■ *Pregledni znanstveni članek*

Matevž Kastrin, Dimitar Hristovski, Andrej Kastrin

Odkrivanje zakonitosti iz literature na področju znanosti o življenju

Povzetek. Odkrivanje zakonitosti iz literature (OZL) je razvijajoče se znanstveno področje, ki se ukvarja s samodejnim odkrivanjem novega znanja iz bibliografskih podatkov. V prispevku najprej naredimo kratek pregled problematike znanstvene ustvarjalnosti, ki je neločljivo povezana s področjem OZL. OZL naslonimo na Mednickovo teorijo ustvarjalnosti in Koestlerjev koncept bisociacij. V nadaljevanju predstavimo glavne teoretične modele OZL. Največji del prispevka posvetimo pregledu aplikacij s področja OZL, ki jih razdelimo glede na osnovni mehanizem odkrivanja novega znanja: aplikacije, ki temeljijo na načelu sopojavnosti, in aplikacije, ki temeljijo na semantičnih povezavah med biomedicinskimi koncepti.

Literature-Based Discovery in the Field of Life Sciences

Abstract. Literature-based discovery (LBD) is a growing scientific field that deals with automatic discovery of new knowledge from bibliographic data. We first make a brief overview of the problem of scientific creativity, which is inextricably associated with the field of LBD. We base LBD on Mednick's theory of creativity and Koestler's concept of bisociations. Next, we present the main theoretical models of LBD. The largest part of the paper is devoted to a comprehensive review of LBD applications. We categorise LBD applications according to the main mechanism that provides knowledge discovery: applications based on the co-occurrence principle, and applications based on semantic relations among biomedical concepts.

■ **Infor Med Slov** 2017; 22(1-2): 22-31

Institucije avtorjev / Authors' institutions: Klinični oddelek za travmatologijo, Kirurška klinika, Univerzitetni klinični center Ljubljana (MK); Inštitut za biostatistiko in medicinsko informatiko, Medicinska fakulteta, Univerza v Ljubljani (DH, AK).

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

Prispelo / Received: 18. 4. 2017. Sprejeto / Accepted: 18. 1. 2018.

Uvod

Pozornost raziskovalcev na področju znanosti o življenju se je pred dobrima dvema desetletjema tudi na eksperimentalni ravni preselila iz zanimanja za posamezne biološke koncepte (npr. gene in proteine) k poskusu razumevanja celih bioloških sistemov. Rezultat tega preskoka je med drugim vedno večja količina podatkov, ki jo je potrebno ne le analizirati, pač pa tudi smiselno interpretirati. To zahteva sposobnost integracije podatkovnih virov z obstoječim znanjem, ki je dostopno v literaturi.

Informacijski viri na področju biomedicine so zelo obsežni in v veliki večini primerov tudi prosto dostopni. MEDLINE, najobsežnejša bibliografska zbirka na področju biomedicine, obsega že več kot 25 milijonov bibliografskih zapisov, s prirastkom od 2.000 do 4.000 zapisov dnevno. Pred raziskovalci je zato zahtevna naloga. Integracija podatkovnih virov z eksperimentalnimi podatki zahteva dobro poznavanje in redno spremljanje različnih raziskovalnih smeri, kar je za človekov spoznavni aparat prehud zalogaj. Orodja, ki omogočajo (pol)samodejno rudarjenje po besedilih znanstvenih člankov so zato primarnega pomena za nadaljnji razvoj znanosti o življenju.

Orodja za rudarjenje besedilnih podatkov nam v grobem pomagajo pri štirih, zaporednih in medsebojno povezanih procesih.[1,2] Prvič, raziskovalcu omogočajo identifikacijo ustreznih virov znanja oz. priklic informacij (angl. *information retrieval*). Drugič, olajšajo identifikacijo (bioloških) entitet oz. konceptov (npr. genov in proteinov) v besedilu (angl. *entity recognition*). Tretjič, pomagajo pri esktrakciji predodločenih entitet in relacij (npr. eksplicitnih relacij med geni) iz besedila (angl. *information extraction*). Četrti sklop procesov se nanaša na sisteme za odkrivanje zakonitosti iz literature (OZL), ki temeljijo na prepoznavi posrednih relacij med posameznimi biološkimi koncepti (angl. *literature-based discovery*). Za razliko od prvih treh, je področje OZL najmanj razvito, polno raziskovalnih izzivov in zato glavni predmet tega pregleda.

Znanstvena ustvarjalnost

Ustvarjalnost je pomembna intelektualna aktivnost, pri kateri tvorimo nove ideje in povezave med njimi, ter obenem temeljno gonilo človeškega razvoja. Stari Grki so sicer zasejali seme dvoma, da je ustvarjalnost povezana z norostjo, vendar take poprepreproščene definicije danes nihče več ne jemlje resno. Brez ustvarjalnosti bi človek ostal lačen napredka in bi pri reševanju problemskih nalog ponavljal vedno enake vzorce vedenja.

Znanost običajno opredelimo v smislu sistematičnega preučevanja naravnih in družbenih pojavov. Hipotetično-deduktivni pristop, ki se dandanes večinoma uporablja v znanosti, je sestavljen iz več medsebojno povezanih korakov:[3] (i) opredelitev problema, (ii) identifikacija raziskovalnega vprašanja, (iii) formulacija raziskovalne domneve, (iv) oblikovanje pričakovanega rezultata, (v) izvedba poskusa, (vi) primerjava pričakovanega in dejanskega rezultata ter (vii) oblikovanje sklepa. Odkritje je pogoj *sine qua non* v znanosti in predstavlja vrh človekove ustvarjalne misli. Odkritja omogočajo raziskovalcem spoznavanje sveta okoli sebe, osmišljajo realnost in prestavljajo meje mogočega. Odkritje se lahko nanaša na rešitev znanega problema ali pa se nanaša na problem, ki se šele kasneje izkaže za pomembnega. Pomembna odkritja se običajno dogajajo na mejah znanega, v prostoru brez domenskih ekspertov in ustreznega raziskovalnega instrumentarija. Možgani raziskovalca v tem prostoru delujejo po principu selektivnih poskusov in napak ter na osnovi hevrističnih pravil, ki so osnovana na preteklih izkušnjah.[4] Generiranje novih idej je ključni element ustvarjalnega procesa znanstvenega spoznavanja realnosti.[5,6]

Sodobno znanstveno raziskovanje je ena od najbolj kompleksnih človeških aktivnosti, ki zahteva različne tipe sposobnosti. Znanstveno ustvarjalnost tako lahko vidimo kot skupek kognitivnih in računskih konceptov, med katere uvrščamo (i) motivacijo za raziskovalno delo, (ii) sposobnost pravilne formulacije raziskovalnega problema znotraj obstoječe zakladnice znanja, (iii) sposobnost konstrukcije izčrpnega iskalnega prostora za rešitev raziskovalnega problema, (iv) sposobnost implementacije hevristik za redukcijo prostora iskanja ter (v) potrpljenje in vzdržljivost.[7] Kljub neustavljivi želji, da bi bili sposobni razumeti dinamiko znanstvenih odkritij ali bili celo sposobni njihovih napovedi, je naše znanje o tem še zelo skopo.[8] Nekatera odkritja so zaradi vidne akumulacije teoretičnih spoznanj in empiričnih dokazov kot na dlani. Dober primer je npr. nedavna potrditev obstoja gravitacijskih valov. Po drugi strani se zdijo nekatera odkritja popolnoma nepredvidljiva; tak primer je npr. odkritje penicilina.

