

Področje dela Odseka za tehnologije znanja so napredne informacijske tehnologije za zajemanje, shranjevanje in upravljanje znanja, ki so praktično uporabne za razvoj informacijske in na znanju temelječe družbe. Uveljavljena področja tehnologij znanja vključujejo inteligentno analizo podatkov, besedil in spleta (strojno učenje, rudarjenje podatkov, odkrivanje zakonitosti v podatkih), semantično rudarjenje podatkov in semantični splet, jezikovne tehnologije in računalniško jezikoslovje, podporo pri odločanju ter upravljanje znanja, novo področje raziskav pa je računalniška kreativnost. Poleg razvoja tehnologij znanja, ki večinoma temeljijo na metodah umetne inteligence, razvijamo tudi aplikacije teh tehnologij na področju znanosti o okolju, medicine in zdravstvenega varstva, bio-medicine in bioinformatike, ekonomije in tržništva, jezikoslovja in digitalne humanistike.

Leta 2014 smo bili vključeni v šest domačih in enajst evropskih projektov 7. OP, dve akciji COST, dva bilateralna projekta, v mrežo Evropske znanstvene fundacije, en infrastrukturni projekt ter dva tržna projekta. Na odseku se je usposabljal deset mladih raziskovalcev.

Na področju inteligentne analize in rudarjenja podatkov smo razvili več novih metod in jih uporabili v raznovrstnih aplikacijah. Razvili smo metodo, ki omogoča semantično in vizualno pojasnjevanje statističnih modelov ter metodo za semantično pojasnjevanje avtomatsko odkritih podskupin v podatkih, ki smo jo objavili v reviji Journal of intelligent information systems. Metodologijo NoiseRank za ansambelsko odkrivanje osamelcev v podatkih ter ViperCharts platformo za vizualni prikaz rezultatov smo objavili v prestižni reviji Data Mining

and Knowledge Discovery. Razvili smo novo metodologijo za večopisno gručenje podatkov, ki je bila uspešno uporabljena na področju analize finančnih trgov. Novo metodologijo za analizo sentimenta v tokovih sporočil twitter smo uspešno implementirali v naši spletni platformi ClowdFlows in jo uspešno uporabili pri napovedovanju cen delnic - delo je bilo objavljeno v dveh revijah z visokim faktorjem vpliva: Information processing & management in Information sciences. Z Odsekom za inteligentne sisteme IJS smo sodelovali pri razvoju interaktivne metode rudarjenja podatkov HMDM (Human-Machine Data Mining).

Razvili smo nove metode za učenje dreves in ansamblov za napovedovanje strukturiranih vrednosti (večciljno klasifikacijo in regresijo ter (hierarhično) večoznačno klasifikacijo) in jih uporabili za iskanje in označevanje slik, modeliranje dinamičnih sistemov in za obravnavanje različnih problemov s področij znanosti o okolju in življenju. Pri tem

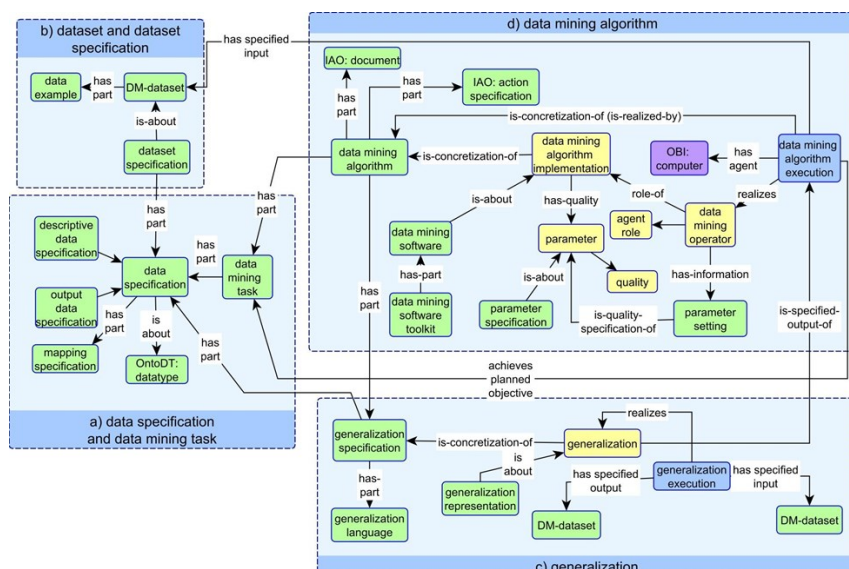
smo pokazali, da z uporabo hierarhij pri hierarhični večoznačni klasifikaciji dobimo boljše rezultate, tako pri napovedovanju strukture združb kot pri drugih nalogah. Enak pristop smo uspešno uporabili pri modeliranju habitata ekstremofilnih gliv, ki lahko nastopajo tudi kot oportunistični patogeni. Razvili smo tudi novo ontologijo podatkovnega rudarjenja OntoDM, ki je v primerjavi z obstoječimi precej globlja in bolj splošna ter dobro opisuje zgoraj omenjene naloge in algoritme za rudarjenje strukturiranih podatkov.

