
Tematika predstavitev znanja je danes v središču pozornosti na področjih kognitivne znanosti in umetne inteligentnosti. Če vzamemo v roke sodoben učbenik kognitivne psihologije[1] ali umetne inteligentnosti,[2] ugotovimo, da so vsebine, povezane s področjem reprezentacij znanja, praviloma obravnavane v samostojnem poglavju. Na podoben način kot v človekovih možganih je potrebno za uspešno reševanje nalog predstaviti znanje tudi stroju. Poznamo več različnih pristopov k reprezentaciji znanja. Tako ločimo med predstavitvami s (i) pravili, (ii) semantičnimi mrežami, (iii) scenariji oz. skripti ter (iv) okviri. Med njimi je najbrž najbolj znana semantična mreža. To je preprost model za reprezentacijo znanja, človeku je lahko razložljiv, enostavno ga je predstaviti tudi računalniku. Opišemo jo z množico entitet in množico relacij med njimi; vsako entiteto običajno opremimo s seznamom lastnosti, ki jo natančneje določajo, pomen relacije pa predstavimo z njenim tipom.

To gradivo ima dva namena. Prvič, v slovenščini želimo predstaviti nekaj osnovnih konceptov, na katerih gradi sodobno proučevanje omrežij znanja, zlasti na področju biomedicine in znanosti o živem. In drugič, čim več bralcev želimo spodbuditi, da tudi sami pokukajo v svet raziskovanja omrežij. Zato mu ponujamo nekaj kazalcev, ki mu utegnejo priti prav pri nadaljnjem študiju. Jedro gradiva predstavljajo trije razdelki, v katerih obravnavamo podatkovne vire, postopek gradnje omrežja ter osvetlimo pristop, ki omogoča enostavno transformacijo relacijskega podatkovja v obliko, ki je primerna za zagon različnih metod statističnega učenja. Zaključimo s pregledom najpomembnejših izzivov za prihodnost.

Omrežje znanja

Za razumevanje notacije v nadaljevanju moramo na hitro ponoviti oziroma vpeljati nekaj osnovnih pojmov iz teorije grafov. Graf je matematična struktura, s katero predstavimo množico entitet in v kateri so izbrane dvojice entitet medsebojno povezane. Entiteto upodobimo z vozliščem (angl. *node*), relacijo med dvema entitetama pa bodisi z usmerjeno (angl. *arc*) bodisi neusmerjeno (angl. *edge*) povezavo. Omrežje je graf, opremljen s podatki.

Besedno zvezo "omrežje znanja" danes povezujemo s širokim naborom aktivnosti. Neposredno se z omrežjem znanja srečujemo ob pregledovanju Wikipedie, uporabi Twitterja ali pri načrtovanju novega biooznačevalca za Alzheimerjevo bolezen. Primer je prikazan na sliki 1.



Slika 1 Izsek iz omrežja znanja Nobelovega nagrajenca Herberta A. Simona (1916–2001), pionirja sodobnega pojmovanja umetne inteligentnosti.

Enoznačne definicije pojma "omrežje znanja" v literaturi ne bomo našli. Paulheim[3] ponuja seznam kriterijev, na podlagi katerih presojamo, ali dejansko obravnavamo omrežje znanja. Najpomembnejša med njimi sta:

- posamezne elemente realnosti (tj. entitete) ter interakcije med njimi (tj. relacije) lahko predstavimo s pomočjo grafa;
- poznamo t. i. metashemo, na osnovi katere lahko opredelimo dovoljene tipe relacij med entitetami.

Ehrlinger in Wöß[4] pravita, da je omrežje znanja "namenjeno integraciji informacij v ontologijo in omogoča luščenje novega znanja". Wang in sodelavci[5] pa eksplicitno definirajo omrežje znanja kot heterogeno omrežje, v katerem lahko vozliščem in povezavam določimo različne tipe.