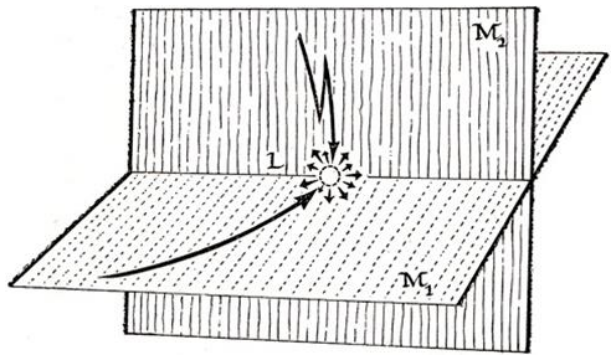
Element novosti je glavni sestavni del znanstvene ustvarjalne misli, vendar je za nas ključnega pomena spoznanje, da so sestavni bloki novih idej pogosto že vključeni v obstoječem znanju. Uzzi[9] namreč na osnovi obširne bibliografske analize dobrih 18 milijonov znanstvenih člankov ugotavlja, da sodobna znanost temelji na novih, a močno konvencionalnih kombinacijah obstoječega znanja. Posledično

ugotavlja, da bo znanstveni članek, ki je zmes novega in konvencionalnega, imel zelo visoko verjetnost, da bo visoko citiran. Znanstveno ustvarjalnost si zato lahko predstavljamo kot kontinuum, kjer na enem koncu domuje novost, na drugem pa konvencionalnost.

Teoretičnih modelov, ki pojasnjujejo človekovo ustvarjalnost, je ogromno, njihov pregled pa bi presegal okvire tega prispevka. Za naše potrebe je dovolj, da iz zgodovinskega spomina obudimo Mednickovo[10] teorijo ustvarjalnosti in Koestlerjev[11] model bisociacij. Mednick je zasnoval teoretičen model ustvarjalnosti, v katerem je izpostavil pomen asociacij pri porajanju novih idej. Proces ustvarjalnega mišljenja definira kot povezovanje asociacijskih elementov v nove kombinacije, ki so na svojstven način uporabne. Bolj kot sta si elementa v asociacijskem prostoru oddaljena, bolj ustvarjalen je proces oziroma njegova rešitev. Do novega odkritja po Mednicku lahko pridemo po slučaju ali pa zaradi podobnosti asociacijskih elementov.

Koestlerjev osnovni koncept bisociacij je predstavljen na sliki 1. Diagram predstavlja dve domeni znanja M_1 in M_2, kot ravnini, ki sta pravokotni druga na drugo. Vsaka od ravnin se nanaša na svoj asociativni kontekst. Koestler definira ustvarjalnost kot bisociacijo, povezavo med elementi iz dveh različnih asociativnih kontekstov. Z drugimi besedami je bisociacija nov kos znanja, ki povezuje dve domeni znanja na doslej še nepovezan način. Formalno, med dvema konceptoma obstaja bisociacija, če: (i) nista koncepta evidentno neposredno povezana, (ii) vsak koncept prihaja iz svoje domene in (iii) nova povezava odkriva nov, še nepoznan, pogled na problemsko domeno. Iz zgodovine je najbolj znan primer Arhimeda, ki je poskušal ugotoviti, ali je zlatu v kroni primešano srebro, ne da bi krono kakorkoli poškodoval. Prvi asociacijski kontekst je predstavljal problem, kako izmeriti prostornino nepravilno oblikovanega predmeta. Drugi asociacijski kontekst je bila Arhimedova ugotovitev, da je prostornina vode, ki se prelije iz kadi, enaka prostornini kopalca. S tem spoznanjem Arhimedu ni bilo težko rešiti naloge s krono, tako da je vzpostavil bisociacijo med obema asociacijskima kontekstoma in tako razkril zlatarjevo goljufijo. Do podobnega »heureka« momenta je prišlo ob iznajdbi metode verižne reakcije s polimerazo, katere odkritju je botrovala premetena kombinacija dotedanjih dobro poznanih tehnik v biokemiji.

Dolgo časa je veljalo, da je racionalen model odkrivanja znanja nemogoč. Raziskovalci na področju umetne inteligentnosti so si prizadevali proces odkrivanja znanja natančno opisati s pomočjo algoritmov in ga čim bolj avtomatizirati. Na nekaterih področjih jim je to celo uspelo. V okviru Microsoftovega projekta Adam so npr. raziskovalci razvili avtomatiziran sistem za prepoznavo slik, ki temelji na metodi globokega učenja (angl. *deep learning*).[12] Pasmo psa je Adam sposoben prepoznati na podlagi enega samega posnetka. Na področju bibliometrije so raziskovalci s pomočjo računalnika poskušali simulirati pomembna zgodovinska odkritja, drugi pa so se usmerili v računsko odkrivanje novega znanja.[13] Čeprav že sam termin »računsko odkrivanje znanja« implicira avtomatiziran proces, bolj podroben pregled literature razkriva, da pri praktično vsakem uspešnem poskusu človek igra pomembno vlogo. Proces računskega odkrivanja znanja lahko podobno kot pri človeku opišemo z naslednjimi koraki: (i) formulacija problema, (ii) reprezentacija problema, (iii) manipulacija podatkov, (iv) manupulacija algoritmov ter (v) filtriranje rezultatov in njihova interpretacija.



Slika 1 Ravnini predstavljata nepovezana asociacijska konteksta (domeni znanja) M_1 in M_2. Koestler definira ustvarjalnost kot bisociacijo, povezavo med dvema elementoma iz različnih asociativnih kontekstov.

