

DATA A ZNALOSTI & WIKT 2018

Sborník konference

Editori

Jaroslav Zendulka
Mária Bieliková
Radek Burget
Zbyněk Křivka

*Fakulta informačních technologií
Vysoké učení technické v Brně*

Brno, Hotel SANTON
Česká republika
11.-12. října 2018

© Vysoké učení technické v Brně, 2018

ISBN 978-80-214-5679-2

Autoři jsou uvedeni v obsahu. Příspěvky jsou publikovány ve tvaru, v jakém je dodali autoři, bez podstatných změn. Každý příspěvek byl recenzován, recenzenti jsou členy programových výborů konferencí.

Předmluva

Informační technologie a data jimi zpracovávaná se staly nedílnou součástí našeho každodenního života. Stejně tak jsme svědky rostoucí snahy o maximální využití netriviálních, skrytých informací, které lze z dat extrahovat a následně využít. Výzkum a vývoj aplikací v oblasti databázových technologií, informačních systémů, datového a znalostního inženýrství má v Česku i na Slovensku poměrně dlouhou tradici. I přes nezpochybnitelnou potřebu prezentovat a konfrontovat dosahované výsledky na mezinárodním poli, mají i národní, resp. česko-slovenská setkání odborníků z univerzit i průmyslové praxe v systému vzájemného informování se a předávání zkušeností své místo. Jedním z takových významných komunitních setkání je i konference Data a znalosti, která se konala společně s tematicky blízkým seminářem WIKT (Workshop on Intelligent and Knowledge oriented Technologies) pod názvem Data a znalosti & WIKT 2018 ve dnech 11. a 12. října 2018 v Brně v hotelu Santon u brněnské přehrady.

V případě konference Data a znalosti šlo o 4. ročník konference, která navazuje na dlouholetou tradici dvou prestižních česko-slovenských konferencí – na konferenci Datakon a konferenci Znalosti. Konference Datakon existovala od roku 2000, kdy navázala na původně databázovou konferenci Datasem s tehdy již dvacetiletou tradicí. Konference Znalosti byla organizována od roku 2001. Seminář WIKT je organizován od roku 2006 na Slovensku s účastníky ze Slovenska i Česka. Jak v případě konference Data a znalosti, tak semináře WIKT tvoří významnou část účastníků studenti doktorských studijních programů prezentující své výsledky a získávající užitečnou zpětnou vazbu ke směřování jejich výzkumu. Poprvé byly obě akce organizovány společně v roce 2016 ve Smolenicích na Slovensku.

Volba hotelu Santon pro letošní ročník konference pořádaný v Brně nebyla náhodná. V letech 1991 až 2009 se právě v tomto hotelu konalo 19 ročníků konference Datasem a Datakon.

Hlavními vypsányi tematickými okruhy konference v letošním roce byly:

- Cloud, technická řešení, aplikační příklady
- Datově centrické pojetí bezpečnosti
- Technologie blockchain
- Grafové databáze
- Integrace relačních a NoSQL databází
- Multi-modelové databázové architektury
- Získávání, ukládání, zpracování a vizualizace rozsáhlých dat (Big Data)
- Tvorba, publikování a využívání otevřených a propojených dat (Linked Data)
- Indexování a vyhledávání textových a multimediálních dat
- Procesy a role ve správě dat
- Podniková architektura
- Strojové učení, hluboké učení, data mining (Data Science)
- Detekce anomálií
- Anonymizace a zachování soukromí při dolování z dat

- Aplikace strojového učení v bioinformatice, v počítačovém vidění a při zpracování řeči
- Modelování uživatelů, adaptivní a personalizované systémy
- Pokročilá uživatelská rozhraní softwarových a informačních systémů
- Systémy pro správu znalostí v organizacích
- Expertní, inteligentní a agentní systémy, počítačová inteligence
- Počítačová lingvistika
- Ontologické a konceptuální modely
- Automatické odvozování a plánování
- Sociální web a jeho aplikace, analýza sociálních sítí
- Personalizovaný web a jeho aplikace, doporučování
- Zpracování informačních zdrojů v českém/slovenském jazyce
- Sémanticky a servisně orientované architektury
- Znalostní technologie a jejich aplikace
- Znalostní báze, usuzování a odvozování

Autoři posílali příspěvky v českém, slovenském nebo anglickém jazyce do následujících kategorií:

- výzkumný příspěvek,
- příspěvek o probíhajícím výzkumu,
- aplikační příspěvek,
- vizionářský příspěvek,
- projektový příspěvek,
- doktorandské symposium.

Navíc bylo pozváno pět autorů k prezentaci zvaných přednášek pojednávajících o použití databáze SAP HANA v podnikových informačních systémech, správě dat v prostředí velké organizace, využívání strojového učení společností Seznam.cz, aplikaci ontologického modelu společností Bisnode a některých aspektech kybernetické bezpečnosti v souvislosti s nařízením GDPR.

Celkem bylo podáno 49 příspěvků. Nejvíce byly zastoupeny výzkumné příspěvky a příspěvky o probíhajícím výzkumu. Všechny příspěvky byly posuzovány dvojicí oponentů z řad členů programového výboru. Přijato bylo 43 příspěvků, z toho 23 bylo prezentováno formou krátkého sdělení, 13 formou úvodu a následné diskuse u posteru a 7 doktorandských příspěvků formou delšího úvodu a diskuse u posteru.

Konference Data a znalosti & WIKT 2018 měla celkem 70 účastníků. Je potěšitelné, že byly zastoupeny všechny výzkumné skupiny z Česka a Slovenska, které na konferencích Data a znalosti a semináři WIKT tradičně prezentují své výsledky.

Děkujeme členům programového výboru, kteří se ochotně podíleli na posuzování příspěvků. Náš velký dík patří i organizačnímu výboru a dalším kolegům z Fakulty informačních technologií VUT v Brně, zejména Radkovi Burgetovi a Zbyňku Křivkovi za zajištění místa konání konference, přípravu sborníku a ostatní organizační činnosti během přípravy a konání konference.

V neposlední řadě patří naše poděkování partnerům konference, kterými byly společnosti SAP ČR, spol. s r.o., Česká společnost pro kybernetiku a informatiku,

Profinit EU, s.r.o. a Red Hat Czech s.r.o. Svou podporou umožnily konání konference za přijatelných finančních podmínek a formou slevy pro studenty doktorských studijních programů podnítily jejich účast.

Naše poděkování za podporu a vstřícnost patří i Fakultě informačních technologií VUT v Brně a jejímu děkanovi prof. Dr. Ing. Pavlu Zemčikovi, který letošní ročník konference zaštitil.

Brno, říjen 2018

Jaroslav Zendulka a Mária Bieliková

Organizace konference

Stálý řídicí výbor konference Data a znalosti

Předseda (Chair): *Dušan Chlapek*, VŠE Praha
Členové (Members): *Mária Bieliková*, FIIT STU Bratislava
Radek Burget, FIT VUT Brno
Tomáš Horváth, UPJŠ Košice
Petr Hujňák, Per Parties Consulting Praha
Pavel Kordík, FIT ČVUT Praha
Karol Matiaško, ŽU Žilina
Ján Paralič, FEI TU Košice
Jaroslav Pokorný, MFF UK Praha
Lubomír Popelínský, FI MU Brno
Jan Rauch, VŠE Praha
Karel Richta, FEL ČVUT Praha
Vojtěch Svátek, VŠE Praha
Petr Šaloun, VŠB-TU Ostrava
Michal Valenta, FIT ČVUT Praha

Stálý řídicí výbor WIKT

Členové (Members): *Mária Bieliková*, FIIT STU Bratislava
Peter Butka, FEI TU Košice
Ladislav Hluchý, ÚI SAV Bratislava
Martin Homola, FMFI UK Bratislava
Tomáš Horváth, PrF UPJŠ Košice
Daniela Chudá, FIIT STU Bratislava
Stanislav Krajčí, PrF UPJŠ Košice
Michal Laclavík, Magnetic, a.s.
Marián Mach, FEI TU Košice
Kristína Machová, FEI TU Košice
Karol Matiaško, FRI ŽU Žilina
Pavol Návrat, FIIT STU Bratislava
Ján Paralič, FEI TU Košice
Viera Rozinajová, FIIT STU Bratislava
Petr Šaloun, FEI VŠB-TU Ostrava
Peter Vojtáš, MFF UK Praha
Jaroslav Zendulka, FIT VUT Brno

Programový výbor

Předseda (Chair): *Mária Bieliková*, FIIT STU Bratislava
Jaroslav Zendulka, FIT VUT Brno

Členové (Members): *František Babič*, FEI TU Košice
Michal Barla, FIIT STU Bratislava
Roman Barták, MFF UK Praha
Vladimír Bartík, FIT VUT Brno
Petr Berka, VŠE Praha
Jan Boháčik, FRI ŽU Žilina
Radek Burget, FIT VUT Brno
Peter Butka, FEI TU Košice
Martin Dostal, ZČU Plzeň
Igor Farkaš, FMFI UK Bratislava
Dalibor Fiala, ZČU Plzeň
Ján Genči, FEI TU Košice
Peter Gurský, PrF UPJŠ Košice
Zdeněk Havlice, TU Košice
Ladislav Hluchý, SAV Bratislava
Martin Holeňa, ÚI AV ČR
Irena Holubová, MFF UK Praha
Tomáš Horváth, PrF UPJŠ Košice
Tomáš Hruška, FIT VUT Brno
Petr Hujňák, Per Parties Consulting Praha
Dušan Chlapek, VŠE Praha
Daniela Chudá, FIIT STU Bratislava
Jana Klečková, ZČU Plzeň
Jiří Kléma, FEL ČVUT Praha
Jakub Klímek, FIT ČVUT Praha
Michal Kompan, FIIT STU Bratislava
Pavel Kordík, FIT ČVUT Praha
Stanislav Krajčí, UPJŠ Košice
Pavel Král, ZČU Plzeň
Petr Křemen, FEL ČVUT Praha
Zbyněk Křivka, FIT VUT Brno
Petr Kučera, Komix
Martin Labský, IBM TJW, Praha
Peter Lacko, FIIT STU Bratislava
Michal Laclavík, ÚI SAV Bratislava
Vitaly Levashenko, FRI ŽU Žilina

Marian Mach, TU Košice
Kristina Machová, TU Košice
Karol Matiaško, ŽU Žilina
Peter Mikulecký, Univerzita Hradec Králové
Martin Molhanec, FEL ČVUT Praha
Roman Mouček, ZČU Plzeň
Pavol Návrat, FIIT STU Bratislava
Giang Nguyen, UI SAV Bratislava
Ján Paralič, FEI TU Košice
Robert Pergl, FIT ČVUT Praha
Tomáš Pitner, FI MU Brno
Jaroslav Pokorný, MFF UK Praha
Lubomír Popelínský, FI MU Brno
Jan Rauch, VŠE Praha
Václav Řepa, FIS VŠE Praha
Karel Richta, FEL ČVUT Praha
Viera Rozinajová, FIIT STU Bratislava
Hana Rudová, FI MU Brno
Pavel Smrž, FIT VUT Brno
Ivan Srba, FIIT STU Bratislava
Josef Steinberger, ZČU Plzeň
Vojtěch Svátek, VŠE Praha
Petr Šaloun, VŠB-TU Ostrava
Jozef Tvarožek, FIIT STU Bratislava
Michal Valenta, FIT ČVUT Praha
Peter Vojtáš, MFF UK Praha
Ondřej Zamazal, VŠE Praha
Martin Zima, ZČU Plzeň

Organizační výbor

Předseda (Chair): *Radek Burget*, FIT VUT Brno
Členové (Members): *Michaela Bílková*, FIT VUT Brno
Zbyněk Křivka, FIT VUT Brno
Jaroslav Zendulka, FIT VUT Brno

Konferenci organizuje



Partneři konference

Sponsored by



PROFINIT



ČSKI

Obsah

Předmluva	3
<i>Mária Bielíková, Jaroslav Zendulka</i>	

Zvané přednášky

Užití sloupcové in-memory databáze SAP HANA v podnikových informačních systémech	17
<i>Radim Benek</i>	
Strojové učení v Seznam.cz	21
<i>Vladimír Kadlec</i>	
Aplikace ontologického modelu pro potřeby mezinárodního produktu nad heterogenními business daty v souladu s GDPR	23
<i>Patrik Kompuš</i>	
Kybernetická bezpečnost a GDPR	25
<i>Jan Kolouch</i>	
Přístupy ke správě dat v prostředí velké organizace	37
<i>Dušan Chlapek</i>	

Vybrané příspěvky

Integrace relačních a grafových databází funkcionálně	43
<i>Jaroslav Pokorný</i>	
Interaktivní a grafická podpora výuky dotazovacího jazyka SPARQL	49
<i>Vojtěch Svátek, Jindřich Mynarz, Marek Dudáš, Jan Kučera</i>	
Selinon – Dynamic Distributed Task Flows	53
<i>Fridolín Pokorný</i>	
Znalostné přístupy pre inteligentnú analýzu veľkých dát	59
<i>Ján Paralič, Peter Bednár, Kristína Machová, Peter Butka, František Babič, Martin Sarnovský, Vladimír Gašpar</i>	

REBELION – odhaľovanie antisociálneho správania na Webe	65
<i>Pavol Návrat, Daniela Chudá, Mária Bielíková, Marián Šimko, Jakub Šimko, Ivan Srba, Michal Kompan, Róbert Móro, Irina Malkin Ondík, Peter Lacko, Alena Martonová, Kristína Machová, Ján Paralič, Peter Butka, Peter Bednár, Martin Sarnovský, Barbora Mesárošová, Radoslav Blaho, Lucia Sabová</i>	
Bioinformatika – Výzvy a príležitosti v proteomike	71
<i>Ján Genčí</i>	
Performing Feature Selection Before Removing Outliers To Increase Classifier's Accuracy	77
<i>Nataliia Petliak, Oleksandra Tmenova, Matúš Námešný, Tomáš Bončo, Luboš Popelínský</i>	
Výber vhodných atribútov pre tvorbu klasifikačných modelov v oblasti medicíny	83
<i>Michal Vadovský, Ján Paralič</i>	
Strojové učení při analýze rozsáhlých astronomických datasetů s časovými řadami	89
<i>David Andrešič, Petr Šaloun, Bronislava Pečíková</i>	
Určovanie zvedavosti používateľov z ich správania v doméne online obchodu	95
<i>Jakub Šimko, Sandra Kostová, Mária Bielíková</i>	
Inferencia typu osobnosti používateľa v elektronickom obchode	101
<i>Robert Móro, Veronika Balážová, Mária Bielíková</i>	
Interaktívny nástroj pre analýzu a vizualizáciu textových dokumentov	107
<i>Miroslav Smatana, Peter Butka, Dominika Maršáleková</i>	
Vyhledávání obrázků v rozšířené realitě	111
<i>Petr Chmelař, Jaroslav Kuchař</i>	
Concepts and Relations Acquisition within Ontology Learning Process for Automotive	115
<i>Ondřej Šebek, Václav Jirkovský, Nestor Rychtycký</i>	
Data Stewardship Wizard for Open Science	121
<i>Marek Suchánek, Robert Pergl</i>	
Semantic Support for Facilitating Production Monitoring	127
<i>Václav Jirkovský, Petr Kadera, Marek Obitko</i>	

Information Modelling in Evidence Based Health Care 133
Václav Řepa

Inteligentná analýza dát na FIIT STU 139
*Viera Rozinajová, Petra Vrablecová, Nadežda Andrejčíková, Anna Bou Ezzeddine
Mária Lucká, Miroslav Blšták, Marek Lóderer, Jaroslav Loeb*

**Adaptívne zložené klasifikačné modely pre detekciu udalostí
v prevádzke IT služieb** 143
Martin Sarnovský, Katarína Oravcová, Ján Marcinko

**Authority and Hub Mining from the Structure of a Social
Network** 149
Kristína Machová, Jozef Kováč

**Stratégie a zámery vyhľadávania v elektronických obchodoch
zameraných na zľavy: Prípadová štúdia** 155
Andrea Hrčková, Miriam Ondrišová, Jakub Šimko, Sandra Kostová

**Influence of the individual's size on the island model
architecture** 163
Štěpán Balcar

Knowledge-Driven Manufacturing 169
Petr Kadera, Petr Novák, Jiří Vyskočil, Václav Jirkovský

**Představení projektu „Moderní zpřístupnění historických
pramenů“** 173
Pavel Král, Karel Halla, Radek Široký, Ladislav Lenc, Jiří Martínek

Postery

**Využitie dostupných otvorených dát na podporu výberu
nehnutelnosti v Slovenskej republike** 177
Pavol Šatala, Tomáš Babič, Peter Butka

**Semiautomatizované porovnávanie certifikačných schém
cloudových služieb** 183
Martin Labaj, Karol Rástočný, Daniela Chudá

**Využitie kombinácie User experience a agilného prístupu
v procese vývoja webovej aplikácie** 187
Viera Pavlišková, František Babič, Ľudmila Pusztová

Behaviorálna biometrická autentifikácia na mobilných zariadeniach	193
<i>Kamil Burda, Daniela Chudá</i>	
Personalizované odporúčanie krátkodobých ponúk	199
<i>Peter Gašpar, Ivan Srba, Michal Barla, Michal Kompan, Mária Bielíková</i>	
Benchmark of LFS and other Item-based Product Recommendation Algorithms on Public Dataset	205
<i>Michal Laclavík, Marek Ciglan</i>	
Project Thoth – A Recommendation Engine for Machine Learning Applications	211
<i>Fridolín Pokorný, Christoph Görn</i>	
Autoencoders vs. others for anomaly detection	215
<i>Kristyna Míklášová, Ondřej Lomič, Lukáš Číсар, Luboš Popelínský, Veronika Krejčířová</i>	
FireProt: web server for automated design of thermostable proteins	221
<i>Miloš Musil</i>	
Inteligentná optimalizácia pokrytia a efektívnosti heterogénnych sietí využitím strojového učenia bez učiteľa	225
<i>Eugen Šlapak, Marcel Vološin, Juraj Gazda</i>	
Dynamic Social Networks, Their Analysis	231
<i>Iveta Mrázová, Peter Zvirinský</i>	
Siamese Convolutional Neural Networks for Recognizing Partial Entailment	237
<i>Martin Víta</i>	
Datová sada pro detekci dezinformačního obsahu – případová studie Novičok v Česku	243
<i>Martin Římnáč</i>	

Doktorandské příspěvky

Self-organizing neural networks for the analysis of country development	249
<i>Ildefonso Padrón Peña, Oto Mráz</i>	

Predikcia úpadku s.r.o. využitím metod pre rozpoznanie odľahlých bodov na nevyvážených dátach	255
<i>Peter Drotár, Martin Zorinčák, Peter Gnip</i>	
SoluProt: Prediction of Protein Solubility	261
<i>Jiří Hon, Martin Marušiak, Tomáš Martínek, Jaroslav Zendulka, David Bednář, Jiří Damborský</i>	
digIS: automated pipeline for detecting distant, novel insertion sequence elements in prokaryotes	267
<i>Janka Puterová, Tomáš Martínek</i>	
Cytometry analysis and its problems	273
<i>Tomáš Jarábek, Mária Lucká</i>	
Webová aplikace statistické nástroje podporující správný výběr a realizaci zvolené statistické metody	279
<i>Lubomír Štěpánek, Čestmír Štuka, Martin Vejražka</i>	
PhD Odyssey – Interweaving PhD stories into the Semantic Web	285
<i>Viet Bach Nguyen</i>	
Rejstřík autorů	289

Užití sloupcové in-memory databáze SAP HANA v podnikových informačních systémech

Radim Benek¹

¹AIS Financials Brno, SAP Labs Czech Republic
Holandská 6/8, 639 00 Brno
radim.benek@sap.com

Abstrakt. SAP HANA je dostupné řešení kombinující sloupcovou DB s in-memory přístupem, které integruje například prostorovou nebo prediktivní analýzu v zákaznickém systému nebo v cloudu. Přednáška popíše základy těchto technologií jednotlivě i to, jak jejich vzájemná integrace umožňuje realizovat řešení, která doposud nebyla technicky možná.

Klíčová slova: SAP HANA DB, in-memory, sloupcová.

1 SAP HANA databáze

1.1 Sloupcová in-memory databáze

Data v databázových strojích se vždy zpracovávají v procesorech. Úzkým hrdlem je načtení dat do CPU cache procesoru. V minulosti se data načítala z disku, kde je přístupová doba řádově 10ms. Zvětšování kapacity a snižování ceny RAM paměti v posledním desetiletí umožnilo využití RAM paměti jako uložistiště dat. Přístupová doba k datům se výrazně zkrátila na 100ns. Další urychlení zpracování dat přišlo s nástupem více jádrových procesorů a blade serveru, které podporují masivní paralelizaci zpracování dat.

Sloupcové uspořádání dat nabízí řadu výhod. Například kompresi dat pomocí slovníku hodnot s kompresním faktorem v řádu jednotek až desítek. Pro rychlé zpracování dat se používá přístup „insert only“, kdy se záznamy pouze vkládají. Záznamy určené ke smazání se označí jako neplatné. Záznamy, které se mají aktualizovat, používají kombinaci vložení nového a smazání původního záznamu. Navíc sloupcové uspořádání dat zjednodušuje použití paralelizace na úrovni jader procesoru, kde každé jádro zpracovává data z jednoho sloupce uložená v paměti na jednom místě.

Jedna z mála nevýhod sloupcového uspořádání dat je čtení dat pro všechny sloupce (celý řádek). V takovém případě může být sloupcové uspořádání o něco pomalejší než při řádkově uložených datech. Sloupcové databáze mají ale vynikající výsledky při čtení agregovaných dat (group-by) nebo při čtení nenaindexovaných sloupců [1].

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 17-20.*

1.2 Praktické příklady a testy

Při praktických testech na interním SAP databázovém systému o velikosti RAM 48 TB a 1100 CPU se ukázal velký kompresní faktor (50) sloupcového uspořádání tak i rychlost čtení (20x rychlejší oproti řádkově uspořádané tabulce).

Rychlost SAP HANA databáze lze proto využít v ERP systémech ke zrychlení zákaznických aplikací. Jedním z příkladu je CO-PA (Profitability Analysis). S použitím sloupcové uspořádání dat pro uložení 550 milionu záznamů bylo potřeba jen 30 GB paměti. Při řádkové uspořádání bylo potřeba 580 GB v nekomprimované nebo 140 GB v komprimované podobě na disku. Celkové zlepšení rychlosti „Period-End Closing“ aplikace (díky rychlejšímu čtení) bylo 2 % – 50 % podle typu použití. Pokud se ale aplikace optimalizovala na čtení ze sloupcové databáze, zlepšení bylo 80 %. Mnohem většího zlepšení se dosáhlo u „KE28 - Top Down Distribution“, kde se zkrátila celková doba zpracování z 28,6 h na 4,3 h.

1.3 Nástroje pro podporu běhu databáze

Aby in-memory databáze byla spolehlivá, implementuje SAP HANA obvyklé funkce pro protokolování (logging), uložení stavu (savepoints), zálohu a obnovu dat z disku. Data jsou tedy bez problému nahrána zpět do paměti po výpadku proudu nebo restartu.

Se stále vzrůstajícím objemem standardních dat, nestrukturovaných dat (audio, video) i nových typů dat (IoT), je potřeba kapacity databázového systému zvětšovat pomocí distribuovaných systémů. Kromě zvětšování kapacity paměti v rámci jednoho stroje (scale-up), lze v distribuovaných systémech mít další databázové stroje, které navzájem spolupracují (scale-out). Kromě toho lze databázový stroj nakonfigurovat jako multi-tenant, to znamená, že obsahuje více izolovaných databázových systémů. Pokud ani distribuované systémy nestačí pro uchování obrovských dat v operační paměti, je možné méně často využívané data (hot/cold data, např. finanční data starší 3 let) odložit na disk. V případě, že se s daty bude pracovat, načtou se opět do operační paměti.

2 SAP HANA Platform

Sloupcová in-memory databáze, která je velmi rychlá pro čtení dat, lze využít jako základ celé platformy dalších služeb.

2.1 Grafy

Efektivní zpracování grafových dat je požadováno různými aplikacemi (řízení dodavatelského řetězce, přeprava, analýza sociálních medií apod.). SAP HANA podporuje zpracování grafových dat přímo v databázovém stroji ve sloupcových tabulkách. Tím je práce s grafovým datovým modelem jednoduchá a velmi rychlá. Grafové struktury jsou v SAP HANA reprezentovány vrcholy (Zeus, Hera, ...), orientovanými hranami a vlastnostmi jak uzlů (bůh), tak hran (manželka, dcera). Tyto vlastnosti lze pak využít pro dotazování (nejkratší cesta, sousedé, ...).

2.2 Prostorová data

Prostorová data lze využít pro dotazy typu: „*Dej mi oblast s nejvyšším počtem zákazníků, kde je průměrná nejbližší vzdálenost k jakémukoli obchodu s potravinami větší než 20 km v oblasti vzdálené 2 km od dálnice*“. Implementace prostorových dat v SAP HANA odpovídá specifikaci Open Geospatial Consortium (OGC), která je podмноžinou ISO specifikace. SAP HANA podporuje prostorová data jako bod, čára, kruh, polygon a něco kolem 70 funkcí jako sjednocení, průnik, rozdíl, atd.

Rychlého zpracování prostorových dat lze využít například v letectví, kde dispečerů leteckých společností potřebují v reálném čase změnit trajektorii letu kvůli nenadálým událostem (např. změny počasí). Pomocí rychlého zpracování prostorových dat je možné trajektorii letu změnit více než 10krát rychleji.

2.3 Další služby

Kromě již zmíněných služeb nabízí SAP HANA platforma další služby:

- Datové řady (IoT)
- Block chain
- Statistické zpracování dat pomocí jazyka R
- Prediktivní analýzy.

Pro případ, že uvedené služby nestačí, může uživatel využít aplikační služby XS advanced. To je aplikační platforma, která podporuje několik programovacích jazyků jako JAVA nebo Node.js. V nich si uživatel vyvine a nechá běžet SAP HANA webové aplikace bez nutnosti instalovat další aplikační server.

2.4 SAP HANA Express Edition

Schopnosti SAP HANA databáze si může vyzkoušet každý na svém osobním počítači jako virtuální stroj, docker nebo binární Linux instalaci. Popřípadě je možnost ji vyzkoušet v cloudu, a to na Google Cloud Platform, Microsoft Azure Cloud nebo Amazon Web Services. Více informací lze nalézt na <http://sap.com/sap-hana-express>.

3 Závěr

Sloupcová in-memory databáze SAP HANA nabízí, kromě využití v podnikových informačních systémech, možnosti, které doposud nebyly myslitelné. Rychlost zpracování dat tak nabízí prostor pro zcela nové aplikace.

Literatura

1. Plattner, H.: A Course in In-Memory Data Management. Springer-Verlag, Heidelberg, 2013.

Annotation:

Usage of in-memory column-based SAP HANA databases in enterprise information systems of Data a znalosti & WIKT 2018

Recent hardware achievements and price reduction enabled to build database engines which store its data in main memory in the column-based form. With such aim, the performance is significantly better than traditional databases. Classical ERP applications benefit from its speed. But it also brings new possibilities for completely new applications which were not possible in the past.

Strojové učení v Seznam.cz

Vladimír Kadlec

Seznam.cz, a.s.
Londýnské náměstí 856/2, 639 00 Brno
vladimir.kadlec@firma.seznam.cz

Abstrakt. Každá služba, kterou Seznam.cz poskytuje, nějakým způsobem využívá strojové učení. V přednášce se dozvíte konkrétní příklady, jak nám v Seznamu strojové učení pomáhá a proč se bez něj už nedokážeme obejít.

Klíčová slova: Strojové učení · Vyhledávání · WWW · Seznam.cz

1 Abstrakt

Do vyhledávacího políčka na Seznam.cz uživatelé zadávají denně miliony dotazů a očekávají nalezení správné informace ukryté v miliardách webových stránek. Strojové učení je nástrojem, který nám umožňuje automaticky zpracovat obrovské množství dat a v jednotkách milisekund analyzovat, zda má danou informaci smysl uživateli zobrazit či nikoliv. Webová služba Mapy.cz v rámci projektu Panorama poskytuje fotografie z téměř každé ulice v každém městě a vesnici v České republice. Každý rok vzniknou řádově stovky terabajtů fotografií, na kterých musíme anonymizovat obličeje kolemjdoucích a registrační značky automobilů. Bez strojového učení by byla tato úloha jen velmi obtížně řešitelná. Nejen na dvou popsanych službách z portfolia Seznam.cz ukážeme, proč je pro nás strojové učení nepostradatelné a kde všude tento přístup používáme.

Aplikace ontologického modelu pro potřeby mezinárodního produktu nad heterogenními business daty v souladu s GDPR

Patrik Kompuš

Bisnode Česká republika
patrik.kompus@bisnode.com

Abstrakt. Společnost Bisnode se zabývá sběrem a zpracováním dat o firmách a osobách. Díky strategickému partnerství s Dun & Bradstreet má k dispozici informace o 250 milionech společností z 240 zemí a regionů světa. Na základě těchto dat a analýz Bisnode svým klientům, prostřednictvím různých produktů a služeb, poskytuje skóringové hodnocení firem, které predikuje pravděpodobnost úpadku v následujících dvanácti měsících, sleduje finanční zdraví a platební morálku firem, stejně tak nabízí efektivní nástroje na prověřování, ale i vyhledávání obchodních partnerů. V rámci transformace se Bisnode rozhodl sjednotit své produkty a služby, které nabízí v 18 zemích Evropy, a vytvořit jednu platformu. Velkou výzvou v tomto ambiciózním projektu je sjednocení pojmů a datových modelů pro zachování všech detailů specifických pro jednotlivé trhy. Dalším požadavkem je možnost sledovat tok dat od jednotlivých zdrojů až ke klientovi tak, abychom nabízeli služby v souladu s evropskou směrnicí GDPR. Přednáška bude zaměřena na aplikaci ontologického modelu pro potřeby mezinárodního produktu nad heterogenními daty o firmách a osobách z různých, veřejných i soukromých zdrojů, kde klíčovou roli hraje GDPR.

Klíčová slova: ontologie, GDPR, skóringové hodnocení firem, data o firmách

Kybernetická bezpečnost a GDPR

Data a znalosti & WIKT 2018

Jan Kolouch

CESNET, z.s.p.o.
Zikova 1903/4, 160 00 Praha 6
jan.kolouch@cesnet.cz

Abstrakt. Článek je zaměřen na některé aspekty kybernetické bezpečnosti v souvislosti účinným s obecným nařízením o ochraně osobních údajů (GDPR). Pozornost je věnována zejména vymezení místní působnosti tohoto nařízení, definování osobního údaje a jeho vztahu k informačním a komunikačním technologiím. V neposlední řadě se článek věnuje i zpracování a zabezpečení osobních údajů.

Klíčová slova: GDPR, osobní údaj, IP adresa, bezpečnost

1 Úvod

Obecné nařízení o ochraně osobních údajů (EU) 2016/679 (anglicky: General Data Protection Regulation neboli GDPR)¹ je jedním z významných mezinárodních právních dokumentů, který s problematikou kybernetické bezpečnosti bezprostředně souvisí, byť není primárně cílen do oblasti ICT.

„GDPR ≠ IT + software.

Nové nařízení o ochraně osobních údajů má 778 řádků a z toho jen 26 se přímo týká IT bezpečnosti. Máte představu, co obsahují ty ostatní?“

Mgr. Eva Škorníčková²

Právě na GDPR a implementaci povinností z tohoto nařízení vyplývajících je možné demonstrovat tu skutečnost, že je vhodné řešit komplexně problematiku bezpečnosti, a ne uměle izolovat povinnosti vyplývající z různých právních norem (v tomto případě ze zákona o kybernetické bezpečnosti a GDPR).

¹ Dále jen GDPR. [online]. Dostupné z: <http://eur-lex.europa.eu/legal-content/CS/TXT/HTML/?uri=CELEX:32016R0679&qid=1488972453767&from=CS>

² ŠKORNIČKOVÁ, Eva. *Jednoduchý test: Jak jste na tom s přípravou na GDPR?* [online]. [cit. 10. 11. 2017]. Dostupné z: <https://www.gdpr.cz/blog/jednoduchy-test-jak-jste-na-tom-s-pripravou-na-gdpr/>

Cílem článku není provést samostatný a komplexní rozbor problematiky GDPR. Na tomto místě budou definovány pouze dílčí pojmy a práva a povinnosti, které z GDPR vyplývají, a zároveň mají přesah do oblasti kybernetické bezpečnosti.

Nařízení GDPR představuje **obecný právní rámec ochrany osobních údajů** platný a účinný na celém území EU, a v určitých případech i mimo toto teritorium. Hlavním cílem GDPR je zajistit komplexní ochranu práv subjektů údajů proti neoprávněnému zacházení s jejich daty a osobními údaji; nastolit rovnováhu mezi oprávněnými zájmy správců, zpracovatelů a subjektů údajů; vytvořit systém jednotné vymahatelnosti práva a jednotného sankčního mechanismu v této oblasti atd.

„Ač ochrana osobních údajů u nás platí od roku 1992, ač úřad zřízený zákonem o ochraně osobních údajů kontroluje povinnosti tímto zákonem uložené už téměř dvě desetiletí, pro mnohé, kteří si se zákonem starosti nedělali, jako by šlo o novou věc.“³

Listina základních práv a svobod v čl. 10 odst. 2 a 3 uvádí:

„Každý má právo na ochranu před neoprávněným zasahováním do soukromého a rodinného života.“

„Každý má právo na ochranu před neoprávněným shromažďováním, zveřejňováním nebo jiným zneužíváním údajů o své osobě.“

Další ochranu osobních údajů zajišťuje zákon č. 101/2000 Sb., o ochraně osobních údajů či například zákon č. 89/2012 Sb., občanský zákoník, zákon č. 372/2011 Sb., o zdravotních službách a podmínkách jejich poskytování (zákon o zdravotních službách), zákon č. 85/1996 Sb., o advokacii, zákon č. 127/2005 Sb., o elektronických komunikacích, ve znění pozdějších předpisů, aj.

Bylo by tedy teoreticky možné konstatovat, že ochrana osobních údajů, jakož i osob, kterých se tyto údaje týkají, zde existovala a existuje. Nicméně rychlý technologický rozvoj a globalizace (zejména dat a služeb v oblasti ICT) s sebou nutně přinesly nové výzvy i pro oblast ochrany osobních údajů.

Rozsah shromažďování a sdílení osobních údajů právě díky informačním a komunikačním technologiím a službám, které jsou na ně navázány, významně vzrostl. Informační a komunikační technologie umožňují jak soukromým společnostem, tak orgánům veřejné moci využívat při provádění jejich činností osobní údaje v nebývalém rozsahu. Na druhou stranu je také možné pozorovat masivní dobrovolné zveřejňování osobních údajů samotnými fyzickými osobami, jichž se tyto údaje týkají.

Informační a komunikační technologie výrazně změnily ekonomiku i společenský život a měly by usnadňovat volný pohyb osobních údajů v rámci Evropské Unie a předávání těchto údajů do třetích zemí a mezinárodním organizacím. Současně by však tyto technologie a procesy s nimi spojené měly zajistit vysokou úroveň ochrany osobních údajů.⁴

³ *GDPR stručně.* [online]. [cit. 7.8.2018]. Dostupné z: <https://www.uouu.cz/gdpr%2Dstrucne/ds-4843/archiv=0&p1=3938>

⁴ Srov. recitál 6 GDPR

Díky výše uvedenému však **vzniká zajímavý paradox**, který spočívá v následujících bodech:

- **fyzické osoby samy a dobrovolně o sobě zveřejňují stále větší množství dat** (fotografie, videa aj.), přičemž k distribuci těchto dat typicky využívají služby informační společnosti, které jsou založeny na EULA⁵ či SLA⁶ mezi uživatelem a poskytovatelem služby
- **nejvíce jsou osobní údaje zveřejňovány v rámci sociálních sítí**, které z podstaty své funkce takovéto zveřejňování předpokládají a zakotvují ve smluvních podmínkách pravidla, na základě kterých je s takovými daty zacházeno
- **fyzické osoby při využívání řady služeb informační společnosti předpokládají a mnohdy i očekávají interakci mezi těmito technologiemi a jejich kyberosobností⁷**
- mezinárodní společenství, stát, ale i **fyzické osoby samy vyžadují vyšší zabezpečení osobních údajů a znemožnění přístupu k těmto údajům jiným** (zpravidla neoprávněným) subjektům, a to vše za podmínky zachování existence prvních tří bodů tohoto paradoxu.

Důsledek tohoto paradoxu je zřejmý. Poskytovatelé služeb informační společnosti⁸ tak musí věnovat vyšší úsilí zabezpečení jednotlivých služeb, které koncovému uživateli poskytují, vyšší úrovni zabezpečení dat vztahujících se k uživateli, modifikaci stávajících smluvních podmínek a zavedení dalších požadavků vyplývajících z GDPR.

⁵ **EULA (End Users Licence Agreement)** je označení pro smluvní podmínky, umožňující využití služby daného poskytovatele služby. EULA představuje smlouvu, která je zpravidla jednostranně vymezena poskytovatelem služby. Uživatel však není nikterak omezován na svých právech, neboť má možnost volby v podobě nevyužití takto jednostranně stanovených smluvních podmínek. V případě souhlasu s využíváním takovýchto služeb je možné obecně konstatovat, že dojde primárně k aplikaci soukromoprávních norem.

Otázkou je, zda si uživatel skutečně uvědomuje, jaké smluvní podmínky odsouhlasil, kdy se pro něj stávají závaznými a jaký možný (legální) zásah do jeho základních lidských práv a svobod takto vyslovený souhlas představuje. Další neopomenutelnou skutečností pak je, že takto poskytovaná služba může ovlivnit práva a oprávněné zájmy (např. bezpečnost IT, důvěryhodnost dat aj.) třetích osob (např. zaměstnavatele aj.), které k využívání předmětné služby explicitně souhlas nevyjádřily.

Smutným faktem zůstává ta skutečnost, že velmi malé procento uživatelů je ochotno číst smluvní podmínky, vztahující se k té které poskytované službě.

⁶ **SLA (Service-level agreement)** označuje smlouvu sjednanou mezi poskytovatelem služby a jejím uživatelem.

⁷ „Kyberosobnost“ zahrnuje identifikaci osoby na síti, jako je e-mailová adresa, IP adresa, číslo telefonu a další.. **Tuto interakci je možné sledovat při využívání polohových a geolokačních služeb** (např. Google maps, Waze, Seznam mapy aj.), neboť fyzická osoba předpokládá, že ji bude počítačový systém schopen lokalizovat a zobrazit jí nejvýhodnější cestu. Stejně tak je ona interakce očekávána např. **u služeb umožňujících prodej a nákup zboží** (např. Letgo – viz doporučené inzeráty dle geolokace či již nakoupeného zboží), **restauračních a ubytovacích službách** (např. Tripadvisor, Booking.com, Airbnb aj.) aj.

⁸ Blíže viz KOLOUCH, Jan. *CyberCrime*. Praha: CZ.NIC, 2016, s. 78 a násl. a s. 109 a násl.

2 Místní působnost GDPR

Někoho by mohlo napadnout, že způsobem, jak se vyhnout GDPR, by bylo přesunout se mimo jeho dosah, tedy mimo teritorium EU. Nařízení GDPR se však uplatní v případech, kdy je:

- **provozovna správce nebo zpracovatele v EU**, bez ohledu na to, zda zpracování probíhá v EU
- **správci nebo zpracovatelé neusazení v EU**
 - zboží nebo služby jsou nabízeny subjektům údajů v EU (bez ohledu na úplatu)
 - monitorováno chování subjektů údajů v rámci EU⁹

Díky takto vymezené místní působnosti má GDPR exteritoriální dosah a de facto se bude vztahovat na všechny služby informační společnosti, ke kterým lze získat přístup z geografického teritoria EU, nebo které monitorují chování subjektů údajů v rámci EU.

3 Osobní údaj

Osobním údajem dle čl. 4 odst. 1 GDPR jsou „**veškeré informace o identifikované nebo identifikovatelné fyzické osobě**. Identifikovatelnou fyzickou osobou je fyzická osoba, kterou lze přímo či nepřímo identifikovat zejména odkaz na určitý identifikátor, například jméno, identifikační číslo, lokační údaje, síťový identifikátor nebo na jeden či více zvláštních prvků fyzické, fyziologické, genetické, psychické, ekonomické, kulturní nebo společenské identity této fyzické osoby.“

Z vlastního znění výše uvedeného článku, ve srovnání s ustanovením § 4 zákona č. 101/2000 Sb., o ochraně osobních údajů vyplývá, že došlo de facto jen ke stylistickým, nikoliv však obsahovým změnám. Nicméně to, co **GDPR** zásadně **změnilo**, je **rozšíření okruhu dat a informací, které je možné považovat za osobní údaje**.

Osobním údajem je dle GDPR **jakákoliv informace** (např. obrazové, písemné, slovní, digitální, genetické, zdravotnické aj.), která **má vztah** (obsahem – např. jméno, adresa, pracovní zařazení, email aj.), **k subjektu údajů**.¹⁰

Z tohoto pohledu a v souladu s výkladem uvedeným v recitálech 30, 34, 35, 38 GDPR¹¹ je třeba za osobní údaj považovat: jméno a příjmení, **identifikační číslo, rodné číslo, lokační údaje (geo-)**, věk a datum narození, pohlaví, osobní stav, občanství, **síťové identifikátory (IP adresa, identifikátory cookies, radio frequency identification tags aj.)**, **fotografie, prvky** fyzické, fyziologické, genetické, psychické,

⁹ Viz Čl. 3 GDPR – Místní působnost

¹⁰ **Subjektem údajů je** dle čl. 4 odst. 1 GDPR identifikovaná nebo identifikovatelná **fyzická osoba**. **Subjekt může být identifikován:**

- přímo
- nepřímo (např. výběr vyčleněním aj.)

¹¹ Recitály jsou ustanovení předcházející vlastnímu textu nařízení GDPR a jsou v některých případech výkladem či do jisté míry důvodovou zprávou k vlastnímu textu nařízení.

ekonomické, **kulturní nebo společenské identity**, osobní či pracovní adresa, osobní či pracovní telefonní číslo, **osobní či pracovní email, ověřovací identifikační údaje**, identifikační čísla vydaná státem.

Tučně vyznačené osobní údaje mají typicky vztah k informačním a komunikačním technologiím a aplikacím, které tyto technologie využívají. Rozšíření okruhu dat, jež je možné považovat za osobní údaje, výrazným způsobem zasahuje do problematiky kybernetické bezpečnosti a zajištění ochrany dat, která jsou spravována v dané organizaci.

Pokud se zaměříme na **položku síťových identifikátorů a ověřovacích identifikačních údajů**, zjistíme, že za osobní údaj může a zřejmě i bude považována **řada dat umožňujících základní fungování počítačového systému v síti**.

Velmi často byla v praxi diskutována otázka - je IP adresa osobním údajem?