Metode za učenje modelov dinamičnih sistemov iz podatkov in domenskega predznanja smo nadalje razvijali v smeri učenja ansamblov več takih modelov.

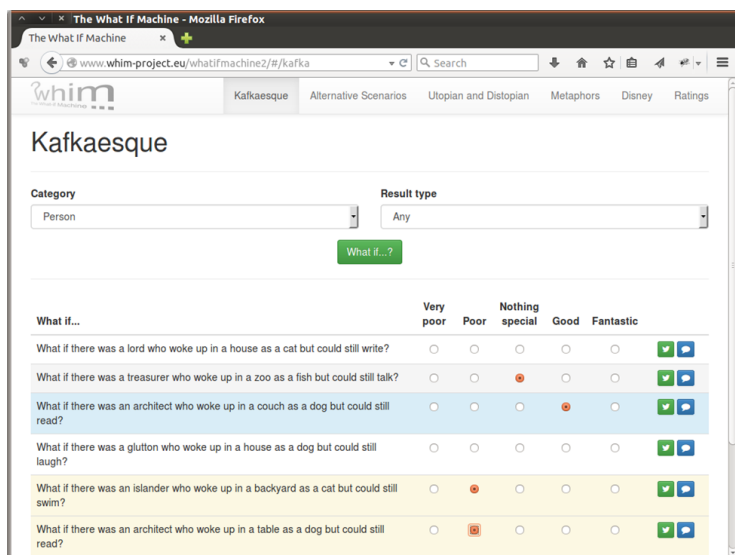


Vodja:
Prof. Nada Lavrač

V aprilu leta 2014 se je Odsek za tehnologije znanja priključil projektu Človeški možgani (Human Brain Project), ki je eden od dveh megaprojektov financiranih v okviru programa EU FET Flagship.



Slika 1: Del ontologije podatkovnega rudarjenja OntoDM-core.



Slika 2: Spletni vmesnik "Kajče stroja" in njegovih komponent za ocenjevanje z množičenjem.

Slednje raziskave so potekale v okviru evropskega projekta 7. OP SUMO (Supermodeling by combining imperfect models), ki je bil uspešno zaključen v letošnjem letu. Razvite metode smo uporabili za modeliranje vodnih ekosistemov in celotnih porečij. Zaključili smo tudi evropski projekt 7. OP REWIRE (Rehabilitative Wayout in Responsive home Environments), kjer smo razvijali sistem za rehabilitacijo pacientov, ki so utrpeli možgansko kap. V REWIRE smo uporabili metode strojnega učenja za analizo podatkov o bolniku s ciljem iskanja povezav med stanjem pacienta ob sprejemu v bolnišnico in kakovostjo njegovega življenja po rehabilitaciji (leto kasneje).

Začeli smo s koordinacijo projekta FET MAESTRA, ki obravnava naloge analize kompleksnih podatkov. Cilj projekta je razvoj nove generacije metod napovednega modeliranja, ki lahko sočasno obravnavajo

različne vidike kompleksnosti podatkov. Ti vključujejo napovedovanje strukturiranih vrednosti v okviru zelo velikih, omrežnih in nepopolno označenih podatkov. Pridružili smo se projektu Človeški možgani (Human Brain Project, FET Flagship), kjer sodelujemo pri razvoju platforme za medicinsko informatiko. Platforma bo omogočala analizo velikih količin podatkov, rutinsko zbranih v bolnišnicah za diagnostične namene. V ta namen razvijamo metode za razvrščanje v skupine s pravili, ki bodo temeljile na pristopih za iskanje podskupin ter napovedno razvrščanje in omogočajo gradnjo razumljivih opisov skupin.