Znanstvena skupnost je danes kompleksen ekosistem, s stotinami bolj ali manj prepletenih raziskovalnih področij, desettisoči raziskovalcev in vrtoglavim številom znanstvenih publikacij. Znanstveni napredek bi bil seveda idealen, če bi imel raziskovalec pregled nad vso relevantno literaturo s svojega znanstvenega področja. Wyatt[14] je že v devetdesetih letih prejšnjega stoletja ugotovil, da se število biomedicinskih revij podvoji približno na vsakih 20 let. Vpogled v PubMed npr. razkriva, da je danes o diabetesu napisanih dobrih 575.000 znanstvenih člankov. Predstavljajte si diabetologa, ki bi vsak dan prebral 20 člankov. Da bi se prebil čez celo to skladovnico člankov, bi potreboval kar 80 let! Danes velja, da je praktično nemogoče biti odličen ekspert na več področjih hkrati. Znanosti o življenju čedalje bolj stremijo k podatkovno intenzivnemu raziskovanju. Kljub temu se še vedno zanašajo na tradicionalne metode

raziskovanja, ki med drugim slonijo na ročnih interpretacijah domenskega eksperta, v upanju, da bo našel uporabne vzorce v podatkih. Avtomatizacija znanstvenega procesa je zato ne samo teoretično, pač pa tudi povsem praktično vprašanje.[15] Glavna predpostavka avtomatskega odkrivanja zakonitosti iz literature je uporabniku omogočiti samodejno iskanje novega, potencialno uporabnega in razumljivega znanja.

Swanson in odkrivanje zakonitosti iz literature

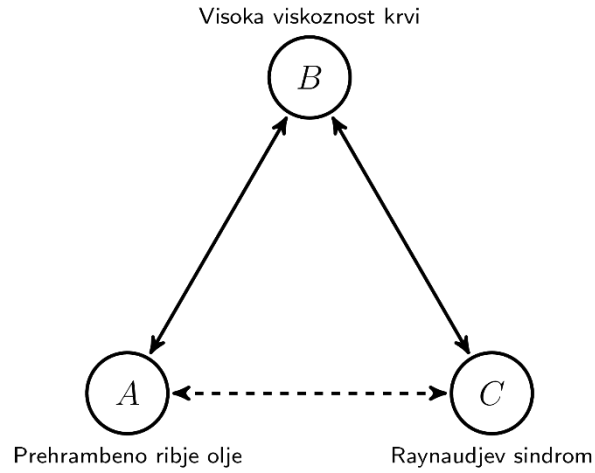
Zelo površno rečeno je znanstvena literatura na področju znanosti o življenju nestrukturiran skupek konceptov, kot so npr. ključne besede, imena genov, zdravil, terapevtskih postopkov in relacij med njimi. Akumulacija znanja temelji na njihovi učinkoviti prepoznavi, ekstrakciji in transformaciji v dobro strukturirano in semantično stabilno obliko.

V sredini 80. let prejšnjega stoletja je Swanson[16] sprožil novo vejo interdisciplinarnega raziskovanja. Pri OZL je šlo za slučajno odkritje, ki je prineslo nov veter v jadra tehnološko osveščenih raziskovalcev na področju znanosti o življenju, po drugi strani pa priložnost za razvoj novih računalniških aplikacij. Podrobnejši opis Swansonovega odkritja sledi v naslednjem razdelku. Formalno OZL opredelimo kot avtomatiziran proces iskanja komplementarnih (vsebinsko dopolnjujočih) struktur med disjunktnimi množicami znanstvene literature.[17]

Teoretično ozadje Swansonovega pristopa k OZL lahko navežemo na Mednickovo[10] teorijo ustvarjalnosti, s posebnim poudarkom na medkontekstnih asociacijah, ki jim Koestler[11] pravi bisociacije. Če bi iz množice člankov po slučaju potegnili dva, bi z veliko verjetnostjo ne bila komplementarna. Za ta namen zato potrebujemo avtomatiziran iskalni proces, ki bo karseda dobro izkoristil človekovo znanje in sposobnost presoje ter ju kombiniral s hitrostjo računalnika in njegovo učinkovitostjo pri obdelavi velikih količin podatkov. Koncept OZL je osnovan na treh medsebojno povezanih predpostavkah:[16] (i) količina znanstvene literature je prevelika, da bi jo en sam raziskovalec lahko pregledal, (ii) raziskovalci so običajno specializirani za eno področje, brez dobrega vpogleda v druga raziskovalna področja in (iii) znanost je fragmentirana, sestavlja jo večje število bolj ali manj tesno povezanih skupnosti.

OZL je bilo doslej, z redkimi izjemami,[18] uporabljeno zgolj na področju znanosti o življenju; Swanson[16] pravi, da bržkone zaradi goste povezanosti konceptov v njem, sami pa pristavljamo, da je glavni vzrok temu predvsem prosta dostopnost in enostavnost strojne manipulacije bibliografske zbirke MEDLINE. Swanson[16] pojmuje OZL primarno kot človeško aktivnost, ki pa nujno potrebuje računalniško podporo. OZL zato vidi predvsem kot orodje za spodbujanje raziskovalčeve ustvarjalnosti pri tvorjenju plavzibilnih in preverljivih domnev.

Najbolj znan in v aplikacijah najpogosteje uporabljen je Swansonov[17] *ABC* model odkrivanja znanja. Model je zasnovan na preprosti ideji, da lahko novo znanje odkrijemo med neposredno nepovezanima konceptoma (*A*) in (*C*) preko vmesnega koncepta (*B*). Shematsko je model prikazan na sliki 2.

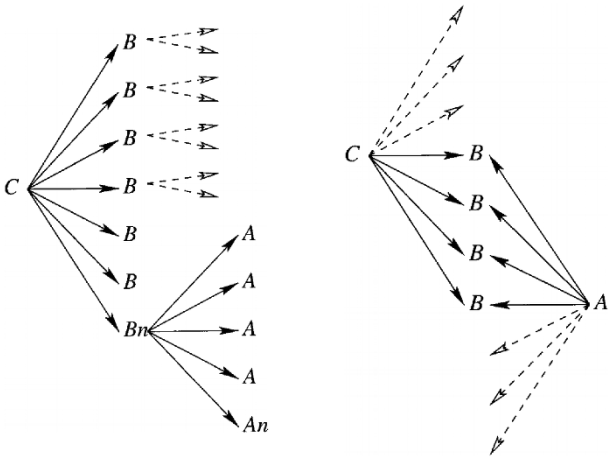


Slika 2 Swansonov *ABC* model odkrivanja znanja. Novo znanje predstavlja povezava med neposredno nepovezanima konceptoma (*A*) in (*C*) preko vmesnega koncepta (*B*).

Model *ABC* se uporablja v dveh izpeljankah:[19] odprti in zaprti. Pri odprtem iskanju je glavni namen generiranje raziskovalne domneve. Običajno imamo na voljo raziskovalni problem. Proces iskanja začnemo z enim konceptom (*A*), kateremu priredimo množico konceptov (*B*). Postopek iskanja ponovimo za vsakega ob (*B*) konceptov. Dobljeni zadetki predstavljajo koncepte (*C*). Te filtriramo tako, da izločimo vse neposredne povezave s konceptom (*A*). Postopek odprtega iskanja je shematsko prikazan na sliki 3. Pri drugi obliki, zaprtem iskanju, raziskovalno domnevo le preverjamo. Začetni (*A*) in ciljni (*C*) koncept poznamo, iščemo pa vmesne koncepte (*B*). Tudi model zaprtega iskanja je shematsko prikazan na sliki 3.