V této věci je vhodné kromě GDPR přihlídnout i k judikatuře Soudního dvora EU, který mimo jiné rozhodoval v kauze: **Patrick Breyer proti Bundesrepublik Deutschland**.¹²

Patrick Bayer se u německých soudů domáhal, aby Německo přestalo uchovávat jeho IP adresy, které získalo při jeho „návštěvách“ několika internetových stránek německých spolkových orgánů, které byly veřejně přístupné. Z pohledu činnosti provozovatelů dotčených webových stránek se jednalo se o klasické logování služeb tímto ISP nabízených.

Německé soudy přerušily řízení a položily předběžnou otázku soudnímu dvoru EU, protože v dané věci neexistuje jednotný výklad práva EU. Jde zejména o to, jestli k tomu, aby nějaký údaj byl osobním údajem, a tedy identifikoval konkrétní osobu, je třeba vycházet z „objektivního“, či „relativního“ kritéria. **„Objektivní“ kritérium znamená, že údaje, jako jsou IP adresy, by mohly být považovány za osobní údaje zpracovávané ISP jiných služeb než připojení (např. provozovatelem internetové stránky), a to i tehdy, pokud by byla schopna identifikovat konkrétního uživatele jen třetí osoba (typicky ISP připojení). „Relativní“ kritérium znamená, že IP adresy by mohly být považovány za osobní údaj u ISP připojení, neboť mu umožňují přesně určit totožnost uživatele, ale už ne u ISP služeb, který disponuje skutečně pouze údajem o IP adrese a nezná jméno návštěvníka.**

Soudní dvůr EU konstatoval, že je nesporné, že **dynamická IP adresa nepředstavuje informaci o „identifikované osobě“**, neboť adresa přímo neodhaluje **totožnost fyzické osoby**, která je majitelem počítače, ze kterého byla navštívena internetová stránka, ani **totožnost jiné osoby, která mohla tento počítač používat**.

Na druhou stranu však Soudní dvůr (druhý senát) také uvedl (následně i rozhodl), že **dynamická adresa internetového protokolu, kterou poskytovatel on-line mediálních služeb uchovává v souvislosti s přístupem osoby na internetovou stránku,**

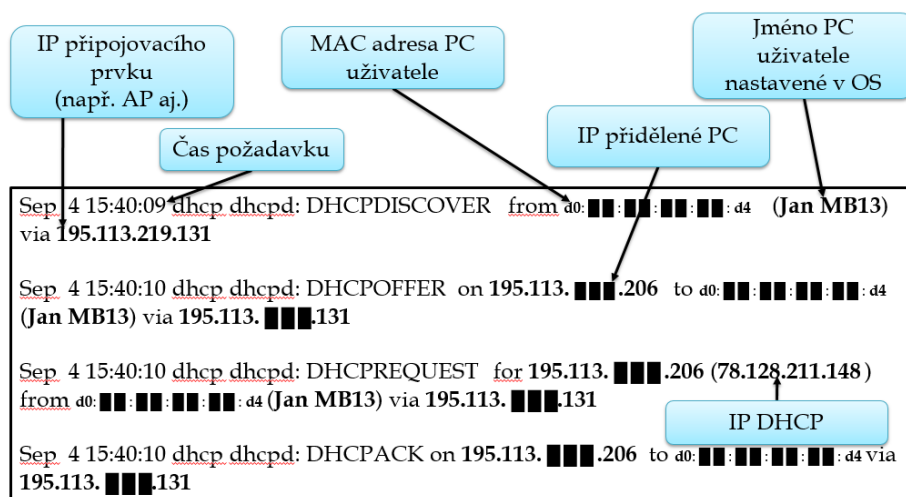
¹² Blíže viz: [online]. Dostupné z: <http://curia.europa.eu/juris/document/document.jsf?text=&docid=184668&pageIndex=0&doclang=cs&mode=lst&dir=&occ=first&part=1&cid=1403270>

kteřou tento poskytovatel zpřístupnil veřejnosti, pro uvedeného poskytovatele **představuje osobní údaj ve smyslu článku 2 písm. a) směrnice Evropského parlamentu a Rady 95/46/ES ze dne 24. října 1995 o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů, pokud má poskytovatel k dispozici právní prostředky, které mu umožňují nechat identifikovat subjekt údajů díky dalším informacím, kterými disponuje poskytovatel internetového připojení tohoto subjektu.**

Dynamická IP adresa může být dle tohoto rozsudku, z 19. října 2016, za určitých okolností osobním údajem.

Dopad té skutečnosti, že **IP adresa, jakožto i další síťové identifikátory mohou být osobním údajem**, demonstrují na dvou příkladech.

Na následujícím obrázku je možné vidět komunikaci PC a jednotlivých prvků sítě (AP, DHCP server) a následné připojení PC do sítě.



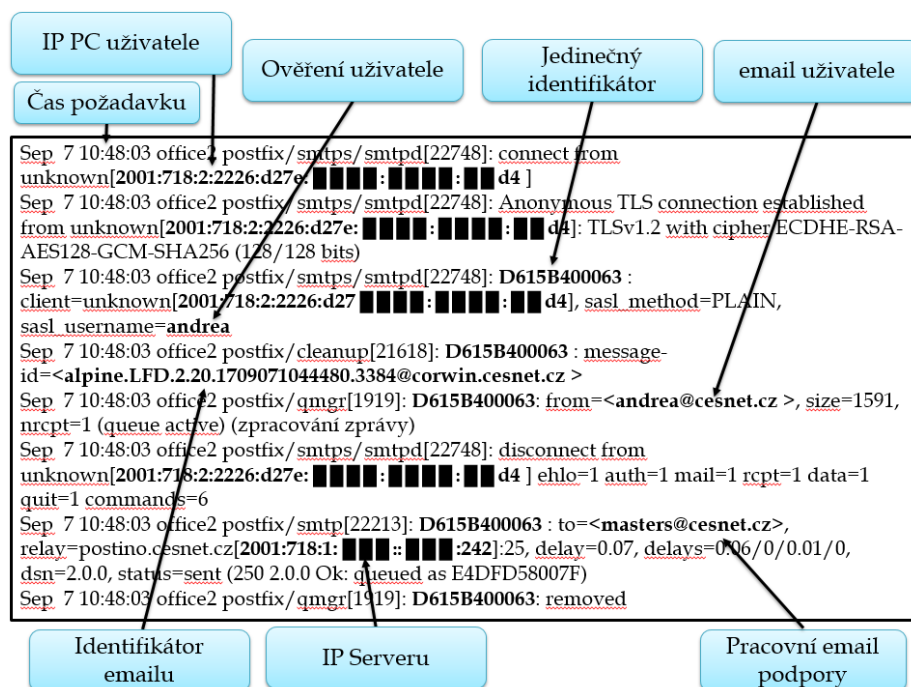
Obrázek 1 - DHCP

Pokud se důsledně zaměříme na **data** (informace), **která mají vztah k subjektu údajů a jsou jej schopny identifikovat**, pak osobním údajem v tomto případě nebude pouze IP adresa připojovacího prvku a IP adresa DHCP serveru.

Osobním údajem je i teoreticky čas daného požadavku, neboť se jedná o stopu, která může být zejména v kombinaci s jedinečnými identifikátory a dalšími informacemi, které servery získávají, použita k identifikaci fyzické osoby.¹³ Zároveň se jedná o velmi podstatnou informaci, neboť bez přesného času není možné identifikovat komu (jakému počítačovému systému) byla přidělena konkrétní IP adresa.

Dalším příkladem zobrazujícím rozsah zpracovávání dat, která je možné považovat za osobní údaje, je zpracování osobních údajů při odeslání emailu skrze SMTP.

¹³ Blíže viz recitál 30 GDPR



Obrázek 2 - SMTP

Pokud se opět důsledně zaměříme na **data** (informace), která mají vztah k subjektu údajů a jsou jej schopna identifikovat, pak osobním údajem v tomto případě nebude pouze IP adresa serveru.

Pracovní email podpory by mohl být opět osobním údajem, pokud k němu budou přiřazeny další identifikátory, které jsou schopné identifikovat fyzickou osobu.

Klíčovou otázkou je, zda jsem v rámci veškerých procesů, které se odehrávají v počítačových systémech (prvcích ICT), které jsou daným subjektem (fyzická či právnická osoba) spravovány, **schopen rozlišit situaci, kdy dochází k přenosu dat čistě mezi počítačovými systémy, bez vztahu k jakékoliv fyzické osobě, a kdy už se do těchto procesů zapojí fyzická osoba, jakožto subjekt údajů dle GDPR.**

Osobně se domnívám, že až na specifické výjimky nebudu schopen vyčlenit procesy, které se odehrávají bez lidské interakce. Na základě tohoto tvrzení pak budu nucen aplikovat požadavky vyplývající z GDPR na veškeré procesy, při nichž dochází k manipulaci s informacemi, které mají vztah k subjektu údajů a jsou jej schopny identifikovat. Zároveň bude třeba přijmout dostatečná bezpečnostní opatření, aby byla dostatečně chráněna jak přenosová soustava, počítačové systémy a aplikace, které s takovýmito informacemi pracují, tak informace (resp. data) samotná.

Vedle shora uvedených osobních údajů definuje GDPR zvláštní kategorie osobních údajů, mezi které patří údaje o rasovém či etnickém původu, vyznání, politických názorech, členství v odborech či jiných organizacích, sexuální orientaci, spáchání

deliktů (trestný čin/přestupek aj.) a potrestání za ně, genetické údaje (DNA & RNA), biometrické údaje, údaje o zdravotním stavu.

4 Zpracování osobních údajů

Zpracováním osobních údajů se dle čl. 4 odst. 2 GDPR rozumí **jakákoliv operace** nebo soubor operací **s osobními údaji** nebo soubory osobních údajů, **který je prováděn pomocí či bez pomoci automatizovaných postupů**, jako je shromáždění, zaznamenání, uspořádání, strukturování, uložení, přizpůsobení nebo pozměnění, vyhledání, nahlédnutí, použití, zpřístupnění přenosem, šíření nebo jakékoliv jiné zpřístupnění, seřazení či zkombinování, omezení, výmaz nebo zničení.

Ochrana subjektu údajů se vztahuje na zpracování osobních údajů, pokud jsou tyto údaje uloženy v evidenci nebo do ní mají být vloženy.¹⁴

Pojem **zpracování** dle GDPR však **nelze chápat jako jakékoli nakládání s osobním údajem. Zpracování osobních údajů je nutné považovat již za sofistikovanější činnost, kterou správce s osobními údaji provádí za určitým účelem a z určitého pohledu tak činí systematicky.**¹⁵

Ze zpracování osobních údajů dle GDPR je mimo jiné **vyňata činnost prováděná fyzickou osobou v rámci čistě osobní povahy nebo činnosti prováděné výhradně v domácnosti, a tedy bez jakékoliv souvislosti s profesní nebo obchodní činností.**¹⁶

V čl. 5 odst. 1 písm. a) GDPR jsou stanoveny zásady zpracování osobních údajů. Mezi tyto zásady dle GDPR patří:

- **zákonnost, korektnost, transparentnost** [čl. 5 odst. 1 písm. a) GDPR] – správce osobních údajů je povinen:
 - informovat subjekt údajů o probíhající operaci zpracování a jejích účelech
 - informovat subjekt údajů o profilování a o jeho důsledcích
 - informovat subjekt údajů, pokud jsou osobní údaje získávány od něj, zda je povinen tyto údaje poskytnout, a o důsledcích jejich případného neposkytnutí
 - **prokázat existenci nejméně jednoho právního důvodu pro zpracování osobních údajů**
 - **dokumentovat** (co, jak, proč zpracovává; souhlas a zákonný důvod; čas, po který zpracovává; **přijaté záruky a bezpečnostní opatření**)
- **omezení účelu** [čl. 5 odst. 1 písm. b) GDPR] – osobní údaje musí být shromažďovány pro určité, výslovně vyjádřené a legitimní účely a nesmějí být dále zpracovávány způsobem, který je s těmito účely neslučitelný,
- **minimalizace údajů** [čl. 5 odst. 1 písm. c) GDPR] – osobní údaje musí být přiměřené a relevantní ve vztahu k účelu, pro který jsou zpracovávány,
- **přesnost** [čl. 5 odst. 1 písm. d) GDPR] – osobní údaje musí být přesné a v případě potřeby aktualizované; musí být přijata veškerá rozumná opatření, aby osobní údaje

¹⁴ Viz recitál 15 GDPR

¹⁵ Blíže viz *Základní příručka k GDPR*. [online]. [cit. 7.8.2018]. Dostupné z: <https://www.uoou.cz/zakladni%2Dprirucka%2Dk%2Dgdpr/ds-4744/archiv=0&p1=3938>

¹⁶ Viz recitál 15 GDPR

je, které jsou nepřesné s přihlédnutím k účelům, pro které se zpracovávají, byly bezodkladně vymazány nebo opraveny,

- **omezení uložení** [čl. 5 odst. 1 písm. e) GDPR] – osobní údaje by měly být uloženy ve formě umožňující identifikaci subjektu údajů jen po nezbytnou dobu pro dané účely, pro které jsou zpracovávány,
- **integrita a důvěrnost** [čl. 5 odst. 1 písm. f) GDPR] – osobní údaje musí být **zpracovávány způsobem, který zajistí náležité zabezpečení osobních údajů, včetně jejich ochrany pomocí vhodných technických nebo organizačních opatření před neoprávněným či protiprávním zpracováním a před náhodnou ztrátou, zničením nebo poškozením technické a organizační zabezpečení osobních údajů.**

5 Zabezpečení osobních údajů

Jednou z oblastí, které se GDPR explicitně věnuje, je i **problematika zabezpečení zpracování osobních údajů**.

V čl. 32 GDPR je uvedeno, že **správce** (případně zpracovatel) **musí** s přihlédnutím ke stavu techniky, nákladům na provedení, povaze, rozsahu, kontextu a účelům zpracování i k různě pravděpodobným a různě závažným rizikům pro práva a svobody fyzických osob, **přijmout vhodná technická a organizační opatření, aby zajistili úroveň zabezpečení odpovídající danému riziku**, případně včetně:

- pseudonymizace a šifrování osobních údajů
- schopnosti zajistit neustálou důvěrnost, integritu, dostupnost a odolnost systémů a služeb zpracování
- schopnosti obnovit dostupnost osobních údajů a přístup k nim včas v případě fyzických či technických incidentů
- procesu pravidelného testování, posuzování a hodnocení účinnosti zavedených technických a organizačních opatření pro zajištění bezpečnosti zpracování

„Při posuzování vhodné úrovně bezpečnosti se zohlední zejména rizika, která představuje zpracování, zejména náhodné nebo protiprávní zničení, ztráta, pozměňování, neoprávněné zpřístupnění předávaných, uložených nebo jinak zpracovávaných osobních údajů, nebo neoprávněný přístup k nim.“¹⁷

Při určování rizika je třeba vycházet zejména z kategorie osobních údajů, které by mohly být porušením zabezpečení dotčeny, charakteru porušení zabezpečení a počtem dotčených subjektů údajů. Vyšší riziko představují „citlivější“ osobní údaje (viz např. zvláštní kategorie osobních údajů), rozsáhlejší soubor osobních údajů, případně údaje, jimiž lze způsobit subjektu údajů újmu či zásah do jeho práv.

Dle čl. 32 odst. 4 GDPR přijmou správce a zpracovatel opatření pro zajištění toho, aby jakákoliv fyzická osoba která jedná z pověření správce nebo zpracovatele a má přístup k osobním údajům, zpracovávala tyto osobní údaje pouze na pokyn správce, pokud jí jejich zpracování již neukládá právo Unie nebo členského státu.

¹⁷ Čl. 32 odst. 2 GDPR

6 Závěr

Vlastní nařízení GDPR obsahuje i další instituty (např. pseudonymizace, požadavky na výmaz či přenositelnost osobních údajů aj.), které se mohou vztahovat k činnosti, jež je prováděna v rámci informačních a komunikačních systémů, a které vyžadují náležitou úroveň zabezpečení a ochrany.

Podstatné je identifikovat vliv (dopad) GDPR na organizaci, na její jednotlivé části a procesy. De facto jde o provedení auditu, kde všude v organizaci, případně u jednotlivce se pracuje s osobními údaji ve vztahu ke GDPR. Následně postup spočívá v modifikaci či tvorbě pravidel a procesů (pokud je to třeba) jak uvnitř organizace, tak ve vztahu k subjektu údajů. Veškerá tato činnost by současně měla respektovat základní principy bezpečnosti.

Stejně jako při zavádění bezpečnostních pravidel obecně, tak při implementaci GDPR či jiných dokumentů a doporučení je třeba si uvědomit, že neexistuje jedno pravidlo, vzor, nástroj, řešení či postup aplikovatelný pro každou organizaci a každou situaci či každou organizaci.

Je třeba přijmout a implementovat vlastní řešení v souladu s GDPR.

Je třeba individualizovat...

Literatura

Reference na knihy:

1. KOLOUCH, Jan. *CyberCrime*. Praha: CZ.NIC, 2016. ISBN 978-80-88168-15-7

Reference internetové zdroje:

2. Nařízení Evropského Parlamentu a Rady (EU) 2016/679, ze dne 27. dubna 2016, o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES. [online]. Dostupné z: <http://eur-lex.europa.eu/legal-content/CS/TXT/HTML/?uri=CELEX:32016R0679&qid=1488972453767&from=CS>
3. ŠKORNIČKOVÁ, Eva. *Jednoduchý test: Jak jste na tom s přípravou na GDPR?* [online]. [cit. 10. 11. 2017]. Dostupné z: <https://www.gdpr.cz/blog/jednoduchy-test-jak-jste-na-tom-s-pripravou-na-gdpr/>
4. *GDPR stručně*. [online]. [cit. 7.8.2018]. Dostupné z: <https://www.uoou.cz/gdpr%2Dstrucne/ds-4843/archiv=0&p1=3938>

Annotation:

Cyber security and GDPR

Data a znalosti & WIKT 2018

The article focuses on some aspects of cyber security in the context of the General Data Protection Regulation (GDPR). Attention is paid in particular to the definition of the local scope of

this Regulation, the definition of personal data and its relation to information and communication technologies. Last but not least, the article deals with the processing and security of personal data.

Přístupy ke správě dat v prostředí velké organizace

Dušan Chlapek

Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3, Česká republika
chlapek@vse.cz

Abstrakt. Ve zvané přednášce budou popsány teoretické i praktické přístupy k vymezení oblasti správy dat jako samostatné oblasti řízení s vlastní strategií a definovanými procesy, rolemi, odpovědnostmi a samostatnými nástroji. Současně bude diskutován vztah procesů a rolí zaměřených na správu dat ve vztahu k informatickým procesům a rolím.

Klíčová slova: správa dat, procesy, role, architektura, řízení IT.

1 Úvod

V posledních letech se většina podniků i institucí více či méně zabývá daty zejména pod vlivem nařízení [1] o ochranně fyzických osoba v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů (zkráceně GDPR) a případně pod vlivem dalšího připravovaného nařízení [2] o respektování soukromého života a ochraně osobních údajů v elektronických komunikacích. A přestože jsou data jedním z nejcennějších aktiv, tak v málokterém podniku a instituci nalezneme data dostatečně dokumentovaná, natož aby byla správa dat rozlišována jako samostatně řízená oblast, která má jasné vymezené procesy, role, odpovědnosti a standardy.

Již několik let se má při záměru budovat informační systém v ČR v oblasti veřejné správy [3], [4] nejprve vytvořit podniková architektura. V roce 2016 byl představen Národní architektonický plán, který je dle [5] chápán jako „*soubor architektonických dat (modelů) a diagramů, udržovaných společně odborem Hlavního architekta eGovernmentu na Ministerstvu vnitra a jednotlivými orgány veřejné moci*“ a v metodickém dokumentu [6] byly definovány pohledy na architekturu pro vyplnění formuláře, kterými orgány veřejné moci žádají o souhlas u odboru Hlavního architekta eGovernmentu. Tato metodika [6] definuje následující pohledy:

- na úrovni byznys architektury pohledy činnostních funkcí a služeb veřejné správy,
- na úrovni aplikační architektury pohledy struktury aplikací a komunikace aplikací,
- na úrovni technologické / platformové architektury pohled struktury IT technologické / platformové architektury,
- na úrovni komunikační / síťové architektury pohled struktury komunikační / síťové architektury,
- pohled služeb v čtyřvrstvé vizi architektury veřejné správy.

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 37-41.*

Podíváme-li se na uvedené pohledy používané k modelování podnikové architektury, vidíme, že mezi uvedenými pohledy chybí pohled zaměřený na popis datové architektury. To jenom dále podtrhuje skutečnost, že přestože jsou data klíčovým aktivem, tak i při zvažování záměrů na budování informačních systémů ve veřejné správě ČR není zatím vyžadováno se zabývat jejich systematickým popisem a modelováním na úrovni podnikové architektury.

2 Správa dat, datová architektura a standardy

Mezi hlavní klíčové standardy v oblasti správy dat jsou řazeny zejména publikace The DAMA Guide to the Data Management Body of Knowledge [7] a Queensland Government Information Management Policy Framework Definitions [9].

Správa dat je v souladu s [7] chápána jako tvorba a provádění politik, projektů, procesů, praktik či procedur, prostřednictvím kterých jsou zajištěny řízení, ochrana, dodávka a zvyšování hodnoty dat a informačních aktiv.

Některé standardy (například [8]) rozlišují a popisují informační cyklus, kdy byznys proces (podporovaný informatickými procesy a službami) vytváří a zpracovává data, která jsou následně transformována na informace a tyto jsou dále transformovány na znalosti, které pak vytváří hodnotu, kterou využívá opět byznys proces. Ale v tomto příspěvku se budeme věnovat pouze správě dat.

Správa dat obsahuje (upraveno dle [7]) následující oblasti:

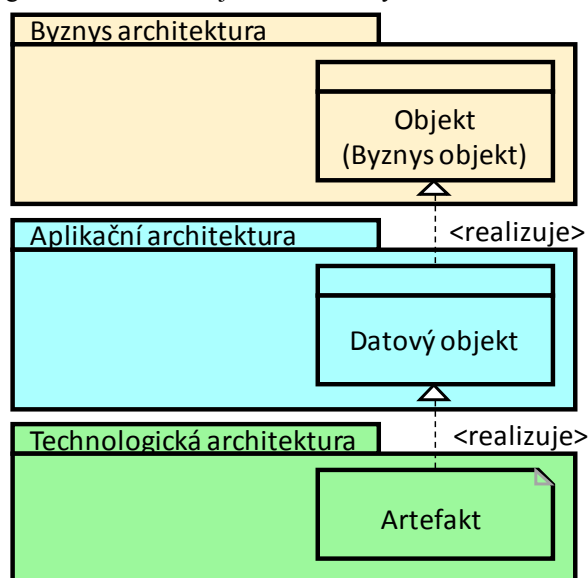
- strategické řízení,
- řízení datové architektury,
- pořizování, provoz a údržba dat, poskytování dat, skartace dat,
- řízení bezpečnosti dat,
- řízení referenčních dat a master dat,
- datové sklady a business intelligence,
- správa dokumentů a obsahu,
- řízení nákladů a přínosů dat,
- řízení kvality dat,
- řízení metadat.

Architektura v podnikové informatice (také podniková architektura) je dle [10] „*chápána jako přístup, koncept, prostředek a nástroj, kterým vyjadřujeme fundamentální uspořádání vztahu mezi byznysem a jeho informačním systémem, které vede k naplnění mise organizace, přičemž respektuje okolní prostředí a konzistentně dodržuje formulované principy návrhu a rozvoje systému*“.

Datová architektura je jednou z dílčích architektur podnikové architektury [11] a měla by obsahovat:

- datový model – na konceptuální úrovni,
- zachycení vazby na další úrovně modelů (logický a fyzický),
- zachycení vazby prvků datové architektury na prvky dalších architektur (byznys procesy, služby, aplikace, technologie, ...).

Prvky datové architektury v notaci ArchiMate dle [6] použité ve vazbě na byznys, aplikační a technologickou architekturu jsou znázorněny na obrázku č. 1.



Obr. 1. Prvky datové architektury ve vazbě na další architektury

3 Implementace standardů v oblasti správy dat

Popisovaný postup zavedení systematického přístupu k řízení správy dat byl realizován v prostředí velké organizace v oblasti veřejné správy. Organizačně již před zahájením prací bylo odděleno řízení informatiky od řízení správy dat. Nebyly však definovány procesy, jejich výstupy a nebyly vymezeny práva a odpovědnosti jednotlivých útvarů.

Postup definování procesů a odpovědností byl rozdělen do následujících kroků:

- 1) Stanovení cílů projektu
- 2) Analýza stávajících řídicích dokumentů a dostupných standardů
- 3) Návrh a přizpůsobení oblastí řízení správy dat. Implementovány byly následující oblasti správy dat:
 - a) Strategické řízení správy dat
 - b) Řízení datové architektury
 - c) Tvorba a údržba metodiky správy dat
 - d) Pořizování, provoz, údržba a poskytování dat
 - e) Řízení kmenových dat a číselníků
 - f) Řízení metadat
 - g) Řízení kvality dat
 - h) Řízení bezpečnosti informací

Pro každou uvedenou oblast v rámci řízení správy dat byly stanoveny cíle oblasti a byla provedena dekompozice na podoblasti a skupiny činností, ke kterým byly specifikovány vstupy, výstupy a odpovědnosti rolí a útvarů.

- 4) Následně byl vytvořený řídicí dokumenty pro „uzákonění“ nově definovaných procesů:
 - a) úpravy organizačního řádu,
 - b) návrh směrnice upravující celou oblast správy dat,
 - c) vytvoření a úpravy pracovních pokynů a standardů upravujících způsob provádění vybraných skupin činností, zejména ve vztahu k dodavatelům aplikací.
- 5) Školení dotčených útvarů a rolí
- 6) Rozvoj a implementace nástrojů pro podporu vybraných oblastí:
 - a) zachycení datové architektury (konceptuální a logické modely),
 - b) správa a řízení metadat – budování datového katalogu, který je propojen na datové katalogy jednotlivých databází,
 - c) řízení datové kvality, kdy v nástroji jsou evidovány pro jednotlivé datové objekty požadavky na datovou kvalitu, výsledky realizovaných testů, navrhovaná opatření a jejich vyhodnocení.
- 7) Vytvoření Strategie v oblasti správy dat do roku 2020. Strategie obsahovala následující části:
 - a) poslání a vize správy dat,
 - b) strategické cíle správy dat,
 - c) zhodnocení současného stavu správy dat,
 - d) definování cílového stavu správy dat a formulace požadavků,
 - e) dílčí cíle, termíny a klíčové ukazatele dosažení cílů.

Celý uvedený postup byl rozložen do tří let. Na definování a implementaci procesů v oblasti správy dat se podílelo deset osob z útvaru zodpovědného za správu dat. Dále se připomínkování vybraných výstupů a diskusí ohledně rozdělení kompetencí účastnilo pět pracovníků za útvar zodpovědný za řízení informatiky.

Mezi nejvíce řešené problémy patřily zejména:

- oddělení správy dat od informatických procesů, výběr a nastavení jednotlivých oblastí, podoblastí a skupin činností v rámci správy dat,
- rozdělení práv a odpovědností za oblast správy dat mezi útvar správy dat a útvar informatiky,
- vyjasnění rolí, kdy z poměrně velkého seznamu rolí byly nakonec implementovány pouze role správce dat, vlastník dat, vlastník procesu. Všechny ostatní odpovědnosti nebyly řešeny rolími, ale byly přiděleny přímo na jednotlivé útvary.

4 Závěr

Popisovaný projekt je zatím v České republice ojedinělý a průkopnický. Ale v kontextu současných trendů, kdy stále více částí informatického řešení (služba, software, aplikace, infrastruktura) je outsourcováno, resp. poskytováno jako služba, a dokonce v poměrně širokém i dlouhém dodavatelském řetězci, bude na jasném vymezení procesů správy dat stále více záležet. Současně bude nutné i díky stále rostoucím legislativním požadavkům řídit datovou architekturu a její vazby k ostatním architektuрам.

Poděkování: Tento článek byl zpracován za podpory prostředků dlouhodobého koncepčního rozvoje vědy a výzkumu Fakulty informatiky a statistiky Vysoké školy ekonomické v Praze.

Literatura

1. NAŘÍZENÍ EVROPSKÉHO PARLAMENTU A RADY: o ochraně fyzických osob v souvislosti se zpracováním osobních údajů a o volném pohybu těchto údajů a o zrušení směrnice 95/46/ES (obecné nařízení o ochraně osobních údajů). Brusel, 2016, číslo 679. [cit. 30.8. 2018] Dostupné z: <http://eur-lex.europa.eu/legal-content/CS/TXT/PDF/?uri=CELEX:32016R0679&from=LT>
2. Návrh NAŘÍZENÍ EVROPSKÉHO PARLAMENTU A RADY: o respektování soukromého života a ochraně osobních údajů v elektronických komunikacích a o zrušení směrnice 2002/58/ES (nařízení o soukromí a elektronických komunikacích). Brusel. 2017. [cit. 30.8. 2018] Dostupné z: <https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX:52017PC0010>
3. Strategický rámec rozvoje veřejné správy České republiky pro období 2014-2020. Dostupné z: <http://database-strategie.cz/cz/mv/strategie/strategicky-ramec-rozvoje-verejne-spravy-ceske-republiky-pro-obdobi-2014-2020> .
4. Usnesení vlády č. 889 ze dne 2.11.2015, Příloha č. 1: Strategie rozvoje ICT služeb veřejné správy a její opatření na zefektivnění ICT služeb. [cit. 30.8. 2018] Dostupné z: <https://apps.odok.cz/zvlady/usneseni/-/usn/2015/889> .
5. Felix, O., Hrabě, P., Šedivý, T.: Prezentace Národního architektonického plánu. Praha, září 2016. Oddíl Architektonické vzory a referenční modely. [cit. 30.8. 2018] Dostupné z: <http://www.mvcr.cz/clanek/agenda-odboru-hlavniho-architekta-egovernmentu.aspx?q=Y2hudW09OA%3d%3d>
6. Odbor Hlavního architekta eGovernmentu, MVČR: Definice pohledů na architekturu pro vyplnění formuláře žádosti OHA verze 2.0. Praha, říjen 2016. Oddíl Metodiky a další dokumenty. [cit. 30.8. 2018] Dostupné z: <http://www.mvcr.cz/clanek/agenda-odboru-hlavniho-architekta-egovernmentu.aspx?q=Y2hudW09Ng%3d%3d>
7. Mosley, Mark a kol., 2010. The DAMA Guide to the Data Management Body of Knowledge. Technics Publications, LLC, 2010. str. 406. ISBN 978-1-9355040-2-3.
8. COBIT 5: A Business Framework for the Governance and Management of Enterprise IT. Rolling Meadows. IL: ISACA, 2012. ISBN 978-160-4202-373.
9. QGCIO, 2017. Queensland Government Information Management Policy Framework Definitions. In: Queensland Government [Online]. August 2017. [cit. 30.8. 2018] Dostupné z: <https://www.qgcio.qld.gov.au/documents/information-management-policy-framework>.
10. Vofříšek, Jiří, a kol. 2008. Principy a modely řízení podnikové informatiky. Praha: Oeconomia, 2008. ISBN: 978-80-245-1440-6.
11. Australian Government Information Management Office. Australian Government Architecture Reference Models. Version 3.0. In: Australian Government. [Online] August 2011 [Citace: 31.8.2018] <https://www.finance.gov.au/sites/default/files/aga-ref-models.pdf>.

Annotation:

Data Management Approaches in a Large Organization Environment

The theoretical and practical approaches to defining the area of data management will be described as separate area of management with their own strategy and defined processes, roles, responsibilities and independent tools.

Integrace relačních a grafových databází funkcionálně

Jaroslav Pokorný

Matematicko-fyzikální fakulta Univerzita Karlova v Praze,
Malostranské nám. 25, 118 00 Praha 1
pokorny@ksi.mff.cuni.cz

Abstrakt. V dnešním světě multimodelových databází integrujeme databáze vyjádřené v různých datových modelech. Aktuální je integrace relačních a NoSQL databází. Cílem článku je ukázat možnosti integrace relačních a grafových databází (GDB) pomocí funkcionálního datového modelu a jeho formálního jazyka - typovaného λ -kalkulu (jazyk LT). Předpokládáme, že existuje databázové schéma jak pro relační databázi tak pro GDB. Relace jsou uvažovány jako charakteristické funkce a atributové grafy jako množiny jednohodnotových, n -ticových a vícehodnotových funkcí. Dotaz nad takovou integrovanou databází je pak možné vyjádřit výrazem LT. Grafová část dotazu se transformuje do dotazovacího jazyka GDB, relační pak např. do SQL. Uživatelsky přívětivější verze této dotazovací podmnožiny LT by v praxi mohla sloužit jako účinný nástroj pro vyhledávání nad tímto typem multimodelové databáze.

Klíčová slova: grafové databáze, relační databáze, integrace databází, funkcionální datový model

1 Úvod

NoSQL databáze rostou v popularitě zejména pro aplikace v oblasti Big Data. Významnou kategorií NoSQL databází jsou *grafové databáze* (GDB) [8]. GDB jsou založeny na teorii grafů, používají *uzly*, *vlastnosti* a *hrany*. Uzel reprezentuje entitu, jako je Uživatel, Film, Objekt, orientovaná hrana představuje vztah mezi dvěma uzly, např. *Je_přítelem*, *Kdo_kupuje_co* atd. Označované uzly a hrany mohou mít různé vlastnosti (atributy) dané páry klíč-hodnota. Mezi dvěma uzly může existovat více hran. Atributový graf je pak multigrafem. GDB může být buď jeden graf nebo množina grafů. Uvažovat budeme první kategorii. *Grafové SŘBD* (GSŘBD) se ukázaly jako velmi účinné a vhodné pro mnoho případů využití dat, kde vztahy hrají klíčovou roli. Dotazování je pak aktuálním problémem v aplikacích využívajících GDB.

Vzhledem k tomu, že každá NoSQL databáze má své vlastní API a typicky nepodporuje standardy jako SQL a JDBC, integrace těchto systémů s podnikovými databázemi či reportovacím softwarem vyžaduje další úsilí. Problémem je integrace relačních databází a NoSQL databází, tj. NoSQL databáze jsou uvažovány v kontrastu s relačními produkty. Jsou však možné i jiné přístupy, např. funkcionální přístup. Koncem 80. let existoval funkcionální jazyk DAPLEX [9]. Jazyk umožňoval pouze vnořené aplikace funkcí. Funkcionální přístup lze snadno uvažovat v prostředí

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 43-48.*

GDB. Z GSŘBD můžeme zmínit Gremlin¹ - funkcionální grafový dotazovací jazyk vyvinutý Apache TinkerPop, který umožňuje vyjádřit složitý průchod grafem. Řada prací využívajících funkcionální přístup ke správě dat je obsažena v [2]. Použijeme funkcionální přístup, ve kterém je atributový graf reprezentován funkcemi - atributy. Inspiraci lze nalézt v databázovém modelu HIT (např. [3]), jako funkcionální alternativy E-R modelu. Funkce budou jednohodnotové, vícehodnotové a n-ticové. Jako manipulační jazyk lze použít typovaný λ -kalkul. Cílem článku je představit funkcionální přístup k modelování relací a atributových grafů. Diskutovány jsou možnosti integrace na úrovni společného dotazovacího jazyka.

V sekci 2 se zabýváme funkcionálním přístupem k modelování a dotazování dat pomocí typovaného λ -kalkulu. Je ukázáno, jak funkcionálně chápat atributové grafy a databázové relace. Použitelnost jazyka pro integraci RDB a GDB je diskutována v sekci 3. Sekce 4 podává závěry a návrhy pro budoucí výzkum a vývoj.

2 Funkcionální přístup k datovému modelování

Náš funkcionální přístup je založen na typovém systému. Budeme používat elementární typy a dva strukturované typy. Typované funkce vhodné pro modelování reálných datových objektů jsou *atributy* vnímané jako empirické typované funkce, které jsou popsány výrazem přirozeného jazyka [3]. Přístup byl studován hlavně v 90. letech v kontextu konceptuálního modelování databází (viz např. [4]).

Předpokládáme existenci některých (*elementárních*) typů $S_1, \dots, S_k$ ($k \geq 1$) tvořících bázi **B**. Složitější typy jsou konstruovány následujícím způsobem.

Pokud $S, R_1, \dots, R_n$ ($n \geq 1$) jsou typy, pak

- (i) $(S:R_1, \dots, R_n)$ je (*funkcionální*) typ,
- (ii) $(R_1, \dots, R_n)$ je (*n-ticový*) typ.

Množina typů **T** nad **B** je nejmenší množina obsahující všechny typy z **B** a ty odvozené pomocí (i) - (ii). Když jsou S_i v **B** interpretovány jako neprázdné množiny, pak $(S:R_1, \dots, R_n)$ označuje množinu všech (totálních nebo partiálních) funkcí z $R_1 \times \dots \times R_n$ do S , $(R_1, \dots, R_n)$ označuje kartézský součin $R_1 \times \dots \times R_n$. Značení A/t říká, že objekt A je typu t . O A je pak možné říci, že jde o t -objekt.

Tento zcela obecný aparát umožňuje definovat složité strukturované typy. Pro naše databázové aplikace jsou podstatné elementární typy a omezení na strukturální složitost na nich vybudovaných funkcí. Pro popis databázových objektů vystačíme s funkcionálními a n-ticovými typy, u kterých S a R_i jsou pouze elementární typy.

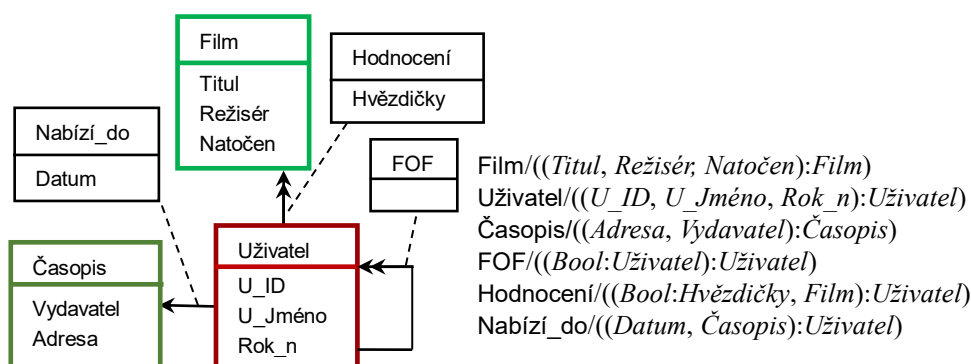
Mezi elementárními typy předpokládáme *String*, *Number* a typ *Bool* = {TRUE, FALSE}. *Bool* umožňuje modelovat množiny a relace jako unární resp. n-ární charakteristické funkce typu $(Bool:S_1, \dots, S_n)$ ($n \geq 1$). Logické spojky a predikáty jsou rovněž typované funkce, např. **and**/ $(Bool: Bool, Bool)$, podobně i kvantifikátory. Aritmetické operace jsou příklady $(Number: Number, Number)$ -objektů, $COUNT_S/((Number:(Bool:S))$.

¹ <https://tinkerpop.apache.org/gremlin.html>

2.1 Atributové grafy funkcionálně

Nechť $\mathbf{B} = \{Uživatel, Film, U_ID, Jméno, Rok_n, \dots\}$. Výraz "filmy hodnocené uživatelem" označuje $((Bool: Film): Uživatel)$ - objekt, tj. (parciální) funkci $f: Uživatel \rightarrow (Bool: Film)$. Báze $\mathbf{B}$ obsahuje *deskriptivní* a *entitní* typy. Pro GDB můžeme entitní typy chápat jako množiny ID uzlů. Deskriptivní jsou *String*, *Number* a *Bool*.

Pojem atributu použitý v GDB může být omezen na atributy typů $(R:S)$ a $((Bool:R):S)$, kde R a S jsou typy entit. Vlastnosti popisující typy entit mohou být typů $((S_1, \dots, S_m):R)$ ($m \geq 1$), kde S_i jsou deskriptivní elementární typy, R je typ entit. Hrany pak lze v atributových grafech definovat pomocí typů $((S_1, \dots, S_m, R_1): R_2)$ nebo $((Bool: S_1, \dots, S_m, R_1): R_2)$ ($m \geq 0$).



Obr. 1. Databázové schéma GDB Filmy

Obr. 2. Funkcionální schéma GDB Filmy

Ač NoSQL databáze většinou nepoužívají databázové schéma, zde musíme mít tyto struktury k dispozici. V modelu zavedeném v [5] je schéma GDB opět vyjádřeno atributovým grafem (obr. 1 – schéma Filmy). Orientované hrany označují funkcionální typy vztahů mezi uzly (typy entit). Dvojitá šipka prozrazuje vztah 1:N mezi dvěma typy uzlů. Může existovat i více typů hran daného typu mezi dvěma typy uzlů, tj. GDB je pak *multigrafem*. Ekvivalentní vyjádření schématu Filmy ve funkcionálním modelu je na obr. 2.

Všimněme si, že v tomto databázovém modelu mají jak uzly, tak hrany pevný počet vlastností. Vlastnosti uzlů i hran jsou modelovány velmi jednoduše, tj. jejich množina je specifikována n-ticí. To může být příliš striktní ve světě NoSQL databází. Na druhé straně vlastnosti jsou atributy – parciální funkce. To znamená, že pro některé uzly resp. hrany GDB nemusí být definovány. Přidávání vlastností lze technicky řešit např. dynamickým rozšiřováním schématu GDB v závislosti na vstupujících hodnotách nových vlastností do GDB. Odstraňování vlastností ze schématu není nutné. V limitním případě daná vlastnost nemusí nabýt hodnoty na žádném uzlu/hraně aktuální GDB. Pro GDB je pak možné hovořit o *rozšiřitelném databázovém schématu*. Relativně neměnnou zůstává jeho část týkající se grafové struktury, dynamická část se týká vlastností.

2.2 Relace funkcionálně

Ve světě RDB, jsou jako elementární typy S_i použity (relační) atributy $A_i:D_i$. S_i jsou pro každý typ relace R neprázdné množiny hodnot, $S_i \neq S_j$ pro $i \neq j$. R je v uvažovaném typovém aparátu ($Bool:S_1, \dots, S_m$)-objekt ($m \geq 0$).

Uvažujeme RDB s relacemi Herec(Jméno, Titul, Role) a Film(Titul, Natočen, Režisér, Žánr), tj. ve funkcionální přístupů atributy

Herci/(*Bool:Jméno, Titul, Role*),
Filmy/(*Bool:Titul, Natočen, Režisér, Žánr*).

Všimněme si, že informace o primárních klíích relací (podtržené atributy v jejich schématech) se v atributovém vyjádření ztrácí.