V tem gradivu bomo omrežje znanja formalno definirali z množico trojčkov ⟨*glava*, *relacija*, *rep*⟩ oz. krajše ⟨*h*, *r*, *t*⟩, s katerimi opišemo relacijo *r* med začetno entiteto *h* in končno entiteto (ali atributom) *t*. Krajša oblika izhaja iz prvih črk angleških izrazov *head*, *relation* in *tail*. Ločimo dva tipa trojčkov: z njimi lahko (i) opišemo relacijo med entitetama, tj. ⟨$entiteta_1$, *relacija*, $entiteta_2$⟩, ali (ii) entiteto opremimo z atributom in pripadajočo vrednostjo, tj. ⟨*entiteta*, *atribut*, *vrednost*⟩. Trojček

⟨*H. A. Simon, področje dela, umetna inteligentnost*⟩

je torej prvega tipa, saj vzpostavlja relacijo med dvema entitetama, trojček

⟨*Univerza v Ljubljani, št. študentov, 39.010*⟩

pa entiteto Univerza v Ljubljani razširi z atributom, na podlagi katerega dobimo informacijo o številu vpisanih študentov. Formalno bomo omrežje znanja predstavili kot $G = (E, R, A, V, T^R, T^A)$, kjer je E množica entitet, R množica relacij, A množica atributov in V množica vrednosti atributov. Relacije med pari entitet bomo potemtakem predstavili z množico $T^R \subseteq E \times R \times E$, z množico $TA \subseteq \mathrm{E} \times \mathrm{A} \times \mathrm{V}$ pa analogno množico entitetam pripisanih atributov.

Na področju biomedicine je bilo prvo odmevno omrežje znanja predstavljeno pred poldrugim desetletjem, ko so Belleau in sodelavci[6] podatke iz prosto dostopnih podatkovnih zbirk – pretežno s področja bioinformatike – prepisali iz klasične tabelarne oblike v zapis RDF (angl. *Resource Description Framework*). Za ilustracijo so v tabeli 1 povzete osnovne lastnosti nekaterih najpogosteje uporabljenih omrežij znanja na področju biomedicine.

Tabela 1 Osnovne lastnosti nekaterih najpogosteje uporabljanih omrežij znanja na področju biomedicine.

Omrežje	Št. entitet	Št. relacij	Št. tipov entitet	Št. tipov relacij	Posodobitev
Hetionet	$47 \cdot 10^3$	$2{,}3 \cdot 10^6$	11	24	2017
DRKG	$97 \cdot 10^3$	$5{,}9 \cdot 10^6$	13	107	2000
BioKG	$105 \cdot 10^3$	$2 \cdot 10^6$	10	17	2017
PharmKG	$8 \cdot 10^3$	$501 \cdot 10^3$	3	29	2017
OpenBioLink	$185 \cdot 10^3$	$4{,}8 \cdot 16^6$	7	30	2017
Clinical Knowledge Graph	$19 \cdot 10^6$	$217{,}3 \cdot 10^6$	36	47	2017

Viri podatkov

Kakovostni viri podatkov so za gradnjo omrežja znanja ključnega pomena. Sledi pregled treh skupin podatkovnih virov, ki so bili doslej uporabljeni pri izdelavi omrežij znanja na področju biomedicine. To so (i) ontologije in terminologije, (ii) zdravstveni ipd. zapisi in (iii) obstoječe zbirke podatkov.

Ontologije in terminologije

Nujen (seveda pa ne zadosten) pogoj za reševanje nalog, povezanih s strojno obdelavo besedil, je zbirka pojmov, ki posamezen termin preslika v ustrezno pojmovno oznako. Na področju biomedicine in znanosti o živem to nalogo opravlja sistem UMLS (angl. *Unified Medical Language System*), ki ga sestavljajo tri komponente (slovenske ustreznice in pojasnila k posameznim komponentam podaja Vintar[7]):

1. Metatezaver (angl. *Metathesaurus*) – v trenutni različici je sestavljen iz približno 4,5 milijona strokovnih pojmov in okrog 17 milijonov izrazov, izluščenih iz 159 virov v različnih jezikih (npr. kontroliranih geslovnikov in klasifikacijskih sistemov, kot so MeSH, SNOMED CT, ICD-10, DSM-IV in Gene Ontology);
2. Semantično omrežje (angl. *Semantic Network*) – omrežje trenutno sestavlja 127 semantičnih tipov (tj. pojmovnih kategorij) in 54 semantičnih relacij (tj. pomenskih razmerij), ki jih lahko vzpostavimo nad semantičnimi tipi;
3. Zakladnica besedišča (angl. *SPECIALIST Lexicon*) – slovar izrazov z različnimi besednimi oblikami, oblikoslovnimi lastnostmi in lemami.