Čeprav je model *ABC* največkrat uporabljen, pa ni edini. Wilkowski[20] je predlagal razširitev modela *ABC* tako, da je enovozliščne (*B*) koncepte nadomestil z večvozliščnimi vmesnimi koncepti. Model ima

oznako *AnC*, kjer je $n = (B_1, B_2, ..., B_m)$. Van der Eijk[21] je za OZL prvi predlagal uporabo modela omrežja ob predpostavki, da koncept sopojavnosti (angl. *co-occurrence*) zagotavlja celosten pogled na kompleksnost povezav med koncepti.



Slika 3 Odprt (levo) in zaprt (desno) proces odkrivanja novega znanja. Polne povezave vodijo do potencialno zanimivih konceptov, črtkane povezave pa do konceptov, ki so se ob naknadnem preverjanju izkazali za nezanimive.

Pregled sistemov za računalniško odkrivanje zakonitosti iz literature

Pristopi, osnovani na načelu sopojavnosti

Swanson[22] je do svojega prvega odkritja prišel po naključju ob študiju literature o inuitski prehrani. Inuiti so staroselsko ljudstvo, ki prebiva na Grenlandiji, severnih obalah Kanade, Aljaski ter skrajnovzhodnem koncu Sibirije. Ob branju več kot 4.000 naslovov člankov v zbirki MEDLINE je prišel do spoznanja, da prehrambeno ribje olje znižuje viskoznost krvi, zmanjšuje strjevanje krvnih ploščic in inhibira odziv žilne stene. Hkrati je ugotovil, da zmanjšana viskoznost krvi in strjevanja krvnih ploščic ob hkratni inhibiciji odziva žilne stene preprečuje nastanek Raynaudovega sindroma. Pri Raynaudovem sindromu gre za občasna skrčenja manjših žilnih odvodnic, najpogosteje v prstih rok, lahko pa tudi na prstih nog, jeziku in nosu. Motnja v prekrvavitvi traja navadno nekaj minut do nekaj ur. Swanson je tako postavil domnevo, da prehrambeno ribje olje preprečuje nastanek Raynaudovega sindroma. Kasnejši klinični eksperiment je povezavo dejansko potrdil.[23] Kasneje je sam oz. s sodelavci opisal še več primerov odkritij. Večino od njih si ogledamo v nadaljevanju prispevka.

Pri iskanju citatov si je Swanson[22] pomagal z iskalnikom Dialog SciSearch, s katerim je preiskal naslove in povzetke bibliografskih zbirk MEDLINE in Embase (Excepta Medica). Okoli 1.000 bibliografskih zapisov se je nanašalo na Raynaudov sindrom, 3.000 zapisov pa na prehrambeno ribje olje. Množico člankov je prečistil in izdvojil 489 citatov, od katerih so le štirje povezovali obe iskalni domeni. Še več, le dva citata sta govorila o Raynaudovem sindromu in ribjem olju, vendar ne v kontekstu Swansonovega odkritja. Swanson je zato skoval termin »neodkrito javno znanje« (angl. *undiscovered public knowledge*), s katerim je opisal fenomen nepovezanih, a logično komplementarnih delov znanja, ki se skrivajo v literaturi.

Dve leti kasneje je Swanson[24] opisal povezavo med migreno in magnezijem, temu pa je dodal še povezavo med somatomedinom C in argininom.[25] S Smalheiserjem[26] je opisal povezave med pomanjkanjem magnezija in različnimi možganskimi funkcijami, med indometacinom in Alzheimerjevo boleznijo,[27] estrogenom in Alzheimerjevo boleznijo[28] ter med od kalcija neodvisno fospolipazo A2 in shizofrenijo.[29] Swanson in Smalheiser[30] sta bila prva, ki sta strokovni javnosti ponudila orodje Arrowsmith, prostodostopno aplikacijo za interaktivno OZL.

Gordon in Lindsay[31] sta poskušala sistematično ponoviti Swansonovo domnevo "Raynaudov sindrom ↔ prehrambeno ribje olje" s pomočno avtomatiziranega postopka odprtega odkrivanja. Pomagala sta si z metodami leksikalne statistike, tako da sta uporabila različne mere, kot so frekvenca posameznih besed, frekvenca dokumentov, v katerih nastopa posamezna beseda, in mera TF×IDF.[32] Za razliko od Swansona sta uporabila celoten zapis, vključno s povzetkom, in ne samo naslova zapisa v MEDLINE. Potrdila sta, da je avtomatizirano odkrivanje znanja na primeru Swansonove domneve možno. Za razliko od sistema Arrowsmith daje njun pristop večjo težo posameznim besedam in ne le dokumentom. Kasneje sta poskušala enako metodologijo uporabiti tudi za potrditev domneve "migrena ↔ magnezij", vendar jima je spodletelo.[33] Uspešna sta bila šele z novim pristopom, kjer sta uporabila model *n*-gramov.

Gordon in Dumais[34] izhajata iz predpostavke, da so v procesu OZL dobri kandidati tisti koncepti, ki so karseda podobni (semantično in statistično) z izhodiščnim konceptom. Kot metodo za odkrivanje podobnih konceptov predlagata latentno semantično indeksiranje (angl. *latent semantic indexing*, LSI). LSI definira vektorski prostor, v katerem semantično sorodni termini ležijo bolj skupaj. Proces iskanja

začnemo tako, da izhodiščnemu konceptu (*A*) s pomočjo LSI poiščemo karseda podobne koncepte (*B*). Urejeni seznam konceptov (*B*) ovrednoti domenski ekspert in izbere ožjo množico konceptov. Proces iskanja nato ponovimo še za iskanje konceptov (*C*). Z izjemo metodologije avtorja ne ponujata ničesar novega. Odkrivanje novega znanja na problemski nalogi odprtega tipa ovrednotita tako, da svoje rezultate primerjata z eksperimentom, ki sta ga izvedla Gordon in Lindsay,[31] in potrdita komplementarnost obeh pristopov.