2.3 Typovaný λ -kalkul

DML pro funkce je tradičně typovaný λ -kalkul (*jazyk termů* – LT), který se skládá z konstant, proměnných, aplikací funkcí, n -tice termů a λ -abstrakcí. Termy jsou typované. Je-li term $M/(R_1, \dots, R_n)$, pak *komponenty* $M[1], \dots, M[n]$ jsou rovněž termy.

Jazyk LT s aplikacemi funkcí, λ -abstrakcemi a komponentami poskytuje silný nástroj pro dotazování nad GDB [6]. Důležitou je λ -abstrakce. Umožňuje vytvářet množiny (relace) jako argument dalších funkcí, např. UNION, COUNT apod. Např. dotaz "Najdi názvy filmů natočených Spielbergem" lze vyjádřit termem

$$\lambda t (\exists m, n \text{ Film}(m)(t, \text{'Spielberg'}, n)) \quad (1)$$

typu (*Bool:Titul*). Dotazování nad typovanou RDB připomíná doménový relační kalkul. Např.

$$\lambda n (\forall t (\exists re, g \text{ Filmy}(t, re, \text{'Spielberg'}, g) \text{ implies } \exists ro \text{ Herci}(n, t, ro))) \quad (2)$$

vyjadřuje dotaz "Najdi herce, kteří hrají v každém filmu režiséra Spielberga."

Bez ohledu na vyjadřovací sílu, LT termy nejsou moc uživatelsky přívětivé. Jak částečně řešit tento problem, naznačují tyto verze dotazů (1) a (2):

$$\lambda t^{Titul} (\exists m^{Film} \text{ Film}(m^{Film})(t^{Titul}, \text{'Spielberg'}^{Režisér}))$$

$$\lambda n^{Jméno} (\forall t^{Titul} (\text{Filmy}(t^{Titul}, \text{'Spielberg'}^{Režisér}) \text{ implies Herci}(n^{Jméno}, t^{Titul})))$$

Eliminovali jsme existenční kvantifikátory a odpovídající proměnné, týkající se pouze jedné aplikace funkce. Explicitní informace o typu proměnné také usnadňuje čtení výrazu dotazu. Komponenty n -tice rovněž není nutné provádět pozici, ale jménem vlastnosti/typu uzlu. Místo $\text{Film}(\text{Id4})[1]$ je čitelnější výraz $\text{Film}(\text{Id4}).\text{Titul}$.

Dotazy lze tedy vyjadřovat vhodnými termy konstruujícími množiny, relace, n -tice, Boolské hodnoty (ANO/NE dotazy), ale i složitější datové struktury. Pro GDB by mohlo jít o grafy. Např. term dotazu

$$\lambda u_1 \lambda u_2 (\text{FOF}(u_1)(u_2) \text{ and } \text{Uživatel}(u_2).\text{Rok_n} > 2000)$$

je typu ($(\text{Bool:Uživatel}):Uživatel$). Jde o vícehodnotovou funkci přiřazující každému uzlu typu *Uživatel* množinu uzlů typu *Uživatel* s vlastností $\text{Rok_n} > 2000$. Jinými

slovy řečeno, odpovědí na dotaz je jistý graf s hranami, které ovšem nemají přiřazenou žádnou značku.

3 Integrace relací a atributových grafů

Kniha [1] nabízí tři způsoby integrace dvou různých světů - relačních a NoSQL databází: *nativní*, *hybridní* a *redukující na jednu možnost* (relační nebo NoSQL). K integraci existují různé přístupy [6]:

- polyglotní perzistence,
- multimodelový přístup,
- víceúrovňové modelování,
- NoSQL relačně a model konverze schématu s následným převodem dat.

Pro integraci RDB a GDB funkcionálně je nejbližší víceúrovňové modelování. To obecně zahrnuje přístupy jako

- a) *speciální abstraktní model*,
- b) *NoSQL-on-RDBMS*, či
- c) *integraci ontologií*.

Náš přístup je blízký přístupu a), kde výchozím modelem je funkcionální datový model. Vzhledem k tomu, že funkcionální modelování grafů poskytuje spíše hybrid mezi konceptuálním a databázovým schématem, je relevantní i přístup c). Dotazy v LT odeslané do integrovaného systému jsou přeloženy do dílčích dotazů kompatibilních s RDB (např. SQL) a GDB (např. Cypher z Neo4j [8]).

Na úrovni dat jsou odpovídající databáze obecně heterogenní. Elementární typ *Titul* filmů nemusí být stejný jako doména atributu *Titul* relace *Film*. Předpokládá by se mohl jejich neprázdný průnik. Pro integraci obou schémat je proto nutné alespoň přejmenovat některé atributy a/nebo relace (zde *Film*). V relačním schématu použijeme *Filmy*. Pak,

$$\lambda u^{Uživatel}, g^{Žánr}, n^{Number} (n^{Number} = \text{COUNT}_{\text{Film}} (\lambda m^{\text{Film}} (\text{Hodnocení}(u^{Uživatel})(m^{\text{Film}}) \text{ and } \exists t^{\text{Titul}} s^{\text{Titul}} \text{Film}(m^{\text{Film}}).t^{\text{Titul}} = s^{\text{Titul}} \text{ and Filmy}(s^{\text{Titul}}, g^{\text{Žánr}})))$$

vyjadřuje dotaz "Najdi pro každého uživatele a žánr počet recenzí, které v žánru provedl". Odpovědí na dotaz bude ternární relace kombinující ID uzlů (*Uživatel*), hodnoty relačního atributu (*Žánr*) a odvozené hodnoty (*Number*).

4 Závěry

V příspěvku jsme se zaměřili na GDB s atributovým grafem v integraci s RDB pomocí funkcionálního přístupu.

Současné výzvy pro databázový výzkum takové infrastruktury zahrnují:

- nalezení vhodné a dostatečně silné podmnožiny LT pokrývající požadavky dotazování GDB integrované s RDB,
- vývoj silné a uživatelsky přívětivé verze dotazovacího jazyka založeného na LT,
- vývoj prototypu využívajícího RDB a Neo4j GSŘBD pro zdrojové databáze.

Další téma je návrh a implementace softwarového prostředku pro návrh rozšířitelného databázového schématu pro GDB v kontextu existujícího GRBD.

Literatura

1. Gaspar, D., Coric, I.: Bridging Relational and NoSQL Databases. IGI Global, 2017.
2. Gray, P.M.D., Kerschberg, L., King, P.J.H., and Poulouvassilje: The Functional Approach to Data Management. Springer, Berlin, 2003.
3. Pokorný, J.: A function: unifying mechanism for entity-oriented database models. In: Batini, C. (Ed.), Entity-Relationship Approach, Elsevier Science Publishers B.V., North-Holland (1989), 165 181.
4. Pokorný, J.: Database semantics in heterogeneous environment. In: Proc. of 23rd Seminar SOFSEM'96: Theory and Practice of Informatics, K.G. Jeffery, J. Král, M. Bartošek (Eds.), Springer-Verlag, (1996) 125 142.
5. Pokorný, J.: Conceptual and Database Modelling of Graph Databases. In: Proc. of IDEAS' 16, B. Desai (Ed.), ACM, (2016) 370 377.
6. Pokorný, J.: Functional Querying in Graph Databases. Vietnam Journal of Computer Science, Springer, 5(2), (2017) 95 105 DOI 10.1007/s40595-017-0104-6
7. Pokorný, J.: Integration of Relational and NoSQL Databases. In: Ngoc Thanh Nguyen, et al (eds), ACIIDS (2), LNCS, vol. 10752, (2018) 35 45.
8. Robinson, I., Webber, J., Eiffrém, E.: Graph Databases. O'Reilly Media, 2013.
9. Shipman, D.W.: The functional data model and the data languages DAPLEX. ACM Transactions on Database Systems (TODS), vol. 6, issue 1, ACM (1981) 140 173.

Poděkování. Práce byla podpořena projektem Q48 programu Progres na UK, Praha.

Annotation:

Integration of relational and graph databases functionally

In today's world of multimodel databases, we often integrate databases expressed in different data models. Recently, an integration of relational and NoSQL databases is required. The aim of the paper is to show the possibility of integration of relational and graphical databases (GDB) using a functional data model and its formal language - a typed λ -calculus (LT language). We assume that a database schema exists for both the relational and the GDB. Relations are considered as characteristic functions and attribute graphs as sets of one-valued, tuple and multivalued functions. A query over such an integrated database can then be expressed into LT. The graph part of the query is transformed into the GDB query language, the relational part then into, e.g., SQL. A more user-friendly version of the query subset of LT could in practice serve as an effective search tool for this type of multimodel database.

Interaktivní a grafická podpora výuky dotazovacího jazyka SPARQL

Vojtěch Svátek¹, Jindřich Mynarz^{1,2}, Marek Dudáš¹ a Jan Kučera³

¹ Katedra informačního a znalostního inženýrství, FIS VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
{svatek|marek.dudas}@vse.cz

² MSD IT Global Innovation Center s.r.o., Svornosti 3321/2, 150 00 Praha 5
jindrich.mynarz@merck.com

³ Katedra informačních technologií, FIS VŠE, Nám. W. Churchilla 4, 130 67 Praha 3
jan.kucera@vse.cz

Abstrakt. Znalost dotazovacího jazyka SPARQL je jednou z klíčových dovedností pro práci s propojenými daty na webu. Pro jeho výuku na VŠE v Praze postupně vznikají dva nástroje: interaktivní cvičebnice SPARQLab a grafický katalog dotazových vzorů PatRQL implementovaný ve formě obohacené myšlenkové mapy.

Klíčová slova: SPARQL, sémantický web, cvičebnice, vzor

1 Úvod

Jazyk SPARQL je v současnosti jedním z hlavních nástrojů, pomocí kterých lze ve webovém prostředí prozkoumávat a zpracovávat strukturovaně publikovaná data. Pro jeho výuku jsou dostupné kvalitní učebnice typu *Learning SPARQL* [2], umožňující seznamovat se s jazykem krok za krokem, včetně dílčích cvičení. Úplný seznam možností podávají oficiální specifikace včetně *referenčních příruček* vydaných jako standardy W3C.[1] Žádný z těchto zdrojů ale sám o sobě nezajišťuje uživateli možnost aktivně si formulovat své pokusy o dotazy a porovnávat je s vzorovým správným řešením,⁴ i když interaktivitu lze částečně simulovat na úrovni *prezentace* odkazující na dotazovací rozhraní třetích stran; takto funguje prezentace *SPARQL By Example*⁵ vytvořená v r. 2013 firmou Cambridge Semantics, která pro případ nefunkčnosti externích zdrojů zahrnuje jako svou součást statickou reprezentaci očekávaných výsledků. Konečně, pro *plně interaktivní* výuku SPARQLu nad dedikovaným rozhraním byla v. 2015 ve Swiss Institute of Bioinformatics vyvinuta aplikace SPARQL Playground.⁶ Ani jeden z těchto přístupů ovšem neposkytuje studentovi zpětnou vazbu, pokud jde o samotný tvar použitého dotazu, ale nanejvýš pokud jde o výsledky

⁴ Toto umožňuje v případě jazyka SQL např. aplikace <https://www.sqlteaching.com/>.

⁵ <https://www.cambridgesemantics.com/blog/semantic-university/learn-sparql/sparql-by-example>

⁶ <http://sparql-playground.sib.swiss>

dotazem vrácené. Srovnání dotazu položeného podle textového zadání se vzorovým tvarem dotazu je úlohou poměrně komplikovanou, protože dotaz vracející stejné výsledky lze zpravidla formulovat více způsoby, přičemž jejich výpočetní náročnost, i syntaktická elegance a srozumitelnost, se mohou lišit.

Pro samotnou tvorbu dotazu jsou pak hlavními možnostmi tvorba “na zelené louce”, z jednotlivých elementárních příkazů a trojic grafového vzoru, nebo úprava již existujícího, hotového dotazu. Je zjevné, že první možnost je pracná, zatímco druhá vede k riziku zavlečení nadbytečných nebo suboptimálních prvků. Stejně jako v programování nebo např. tvorbě ontologií se jeví jako vhodné podpořit tvorbu dotazů pomocí katalogů vzorů. Jediný autorům známý výzkumný projekt věnovaný vzorům SPARQL dotazů se však zaměřil jen na empirickou analýzu výskytu vzorů v rámci rozsáhlých logů populárních serverů se SPARQL rozhraním [3]. Vzory zkoumané v této analýze jednak reprezentují pouze dotazy pokládáné do veřejných rozhraní (jejichž výpočetní složitost je tudíž nutně omezena), jednak jsou zjednodušeny o prvky, které autoři pokládali za pro vzory nepodstatné (např. modifikátory ORDER BY nebo DISTINCT), ale do výuky SPARQL patří. Uvedený seznam častých vzorů (potenciálně využitelný jako výukový materiál) také postrádá provázání na další pojmy a doporučení.

Na VŠE je jazyk SPARQL vyučován již od r. 2006. Podkladovými materiály však zpočátku byly jen webové prezentace a později zahraniční učebnice. Potřeba vyplnit mezery popsané výše byla od r. 2015 motivací pro návrh a řešení dvou projektů Interní rozvojové soutěže, v rámci kterých byly vytvořeny dva nástroje: SPARQLab a PatRQL; oba jsou i nadále vylepšovány a rozšiřovány.

2 Cvičebnice SPARQLab

SPARQLab⁷ je webová aplikace volající rozhraní SPARQL (tzv. endpoint) s vystaveným reálným datasetem. Aplikace předkládá uživateli sadu (aktuálně již přes 90) textově zapsaných cvičných úloh na formulaci SPARQL dotazů, pro které má uložena vzorové řešení, automaticky s nimi zadané dotazy porovnává a upozorňuje na odlišnosti. Díky expresivitě SPARQLu mohou být správná řešení cvičení formulována mnoha způsoby. Abychom snížili podíl správných řešení vyhodnocených jako nesprávná, SPARQLab při porovnávání se vzorovým řešením abstrahuje od nepodstatných detailů syntaktické formulace dotazu. Dotazy jsou porovnány na úrovni SPARQL algebry [1, sekce 18] po aplikaci optimalizací sjednocujících zaměnitelné konstrukty jazyka SPARQL a bez ohledu na pojmenování proměnných. Pokud se dotazy v tomto srovnání neshodují, provede se porovnání jejich výsledků. Výsledky SELECT dotazů jsou porovnány jako množiny množin pro abstrakci od pořadí, pokud dotaz nevyžaduje ORDER BY. ASK dotazy, u kterých je velké riziko falešné shody (binárního) výsledku, jsou transformovány na SELECT dotazy omezené na vzorek výsledků. V případě CONSTRUCT a DESCRIBE dotazů je pro srovnání použit algoritmus grafového isomorfismu.

⁷ Zdrojový kód je na <https://github.com/jindrichmynarz/sparqlab>, aktuální instance je vystavena na <http://doc.lmcloud.vse.cz/sparqlab>.

Data, nad kterými SPARQLab operuje, byla převzata z portálu otevřených dat České správy sociálního zabezpečení (ČSSZ). Tato data používají principy multidimenzionálního modelování pomocí slovníků Data Cube Vocabulary⁸ a SKOS,⁹ což je v současnosti nejlepší praxe pro statisticky zaměřené datasety publikované veřejnou správou jako propojená data na webu.

3 Katalog vzorů PatRQL

PatRQL je grafický model vyvíjený v prostředí editoru konceptových map ContextMinds. Předkládá provázání mezi jednotlivými konstrukty SPARQLu, na nich založenými komplexnějšími vzory, kategoriemi SPARQL dotazů, a problémy, které má dotaz/vzor vyřešit. Podkladem pro formulaci vzorů byla jak vlastní zkušenost autorů, tak i zmínky vybrané z literatury, zejména z knihy *Learning SPARQL* [2]. ContextMinds¹⁰ je webový nástroj pro tvorbu studijních materiálů v podobě konceptových (pojmových) map. V principu je podobný nástroji CmapTools [4], ve vývoji v rámci “startupového” projektu jednoho z autorů tohoto příspěvku jsou však pokročilé funkce založené na technologiích sémantického webu, které by měly tvorbu map výrazně usnadnit. Konceptová mapa reprezentuje znalosti formou grafu, kde uzly představují koncepty a hrany mezi nimi znázorňují vztahy. Dlouhodobé experimenty se zapojením konceptových map ve výuce prokázaly výrazné zlepšení studijních výsledků [5].

Na Obr. 1 je výřez části katalogu s okolím jednoduchého vzoru. Pro řešení problému popsaného boxem vpravo nahoře je doporučen vzor kombinující konstrukty COUNT a GROUP BY. Jedná se o vzor, který lze vcelku chápat jako tzv. explorační dotaz; pro tuto kategorii dotazů je pak typické využívání pojmenovaných grafů (konstrukt GRAPH), aby nebyla do explorační primárních dat zahrnuta i data z pomocných grafů. Jeden z konstruktů, GROUP BY, lze dále v určitém kontextu chápat jako alternativní ke konstruktu DISTINCT, protože oba zajišťují deduplikaci výsledků. Při najetí myši na některý uzel jsou zobrazeny meta-informace: v tomto případě se jedná o textový popis “logiky” vzoru. Aktuální podoba katalogu je dostupná na webu.¹¹

4 Závěr a perspektivy

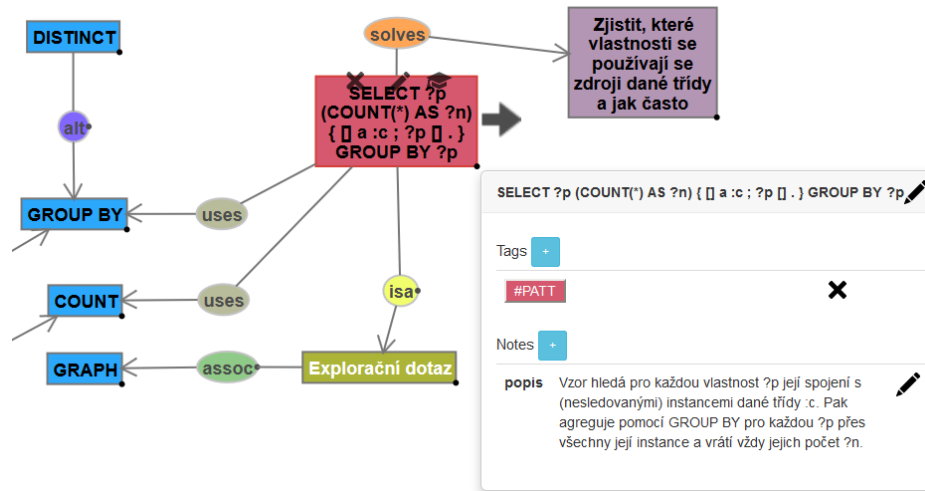
V příspěvku byly představeny dva nástroje, které by měly společně přispět k usnadnění pochopení dotazovacího jazyka SPARQL. Cvičebnice SPARQLab se již osvědčila ve výuce předmětu *Propojená data na webu* na FIS VŠE, a přinejmenším ukázkově je využívána v obdobných předmětech i na dalších českých školách (např. MFF UK nebo FIT ČVUT). Katalog vzorů PatRQL je v ranější fázi zralosti, a jeho nasazení ve výuce proběhne až v říjnu 2018. V dlouhodobém

⁸ <https://www.w3.org/TR/vocab-data-cube>

⁹ <https://www.w3.org/2004/02/skos/>

¹⁰ Prototyp nástroje je dostupný na <http://app.contextminds.com>.

¹¹ <https://bit.ly/2LOLpDG>



Obr. 1. Fragment katalogu PatRQL

horizontu se předpokládá větší propojení obou nástrojů. U PatRQLu, resp. platformy ContextMinds, je třeba dorešit i otázku modularizace rozsáhlých grafů.

Literatura

1. SPARQL 1.1 Overview. W3C Recommendation 21 March 2013. Online <https://www.w3.org/TR/sparql11-overview>.
2. DuCharme, B.: Learning SPARQL. Querying and Updating with SPARQL 1.1. O'Reilly Media (2011).
3. Stegemann, T., Ziegler, J.: Pattern-Based Analysis of SPARQL Queries from the LSQ Dataset. In: Nikitina, N. et al. (eds.) International Semantic Web Conference (Posters, Demos & Industry Tracks). CEUR-WS.org (2017).
4. Canas, A. J., Novak, J. D.: Concept mapping using CmapTools to enhance meaningful learning. In: Knowledge cartography. Springer (2008). p. 25-46.
5. Novak, J. D.: Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations. Taylor & Francis (2010).

Poděkování: Výsledky popisované v tomto textu vznikly s podporou rozvojových projektů 6/2016 a 50/2017 řešených na FIS VŠE v Praze.

Annotation:

Interactive and graphical support of SPARQL query language education

The SPARQL language is one of the semantic web and linked data cornerstones. Two tools are described that aim to support SPARQL education: the interactive web exercise book SPARQLab, and the graphical query pattern catalogue PatRQL, implemented as an enriched mindmap.

Selinon – Dynamic Distributed Task Flows

Fridolín Pokorný

Red Hat Czech s.r.o., Purkyňova 99, 612 00 Brno
`fridolin@redhat.com`

Abstract. There has been a significant growth in volume of data produced and available publicly on the Internet as well as in private sources. Processing such volume of data can be challenging and in many cases time consuming for a single computer. In this paper we present a distributed system made of simple tasks that can be executed with high level of parallelism using multiple cluster nodes. System behaviour is defined via a simple configuration file. This declarative abstraction model and distributed architecture allows to model complex data flows while assuring stability and resilience even during online production redeployment.

Keywords: big data · data processing · distributed systems · Python

1 Introduction

In the upcoming sections there will be discussed core concepts of distributed task queues, briefly introduced project Celery that is used by Selinon as a task scheduling backend, Celery's disadvantages and basic abstractions used in project Selinon [1,2] we developed that helped us horizontally scale a self-sustaining system able to process big data.

2 Distributed Task Queues

The idea of distributed systems is definitely not novel. There exist multiple implementations of distributed task queues – in the Python ecosystem, there can be found namely project Celery [3], Spotify Luigi [8], Apache Airflow [7] and others. Most of these solutions require a broker and workers. A broker (also known as message broker or message queue) handles queueing messages which carry information what should be computed. Subsequently, there are connected workers to the broker which consume messages and process them. If a message is processed, it is removed from message broker and the worker that processed the given message is available to process other messages queued on broker. Most often, there is also available a database or storage to which computed results are stored to. The described architecture can be seen on Figure 1.

In many scenarios, having a single task for a request described by a message and processed by workers is not desired. In order to take advantage of horizontally scalable solution the request can be split into standalone units, called

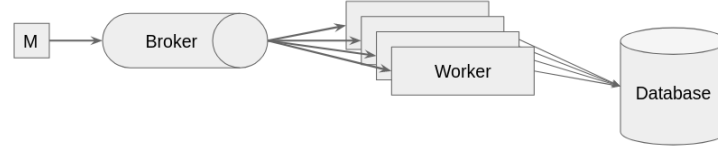


Fig. 1: An architecture of a distributed task queue. Message M is queued on the message broker, subsequently picked by workers which persist computed results into database.

tasks, that compute one type of a result which can be additionally processed by other standalone tasks (also known as task flows, see below). The advantage of such logical split is mostly in possible parallelism and task retries in case of task failures. If a task fails, it can be retried without any negative impact on other independent tasks that can meanwhile run in parallel on different workers.

3 Using Celery for Distributed Task Queue

All of the objectives mentioned in the Section 2 can be accomplished with stated open source solutions from Python ecosystem. In fact, we used Celery to accomplish desired behaviour. Celery offers "Celery primitives" that allow to model task flows—dependencies between tasks that can be either data dependencies (waiting for results of a task in order to start other tasks) or time dependencies (waiting on external events to proceed to next tasks). Celery primitives allow grouping tasks so tasks are executed in parallel or sequentially possibly on different workers in parallel.

An example of task flow modeling using Celery primitives can be seen on Figure 2 (the example is intentionally simplified, see Celery documentation [4] for more in depth Celery primitives explanation). With Celery primitives we can create "phases" which group tasks that can be executed in parallel as there are no inter-task dependencies inside phases. Phases need to be executed one after another sequentially due to time or data dependencies of tasks that create these phases.

The task flow made of phases using Celery's primitives is hard-coded inside application logic that makes application harder to maintain, harder to extend, more error prone with task or phase updates and adding a new task into an existing flow requires knowledge of internal dependencies of tasks that could be harder to see from source code. In the example above, the long running tasks block execution of tasks that could be already started respecting task flow dependencies.

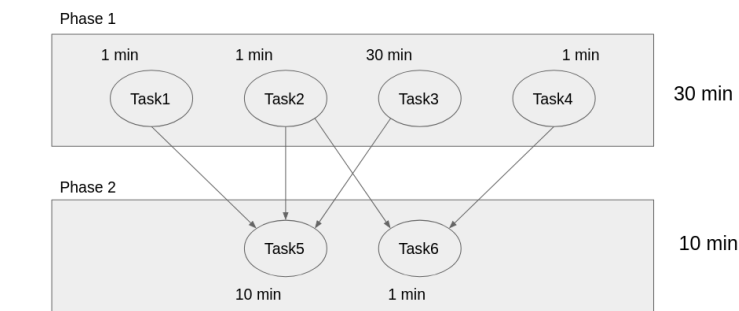


Fig. 2: Phases created using Celery’s primitives for tasks. Task **Task6** could be executed after 1 minute assuming **Task2** and **Task4** finish in one minute, but **Task3** blocks execution of **Task6**.

4 Orchestrating Task Flows with Selinon

As we experienced drawbacks when using raw Celery, we managed to design a task flow orchestration tool on top of Celery which removes phase borders and all of the disadvantages stated above.

In the upcoming subsections, there are introduced Selinon abstractions. These abstractions create core components of Selinon design that act like a loosely coupled entities which, based on configuration, can create complex systems capable of serving large workload requests.

4.1 Tasks and Task Flows

A Selinon task is abstracted into a Python class respecting Selinon task interface. Implemented Selinon tasks are subsequently registered in the Selinon YAML [5] configuration file and used to create task flows. The configuration of a flow states edges (or transitions) that model data or time dependencies between tasks. An example of such flow configuration can be seen on Figure 3.

In the example shown in Figure 3, Selinon starts **Task0** and **Task1**. Once **Task1** finishes, there are started tasks **Task2** and **Task3** that can run in parallel without possibly waiting for **Task0** to finish or fail. The `condition` section restricts **Task2** and **Task3** execution only for production systems (environment variable `PRODUCTION` is present). The conditional execution of tasks make the whole system dynamic – these conditions are evaluated each time an edge is fired and they can perform various actions such as checking external events or inspecting results of previous tasks. Their main purpose is to abstract away conditional task execution from actual task implementation logic as well as an optimization to prevent from queueing overhead needed to schedule unwanted tasks.

Besides scheduling tasks directly, Selinon is capable of creating nested flows – scheduling flows from within a flow. Moreover, the scheduling algorithm that

```

flow-definitions:
  - name: flow1
    - from:
      to:
        - Task0
        - Task1
    - from: Task1
      to:
        - Task2
        - Task3
    condition:
      name: envExist
      args:
        name: PRODUCTION

```

Fig. 3: An example of Selinon flow configuration.

Selinon offers removed a restriction on directed acyclic graph (DAG) in flow design which other open source alternatives have.

4.2 Scheduling Algorithm

The scheduling of tasks is done by a special message that is created per flow. This message keeps track of state in the flow. Each time a task or sub-flow is scheduled, the message carrying flow state information is updated and rescheduled with adjusted flow state. This mechanism allows to treat nested flows as any other task as the nested flow state is tracked in a separate newly created message referenced from the parent flow.

These special messages carrying flow state allow to sample state in the distributed environment as they are periodically scheduled and processed by workers.

4.3 Storage and Database Adapters

If there is a requirement for storing computed results there can be assigned storage adapters that handle task result storing and possible task result retrieval. In order to use an adapter, the adapter needs to be registered into the system in the YAML configuration file and assigned to a task.

This separation of task logic (how results are computed) from actual storing and retrieval mechanism allowed us to use transparently different storage and database solutions to store data in a way they could be later easily queryable based on data characteristics and query types we wanted to perform on computed results.

4.4 Application Monitoring and Healthiness

Selinon offers a built in support for application monitoring using flow tracepoints. Tracepoints are events that occur in cluster (such as a flow has started, a task has started, a task has failed) that help to monitor cluster behaviour. Selinon comes with a built-in support of error tracking software (such as Sentry [9]) as well as structured well-defined logs that can be aggregated and further inspected using cluster-wide aggregated logging.

5 Conclusion

Selinon is an open source project, it's source code is available on GitHub [2]. It is used in the OpenShift.io [6] project for gathering and processing big data about packages in various ecosystems. With Selinon, we were able to create complex task flows that processed large amount of data stored in different database systems based on requirements. The configuration-centric design of loosely coupled components allowed us to perform system checks before actual deployment and gave us an ability to create easy to monitor dynamic system which behaviour can be changed even in online production systems without affecting already queued workload.

References

1. Selinon project homepage, <https://selinon.readthedocs.io/>. Last accessed 28 May 2018
2. Selinon project on GitHub, <https://github.com/selinon/>. Last accessed 28 May 2018
3. Celery project homepage, <https://celeryproject.org>. Last accessed 28 May 2018
4. Celery project documentation, docs.celeryproject.org/. Last accessed 28 May 2018
5. The Official YAML Web Site, <http://yaml.org/>. Last accessed 28 May 2018
6. OpenShift.io: An Open Online Development Environment, <https://www.openshift.io>. Last accessed 28 May 2018
7. Apache Airflow documentation, <https://airflow.apache.org>. Last accessed 28 May 2018
8. Spotify Luigi project on GitHub, <https://github.com/spotify/luigi>. Last accessed 28 May 2018
9. Sentry: Error Tracking Software, <https://sentry.io>. Last accessed 28 May 2018

Znalostné prístupy pre inteligentnú analýzu veľkých dát

Ján Paralič, Peter Bednár, Kristína Machová, Peter Butka, František Babič,
Martin Sarnovský, Vladimír Gašpar

¹Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach, Letná 9/B, 042 00 Košice, Slovenská republika
{jan.paralic, peter.bednar, kristina.machova, peter.butka,
frantisek.babic, martin.sarnovsky, vladimir.gaspar}@tuke.sk

Abstrakt. Tento príspevok prezentuje základné východiská a ciele v minulom roku začatého projektu APVV riešeného v spolupráci výskumného tímu z FEI TU v Košiciach a FIIT STU v Bratislave. Projekt sa zameriava na výskum a vývoj inteligentných metód umožňujúcich spracovanie a analýzu veľkých dát v širšom znalostnom kontexte, ktorý popisuje a charakterizuje takéto prostredie. Ide nielen o veľké dáta statického charakteru, ale aj dynamicky sa meniace prúdy dát v čase, ktoré môžu byť reprezentované dátami z prevádzky rôznych druhov kritických sieťových infraštruktúr, dátami zo sociálnych sietí alebo z rôznych sietí senzorov a inteligentných zariadení. Hlavným cieľom projektu je návrh a verifikácia nových adaptívnych metód pre analýzu veľkých dát v dynamicky sa meniacom prostredí, ktoré budú schopné extrahovať nové znalosti a integrovať ich so znalostným modelom prostredia.

Kľúčové slova: veľké dáta, analýza veľkých dát, dynamické znalostné modely, predikcia, zhľukovanie, prúdy dát.

1 Úvod

V 4-ročnom projekte s názvom „Znalostné prístupy pre inteligentnú analýzu veľkých dát“ (APVV-16-0213) financovanom Agentúrou na podporu výskumu a vývoja (APVV) Ministerstva školstva Slovenskej republiky sme sa rozhodli venovať veľkým dátam dynamického ale aj statického charakteru. Prvý smer reprezentuje oblasť veľkých dát prúdového charakteru [4], t.j. spracovanie a analýza kontinuálne prebiehajúceho a často veľmi rýchleho toku dát väčšinou heterogénneho charakteru. Príkladom môžu byť dáta z prevádzky rôznych druhov kritických sieťových infraštruktúr [7], dáta reprezentujúce vyhľadávania na internete [2], dáta zo sociálnych sietí [6] alebo dáta z rôznych senzorov [8]. Druhý smer reprezentujú veľké objemy statických dát charakterizujúce nejaký stav ako napr. dáta z oblasti bioinformatiky.

Jednou z aplikačných domén, na ktorej bude navrhnutý prístup experimentálne testovaný, predstavuje aj medicínska diagnostika a s ňou spojená podpora rozhodovania [11], [12]. Cieľom je do štandardného analytického procesu efektívnejšie integrovať znalosti expertov a informácie reprezentujúce stav alebo situáciu, ktoré boli zdrojom

spracovaných dát. Inšpiráciou sú napr. práce výskumných kolektívov prof. Andreea Holzingera z Rakúska [9] alebo prof. Wila van der Aalsta z Holandska [10].

Dôležitým faktorom pri spracovaní prúdových dát je prekonanie značnej dátovej heterogenity v prípade, že je potrebné korelovať udalosti z viacerých zdrojov [5]. Ako vhodnou metódou sa ukazuje použitie sémantických modelov založených na ontológiách, ktoré slúžia na harmonizáciu dát a na sémantické interpretovanie informácií obsiahnutých v udalostiach. Výhodou ontológií je, že umožňujú jednotne modelovať jednak význam jednotlivých udalostí a zároveň reprezentovať kontext, v ktorom sa udalosti vyskytujú (napr. pri aplikáciách Internetu vecí je pomocou ontológií možné interpretovať jednak namerané dáta, ale aj sémanticky reprezentovať topológiu senzorickej siete a ďalšie znalosti o meranom prostredí). Okrem konceptuálneho modelovania znalostí, sémantické technológie ponúkajú aj možnosť definovať bázu znalostí v podobe inferenčných pravidiel, pomocou ktorých je možné odvodiť zo známych informácií nové fakty, resp. použiť ich na detekciu zaujímavých udalostí. Súčasný výskum sa zameriava hlavne na spojenie sémantických technológií a metód pre komplexné spracovanie udalostí (tzv. SCEP – Semantic Complex Event Processing) [3].

V našom projekte sa zameriavame na také úlohy, kedy spracovanie a analýza veľkých dát (či už prúdových, statických, resp. Smart Data) je postavené na širšom znalostnom kontexte, ktorý popisuje prostredie, v ktorom sú dáta generované, detegované a využívané pre rozhodovanie [1]. Hlavným cieľom projektu je návrh a verifikácia nových adaptívnych metód pre analýzu veľkých dát v dynamicky sa meniacom prostredí, ktoré by boli schopné extrahovať nové znalosti a integrovať ich so znalostným modelom prostredia. Interakcia medzi vstupnými dátami a prostredím zároveň determinuje aj dekompozíciu hlavného problému na nasledujúce čiastkové výskumné úlohy:

1. Návrh dynamických znalostných modelov, ktoré budú reprezentovať kľúčové charakteristiky a závislosti skúmaných prostredí typických pre riešené výskumné úlohy.
2. Návrh a verifikácia adaptívnych inkrementálnych metód pre predikciu udalostí na základe rozšíreného znalostného kontextu.
3. Návrh a verifikácia metód pre zhlukovanie, popisnú analýzu a vizualizáciu, ktoré umožnia hľadanie nových kauzálnych vzťahov v dátach na základe príslušného dynamického znalostného modelu.

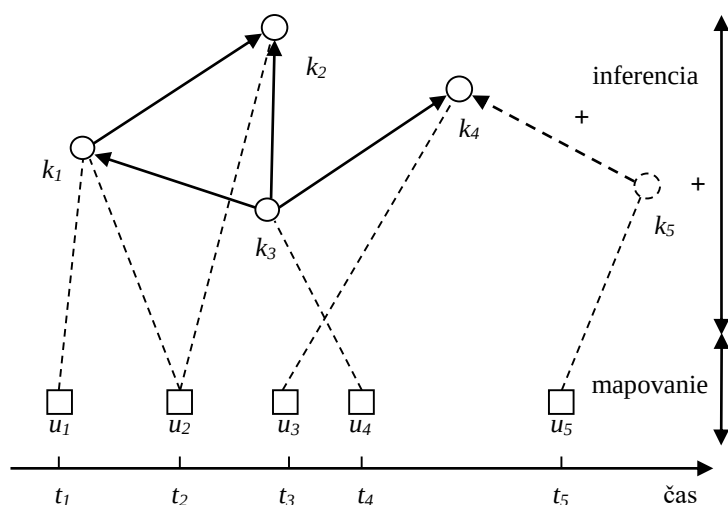
2 Navrhovaný koncept riešenia

Pre popis navrhovaného riešenia sme zaviedli pojem *dynamického znalostného kontextu*, ktorý všeobecne popisuje proces generovania dát v skúmanom prostredí. Pri spracovaní dát môžu mať samotné dáta statický alebo prúdový charakter, ale vo všeobecnosti budeme uvažovať, že sú generované v dynamickom kontexte.

Formálne je možné definovať dynamický znalostný kontext ako neohraničenú postupnosť diskretných udalostí $u_1, u_2, \dots$ ktoré sa vyskytli v čase $t_1, t_2, \dots$. Keďže aj spojité veličiny (napr. teplota, spotreba elektrickej energie a pod.) sa merajú diskretné pre zvolenú vzorkovaciu frekvenciu, dáta majú vždy diskretný charakter. Vo všeobecnosti sa nové udalosti vyskytujú asynchrónne, t.j. rozdiely medzi časmi $t_1, t_2, \dots$ nie sú konštantné. Každá z udalostí je popísaná vlastnosťami a nad postupnosťou udalostí a ich

vlastností je možné definovať zložité vzory, ktoré klasifikujú udalosti do rôznych typov (napr. rôzne diagnostické chyby, rôzne aktivity používateľov a pod.), resp. ktoré vymedzujú v dátach časové okná, nad ktorými je možné vypočítať agregácie. Kontexty je možné jednoducho spájať a vo vzoroch korelovať udalosti rôznych typov. Akciou po detegovaní vzoru môže byť napr. vytvorenie novej udalosti smerovanej do odvodeného kontextu, t.j. znalostné kontexty je možné hierarchicky usporiadať.

Hlavným problémom súčasných systémov je to, že detegovanie vzorov je závislé od kontextu jednotlivých udalostí, pričom do kontextu sú zahrnuté iba predchádzajúce udalosti a ich vlastnosti. Pri praktických aplikáciách je však potrebné kontext udalostí rozšíriť aj o ďalšie znalosti. Aj keď existuje viacero reprezentácií znalostí, medzi najflexibilnejšie patrí reprezentácia v podobe orientovaných hranovo a uzlovo ohodnotených grafov (tzv. sémantických sietí). Jednotlivé vrcholy grafu $k_1, k_2, \dots, k_n$ a ich vlastnosti zodpovedajú konceptom modelovanej domény (uzly môžu reprezentovať napr. profily používateľov, senzory, prvky monitorovanej sústavy a pod.) a hrany zodpovedajú sémantickým vzťahom medzi konceptmi (napr. sociálnym vzťahom medzi používateľmi, topologickým vzťahom v senzorickej sieti, a pod.). Pri detegovaní vzorov teda dochádza k interakcii a) informácií, ktoré sú súčasťou samotnej udalosti, b) informácií o predchádzajúcich udalostiach s c) doménovými znalosťami v znalostnom modeli. Tieto interakcie sú zachytené na nasledujúcom obrázku (Obr. 1).



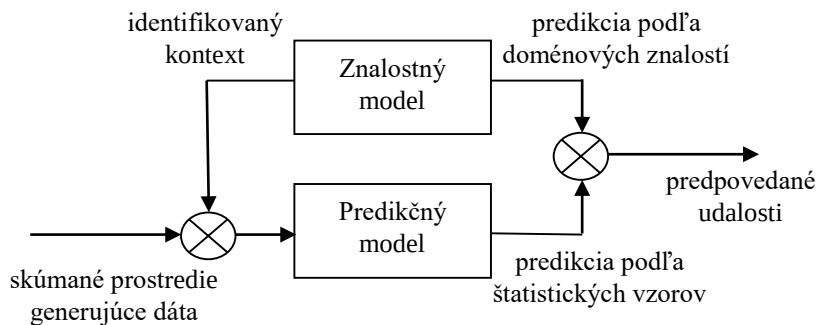
Obr. 1. Interakcia medzi udalosťami a znalostným modelom

Jedným z problémov pri navrhovanom koncepte riešenia je mapovanie dát jednotlivých udalostí na znalostný model. Pri tomto probléme je potrebné riešiť problematiku heterogénnych dát a extrahovania informácií z neštruktúrovaných dát ako napr. z textov v prirodzenom jazyku. Ďalšou dôležitou vlastnosťou nášho návrhu je to, že znalostný

model sa môže v čase meniť, t.j. napr. po detegovaní udalosti u_5 je pridaný do znalostného modelu nový koncept k_5 a nová relácia, ktorá ho spája s už existujúcim konceptom k_4 . Keďže detegovanie udalostí využíva rozšírený kontext a inferenciu v znalostnom modeli, dynamická zmena znalostného modelu sa môže prejavovať ako dynamická adaptácia vzorov pre detegovanie nasledujúcich udalostí. Aby sa dosiahla potrebná škálovateľnosť nášho návrhu, bude potrebné distribuovať nielen spracovanie dát ale aj inferenciu v samotnom znalostnom modeli.

Uvedený konceptuálny návrh projektu ďalej originálnym spôsobom rozširuje aktuálny stav výskumu v nasledujúcich aspektoch. Súčasný výskum v oblasti sémantického komplexného spracovania udalostí (SCEP) je zameraný na ontologické modely založené na výrokovej logike. Náš konceptuálny návrh zahŕňa všeobecnejšiu reprezentáciu znalostí - sémantické siete, ktoré umožňujú aj iné typy inferencie ako napr. pravdepodobnostné metódy, alebo využitie fuzzy logiky. Výhodou je podpora uvažovania pri neurčitých informáciách, ktoré často vznikajú napr. pri extrahovaní informácií z neštruktúrovaných dát, pri spracovaní dát zo sociálnych sietí a v zašumenom prostredí.

Výskum v oblasti učenia predikčných modelov je založený na predpoklade, že všetky informácie, ktoré ovplyvňujú budúce udalosti, sú obsiahnuté v aktuálnom kontexte prostredia. Navrhovaný konceptuálny model umožní a) rozšíriť kontext vstupných udalostí o nové príznaky odvodené zo znalostného modelu a b) rozšíriť predikciu pravidlami, ktoré sú získané na základe explicitných doménových znalostí vyjadrených v znalostnom modeli. Tieto doménové pravidlá môžu napr. pokryť detekciu výnimočných prípadov, ktoré by bolo náročné štatisticky odvodiť z histórie skúmaného prostredia. Interakcia medzi predikčným štatistickým modelom a znalostným modelom je znázornená na nasledujúcej schéme (Obr. 2).



Obr. 2. Schéma využitia znalostného modelu pri predikcii

Podobne ako pre predikčné modely, pri popisnej analýze dát rozšíri znalostný model kontext príslušných dát t.j. napr. popisné metódy pre zhľukovanie môžu do popisu zhľukov zahrnúť všetky relácie a kauzálne vzťahy odvodené zo znalostného modelu.