Zdravstveni zapisi

Druga skupina virov združuje podatkovja, ki jih glede na stopnjo urejenosti poznamo pod oznako nestrukturirani podatki. Najpogosteje so to (elektronski) zdravstveni zapisi, povzetki kliničnih raziskav in laboratorijski izvidi. V tem okviru gre izpostaviti omrežje znanja, zgrajeno nad elektronskimi zdravstvenimi zapisi več kot 260 tisoč pacientov, ki omogoča pregledovanje omrežja po tipih izluščenih entitet (bolezen, zdravilo, postopek

obravnave in uporabljen pripomoček).[8] Na podobnem obsegu zapisov so omrežje znanja zgradili tudi Rotmensch in sodelavci,[9] ki so ugotovili visoko stopnjo točnosti avtomatskega luščenja trojčkov v primerjavi z domenskim ekspertom. Zhao in sodelavci[10] poročajo, da se omrežje znanja, zgrajeno na osnovi kliničnih zapisov, ponaša s topološkimi lastnostmi, ki so sicer značilna za kompleksna omrežja (npr. majhen premer omrežja, brezlestvičnost in visoka stopnja gručenja).

Obstoječe zbirke podatkov

V to skupino sodijo bibliografski viri in druge sorodne zbirke podatkov. Med bibliografskimi zbirkami prvenstvo zaseda zbirka MEDLINE/PubMed, ki trenutno obsega več kot 35 milijonov bibliografskih zapisov, s povprečnim dnevnim prirastom okrog 3000 zapisov. Zbirka je prosto dostopna in enostavno strojno berljiva, kar je bržkone glavni razlog, da jo za preizkušanje novih metod besedilnega rudarjenja uporabljajo številni raziskovalci. Prvi resen poskus preslikave celotne zbirke MEDLINE/PubMed v strukturirano obliko predstavlja izgradnja omrežja SemKG. Avtorji so uporabili vrsto prosto dostopnih orodij, s katerimi so iz množice slabih 23 milijonov povzetkov izluščili približno milijon entitet in več kot 14 milijonov relacij. Tako entitete kot posamezne relacije so opremili s semantičnimi tipi oz. tipi semantične relacije. Obseg omrežja so kasneje pomembno razširili, ko so v omrežje dodali tudi avtorje sestavkov z ustrezno razdvoumljenimi imeni, imena institucij, na katerih so zaposleni, ter raziskovalne projekte, pri katerih sodelujejo.[11]

Pandemija covida-19 je pustila sled tudi v objavi množice orodij in zbirk za rudarjenje besedil. Od začetka pandemije je raziskovalna skupnost gradila korpuse znanstvenih sestavkov s tematiko novega koronavirusa, med katerimi gre izpostaviti CORD-19[12] in LitCovid.[13] Oba korpusa sta temeljna kamna, na osnovi katerih je bilo zgrajeno – in nedavno objavljeno – omrežje CovidPubGraph, ki ponuja trenutno najbolj celovit pregled védenja o virusu SARS-CoV-2.[14] Trenutna različica omrežja je sestavljena iz več kot 268 milijonov trojčkov.

Gradnja omrežja znanja

Gradnja omrežja znanja je povezana s številnimi metodološkimi izzivi in zahteva interdisciplinarna znanja. V nadaljevanju izpostavimo tri korake, s katerimi se srečamo v postopku konstrukcije slehernega omrežja: (i) luščenje entitet in relacij; (ii) normalizacijo terminov ter (iii) integracijo in zlivanje znanja.