Weeber, Klein, denBerg in Vos[19] so kot pomoč uporabniku pri odkrivanju znanja razvili dvostopenjski sistem DAD (angl. *Disease – Adverse drug reaction – Drug*), ki omogoča odprti in zaprti način odkrivanja znanja. DAD po mnenju njegovih avtorjev pomaga raziskovalcu na treh področjih: (i) močno zmanjša prostor iskanja novega znanja, (ii) pomaga pri interpretaciji rezultatov iskanja s pomočjo semantične analize in (iii) omogoča vpogled v tekstovni kontekst dane domneve. Eksperimentalno okolje predstavlja celotna zbirka MEDLINE zapisov. DAD uporablja sistem MetaMap,[35] ki prosto besedilo naslova in povzetka bibliografskega zapisa preslika v biomedicinske koncepte iz metatezavra UMLS. Glavna prednost uporabe konceptov je, da se različne jezikovne različice, sinonimi in izpeljanke preslikajo na eno entiteto; npr. simboli in termini IL12, IL-12 in interleukin 12 se vsi preslikajo v koncept Interleukin-12. DAD prav tako omogoča filtriranje konceptov po semantičnih tipih, kot jih najdemo v orodju UMLS Semantic Network. Mero povezanosti med koncepti (*AB*) in (*BC*) predstavlja število sopojavnosti danih konceptov v posameznih stavkih besedila. Sistem so ovrednotili s pomočjo Swansonovih domnev "Raynaudov sindrom ↔ ribje olje" in "migrena ↔ magnezij", vendar niso predstavili statističnih mer ustreznosti sistema. Glavna slabost sistema DAD je problem razreševanja dvoumnih konceptov; MetaMap npr. zmotno preslika termina »mg« (miligram) in »Mg« (magnezij) v koncept »magnezij«.

Stegmann in Grohmann[36] sta bazo znanja zgradila na podlagi sopojavnosti MeSH terminov. MeSH je kontroliran geslovnik, ki vsebuje biomedicinske izraze (deskriptorje) na različnih nivojih specifičnosti. Za odkrivanje posrednih povezav med koncepti sta uporabila metodologijo strateškega diagrama, ki jo je prvi predlagal Callon.[37] Dvojice MeSH terminov najprej po podobnosti razvrstita v skupine, nato pa za vsako skupino izračunata njeno gostoto (moč povezav med MeSH termini znotraj ene skupine) in središčnost (moč povezav med MeSH termini med različnimi skupinami). Na primerih domnev "Raynaudov sindrom ↔ ribje olje" in "migrena ↔ magnezij" sta odkrila, da vmesni koncepti (*B*) nastopajo v območju pod mediano gostote in središčnosti. Na novo sta tako odkrila povezavo med prioni in manganom, pri kateri kot vmesni koncept nastopijo fuzijski proteini.

Pratt in Yetisgen-Yildiz[38] sta predstavila sistem LitLinker, ki za odkrivanje novega znanja uporablja kombinacijo tehnik procesiranja naravnega jezika in algoritmov rudarjenja podatkov. Njun cilj je bil ponuditi sistem za odprto odkrivanje znanja, ki bo čim bolj avtomatiziran in neodvisen od uporabnika. Za ekstrakcijo znanja sistem uporablja le naslove citatov, katerih besede s pomočjo sistema MetaMap preslika v UMLS koncepte. Sistem množico konceptov nato prečisti, tako da odstrani (i) neinformative koncepte, (ii) pomensko preveč sorodne koncepte ter (iii) koncepte, ki z vidika preučevanjega problema niso zanimivi. Slednje naredi s pomočjo orodja UMLS Semantic Network, tako da za nadaljnjo analizo uporabi le tiste biomedicinske koncepte, ki ustrezajo podanemu semantičnemu tipu. Za razliko od prejšnjih sistemov, LitLinker za identifikacijo povezav med koncepti uporablja asociacijska pravila, natančneje algoritem Apriori. Ciljne koncepte grupira glede na število vmesnih konceptov. Dobljena skupina konceptov prevzame ime koncepta, katerega ime je najkrajše. Namen tega koraka je grupirati podobne koncepte tako, da lažje poiščemo ciljne koncepte. Za ovrednotenje sistema sta avtorja uporabila Swansonovo domnevo "migrena ↔ magnezij", vendar brez statističnega ovrednotenja. Glavna prednost sistema je iskanje vmesnih (*B*) konceptov, kjer uporabnikova pomoč ni potrebna.

Srinivasan[39] namesto surovih UMLS konceptov za odkrivanje znanja uporabi MeSH termine. Glavni namen njenega pristopa je čim bolj avtomatizirati proces odkrivanja novega znanja, brez pomoči uporabnika. Za izbran izhodiščni koncept (*A*) algoritem sestavi MeSH profil za izbrane semantične tipe. Profil predstavlja relativno pomembnost različnih MeSH terminov za množico izbranih semantičnih tipov. Za vsak semantični tip nato izberemo *n* najbolj zanimivih MeSH terminov; ti termini predstavljajo koncepte (*B*). Termine (*C*) dobimo tako, da enak postopek neodvisno ponovimo nad vsakim od terminov (*B*). Avtorica je uspešno ponovila iskanje nad dvema odprtima in petimi zaprtimi problemi. Glavna očitka temu pristopu sta: (i) nekateri, zlasti pa novi, bibliografski zapisi (še) nimajo pripisanih MeSH terminov in (ii) MeSH termini so pogosto bolj splošni od prostega besedila v naslovih in povzetkih bibliografskih zapisov ter od

UMLS konceptov, kar lahko močno ovira odkrivanje znanja.

Van der Eijk in sodelavci[21] so uporabili preslikavo iz omrežja sopojavnosti v asociativni prostor konceptov (angl. *associative concept space*). Gre za večrazsežni evklidski prostor, v katerem so koncepti s pomočjo Hebbovega učnega algoritma razvrščeni tako, da so pri visoki sopojavnosti koncepti narisani skupaj in obratno, koncepti, ki redko nastopajo skupaj, pa so narisani daleč vsaksebi. Kombinacija omrežja sopojavnosti in prostorske predstavitve omogoča hitro iskanje poti v omrežju, saj lahko dolžino poti med dvema vozliščema ocenimo na podlagi evklidske razdalje v prostoru. Daljše poti (več kot en skok med konceptoma) kažejo na posredne relacije med koncepti in predstavljajo kandidatke za nove raziskovalne domneve. Sestavni del pristopa je vizualizacija rezultatov, kar predstavlja pomembno prednost pred ostalimi predhodno razvitimi pristopi. Avtorji so uspešnost algoritma pri odkrivanju novega znanja preverili na simuliranih in realnih podatkih iz bibliografske zbirke MEDLINE. Podobno kot pri ostalih, tudi tukaj pogrešamo poglobljeno statistično evalvacijo sistema.