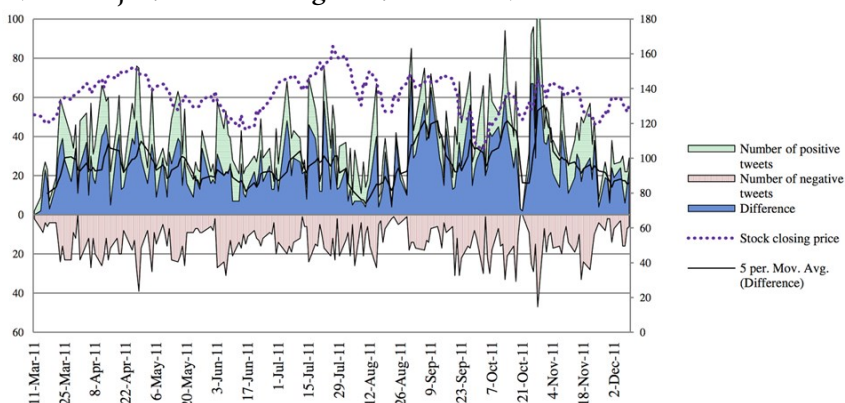
Na področju računalniške kreativnosti smo metodo ukrivljenih matrik (banded matrices), ki smo jo uporabili na področju ustvarjanja pripovedi iz stavkov iz različnih korpusov, ovrednotili v primerjavi z drugimi metodami za ideacijo izmišljenih pripovedi. V sodelovanju s podjetjem XLAB

Metoda "NoiseRank" za odkrivanja in rangiranja šuma in anomalij v podatkih je bila s strani Znanstvenoraziskovalnega sveta za tehniške vede ARRS prepoznana kot izjemen znanstveni dosežek in je bila predstavljena na dogodku "Odlični v znanosti 2013" v sklopu 9. Slovenskega foruma inovacij 2014.

smo razvili metodologijo za avtomatsko ustvarjanje sloganov, ki temelji na uporabi kombinacije računalniškega jezikoslovja, semantičnih virov in genetskih algoritmov.

Nadaljujemo sodelovanje v evropskem projektu 7. OP s področja računalniškega razumevanja jezika MUSE (Machine Understanding for interactive Storytelling), katerega cilj je prikaz besedil v animirani obliki 3D. Uspešno smo sodelovali tudi v več

evropskih projektih 7. OP s področja računalniške kreativnosti: že od leta 2012 smo vključeni v evropski projekt 7. OP PROSECCO, katerega cilj je promocija aktivnosti na področju računalniške kreativnosti, od leta 2013 dalje pa sodelujemo v projektih ConCreTe (Concept Creation Technologies) in odlično ocenjenem projektu WHIM (The What-If Machine). V letu 2014 smo organizirali Peto mednarodno konferenco s področja računalniške kreativnosti (The Fifth International Conference on Computational Creativity, ICC3 2014) v Ljubljani in Sedemnajsto mednarodno konferenco o znanosti odkrivanja (The Seventeenth International Conference on Discovery Science, DS 2014) na Bledu, vsake se je udeležilo okrog sto udeležencev.



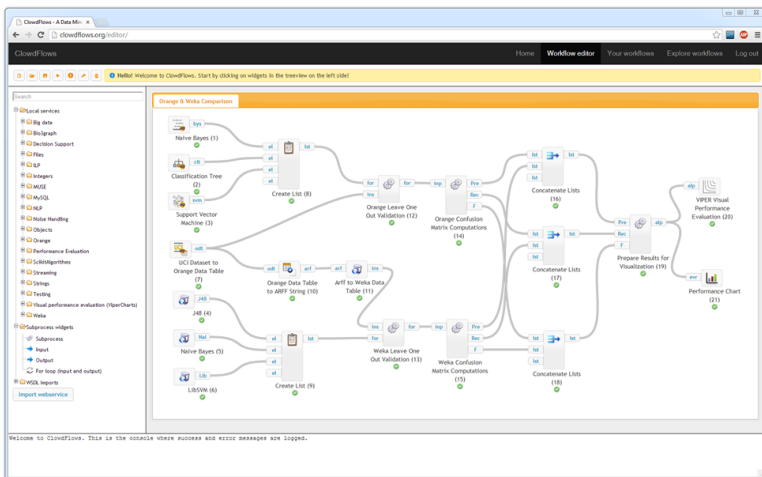
Slika 3: Grafični prikaz sentimeta Twitter sporočil, ki omenjajo delnice podjetja Baidu.