Luščenje entitet in relacij

Osnovno orodje za besedilno rudarjenje biomedicinskih besedil je MetaMap, ki deluje kot označevalnik biomedicinskih izrazov in omogoča, da prosto besedilo (npr. naslov in/ali povzetek zapisa MEDLINE/PubMed) preslikamo v ustrezno pojmovno oznako (t. i. biomedicinski koncept) iz metatezavra UMLS.[15] V tem koraku še ničesar ne vemo o pomenskem razmerju med dvema izluščenima konceptoma. Slednjemu je namenjeno orodje SemRep za procesiranje naravnega jezika, ki na osnovi leksikalnih pravil in zgoraj omenjenega semantičnega omrežja iz sistema UMLS identificira tudi pomensko razmerje med obema konceptoma.[16] Alternativno orodje je PKDE4J, ki je prav tako namenjeno luščenju biomedicinskih entitet in relacij, a zahteva ročno dodajanje terminoloških slovarjev.[17] Pred kratkim so bila raziskovalni skupnosti ponujena tudi orodja za luščenje entitet, ki temeljijo na modelu globokih nevronskih mrež, npr. HunFlair[18] in BERN2.[19] Ti orodji se v primerjavi s prej omenjenima SemRep in PKDE4J ponašata s pomembno višjima merama natančnosti in priklica. Na voljo je tudi že nekaj aplikacij globokega učenja za luščenje relacij, a so prilagojene le za kitajščino.[20] Nemalo težav na področju globokega učenja povzroča slaba interpretatibilnost rezultatov, saj algoritmi nevronskih mrež v dobršni meri delujejo po principu črne škatle. Smiselna razlaga rezultatov je zato – v primerjavi s sistemi, ki temeljijo na leksikalnih pravilih – pogosto nemogoča.

Normalizacija terminov

V avtomatiziranem postopku gradnje omrežja znanja je poleg gole prepoznave entitete pomemben korak tudi normalizacija, ki različne jezikovne različice, sinonime in izpeljanke poveže z eno entiteto. S problemom normalizacije se v biomedicinskih besedilih najpogosteje srečamo pri obravnavi imen in simbolov genov (npr. različne simbole in termine, kot so IL12, IL-12 in interleukin 12, je potrebno preslikati v pomensko entiteto Interleukin-12). Še pred nedavnim se je normalizacija v pretežni meri opravljala s pomočjo sistema UMLS in geslovnika MeSH. Danes lahko v ta namen uporabimo označevalnik PubTator.[21]

Integracija in zlivanje znanja

V postopku gradnje omrežja znanja razlikujemo med integracijo in zlivanjem podatkov. Pojem "integracija" se nanaša na povezovanje različnih podatkovnih zbirk, s pojmom "zlivanje" pa merimo na dopolnjevanje modalnosti podatkov. Dober primer integrativnega pristopa h gradnji omrežja znanja je

storitev PreMedKB.[22] Razpršenosti in heterogenost biomedicinskih zbirk pogostokrat botruje situaciji, ko "zaradi dreves ne vidimo gozda" in onemogoča celosten vpogled v mehanizme delovanja kompleksnih, multifaktorskih bolezni in načine njihovega zdravljenja. Z orodjem PreMedKB so avtorji pokazali, da lahko z razmeroma preprosto uporabo metapodatkovnih shem in sistema UMLS integriramo večje število sicer heterogenih zbirk podatkov.

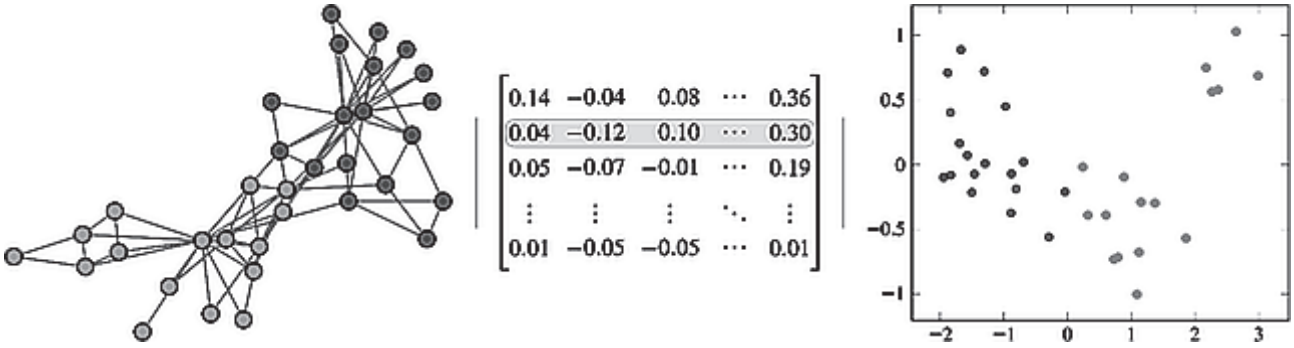
Uporaba omrežij znanja

Omrežja so kompleksne strukture, predvsem ki jih ni lahko razumeti in interpretirati. V skupnosti, ki se ukvarja z analizo omrežij, se je pred dobrima dvema desetletjema porojilo – v zadnjem desetletju pa močno intenziviralo – področje reprezentacijskega učenja, ki omogoča enostavno preslikavo relacijskega podatkovja v vektorsko obliko (govorimo o t. i. vložitvi omrežja), ki ohrani karseda veliko strukturnih lastnosti izvornega omrežja.