Wren in sodelavci[40] izhajajo iz predpostavke, da je število konceptov (C) v klasičnem odprtem modelu odkrivanja znanja običajno zelo veliko. Relacije $(A) \leftrightarrow (C)$ so zato rangirali s pomočjo modela slučajnega omrežja, tako da so izračunali razmerje med številom dejanskih povezav in pričakovanim številom povezav. Relacije, katerih količnik je presegel pražno vrednost, so predstavljale nove raziskovalne domneve, potencialno novo znanje. Omrežje povezav so zgradili tako, da so izdvojili koncepte, ki se nanašajo na gene, bolezni, fenotipe in kemične učinkovine. Z uporabo opisanega pristopa so avtorji predpostavili domnevo o obstoju vzročne povezave med srčno hipertrofijo in klorpromazinom. Klinični eksperiment na miših je domnevo potrdil. Statistične evalvacije sistema avtorji ne podajo.

Hristovski s sodelavci[41] namesto besed iz naslovov in povzetkov bibliografskih zapisov za odkrivanje novih relacij uporabi MeSH deskriptorje. Za opisovanje znanih in napovedovanje potencialnih novih relacij avtorji uporabijo asociacijska pravila. Kot prvi v vrsti sistemov za OZL predstavijo tudi izčrpno statistično evalvacijo nad 10 izbranimi boleznimi. Kasneje Hristovski s sodelavci[42] predstavi spletno aplikacijo BITOLA za interaktivno odkrivanje zakonitosti iz literature. Sistem je posebej prilagojen odkrivanju kandidatnih genov za bolezni, vendar se ga lahko uporablja tudi kot splošni generator novih raziskovalnih domnev na področju biomedicine. Baza znanja je zgrajena na osnovi naslovov in povzetkov člankov, pripadajočih MeSH terminov ter imen genov. Število relacij $(A) \leftrightarrow (C)$ je omejeno s filtriranjem po semantičnih tipih in kromosomski lokaciji gena. Statistične evalvacije sistema avtorji ne podajo, predstavijo pa praktični primer odkrivanja genov, kandidatov za polimikrogirijo.

Sistem Telemakus bazo znanja zgradi na osnovi ekstrakcije UMLS konceptov iz biomedicinske literature, tako da uporabi podatke iz tabel in slik polnih dokumentov.[43] Prednost sistema je velika natančnost, saj je del konceptov ekstrahiran ročno. V tem pa se skriva tudi slabost pristopa, saj je ročna ekstrakcija počasna in draga. Poleg tega je sistem omejen le na dve domeni: na omejevanje kalorij in biologijo staranja, vendar je po zagotovilih avtorjev razširljiv tudi na druge domene.

Na koncu tega razdelka ne smemo pozabiti na metodo RaJoLink, ki je plod domačega znanja.[44] Metoda implementira Swansonov model ABC tako, da samodejno predlaga kandidatne koncepte (A), ki so logično povezani s proučevanim fenomenom (C). Kandidatne koncepte izbira po načelu redkosti pojavljanja. Avtorji žal ne podajo statistične evalvacije sistema, prav tako ni dostopna računalniška aplikacija.

Pristopi osnovani na semantičnih povezavah

Doslej so vsi predstavljeni sistemi kot glavni mehanizem odkrivanja novih relacij uporabljali preprosto mero sopojavnosti. Sopojavnost ima vrsto pomanjkljivosti, med drugim (i) generira veliko število napačno pozitivnih relacij $(A) \leftrightarrow (C)$ ter hkrati (ii) ničesar ne pove o vsebini relacije. Raziskovalce namreč poleg golega obstoja interakcije med koncepti zanimajo tudi mehanizmi, ki ležijo v ozadju ter vzročno-posledični odnosi med koncepti.[45]

Hu s sodelavci[46] je bil prvi, ki je pri rudarjenju povezav med koncepti upošteval tudi semantiko. Razvili so sistem Bio-SbKDS, ki z uporabo semantičnega znanja močno zmanjša prostor preiskovanja plavzibilnih domnev v procesu odkrivanja novega znanja. V primerjavi s prejšnjimi pristopi Bio-SbKDS uporabi semantično omrežje iz paketa UMLS, s katerim poišče ustrezne relacije med koncepti, le te pa filtrira s pomočjo semantičnih tipov.

Hristovski s sodelavci[47] nato uvede inovativen koncept "vzorec odkrivanja", ki vsebuje množico pogojev, katerim mora zadoščati novo odkrita raziskovalna domneva. Pogoji so sestavljeni iz kombinacij semantičnih relacij med koncepti, ekstrahiranimi iz bibliografske zbirke MEDLINE. Semantične relacije med koncepti so pridobljene s

pomočjo orodij BioMedLee[48] in SemRep.[49] BioMedMee je zbirka modulov za ekstrakcijo genotipskih in fenotipskih entitet iz besedila. S kombinacijo slovničnih pravil ter slovarja posameznim besedam in besednim zvezam pripiše ustrezno semantično in sintaktično kategorijo. SemRep je sistem za procesiranje naravnega jezika, ki je namenjen identifikaciji semantičnih relacij iz biomedicinskih besedil. Semantična relacija ima v splošnem obliko *(subjekt, predikat, objekt)*, kjer predikat izraža vsebino odnosa med subjektom in objektom. Uporabo vzorcev iskanja ilustrirajo na primeru Swansonove domneve "Raynaudov sindrom ↔ ribje olje", hkrati pa najdejo novo povezavo med Huntingtonovo boleznijo in inzulinom. Kasneje so iskanje ponovili še na primeru Parkinsonove bolezni in predlagali izbor antiepileptikov, ki vplivajo na povišanje funkcije GABA, za katero je znano, da je pri teh bolnikih močno znižana.[50] Ahlers in sodelavci[51] so enako metodologijo uporabili v zaprtem modelu iskanja, kjer so poskušali ugotoviti vmesne koncepte med antipsihotiki (angl. *antipsychotic agents*) in zdravljenjem raka. Odkrili so pet biomolekul, opisali možne interakcije z obema začetnima konceptoma, evalvacije pa ne podajo. Možnost odkrivanja novega znanja z odprtim in zaprtim modelom iskanja in uporabo semantičnih relacij je v sistemu EpiphaNet predstavil Cohen (2010) s sodelavci.[52] Gre za interaktivno aplikacijo, ki rezultate predstavi grafično v obliki omrežja konceptov. Zbirko znanja EpiphaNet zgradi na osnovi semantičnih predikatov iz baze SemMedDB.[53] Sistem lahko napoveduje sopojavnost parov konceptov ter s tem simulira proces odkrivanja novega znanja. EpiphaNet temelji na metodah distributivne semantike: povezave med pari konceptov napoveduje na osnovi metodologije reflektivnega slučajnega indeksiranja (angl. *reflective random indexing*). Wilkowski s sodelavci[20] je predstavil inovativen koncept pregledovalnega odkrivanja znanja, katerega namen je interaktivno odkrivanje znanja. Metoda uporabnika vodi skozi serijo vmesnih konceptov (*B*), kjer je ustreznost koncepta določena z njegovo stopnjo v omrežju. Znanje je opisano v obliki omrežja, v katerem vozlišča predstavljajo posamezne biomedicinske koncepte, povezave med vozlišči pa so opisane s semantičnimi relacijami iz sistema SemRep.