Na področju analize besedil, spletnih podatkov in heterogenih informacijskih omrežij smo uspešno zaključili FET projekt FOC (Forecasting Financial Crises), ki je bil ocenjen kot odličen. Začeli smo delo na projektih SIMPOL (Financial Systems Simulation and Policy Modelling) in MULTIPLEX (Foundational Research on MULTilevel complex networks and systems).

Namen projekta SIMPOL je podpora odločevalcem in regulatorjem pri modeliranju politik in analizi njihovih posledic, predvsem na področju financiranja vzdržnih in okolju prijaznih projektov. Razvili smo

platformo, ki podpira različne metode zbiranja in analize relevantnih podatkov. Predstavitev podatkov temelji na grafih in na principih semantičnega spleta (Semantic Web). Podatki se shranjujejo v bazi Neo4j, ki je zmogljiva baza, ki ne temelji na SQL, pač pa je razvita za učinkovito manipulacijo in poizvedovanje po grafih. Implementirali smo spletni portal <http://simpol.ijs.si/>, ki uporablja Neo4j v ozadju. Portal podpira vse predvidene aktivnosti za izgradnjo socio-ekonomskih omrežij: pregled, izbiro in vpogled v obstoječa omrežja v bazi, ekstrakcijo omrežij iz povezanih odprtih podatkov (Linked Open Data), uvoz in izvoz v platformo za množičenje (crowdsourcing), spremljanje in ekstrakcijo podatkov iz novic, spletnih dnevnikov ter socialnih omrežij.

Cilj projekta MULTIPLEX je razvoj matematičnega ogrodja kompleksnih omrežij in algoritmov, ki bi vzpostavil teoretično osnovo za razumevanje, napovedovanje in (morda) kontrolo kompleksnih omrežij. Naša vloga v projektu je konstrukcija večnivojskih omrežij iz tokov besedilnih podatkov, t.j. iz novic, spletnih dnevnikov in Twitterja. Iz tvitov povezanih z okoljsko problematiko smo izdelali "retweet" omrežje, označili uporabnike z njihovim vplivom in njihove tvite s sentimentom, ter identificirali glavne skupnosti vplivnih akterjev. Na osnovi sentimenta prirejenega skupnostim smo analizirali njihov odnos do posameznih okoljskih tematik. Iz finančnih novic smo ustvarili časovno



Slika 4: Tipični delotok za rudarjenje podatkov v spletni platformi CloudFlows.

V letu 2014 je Odsek organiziral tri mednarodne konference: Peto mednarodno konferenco s področja računalniške kreativnosti (The Fifth International Conference on Computational Creativity, ICC3 2014), Sedemnajsto mednarodno konferenco o znanosti odkrivanja (The Seventeenth International Conference on Discovery Science, DS 2014) in Deveto konferenco o jezikovnih tehnologijah v okviru multikonference Informacijska družba (The Ninth Language Technologies Conference at the Information Society Multiconference, LT@IS 2014).

korpus starejšega slovenskega jezika. Izvedli smo študijo usmerjeno na uporabniške izkušnje s temi spletnimi jezikovnimi viri, ki je npr. pokazala, da je večina uporabnikov (predvsem učitelji, profesorji in študentje) mnenja, da so ti viri pomembni tako za njihovo delo, kot v splošnem za slovenščino, pridobili pa smo tudi koristne nasvete za izboljšave. Nadaljevali smo tudi sodelovanje s SAZU na izdelavi njihovega Slovenskega biografskega leksikona.