Na področju analize kompleksnih omrežij sta v vrsti različnih pristopov k reprezentacijskemu učenju, najpogosteje uporabljena algoritma DeepWalk in node2vec. Osnovna ideja vložitve posameznih vozlišč je ilustrirana na sliki 2. Algoritem DeepWalk vložitev vozlišč opravi na osnovi modela skip-gram s prirezanimi slučajnimi sprehodi,[24] node2vec pa z maksimizacijo pogojne verjetnosti nad sosedstvi vozlišč.[25] Dober vpogled v različne družine pristopov k reprezentacijskemu učenju nad homogenimi omrežji nudita pregledna članka.[5, 26]

Spomnimo se, da lahko s heterogenim omrežjem predstavimo različne tipe relacij med vozlišči.[27] Ustrezen pristop za reprezentacijsko učenje nad omrežjem znanja mora zato upoštevati tako tip entitete kot tip relacije. Paleta možnosti za obravnavo vložitev omrežja znanja je široka. Chang in sodelavci[28] so orali ledino in predlagali arhitekturo globokega učenja za obravnavo heterogenih interakcij v omrežju. Odmeven je bil tudi prispevek avtorjev algoritma metapath2vec, ki za okolico vozlišč najprej preišče z vnaprej definiranimi vzorci slučajnih sprehodov, nato pa pripravi vložitev s pomočjo modela skip-gram.[29] Pregled literature razkrije tri družine algoritmov za reprezentacijsko učenje nad omrežji znanja:[30] (i) modeli translacije v vektorskem prostoru; (ii) semantični modeli in (iii) modeli na osnovi globokih nevronskih mrež.

- Osnovna zamisel modelov, ki temeljijo na konceptu vzporednega premika v vektorskem prostoru je, da v trojčku $\langle h, r, t \rangle$ relacijo r obravnavamo kot translacijo iz izhodiščnega vozlišča h v končno vozlišče t, torej $h + r' \cong t$. Najenostavnejši algoritem TransE vektorsko vložitev vozlišč in povezav pripravi na osnovi modela nevronske mreže, v katerem minimiziramo kriterijsko funkcijo $f(h, r, t) = ||\mathrm{h} + \mathrm{r} - \mathrm{t}||$.[31] TransE odpove v primeru večrelacijskega omrežja (tj. v kardinalnostih ena-proti-mnogo in mnogo-proti-mnogo). To pomanjkljivost odpravlja model TransR, v katerem entitete in relacije vlagamo v ločena latentna prostora.[32]
- Osnova semantičnih modelov je koncept razdalje. Algoritem RESCAL je bil razvit na predpostavki, da sta si entiteti podobni, če se s podobnimi entitetami povezujeta s podobnimi relacijami. V to družino spadata še algoritma DistMult[33] in ComplEx.[34]
- Tretja družina algoritmov za reprezentacijsko učenje nad omrežji znanja temelji na pristopu globokih nevronskih mrež. V tem okviru omenjamo dva konvolucijska modela, ConvE[35] in ConvKB.[36] Njuna glavna slabost je, da pri pripravi vložitev posamezne trojčke obravnavata ločeno. Pomanjkljivost lahko odpravimo z vključitvijo mehanizma pozornosti, na katerem je osnovan model HRGAT.[37]



Slika 2 Vložitev vozlišč homogenega omrežja. Vozlišča omrežja (levo) predstavimo v vektorski obliki (sredina), pri čemer težimo k ohranitvi kar največ njihovih (strukturnih) lastnosti. Število vrstic matrike ustreza številu vozlišč omrežja. Osenčeno je predstavljena vektorska vložitev za izbrano vozlišče. Dolžino vektorjev v praksi izberemo na podlagi kompromisa med natančnostjo reprezentacije in sprejemljivo kompleksnostjo. Končno lahko nad matriko vložitev uporabimo nalogi primeren postopek statistične analize. Za grafični prikaz smo matriko vložitev dodatno skrčili z analizo glavnih komponent. Opazimo, da sta skupnosti vozlišč v omrežju (levo) lepo razvidni tudi v prostoru, ki ga razpenjata le prvi dve glavni komponenti (desno).