V zlati dobi funkcijske genomike se je več raziskovalcev ukvarjalo z izdelavo sistemov za pomoč pri interpretaciji eksperimentov DNA mikromrež.[2] V tem kontekstu zato omenjamo sistem SemBT, ki integrira rezultate poskusa z DNA mikromrežami s semantičnimi relacijami iz literature kot jih predlaga SemRep.[54] Mikromreže nam dajo vpogled v statistično značilno različno izražene gene pri danem eksperimentalnem pogoju, iz semantičnih relacij pa lahko razberemo interakcije genov z ostalimi biomedicinskimi koncepti. V tem okviru nam SemBT lahko pomaga pri interpretaciji rezultatov mikromrež ali pa pri načrtovanju novih terapevtskih postopkov za proučevano bolezen. Generiranje domnev v zadnjem primeru temelji na dveh vzorcih iskanja: "inhibiraj prekomerno izražen gen" ter "stimuliraj premalo izražen gen". Ideja v ozadju prvega vzorca iskanja je najti tak koncept (npr. učinkovino), ki bo inhibiral gen, ki je prekomerno izražen. Avtorji v nadaljevanju ponujajo več raziskovalnih domnev, ki se navezujejo na zdravljenje Parkinsonove bolezni. Kot možni učinkovini za zdravljenje Parkinsonove bolezni se ponujata laclitaxel in quecetin, saj se je ob pregledu literature izkazalo, da inhibirata gen HSP27 (HSPB1), ki se je v eksperimentalni študiji izkazal za prekomerno izraženega v testni skupini.

Semantični MEDLINE je spletna aplikacija, ki omogoča iskanje po predikatih, ekstrahiranih s pomočjo sistema SemRep.[53] Avtorji aplikacije zagotavljajo bolj ali manj svežo posodobitev ekstrahiranih predikatov. Za dani koncept sistem ekstrahira ustrezne predikate in izloči neinformativne koncepte, kot je npr. "Human". Nato izbere najpogostejše predikate in jih uporabniku predstavi v obliki neusmerjenega omrežja, v katerem vozlišča predstavljajo argumente (objekte oz. subjekte), povezave pa se navezujejo na semantične relacije med njimi. Semantični MEDLINE je za potrebe OZL prvi uporabil Miller s sodelavci.[55] Uporabili so zaprt model odkrivanja znanja iz literature, v katerem so vnaprej predpostavili povezavo (*A*) ↔ (*C*), da je raven testosterona povezana s spanjem, tako da vpliva na etiologijo slabega spanca pri odraslih moških. S pomočjo pregledovanja semantičnih relacij so nato poskušali najti vzročni mehanizem, ki bi to povezavo pojasnil. Kot možna rešitev se je izkazal kortizol, s katerim lahko pojasnimo manjkajočo korelacijo med nizko ravnijo testosterona pri moških in slabo kvaliteto njihovega spanja.

SemPathFinder temelji na semantični analizi poti v omrežju, ki omogoča »pripovedovanje zgodb«.[56] Pripovedovanje zgodb je potrebno razumeti v smislu sodobnega pristopa k predstavitvi podatkov, kot sprehajanje po poteh v omrežju, tako da skupaj povezujemo tiste pare nepovezanih konceptov, ki so si semantično blizu. SemPathFinder model *ABC* iskanja razširi, tako da namesto enega vmesnega koncepta vpelje serijo (*B*) konceptov. Izbira najustreznejše poti je še vedno prepuščena domenskemu ekspertu. Sistem je namenjen za odprto in zaprto odkrivanje znanja. Sistem so avtorji preizkusili na zaprtem modelu na domnevah

"Raynaudov sindrom ↔ ribje olje" in "migrena ↔ magnezij" ter potrdili uspešnost iskanja.

Zaključek

OZL je danes pomembna raziskovalna disciplina na področju rudarjenja tekstovnih podatkov. Kljub precejšnjemu napredku je glavni očitek vsem predstavljenim sistemom ta, da ne omogočajo popolnoma samodejnega rudarjenja in potrebujejo posredovanje človeka. Druga pomembna pomanjkljivost je uporaba metodologije za ustrezno ovrednotenje procesa OZL; večina sistemov ovrednotenje opravi kar na že preverjenih domnevah in zanemari statistično preverjanje veljavnosti.

Pregled, ki je za nami, poskuša biti celovit, vendar nikakor ni izčrpen. Zajeti poskuša tiste segmente OZL, ki so po mnenju avtorjev nujni za temeljit pregled vsebine. Menimo, da smo vanj vključili vse pomembne sklice na literaturo, ki bodo bralcu lahko v pomoč pri morebitnem nadaljnjem raziskovanju področja OZL.