Sodelovali smo v razvoju koncepta načrtovanega novega slovarja slovenskega jezika, v spodbujanju odprtega dostopa do jezikovnih virov kot znanstvenih podatkov ter v izgradnji slovenske raziskovalne infrastrukture CLARIN.SI. Ta infrastruktura je sedaj organizirana kot konzorcij s, trenutno, devetimi partnerji: tremi slovenskimi univerzami (Ljubljana, Maribor, Primorska), dvema raziskovalnima inštitutoma (ZRC SAZU, IJS), kot tudi društvu (SDJT, DZDR), zavodu (Trojina) in podjetji (Amebis, Alpineon), ki se ukvarjajo z jezikovnimi viri in tehnologijami. Infrastruktura CLARIN.SI sedaj izpolnjuje organizacijske in tehnične pogoje, da v 2015 postane članica evropske infrastrukture CLARIN ERIC, Sloveniji pa naj bi dobila stabilen in trajen repozitorij jezikovnih virov ter platformo za spletne jezikovnotehnološke servise.

Oktobra 2014 smo, v sklopu krovne konference Informacijska družba, organizirali Deveto konferenco Jezikovne tehnologije. Ta srečanja so glavni forum za predstavitev raziskav in razvoja jezikovnih tehnologij v Sloveniji kot tudi na

spremenljiva omrežja sopojavitev 50 različnih držav. Ta omrežja smo primerjali s finančnimi, trgovskimi in geografskimi omrežji med državami. Izkazuje se, da novice najbolj odražajo geografsko bližino držav. Če pa se omejimo na novice, ki imajo pozitiven sentiment, pa so tako konstruirana omrežja najbolj podobna trgovskim omrežjem.

Na področju jezikovnih tehnologij in digitalne humanistike delamo predvsem na izdelavi jezikovnih virov in metod za jezikoslovno označevanje besedil, s poudarkom na slovenskem jeziku. V 2014 smo začeli z delom na projektu JANES (Jezikoslovna analiza nestandardne slovenščine), <http://nl.ijs.si/janes/>, ki ga vodi Oddelek za prevajalstvo Filozofske fakultete Univerze v Ljubljani. Izdelali smo slWaC 2.0, nov korpus slovenščine zajete s spleta, ki vsebuje čez milijardo besed. Razvili smo tudi orodje, imenovano TweetCaT, za zbiranje tvitov v ciljnih jezikih in metodo za normalizacijo nestandardnega jezika, ki temelji na statističnem strojnem prevajanju. S temi orodji smo izdelali prvi velik jezikoslovno označen korpus slovenskih tvitov, ki vsebuje preko 35 milijonov besed. Začeli smo tudi z delom na zajemu komentarjev, forumov in blogov s spleta.

Pred kratkim zaključen projekt IMP, <http://nl.ijs.si/imp/>, je izdelal digitalno knjižnico, slovar in

Hrvaškem. Zbornik konference 2014 vsebuje dve vabljeni predavanji tujih znanstvenikov in 29 recenziranih prispevkov s 180 stranmi. Jezika konference sta tako slovenski kot angleški, vsak prispevek pa ima povzetka v obeh jezikih.

Nadaljevali smo delo v sklopu ESF mreže NetWordsS (European Network on Word Structure), in akcije COST PARSEME (Parsing and Multi-word Expressions), začeli pa delo na dveh bilateralnih projektih, eden s Hrvaško in drugi s Srbijo. Sodelovali smo s SIST, Slovenskim institutom za standardizacijo, v tehničnem odboru SIST/TC IDT »Informatika, dokumentacija in splošna terminologija« in v tehničnem odboru ISO TC37 »Terminology and other language and content resources«, kjer v SC4 »Language resource management« sodelujemo pri pripravi standardov s področja zapisa jezikovnih virov. Pri Ministrstvu za kulturo RS smo pomagali v pripravi Akcijskega načrta za izvajanje nacionalnega programa za jezikovno politiko 2014-2018.