3 Záver

V tomto článku bol predstavený hlavný cieľ a základný koncept riešenia projektu APVV riešeného v spolupráci s FIIT STU. Prvé výsledky riešenia projektu a orientácia ďalšieho výskumu za skupinu z FIIT STU je prezentovaná v ďalšom, samostatnom príspevku. V nasledujúcom období sa zameriame na integráciu pojmov z doménových a dátovo-analytických ontológií a na návrh jednotného sémantického rámca ktorý by umožňoval jednoduchú rozšíriteľnosť na ďalšie domény. Budeme sa tiež venovať návrhu a implementácii znalostných modelov z viacerých domén, ktoré využijeme pre presnenie adaptívnych klasifikačných, regresných a popisných metód dolovania v dátach.

Literatúra

1. Bednár, P.: Vocabulary matching for information extraction language. In: SAMI 2017. Danvers, IEEE (2017) 149-152.
2. Butka, P., Low, T., Kotzyba, M., Haun, S., Nurnberger, A.: Exploration of Web Search Results Based on the Formal Concept Analysis. In: Semantic Keyword-Based Search on Structured Data Sources. Springer (2018), 123-135.
3. Keskiärrä, R.: Semantic Complex Event Processing for Decision Support. ISWC'14 Proc. of 13th Int. Semantic Web Conference - Part II, Springer, New York (2014), 529-536.
4. Sarnovský, M., Bednár, P., Smatana: Big Data Processing and Analytics Platform Architecture for Process Industry Factories. In: Big Data and Cognitive Computing. Vol. 2, no. 1 (2018), 3-15.
5. Sarnovský, M., Bednár, P., Smatana, M.: Data integration in scalable data analytics platform for process industries. In: INES 2017. Danvers: IEEE (2017), 187-191.
6. Sarnovský, M., Butka, P., Paulina, J.: Social-media data analysis using tessera framework in the hadoop cluster environment. In: Advances in Intelligent Systems and Computing: ISAT 2016. Springer, Vol. 522 (2017) 239-251.
7. Šátala, P., Gašpar, V., Butka, P.: Using IoT devices for movement detection in medical environment - proof of concept. In: SAMI 2018. Danvers, IEEE (2018), 61-65
8. Šátala, P., Gašpar, V., Butka, P.: Chytré zariadenia v priemysle (1) až (4). Série článkov v časopise ATP Journal. č. 1 až 4 (2018).
9. Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Springer Brain Informatics (BRIN), 3, (2), 119-131
10. P. M. Dixit ; J. C. A. M. Buijs ; W. M. P. van der Aalst. 2018. ProDiGy: Human-in-the-loop process discovery. 12th International Conference on Research Challenges in Information Science (RCIS), Nantes, France
11. Šabanović Š., Ljiljana MT., Babič F., Vadovský M., Paralič J., Včev A., Holzinger A.: Metabolic syndrome in hypertensive women in the age of menopause: a case study on data from general practice electronic health records. BMC Med Inform Decis Mak. 2018 Apr 2;18(1):24.
12. Babič F., Vadovský M., Paralič J.: Medical data analytics is not a simple task. In: MIE 2018. Amsterdam: IOS Press, 2018 p. 371-375

PodĎakovanie: Tento článok vznikol vďaka podpore Agentúry na podporu výskumu a vývoja na základe Zmluvy č. APVV-16-0213.

Annotation:*Knowledge-based approaches for intelligent analysis of big data*

This paper presents basic ideas of a Slovak research project focused on research and development of intelligent methods enabling the processing and analysis of big data in the broader knowledge context, which describes and characterizes such an environment. We focus not only on big data with static character, but also on dynamically changing data streams, which can be represented by the operation data from the different types of critical network infrastructure, data from social networks or networks of different sensors and smart devices. The main objective of the project is to design and verify new adaptive methods for analyzing big data in a dynamic environment, able to extract new knowledge and to integrate them with the knowledge model of the environment.

REBELION – odhaľovanie antisociálneho správania na Webe

Pavol Návrat¹, Daniela Chudá¹, Mária Bieliková¹, Marián Šimko¹, Jakub Šimko¹,
Ivan Srba¹, Michal Kompan¹, Róbert Móro¹, Irina Malkin Ondik¹, Peter Lacko¹,
Alena Martonová¹, Kristína Machová², Ján Paralič², Peter Butka², Peter Bednár², Mar-
tin Sarnovský², Barbora Mesárošová³, Radoslav Blaho³, Lucia Sabová³

¹Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave

Ilkovičova 2, 842 16 Bratislava

{meno}. {priezvisko}@stuba.sk

²Katedra kybernetiky a umelej inteligencie, FEI, Technická univerzita v Košiciach

Letná 9, 04200 Košice

{meno}. {priezvisko}@tuke.sk

³Katedra psychológie, FiF, Univerzita Komenského v Bratislave

Šafárikovo námestie 6, 81499 Bratislava

{meno}. {priezvisko}@uniba.sk

Abstrakt. V tomto príspevku opisujeme začínajúci výskumný projekt REBELION, v ktorom sa riešiteľský kolektív zameriava na výskum nových modelov a metód pre automatizované rozpoznávanie antisociálneho správania v online komunitách. Antisociálne správanie v online prostredí patrí medzi najaktuálnejšie a najvážnejšie problémy, ktoré výrazným spôsobom ohrozujú nielen princípy, na ktorých bol web vybudovaný, ale majú aj kritický presah do fungovania spoločnosti. Typickým príkladom je šírenie falošných informácií prostredníctvom sociálnych sietí, ktoré vplývajú na budovanie názorov ľudí a ovplyvňujú tak ich následné rozhodnutia, či šírenie neznášanlivých postojov, prípadne priame útočenie na osoby v online priestore. Tieto negatívne situácie nechtiac umožnil rozvoj informačných technológií.

Kľúčové slová: antisociálne správanie, online komunity, sociálny web, strojové učenie, spracovanie prirodzeného jazyka, umelá inteligencia.

1 Úvod

Súčasný web môžeme charakterizovať ako priestor, kde každý môže voľne a takmer bez obmedzení vytvárať, šíriť a prijímať informácie. Zatiaľ čo väčšina používateľov webu sa snaží udržať web bezpečný a účelný, existuje skupina používateľov, ktorí vykazujú tzv. antisociálne správanie. Antisociálne správanie môžeme na najvyššej úrovni rozdeliť do dvoch foriem [1]: šírenie dezinformácií (napr. hoax alebo falošné správy) a reakcie antisociálnych používateľov (napr. hejtovanie, manipulovanie diskusií alebo kyberšikana).

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 65-69.*

Následky a rozsah antisociálneho správania nie sú zanedbateľné, práve naopak v niektorých oblastiach sú až alarmujúce.

- Dezinformácie na webe majú výrazný presah do reálneho sveta. Je to možné ilustrovať na vytváraní osobných názorov a rozhodovaní počas politických volieb. Až 61% ľudí narodených medzi rokmi 1977 a 1995 považuje sociálnu sieť Facebook za primárny zdroj politických správ [2], pričom v prípade amerických prezidentských volieb v roku 2016 počet falošných správ šírených touto sociálnou sieťou prekonal počet pravdivých správ [3].
- Dezinformácie slúžia na účelovú manipuláciu ľuďmi. Ako príklad môžeme uviesť štúdiu [4] na systéme Yelp, ktorá ukázala, že až 16% recenzií je falošných.
- Vysoký výskyt používateľov šíriacich nevhodný obsah degraduje úroveň diskusie a odrádza ostatných používateľov od zapájania sa do diskusií a zákonite aj od ďalšieho navštevovania online portálov/komunit.
- Až 40% používateľov webu zažilo obťažovanie v online priestore¹.
- Až 18% európskej mládeže bolo niekedy v ich živote obeťou kyberšikany, čo v mnohých prípadoch viedlo k silným depresiám a dokonca pokusom o samovraždu².

V svetovom meradle bol prijatý konsenzus, že na antisociálne správanie v online priestore nie je možné naďalej reagovať len ignorovaním, ktoré donedávna predstavovalo najčastejšiu reakciu na prejavy antisociálneho správania. V rámci práve začínajúceho projektu REBELION nasledujeme túto motiváciu s cieľom prispieť do existujúceho výskumu zameraného na detekciu a prevenciu antisociálneho správania na Webe.

2 Súčasný stav poznania

Napriek vážnym následkom a širokému rozsahu antisociálneho správania súčasný stav poznania neposkytuje efektívne riešenia, ako regulovať a eliminovať antisociálne správanie. Pomerne priamočiara je snaha redukovat' antisociálne správanie prostredníctvom využívania moderátorov. V mnohých komunitách sa navyše využíva čerpanie z davu - moderátorom pomáhajú ostatní členovia komunity tým, že označujú podľa nich nevhodné príspevky alebo používateľov. Takáto ľudská "manuálna" kontrola však naráža na prirodzené obmedzenia, medzi ktoré patrí predovšetkým subjektívnosť a nemožnosť škálovať na množstvo obsahu vytváraného ľuďmi v rámci online komunit. Prípadová štúdia v CQA systéme Yahoo! Answers [5] ukázala, že len 50% používateľov, ktorých príspevky boli nahlásené ako nevhodné, bolo moderátormi nakoniec skutočne vylúčených. Navyše až 40% z vylúčených používateľov nebolo nahlásených komunitou ani raz. Svetový výskum preto hľadá odpoveď, ako dokážeme proces regulácie a eliminácie antisociálneho správania podporiť informačnými technológiami. Aktuálne výskumné

¹ Online harassment, pew research center. <http://www.pewinternet.org/2014/10/22/online-harassment>, 2014.

² EU COST Action IS0801 on Cyberbullying (<https://sites.google.com/site/costis0801/>)

výsledky ukazujú, že s využitím metód a techník umelej inteligencie, špeciálne strojového učenia, je možné vylepšiť úspešnosť manuálnej detekcie a v niektorých prípadoch ju dokonca prekonať [6].

Pre polozenie silných základov projektu REBELION sme vykonali rozsiahlu prehľadovú štúdiu, v rámci ktorej sme vyhľadali a systematicky zakategorizovali 77 výskumných článkov zaoberajúcich sa antisociálnym správaním. Prehľadová štúdia ukázala enormný nárast záujmu výskumníkov o túto oblasť v roku 2017, v ktorom sa počet publikácií oproti roku 2016 zvýšil trojnásobne, pričom rovnaký trend pokračuje aj v roku 2018.

Definície niektorých typov antisociálneho správania sa často prekrývajú a jeden používateľ môže naraz preukazovať znaky viacerých z nich. Celkovo 44 príspevkov sa zameriavalo na šírenie dezinformácií:

- Falošné a skreslené správy (angl. *fake and biased news*, 32 publikácií).
- Falošné recenzie (angl. *fake reviews*, 7 publikácií).
- Hoaxy (5 publikácií).

Približne rovnaký počet publikácií (39) sa zameriaval na reakcie používateľov.

- Trolovanie (angl. *trolling*, 13 publikácií).
- Manipulovanie diskusií (angl. *sockpuppeting*, 6 publikácií).
- Zneužívanie komunity (angl. *community abuse*, 5 publikácií).
- Vandalizmus (angl. *vandalism*, 4 publikácie).
- Záškodnícke správanie hráčov online hier (angl. *griefing*, 4 publikácie).
- Kyberšikana (angl. *cyberbullying*, 4 publikácie).
- Hejtovanie (angl. *hating*, 2 publikácie).
- Spamovanie (angl. *spamming*, 1 publikácia).

Antisociálne správanie je študované v celom spektre online komunít, predovšetkým sú to sociálne siete (24 publikácií), novinové portály (13 publikácií), systémy pre odpovedanie na otázky (angl. *Community Question Answering* - CQA, 6 publikácií), online hry (6 publikácií), online encyklopédie (6 publikácií) a diskusné fóra (4 publikácie).

3 Inovativnosť projektu a použitých metód

Na základe podrobnej analýzy súčasného stavu poznania sme identifikovali otvorené problémy, ktoré charakterizujú aktuálny stav výskumu antisociálneho správania a jeho automatického rozpoznávania. Tieto otvorené problémy sú zároveň východiskom pre smerovanie a zabezpečenie inovativnosti riešeného projektu.

1. Metódy strojového učenia sú aplikované bez hlbšieho porozumenia skúmaného antisociálneho správania. Napriek silne multidisciplinárnemu charakteru oblasti, existujúce metódy pre detekciu a predikciu antisociálneho správania sú vo veľkej miere dátovo orientované a vychádzajú výlučne z poznatkov oblasti informačných technológií. Nevyužívajú poznatky zo súvisiacich odborov, najmä z oblasti psychológie.

2. Nevyužívanie celého spektra dostupných informácií o obsahu a používateľoch. Zapojenie viacerých zdrojov dát pritom môže významne prispieť k úspešnosti detekcie a predikcie antisociálneho správania.
3. Absencia multiplatformovej detekcie antisociálneho správania. Súčasné metódy sa zameriavajú takmer vždy na skúmanie antisociálneho správania len v jednom type komunity a navyše len v rámci jedného systému. Pri viacerých typoch antisociálneho správania je však možné stavať na prenosnosti naučených modelov.
4. Potreba skoršej detekcie/predikcie. Ako vyplynulo z našej prehľadovej štúdie, len minimálny počet príspevkov sa zameriava na skorú detekciu antisociálneho správania (t.j. detekcia antisociálneho správania používateľa už v jeho začiatkoch). Takáto skorá detekcia/predikcia má pritom veľmi praktické využitie, keďže môže výrazne napomôcť redukcii negatívnych dopadov antisociálneho správania.

Hlavným cieľom projektu je priniesť nové poznatky v oblasti informatických vied a v informačných technológiách, najmä:

1. S využitím analýzy dát a kvalitatívnych prístupov **vyskúmať do hĺbky fenomény antisociálneho správania ľudí v online komunitách** a priniesť nové poznatky spojené s takýmto správaním sa človeka v online priestore; predmetom skúmania budú v menšej miere procesy, ako toto malígne správanie vzniká (porozumenie dôvodom vzniku fenoménu pomôže v konečnom dôsledku aj jeho identifikácii); vo väčšej miere budeme skúmať prejavy antisociálneho správania, na ktoré sa chceme pozrieť ľudským pohľadom s ambíciou k nemu priblížiť “pohľad strojový”.
2. S využitím dolovania v dátach navrhnuť a verifikovať **nové modely, ktoré reprezentujú dátový pohľad na antisociálne správanie ľudí** v online priestore; modely budú vyplývať z dvoch zdrojov: (i) dátami vedenej kvantitatívnej analýzy a (ii) antropocentrickej kvalitatívnej analýzy (cieľ 1); tieto zdroje, odlišné svojím charakterom, chceme na úrovni modelov prepojiť.
3. S využitím strojového učenia navrhnuť a **verifikovať nové metódy umelej inteligencie pre automatizovanú detekciu a predikciu** rôznych typov antisociálneho správania a návrh využitia ich výsledkov pre prevenciu a elimináciu negatívnych dopadov antisociálneho správania v online komunitách. Využíjúc získané modely (cieľ 2) vytvoríme metódy na báze strojového učenia, schopné automatizovanej detekcie antisociálneho správania, pričom cieľ dosiahneme predovšetkým prepracovaným návrhom črt.

Projekt orientujeme na automatizovanú detekciu a predikciu antisociálneho správania sa človeka v online priestore predovšetkým po stránke:

- **analýzy dát** (kvantitatívnymi a kvalitatívnymi prístupmi informatických vied a psychológie), ako aj po stránke
- **technologickej**, kde sa do popredia dostávajú nové prístupy a technológie v umelej inteligencii, špeciálne strojovom učení (napr. neurónové siete a využitie hlbokého učenia, spracovanie prirodzeného jazyka, modelovanie používateľa), ktoré vhodne doplníme metódami založenými na sledovaní pohľadu.

4 Závěr

Pokým antisociálně správání nechtiac umožnil rozvoj informačných technológií, tak informačné technológie sa môžu stať aj prostriedkom pre lepšie porozumenie a vysporiadanie sa s týmto fenoménom. Dosiahnutie tohto cieľa však vyžaduje multidisciplinárny výskum, ktorý adresujeme v projekte REBELION. Primárnym zameraním projektu je výskum modelov a metód pre automatizované rozpoznávanie antisociálneho správania. Explicitné modely postavíme na dátovej analýze ale aj na kvalitatívnom porozumení malígneho obsahu a správania. Metódy rozpoznávania budú na týchto modeloch stáť a budú sa opierať o princípy strojového učenia, pokročilú automatickú analýzu textov v prirodzenom jazyku a modelovanie používateľa.

Literatúra

1. Kumar, S., Cheng, J., & Leskovec, J. (2017). Antisocial Behavior on the Web: Characterization and Detection. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 947–950).
2. Mitchell, A., Gottfried, J., and Matsa, K.E. Facebook top source for political news among millennials. Pew Research Center, June 1, 2015;
3. Silverman, C. This analysis shows how viral fake election news stories outperformed real news on Facebook. BuzzFeed (Nov. 16, 2016);
4. Luca, M., & Zervas, G. (2016). Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12), 3412–3427.
5. Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., & Bonchi, F. (2015). The Social World of Content Abusers in Community Question Answering. In Proceedings of the 24th International Conference on World Wide Web - WWW '15 (pp. 570–580).
6. Mihalcea, R. (2017). Deception Detection: When Computers Become Better than Humans. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management - CIKM '17 (pp. 3–3). New York, New York, USA: ACM Press.

PodĎakovanie: Tento článok vznikol vďaka podpore projektu APVV-17-0267.

Annotation:

REBELION - Recognition of Antisocial Behaviour on the Web

In this paper, we describe the starting project REBELION, which is focused on research of new models and methods for automatic recognition of antisocial behavior in online communities. Antisocial behavior in online environment is one of the most actual and serious problems, which significantly threatens not only the principles, on which the web was built, but also has critical overreach to society. As a typical example, we can take the spread of misinformation through social networks, which influences opinions of people and consequently, their decisions, spread of hate speech or direct attacks on persons in online space. These negative situations have been unintentionally enabled by the rise of information technologies.

Bioinformatika - Výzvy a príležitosti v proteomike

Ján Genči

Katedra počítačov a informatiky, Technická univerzita v Košiciach
Letná 9, Košice
genci@tuke.sk

Abstrakt. V oblasti bioinformatiky, popri genomike je asi najzaujímavejšia oblasť proteomika. Kým genomika predstavuje súbor všetkých možných ciest vývoja organizmu, proteomika charakterizuje jeho aktuálny stav – súbor aktuálnych proteínov (proteóm) určuje, čo sa v organizme deje, resp. čo sa bude diať v budúcnosti.

Proteíny, podobne ako genóm, majú „digitálnu podstatu“ (20 základných aminokyselín). Mapovanie procesov prebiehajúcich v organizme preto predstavuje výzvu aj v oblasti spracovania dát. V prvom rade ide o identifikáciu proteínov (aktuálneho proteómu v danom čase). Následne predstavuje výzvu pochopenie mechanizmov fungovania proteómu (tzv. protein-protein interactions, funkčná proteomika). Cieľom príspevku je naznačiť niektoré výzvy a príležitosti vyplývajúce z proteomiky pre oblasť informatiky a budovania tzv. informačných modelov, ktoré umožnia pochopenie fungovania živých organizmov.

Kľúčová slova: bioinformatika, proteomika, spracovanie proteomických dát, informačné modely v proteomike

1 Úvod

Objav štruktúry DNA¹ preukázal digitálnu podstatu biologického života. Metódy sekvenovania DNA [1] prvotne vyvinuté v projekte mapovania ľudského genómu [2] by bez počítačového spracovania boli neuskutočniteľné. Následné podstatné zdokonalenie metód sekvenovania viedlo k nesmiernej popularite genomiky aj v oblasti spracovania (genomických) dát. Popri genomike sa však rozvíjali aj ďalšie oblasti spracovania - tzv. ‘omics’² prístupy (metabolomics, lipidomics, ...), medzi ktorými je asi najzaujímavejšia proteomika. Kým genomika predstavuje súbor všetkých možných ciest vývoja organizmu, proteomika charakterizuje jeho aktuálny stav – súbor aktuálnych proteínov (proteóm) určuje, čo sa v organizme deje, resp. čo sa bude diať v budúcnosti.

¹ https://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/index.html

² <https://www.nature.com/omics/about/index.html>

Predpokladá sa, že ľudský proteóm sa skladá z cca milióna proteínov. Úlohou týchto proteínov je napr. zabezpečiť formovanie organizmu, zabezpečiť jeho fungovanie (komunikácia), riadiť tzv. expresiu génov.

Proteíny, podobne ako genóm, majú „digitálnu podstatu“ (20 základných aminokyselín) [5]. Mapovanie vyššie uvedených procesov preto predstavuje výzvu v oblasti širokej škály spracovania dát. V prvom rade ide o identifikáciu proteínov (aktuálneho proteómu v danom čase). Následne predstavuje výzvu pochopenia mechanizmov fungovania proteómu (tzv. protein-protein interactions, funkčná proteomika). V príspevku sa venujeme výzvam a príležitostiam, ktoré vyplývajú z procesov identifikácie proteínov a naznačíme problematiku budovania informačných modelov pre oblasť proteomiky.

2 Identifikácia proteínov

Historicky bolo vyvinutých niekoľko metód identifikácie proteínov [3]. V súčasnosti najrozšírenejšou sa javí použitie hmotnostnej spektrometrie (MS – mass spectrometry) [4], špeciálne upravenej pre potreby proteomiky. V tomto prípade sa skúmaná biologická vzorka najprv biologicky spracuje tak, že sa proteíny, nachádzajúce sa vo vzorke, enzymaticky rozštiepia na menšie časti (peptidy) a tieto sa následne prostredníctvom kvapalinovej chromatografie vhodne separujú. Takto vyseparované biomolekuly (peptidy) postupne vstupujú do hmotnostného spektrometra, ktorý najprv biomolekuly ionizuje a následne na báze fyzikálneho princípu identifikuje ich intenzitu pre každý pomer hmotnosti a náboja biomolekuly (tzv. m/z ratio). Táto fáza identifikácie, nazývaná aj fáza MS1, môže byť doplnená o fázu identifikácie štruktúry biomolekuly (fáza MS2) – rozrušovaním väzieb medzi aminokyselinami a následnou identifikáciou pomeru m/z pre jednotlivé novozískané fragmenty. Na základe takto získaných dát dochádza následným softvérovým spracovaním k pokusu o identifikáciu konkrétnych peptidov a v ďalšom kroku, na báze dostupných proteomických databáz, k pokusu o identifikáciu konkrétnych proteínov, spolu s určením ich koncentrácie. Biológovia, na základe zmeny proteómu resp. koncentrácie jednotlivých proteínov vo vzorkách skupín pacientov sa snažia určiť preteomické príznaky rôznych ochorení.

2.1 Problémy pri identifikácii proteínov MS metódami

Uvedený postup spracovania nesie v sebe celý rad problémov. Úplne prvým je nedeterminickosť procesu merania - niektoré molekuly môžu ostať pre MS

neviditeľné. Z uvedeného dôvodu sa každá vzorka meria opakovane a z nameraných hodnôt je potrebné zrekonštruovať štatistickú reprezentáciu vzorky. Ďalší problém - v procese merania sa nemusí podariť určiť štruktúru identifikovaného peptidu, či na základe identifikovaného peptidu nie je možné určiť o aký proteín sa jedná (napr. preto, lebo daný proteín sa nenachádza v dostupnej proteomickej databáze). To všetko vedie k tomu, že biologická interpretácia získaných výsledkov môže byť veľmi obtiažna.

Celý proces navyše komplikuje fakt, že počet dostupných vzoriek (pacientov s daným ochorením) a aj množstvo odobraného biologického materiálu (teda počet možných opakovaných meraní jednej vzorky) je veľmi limitované.

2.2 Výzvy a príležitosti

Spracovanie vyššie uvedených dát predstavuje celý rad výziev, ale aj príležitostí na rozpracovanie nových prístupov. Prvou výzvou je pokus o nahradenie klasického štatistického spracovania získaných dát napr. metódami strojového učenia. Základným problémom v tomto prípade je dostupné množstvo vzoriek (počet pacientov s danou diagnózou je limitovaný, obmedzené možnosti spracovania). Z druhej strany, množstvo dát získaných z jedného merania je veľmi veľké – môže sa jednať o desiatky, či stovky tisíc identifikovaných (zaznamenaných) biomolekúl.

Ďalšou výzvou, z nášho pohľadu, je zmeniť „uhol pohľadu“ biológov na získané dáta. Biovedci potrebujú identifikovať proteín, aby mohli urobiť akceptovateľné závery. Myslíme si však, že niektoré závery je možné urobiť aj na základe vzorov identifikovaných na báze informácií o peptidoch, či dokonca na základe prvotných informácií o identifikovaných biomolekulách.

3 Budovanie informačných modelov

Skúmaniu podstaty biologického života sa venuje veľké množstvo laboratórií a v nich ešte väčší počet vedeckých tímov. Skúmajú sa rôzne aspekty problematiky s dopadmi na rôzne oblasti života (v prípade človeka to je napr. medicína). Väčšina takýchto tímov má ambíciu výsledky svojej práce publikovať vo vedeckých časopisoch, hoci väčšinou ide iba o malý príspevok v oblasti globálneho ľudského poznania. O množstve odvedenej práce môže svedčiť napr. fakt, že len pre pojem rakoviny prsníka (breast cancer) databáza Pub-

med³ eviduje od začiatku roku 2014 takmer 70 tisíc publikovaných článkov, od začiatku roka 2018 do polovice júna je to takmer 10 tisíc článkov.

Je zrejmé, že je prakticky nemožné, aby ktokoľvek prečítal takéto množstvo publikácií a bol schopný informácie v nich uvedené spracovať.

3.1 Výzvy a príležitosti

Hlavnou výzvou v tomto smere je identifikácia vhodných prístupov na budovanie informačných modelov, ktoré by vhodným spôsobom prezentovali podstatné informácie prezentované v jednotlivých článkoch. V prvom rade ide o identifikáciu konkrétnych informácií a znalostí (informačný model), ktoré chceme prezentovať. Následne je potrebné zvoliť vhodný spôsob prezentácie daného informačného modelu tak, aby záujemcom poskytol nimi požadované informácie. Uvedený prístup si vyžaduje predovšetkým dostup k zdrojom, následne extrakciu relevantných informácií z dostupných textov a databáz, spracovanie extrahovaných informácií a naplnenie informačného modelu.

Ako príklad v oblasti proteomiky môžeme uviesť budovanie tzv. PPI (protein-protein interaction) sietí alebo modely identifikujúce proteíny, ktoré ovplyvňujú expresiu určitých génov.

4 Záver

Cieľom príspevku bolo naznačiť výzvy a príležitosti, ktoré stoja pred informatickou komunitou v oblasti spracovania proteomických dát. Hlavný dôraz je v príspevku kladený na spracovanie proteomických dát získaných metódami hmotnostnej spektrometrie, avšak nezanedbateľný prínos môže predstavovať aj budovanie informačných modelov určených na poskytovanie komplexných informácií na báze zverejnených čiastočných poznatkov.

Literatúra

1. James M. Heather, Benjamin Chain: The sequence of sequencers: The history of sequencing DNA, *Genomics*, Volume 107, Issue 1, (2016), ISSN 0888-7543, Pages 1-8
2. Chial, H. DNA sequencing technologies key to the Human Genome Project. *Nature Education* 1(1):219, (2008)

³ <https://www.ncbi.nlm.nih.gov/pubmed/>

3. Kris Gevaert, Joël Vandekerckhove: Protein identification methods in proteomics. ELECTROPHORESIS (2000), Volume21, Issue 6, ISSN: 0173-0835
4. Hoffmann, and Vincent Stroobant. *Mass spectrometry : principles and applications*. Chichester, West Sussex, England Hoboken, NJ: J. Wiley, 2007.
5. Benjamin Lewin: Essential Genes. Pearson Education. 2005

Pod'akovanie: Tento článok vznikol vďaka čiastočnej podpore projektu VEGA 1/0433/16

Annotation:

Bioinformatics - Challenges and opportunities in proteomics

In the area of bioinformatics, besides genomics, perhaps the most interesting area is proteomics is. While genomics represents a set of all possible pathways to the development of an organism, proteomics characterizes its current state - a set of current proteins (proteome) determines what is happening in the organism or what will happen in the near future. Proteins, like the genome, have a "digital essence" (20 basic amino acids). Therefore, identification of processes ongoing in a body is also a challenge in the field of data processing. First of all, we are speaking about identification of proteins (the current proteome at a given time). Consequently, it is challenging to understand protein-protein interactions or functional proteomics. The aim of the paper is to highlight some of the challenges and opportunities arising from proteomics in the field of informatics and building the so called information models, which will allow us to understand functioning of living organisms.

Performing Feature Selection Before Removing Outliers To Increase Classifier's Accuracy

Nataliia Petliak, Oleksandra Tmenova, Matúš Námešný, Tomáš Bončo, and
Luboš Popelínský

KD Lab FI MU
Botanická 68a, 60200 Brno
popel@fi.muni.cz

Abstrakt This work addresses the problem of feature selection for boosting the performance of outlier detectors in the context of supervised classification. Different feature selection and outlier detection methods are applied to four datasets used in the experiment and a comparative analysis between combinations of these methods is reported. We present combinations producing the best accuracy of a classifier and show the optimal number of outliers to be removed.

1 Introduction

Outlier detection is a well-known data mining task that focuses on finding anomalies (an outlier or a series of outliers) in data. The context of the problem we address in this paper is detecting class-based outliers. Class-based outliers are those cases that look anomalous when the class labels are taken into account, but they do not have to be anomalous when the class labels are ignored [3]. Two outlier detection methods that we are testing for finding class-based outliers are RF-OEX [3], based on Random Forests, and CODB, a distance-based approach [1].

Another well-known problem that we are covering in this paper, is feature selection. In case of a supervised learning task, this problem consists in selecting a subset of features from a given set of labeled fixed-length feature vectors, thereby reducing the number of dimensions and improving a learning process, since used data becomes less noisy, that is it doesn't contain irrelevant and/or redundant attributes.

In this paper, we tackle the problem of selecting a subset of features in order to obtain a better performance of a classifier after removing outliers from data. We want to find out whether dimensional reduction helps improving the result of outlier detection and if so, what are the best combinations of feature selection and outlier detection methods one should apply to our problem of interest. In addition, we want to find out which number of outliers it is better to remove to have higher accuracy.

The rest of the article is organized as follows. In section two a brief overview of tools we used is exposed. In section three our methodology is presented. In

section four the experimental results are reported. Finally, in section five the article concludes.

2 Tools

To test the relationships between feature selection, outlier detection and classifier accuracy we chose 4 datasets, 4 feature selection methods, 4 classifiers and 2 outlier detection methods. Since the majority of above-mentioned methods are implemented in WEKA [2], we used it for our analysis.

The 4 datasets were diabetes, optical, quisclas and satimage and were obtained from project MetaL. The diabetes dataset has 768 examples and 8 features. The optical dataset has 5620 examples and 64 features. The quisclas dataset has 5891 examples and 18 features. The satimage dataset has 4435 examples and 36 features.

For feature selection, we used different attribute evaluation methods and different search methods giving us the total of 4 combinations. The first two use CfsSubsetEval as evaluation method and either ScatterSearchV1 or GreedyStepwise as search methods. The second two combinations use either GainRatioAttributeEval or InfoGainAttributeEval as evaluation methods but they both use Ranker as search methods.

The classifiers were chosen to cover a wide variety of classifiers. Therefore we have chosen J48 (a Java implementation of C4.5 decision tree algorithm), Naive Bayes (NB), Support Vector Machine (SMO) and K-nearest Neighbors (IB1).

To detect outliers in our data we used two different outlier detection algorithms. The first algorithm is CODB [1]. It measures the distance to an example's nearest neighbors. The second algorithm is RF-OEX [3] which is based on the random forest.

3 Methodology

3.1 Datasets

As RF-OEX was performing slow on large datasets, we sampled three of our datasets, to have between 700 to 900 instances with approximately the same distribution of the classes as our original datasets. We performed all our tests on sampled quisclas, optical and satimage datasets and on the original diabetes dataset.

3.2 Process

At the beginning, we performed feature selection on four datasets and kept the attributes with the most gain for each dataset. Then we ran RF-OEX and CODB outlier detection methods for each dataset for each feature selection method. Then we removed 1 to 10 most significant outliers from a dataset and classified data using 4 different classifiers - J48 (C4.5), NaiveBayes, SMO, and

IB1. That gave us a result for each combination of a dataset, a feature selection method, an outlier detection method and a number of removed outliers. We also performed these tests without applying feature selection method. After that, we aggregated all these results and started to look at relationships between them. For comparison between CODB and RF-OEX, we compared their best results, their worst results, average ones, as well as their medians.

4 Results

We ran over 1600 tests and received the accuracy of each classifier with every combination of the feature selection method (or none) and outlier detection method, as well as the baseline. The results can be found at <https://www.fi.muni.cz/~popel/644892/>.

Tab. 1. RF-OEX vs CODB

dataset	OD	max	min	avg	median
diabetes	CODB	2.0326	0.0246	0.7636	0.5509
diabetes	RF-OEX	2.7728	0.0759	1.1566	1.2379
quisclas	CODB	2.1992	0.1166	0.634	0.4421
quisclas	RF-OEX	2.9738	0.575	1.2897	1.2897
optical	CODB	3.9111	0.0182	1.7632	1.8346
optical	RF-OEX	4.6674	0.0129	1.8542	2.1798
satimage	CODB	3.0037	0.0405	1.1963	1.2298
satimage	RF-OEX	4.0472	0.2137	1.3926	1.4105

We compared the results of the baseline and two outlier detection methods. Removing outliers from the dataset always increased the accuracy up to 4.6 %. Comparing CODB and RF-OEX (Table 1), the latter gave the best results for every dataset. Applying feature selection methods before removing outliers did not always give higher accuracy compared to removing outliers without previous feature selection. It depended on the dataset and the classifier, as for diabetes and quisclas datasets using RF-OEX without applying feature selection methods provided the best results out of all combinations and increased the accuracy by 2.7728 % and 4.6674 % accordingly. For both optical and satimage datasets InfoGainAttributeEval+Ranker in combination with RF-OEX and J48 classifier boosted the accuracy by 2.9738 % and 4.0472 % respectively.

We tested every classifier 10 times for each of the datasets to find the efficient number of outliers to be removed. We removed outliers from 1 to 10 according to their rank and got the following results. In diabetes dataset removing 10 outliers provided the best results in 3 out of 4 classifiers. The best accuracy for quisclas dataset was achieved by removing 5 outliers, however, 10 was also a sufficient number for IB1 and SMO classifiers. In optical dataset 8 removed outliers increased the accuracy the most in different combinations of FS methods

and J48 and IB1 classifiers. For the satimage dataset removing 9 outliers gave the best result for J48 and for both IB1 and NB 10 was the most optimal number. Then we asked the question what is generally the optimal number of

Tab. 2. Aggregated results

OD	Classifier	Best number of outliers	Avg. gain
CODB	J48	7	0.113 %
CODB	IB1	5	0.0646 %
CODB	NB	10	-0.0108 %
CODB	SMO	5	-0.0254 %
RF-OEX	IB1	10	0.1277 %
RF-OEX	J48	10	0.1228 %
RF-OEX	NB	7	0.043 %
RF-OEX	SMO	10	0.021 %

outliers that should be removed in order to get the best results regardless of a dataset. Results can be found in Table 2. Based on them we can recommend using the combination of RF-OEX and IB1 with 10 removed outliers. We also do not recommend using the combination of CODB and SMO, because even for the best case (5 removed outliers) it performed worse (in average) than dataset without additional processing.

5 Conclusion

Based on our data we can conclude that removing class-based outliers does improve the overall accuracy. However, the optimal number of outliers to remove varies a lot and depends on the dataset. Out of the two outlier detection methods, RF-OEX proved to be more successful.

Feature selection in combination with outlier removal does not always improve the overall accuracy as compared to outlier removal only. However, there are feature selection methods that do give higher accuracy but they depend on the dataset they are applied to.

As future work, we would suggest extending the analysis on more high-dimensional and larger datasets (as we were using sampling) to better study the correlation between feature selection and outlier detection.

Reference

1. Nabil M. Hewahi, Motaz K. Saad *Class Outliers Mining: Distance-Based Approach*. International Journal of Intelligent Systems and Technologies 2;1
2. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). *The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

3. Leona Nezvalova, Lubos Popelinsky, Luis Torgo, and Karel Vaculik *Class-based outlier detection: staying zombies or awaiting for resurrection?* In Elisa Fromont, Tijl De Bie, Matthijs van Leeuwen. Advances in Intelligent Data Analysis XIV - 14th International Symposium, IDA 2015. Neuveden: Springer, 2015. s. 193-204.

Acknowledgements. This work has been supported by Faculty of Informatics, Masaryk University Brno.

Výber vhodných atribútov pre tvorbu klasifikačných modelov v oblasti medicíny

Michal Vadovský, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00 Košice, Slovenská republika
michal.vadovsky@tuke.sk, jan.paralic@tuke.sk

Abstrakt. Predovšetkým v medicíne je výber správnych atribútov pre klasifikačnú úlohu často veľmi komplikovaný. Význam niektorých atribútov je často ťažko pochopiteľný pre bežných analytikov, ktorí môžu mať problémy určiť ich relevantnosť pre ďalšie analýzy. V tomto článku sme sa zamerali na popis rôznych spôsobov výberu atribútov, napríklad pomocou experta, štatistických testov, odstránením kolinearít alebo metódami Backward/Stepwise Selection a regresnou analýzou LASSO. Dosiahnuté výsledky jednotlivých metód výberu atribútov pri použití rôznych klasifikačných modelov prezentujeme na zredukovaných dátach so zameraním na Parkinsonovu chorobu, Metabolický syndróm a Mierne kognitívne zhoršenie. V závere tohto článku je spomenuté taktiež ďalšie zameranie našej práce.

Kľúčové slova: medicína, redukcia dimenzií, klasifikačné modely.

1 Úvod

V posledných rokoch sa čoraz častejšie zvyšuje množstvo vyprodukovaných dát, či už v podobe veľkého množstva záznamov a/alebo atribútov (dimenzií). Tento problém sa vyskytuje napríklad v medicíne, inžinierstve, riadení vzťahov so zákazníkmi a v mnohých ďalších oblastiach. Je to výzva, ktorej čelia súčasné a budúce výskumy, ktoré uplatňujú techniky dolovania v dátach na skutočné problémy vo svete. Dáta s veľkým množstvom dimenzií môžu obsahovať nadbytočné, redundantné alebo irelevantné informácie, ktoré môžu v konečnom dôsledku zhoršiť výkonnosť algoritmov [1]. Pomocou minimálneho množstva dimenzií môžeme získať modely, ktoré vyžadujú omnoho menej času na výpočet a sú jednoduchšie pre interpretovateľnosť a pochopenie. Mnoho algoritmov strojového učenia by malo byť navrhnutých tak, aby vybrali vhodnú podmnožinu atribútov, ktorá s najväčšou pravdepodobnosťou dokáže predpovedať cieľovú triedu. Nedávne výskumy ukázali, že na bežné algoritmy strojového učenia nepriaznivo vplýva výskyt irelevantných a nadbytočných informácií v trénovacej množine, preto je potrebné pri problematickom množstve dimenzií vykonať vhodné odstránenie nepotrebných atribútov (dimenzií) [2].

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 83-88.*

V našom článku sme sa zamerali na popis rôznych spôsobov pre výber dôležitých atribútov a ich následné použitie pre tvorbu klasifikačných modelov vhodných na diagnostiku primárne Parkinsonovej choroby (PCH), ale aj ochorení ako Metabolický syndróm (MS) a Mierne kognitívne zhoršenie (MCI).

2 Súčasný stav

Pri určovaní štádia alebo závažnosti PCH sa najčastejšie používa škála MDS-UPDRS, ktorá je vo forme dotazníka rozdelená do 4 základných častí (I: Nemotorické aspekty denného života, II: Motorické aspekty denného života, III: Vyšetrenie motoriky, IV: Motorické komplikácie). Za štandard pre hodnotenie kvality života u pacientov trpiacich PCH sa považujú dotazníky PDQ8 a PDQ39 (*Parkinson's Disease Questionnaire*), ktoré sa odlišujú iba v počte otázok. Tieto dotazníky hodnotia subjektívny zdravotný stav pacientov, pričom PDQ39 pozostáva z 39 otázok zoskupených do 8 podskupín a PDQ8 (kratšia verzia PDQ39) sa skladá z 8 otázok, ktoré predstavujú práve podskupiny otázok z PDQ39. Pri oboch dotazníkoch sú odpovede pacientov zaznamenané pomocou Likertovej škály s hodnotami od 0 (nikdy) po 4 (vždy). Celkové skóre môže dosiahnuť hodnoty od 0 po 100, pričom vyššie skóre znamená nižšiu kvalitu života.

V publikácii [3] sa Škorvánek s kol. zamerali na analýzu vzťahov medzi položkami v dotazníkoch MDS-UPDRS a PDQ8. V prvej časti sledovali vzťahy medzi položkami oboch dotazníkov pomocou **Spearmanovho korelačného koeficientu**. Ďalej sa zamerali na **hierarchickú regresnú analýzu**, kde sledovali vplyv 4 častí MDS-UPDRS ku celkovému skóre PDQ8, pričom brali do úvahy aj vek, pohlavie, dĺžku štúdia, celkovú dennú dávku Levodopy a dĺžku ochorenia. Výsledkom tejto hierarchickej regresnej analýzy bolo zistenie, že až 83,3% modelu bola vysvetlená pomocou údajov z častí II ($\Delta R^2/R^2 = 68\%$) a I ($\Delta R^2/R^2 = 10,4\%$). Ostatné časti MDS-UPDRS už významne neprispievali k vysvetleniu modelu. Preto sa v ďalšej časti hierarchickej regresnej analýzy zamerali už len na položky z týchto 2 častí. Z časti II najviac vplývali na výsledný model položky: Únava, Depresívna nálada a Bolesť a iné pocity. Pri časti I to boli položky: Obliekanie, Záľuby a iné aktivity a Stuhnutie. Nakoniec vykonali ešte **LASSO** analýzu, kde sledovali vzťahy medzi všetkými položkami MDS-UPDRS a výsledným skóre PDQ8. Hodnotu regresného koeficientu vyššiu ako 1 dosiahlo 6 položiek z časti I, 7 položiek z časti II, 3 položky z časti III a 1 položka z časti IV. Z piatich položiek s najvyšším vplyvom na skóre PDQ8 boli 4 položky z časti I (nemotorové) – Depresívna nálada, Apatia, Bolesť a iné pocity a Únava.

3 Metódy a dosiahnuté výsledky

V doterajších výskumoch zameraných na ochorenia PCH, MCI a MS sme používali rôzne prístupy pre výber vhodných atribútov, ako napríklad výber založený na expertovi, štatistických testoch, odstránení vzájomných závislostí (kolinearity) a použití rôznych metód na redukcii dimenzií (Backward a Stepwise Selection, regresná analýza typu LASSO).