Zaključek

Predstavljeni pregled področja omrežij znanja nikakor ni izčrpen. Upamo pa, da ponuja dovolj celovit vpogled v to obširno tematiko. Različne načine predstavitve znanja smo naslovili le bežno; zgolj toliko, da smo poudarili dolgo preteklost raziskav na tem področju. Prav tako smo navedli le tiste vire podatkov, ki jih najpogosteje navaja znanstvena literatura. Pri tem smo namenoma – zaradi aktualnosti – poudarili gradnjo omrežij iz nestrukturiranih podatkov. Obravnava reprezentacijskega učenja bi zahtevala samostojen prispevek, zato vabimo bralce, da sežejo po dodatni literaturi.

Zahvala

Prispevek je nastal v okviru raziskovalnega projekta J5-2552, ki ga financira Agencija za raziskovalno dejavnost Republike Slovenije. Posebna zahvala – za potrpežljivost in vsebinske pripombe – gre glavnemu uredniku.

Reference

1. McBride DM, Cutting JC, Zimmerman C. *Cognitive psychology: theory, process, and methodology* (3rd ed.). Thousand Oaks 2023: Sage.
2. Russell SJ, Norvig P. Artificial intelligence: a modern approach (4th ed.). Hoboken 2020: Pearson.
3. Paulheim H. Knowledge graph refinement: a survey of approaches and evaluation methods. *Semant Web* 2017; 8(3): 489-508. https://doi.org/10.3233/SW-160218
4. Ehrlinger L, Wöß W. Towards a definition of knowledge graphs. In: Martin M, Cuquet M, Folmer E (eds.). *SEMANTiCS (posters, demos, SuCCESS) 2016*. Leipzig 2016: CEUR-WS.org; 4. https://ceur-ws.org/Vol-1695/paper4.pdf
5. Wang Q, Mao Z, Wang B, Guo L. Knowledge graph embedding: a survey of approaches and applications. *IEEE Trans Knowl Data Eng* 2017; 29(12): 2724-2743. https://doi.org/10.1109/TKDE.2017.2754499
6. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *J Biomed Inform* 2008; 41(5): 706-716. https://doi.org/10.1016/j.jbi.2008.03.004
7. Vintar Š. Označevanje in odkrivanje pomenskih razmerij v medicinskih besedilih. *Infor Med Slov* 2005; 10(1): 9-18.
8. Finlayson SG, LePendu P, Shah NH. Building the graph of medicine from millions of clinical narratives. *Sci Data* 2014; 1(1): 140032. https://doi.org/10.1038/sdata.2014.32
9. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. *Sci Rep* 2017; 7(1): 5994. https://doi.org/10.1038/s41598-017-05778-z
10. Zhao C, Jiang J, Xu Z, Guan Y. A study of EMR-based medical knowledge network and its applications. *Comput Methods Programs Biomed* 2017; 143: 13-23. https://doi.org/10.1016/j.cmpb.2017.02.016
11. Xu J, Kim S, Song M, et al. Building a PubMed knowledge graph. *Sci Data* 2020; 7(1): 205. https://doi.org/10.1038/s41597-020-0543-2
12. Wang LL, Lo K, Chandrasekhar Y et al. CORD-19: the COVID-19 open research dataset (v4). *arXiv* 2020: 2004.10706. https://doi.org/10.48550/arXiv.2004.10706
13. Chen Q, Allot A, Leaman R et al. LitCovid in 2022: an information resource for the COVID-19 literature. *Nucleic Acids Res* 2023; 51(D1): D1512-D1518. https://doi.org/10.1093/nar/gkac1005
14. Pestryakova S, Vollmers D, Sherif MA et al. COVIDPUBGRAPH: a FAIR knowledge graph of COVID-19 publications. *Sci Data* 2022; 9(1): 389. https://doi.org/10.1038/s41597-022-01298-2
15. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001: 17-21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243666/

16. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform* 2003; 36(6): 462-477. https://doi.org/10.1016/j.jbi.2003.11.003
17. Song M, Kim WC, Lee D, Heo GE, Kang KY. PKDE4J: entity and relation extraction for public knowledge discovery. *J Biomed Inform* 2015; 57: 320-332. https://doi.org/10.1016/j.jbi.2015.08.008
18. Weber L, Sänger M, Münchmeyer J, Habibi M, Leser U, Akbik A. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. *Bioinform* 2021; 37(17): 2792-2794. https://doi.org/10.1093/bioinformatics/btab042
19. Sung M, Jeong M, Choi Y, Kim D, Lee J, Kang J. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinform* 2022; 38(20): 4837-4839. https://doi.org/10.1093/bioinformatics/btac598
20. Yang Y, Lu Y, Yan W. A comprehensive review on knowledge graphs for complex diseases. *Brief Bioinformatics* 2023; 24(1): bbac543. https://doi.org/10.1093/bib/bbac543
21. Wei C-H, Allot A, Leaman R., Lu Z. PubTator Central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 2019; 47(W1): W587-W593. https://doi.org/10.1093/nar/gkz389
22. Yu Y, Wang Y, Xia Z et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res* 2019; 47(D1): D1090-D1101. https://doi.org/10.1093/nar/gky1042
23. Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. To embed or not: network embedding as a paradigm in computational biology. *Front Genet* 2019; 10: 381. https://doi.org/10.3389/fgene.2019.00381
24. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: online learning of social representations. In: Krishnapuram B et al. (eds.). *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD '14*. New York 2014: Association for Computing Machinery; 701-710. https://doi.org/10.1145/2623330.2623732
25. Grover A, Leskovec J. Node2vec: scalable feature learning for networks. In: Krishnapuram B et al. (eds.). *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York 2016: Association for Computing Machinery; 855-864. https://doi.org/10.1145/2939672.2939754
26. Yi, H-C, You Z-H, Huang D-S, Kwoh CK. Graph representation learning in bioinformatics: trends, methods and applications. *Brief Bioinformatics* 2022; 23(1): bbab340. https://doi.org/10.1093/bib/bbab340
27. Shi C, Li Y, Zhang J, Sun Y, Yu PS. A survey of heterogeneous information network analysis. *IEEE Trans Knowl Data Eng* 2017; 29(1): 17-37. https://doi.org/10.1109/TKDE.2016.2598561
28. Chang S, Han W, Tang J, Qi G-J, Aggarwal CC, Huang TS. Heterogeneous network embedding via deep architectures. In: Cao L (ed.). *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York 2015: Association for Computing Machinery; 119-128. https://doi.org/10.1145/2783258.2783296
29. Dong Y, Chawla NV, Swami A. Metapath2vec: scalable representation learning for heterogeneous networks. In: Matwin S, Yu S. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '17*. New York 2017: Association for Computing Machinery; 135-144. https://doi.org/10.1145/3097983.3098036
30. Ji S, Pan S, Cambria E, Marttinen P, Yu PS. A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans Neural Netw Learn Syst* 2022; 33(2): 494-514. https://doi.org/10.1109/TNNLS.2021.3070843
31. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In: Burges CJC et al. (eds.). *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2*. Austin 2013: AAAI Press; 2787-2795. https://dl.acm.org/doi/10.5555/2999792.2999923
32. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. *Proceedings of the AAAI Conference on Artificial Intelligence* 2015; 9(1): 2181-2187. https://doi.org/10.1609/aaai.v29i1.9491
33. Yang B, Yih W-T, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. *arXiv* 2015: 1412.6575. https://doi.org/10.48550/arXiv.1412.6575
34. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. In: Balcan MF, Weinberger KQ (eds.). *ICML'16: Proceedings of the 33rd International Conference on Machine Learning – Volume 48*. New York 2016: JMLR.org; 2071-2080. https://proceedings.mlr.press/v48/trouillon16.html
35. Dettmers T, Minervini P, Stenetorp P, Riedel S. Convolutional 2D knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 2018; 32(1): 1811-1818. https://doi.org/10.1609/aaai.v32i1.11573
36. Nguyen DQ, Nguyen TD, Nguyen DQ, Phung D. A novel embedding model for knowledge base completion based on convolutional neural network. In: Walker M, Ji H, Stent A (eds.). *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans 2018: Association for Computational Linguistics; 327-333. https://doi.org/10.18653/v1/N18-2053
37. Zheng S, Rao J, Song Y et al. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief Bioinformatics* 2021; 22(4): bbaa344. https://doi.org/10.1093/bib/bbaa344