Dolgoročni cilj na področju podpore pri odločanju je razvijati metode in tehnike odločitvenega modeliranja, jih podpreti z računalniškimi orodji ter jih povezovati s sistemi za rudarjenje podatkov. V letu 2014 smo posebno pozornost posvetili nadgradnji kvalitativne večparametrsk metode DEX z možnostmi obravnavanja numeričnih kriterijev, verjetnostnih porazdelitev in relacijsko opredeljenih odločitvenih alternativ ter razvoju ustreznih podpornih orodij. Posodobili smo računalniški program za večparametrsko modeliranje DEXi in ga dopolnili z možnostmi grafičnega pregledovanja funkcij koristnosti. Izvedli smo raziskavo, v kateri smo preverili možnost uporabe združevalno-razdruževalnih metod UTA in ACUTA za prikaz funkcij koristnosti v metodi DEX. Na področju uporabe omenjenih metod smo uspešno zaključili dva večja projekta: projekt 7. OP FIRST, kjer smo razvili sistem za ocenjevanje ugleda, in domači projekt OVJE, kjer smo razvili odločitvene modele in sistem za ocenjevanje vzdržnosti za razvoj energetike v Sloveniji do leta 2030 s poudarkom na jedrski tehnologiji. Druge praktične aplikacije se nanašajo na upravljanje v kriznih razmerah, strateško upravljanje v podjetjih, podporo pri odločanju v zavarovanih območjih in proizvodnjo igralniških naprav.

V okviru projekta EVADIFF (Evaluation et développement de modèles et outils d'aide à la décision utilisés pour la prévention des pollutions diffuses par les produits phytopharmaceutiques), naročnik je francoski institut ARVALIS (Institut du végétal) iz Francije, smo razvili sistem za podporo odločanja glede izbire ukrepov za zmanjševanje

IZBRANE OBJAVE

1. Ikonovska, E., Gama, J., Džeroski, S. Online tree-based ensembles and option trees for regression on evolving data streams. *Neurocomputing*, ISSN 0925-2312, 2015, vol. 150, part. 150, pp. 458-470, doi: 10.1016/j.neucom.2014.04.076.
2. Levatić, J., Kocev, D., Džeroski, S. The importance of the label hierarchy in hierarchical multi-label classification. *Journal of intelligent information systems*, ISSN 0925-9902, [in press] 2014, 25 pp., doi: 10.1007/s10844-014-0347-y.
3. Miljković, D., Depolli, M., Stare, T., Mozetič, I., Petek, M., Gruden, K., Lavrač, N. Plant defence model revisions through iterative minimisation of constraint violations. *International journal of computational biology and drug design*, ISSN 1756-0756, 2014, vol. 7, no. 1, pp. 61-79, doi: 10.1504/IJCBD.2014.058588.
4. Panov, P., Soldatova, L. N., Džeroski, S. Ontology of core data mining entities. *Data mining and knowledge discovery*, ISSN 1384-5810, [in press] 2014, 44 pp., doi: 10.1007/s10618-014-0363-0.
5. Piškorec, M., Antulov-Frantulin, N., Kralj Novak, P., Mozetič, I., Grčar, M., Vodenska I., Šmuc, T. Cohesiveness in Financial News and its Relation to Market Volatility, *Scientific Reports* 4, 5038-1-8, 2014.
6. Popovič, M., Štefančič, H., Sluban, B., Kralj Novak, P., Grčar, M., Mozetič, I., Puliga, M., Zlatić, V. Extraction of Temporal Networks from Term Co-occurrences in Online Textual Sources, *PLoS ONE* 9(12), e99515, 2014.
7. Ramšak, Ž., Baebler, Š., Rotter, A., Korbar, M., Mozetič, I., Usadel, B., Gruden, K. GoMapMan: integration, consolidation and visualisation of plant gene annotations within the MapMan ontology, *Nucleic Acids Research* 42(D1), D1167-D1175, 2014.
8. Sluban, B., Gamberger, D., Lavrač, N. Ensemble-based noise detection: noise ranking and visual performance. *Data mining and knowledge discovery*, ISSN 1384-5810, 2014, vol. 28, no. 2, pp. 265-303.
9. Smailović, J., Grčar, M., Lavrač, N., Žnidaršič, M. Stream-based active learning for sentiment analysis in the financial domain. *Information sciences*, ISSN 0020-0255, 2014, vol. 285, pp. 181-203, doi: 10.1016/j.ins.2014.04.034.
10. Vavpetič, A., Podpečan, V., Lavrač, N. Semantic subgroup explanations. *Journal of intelligent information systems*, ISSN 0925-9902, 2014, vol. 42, no. 2, pp. 233-254, doi: 10.1007/s10844-013-0292-1.