3.1 Expert

V situácii, ak existuje vzájomná komunikácia medzi dátovými analytikmi a lekármi, ktorí dáta zozbierali, je možné ponechať výber atribútov práve na lekára. V našom prípade [4] nám dáta o pacientoch trpiacich MS poskytli lekári z Chorvátska, ktorí sami určili 4 rôzne podmnožiny atribútov vhodné pre vykonanie ďalších analýz. Pre tieto podmnožiny atribútov sme pomocou **logistickej regresie** vypočítali parametre ako **z-hodnota**, **pomer šancí (OR)** a **interval spoľahlivosti (CI 95%)**. Tieto parametre sú často používané v medicínskej oblasti, pretože podávajú informácie o dôležitosti jednotlivých atribútov pre finálnu diagnostiku a sú taktiež ľahko interpretovateľné. Najlepší model dosiahol hodnotu $R^2 = 0,41$, čo znamená že model má určitú predikčnú silu k diagnostike MS. Na úrovni spoľahlivosti 95% sme identifikovali atribúty ako *užívanie statín* (z-hodnota = 5,9; OR = 15,8; CI = <7,3, 34,2>), *Metformínu* (3,5; 44,5; <7,6, 259,5>) a *beta-blokátorov* (2,04; 2,7; <1,2, 5,9>) [5].

3.2 Štatistické testy

Pomocou štatistických testov dokážeme v dátach sledovať závislosti vstupných atribútov rôzneho dátového typu ku cieľovému atribútu, ktorý v medicíne väčšinou podáva informáciu o tom, či pacient má alebo nemá dané ochorenie. V našom prípade sme tieto štatistické testy používali skôr len vo fáze pochopenia dát, aby sme vedeli vplyvy atribútov ku cieľovej diagnostike. Pri numerických atribútoch bolo najprv potrebné testovať normalitu (**Shapiro-Wilk test**) a následne podľa výsledku vybrať **Welchov** alebo **Wilcoxon test**. Pri nominálnych atribútoch sme sledovali závislosť pomocou **Chi-kvadrát** testu, resp. **Fisherovho testu**. Určením hladiny významnosti na úrovni napr. $\alpha = 0,05$ môžeme podľa **p-hodnoty** daného testu vybrať štatisticky významné atribúty a pracovať už len s touto podmnožinou.

3.3 Kolinearita

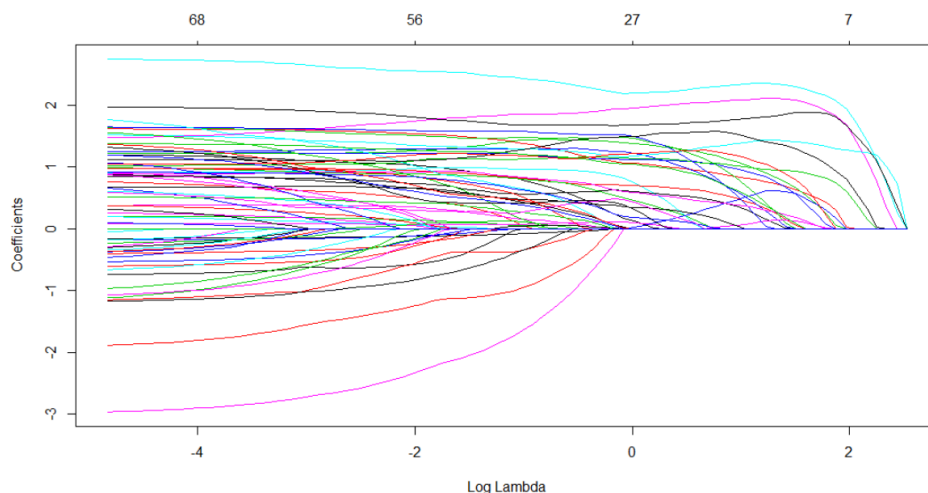
Kolinearita [6] poukazuje na situáciu, v ktorej sú dva alebo viac predikujúcich premenných navzájom závislých, čo môže spôsobiť skreslenie dosiahnutých výsledkov. Pre určenie kolinearity je pre každý atribút potrebné vypočítať **inflačný faktor rozptylu (VIF)**, ktorý ak presahuje hodnotu 5 popri prípade 10, tak hovoríme o silnej kolinearite. V publikácii [7] sme analyzovali voľne dostupné dáta zaznamenávajúce reč pacientov s PCH. Výsledne klasifikačné modely sme vytvorili buď na množine dát so všetkými 25 atribútmi, alebo na dátach očistených od kolinearity ($VIF > 5$), kde sa počet atribútov znížil na 13. Celková presnosť týchto modelov bola testovaná pomocou metód **10-CV** (*10-násobná krížová validácia*) a **LOO** (*Leave One Out*). Pri LOO sa klasifikačný model vytvára na $n-1$ záznamoch a testuje sa len na 1 zázname, pričom sa celý proces opakuje n -krát. S výnimkou jedného prípadu sa nám odstránením atribútov s vysokou kolinearitou zvýšila aj presnosť modelov klasifikujúcich pacientov do jednej zo 7 štádií PCH. Pri použití 10-CV dosiahol najvyššiu presnosť na úrovni 86,2% algoritmus rozhodovacieho stromu C5.0.

S cieľom zvýšiť presnosť modelov sme sa zamerali aj na sofistikovanejšie metódy ako **RandomForest**, **Bagging** a **Boosting**. Najvyššia bola dosiahnutá práve pomocou metódy **Boosting**. Táto metóda vygeneruje m rozhodovacích stromov, kde každý ďalší vygenerovaný rozhodovací strom berie do úvahy aj informácie z toho predchádzajúceho s cieľom minimalizovať vzniknutú chybu. Taktiež príkladom, ktoré boli klasifikované nesprávne, je v ďalšej iterácii priradená väčšia váha. V našom experimente sme pri metóde **Boosting** vygenerovali 50, 100 a 150 rozhodovacích stromov, pretože ich rastúcim počtom sa presnosť väčšinou ustáli, poprípade rastie už len minimálne, avšak čas potrebný na vygenerovanie týchto stromov rastie exponenciálne [8]. Pri pokuse orezať vygenerovaný rozhodovací strom alebo obmedziť jeho maximálnu hĺbku došlo v našom prípade k zníženiu presností. Vygenerovaním 150 rozhodovacích stromov bola pri klasifikácii pacientov s PCH pomocou metódou **Boosting** dosiahnutá najvyššia presnosť na úrovni 95,77%.

3.4 Metódy na redukciu dimenzií

Výber vhodných atribútov je možné realizovať aj metódami **Backward Stepwise Selection** (BSS), **Forward Stepwise Selection** (FSS) a regresnou analýzou typu **LASSO** [6]. Pri metóde FSS sa začína s modelom, ktorý neobsahuje žiadny atribút, pričom postupne sa atribúty pridávajú a hľadá sa ten, ktorý najviac dokáže zlepšiť model na základe ukazovateľov RSS (reziduálny súčet štvorcov) alebo R^2 (koeficient determinácie). Metóda BSS je presným opakom FSS, pretože na začiatku sú v modeli zahrnuté všetky atribúty a postupne sú odstraňované tie, ktoré sú najmenej dôležité. Pre zistenie optimálneho počtu atribútov pri použití BSS a FSS je možné použiť rôzne ukazovatele ako napr. *Adjusted R^2* (korigovaný koeficient determinácie), *Mallows's C_p* alebo *BIC* (Swarz-Bayesovo kritérium). **LASSO** je metóda regresie, ktorá v sebe zahŕňa regularizáciu. Miera penalizácie veľkosti absolútnych hodnôt regresných koeficientov je regulovaná veľkosťou koeficientu λ , vďaka čomu môžu byť niektoré odhady koeficientov nulové. Čím je hodnota λ vyššia, tým viac regresných koeficientov nadobúda nulovú hodnotu. Optimálnu hodnotu λ sme vybrali na základe minimálnej chybovosti modelu pri použití 10-násobnej krížovej validácie.

Pri analýze dát z chôdze pacientov [9] s PCH sme pomocou týchto metód zredukovali celkový počet 166 atribútov na 68 pri FSS, 81 pri BSS a nakoniec 120 použitím **LASSO**. Následne sme vo fáze modelovania použili naivný Bayesovský klasifikátor a rôzne algoritmy pre tvorbu rozhodovacích stromov, kde sme pomocou 10-násobnej krížovej validácie sledovali *presnosť*, *senzitivitu* a *špecificitu* modelov na testovacej množine. Z pohľadu algoritmu bol najlepší **RandomForest**, ktorý dosiahol najvyššiu presnosť 99,49% a senzitivitu 99,65% na množine získanej metódou **LASSO** a najvyššiu špecificitu 99,5% na množine získanej BSS. Pomocou metód FSS a BSS sa nám pri MCI podarilo vybrať zhodne po 7 najdôležitejších atribútov z celkového počtu 62. Na Obr.1 si môžeme všimnúť ukážku zmenšovania koeficientov smerom k nule. V tomto prípade sme použili **LASSO** regresiu pre sledovanie vzťahu medzi položkami v dotazníku MDS-UPDRS a výsledným skóre PDQ39 informujúce o kvalite života pacientov s PCH. Z celkového počtu 70 atribútov sa ich počet zmenšil na 48, pri optimálnej hodnote $\lambda = 0,2764$, pričom len 15 atribútov dosiahlo hodnotu koeficientu > 1 .



Obr. 1. Regresná analýza LASSO

4 Záver a budúca práca

V tomto článku sme sa zamerali na metódy výberu vhodných atribútov pre následné vytvorenie klasifikačných modelov, ktoré môžu lekárom pomôcť pri diagnostikovaní určitej choroby. Pre doktorov je niekedy veľmi zložitá vybrať vhodnú množinu atribútov a práve tieto metódy by im v tom mohli byť nápomocné. Rovnako vo fáze prípravy dát je veľmi dôležité odstrániť v dátach kolinearitu, ktorá môže spôsobiť skreslenie výsledkov. Za účelom zvýšenia dosiahnutých presností sme použili sofistikovanejšie metódy, ktoré používajú stromy ako stavebné bloky na vytvorenie silnejších predikčných modelov. Výsledná klasifikácia do triedy je potom zvolená hlasovaním všetkých vygenerovaných stromov. Vo forme SWOT analýzy sme taktiež v článku [10] popísali naše doterajšie skúsenosti s rôznymi vzorkami lekárskeho údajov.

V budúcej práci sa chceme zamerať na ďalšiu analýzu dát z [3], ktoré obsahujú vzorku 3206 pacientov trpiacich PCH popísaných 207 atribútmi. Na týchto dátach plánujeme vykonať analýzu zvyšných dotazníkov (MDS-UPDRS, PDQ8, PDQ39, EQ-5D-3L), ich vzájomné porovnanie a hľadanie skrytých vzťahov medzi položkami pomocou hierarchickej regresnej analýzy a LASSO. Dotazník EQ-5D-3L je štandardizovaný nástroj na meranie všeobecného zdravotného stavu pacienta. Jeho veľkou výhodou je to, že je vhodný a spoľahlivý v rôznych zdravotných podmienkach. Pacient odpovedá na 5 otázok (mobilita, starostlivosť o seba, zvyčajné činnosti, bolesť/nepohodlie, úzkosť/depresia) pri ktorých môže vybrať 1 z 3 možností (žiadne, nejaké, extrémne problémy). Okrem toho sa v tomto dotazníku nachádza ešte jedna mierka, pri ktorej má pacient vyjadriť ako dobre alebo zle sa cíti v danom dni (podľa jeho názoru) v rozsahu od 0-100 (najhorší – najlepší zdravotný stav). Porovnanie dotazníkov MDS-UPDRS a PDQ39 sme už vykonali, avšak výsledky ešte nie sú vyhodnotené expertom (lekárom).

PodĎakovanie: Táto publikácia vznikla vďaka podpore Agentúry na podporu výskumu a vývoja na základe Zmluvy č. APVV-16-0213 a Vedeckej grantovej agentúry MŠVVaŠ SR a SAV, projekt č. 1/0493/16.

Literatura

1. Liu, H., Motoda, H.: Feature selection for knowledge discovery and data mining, Springer, 1998.
2. Santos, V. et al.: Ensemble Feature Ranking Applied to Medical Data. *Procedia Technology* 17 (2014), pp. 223-230.
3. Škorvánek, M. et al.: Relationship between the MDS-UPDRS and Quality of Life: a large multicenter study of 3206 patients. *Movement Disorders*, vol. 31, no. 2 (2016), 31 p.
4. Sabanovic, S. et al.: Metabolic syndrome in hypertensive women in the age of menopause: A case study on data from General Practice Electronic Health Records. *BMC Medical Informatics and Decision Making*. vol. 18, no. 1:24 (2018), 18 p.
5. Babič, F., Vadovský, M., Paralič, J.: Exploračná analýza medicínskych záznamov. In: *Data a Znalosti*, Plzeň: Západočeská univerzita v Plzni, 2017. pp. 166-170.
6. G. James et al.: *An Introduction to Statistical Learning: with Applications in R*. New York: Springer-Verlag, 2013, p.426.
7. Vadovský, M., Paralič, J.: Použitie spracovaných záznamov reči pacientov pre určenie štádia Parkinsonovej choroby. In: *Data a Znalosti*, Plzeň: Západočeská univerzita v Plzni, 2017. pp. 215-220.
8. Oshiro, T. M., et al.: How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, vol. 7376 (2012), pp. 154-168.
9. Vadovský, M., Paralič, J., Tocimáková, Z.: Spracovanie a analýza dát z mobilnej aplikácie na sledovanie pohybových príznakov Parkinsonovej choroby. In: *Electrical Engineering and Informatics 8*, Košice: FEI TU, 2018. pp. 353-358.
10. Babič, F., Vadovský, M., Paralič, J.: Medical data analytics is not a simple task. In: *Medical Informatics Europe 2018*, Amsterdam: IOS Press, 2018. pp. 371-375.

Annotation:

Choosing the right attributes for creating classification models in the field of medicine

Especially in medicine, choosing the right attributes for the classification task is often very complicated. The meaning of some attributes is often difficult to understand for regular analysts who may have difficulty determining their relevancy for further analysis. In this article, we've focused on describing the various ways of selecting attributes such as expert's selection, statistical tests, removing collinearity or Backward/Stepwise Selection and LASSO regression methods. In addition, we showed some of the achieved results of various classification models on reduced data focusing on Parkinson's disease, Metabolic Syndrome and Mild Cognitive Impairment. In conclusion, we describe the future direction of our work.

Strojové učení při analýze rozsáhlých astronomických datasetů s časovými řadami

David Andrešič¹, Petr Šaloun¹ a Bronislava Pečíková²

¹ Katedra informatiky

Fakulta elektrotechniky a informatiky, VŠB-TU Ostrava

17. listopadu 15/2172, 708 33 Ostrava - Poruba

{david.andresic.st, petr.saloun}@vsb.cz

² Fakulta informatiky a informačních technologií STU v Bratislave

Ilkovičova 2, 842 16 Bratislava 4

xsuchanovab@stuba.sk

Abstrakt. Analýza časových řad je stále častější úlohou v mnoha oborech. Jedním z nejčastějších cílů je hledání skrytých period obsažených v datových sadách. V této práci stručně popisujeme možnosti řešení tohoto problému pro astronomické časové řady pocházející z několika přístrojů a to zejména s využitím umělých neuronových sítí. S ohledem na obvyklou velikost těchto datových sad se snažíme najít plně automatizované řešení, kde jedním z nejdůležitějších kritérií je rychlost hledání alespoň na úrovni binární klasifikace (tj. zda daná časová řada obsahuje periodu, nebo ne). Námi navrhované možnosti proto využívají kombinaci klasických statistických postupů a umělých neuronových sítí (multi-layer perceptron a jeho varianta s různými aktivačními funkcemi) doplněné o srovnání s řešením využívajícím výhradně umělé neuronové sítě (LSTM). V tomto článku popisujeme prozatímni výsledky posledně zmiňované metody na datové sadě pocházející z projektu BRITe.

Klíčová slova: Časové řady, velká data, perioda, umělé neuronové sítě, multi-layer perceptron, LSTM, astronomie, statistika, klasifikace

1 Úvod

Analýza časových řad pocházejících z mnoha různých zařízení nachází uplatnění v celé řadě oborů. My se v této práci zaměřili na časové řady z oblasti astronomie, kde mohou sloužit k detekci proměnných hvězd, analýze hvězdných systémů i objevům planet. Tyto světelné křivky obsahují časovou osu (nejčastěji ve formě Juliánského, příp. Modifikovaného Juliánského data) a osu jasnosti, světelného toku, příp. jiné, potenciálně proměnné veličiny.

K tomuto účelu lze využít metody hledání period ukrytých v rozsáhlých datových sadách časových řad. Konvenčním přístupem mohou být např. různé statistické postupy jako např. analýza rozptylu, nebo algoritmus Lomb-Scargle.

Jelikož se zde pohybujeme v oblasti velkých dat, hledali jsme způsoby, jak tyto statistické procesy automatizovat prostřednictvím umělých neuronových sítí

(ANN), které by mohly časové řady klasifikovat alespoň na úrovni rozhodnutí, zda obsahují periodu, či ne (viz část 3.1). Jednou možností, jak tuto klasifikaci dále vylepšit, může být např. vlastní aktivační funkce (AF) jednotlivých neuronů v závislosti na datech z učicí fáze (viz část 3.2). Jako slibný vypadal také algoritmus LSTM využívající výhradně ANN (viz část 3.3).

2 Datové sady

Naše datová sada obsahuje veřejně dostupné časové řady z celkem 4 astronomických projektů: Kepler (Mise K2), MOST, BRITE, ASAS. Všechna data jsou ve formátu podobném CSV.

Z důvodu rozdílů v datových sadách (např. v časových jednotkách, naměřených veličinách i identifikátorech hvězd) jsme se pro tuto fázi rozhodli provádět klasifikaci pouze po dílčích datových sadách namísto jejich agregace.

Prozatímní výsledky popisované v této práci se týkají datové sady z projektu BRITE (viz část 4). Data jsme spojili s daty z katalogu proměnných hvězd GCVS³, který obsahuje informace o 52011 hvězdách, které byly do roku 2015 identifikovány jako proměnné [1]. Tato data použijeme jako popisky během učicí fáze pro stanovení chyby.

3 Možnosti řešení

Zvažovali jsme možnosti řešení od těch „naivních“ jako prostá automatizace statistických metod s využitím umělé inteligence až po řešení postavená čistě na umělých neuronových sítích. Zde přinášíme jejich přehled.

3.1 Klasifikace výsledků statistické analýzy pomocí umělé neuronové sítě

Klasickým způsobem, jak ověřit, zda daná světelná křivka obsahuje periodu, či nikoliv, je podrobit ji statistické analýze (např. analýza rozptylu, nebo algoritmus Lomb-Scargle) a sestavit tzv. periodogram. Ten obsahuje periody z předem daného seznamu a pravděpodobnost, že daná světelná křivka danou periodu obsahuje.

Tento postup lze automatizovat a periodogramy klasifikovat pomocí umělých neuronových sítí, konkrétně multi-layer perceptronu (MLP). Cílem je získat uspokojivou klasifikaci alespoň na úrovni, zda daná časová řada obsahuje periodu, či nikoliv.

Výhodou této metody je rychlost klasifikace: výpočet periodogramu je zpravidla záležitostí několika vteřin. Velkou nevýhodou však představuje fakt, že periodogram je generován pro předem danou množinu period (např. 1–30 dní). Delší periody by tak byly vynechány z výpočtu, ačkoliv dle GCVS jde statisticky o malou část hvězd.

³ <http://www.sai.msu.su/gcvs/gcvs/>

V současnosti existují nástroje, které implementují různé statistické algoritmy pro sestavení periodogramu, přičemž kromě těchto funkcí obsahuje také algoritmy pro hledání korelací v šumu (např. VARTOOLS⁴). I na tyto výsledky lze přizpůsobit MLP.

3.2 Evolucí získaná aktivační funkce perceptronu

Předchozí postup lze dále optimalizovat s využitím jiné AF jednotlivých perceptronů MLP. Zajímavé může být už srovnání výsledků s využitím klasických AF, jako např. lineární, Heavisideova, sigmoid, bipolární sigmoid, hyperbolický tangens, arcus tangens, ELU, ISRLU SoftSign apod.

My se hodláme soustředit především na „šlechtění“ AF evoluční metodou s využitím reálných dat učicí sady. Inspiraci nacházíme v [2], kde je aktivační funkce získána učením zvláště pro každý umělý neuron pomocí klesání podle gradientu. Takto získaná AF „na míru“ datové sadě (resp. učicí sadě) by měla vykazovat lepší výsledky při následné klasifikaci.

3.3 Long Short Term Memory

Tento přístup využívá výhradně ANN, konkrétně Long Short Term Memory (LSTM). Ta je evolucí tzv. Recurrent Neural Network (RNN) [3], se kterou řeší obecný problém MLP: neschopnost reagovat dle předchozích vstupů. RNN toto řeší rozšířením MLP o vnitřní paměť (tzv. *context units*), která obsahuje kopii předchozí skryté vrstvy a je prezentována MLP v aktuálním kroku. Tento mechanismus činí RNN vhodnou k analýze časových řad, jelikož je ANN díky němu schopna si „zapamatovat“ sekvence obsahující dlouhé vzory předem neznámé délky. RNN však trpí problémem zvaným *Vanishing and Exploding Gradients* [4] [5], které činí učení RNN obtížným. Z tohoto důvodu byl v [6] navržen algoritmus LSTM.

LSTM je používána k predikci časových řad, v [7] však autoři ukazují, že je možné je použít také pro vyhledávání anomálií v časových řadách (LSTM-AD). Problém může představovat šum, což je včetně návrhu řešení s využitím samoorganizujících map rozebráno v [8].

My se pokoušíme pomocí LSTM „naučit“ ANN vzory časových křivek, které by dle katalogu GCVS měly obsahovat periodu.

4 Naše aktuální výsledky s využitím metody LSTM

Výsledky zatím máme pro metodu LSTM nad datovou sadou BRITE. Nejprve jsme klasifikovali 900 časových řad do 43 tříd proměnnosti (dle GCVS - např. různé podskupiny eruptivních, pulzujících, rotujících, nebo kataklyzmatických proměnných hvězd, popř. zákrytových vícehvězdných systémů, intenzivních zdrojů rentgenového záření, příp. nové, zatím více nespecifikované typy

⁴ <http://www.astro.princeton.edu/jhartman/vartools.html>

hvězdné variability). Tyto byly náhodně rozděleny v poměru 70:30 na data učící a testovací. Každá časová řada měla až 66500 měření. Kratší byly doplněny na tuto délku číslem -1 .

S těmito daty byla úspěšnost klasifikace jen 15%, přičemž úpravy parametrů LSTM (počet skrytých vrstev, jejich neuronů a epoch) neměly žádný vliv. Dále jsme klasifikovali do abstraktnějších tříd proměnnosti, až jsme skončili o pouhého zjišťování, zda daná časová řada obsahuje periodu. Úspěšnost byla pouze lehce nad 50%, což poukazuje spíše na náhodu. Data jsme dále normalizovali do intervalu $< 0; 1 >$ s 10 platnými číslicemi, což zvýšilo úspěšnost klasifikace do dvou tříd přes 60% a řádově zmenšilo chybu. V rámci experimentování jsme se pokusili klasifikovat i učící sadu jako takovou, nicméně zde byla úspěšnost pouhých 65%.

Výsledky nás vedou k úvaze, že se LSTM bylo schopno částečně naučit vzory v časových řadách, což by bylo možné vylepšit rozšířením učící sady.

5 Závěr

Popsali jsme tři možné přístupy k automatizaci hledání period v astronomických časových řadách s využitím ANN, příp. v kombinaci s konvenčními statistickými metodami sestavujícími periodogramy. S ohledem na omezení statistických metod pouze na předem definované periody se jako lepší volba jeví algoritmus LSTM. Ten však obecně dosahuje horších výsledků u časových řad obsahujících šum (v astronomii běžný jev), což by se dalo vyřešit jeho automatických odstraněním, nicméně i tento šum může obsahovat skryté (a informačně hodnotné) periody. Toto lze řešit např. separátním zpracováním časových řad s a bez redukce šumu, porovnáním výsledků a případným označením časové řady pro obsluhu.

Dále plánujeme více experimentů se zmiňovanými postupy, abychom našli ten nejvíce přesný a rychlý zároveň. To zahrnuje experimentování a porovnání různých druhů AF v porovnání s „vyšlechtenými“ v případě postupů využívajících statistické metody. Výsledky srovnáme s výsledky algoritmu LSTM.

Ve výsledku bychom tedy měli být schopni automaticky klasifikovat neznámé astronomické časové řady s vyšší přesností. Zároveň odfiltrovat ty, které neobsahují periody a zbývající řady předat obsluze k posouzení a hlubší analýze pomocí časově náročnějších metod.

Literatura

1. Durevich O.V. Kireeva N.N. Pastukhova E.N. Samus N.N., Kazarovets E.V. 2017. General Catalogue of Variable Stars: Version GCVS 5.1. *Astronomy Reports* 61, 1 (2017), 80–88.
2. Sadowski P. Baldi P. Agostinelli F., Hoffman M. 2015. Learning Activation Functions to Improve Deep Neural Networks. In *ICLR 2015 workshop*.
3. Elman, J.L., Finding structures in time. *Cognitive Sci.*, 1990, 14, 179–211.
4. Hochreiter, Sepp. (1991). Untersuchungen zu dynamischen neuronalen Netzen.

5. John F. Kolen; Stefan C. Kremer, „Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies,“ in A Field Guide to Dynamical Recurrent Networks , 1, Wiley-IEEE Press, 2001, pp.464- doi: 10.1109/9780470544037.ch14
6. Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.
7. Malhotra, Pankaj & Vig, Lovekesh & Shroff, Gautam & Agarwal, Puneet. (2015). Long Short Term Memory Networks for Anomaly Detection in Time Series.
8. Lee Giles, C & Lawrence, Steve. (2001). Noisy Time Series Prediction using Recurrent Neural Networks and Grammatical Inference. Machine Learning. 44. 161-183. 10.1023/A:1010884214864.

Poděkování: Tento článek by nevznikl bez finančního přispění následujících grantů: Grant of SGS No. SGS 2018/177, VSB-Technical University of Ostrava a TA ČR No. TL01000299.

Annotation:

Time series analysis is a growing inssue accross multiple fields of science. One of the most common tasks is finding hidden periods in data sets. In this paper, we describe possible approaches for astronomical time series using artificial neural networks. With respect to a general size of the data, we are looking for automated solution with focus on speed and accuracy at least on a "binary classification" level. Our suggested approaches incudes combination of classic statistic approaches supported by artificial neural networks and pure artifical neural networks (LSTM).

Určovanie zvedavosti používateľov z ich správania v doméne online obchodu

Jakub Šimko¹, Sandra Kostova¹, Mária Bieliková¹

Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave

Ilkovičova 2, 842 16 Bratislava

`jakub.simko@stuba.sk`, `kostova.sandra@gmail.com`, `maria.bielikova@stuba.sk`

Abstrakt. Zvedavosť je jednou z osobnostných charakteristík používateľa, ktorú je výhodné modelovať. Dopad má predovšetkým na spotrebiteľské správanie, napr. v doméne online obchodu. Zisťovaniu zvedavosti automatickou analýzou správania na webe sa však doteraz nevenovala výskumná pozornosť. V našom príspevku prezentujeme experimenty nad vzorkou 2130 používateľov, ktorých správanie na portáli webového obchodu máme k dispozícii v rozsahu niekoľkých mesiacov. Pomocou dotazníka CEI-II sme získali referenčnú informáciu o zvedavosti týchto používateľov. Skonstruovali sme črty správania s potenciálom zvedavosti charakterizovať a natrénovali s ich pomocou klasifikačný model, ktorý zaraďuje používateľov podľa miery ich zvedavosti do troch tried so správnosťou 65%.

Kľúčové slova: Modelovanie používateľa · Zvedavosť · Správanie na Webe · Strojové učenie

1 Úvod

Modelovanie používateľa na základe jeho správania na webe je dnes širokou a intenzívne skúmanou disciplínou. Pozornosť sa venuje, okrem iného, dlhodobým charakteristikám používateľa, medzi ktoré patria aj jeho osobnostné črty. Existujúce výskumy ukazujú, že osobnostné charakteristiky ovplyvňujú správanie používateľov na webe [2]. Pochopiteľne, informácie o svojich charakteristikách nám používateľa neodovzdávajú explicitne: potrebujeme ich odvodzovať nepriamo z ich správania.

Jednou z dlhodobějších vlastností človeka (používateľa) je miera jeho *zvedavosti*. Zvedavosťou rozumieme mieru záujmu používateľa o hľadanie nových poznatkov, získavanie nových skúseností a ochotu človeka užívať si nové, nepredvídateľné udalosti. Informácia o úrovni zvedavosti v používateľskom modeli má potenciálne veľkú využiteľnosť. Môže napomôcť vysvetliť správanie používateľa alebo môže byť použitá pri odporúčaní. Modelovaniu zvedavosti (na základe správania používateľa) sa zatiaľ pozornosť výskumu nevenovala.

Tento príspevok sumarizuje našu prácu, v ktorej sme skúmali možnosti určovania zvedavosti používateľa na základe jeho správania v prostredí webu. Zamerali sme sa na správanie v doméne online obchodu. Pre túto doménu sme mali k

dispozícii veľkú dátovú vzorku obsahujúcu správanie používateľov. Doména zároveň patrí k oblastiam, kde môže nájsť informácia o zvedavosti svoje využitie. K vzorke sme zároveň získali referenčnú informáciu o zvedavosti používateľov pomocou dotazníka.

Pre určovanie zvedavosti používateľa sme natrénovali model využívajúci črty, ktoré je bežne možné skonštruovať v doméne online obchodu.

2 Charakteristika dátovej vzorky

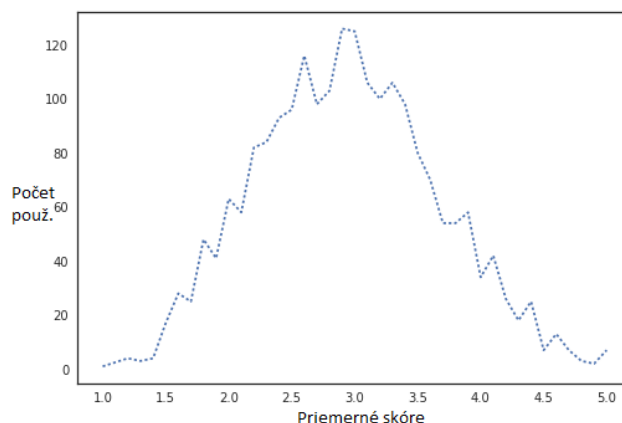
Dátová vzorka, s ktorou sme pracovali, bola vytvorená v rámci projektu HIBER [1]. Ide o záznamy správania vzorky 2130 používateľov slovenského zľavového portálu za obdobie niekoľkých mesiacov. Okrem uskutočnených transakcií (nákupov) zachytáva vzorka interakciu používateľov s rozhraním portálu (napr. zadanie vyhľadávacieho dopytu, prezeranie zoznamu produktov, prezeranie produktu, vloženie produktu do košíka). Vzorka tiež zahŕňa obsah ponúkaných produktov ako aj obsahové meta-informácie (napr. kategórie produktov).

Vzorka obsahuje vyše 300 000 udalostí (akcií). Na jedného používateľa pripadá v priemere 143 udalostí, z toho 8 nákupov (distribúcia je však nerovnomerná – má mocninový charakter). Pre náš výskum je dôležité zaradenie udalostí do sedení, hranice ktorých definujeme 30 minútami neaktivity používateľa. Takýchto sedení sa vo vzorke nachádza 53 432, čo predstavuje približne 25 sedení na používateľa.

Referenčnú informáciu o reálnej miere zvedavosti používateľov sme získali pomocou dotazníkového prieskumu. Dotazník bol súčasťou kampane, v rámci ktorej používatelia obsiahnutí vo vzorke vyplnili v lete 2017 sériu dotazníkov zameraných na rozličné osobnostné charakteristiky [1].

Dotazník pre zisťovanie zvedavosti, ktorý sme pre kampaň preložili do slovenského jazyka, bol založený na existujúcom dotazníku CEI-II (Curiosity and Exploration Inventory) [3]. Obsahuje 10 výrokov. Výroky sa týkajú črt osobnosti, ktoré sú prejavom zvedavosti, napr. *“všade kde chodím, obzerám sa po nových veciach a zážitkoch”* alebo *“som ten typ človeka, ktorý obľubuje neznámych ľudí, udalosti a miesta”*. Respondent ohodnotí každý výrok na 5-stupňovej škále podľa toho, ako veľmi súhlasí s jeho videním samého seba. Výsledkom dotazníka je celkové skóre s , vypočítané ako priemer hodnotení jednotlivých výrokov (niektoré výroky prispievajú do priemeru negatívne v dôsledku opačnej formulácie).

Pri vyhodnotení dotazníku sa ukázalo, že je naša populácia podľa celkového skóre normálne rozdelená ($\mu = 2,9$ $\sigma = 0,5$) (pozri obr. 1). Klasifikácia do dvoch tried (zvedavý/nezvedavý) v tomto prípade nedáva veľký zmysel v dôsledku veľkej skupiny priemerne zvedavých používateľov. Rozhodli sme sa preto pre klasifikáciu používateľov do troch tried: s malou, strednou a vysokou zvedavosťou. Rozhodli sme sa pre relatívne prísne kritériá zaradenia do krajných tried. Do krajných tried sme zaradili participantov z prvého ($s < 2,4$) resp. štvrtého ($s > 3,4$) kvartilu, pokiaľ aspoň dva výroky ohodnotili extrémnymi hodnoteniami príslušnej valencie. Výsledkom bola skupina 349 vysoko zvedavých, 340 málo zvedavých a 1441 priemerne zvedavých používateľov.



Obr. 1. Používatelia v našej vzorke ($N = 2130$) sú podľa skóre zvedavosti zistenom dotazníkom normálne rozdelení.

3 Model

Pred vytvorením klasifikačného modelu sme navrhli sadu črt (pre prehľadnosť uvádzame v pomenovaných skupinách). Využili sme črty bežne využívané pri iných klasifikačných úlohách no konštruovali sme aj také, ktoré vychádzali z kvalitatívnej analýzy zvedavého správania:

1. *Demografia.* Zahŕňa vek, pohlavie a príjmovú skupinu (odhad).
2. *Celková aktivita používateľa.* Zahŕňa počet udalostí, počet sedení, podiel časových kategórií udalostí (ranné, obedné, večerné ...), podiel typov udalostí (návštevy produktov, vyhľadávania, prezeranie zoznamu).
3. *Spôsob začiatku sedení.* Zahŕňa podiely spôsobov, akými používatelia začínali jednotlivé sedenia. Môže ísť o priamu návštevu titulky, vstup z iného webového obchodu (napr. meta-vyhľadávača), vstup zo sociálnej siete alebo vstup z neznámej stránky. Týmto črtami sme chceli zachytiť náchylnosť používateľa reagovať na odporúčania (a vo všeobecnosti rozptýlenia).
4. *Informácie o sedeniach.* Zahŕňa priemernú dĺžku sedenia, počet udalostí v sedení, priemerný čas medzi sedeniami či pravidelnosť sedení.
5. *N-gramy.* Zahŕňa najčastejšie sa vyskytujúce sekvencie typov udalostí používateľov (prakticky na najmä na úrovni bi- a tri-gramov).
6. *Práca s kategóriami.* Zahŕňa priemerný počet navštívených kategórií v rámci sedenia (charakteristika, ktorá by mala hovoriť o miere zvedavosti) či mieru prezerania pod-kategórií produktov.
7. *Práca so zoznamami produktov.* Zahŕňa podiel prípadov, kedy sa používateľ dostal na ďalšiu stránku výsledkov vyhľadávania alebo zoznamu produktov.

Ako klasifikačný model sme zvolili náhodný les. Model sme trénovali v 4-skladovej krížovej validácii. Keďže náš dataset bol vzhľadom na mohutnosť tried nevyvážený, delenie na testovaciu a trénovaciu sadu prebehlo pre každú triedu

Tab. 1. Výsledky tréovania náhodného lesa s rôznymi stratégiami výberu črt.

Spôsob výberu črt (počet črt)	Správnosť	Presnosť	Pokrytie	Náhoda (baseline)
Všetky črty	65,7 (+/-2,24)	56,2	64,5	44,4
Feature importance (7)	65,1 (+/- 2,27)	55,6	63,6	44,4
PCA (11)	64,8 (+/- 1,84)	55,7	63,8	44,4
Dvojica črt: príchod zo sociálnej siete, uskutočnenie nákupu	70,0 (+/-3,69)	68,4	48,7	44,4

zvlášť. Využili sme ladenie hyper parametrov. Zároveň sme experimentovali s viacerými metódami výberu atribútov (random forest feature importance, PCA). Taktiež sme skúšali trénovať veľmi jednoduché modely s dvojicami vybraných atribútov. Výsledky zhŕňa tabuľka 1.

4 Záver: úspešnosť odhadu zvedavosti je sľubná

Cieľom našej práce bolo natrénovať model (klasifikátor), ktorý by bol schopný určovať *zvedavosť* používateľov na základe ich správania na webe. Experimentovali sme v doméne *online obchodov*, v ktorej je znalosť o používateľovej zvedavosti veľmi užitočná, či už pri analýze nákupného správania alebo odporúčaní. Zároveň sme pre túto doménu mali k dispozícii rozsiahlu dátovú vzorku správania.

Pre 2130 používateľov z dátovej vzorky sme pomocou dotazníka zistili referenčnú mieru ich zvedavosti. Používatelia sú podľa tejto miery rozdelení normálne a preto sme sa pre účely klasifikácie rozhodli rozdeliť ich do troch skupín: veľmi, stredne a málo zvedavých.

Tvorbe klasifikátora predchádzal návrh črt, ktoré by zvedavosť používateľov mohli pomôcť rozlíšiť. Využili sme črty bežne využívané pri iných klasifikačných úlohách no konštruovali sme aj také, ktoré vychádzali z kvalitatívnej predstavy zvedavého správania. Viaceré črty sa opierali o charakteristiky používateľských sedení, ktoré sme definovali 30 minútovými odstupmi neaktivity.

So sľubnou úspešnosťou sa nám podarilo natrénovať klasifikátor na báze náhodného lesa. Model klasifikuje do troch tried so správnosťou 65% pri všetkých atribútoch, čím výrazne prekonáva náhodnú klasifikáciu. Dosahuje dokonca 70% úspešnosť, ak ako črty vyberieme správnu dvojicu atribútov kombinujúcu informáciu o tom, či si používateľ niečo kúpil po tom, čo sa na odkaz o produkte dostal zo sociálnej siete.

Výskum zisťovania zvedavosti zo správania považujeme za sľubne rozpracovaný. Uskutočnili sme prvú analýzu dát. Budúca práca bude vyžadovať predovšetkým konštrukciu sofistikovanejších črt, resp. analýzu prípadov nesprávnej klasifikácie v súčasných experimentoch.

PodĎakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov VG 1/0667/18, VG 1/0646/15 a APVV-15-0508.

Literatúra

1. Bielikova, M., Heretik, A., Schraggeova, M., Simko, J., Srba, I., Kompan, M., Turova, D., Gaspar, P., Gondova, V., Svetsky, V., Moro, R.: Rozsiahla dátová vzorka prepájajúca správanie používateľov e-obchodu s explicitne zistenými osobnostnými charakteristikami. In: Proceedings of the 12nd Workshop on Intelligent and Knowledge oriented Technologies – WIKT 2017. Technická univerzita v Košiciach (2017)
2. Hu, R., Pu, P.: Exploring personality's effect on users' rating behavior. In: CHI '14 Extended Abstracts on Human Factors in Computing Systems. pp. 2599–2604. CHI EA '14, ACM, New York, NY, USA (2014)
3. Kashdan, T.B., Gallagher, M.W., Silvia, P.J., Winterstein, B.P., Breen, W.E., Terhar, D., Steger, M.F.: The curiosity and exploration inventory-ii: Development, factor structure, and psychometrics. *Journal of research in personality* **43**(6), 987–998 (2009)

Annotation:

Assessment of User's Curiosity Based on Behavior in Online Shop

Curiosity is one of the user personal characteristics useful to model. It impacts the consumer behavior. In this work, we automatically assess the user's curiosity through a model based on behavioral features, constructable from common online shop logs. These features include demographics, overall activity of the users, type of their initial engagement into sessions or product visit sequences. The model is trained on a sample of 2130 users. For each user, we have several months of records of activity on a Slovak online shopping portal. The ground truth curiosity labels were acquired through online questionnaire, issued to all users in the sample. Our model classifies the users into three categories (curious, neutral, anti-curious) with accuracy of 65%.

Inferencia typu osobnosti používateľa v elektronickom obchode

Róbert Móro, Veronika Balážová, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava
{robert.moro, xbalazovav, maria.bielikova}@stuba.sk

Abstrakt. Typ osobnosti nám pomáha vysvetliť, ako používatelia pracujú s webovou stránkou. Automatická inferencia osobnostných charakteristík používateľa, ako je napr. typ jeho osobnosti, môže byť užitočná pri personalizácii aplikácií či odporúčaní. Navrhli sme metódu inferencie typu osobnosti používateľa, pričom osobnosť modelujeme pomocou päťfaktorového modelu (angl. Big Five). Metódu sme overili na dátach z elektronického obchodu, ktoré zachytávajú aktivitu a typ osobnosti 2168 používateľov.

Kľúčové slová: Modelovanie používateľa, typ osobnosti, NEO-FFI, elektronický obchod

1 Úvod a existujúce prístupy

Adaptívne systémy modelujú tradične záujmy či vedomosti používateľa, na základe ktorých sa snažia prispôbovať ponúkaný obsah alebo spôsob, akým je možné obsah vyhľadávať alebo prezerat' [2]. Čoraz častejšie je však trendom modelovať aj ďalšie osobnostné [5] či kognitívne charakteristiky [7] používateľov, keďže tieto majú vplyv na spôsob, akým používatelia interagujú so systémom.