Reference

1. Jensen LJ, Saric J, Bork P: Literature mining for the biologist: From information retrieval to biological discovery. *Nat Rev Genet* 2006; 7: 119-29.
2. Faro A, Giordano D, Spampinato C: Combining literature text mining with microarray data: Advances for system biology modeling. *Brief Bioinform* 2012; 13: 61-82.
3. Lawson AE: What does Galileo's discovery of Jupiter's moons tell us about the process of scientific discovery? *Sci Educ* 2002; 11: 1-24.
4. Simon HA, Newell A: Human problem solving: The state of the theory in 1970. *Am Psychol* 1971; 26: 145-159.
5. Clement JJ: *Creative model construction in scientists and students: The role of imagery, analogy, and mental simulation.* New York, NY 2008: Springer.
6. Nersessian N: *Creating scientific concepts.* Cambridge, MA 2008: MIT Press.
7. Kocabas S: *Elements of scientific creativity.* AAAI Technical Report SS-93-01 1993.
8. Clauset A, Larremore DB, Sinatra R: Data-driven predictions in the science of science. *Science* 2017; 355: 477-80
9. Uzzi B, Mukherjee S, Stringer M *et al.*: Atypical combinations and scientific impact. *Science* 2013; 342: 468-72.
10. Mednick AM: The associative basis of the creative process. *Psychol Rev* 1962; 69: 220-232.
11. Koestler A: *The art of creation.* London 1964: Penguin Books.
12. Microsoft research blog: *On Welsh Corgis, computer vision, and the power of deep learning.* http://bit.ly/2pp9ZQB (11.4.2017).
13. Langley P: Computational support of scientific discovery. *Int J Hum Comput Stud* 2000; 53: 393-410.
14. Wyatt J: Use and sources of medical knowledge. *Lancet* 1991; 338: 1368-73.
15. King RD, Whelan KE, Jones FM *et al.*: Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* 2004; 427: 247-2252.
16. Swanson DR: Literature-based knowledge discovery? The very idea. V: Bruza P, Weeber M (ur.), *Literature-based discovery.* Berlin 2008: Springer; 3-11.
17. Swanson DR: Undiscovered public knowledge. *Libr Q* 1986; 56: 103-118.
18. Gordon M, Lindsay RK, Fan W: Literature-Based Discovery on the World Wide Web. *ACM Trans Internet Tech* 2002; 2: 261-275.
19. Weeber M, Klein H, De Jong-Van Den Berg LTW *et al.*: Using concepts in literature-based discovery: Simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J Am Soc Inform Sci Tech* 2001; 52: 548-57.
20. Wilkowski B, Fiszman M, Miller CM *et al.*: Graph-based methods for discovery browsing with semantic predications. *AMIA Annual Symposium Proceedings* 2011, 1514-23.
21. van der Eijk CC, van Mulligen EM, Kors JA *et al.*: Constructing an associative concept space for literature-based discovery. *J Am Soc Inform Sci Tech* 2004; 55: 436-44.
22. Swanson DR: Fish Oil, Raynaud's syndrom, and undiscovered public knowledge. *Perspect Biol Med* 1986; 30: 7-18.
23. Digiacomo RA, Kremer JM, Shah DM: Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *Am J Med* 1989; 86: 158-64.
24. Swanson DR: Migraine and Magnesium: Eleven Neglected Connections. *Perspect Biol Med* 1988; 31: 526-557.
25. Swanson DR: Somatomedin C and arginine: Implicit connections between mutually isolated literatures. *Perspect Biol Med* 1990; 33: 157-86.
26. Smalheiser NR, Swanson DR: Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease. *Neurosci Res Commun* 1994; 15: 1-9.
27. Smalheiser NR, Swanson DR: Indomethacin and Alzheimer's disease. *Neurology* 1996; 46: 583.
28. Smalheiser NR, Swanson DR: Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology* 1996; 47: 809-10.
29. Smalheiser NR , Swanson DR: Calcium-independent phospholipase A2 and schizophrenia. *Arch Gen Psychiatry* 1998; 55: 752-3.
30. Swanson DR, Smalheiser NR: An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artif Intell* 1997; 91: 183-203.
31. Gordon MD, Lindsay RK: Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *J Am Soc Inform Sci* 1996; 47: 116-28.

32. Baeza-Yates R, Ribeiro-Neto B: *Modern information retrieval: The Concepts and Technology behind Search*. New York, NY 1999: ACM Press.
33. Lindsay RK, Gordon MD: Literature-based discovery by lexical statistics. *J Am Soc Inf Sci* 1999; 50: 575-87.
34. Gordon MD, Dumais S: Using latent semantic indexing for literature based discovery. *J Am Soc Inf Sci* 1998; 49: 674-85.
35. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *AMIA Annual Symposium Proceedings* 2001, 17-21.
36. Stegmann J, Grohmann G: Hypothesis generation guided by co-word clustering. *Scientometrics* 2003; 56: 111-35.
37. Callon M, Courtial JP, Laville F: Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics* 1991; 22: 155-205.
38. Pratt W, Yetisgen-Yildiz M: (2003). LitLinker: Capturing connections across the biomedical literature. In: Gennari JH, Porter BW, Gil Y (ur.), *Proceedings of the 2nd International Conference on Knowledge Capture*, Sanibel Island, FL, oktober 23-25. New York, NY, 2003: ACM; 105-112.
39. Srinivasan P: Text mining: Generating hypotheses from MEDLINE. *J Am Soc Inform Sci Tech* 2004; 55: 396-413.
40. Wren JD, Bekeredjian R, Stewart JA *et al.*: Knowledge discovery by automated identification and ranking of implicit relationships. *Bioinformatics* 2004; 20: 389-98.
41. Hristovski D, Stare J, Peterlin B *et al.*: Supporting discovery in medicine by association rule mining in Medline and UMLS. In: Patel V, Rogers R, Haux R (eds.), *MEDINFO 2001 Proceeding of the 10th World Congress on Medical Informatics*, London, UK, September 2-5. Amsterdam, NL, 2001: IOS Press; 1344-1348.
42. Hristovski D, Peterlin B, Mitchell JA *et al.*: Using literature-based discovery to identify disease candidate genes. *Int J Med Informat* 2005; 74: 289-98.
43. Fuller SS, Revere D, Bugni PF *et al.*: A knowledgebase system to enhance scientific discovery: Telemakus. *Biomed Digit Libr* 2004; 1: 2.
44. Petrič I, Urbaničič T, Cestnik B *et al.*: Literature mining method RaJoLink for uncovering relations between biomedical concepts. *J Biomed Inform* 2008; 42: 219-27.
45. Cameron D, Bodenreider O, Yalamanchili H *et al.*: A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *J Biomed Inform* 2013; 46: 238-51.
46. Hu X, Li G, Yoo I *et al.*: A semantic-based approach for mining undiscovered public knowledge from biomedical literature. In: Hu X, Liu Q, Skowron A, Lin TY, Yager RR, Zhang B (eds.), *Proceedings of the 2005 IEEE International Conference on Granular Computing*, Beijing, China, July 25-27. Piscataway, NJ, 2005: IEEE; 22-27.
47. Hristovski D, Friedman C, Rindflesch TC *et al.*: Exploiting semantic relations for literature-based discovery. *AMIA Annual Symposium Proceedings* 2006, 349-53.
48. Lussier Y, Borlawsky T, Rappaport D *et al.*: PhenoGO: Assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput* 2006; 64-75.
49. 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: 462-477.
50. Hristovski D, Friedman C, Rindflesch TC *et al.*: Literature-based knowledge discovery using natural language processing. In: Bruza P, Weeber M (eds.), *Literature-based discovery*. Berlin 2008: Springer; 133-52.
51. Ahlers CB, Hristovski D, Kilicoglu H *et al.*: Using the literature-based disccovery paradigm to investigate drug mechanisms. *AMIA Annual Symposium Proceedings* 2007, 6-10.
52. Cohen T, Whitfield GK, Schvaneveldt RW *et al.*: EpiphaNet: An interactive tool to support biomedical discoveries. *J Biomed Discov Collab* 2010; 5: 21-49.
53. Kilicoglu, H, Shin D, Fiszman M *et al.*: SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics* 2012; 28: 3158-60.
54. Hristovski D, Kastrin A, Peterlin B *et al.*: Combining semantic relations and DNA microarray data for novel hypotheses generation. V: Blaschke C, Shatkay H (ur.), *Linking Literature, Information, and Knowledge for Biology*. Berlin 2010: Springer; 53-61.
55. Miller CM, Rindflesch TC, Fiszman M *et al.*: A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep* 2012; 35: 279-285.
56. Song M, Heo GE, Ding Y: SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge. *J Informetr* 2015; 9: 686-703.