V našej práci sa zameriavame na inferenciu typu osobnosti používateľa v doméne elektronických obchodov, ktorá doteraz nebola dostatočne preskúmaná. Existujúce práce sa zamerali predovšetkým na doménu sociálnych sietí, ako sú napr. Twitter [5] alebo Facebook [6], prípadne na doménu odporúčania filmov [9], v ktorej odporúčač so zapojením typu osobnosti prekonal porovnávané prístupy kolaboratívneho odporúčania. Hoci existuje viacero modelov (taxonómií) typov osobnosti [10], v prevažnej väčšine týchto prác sa osobnosť modeluje pomocou tzv. Big Five modelu, v ktorom je osobnosť vyjadrená pomocou piatich faktorov (neurotizmus, extravergia, otvorenosť, prívetivosť a svedomitosť). Na určovanie typu osobnosti podľa Big Five modelu sa využíva dotazník NEO-FFI (angl. NEO Five Factor Inventory, kde NEO odkazuje na prvé písmená faktorov neurotizmus, extravergia a otvorenosť, keďže len tieto boli uvažované v prvej verzii dotazníka) od Costu a McCraea [4], v ktorom respondenti vyjadrujú svoj súhlas s predloženými tvrdeniami, ktoré vyjadrujú, ako sa sami vnímajú.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 101-105.*

V doméne elektronických obchodov, na ktorú sa zameriavame v našej práci, existujúce štúdie skúmali súvislosť medzi faktormi osobnosti a spôsobom online nakupovania (hedonistické vs. utilitárne správanie) [8]. Zistili napríklad, že používatelia s väčším stupňom neurotizmu, prívetivosti či otvorenosti sa správajú viac utilitárne, t. j. snažia sa nakúpiť čo najrýchlejšie a najefektívnejšie. Znalosť typu osobnosti používateľa v doméne elektronického nakupovania by sa tak dala využiť napr. na prispôsobenie prezentácie odporúčaní (akým spôsobom sa odporúčané položky zobrazia a prípadne aj ich počet), prípadne sa môže využiť na lepšie interpretovanie implicitnej spätnej väzby používateľa.

Navrhli sme preto metódu automatickej inferencie typu osobnosti používateľa založenú na Big Five modeli. K problému pristupujeme ako ku klasifikačnej úlohe, pričom používateľov klasifikuje pre každý faktor do dvoch, resp. troch tried. Navrhnutú metódu sme overili na rozsiahlej dátovej vzorke z existujúceho slovenského elektronického obchodu – zľavového portálu [1].

2 Metóda inferencie typu osobnosti

Navrhnutá metóda pozostáva z troch základných krokov:

1. *Zaznamenávanie aktivity používateľa na evidenčnej vrstve* – zaznamenávame informácie z profilu používateľa (demografické údaje) a tiež jednotlivé udalosti (prezretie ponuky v elektronickom obchode, pridanie produktu do košíka a pod.)
2. *Odvádzanie atribútov na medzivrstve* – demografické údaje a udalosti používateľa zachytené na evidenčnej vrstve spracúvame a vytvárame z nich celkovo 105 atribútov. Tieto je možné zaradiť do troch skupín; okrem atribútov odvodených z *demografických údajov* ide o *atribúty aktivity* (napr. množstvo aktivity používateľa v častiach dňa, resp. týždňa, čas medzi dvomi po sebe idúcimi aktivitami, resp. nákupmi, množstvo minútých peňazí, počet prezretých, resp. zakúpených ponúk a pod.) a *atribúty sedení* (priemerné trvanie sedenia, počet udalostí v sedení, počet zakúpených ponúk a pod.). Sedenie pritom definujeme ako úsek aktivity používateľa, ktorý je od druhého sedenia oddelený aspoň štyrmi hodinami neaktivity. Pri atribútoch odvádzame aj ich priemerné, minimálne, maximálne hodnoty a hodnoty štandardnej odchýlky od priemeru.
3. *Napĺňanie modelu používateľa* reprezentovaného 5-zložkovým binárnym, resp. ternárnym vektorom korešpondujúcim jednotlivým faktorom Big Five modelu osobnosti (napr. používateľ *je / nie je neurotický* v prípade zložky binárneho vektora alebo *je málo neurotický / je stredne neurotický / je veľmi neurotický* v prípade zložky ternárneho vektora). Napĺňanie je realizované natrénovanými klasifikátormi (pre každý faktor osobnosti jeden), ktorých vstupom sú odvodené atribúty na medzivrstve. Pri klasifikácii sme v práci porovnali viacero metód strojového učenia, konkrétne metódu podporných vektorov SVM a náhodný les (v práci však uvádzame len najlepšie výsledky, ktoré boli dosiahnuté pomocou SVM [3]).

Reálne dáta z elektronického obchodu bývajú spravidla riedke a používatelia sa môžu značne odlišovať v množstve vykonanej aktivity. Metóda musí byť schopná sa s týmto vysporiadať; realizujeme preto normalizáciu údajov počtom udalostí pre každého používateľa a tiež dopĺňame chýbajúce hodnoty priemerom pre daný segment.

Keďže pri odvádzaní atribútov na medzivrstve vzniká značný počet atribútov, realizujeme v procese trénovania ich automatický výber. Postupujeme tak, že si vyberieme reprezentantov dvoch tried (napr. extravertní vs. introvertní) z opačných koncov spektra a pre nich realizujeme *štatistické testovanie pomocou t-testu* na určenie, ktoré atribúty sú rozlišujúce, teda rozdiely v ich hodnotách sú pre zvolené skupiny štatisticky významné. Do trénovania klasifikátora následne vstupujú už len takto vybrané významné atribúty.

3 Overenie na rozsiahlej dátovej vzorke

Overenie navrhutej metódy sme realizovali na rozsiahlej dátovej vzorke zo slovenského elektronického obchodu (zľavového portálu) na vzorke 2168 používateľov (70% žien, 30% mužov), ktorí vyplnili osobnostný NEO-FFI dotazník [1]. Dokopy bolo pre nich zaznamenaných okolo 660 000 udalostí rôznych typov, ako napr. prezeranie konkrétnej ponuky (ktoré sú zastúpené v najväčšom počte), prezeranie kategórie ponúk, vyhľadávanie, pridanie ponuky do košíka či jej hodnotenie používateľom. Hoci priemerne spravili používatelia vyše 140 udalostí, ich počet zodpovedá mocnínovej distribúcii, t. j. pre veľa používateľov je zaznamenaný veľmi malý počet udalostí. Zo vzorky sme preto odfiltrovali takých používateľov, ktorí mali menej ako 10 zaznamenaných udalostí.

V prvom kroku sme realizovali binárnu klasifikáciu, t. j. pre každý faktor sme určovali, či má používateľ danú osobnostnú črtu (napr. otvorenosť), alebo nie. Aby sme overili, či je možné medzi dvomi skupinami rozlíšiť, vybrali sme pre každý faktor horný a dolný kvartil a trénovanie a testovanie sme vykonávali len nad takouto obmedzenou vzorkou (s rovnomerným zastúpením oboch tried; klasifikátor majoritnej triedy by tak mal správnosť 0,5). Najlepšie dosiahnuté výsledky (SVM s výberom 22 rozlišujúcich atribútov podľa t-testu) sú zobrazené v tab. 1.

Tabuľka 1. Výsledky binárnej klasifikácie.

Faktor	Správnosť (Acc.)	Presnosť (P)	Úplnosť (R)	F_1 skóre
Otvorenosť	0,66	0,43	0,51	0,47
Svedomitosť	0,55	0,55	0,57	0,56
Extraverzia	0,79	0,66	0,62	0,64
Prívetivosť	0,68	0,56	0,68	0,61
Neurotizmus	0,65	0,66	0,54	0,59

Ako najlepšie atribúty pre inferenciu sa ukázali byť počty kliknutí na odporúčanie v systéme, pomer medzi počtom odporúčaných a počtom kúpených pro-

duktov, ako aj pomer medzi počtom prezeraných a zakúpených produktov. Ako rozlišujúci atribút sa tiež ukázal príchod na stránku, t. j. či používateľia prišli na stránku priamo alebo z iného sídla (napr. Facebooku). Z výsledkov tiež môžeme vidieť, že rôzne faktory osobnosti je možné určovať s rôznou správnosťou; najmenej sa v akciách používateľa odrážala jeho svedomitosť, najviac naopak extravertzia.

Vyskúšali sme tiež ternárnu klasifikáciu; pridali sme používateľov v strede distribúcie pre každý faktor (druhý a tretí kvartil) a natrénovali sme klasifikátory pre tri triedy. Priemerná správnosť podľa očakávania klesla (0,43 pri ternárnej klasifikácii vs. 0,67 pri binárnej), stále sme však dosahovali výsledky nad hranicou náhody.

4 Záver

V práci sme navrhli a overili metódu automatickej inferencie typu osobnosti používateľa. Hlavný prínos spočíva v návrhu a spôsobe určenia rozlišujúcich atribútov pre jednotlivé faktory osobnosti, ako aj v tom, že sme na rozdiel od existujúcich prác ukázali možnosti aplikácie automatickej inferencie v doméne elektronických obchodov.

V ďalšej práci by sme chceli preskúmať a pri trénovaní zapojiť súvislosti medzi jednotlivými faktormi osobnosti, resp. zostaviť komplexnejší klasifikátor na princípe učenia súborom metód. Pre ďalšiu prácu zostáva aj zapojenie automaticky odvodeného typu osobnosti pre prispôbenie obsahu elektronického obchodu, napr. v podobe personalizovaného odporúčania.

Literatúra

1. Bieliková, M., Heretik, A., Schraggeová, M., Šimko, J., Kompan, M., Turoňová, D., Gašpar, P., Svetský, V., Móro, R.: Rozsiahla dátová vzorka prepájajúca správanie používateľov e-obchodu s explicitne zistenými osobnostnými charakteristikami. In: Proceedings of 12th Workshop on Intelligent and Knowledge Oriented Technologies - WIKT 2017. p. 4 (2017)
2. Brusilovsky, P., Millán, E.: User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: The Adaptive Web, pp. 3–53. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
3. Cortes, C., Vapnik, V.: Support-vector networks. *Machine learning* **20**(3), 273–297 (1995)
4. Costa, P., McCrae, R.: Neo five-factor inventory (neo-ffi). Odessa, FL: Psychological Assessment Resources (1989)
5. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. pp. 149–156 (2011)

6. Ortigosa, A., Carro, R.M., Quiroga, J.I.: Predicting user personality by mining social interactions in facebook. *J. of Comp. and Sys. Sci.* **80**(1), 57–71 (2014)
7. Steichen, B., Carenini, G., Conati, C.: User-adaptive information visualization - Using eye gaze data to infer visualization tasks and user cognitive abilities. In: *Proceedings of the 2013 International Conference on Intelligent User Interfaces - IUI '13*. pp. 317–328. ACM Press, New York, USA (2013)
8. Tsao, W.C., Chang, H.R.: Exploring the impact of personality traits on online shopping behavior. *African Journal of Business Management* **4**(9), 1800–1812 (2010)
9. Wu, W., Chen, L.: Implicit acquisition of user personality for augmenting movie recommendations. In: Ricci, F., Bontcheva, K., Conlan, O., Lawless, S. (eds.) *User Modeling, Adaptation and Personalization*. pp. 302–314. Springer International Publishing, Cham (2015)
10. Zuckerman, M., Kuhlman, D.M., Joireman, J., Teta, P., Kraft, M.: A comparison of three structural models for personality: The big three, the big five, and the alternative five. *Journal of personality and social psychology* **65**(4), 757–768 (1993)

Pod'akovanie: Tento článok vznikol vďaka čiastočnej podpore projektov APVV-15-0508, VG 1/0667/18 a VG 1/0646/15.

Annotation:

Inference of Personality Type of E-Shop User

Personality type helps us to explain, how the users work with a web page. Automatic inference of user characteristics, such as his or her personality type, can be useful for personalization of applications or in recommender systems. We proposed a method of user personality type inference using a five factor model of personality (so called Big Five model). We evaluated the proposed method on data from an e-hop, which contains activity and personality types of 2168 users.

Interaktívny nástroj pre analýzu a vizualizáciu textových dokumentov

Miroslav Smatana, Peter Butka, Dominika Maršaleková

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach

Letná 9, 042 00 Košice, Slovenská republika

miroslav.smatana@tuke.sk, peter.butka@tuke.sk,
dominika.marsalekova@student.tuke.sk

Abstrakt. Digitalizované dáta sú v dnešnej dobe všade okolo nás. Sú vo forme dokumentov, novín, blogov či rôznych článkov. Ľudia sa v tomto veľkom množstve dát strácajú a majú problém v nich nájsť relevantné informácie. To vedie k potrebe ich spracovania, analýzy a vhodnej vizualizácie, ktorej ľudia rozumejú. Preto sme sa v tejto práci rozhodli predstaviť nástroj pre analýzu a vizualizáciu informácií extrahovaných z digitálnych dokumentov. Analýza dát je vykonávaná pomocou 2 metód analýzy textových dokumentov: modelovanie tém, analýza sentimentu a získané informácie sú používateľovi predstavené pomocou interaktívnej vizualizácie.

Kľúčové slova: analýza dát, modelovanie témy, analýza sentimentu, vizualizácia

1 Úvod

V dnešnej dobe sa dáta digitalizujú do rôznych foriem, či už sú to noviny, blogy, sociálne médiá alebo rôzne články. Ľudia vedia k týmto dátam pristupovať, no vďaka ich veľkému množstvu s nimi nevedia efektívne pracovať, čím vzniká potreba nových prístupov, ktoré by dokázali nejakým spôsobom tieto dáta spracovať, analyzovať a následne z nich extrahovať potencionálne užitočné informácie a tak zjednodušiť vyhľadávanie požadovaných informácií. Takto extrahované informácie môžu byť užitočné v mnohých smeroch napr. dáta (príspevky) z diskusných fór môže používateľ využiť pri rozhodovaní či kúpiť alebo nekúpiť nejaký produkt.

Avšak analýza dát je len prvým krokom spracovania týchto dát. Ďalším oveľa dôležitejším krokom je ich vhodná vizualizácia koncovému používateľovi, aby týmto výsledkom dokázal porozumieť a čo najrýchlejšie sa na ich základe rozhodnúť.

Preto sme sa v tejto práci zamerali na vytvorenie nástroja, ktorý umožňuje analyzovať vstupné textové dokumenty a získané výsledky vizualizuje s využitím viacerých vizualizačných techník. Na analýzu dát využívame 2 metódy a to modelovanie témy a analýzu sentimentu.

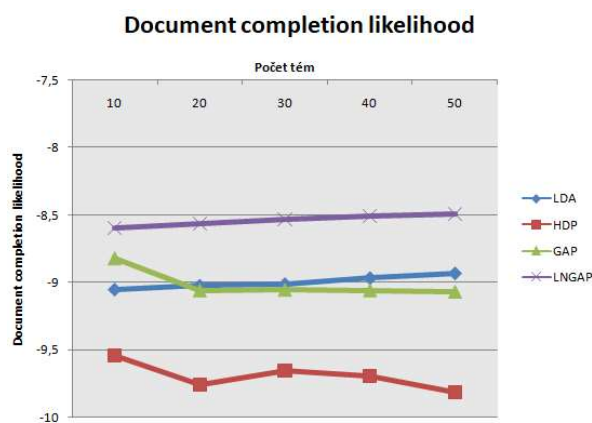
*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 107-110.*

2 Modelovanie témy

Cieľom modelovania tém je vytváranie skupín (tém), ktoré sa spolu vyskytujú dostatočne často, pomocou hľadania skrytých tematických štruktúr vo vstupnej kolekcii dokumentov. Táto metóda analýzy dát ponúka nové možnosti prehľadávania, vyhľadávania a sumarizácie dát.

Ako jednou z prvých metód, ktorá sa pokúša riešiť problém modelovania tém je Latentná sémantická alokácia (LSA) [1]. Avšak formálne nie je predstavovaná ako metóda modelovania tém, ale tvorí základ pre množstvo iných metód modelovania tém ako napr. latentná dirichletová alokácia (LDA) [2]. Medzi ďalšie patria napr. hierarchický dirichletov process (HDP) [3], gamma poissonov model (GAP) [4] alebo lognormálový gamma poissonov model (LNGAP) [5].

Na Obr. 1 je možné vidieť porovnanie vybraných metód modelovania tém použitím metódy "Document completion likelihood" [6], kde čím väčšia hodnota reprezentuje kvalitnejšie výstupné témy. Porovnanie bolo vykonané na množine 8000 ekonomických dokumentov¹.



Obr. 1 Porovnanie metód modelovania tém

3 Analýza sentimentu

Analýza sentimentu, tiež známa pod pojmom dolovanie názoru, je typ štúdie, ktorá si kladie za úlohu automatickú analýzu názorov ľudí, ohodnotenia, postoje a emócie k entitám ako sú produkty, služby, organizácie, problémy, udalosti, témy a ich atribúty. Tieto názory sú kľúčové vo väčšine ľudských aktivít pretože majú najväčší vplyv na naše rozhodovanie (vo väčšine prípadov ak robíme nejaké rozhodnutie chceme poznať názor aj ostatných ľudí). Firmy môžu tieto názory využiť napríklad pri zavedení nového produktu na trh a zistiť, či sa ľuďom páči alebo aké má chyby a na druhej strane ľudia sa na základe nich môžu rozhodnúť či daný produkt kúpiť alebo nie. Hlavným

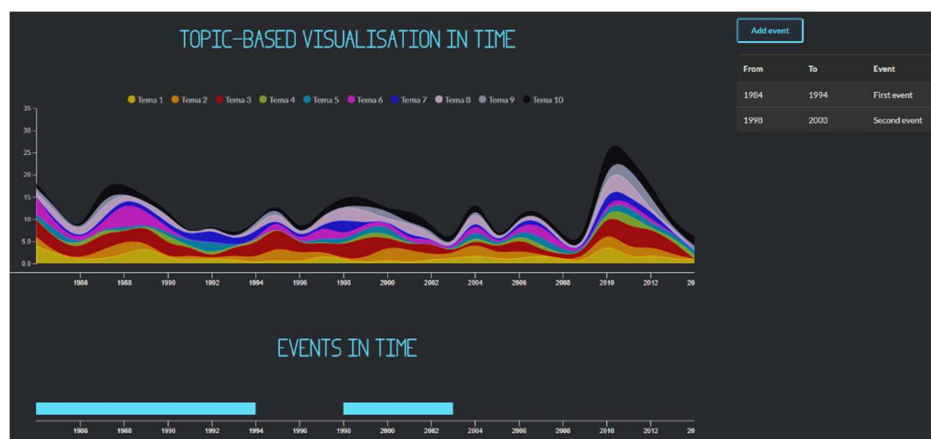
¹ <https://data.world/crowdflower/economic-news-article-tone.60>

cieľom analýzy sentimentu je klasifikácia, či vstupný dokument dokument alebo príspevok vyjadruje pozitívny alebo negatívny názor [7].

4 Navrhovaný nástroj

Cieľom navrhovaného systému je umožniť jednoduché a intuitívne spracovanie textových dokumentov za účelom ich analýzy a následnej možnosti využitia získaných informácií v procese rozhodovania koncovým používateľom.

Funkcionalitu nástroja je možné predstaviť pomocou jednoduchého príkladu použitia. Predstavme si, že spoločnosť zaoberajúca sa predajom mobilných telefónov chce získať údaje z príspevkov ich používateľov na internete. Najprv musia do systému nahradiť svoje dáta a následne zadať dopyt (napr. "Iphone"), ktorý ich v danej množine príspevkov zaujíma. Následne na Obr. 2 je možné vidieť výsledok analýzy danej množiny príspevkov. V hornej časti je zobrazený vývoj tém v čase a v spodnej časti si môže používateľ pridať dôležité udalosti pre dané témy (napr. kedy bol telefón uvedený na trh, marketingovú kampaň a pod.). Na základe týchto informácií môže spoločnosť vidieť napr. či marketingová kampaň fungovala alebo čo ľudia najviac diskutujú o danom telefóne.



Obr. 2 Detail získaných tém na základe vstupného dopytu

Následne je možné po kliknutí na konkrétnu tému získať jej detailnejšie informácie ako kľúčové slová, najinformatívnejšie príspevky a vývoj sentimentu v čase pre túto tému. Tento detail je zobrazený na Obr. 3.

5 Záver

V práci bol prezentovaný základný popis vybraných metód určených na automatickú analýzu textových dát a taktiež nástroj, ktorého úlohou je automatická analýza vstupnej množiny dokumentov a následná vizualizácia získaných informácií. Využitie toh-

to nástroja bolo prezentované na jednoduchom príklade použitia. Ako je možné vidieť tento nástroj môže byť potenciálne užitočný v procese rozhodovania firiem ako aj bežných používateľov.



Obr. 3 Detail témy

Literatúra

1. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), 259-284
2. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.
3. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2012). Hierarchical dirichlet processes. *Journal of the american statistical association*.
4. Ogura H., Amano H., Kondo M. (2013). Gamma-Poisson Distribution Model for Text Categorization.
5. Valverde J. A. (2013). Full Bayes Poisson gamma, Poisson lognormal and zero inflated random effects models: Comparing the precision of crash frequency estimates.
6. Wallach H. M. Evaluation Methods for Topics Models.
7. Liu B. (2012). Sentiment Analysis and Opinion Mining.

Pod'akovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16 a APVV projektu č.APVV-16-0213.

Annotation:

Interactive tool for analysis and visualization of textual documents

In present, digital data are all around us. A lot of them are in form of documents (news, blogs, articles, social media). In such amount of data is hard to find relevant information. It leads to finding new ways of its processing, analysis and appropriate visualization, which is suitable for end users. For that reason we present in this paper tool for automatic analysis and visualization of information extracted from input digital documents. Data analysis is done using 2 methods: topic modeling and sentiment analysis.

Vyhledávání obrázků v rozšířené realitě

Petr Chmelař, Jaroslav Kuchař

Fakulta informačních technologií
České vysoké učení technické v Praze
Tháková 9, 160 00 Praha 6, Česká Republika
{chmelpe7|jaroslav.kuchar}@fit.cvut.cz

Abstrakt. Aplikace pro mobilní telefony se staly každodenní součástí našich životů. Jejich pole působnosti je ale zatím ve většině případů striktně limitované displejem daného zařízení. Jedním ze současných trendů ve vývoji mobilních aplikací je právě rozšiřování možností aplikace za hranice samotného přístroje – do prostředí virtuální, potažmo rozšířené reality. Cílem tohoto projektu je vytvoření systému, který umožní obohacení mobilních aplikací o vyhledávání v prostředí rozšířené reality. Hlavní komponentou systému je serverová aplikace poskytující službu k vyhodnocení shod přijatého obrázku s definovanou databází obrázků. Pro usnadnění implementace vyhledávání v rozšířené realitě je k dispozici knihovna pro platformu iOS. Správa databází obrázků je umožněna prostřednictvím REST API nebo pomocí jednoduchého webového rozhraní.

Klíčová slova: počítačové vidění, rozšířená realita, OpenCV, JavaScript

1 Úvod

Hranice mezi reálným světem a světem informačních technologií se neustále ztenčuje. Současná mobilní zařízení jsou schopna zachytit okolní prostředí, porozumět jeho kontextu a zobrazit realitu rozšířenou o definovanou přidanou hodnotu, to vše v reálném čase. Hovoříme o tzv. rozšířené realitě. Ukázkovým případem užití je například přehrání filmové upoutávky po rozpoznání filmového plakátu. Pro zmíněné porozumění kontextu často dochází k zapojení netriviálních disciplín, jako je počítačové vidění, strojové učení nebo umělá inteligence.

Cílem projektu [1] je vytvoření takového systému, který poskytne vývojářům mobilních aplikací potřebné služby a nástroje pro obohacení aplikací o vyhledávání obrázků v rozšířené realitě. Pro správu obrázků určených k rozpoznávání je k dispozici webové rozhraní. Samotná interakce s rozšířenou realitou je zjednodušena využitím poskytované knihovny. Systém tak umožňuje implementovat danou funkcionalitu i bez komplexní znalosti problematiky.

2 Architektura systému

Proces vyhledávání obrázků v rozšířené realitě sestává celkem ze tří částí. První částí je nepřetržité zachycování okolního prostředí fotoaparátem mobilního

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 111-114.*

zařízení. Dále je nutné porovnat zachycené snímky oproti předem definované databázi obrázků, které chceme rozpoznávat a vyhodnotit případné shody. Poslední částí je vyvolání příslušné akce v prostředí rozšířené reality.

Pro zajištění škálování a jednoduché modifikace databáze obrázků bylo rozhodnuto o implementaci tzv. kontinuální komunikace. Serverová aplikace udržuje databáze obrázků a poskytuje REST API pro jejich modifikaci a vyhodnocení shod. Mobilní zařízení kontinuálně odesílá zachycené snímky na server kde dochází k vyhodnocení shody s určenou databází a odeslání odpovědi.

Hlavní komponentou systému je serverová aplikace implementovaná v programovacím jazyce JavaScript za využití běhového prostředí Node.js. Pro samotnou analýzu obrázků a vyhodnocení shod je využita knihovna OpenCV. Její spuštění v prostředí Node.js je zajištěno využitím knihovny opencv4nodejs [2]. Součástí projektu bylo také rozšíření této knihovny o potřebnou chybějící funkcionalitu.

Vývojáři mobilních aplikací mohou využít knihovnu pro platformu iOS, která zajišťuje proces kontinuální komunikace se serverem. Při detekci shody vyrozumí mobilní aplikaci o výsledku vyhledávání, který je možný dále zpracovat. Databáze obrázků je možné spravovat prostřednictvím REST API nebo využít jednoduchou webovou aplikaci implementovanou pomocí knihovny React.

3 Analýza obrázků a vyhodnocení shod

Pro vyhodnocení shody dvojice obrázků je nejprve nutné provést jejich analýzu. Byla zvolena metoda extrahování lokálních vlastností. Jejím základem je identifikace malých oblastí obrázku, které splňují definovanou charakteristiku. Zpravidla se jedná o hrany, rohy, bloby nebo vysoce kontrastní místa. Identifikované lokální vlastnosti se obvykle skládají z klíčového bodu určujícího pozici a deskriptoru, který reprezentuje vizuální popis dané vlastnosti.



Obr. 1. Příklad porovnání (vlevo obrázek z databáze, vpravo přijatý snímek)

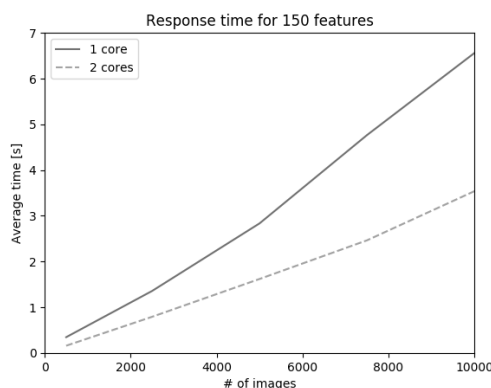
Knihovna OpenCV obsahuje celou řadu algoritmů pro extrakci lokálních vlastností. S ohledem na požadovanou rychlost odezvy systému byl zvolen algoritmus ORB [3]. Deskriptory jsou reprezentované binárními řetězci, které snižují paměťovou složitost. K jejich porovnání je možné využít vysoce efektivní výpočet Hammingovy vzdálenosti.

Vzájemnou shodu dvou obrázků lze určit porovnáním jejich deskriptorů. Využitím algoritmu k nejbližších sousedů jsou nejprve pro každý deskriptor z prvního obrázku nalezeny dva deskriptory z druhého obrázku s nejmenší vzdáleností. Pomocí tzv. poměrového testu [4] následně dochází k odfiltrování nevalidních dvojic. Výsledek je dále zpřesněn výpočtem homografie a ověřením zachování orientace obrázku (Obr. 1).

Uvedený proces je spuštěn na serveru po přijetí požadavku k vyhodnocení shod. Přijatý snímek je porovnán se všemi obrázky v určené databázi. Informace o obrázku s nejvyšší konfencí shody jsou vráceny jako pozitivní odpověď na požadavek. Pokud žádný obrázek nesplňuje minimální míru konfidence, shoda neexistuje a je vrácena negativní odpověď.

4 Testování systému

Součástí projektu bylo automatické otestování implementovaného systému pomocí skriptů. Byla vytvořena testovací databáze obsahující 500 log z volně dostupného souboru dat LLD [5] a dále 500 testovacích scén ve dvou modifikacích představující ideální a zhoršené podmínky.



Obr. 2. Průměrná doba vyhledávání dle počtu obrázků v databázi

Systém dosahuje přesnosti 99 – 100 % při úplnosti 87 % pro ideální podmínky a 65 % pro zhoršené podmínky. Nižší úplnost vyhledávání je ovlivněna náhodným výběrem testovacích obrázků z datasetu, který obsahuje i obrázky nekvalitní, případně obrázky nevhodné pro vyhledávání. Při reálném využití lze tedy očekávat vyšší hodnoty.

Pro zjištění možností škálování systému byla implementována experimentální paralelní verze vyhledávání. Principem je rovnoměrné rozdělení databáze obrázků mezi více procesů. Výsledky testování ukazují téměř dvojnásobné zrychlení při spuštění na dvoujádrovém procesoru (Obr. 2). Dalšího zrychlení by bylo možné dosáhnout vhodným indexováním databáze nebo využitím optimalizované verze knihovny OpenCV určené pro běh na grafických procesorech.

5 Závěr

Využitím open-source technologií byl implementován funkční prototyp systému umožňující obohacení mobilních aplikací o vyhledávání obrázků v rozšířené realitě. Jednovláknová implementace dokáže obsloužit databáze obsahující zhruba 5000 obrázků. Testování ukázalo, že při paralelizaci vyhledávání je možné dosáhnout řádově vyšších limitů, které jsou srovnatelné s komerčními službami. Významným přínosem projektu je také rozšíření knihovny opencv4nodejs o chybějící funkcionalitu.

Literatura

1. Chmelař, P.: Systém pro vyhledávání obrázků v rozšířené realitě. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2018.
2. Mühler, V.: opencv4nodejs [online]. 2018, [cit. 2018-06-21]. Dostupné z: <https://github.com/justadudewhohacks/opencv4nodejs>
3. Rublee, E., Rabaud, V., Konolige, K.: ORB: An Efficient Alternative to SIFT or SURF. International Conference on Computer Vision, Barcelona, 2011.
4. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, ročník 60, č. 2, 2004, s. 91–110.
5. Sage, A., Agustsson, E., Timofte, R.: LLD – Large Logo Dataset [online]. 2017, [cit. 2018-06-22]. Dostupné z: <https://data.vision.ee.ethz.ch/sagea/lld/>

Poděkování: Tento článek vznikl díky podpoře Fakulty informačních technologií, České vysoké učení technické v Praze.

Annotation:

Image Search in Augmented Reality

Mobile applications have become a daily part of our lives. One of the current trends in mobile application development is the expansion beyond the device itself – into the virtual or augmented reality. The goal of this project is to create a system that will allow developers to enhance mobile applications with image search in augmented reality. The main part of the system is a server application that provides a service to evaluate the match of the received image with the defined image database. An iOS library is available to simplify the process of image search in augmented reality. Administration of image databases is available through the REST API or through a simple web interface.

Concepts and Relations Acquisition within Ontology Learning Process for Automotive

Ondřej Šebek¹, Václav Jirkovský¹, and Nestor Rychtycký²

¹ Czech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague

Jugoslávských Partyzánů 3, 160 00 Prague, Czech Republic
{ondrej.sebek, vaclav.jirkovsky}@cvut.cz

² Ford Motor Company
Dearborn, MI, USA
nrychtyc@ford.com

Abstract. Flexible manufacturing as the key target of Industry 4.0 depends on precise knowledge of available human as well as machine resources and a capability to flexibly manage them. Integration of the various resources is a very complex task and represents a significant obstacle for a perfect understanding of the required knowledge. An ontology may facilitate the integration, but the problem occurs when the ontology does not comprise required concepts or relations. External data sources may be exploited for the ontology extensions. Such tasks may be hard to achieve purely by human given the volume of data involved in industrial applications. In this paper, we investigate an approach of acquiring the necessary information for ontology learning using a web mining method. The proposed approach was experimentally verified on the integration of Excel document containing spare parts and Ford Supply Chain Ontology.

Keywords: Ontology · Ontology Learning · Automotive · Web Mining

1 Introduction

Flexible manufacturing as the key target of Industry 4.0 depends on many aspects including precise knowledge of available resources and a capability to manage resources for ensuring highly customized production. Thus, it is required to integrate various data-sources representing various characteristics of the flexible manufacturing needs. Unfortunately, many of required knowledge (for demanded integration) is not expressed explicitly but is hidden in the corresponding text files, hard-coded in programs, or only hidden in the minds of designers of corresponding data models of resources.

In recent years ontology proved to be a suitable tool that enables to build a common knowledge model incorporating information from various sources in a consistent way — especially light-weight ontologies for linked-open-data solutions. The problem of the integration is in the missing knowledge which should be solved using some existing external data sources. Algorithms which form or extend ontologies belong to ontology learning.

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 115-120.*

This paper is organized as follows: First, we introduce basic characteristics of ontology learning. Next, we present the possible solution for forming the new information required for the integration. Finally, the paper concludes with a summary.

2 Ontology Learning

The (semi-)automatic methods for ontology construction are typically referred to as ontology learning altogether. This term was coined by A. Madche and S. Staab in 2001 [5]. In other words, ontology learning deals with the acquisition of a domain model from data. The ontology learning process consists of several separated tasks:

1. Acquisition of the relevant terminology
2. Identification of synonym terms
3. Formation of concepts
4. Hierarchical organization of the concepts
5. Learning relations, properties or attributes
6. Hierarchical organization of the relations
7. Instantiation of axiom schemata
8. Definition of arbitrary axioms

There are several already proposed methods for solving some particular steps of the ontology learning tasks. One of the approaches to the problem of semi-automated ontology learning is to employ structured datasets such as WordNet [6] or DBpedia³. Examples of such efforts may be found in [9], [4], [1] [2]. Web mining as an unsupervised learning strategy gains in popularity among researchers in the last years, e.g., [3], [7], [8]. In this paper, we focused on the first five tasks of the ontology learning process and we exploit structured datasets. Furthermore, web mining together with lexico-syntactic patterns is employed due to specific application domain which is not covered in WordNet and DBpedia.

3 Concepts and Relations Acquisition

In this paper, we deal with a problem of importing new information into existing ontology from an external data source. In our case the input data is extensive and detailed Excel document where records for various automobile spare parts are saved. The goal of performed ontology learning process is to incorporate each of the spare part records into the existing ontology, either by assigning the spare part to an existing concept or by defining new concepts and relations to other concepts in the ontology. Every record of a spare part is characterized by abbreviated part label (e.g., *SE CSHAFT RR OIL*) which must be decoded using an internal database of acronyms. Within this database, every token from abbreviated part label can have more than one expansion, e.g., *SE=sealer, sealing*,

³ <http://wiki.dbpedia.org>

september, etc, resulting in ambiguity of decoded spare part label. Output of the preprocessing is a set of translated part labels which are subsequently used as an input of the ontology learning process.

In the first phase, we examined application of structured datasets, such as DBpedia or WordNet, as a resource for the ontology learning process. The obvious advantage of employing these datasets is that they contain valuable information about the domain in the form of semantic relations to other concepts. Parameters of the search in DBpedia or WordNet must be carefully tailored to a specific application to ensure the required level of successful mappings as well as their meaningfulness. In this test we limit our search in DBpedia to concepts related to a super concept *Vehicle technology* by semantic relations “skos:broader”⁴ and “dc:subject”⁵. We constrain the search in WordNet to concepts related to a super concept *Automotive vehicle* by hypernymy, hyponymy, and meronymy relations.

For the purpose of search in the structured datasets we considered all possible translations of the abbreviated spare part label as well as their permutations and subsets. Table 1 lists few examples of mappings between spare part labels and concepts from DBpedia and WordNet. Obviously, there are a lot of reasonable mappings, but some other mappings are doubtful. The disadvantage of the DBpedia is that it contains rather general semantic relations which don’t constrain very well the type of resulting concept. For instance, the same relation “is Subject of” (i.e., is dc:subject of) associates the concept *Engine technology* to concepts *Crankshaft* (i.e., an artifact, part of an engine) and *Oil pressure* (i.e., physical quantity) in DBpedia. While the latter concept is undoubtedly related to the domain of automotive technology, it doesn’t seem to be useful to include such concept in the context of supply chain ontology. In this sense, the WordNet seems to be more promising as it contains the meronymy relation directly associating some components of the vehicle to others.

Table 1. Examples of mappings between spare part labels and DBpedia and WordNet concepts.

Spare part label	DBpedia	WordNet
<i>O/RG-RR AX WHL SHFT</i>	<i>Axle</i>	<i>Axle</i>
	<i>Wheel</i>	<i>Wheel</i>
	<i>Rear axle</i>	<i>Shaft</i>
	<i>Outrigger</i>	
<i>SW ASY-OIL PRESS SDR</i>	<i>Station wagon</i>	<i>Switch</i>
	<i>Oil pressure</i>	
<i>BLT FL/WHL TO CSHAFT</i>	<i>Crankshaft</i>	<i>Crankshaft</i>
	<i>Flywheel</i>	<i>Bolt</i>

⁴ <https://www.w3.org/TR/skos-reference/#broader>

⁵ <http://dublincore.org/documents/dces/>

Furthermore, it can be seen that the mappings between spare part labels and DBpedia or WordNet concepts are found mostly for subsets of the full label. Moreover, individual mappings for given label may be unrelated as in the case of spare part label mapping to *Station wagon* and *Oil pressure*. This is because the domain in interest is very specialized while both DBpedia and WordNet contain rather general concepts. For this reason we propose more sophisticated approach to the ontology learning task which employs web mining tools. Overall work-flow of the proposed solution is illustrated in Fig. 1. In this solution DBpedia and WordNet are used to define more generalized concepts forming a backbone of the resulting ontology while web mining method is used to define more specialized concepts and to find their relations to the general concepts.

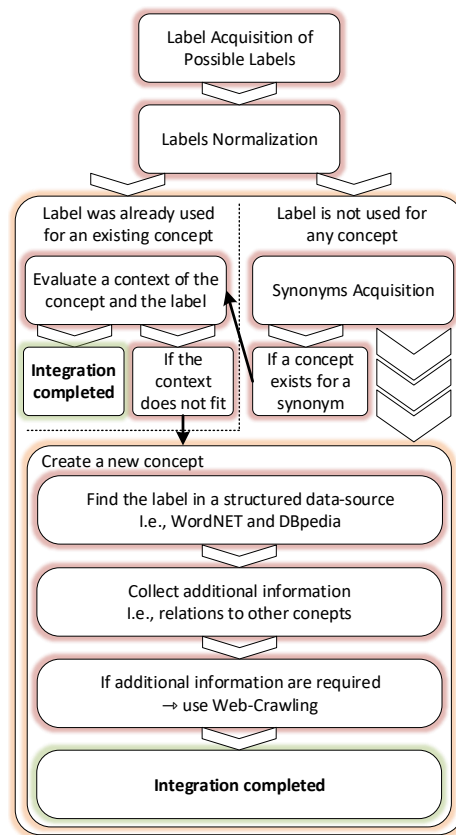


Fig. 1. Work-flow of the proposed solution

Web mining is used in the proposed solution for several tasks. First, in order to resolve the ambiguity of translated spare part label web mining is used to select the most appropriate concept name among the possible spare part labels.

The straightforward algorithm is used for this task — frequency of occurrence within web documents, which are searched using the given term, is assigned to the term (the possible name). The candidate concept name with the highest frequency is selected as the name for the new concept (if it is not already defined in the ontology). Using this method it is evaluated, for example, that the most probable translation of the spare part label *SE CSHAFT RR OIL* is crankshaft rear oil seal.

Furthermore, web mining is used for derivation of relationships between concepts using lexico-syntactic patterns. Used patterns are divided into four categories according to semantic relationship — hypernymy (e.g., “is a”), hyponyms (e.g., “called”), meronymy (e.g., “as part of”), holonymy (e.g., “consists of”). These patterns are searched in web documents together with pairs of ontology concepts.

4 Conclusions

In this paper, a solution for building ontology elements (concepts, their instances, and their relations) has been introduced. This solution may be understood as facilitating the integration of required resources into an existing ontology. The proposed solution was verified on the integration of spare parts database into Ford Supply Chain Ontology.

The proposed solution combines identification of required elements in structured resources, i.e., DBpedia and WordNet, and acquisition of additional knowledge by means of web mining technique. This approach aims to overcome problems when a given domain or given resources are very specific so that the existing structured data-sources are not convenient.

Acknowledgment

This work is supported through the Ford Motor Company University Research Proposal (URP) program and by institutional resources for research by the Czech Technical University in Prague, Czech Republic.

References

1. Booshehri, M., Luksch, P.: Towards adding linked data to ontology learning layers. In: Proceedings of the 16th International Conference on Information Integration and Web-based Applications. pp. 401–409 (2014)
2. Booshehri, M., Luksch, P.: An ontology enrichment approach by using dbpedia. In: Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics (2015)
3. Dong, H., Hussain, F.K.: Self-adaptive semantic focused crawler for mining services information discovery. *IEEE Transactions on Industrial Informatics* **10**(2), 1616–1626 (May 2014). <https://doi.org/10.1109/TII.2012.2234472>

4. Luong, H.P., Gauch, S., Speretta, M.: Enriching concept descriptions in an amphibian ontology with vocabulary extracted from wordnet. 2009 22nd IEEE International Symposium on Computer-Based Medical Systems pp. 1–6 (2009)
5. Maedche, A., Staab, S.: Ontology learning for the semantic web. *IEEE Intelligent systems* **16**(2), 72–79 (2001)
6. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41 (1995)
7. Seitner, J., Bizer, C., Eckert, K., Faralli, S., Meusel, R., Paulheim, H., Ponzetto, S.P.: A large database of hypernymy relations extracted from the web. In: *LREC* (2016)
8. Young, J., Basile, V., Kunze, L., Cabrio, E., Hawes, N.: Towards lifelong object learning by integrating situated robot perception and semantic web mining. In: *Proceedings of the European Conference on Artificial Intelligence (ECAI) 2016 conference* (2016)
9. Zhou, W., Liu, Z., Zhao, Y., Xu, I., Chen, G., Wu, Q., Huang, M.I., Qiang, Y.: A semi-automatic ontology learning based on wordnet and event-based natural language processing. In: *Information and Automation* (2006)

Data Stewardship Wizard for Open Science

Marek Suchánek¹ and Robert Pergl²

Faculty of Information Technology
Czech Technical University in Prague
Thákurova 9, 160 00 Praha
{suchama4¹, perglr²}@fit.cvut.cz

Abstract. Every year, the amount of data (in science) grows significantly as information technologies are used more intensively in various domains of human activities. Biologists, chemists, linguists, and others are not data experts but often just regular users who need to capture and process some huge amount of data. This is where serious problems emerge – bad data management leading to losing important data, producing unverifiable results, wasting funds, and so on. Thousands of qualified data stewards will be needed in following years to deal with this issues. At the Faculty of Information Technology, CTU in Prague, we participate in the European platform ELIXIR in which we work on the Data Stewardship Wizard to help researchers and data stewards with building high-quality FAIR data management plans that are accurate and helpful to their projects. We cooperate on this challenging project with our colleagues from other ELIXIR nodes.

Keywords: Data Stewardship · FAIR Data · Wizard · DM Plan

1 Introduction

Data Management comprises all disciplines related to managing data as a valuable resource. Lately, a new term, Data Stewardship, has started to be used to emphasize not just doing data management during a scientific project, but to cover capturing, processing, (long time) preserving, and re-using data. All of these from the planning stage of the project to the stage when the project is already finished. In organizations, there should be designated data stewards as experts that have the knowledge how to deal with data in a correct way and help researchers to organize their projects appropriately.

Data stewards have complicated tasks to stay well trained themselves and supervise data management of projects in the organization, i.e. communicate and collaborate intensively with researchers and help them with the planning. Data Stewardship Wizard helps with this whole process resulting in successful scientific projects. Moreover, data management plans are more and more often required as an attachment to funding applications. Example of this practice is the European Union. EU and many data experts define Data Management Plans (DMPs) as a key element of good data management. [1]

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 121-125.*

2 ELIXIR

ELIXIR is an international organization with the main goal to coordinate these resources so that they form a single infrastructure. It operates mainly in European states, since 2014 it has gathered 21 members (including the Czech Republic, but for example also Israel) and over 180 research organizations. Each member state represents a node and there is a central point called Hub that is located in the UK. Activities of ELIXIR are separated into five *Platforms*: Data, Tools, Interoperability, Compute, and Training. [2]

ELIXIR and its nodes organize a lot of events to gather scientists and to share knowledge. Also, ELIXIR provides funding through so-called Implementation Studies that are an important part of activities of a particular Platform. They are proposed by Platforms, then must be agreed with the ELIXIR Heads of Nodes committee and approved by the ELIXIR Board. [2]

Our project targets the Interoperability and Training platforms since the Data platform is about specific data resources. FIT CTU in Prague is part of ELIXIR CZ. We tightly cooperate with IOCB that is the Czech ELIXIR national node and also with our colleagues from DTL (Dutch Techcentre for Life Sciences) that acts as the Dutch node of ELIXIR. Thanks to the ELIXIR infrastructure and events, we have a lot of feedback and attention across Europe. [3,4]

3 FAIR Data Principles and Metrics

To do a good and precise data stewardship and to promote open science, a broad international community has developed the FAIR Data principles. FAIR is a nice abbreviation of four pillars: Findable, Accessible, Interoperable, and Reusable. Each of those is representing a set of three or four principles (15 in total) that are stating the requirements. It is clear that these principles are applicable not just to life science datasets but to data and projects in general. [5]

The GO FAIR Initiative focuses on the development of the Internet of FAIR Data & Services (including the European Open Science Cloud) through many activities such as trainings, tools, certifications, and so on. The problem is the figuring out when you can say "my data are FAIR", for this reason, the FAIR Metrics are prepared so you can measure your FAIRness. [6]

There are 14 well-defined metrics developed by the FAIR Metrics Group. Each metric is related to FAIR principle, has own identifier, name, description, information about the measuring and validation, and examples. In most cases, evaluation is a process of retrieving metadata from entered URI and running validation with it. For that purpose, the Metrics Evaluator made mainly by Mark Wilkinson is also placed together with complete metrics description and manuals in the repository <https://github.com/FAIRMetrics/Metrics>. [7]

4 Project: Data Stewardship Wizard

After introducing the background, ELIXIR, and FAIR, the Data Stewardship Wizard (<https://dsw.fairdata.solutions>) can be described.

4.1 Goals and Visions

The goal is to develop an open-source solution that allows data stewards to work with the knowledge in order to build a structured questionnaire with related information. After filling-in project information into this questionnaire, the researchers will get high-quality data management plan (DMP). Again, produced DMP contains detailed information such as advices, contacts, and metrics.

There are two most important differences from other solutions (such as DMPonline or DMPTool). First, the knowledge model is easily extensible and thus we are able to incorporate changes as well as integrate totally new concepts and related applications. Next, other solutions try to simplify the creation of DMPs by reducing the number of questions. Our vision is strictly against this. We want to ask a lot of questions so the DMP is fully tailored to the project. On the other hand, we don't want to overwhelm the researcher with 600 questions but show very few at the beginning and guide through only relevant ones.

4.2 Initialization and Participants

All of this started with data stewardship mind map produced by Rob Hooft (DTL) and related book [8] by Barend Mons (GO FAIR). At first, we encoded the mind map in a machine-readable format – JSON files, provided JSON schema, and published it on GitHub together with colleagues from the Netherlands. Then we developed at our faculty the first version of the Wizard as an interactive browser of knowledge model following our visions. New functionality was being added continually as we were getting user feedback (registrations, saving plans, multiple plans per account, etc.). It has been developed in Haskell with high-quality evolvable architecture allowing such extensions easily.

4.3 New Generation Wizard

This year, the Wizard is being developed as part of an 18-month ELIXIR Implementation Study of the ELIXIR Training Platform. Our team at FIT CTU gained new two skilled members. The study also includes training materials and setting up workshops. These days we already have a new extensible portal solution with Haskell backend and Elm frontend that consists of user and settings management, knowledge model editor and migrations, and finally, the Data Stewardship Planner that produces DMP from a filled questionnaire.

It basically works as depicted in Figure 1. Data steward uses the editor to produce a knowledge model (KM) packages by adding, deleting, or editing entities such as chapters, questions, answers, references and so on. A KM can be updated in this way and have different versions using semantic versioning. Also, it can be based on another parent KM as a customization. If parent KM is updated, there is a migration tool that allows updating child KMs. Finally, in the DS Planner, a KM can be turned into the questionnaire to be answered by a researcher. DMP can be exported to various formats. A researcher is able to provide a feedback regarding any question so data steward can continuously improve the KM according to needs of researchers.

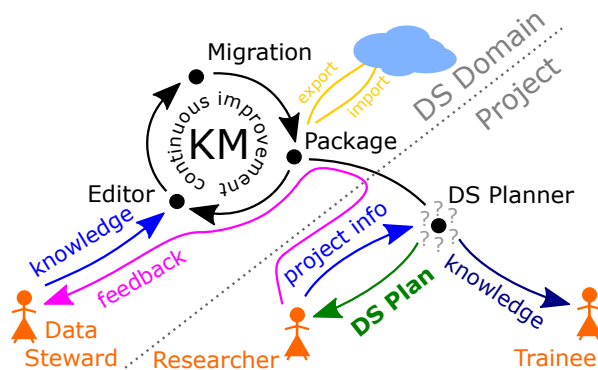


Fig. 1. Data Stewardship Wizard – knowledge flow

4.4 Architecture and Deployment

Currently, there is one instance of the Wizard hosted by ELIXIR CZ, but any organization or individual can deploy their own instance in a very simple way thanks to Docker. Of course, the Wizard provides export and import functionality of KMs as JSON files. JSON schema to describe these files is public together with a guide on how to publicly share KMs from the Wizard. Such setup together with the hierarchy of KMs is depicted in Figure 2.

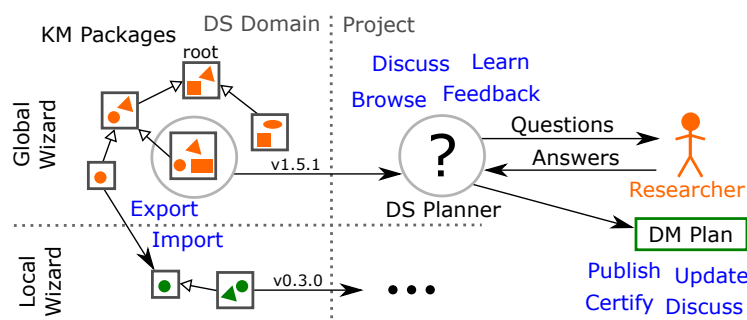


Fig. 2. Data Stewardship Wizard – architecture

MongoDB is used as the storage for all the data in the Wizard. It can be also running from Docker, locally, or remotely. There are also other integrations, for example, giving question-related feedback as GitHub issues directly from the Wizard, sending emails, or subscribing into a newsletter. In the future, the number of such integrations is expected to grow significantly.

4.5 Roadmap

Aside from the tasks that are defined in the Implementation Study, we are focused on the community needs and we align our development plans according to them. We get a lot of valuable feedback from non-commercial as well as the commercial environment through various channels. Our plans and projects are publicly available on <https://github.com/DataStewardshipWizard>.

For Fall 2018, we are ready to deliver new features such as metrics (FAIR but also others) connected to questions and answers, desirability phases of questions, extended types of references, on-the-fly questionnaire indications, summary reports, and DMP exports in various formats. In the following period, we will focus on already mentioned integration, especially in term of evaluating the questionnaire (e.g., connecting to FAIRness evaluator or cost evaluator), but also with other ELIXIR services like Training eSupport System (TeSS), EDAM Ontology, Bioschemas, or FAIRsharing.org.

5 Conclusion

The Data Stewardship Wizard is smart but simple to use solution covering the whole process of data management planning. A broad community is using, testing, and exploring details of the Wizard and providing valuable feedback that we are always very thankful for. We are happy that we do something to help people and science in general and that people are really interested in our work. It still will be a long or even never-ending way of improvements and adjustments, but unpretentious thanks to the used architecture and technologies.

References

1. EUROPEAN COMMISSION: Open Science. Last accessed 20 Jun 2018, <https://ec.europa.eu/research/openscience>
2. ELIXIR: About us, <https://www.elixir-europe.org/about-us>. Last accessed 15 Jun 2018
3. ELIXIR-CZECH: Organisation/Structure, <https://www.elixir-czech.cz/about-elixir-cz/organisation-structure>. Last accessed 15 Jun 2018
4. DTL: ELIXIR-NL, <https://www.dtls.nl/elixir-nl/>. Last accessed 16 Jun 2018
5. WILKINSON, Mark D. et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, vol. 3
6. GO FAIR: GO FAIR Initiative, <https://www.go-fair.org/go-fair-initiative/>. Last accessed 17 Jun 2018
7. FAIR METRICS GROUP: FAIR Metrics, <http://fairmetrics.org>. Last accessed 15 Jun 2018
8. MONS, Barend, Data Stewardship for Open Science: Implementing FAIR Principles, 1st edn. CRC Press, 2018, ISBN 9781498753173

Acknowledgements: The project has been supported by ELIXIR and ELIXIR CZ. This paper was supported by CTU research grant SGS17/211/OHK3/3T/18.

Semantic Support for Facilitating Production Monitoring

Václav Jirkovský¹, Petr Kadera¹, and Marek Obitko²

¹ Czech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague
Jugoslávských Partyzánů 3, 160 00 Prague, Czech Republic
`{firstname.lastname}@cvut.cz`

² Rockwell Automation Research and Development Center
Argentinská 1610/4, 170 00, Prague, Czech Republic
`mobitko@ra.rockwell.com`

Abstract. Flexible manufacturing together with other aspects of Industry 4.0 brings many benefits and means an important competitive advantage for producers within various domains. This trend is highly dependent on proper understanding, processing, and interpretation of manufacturing processes, machine capabilities as well as real-time and historical shop floor data. New and specific means are required for the flexible manufacturing establishment. In this paper, we discuss and describe possibilities how to utilize a combination of reasoning and Semantic Web Rule Language (SWRL) for facilitating production monitoring. The reasoning offer ways how to classify observed data into available concepts. However, it has certain limitations for example resulting from the Open World Assumption of ontologies. Thus, SWRL may be a suitable supplement for enabling production monitoring.

Keywords: Ontology · Reasoning · Semantic Web Rule Language · Production Monitoring

1 Introduction

The fourth industrial revolution (also known as Industry 4.0) represents flexible manufacturing systems which bring many benefits, e.g., they are enablers for highly customized productions. However, many requirements have to be ensured for enabling matured flexible manufacturing systems — proper understanding, processing, and interpretation of manufacturing processes, machine capabilities as well as real-time and historical shop floor data. Many of legacy systems as well as existing approaches are not applicable for flexible manufacturing because of flexible systems are based on not exactly defined sequence of operations.

Very important component of manufacturing systems is a production monitoring. The monitoring is essential for tasks comprising, for example, a production planning or scheduling. Thus, the creation of flexible and versatile production monitoring is the core requirement for enabling flexible manufacturing. In other

words, a controlling of operations within flexible manufacturing systems is not strictly pre-defined, but operations are executed according to a current state of a given (semi-)product and currently available resources. Thus, a suitable production monitoring algorithm should be able to transform observable information into a classification of a production state.

One of the possible approaches is the utilization of a reasoning [3] and knowledge representation in an ontology. These semantic web technologies may be exploited for a transformation of particular partial observations into demanded form/knowledge, i.e., observed situation and corresponding data represent an individual which belongs to a certain concept. The question, if an individual is an instance of a given concept may be answered by a reasoner.

The utilization of reasoning separately has limitations from its nature, and it is suitable to complement the reasoning with the additional suitable tool. Therefore, Semantic Web Rule Language (SWRL) [2] is employed in the flexible production monitoring approach.

This paper is organized as follows: First, we present the definition of concepts for the demonstration of the classification (monitoring). Next, we introduced the classification of the “anonymousProduct” individual followed by the work-flow of production monitoring. Finally, the paper concludes with the summary and presents the benefits and drawbacks of the introduced solution.

2 Production Monitoring

The production monitoring is a complex task which is dependent on many factors as well as influences various processes and systems. The proposed solution presented in this paper is further development of previous research introduced in [4].

The problem of the production monitoring may be expressed as a transformation of observations, which form an individual (from the point of view of Web Ontology Language [1]), into the corresponding concept. In other words, there may be an observable product in some “level of completeness” (i.e., semi-product) and it is required to classify it to corresponding concept. The level of completeness of a product also represents a set of accomplished operations and determine subsequent operations.

The suitable ontology design has to be taken into consideration for enabling the production monitoring. In many cases, ontologies cannot be simply re-used but have to be extended by appropriate axioms, e.g., covering axioms for concepts. Without these axioms, there is no closure for concepts and the classification using reasoning could be impossible. The close of value partitions may be facilitated by Protege editor³, which is capable of adding covering axioms for defined concepts automatically.

The example of the concepts definition for the production monitoring is illustrated in Fig. 1. The figure illustrates the description of the concept “skoda”

³ <https://protege.stanford.edu/>

and related engines represented by the concept “TSI”. For simplicity and clarity, we assigned TDI engines only for “octavia combi” models and TSI only for “octavia” models.

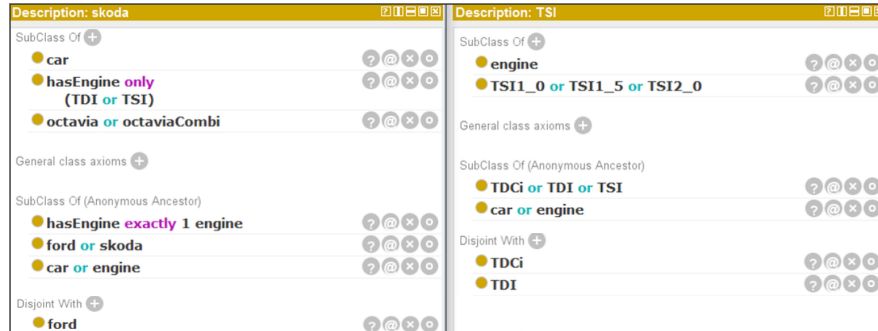


Fig. 1. Skoda and TSI concepts definition in Protege editor

Next, if an individual is defined as an individual of the general car concept with “TSI 1.5” engine, then a reasoner is able according to these observations and definitions classify the individual into the “octavia combi” concept.

An important feature of the concept of ontologies (the Open World Assumption [5]) means a limitation for the proposed solution. The Open World Assumption is reflected in the distributed nature of OWL ontologies, and it may cause problems for the production monitoring. The problem is not relevant for the introduced example, but it is related to “the quantitative” definition of concepts. The example may be defined as a production of a column from Lego bricks, where there are three different columns — two-brick, five-brick, and seven-brick column. The example of seven-brick column definition is illustrated in Fig. 4.

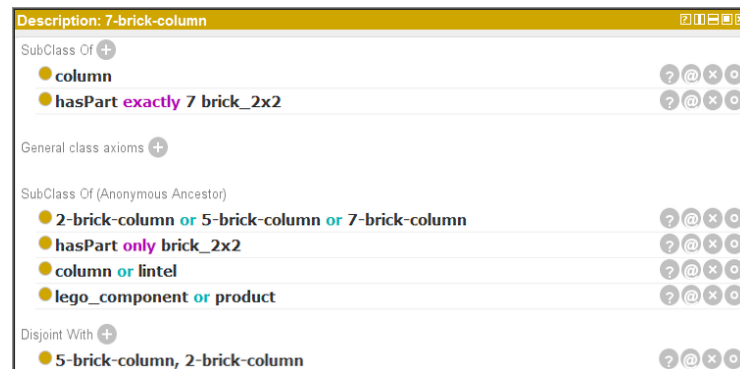


Fig. 2. Definition of 7-brick column in Protege editor

Then, if there is an observable product consisting of two bricks then we would like to know if the observable semi-finished product (an individual) is unfinished two-brick, five-brick, or seven-brick column. In this case, the reasoner cannot naturally decide if the observable product is the column because of the Open World Assumption. We can see completed two-brick-column, but reasoner expects a possible definition of other bricks for this product. Therefore, the individual is classified as more general concept, i.e., the column concept. The example of the classification of the anonymous product consisting of two bricks is illustrated in Fig. 3.

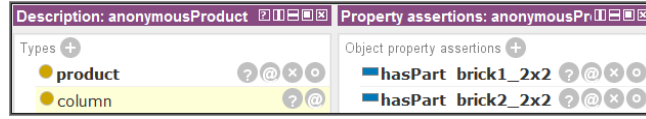


Fig. 3. Classification of the individual consisting of two bricks

If there is another observable product consisting of six bricks then the reasoner classify the product as 7-brick column. The example of the classification of the anonymous product consisting of six bricks is illustrated in Fig. 4.

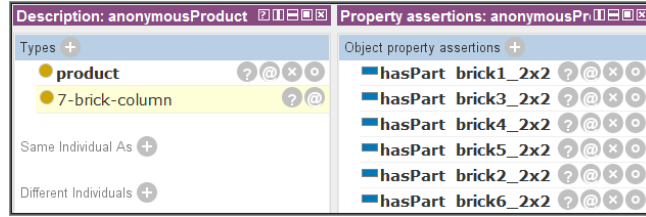


Fig. 4. Classification of the individual consisting of six bricks

This behavior of the reasoner is not an error, and even human cannot classify the product without any additional information properly, e.g., following production steps. On the other hand, we would like to find out all possible products which would be created from the observable individual and information. The problem cannot be solved by accepting the more general concept (i.e., column) as a set of possible target products — because if the individual consists of three different bricks (i.e., disjoint individuals) at a given moment then it is still classified as the column concept (but it cannot be two-brick column anymore).

Thus, we decided to exploit SWRL for the production monitoring. SWRL rules are able to “switch evaluation” into the Close World Assumption. Furthermore, this approach keeps SWRL rules together with a given ontology separated from the control code. In other words, we may directly define that a column

consisting of less than five Lego bricks is semi-finished five-brick-column. The example of the rule for an unfinished product classification as the five-brick column is as follows:

$$\begin{aligned} &product(?p) \wedge brick_2x2(?b) \wedge hasPart(?p, ?b) \circ sqwrl:makeSet(?bricks, ?b) \\ &\quad \circ sqwrl:size(?size, ?bricks) \wedge swrlb:lessThanOrEqual(?size, 5) \\ &\quad \longrightarrow sqwrl:select(?p, ?size) \end{aligned}$$

The proposed solution is not only based on the mentioned algorithm combining OWL reasoning and SWRL rules but also has to reflect the complex environment of production systems. Thus, we may summarize the essential steps of production monitoring work-flow as follows:

- First of all, operations have to be defined as well as concepts representing products and their components. Production monitoring solution should reflect general operations but only feasible operations of a given production system.
- There has to be way how to sense data and recognize the suitable situation for the classification.
- A classification of semi-product is derived from observed information, a given ontology, SWRL rules, and with the help of a reasoner.

The overall work-flow is illustrated in Fig. 5. The setup of the system is not straightforward, but it is iterative process — definitions of operations are adapted according to SCADA⁴ system, and finish product may be controlled by MES or ERP⁵ system.

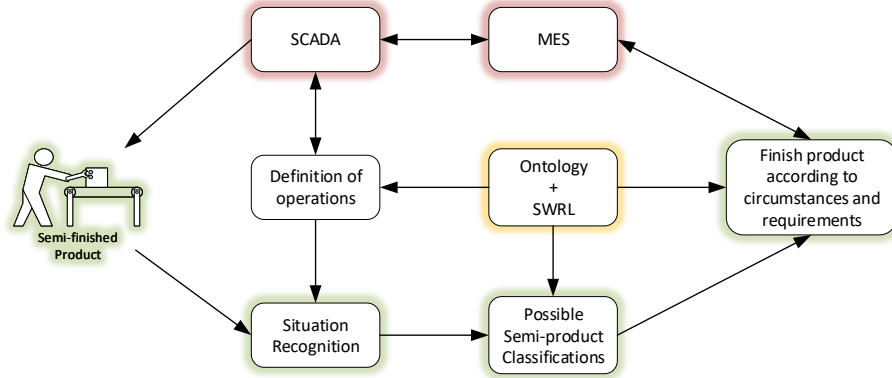


Fig. 5. Workflow of monitoring classification

⁴ <https://cs.wikipedia.org/wiki/SCADA>

⁵ https://en.wikipedia.org/wiki/Enterprise_resource_planning

The solution of the production monitoring has to control also unambiguous situations, i.e., the semi-finished product is classified into more than one concept. The system should decide the best fitting concept or provide all of the possible concepts to the user, etc. Moreover, a situation when there is more than one applicable SWRL rule is possible, and the system should take it into consideration.

Next responsibility of the production monitoring system are notifications of “a level of completeness” to a user. It may be difficult if a classification is unambiguous. Furthermore, the system should decide when and how to add new axioms defining the product individual into an ontology.

3 Conclusions

In this paper, the solution of the flexible production monitoring has been shortly described. The solution benefits from ontology reasoning supplemented by SWRL rules. This approach differs from “hard-coded” solutions and therefore offers possibilities for easy re-usage.

The solution is strongly dependent on the quality of a given ontology design which is related to the definition of production operations. These facts directly influence the functionality of the proposed solution and may be demanding for many applications or/and many users. On the other hand, this obstacle may be overcome by the employment of expanding already existing shared ontologies.

Furthermore, the utilization of knowledge representation in the form of an ontology provides a clearer interpretation of solution outcomes.

Acknowledgment

This research has been supported by Rockwell Automation Laboratory for Distributed Intelligent Control (RA-DIC), by H2020 project DIGICOR, and by institutional resources for research by the Czech Technical University in Prague, Czech Republic.

References

1. Bechhofer, S.: Owl: Web ontology language. In: Encyclopedia of database systems, pp. 2008–2009. Springer (2009)
2. Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosz, B., Dean, M., et al.: Swrl: A semantic web rule language combining owl and ruleml. W3C Member submission 21, 79 (2004)
3. Horrocks, I., Sattler, U.: Ontology reasoning in the shq (d) description logic. In: IJCAI. vol. 1, pp. 199–204 (2001)
4. Jirkovský, V., Kadera, P., Obitko, M., Vrba, P.: Diagnostics of distributed intelligent control systems: Reasoning using ontologies and hidden markov models. IFAC Proceedings Volumes 45(6), 1315 – 1320 (2012), <http://www.sciencedirect.com/science/article/pii/S147466701633333X>
5. Vrandečić, D.: Ontology evaluation. In: Handbook on ontologies, pp. 293–313. Springer (2009)

Information Modelling in Evidence Based Health Care

Václav Řepa¹,

¹University of Economics, Prague
Nám. W.Churchilla 4, 130 67 Praha
repa@vse.cz

Abstract. The paper introduces the idea to use modelling methods from informatics for the further development of the Evidence Based Health Care (EBHC) in terms of both: its technology support as well as an aid for the development of its methodology.

In the paper we briefly introduce the contents of the EBHC and its predecessor Evidence Based Medicine (EBM) and its main limitations and consequent challenges. Using the example, based on existing EBHC practical guidelines for febrile seizures, we then illustrate the possibilities of the use of modelling techniques for an improvement of the EBHC processes and also the related possibilities of the exploitation of modelling techniques from informatics for further development of the EBHC methodology. In conclusions section we then summarize main opportunities of the application of modelling techniques from informatics in the field of EBHC and outline the needed future work.

Keywords: Evidence Based Health Care, Ontology, Conceptual Modelling, Object Life Cycle Modelling, MMABP, Unified Modelling Language.

1 Introduction - Evidence Based Health Care and Information Modelling

The term "Evidence-based medicine" was introduced in the medical literature in 1991 [5]. According to [3] an original definition suggested the process was "an ability to assess the validity and importance of evidence before applying it to day-to-day clinical problems" [10],[6]. The work presented in this paper aims to the enrichment of the EBHC practices with the power of modelling techniques from informatics. The work presented in this paper is based on the methodology MMABP [13]. MMABP is a methodology for modelling business systems, based on the formal meta-model [2] and continuously being developed [9]. MMABP aims to describe the Real World domain as an equilibrium of intention and causality [12]. For all models MMABP uses standard modelling languages; UML [16] for the ontological / conceptual model and BPMN [1] for the process model. The work, presented in this paper, is currently oriented just on the ontological dimension of the modeled domain, nevertheless we plan also the use of process descriptions in further development of the research.

Particular object classes in the conceptual model represent concepts which identify possible real world objects from the given domain. Relationships among object

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 133-138.*

classes then identify possible links among real objects in the domain. In principle, this “system-oriented” view of objects can simply recognize their existence and mutual context, not their dynamic details. Nevertheless, both real objects and their mutual links are naturally dynamic in the real world; they are changing in time. These dynamic aspects cannot be described in the Class Diagram as it principally describes the system of objects, i.e. the system of mutually parallel processes of changes of linked objects. To model the object's dynamics it is necessary to complete the Class Diagram with the set algorithmic models, each focused on just one particular object in terms of its so-called “life cycle”. Object life cycle expresses the internal dynamics of each object of given class. It describes the mechanism of the object evolution during the time. As the tool for the Object life cycle description, the MMABP methodology uses the State Chart diagram from the UML [16]. Thinking about the objects generally (i.e. in terms of their classes), one has to strongly distinguish between the class and its instance. In the case of the object life this requires determining those points in the life of all objects of the same class, which we will be able to identify, and which it is necessary to identify in order to describe the synchronization of the object life with life cycles of other objects. Such points of the object life are its states. So each object instance “lives its own life” while the common, general structure of lives of all instances of the same class is described as a life cycle of the object class. Each described life cycle has to correspond to the particular object class in the Class Diagram. In such way the State Chart specifies the general mechanism of the life of all possible instances of the given class. Described states and transitions among them consequently correspond to the attributes and methods of the class. Life cycle states represent some kind of the special attribute of the class. This “state” attribute exists implicitly, by the definition, and therefore it should not be expressed in the conceptual model as an attribute of the class, similarly to the object identifier which follows from the definition of the meta-concept class where the identity of the object is defined as implicit. Like the “value of identity” represents particular object (class instance), the “value of state” represents the particular state in the object life cycle in which the given object stands at the moment. Each transition between life cycle states then represents the use of the particular class method.

The model at Figure 2 can be used as an example of the description of class life cycle with the State Chart. Every transition between states is described there by the attribute which represents the reason for the transition. The reason for the transition is an event, i.e. some external influence of the given object.

Combination of the traditional conceptual model in the Class Diagram with the description of life cycles of selected objects increases the traditional extent of the conceptual model, focused on just static modality with the description of the dynamics in modality. In other words, this way we are able to describe also some part of the real world causality which is not possible in the traditional conceptual modelling.

2 Example – Febrile Seizures

For the example of the possible application of modelling techniques from informatics we use the problem of Febrile seizures; one of frequent topics processed in the form of practical guidelines [4],[11]. Febrile seizures are defined as seizures occurring in association with a febrile illness in children between the ages of six months and five years. It is important to differentiate febrile seizures with fever from complex febrile seizures secondary to central nervous system infections. Febrile seizures are relatively benign as compared to complex febrile seizures that have serious sequelae, longer duration, and require further medical follow-up [14]. The motivation for practical guidelines in this field underlines also the fact that simple febrile seizures are relatively prevalent conditions that often causing avoidable anxiety to the children and their parents [7].

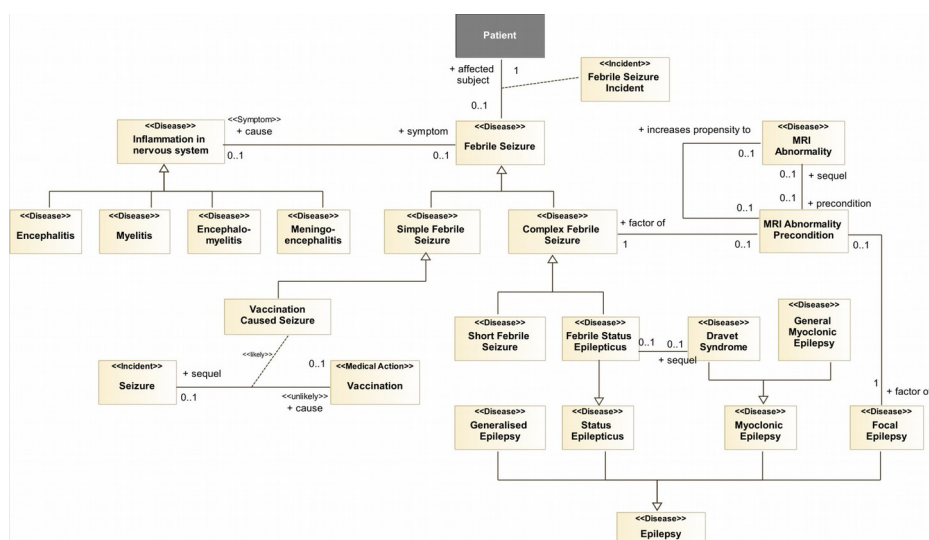


Fig. 1. Febrile Seizures conceptual model - a fragment.

The model at Figure 1 expresses the information about febrile seizures, and some other closely related phenomena. The information has been gathered from several resources, primarily the Simple Febrile Seizures Evidence-Based Guidelines [4]. Additional resources have been used for completion of the contextual information about related topics of febrile seizures, epilepsy, diseases of the nervous system in general, etc [15],[7],[8]. For various reasons, this model contains just a fragment of the complex information about febrile seizures. First, by means of conceptual modelling which description possibilities are naturally limited it is impossible to make a complete picture of such a complex phenomenon, as any medical field is. Second, the model describes a subset of internal relationships among used concepts, selected according to their relevance to the given purpose: to address primarily the problem of simple febrile seizures which is a subject of the source guidelines. Nevertheless,

despite the above discussed limitations, it is obvious that such way of description of the “modal logic of the given field” is much more powerful than the traditional text description currently used as a basic form of Evidence-Based Guidelines.

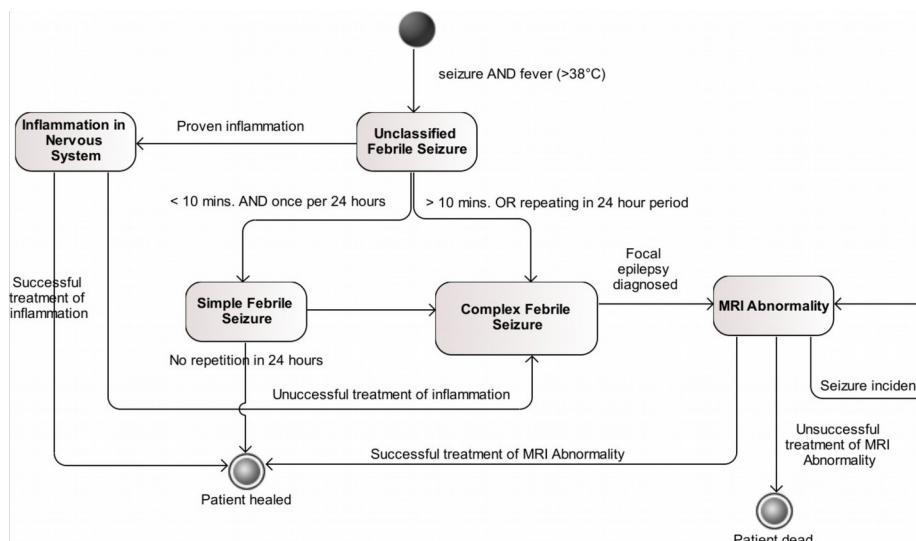


Fig. 2 Febrile Seizure Incident Life Cycle Model

Model at Figure 2 shows the life cycle of possible objects of the class Febrile Seizure Incident. In the conceptual model at Figure 1 it can be found that this class is an association class, it expresses the association between the essential general class Patient and the class Febrile Seizure as a disease (stereotype <<Disease>>). Febrile Seizure Incident thus represents an incident of febrile seizure which happens to the particular patient. The whole structure of mutually associated other classes is connected with the patient just by this association. The life cycle model at Figure 2 thus describes possible ways of the development of the disease starting with the (yet unclassified) febrile seizure according to the modal logic of the domain, following from the conceptual model at Figure 1. The life cycle model is focused on the essential causal relationships, i.e. possible time dependencies of relevant events, associated with the initial incident. The model should be interpreted as follows: the very first state of the life cycle is Unclassified Febrile Seizure. This state occurs after the febrile seizure attack what is characterized in the model as an input event seizure AND fever >38°C. In the very first state the model expects three important possible events. Either the seizure ends within 10 minutes and it is not repeating within 24 hours; in such case the initial incident is classified as a Simple Febrile Seizure and probably does not mean anything serious. Otherwise it is classified as a Complex Febrile Seizure which represents completely different, more serious, possibilities of the further development. The third possible event in this state is the detection of the Inflammation in Nervous System which can explain both the seizure as well as the fever. If successfully treated this state can also result in nothing serious, nevertheless

otherwise it can lead to the state of Complex Febrile Seizure. The model also mentions possible state of MRE Abnormality which can follow the Complex Febrile Seizure if the focal epilepsy is diagnosed. This state can be progressively worsened by subsequent seizure incidents which is modeled here by the self-transition, representing the causal view of the cyclic associations between the concepts MRI Abnormality and MRI Abnormality Precondition in the conceptual model at Figure 1.

3 Conclusions

Unlike most of approaches to apply informatics in the field of healthcare our approach is not oriented just on the support of the evidence of facts. It is aimed on the use of modelling methods from informatics for the further development of the methodology in the field of EBHC. This way our approach follows the use of ontology engineering which also, even indirectly, contributes to the medical research methods and practices forcing them to use the systematic approach to the terminology and other connected aspects of the research. Topics discussed in this paper also represent serious challenges for the future development of the methodology in the field of EBHC. For instance, MMABP technique for modelling the object life cycle includes the rules for achieving both completeness and correctness of the life cycle description. According to these rules all possible state transitions, including the transitions which are not present in the model, should be taken into the account and evaluated. This practice strongly evokes many new research questions which answering can significantly move forward the current knowledge. For instance, in the model at Figure 2 some possible relevant questions may be: (1) Can MRI Abnormality cause Complex Febrile Seizure? If so, under which conditions, what is the related event, what are related medical actions? (2) Can Complex Febrile Seizure turn back to Simple Febrile Seizure? If so, under which conditions, what is the related event, what related medical actions should be performed? (3) Can Inflammation in Nervous System cause just the Simple Febrile Seizure? This question is related to the problem of potential neglect of possible cause of consequent complex febrile seizure.

Our current work is oriented just on the ontological dimension of the modelled domain. In the future we plan to take into the account also the process descriptions of the field of healthcare. Ontological conceptual models can be used for basic definitions of healthcare concepts and their mutual relationships, in terms of the modality of the domain, including also its causality in the form of objects' life cycles. Nevertheless, the healthcare domain, even the field of Evidence Based Healthcare, naturally contains also the process (intentional) dimension. So-called "best practices", an essence of the EBHC, are the best examples of the need to model also the intentionality (recommended practices, guidelines) of this domain. MMABP puts the great emphasis on the consistent integration of both ontological and process models of the Real World. Therefore, in the future steps of this research, we can expect, in this way, yet further significant methodological contribution to the field of EBHC.

References

1. Business Process Model and Notation (BPMN), 2011. OMG Document Number: formal/2011-01-03, Standard document URL: <http://www.omg.org/spec/BPMN/2.0>
2. Business System modelling Specification: <http://opensoul.panrepa.org/meta-model.html>
3. Dawes, M. et al.: Sicily statement on evidence-based practice, BMC Medical Education 5 / 2005.
4. Febrile Seizures: Guideline for the Neurodiagnostic Evaluation of the Child With a Simple Febrile Seizure, Subcommittee on Febrile Seizures, Pediatrics, 2011.
5. Guyatt G: Evidence-Based Medicine., ACP J Club, 1991.
6. Guyatt GH, Rennie D: Users' guides to the medical literature [editorial]. JAMA 1993, ISBN 270:2096-2097.
7. Joanna Briggs Institute: Management of the child with fever. Best Practice, Evidence Based Practice Information Sheets for Health Professionals, Volume 5, Issue 5, 2001.
8. Morris Z., Whiteley W.N., Longstreth W.T., Weber F., Lee Yi-Chung, Tsushima Yoshito et al.: Incidental findings on brain magnetic resonance imaging: systematic review and meta-analysis BMJ Research Paper, 2009.
9. OpenSoul project: <http://opensoul.panrepa.org>.
10. Oxman AD, Sackett DL, Guyatt GH: Users' guides to the medical literature. I. How to get started. The Evidence-Based Medicine Working Group. JAMA 1993, ISBN 270:2093-2095.
11. Patel, A.D., Perry, S.M.: Febrile Seizures: Evaluation and Treatment, JCOM, vol. 24, no. 7, July 2017.
12. Řepa, V. Business system as an equilibrium of intention and causality. E+M Economics and management.XX, 2017, vol.4, pp.219--235. ISSN 1212-3609.
13. Řepa, V.: Business System modelling Specification. In Chu, H., Ferrer, J., Nguyen, T., Yu, Y. (ed.). Computer, Communication and Control Technologies (CCCT '03). Orlando: IIIS, 2003, s. 222--227. ISBN 980-6560-05-1.
14. Sadleir, L., Scheffer, I.: Febrile seizures.in British Medical Journal, 334, 2007, pp. 307-311.
15. Swansen, R.D.: Nursing Interventions in the Prevention of Febrile Seizures, University of Northern British Columbia, 2008.
16. Unified Modelling Language™ (OMG, UML) Infrastructure Specification, version 2.4.1, Object Management Group (OMG) (<http://www.omg.org>).

Acknowledgments: The paper was processed with financial contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague. The work is professionally supported by the Czech Republic (Middle European) Centre for Evidence-Based Healthcare: A Joanna Briggs Institute Centre of Excellence at the Masaryk University in Brno.

Inteligentná analýza dát na FIIT STU

Viera Rozinajová, Petra Vrablecová, Nadežda Andrejčíková, Anna Bou Ezzeddine,
Mária Lucká, Miroslav Blšták, Marek Lóderer, Jaroslav Loeb1

Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava

{viera.rozinajova, petra.vrablecova, nadezda.andrejcikova,
anna.bou.ezzeddine, maria.lucka, miroslav.blstak, marek_loderer,
jaroslav.loeb1}@stuba.sk

Abstrakt. Príspevok podáva informáciu o prebiehajúcom výskume v rámci projektu „Znalostné prístupy pre inteligentnú analýzu veľkých dát“, ktorý rieši Fakulta informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave (FIIT STU) v kolaborácii s Fakultou elektrotechniky a informatiky Technickej univerzity v Košiciach. Projekt je zameraný na získavanie a využitie znalostí a znalostných modelov za účelom vylepšenia výsledkov úloh dolovania v dátach. FIIT STU sa podieľa na výskume inteligentných metód analýzy dát v doménach energetika či inteligentné siete a bioinformatika. Ako riešiteľ sa sústreďí najmä na inkrementálne a adaptívne predikčné a zhukovacie metódy, ktoré sú schopné sa prispôbovať dynamicky meniacim sa dátam (a znalostiam v dátach). Zaoberá sa aj výskumom metód na získavanie znalostí z textu.

Kľúčové slová: znalosti, predikcia, zhukovanie, analýza dát.

1 Úvod

Od októbra 2017 sa Fakulta informatiky a informačných technológií Slovenskej technickej univerzity v Bratislave (FIIT STU) podieľa na riešení 4-ročného projektu financovaného Agentúrou na podporu výskumu a vývoja (APVV) Ministerstva školstva Slovenskej republiky s názvom „Znalostné prístupy pre inteligentnú analýzu veľkých dát“ (APVV-16-0213). Hlavným riešiteľom projektu je Technická univerzita v Košiciach. Projekt je orientovaný na získavanie a integrovanie znalostí do modelov dát meniacich sa v čase za účelom zvýšenia ich presnosti a obohatenia znalostnej bázy zvolenej domény v kontexte úloh dolovania v dátach ako predikcia a zhukovanie.

Znalosti v súčasnom svete zastávajú čoraz významnejšiu úlohu, nakoľko ich aplikácia umožňuje firmám a organizáciám zlepšiť svoje výsledky a získať dominantné postavenie na trhu. Zo sémantického pohľadu môžu byť znalosti o určitej doméne reprezentované množinou konceptov, ktoré sú poprepájané rôznymi vzťahmi. Takúto reprezentáciu môžeme vnímať ako „sémantickú sieť“ alebo ontológiu. V závislosti od vyjadrovacej a odvodzovacej sily tejto reprezentácie, rozlišujeme škálu ontológií od tzv. ľahkých (napr. slovník, hierarchia konceptov) po ťažké (napr. dátový model, taxonómia, deskriptívna logika).

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 139-142.*

Manažment znalostí (zbieranie, organizovanie, spracovávanie, zdieľanie) je výzva, pretože v súčasnosti množstvá zbieraných dát, z ktorých sú znalosti dolované:

- dosahujú enormné veľkosti, neustále pribúdajú, a
- pochádzajú z dynamického prostredia, ktoré sa mení v čase a ovplyvňuje ho veľa faktorov, preto vydolované znalosti majú obmedzenú dobu platnosti.

Preto je žiadúce, aby dáta boli spracovávané *inkrementálne* bez potreby uchovávať všetky v pamäti a aby metódy dátovej analýzy boli *adaptívne*, t. j. vedeli sa prispôbovať meniacim sa vlastnostiam dát v čase a reagovať na nové udalosti. Na FIIT STU sa už dlhodobo venujeme inkrementálnym a adaptívnym predikčným metódam (pozri časť 2), ktoré ale nepracovali s explicitne zadefinovaným doménovým modelom zachytávajúcim doménové znalosti (ktoré by sa navyše dynamicky menili). Ich jediným vstupom bol prúd dát (časový rad) a výstupom predikcia jeho budúcich hodnôt. V rámci tohto projektu overujeme hypotézu, či rozšírenie kontextu vstupných udalostí o nové príznaky odvodené zo znalostného modelu a doménové pravidlá z neho odvodené zlepšia presnosť analytických prístupov aplikovaných na príslušné dátové vzorky.

Okrem skúmania vplyvu zapojenia dynamického znalostného modelu do predikcie, je projekt zameraný aj na skúmanie jeho vplyvu na získavanie znalostí z neštruktúrovaných dát a vplyvu na výsledky metód popisnej analýzy dát. V týchto oblastiach sa čiastočne realizujeme v úlohách dolovanie znalostí z textu a zhlukovanie (pozri časť 3).

2 Adaptívne inkrementálne predikčné metódy

Pri skúmaní predikčných metód vhodných pre inteligentnú analýzu veľkých objemov dát sme sa na metódy pozerali z dvoch hlavných hľadísk: *spôsob spracovania dát* a *spôsob adaptácie na zmeny* v dátach.

Z hľadiska spracovania dát sme rozlišovali dávkové (angl. offline) metódy, inkrementálne metódy s čiastočnou pamäťou (napr. klzavé okno) a online metódy, ktoré spracovávajú dáta jeden-po-jednom a dokážu zabúdať neaktuálne a nepotrebné záznamy. Sústredili sme sa najmä na možnosti úprav existujúcich dávkových metód na online alebo minimálne inkrementálne metódy s menšou pamäťou a výpočtovou zložitou [7,8].

Z hľadiska adaptácie na zmeny v dátach sme skúmali metódy s priamou a nepriamou adaptáciou. Metódy s priamou adaptáciou prispôbujú predikčný model alebo jeho časti len ak detegujú zmenu v prúde dát [7]. Naopak, metódy s nepriamou adaptáciou sa prispôbujú priebežne v intervaloch alebo po príchode každej novej dátovej vzorky bez explicitného sledovania zmien [2,3,8]. Tento prístup sme využili najmä v prístupoch, kedy sme tvorili predikcie váhovanou kombináciou predikcií väčšieho množstva (súboru) rôznorodých modelov. Pre výpočet optimálnych váh sme využívali biologicky inšpirované optimalizačné algoritmy [2,3].

V súčasnosti skúmame možnosti obohatenia týchto modelov o znalostný doménový model – či už pri detekcii zmien v prúde dát alebo v samotnej predikcii. Výskum týchto metód sme doteraz realizovali v doméne energetiky pri predikcii spotreby elektrickej energie, kde je možné merania spotreby doplniť o údaje o počasi, spotrebiteľoch a pod.

3 Aplikačné domény skúmané na FIIT STU

Doména, ktorej sa venujeme na fakulte dlhodobo, je *energetika a inteligentné siete* [6]. Ako sme spomenuli na konci časti 2, v tejto doméne sme riešili najmä úlohu predikcie sumárnej spotreby elektrickej energie skupiny spotrebiteľov. Predikciu sme sa snažili zlepšiť nielen úpravou predikčných metód, ale aj zhlukovaním spotrebiteľov s podobnou spotrebou [4]. Zaoberali sme sa však aj ďalšími komponentami inteligentnej siete, napr. fotovoltikou, veľkokapacitnými batériami či mikrogridmi. Riešili sme úlohy z pohľadov rôznych účastníkov trhu s elektrinou (napr. zákazník, dodávateľ, distribútor) ako predikcia výroby, cien energie, odporúčanie vhodných fotovoltických panelov, veľkokapacitných batérií a ich parametrov, riadenie správania batérie pre dosiahnutie optimálneho odberu zo siete, ceny za elektrinu či výkyvy v sieti, skúmanie vzťahov medzi týmito kritériami, odporúčanie spotreby, motivovanie zákazníka k šetreniu energie prostredníctvom rôznych aplikácií a pod. [5].

V súčasnosti sa začíname intenzívnejšie venovať doméne *bioinformatika*, kde sme doteraz riešili analytické úlohy súvisiace so sekvenovaním DNA a bunkovou cytometriou. Bioinformatické dáta sa vyznačujú najmä svojou mnohorozmernosťou a môžu obsahovať stovky až tisíce atribútov o jednej inštancii. V spolupráci s Biomedicínskym centrom Slovenskej akadémie vied chceme analyzovať dáta o onkologických pacientoch a hľadať korelácie medzi atribútmi, ktoré by mohli indikovať zvýšené riziko ochorenia. Obdobnú úlohu riešime aj v oblasti výskumu Alzheimerovej choroby, kde hľadáme možnosti včasnej diagnostiky, predikcie a prevencie. Znalostný model v týchto oblastiach obsahuje najmä vedomosti o jednotlivých atribútoch, ich prípustných hodnotách a vzťahoch medzi nimi.

Doména, v ktorej je znalostný model neoddeliteľnou súčasťou, sú *výučbové systémy*, v ktorých model reprezentuje jednak znalosti obsiahnuté vo výučbových textoch a takisto znalosti, ktoré si už osvojil študent. Na vytvorenie znalostného modelu sa používajú metódy dolovania v textoch a spracovanie prirodzeného jazyka. V poslednej dobe sme skúmali využitie týchto techník pre generovanie otázok z textu [1].

Všetky uvedené doménové oblasti majú potenciál využitia znalostí a znalostných modelov v úlohách dolovania v dátach. Avšak pre niektoré z nich neexistuje štandardný doménový model, a preto je výzvou nielen ho zakomponovať do predikcie, zhlukovania a pod., ale aj vytvoriť a spravovať/adaptovať v čase.

4 Záver

V aktuálnom projekte v spolupráci s Technickou univerzitou v Košiciach sa FIIT STU zameriava na výskum adaptívnych a inkrementálnych predikčných modelov, ktoré sú schopné spracovávať kontinuálne veľké objemy dát a prispôbovať sa zmenám v dátach, ktoré vyplývajú z dynamickej povahy prostredia. V súčasnosti skúmame možnosti a vplyv obohatenia modelov o znalostný model domény v kontexte prediktívnych a deskriptívnych úloh dolovania v dátach v doménach energetika, bioinformatika či výučbové systémy.

Literatúra

1. Blšták, M., Rozinajová, V.: Machine Learning Approach to the Process of Question Generation. In: *Text, Speech, Dialogue. TSD 2017. LNCS 10415*, K. Ekštejn, V. Matoušek (Eds.), Springer, Cham, Switzerland (2017), 102–110.
2. Grmanová, G. et al.: Application of Biologically Inspired Methods to Improve Adaptive Ensemble Learning. In: *Adv. Nat. Biol. Inspired Comput. AISC 419*, N. Pillay et al. (Eds.), Springer, Cham, Switzerland (2016), 235–246.
3. Halaš, P. et al.: Prediction of electricity consumption using biologically inspired algorithms. In: *2017 IEEE 14th Int. Sci. Conf. Informatics*, V. Novitzká et al. (Eds.), IEEE, Piscataway, NJ, USA (2017), 98–103.
4. Laurinec, P. et al.: Adaptive Time Series Forecasting of Energy Consumption Using Optimized Cluster Analysis. In: *2016 IEEE 16th Int. Conf. Data Min. Work. ICDMW*, C. Domeniconi et al. (Eds.), IEEE, Piscataway, NJ, USA (2016), 398–405.
5. Loeb, J. et al.: Multi-objective Optimization for Power Load Recommendation Considering User's Comfort. In: *Data Anal. Renew. Energy Integr. Informing Gener. Distrib. Renew. Energy. DARE 2017. LNCS 10691*, W. Woon et al. (Eds.), Springer, Cham, Switzerland (2017), 26–32.
6. Rozinajová, V. et al.: Computational intelligence in smart grid environment. In: *Comput. Intell. Multimed. Big Data Cloud with Eng. Appl.*, 1st ed., A.K. Sangaiah et al. (Eds.), Academic Press, Cambridge, MA, USA (2018).
7. Vrabecová, P. et al.: Incremental Adaptive Time Series Prediction for Power Demand Forecasting. In: *Data Min. Big Data LNCS 10387*, Y. Tan et al. (Eds.), Springer, Cham, Switzerland (2017), 83–92.
8. Vrabecová, P. et al.: Smart Grid Load Forecasting Using Online Support Vector Regression. *Comput. Electr. Eng.* 65 (2018) 102–117.

Pod'akovanie: Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0213.

Annotation:

Intelligent analysis of data at FIIT STU

The research at Faculty of Informatics and Information Technologies of Slovak University of Technology in Bratislava (FIIT STU) related to the ongoing project called “Knowledge-based Approach to Intelligent Big Data Analysis”, financed by Slovak Research and Development Agency, is presented. The 4-year project that begun in October 2017 is led by Technical University of Košice. It is oriented on acquisition and integration of knowledge in data mining tasks, such as prediction and clustering, that should improve their accuracy and result in intelligent analysis of data. The research at FIIT STU is focused mostly on adaptive incremental prediction and clustering methods in several application domains, such as smart grid, power engineering and bioinformatics. Partly, it also deals with knowledge acquisition from text and text mining.

Adaptívne zložené klasifikačné modely pre detekciu udalostí v prevádzke IT služieb

Martin Sarnovský ¹, Katarína Oravcová ¹, Ján Marcinko ¹

¹Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická Univerzita v Košiciach, Letná 9, 042 00 Košice
matin.sarnovsky@tuke.sk, katarina.oravcova@tuke.sk,
jan.marcinko@student.tuke.sk

Abstrakt. Cieľom tohto článku je podať prehľad v oblasti spracovania a analýzy dátových prúdov, ktoré sa vyznačujú konceptovým driftom pomocou adaptívnych metód. Naším cieľom je aplikovať tieto metódy na dátach z domény manažmentu IT služieb pri úlohách detekcie rôznych typov udalostí v IT infraštruktúre. Článok prezentuje výsledky získané použitím adaptívnej metódy na dátach záznamov zo sieťovej prevádzky IT infraštruktúry. Dáta sa vyznačovali konceptovým driftom náhleho typu, ktorý navrhnutý adaptívny klasifikačný model kompenzuje. Použili sme adaptívny zložený klasifikátor založený na baggingu s implementovaným mechanizmom pre váhovanie čiastkových klasifikátorov. Výsledky dosiahnuté týmto modelom porovnávame s pôvodným adaptívnym baggingom. V závere práce sa venujeme zhodnoteniu výsledkov a sumarizujeme možné pokračovanie výskumu v tejto oblasti.

Kľúčové slova: konceptový drift, zložené klasifikátory, prúdové dáta.

1 Úvod

V súčasnosti sú dáta generované v obrovských objemoch, ktoré je následné potrebné analyzovať. Dáta pribúdajú v čase, môžu pochádzať z rôznych zdrojov a môžu mať rôznu štruktúru. Rýchle spracovanie a následná analýza predstavuje výzvu nie len z technického ale aj z algoritmického hľadiska. Jedným z aspektov veľkých dát je rýchlosť generovania a zároveň rýchlosť spracovania týchto údajov [2]. Prichádzajúce dáta často nie sú statické ale prichádzajú vo forme dátových prúdov [1, 3]. Dátový prúd predstavuje usporiadanú sekvenciu dátových položiek, ktoré môžu pozostávať iba z jednoduchého páru atribút a hodnota, ale môžu mať aj komplexnejšiu štruktúru [4, 5]. Jedným z problémov, s ktorým je potrebné sa vysporiadať pri analýze dátových prúdov je tzv. konceptový drift [6]. Tento fenomén sa vyskytuje v nestacionárnom a dynamicky sa meniacom prostredí, kde môžu nastať zmeny v distribúcii dát alebo môže dôjsť k zmene vzťahu vstupných dát k cieľovému atribútu v priebehu času. Dôsledkom týchto zmien je fakt, že natrénovaný model už nebude dosahovať relevantné výsledky na nových dátach. Na zachytenie driftu používame detektory, ktoré zachytávajú signály o zmenách z dátovej distribúcie. Pokiaľ detektor zachytí signál o detekcii driftu je potrebné čo najskôr nahradiť aktuálny model novým [5]. Detektory driftu sa potom

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 143-147.*

zvyknú používať v kombinácii s adaptívnymi klasifikátormi, ktoré si kladú za cieľ prispôbiť sa driftujúcim dátam v čase.

Zaujímavou aplikačnou doménou môže byť oblasť manažmentu IT služieb. Zariadenia IT infraštruktúry v podnikoch vyprodukujú množstvo dát, rovnako ako aj používatelia služieb. Dáta majú často prirodzene prúdový charakter (napr. logy z jednotlivých zariadení alebo záznamy o aktivitách používateľov), prostredie sa vyznačuje vysokou dynamikou vzorov v dátach. Z tohoto hľadiska sa oblasť adaptívnych modelov pre klasifikáciu javí ako relevantná. Prediktívne modely sa v IT manažmente používajú hlavne pre detekciu rôznych typov udalostí, ktoré ovplyvňujú prevádzku služieb.

2 Adaptívne zložené klasifikátory

Existuje viacero členení algoritmov pre učenie klasifikátorov na dátových prúdoch. Z pohľadu konceptového driftu je možné rozlíšiť dva základné prístupy - *aktívny* a *pasívny* [6]. V aktívnych prístupoch je model aktualizovaný až potom, ako detektory detegujú zmenu v dátovom prúde. Na druhej strane klasifikátory využívajúce pasívne prístupy aktualizujú model v pevne definovaných intervaloch pravidelne a nezávisle na tom, či došlo ku driftu v dátovom prúde. Okrem adaptácie modelov na nové príklady je potrebné zvážiť aj to, ktoré dáta použiť pri následnom preučení modelu [8].

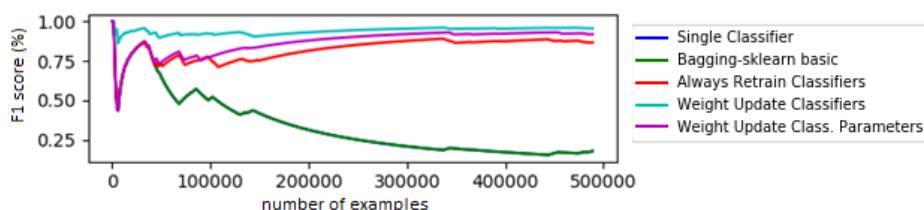
Zložené klasifikátory hrajú kľúčovú rolu pri klasifikácii dátových prúdov, pretože sú jednoducho adaptovateľné na nestacionárne prostredie. Sú reprezentované množinou niekoľkých klasifikátorov, ktorých jednotlivé predikcie sú kombinované do finálneho výsledku klasifikácie. Kľúčovým aspektom zložených modelov je správny výber jednotlivých klasifikátorov, optimálny riešením je výber vzájomne sa dopĺňajúcich čiastkových modelov [5]. Čiastkové predikcie jednotlivých klasifikátorov sú zvyčajne váhované a skombinované do výsledného rozhodnutia. Pri zložených metódach je dôležité spravovanie dynamickej množiny klasifikátorov. Zložený klasifikátor umožňuje začleniť nový komponent, aktualizovať niektorý z existujúcich komponentov alebo aktualizovať váhy existujúcich komponentov [7]. Pokiaľ dôjde k znižovaniu výkonnosti zloženého klasifikátora, niektoré z komponentov môžu byť naopak odstránené [9]. Zložené klasifikátory pracujúce s prúdovými driftujúcimi dátami vieme rozdeliť na:

- a) *Chunk based* - pracujú s pevným počtom trénovaniach dát v bloku, ktoré sú vopred známe. Modelovanie a aktualizácie klasifikátorov prebiehajú, až keď sú známe všetky príklady z nového bloku.
- b) *Online* - trénovanie prebieha stále v čase, keď sa objaví nový príklad. Tieto modely sú vhodné pre spracovanie dát v reálnom čase.

3 Zložené modely pre predikciu udalostí v prevádzke IT

Naším cieľom bolo navrhnúť a implementovať spôsob adaptácie čiastkových klasifikátorov v adaptívnych zložených modeloch. Rozšírili sme základnú adaptívnu bagging metódu o aktualizáciu čiastkových klasifikátorov a kombinovanie nových dát s historickými dátami pri aktualizácii modelu. Navrhnutý algoritmus v každej iterácii

vyhodnocuje prichádzajúce prúdové dáta a aktualizuje čiastkové modely, ktorými reaguje na možný výskyt zmeny konceptov v nových dátach. Na začiatku metóda vytvára viacero rovnako veľkých podmnožín dát, ktoré slúžia ako trénovacie údaje pre jednotlivé čiastkové klasifikátory. Po inicializácii modelu a úvodnom natrénovaní na počiatkovej množine historických dát nasleduje proces iterácie na postupne pribúdajúcich dátach z prúdu po oknách definovanej veľkosti. V každej iterácii sa najprv natrénujú čiastkové klasifikátory, kde potom každý samostatne vyhodnotí súbor aktuálnych dát z poslednej iterácie. Predikované hodnoty sa uložia a porovnajú sa so skutočnými hodnotami z dátového prúdu. Posledný proces, ktorý sa vykonáva v cykle je aktualizácia trénovacej množiny aktuálnymi príkladmi. Ak je trénovacie okno (rozsah prúdových dát, ktoré sa použije na pretrénovanie modelu) väčšie ako iteračné (rozsah mikro-dávok v ktorých sú spracovávané prúdové dáta), tak do nového výberu trénovaných dát je zahrnutá celá aktuálna množina dát z poslednej iterácie. Zvyšok je vybraný z dát, ktoré sa už v trénovanej množine nachádzali. Takto zabezpečíme kombináciu starších dát s aktualizovanými dátami pre pretrénovanie klasifikátorov v nasledujúcej iterácii.



Obr. 1. Vývoj metriky F1

Pri experimentoch sme použili dátový súbor z KDD Cup 1999 [10]. Tento dátový súbor predstavuje výpisy záznamov zariadení v LAN sieti zozbierané za obdobie 9 týždňov. Každý záznam predstavuje jednu aktivitu, záznamy sú zoradené za sebou vo frekvencii vzorkovania 2 sekundy. Rovnako boli používané v simulovanom prúde týchto dát pre vyhodnotenie adaptívneho modelu. Vzorka obsahuje 494 021 záznamov, 41 atribútov a cieľový atribút, ktorý určuje či záznam je normálny údaj alebo nejaký typ sieťového útoku. Cieľový atribút sme transformovali na binárne hodnoty indikujúce, či sa jedná o útok alebo nie. Modely boli otestované na viacerých kombináciách veľkosti trénovacej množiny a iteračného okna. Ako základné klasifikátory pre bagging boli použité rozhodovacie stromy. Obr. 1 porovnáva vývoj F1 metriky kombinujúcej presnosť a návratnosť. Implementovaný klasifikátor (uvedený ako Always Retrain Classifier) je porovnávaný s nemodifikovanými metódami a ďalšími variantami adaptívneho baggingu, ktoré rozširujú predstavený prístup o rôzne spôsoby aktualizácie váh čiastkových modelov [11]. Tab. 1 sumarizuje priemerné metriky vzhľadom na rôzne veľkosti iteračného okna.

Iteračné okno	Metrika	Základný Bagging	Adaptívny Bagging
50	Presnosť	70.58	82.05
	Návratnosť	66.39	75.29
100	Presnosť	62.13	80.42
	Návratnosť	59.58	74.02
500	Presnosť	66.36	76.60

	Návratnosť	62.99	73.68
1000	Presnosť	65.41	66.67
	Návratnosť	62.19	63.49

Tab. 1. Výsledky pri rôznych veľkostiach iteračného okna

4 Zhrnutie

Cieľom článku bolo zhrnúť oblasť adaptívnych zložených metód pre klasifikáciu na dátových prúdoch. Predstavili sme jednoduchú modifikáciu bagging metódy a použili ju na dáta z domény prevádzky IT s náhlym konceptovým driftom. Úvodné experimenty ukázali pre adaptívne zložené metódy dobré výsledky. Ich presnosť a rýchlosť aktualizácie závisela od toho, aká veľká skupina dát bola použitá v každej iterácii a akú veľkú mali trénovaciu množinu pre aktualizácie klasifikátorov. Z pohľadu ďalšieho výskumu v tejto oblasti, existuje viacero smerov, ktoré predstavujú výzvu. Pre efektívnejšie spracovávanie prúdových dát sa ako zaujímavá oblasť javí využitie kontextuálnej informácie alebo znalostných modelov v klasifikácii. Tá predstavuje expertné znalosti z danej domény, ktoré by mohli byť použité pre spresnenie klasifikácie alebo včasnú detekciu driftu. V doméne prevádzky IT služieb by takýto znalostný model by mohol zachytiť rôzne vzťahy medzi udalosťami, ku ktorým v IT infraštruktúre dochádza a umožnil odvodiť informácie, ktoré nie sú explicitne zrejmé zo samotných dát a môžu byť relevantné k riešenej úlohe. Pomocou takýchto modelov by bolo možné detegovať sekvencie rôznych typov udalostí, ktoré definujú neštandardné vzory správania sa alebo práve naopak, bežnú prevádzku. Využitie existujúcich a tvorba nových doménových modelov a extrakcia znalostí z nich pre potreby klasifikácie teda predstavujú zaujímavé smery ďalšieho výskumu v tejto oblasti.

Literatúra

- [1] Chen, M., Mao, S., Liu, Y., 2014. Big data: A survey, in: *Mobile Networks and Applications*. Kluwer Academic Publishers, pp. 171–209.
- [2] Ebner, K., Bühnen, T., Urbach, N., 2014. Think big with big data: Identifying suitable big data strategies in corporate environments, in: *Proceedings of the Annual Hawaii International Conference on System Sciences*. IEEE Computer Society, pp. 3748–3757.
- [3] Laha, A. K., 2017. Analytically Yours: Analysis of Data Streams, in: *International Journal of Business Analytics and Intelligence*, Indian Institute of Management, 1–2
- [4] Gama, J., 2010. *Knowledge Discovery from Data Streams*. CRC Press
- [5] Krawczyk, B., Minku, L.L., Gama, J., Stefanowski, J., Woźniak, M., 2017. Ensemble learning for data stream analysis: A survey. *Information Fusion* 37, 132–156.
- [6] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A., 2014. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4), 1–37
- [7] Brzezinski, D., Stefanowski, J., 2017, Stream Classification, In: *Encyclopedia of Machine Learning* (Springer). Poznan University of Technology, Poznan, Poland

- [8] Bifet, A., Gavalda, R., 2007. Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM international conference on data mining (pp. 443-448). Society for Industrial and Applied Mathematics.
- [9] Wozniak, M., Grana, M., Corchado, E., 2014. A survey of multiple classifier systems as hybrid systems. Information Fusion, 16:3– 17, 2014.
- [10] KDD Cup 1999 Data: Abstract [online]. [2018-01-14]. Available from: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
- [11] Marcinko, J., 2018. Analýza dát pomocou adaptívnych metód, Diplomová práca. Technická univerzita v Košiciach.

Poděkování: Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0213 a VEGA č. 1/0493/16

Annotation: The aim of this paper was to provide an overview in the field of processing data and analysis data streams using adaptive methods. The goal was to develop an evaluate the adaptive bagging method modification and to discover its behaviour on dataset containing the concept drift from IT management domain. We designed and implemented an ensemble adaptive method that we compared with basic model. Finally, we present several options of future work and possibilities of related research. One of the possible goals is to combine adaptive classifiers with domain knowledge obtained from domain-specific semantic models to improve the classification performance.

Authority and Hub Mining from the Structure of a Social Network

Kristina Machova¹ and Jozef Kováč²,

¹ Dept. of Cybernetics and Artificial Intelligence, Technical University, Letná 9,
042 00, Košice, Slovakia

kristina.machova@tuke.sk

² Exponea, Twin City B, Mlynské Nivy 12, 821 09 Bratislava, Slovakia

jozo.kovac@exponea.com

Abstract. The paper introduced a problem of authoritative users searching in a structure of a special social network – the telecommunication network. The focus of this paper is on a classification of all actors of the network into several different groups (casual, member, low hub, strong hub, low authority and strong authority), not only into authoritative or not authoritative ones. This approach is based on a decision tree generation and CRISP-DM methodology. It can be used by a telecommunication association for finding special and valuable customers.

Keywords: Authority identification, telecommunication network, mining in the structure, customers roles

1 Introduction

Various services of different platforms of the social web create a big amount of texts, which offers interesting possibilities for a mining useful information from a conversational content and also structure. A way how to dedicate useful information from newsgroup discussions is presented in [3]. Our approach is based on using methods of Social Network Analysis (SNA) utilizing previous experiences [4] and also on methods of machine learning [6]. SNA involves various approaches: searching suitable metrics for networks analysis, prediction of changes in social networks, developing algorithms for social networks monitoring, visualization of social networks, research of social networks and their time characteristics, etc.

Our focus is on the telecommunication social network. Nowadays, also telecommunication organizations need to identify and understand their customers to offer them services, which the customers want and to prevent their leaving for a competitor. Therefore, it is very important to identify an authority or a hub user, which attract other common users (their friends or family members) to remain in the given telecommunication society. Work [1] introduces an approach to identifying authoritative actors in question-answering forums. Our approach is based on mining in the structure of the telecommunication network.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 149-154.*

According to the philosophic encyclopaedia, an authority is an accomplished ability of some person (society, organization, state etc.) to affect somebody. The authority can be formal and informal [2]. We are interested in informal authorities. The authority mining within various types of networks is important for corporations, which need to understand and know their customers. Therefore, they collect records about customers. The accurate analysis of these data is a matter of principle [5]. We carried out an analysis of data of a telecommunication network with the aim to classify customers into previously defined classes.

2 Data for Authority Mining

We had for disposal two text files containing data from a telephone central, one with data about actors of the network and second one with data about realized calls (relations between actors). This input data set was divided into training and testing set in a ratio 80:20. These data were labelled and feed to input of program R, which analysed these data. Consequently a decision tree was generated for a classification of all actors of network into given categories. Our approach was modelled on the basis of the standard CRISP-DM methodology in following steps: goal understanding, data understanding, data preparation, data analysis, modelling, evaluation and testing, and deployment.

The telecommunication society had to know its customers to offer them satisfactory and valuable services. It was important for them to offer appropriate targeted products. Other reasons for our solution were: to improve a reputation of the corporation, decrease of expenses, increase sales and increase a satisfaction of customers. These reasons had to pass all phases of our solution from the *goal understanding* to employment of the solution in praxis.

For solving this goal, an appropriate data had to be selected on the basis of *data understanding*. The appropriate data were selected in the table format CSV (Comma Separated Values). The first file contained records about realized calls, which represented relations between actors. We were interesting in the following attributes within this first file: call id, caller (who called), called (who received the call), start (date and time) and duration (length of the call). The second file contained records about actors of all calls with the following attributes: name, address and telephone number in the international format. These data contained errors, missing values, different format. They had to be reduced, aggregated, and modified.

Data Preparation: data obtained from the telecommunication society were transformed into the MySQL database. Subsequently, new data were cleaned. The last step was the creation of a table with all useful information about customers and their normalization using the min-max normalization. This pre-processing of data was the most time consuming phase.

3 Telecommunication Network Analysis

From the pre-processed data of training and testing sets, a network was modelled. This network contained actors (represented by nodes) and relationships between them (represented by edges). Such network was visualized using program R. The visualization of our training set (246 nodes and 494 edges) and testing set (48 nodes and 62 edges) is illustrated in Fig.1. in the form of K-core decomposition. It is clear from the Fig.1, that nodes with a greater value of the Core are situated nearer to the centre of the graph.

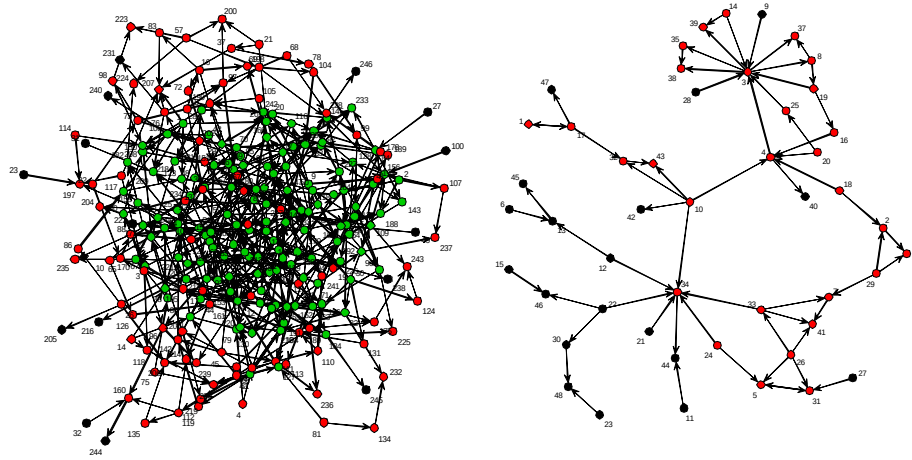


Fig. 1 K-core decomposition of the training set (left) and the testing set (right). All nodes with Core = 1 have black color and belong to a chain. Nodes with Core = 2 have red color and belong to a circle. Nodes with Core = 3 have green color and belong to quaternion.

For a graph visualization, also Fruchterman-Reingold method [5] was used. It solves an optimal location of nodes within the modelled network. It enable to define some smaller sub networks in the process of modularization of a network. The process of modularization of our training and testing sets by program R is illustrated in Fig.2.

5 Authority Modelling

The pre-processed data were not annotated before an input to a modelling phase. So, we represented the goal attribute by values of subnetworks obtained in the previous modularization of the network. At first, an appropriate number of subnetworks - segments had to be stated. Too high number of segments can cause that the clear description of these segments will be difficult. On the other hand, too low number of segments can lead to an impossibility to capture the nature of these segments. The recommended number of segments is from 5 to 12. We stated the number of segments to be 6 on the basis of data analysis. The labelling of data was done with a help of expert.

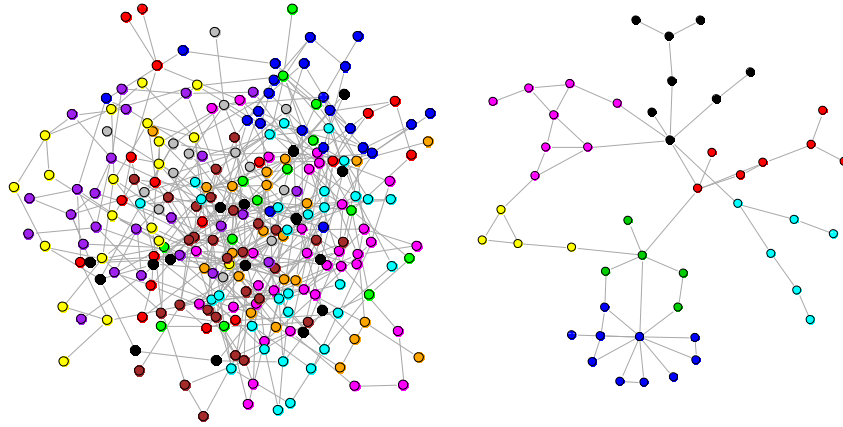


Fig. 2 Modularity of the training set (left, 11 subnetworks) and of the testing set (right, 11 subnetworks). Each colour represents one subnetwork.

So we have defined following particular segments on the basis of previous statistical analysis and in a cooperation with expert:

- CASUAL – actor has a low number of friends and so he/she is connected to a low number of sub-networks through his/her friends.
- MEMBER – actor has more friends than “causal” and all his/her friends are from the same sub network.
- STRONG HUB - actor has a low number of friends but his/her friends belong to many various sub-networks.
- LOW HUB - actor has an average number of friends and the great majority of his/her friends are from the same sub-network as the actor but some friends belong to other subnets.
- STRONG AUTHORITY - actor is characterised by a great number of friends who mainly belong to the same sub-network as the actor.
- LOW AUTHORITY - actor is characterised by a great number of friends who do not belong to the same sub-network as the actor.

Our model for classification of actors of the telecommunication society is based on a decision tree [6], generated by the C4.5 machine learning algorithm from the training set, using the programming language R and the library RWeka. An illustration of the decision tree is in Fig. 3. The generated decision tree in the program R follows:

```

con_net_norm <= 0
| all_friends_norm <= 0: CASUAL (15.0)
| all_friends_norm > 0: MEMBER (41.0)
con_net_norm > 0
| all_friends_norm <= 0.625
| | all_friends_norm <= 0.25
| | | all_friends_norm <= 0.125: STRONG HUB (17.0)
| | | all_friends_norm > 0.125
| | | | con_net_norm <= 0.2: LOW HUB (25.0)
| | | | con_net_norm > 0.2: STRONG HUB (12.0)
| | all_friends_norm > 0.25: LOW HUB (108.0)
| all_friends_norm > 0.625
| | con_net_norm <= 0.4: STRONG AUTHORITY (5.0)
| | con_net_norm > 0.4: LOW AUTHORITY (23.0)

```

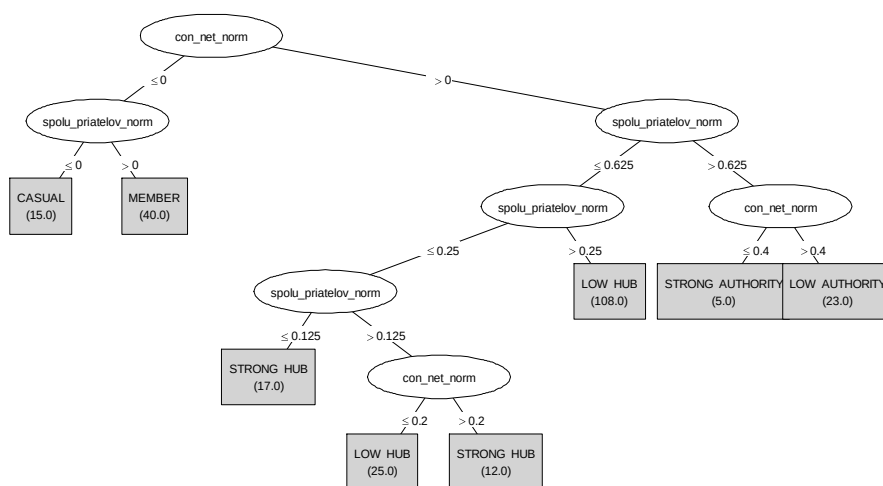


Fig. 3 Decision tree for classification of telecommunication customers

6 Evaluation and Testing

The generated classification model was tested on precision. What was really surprising for us, it was zero error of classification on testing set (see Table 1). Not so much authorities (low and strong) were found. Better situation was in hubs searching. The true is that a big number of found authorities and hubs was not the goal of our approach. The goal was to find real authorities and real hubs, which could attract other actors and users to remain the members of the telecommunication network.

Table 1. Precision of testing results of classification model on the testing set.

	Member	Low authority	Low hub	Casual	Strong authority	Strong hub
Member	22	0	0	0	0	0
Low authority	0	0	0	0	0	0
Low hub	0	0	4	0	0	0
Casual	0	0	0	13	0	0
Strong authority	0	0	0	0	1	0
Strong hub	0	0	0	0	0	8

7 Conclusions

The paper introduces the approach to the authority mining within telecommunication networks based on a structure analysis of this network. The aim was to search for critical – authoritative users of the network. The model was developed and used for the classification of customers of the telecommunication network into following classes: causal, member, strong hub, low hub, strong authority and low authority. The greatest attention was paid to the identification of authorities and hubs. The telecommunication society obtained important information who were the worst customers to be motivated to remain among its clients. This approach is surprisingly precise.

Acknowledgements. The work was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under VEGA grant No. 1/0493/16 and by the Slovak Research and Development Agency under the contract No. APVV-16-0213 and the contract No. APVV-17-0267.

References

1. Bouguessa, M., Dumoulin, B., Wang, S.: Identifying authoritative actors in question-answering forums – the case of Yahoo! Answers. Department of Computer Science, University of Sherbrooke, Quebec, CA, 2008, 1-9.
2. Lankers, R. D.: Credibility on the internet: shifting from authority to reliability. Journal of Documentation, 2008, Vol. 64 Iss: 5, 667 – 686, ISSN: 0022-0418.
3. Mach, M., Lukáč, G.: A Dedicated Information Collection as an Interface to Newsgroup Discussions, IIS 2007 - 18th international conference on Information and Intelligent Systems, September 12-14, 2007, Varazdin, Croatia, 163-169, ISBN 978-953-6071-30-2.
4. Machová, K., Rakuščinec, T.: Dynamic Analysis of Social Networks. In: American Journal of Intelligent Systems Vol. 2, no. 6(2012), p. 148-156, ISSN 2165-8978
5. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge, Press Syndicate of the University of Cambridge, 1994. 1-833, ISBN 0 521 38707 8.
6. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2005, 1-558, ISBN: 0-12-088407-0.

Stratégie a zámery vyhľadávania v elektronických obchodoch zameraných na zľavy: prípadová štúdia

Andrea Hrčková¹, Miriam Ondrišová², Jakub Šimko³, Sandra Kostová⁴

^{1, 2} Katedra knižničnej a informačnej vedy, Ff UK v Bratislave
Šafárikovo námestie 6, 814 99 Bratislava
andrea.hrckova@uniba.sk
miriam.ondrisova@uniba.sk

^{3, 4} Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU
v Bratislave
Ilkovičova 2, 842 16 Bratislava
jakub.simko@stuba.sk
kostova.sandra@gmail.com

Abstrakt. Sledovanie zámeru a spôsobu interakcie s elektronickými obchodmi je dôležité najmä pre účely skúmania a uspokojovania individuálnych potrieb ich používateľov. Pri analýze týchto dát využívame metódu analýzy transakčných logov. Ukazuje sa, že v prípade sledovaného elektronického obchodu je najčastejšou prieskumovou stratégiou prehliadanie (browsing), čo naznačuje intuitívny a náhodný spôsob vyhľadávania na základe vizuálnych podnetov. Väčší pomer transakcií v sledovanom e-obchode však vykonali používatelia, využívajúci stratégiu krátkého prieskumu. Zámer používateľa (informačný, navigačný a transakčný) závisí z veľkej miery od témy obsahu. V príspevku preto taktiež kategorizujeme zámery používateľov v kontexte vyhľadávania zliav. Tieto poznatky možno ďalej využiť pri návrhoch zlepšení informačnej architektúry elektronického obchodu aj personalizovaného odporúčania informácií jeho používateľom.

Kľúčová slová: Prieskumová stratégia, zámer vyhľadávania, vyhľadávanie v elektronických obchodoch, informačné správanie

1 Úvod

Informačné správanie používateľov možno sledovať na základe ich zámeru a zvolenej prieskumovej stratégie. Prieskumovú stratégiu definujeme ako “*voľbu krokov a metód, ktoré si používateľ určuje pri riešení informačného problému*” [1]. Používatelia v praxi kombinujú štyri najznámejšie prieskumové stratégie: prezeranie (browsing), krátky prieskum reprezentovaný zadaním vyhľadávacieho dotazu, monitoring a berrypicking, teda exporatívne a intuitívne zbieranie kúskov informácií z rôznych zdrojov.

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 155-161.*

Zámer vyhľadávania je vyjadrením afektívneho, kognitívneho a situačného cieľa alebo potreby používateľa pri interakcii s vyhľadávacím nástrojom prostredníctvom dotazu [2].

V oblasti klasifikácie vyhľadávacích dotazov na základe zámeru sa využíva najmä Broderova taxonómia [3], ktorá delí dotazy na informačné, navigačné a transakčné. Navigačný zámer implikuje zadanie konkrétnej URL adresy alebo značky produktu či služby pri vyhľadávaní. Informačný zámer charakterizuje získavanie informácií s cieľom ich porovnávaní. Je teda možné ho zistiť z charakteru interakcie a obsahu prezeranej informácie v kombinácii s informáciou o dokončení transakcie. Transakčný vyhľadávací zámer vyžaduje ďalší krok, aby bola aktivita pre používateľa užitočná. Spomínané delenie dotazov na základe zámerov používateľa využívajú aj Jansen a Booth [2], ktorí sledovali najčastejšie vyhľadávané oblasti na internete prostredníctvom vyhľadávacieho nástroja AOL. Sledovanie zámerov používateľov elektronického obchodu môže okrem iného pomôcť zistiť nedostatky a príležitosti v oblasti informačnej architektúry webového sídla, najmä v oblasti použiteľnosti a optimalizácie obsahu pre vyhľadávacie nástroje. To v konečnom dôsledku zanechá v zákazníkoch pocit spokojnosti, ktorý vedie k vyšším transakciám.

2 Ciele a metódy výskumu

Pri analýze informačného správania používateľov sledovaného elektronického obchodu sme využili kvantitatívnu metódu analýzy transakčných logov a kvalitatívnu metódu analýzy vyhľadávacích logov [4], vybraných z daného elektronického obchodu. Konkrétne sme mali k dispozícii 30k akcií (z 8k sedení), z toho takmer 2k vyhľadávacích dotazov na portáli. Rozsiahly dataset, zachytávajúci správanie používateľov v reálnom prostredí, bol získaný vďaka spolupráci akademického a súkromného sektoru. Surové dáta, zachytávajúce správanie, boli nízkoúrovňové a bolo potrebné ich interpretovať. Hoci používateľské akcie, nachádzajúce sa v záznamoch, samy o sebe nehovoria o motívoch, cieľoch a pocitoch používateľov [5], je možné z nich vyčítať isté informácie o ich vyhľadávacích zámeroch a stratégiách.

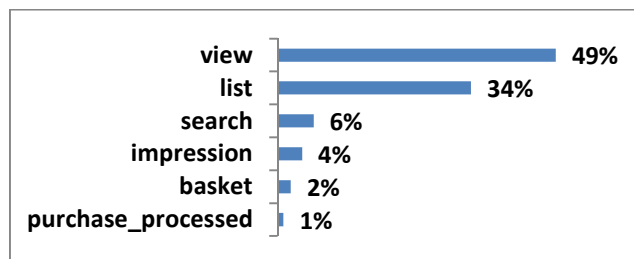
Správanie používateľov bolo zaznamenané v podobe uskutočnených transakcií a jednotlivých udalostí na portáli [6]. Boli vybrané sedenia používateľov spolu s automaticky priradeným typom udalosti, vykonanej na portáli. Taktiež sme mali k dispozícii dotazy používateľov a prezerané produkty. K niektorým produktom boli automaticky priradené kategórie na základe ich zaradenia v elektronickom obchode, avšak v mnohých prípadoch kategórie absentovali, prípadne boli priradené chybné a bolo potrebné ich priradiť a vyhodnocovať manuálne. Vzorku sedení používateľov sme analyzovali zo sedení v čase medzi 9.1.2016 - 18.9.2017, možno ju teda považovať za reprezentatívnu, pretože zachytáva rôzne časové obdobia interakcie používateľov s portálom. Čo sa týka geografického a kultúrneho rozloženia, ide o slovenských, eventuálne českých návštevníkov. Týmto výskumom chceme odpovedať na nasledujúce výskumné otázky: Používa sa pri vyhľadávaní zliav častejšie prieskumová stratégia prezerania alebo krátkeho prieskumu a ktorá z nich najpravdepodobnejšie povedie ku transakcii? Ktoré kategórie dotazov v elektronickom obchode, zameranom na zľa-

vy vedú najčastejšie k transakčnému zámeru používateľa? Aké sú zábery používateľov v jednotlivých kategóriách dotazov a na základe akých aspektov v dotazoch používateľa vyhľadávajú?

3 Závery z analýzy dát

Prieskumovú stratégiu používateľov sme analyzovali podľa vykonaných akcií. Z dostupných údajov sme identifikovali stratégiu prezerania a krátkého prieskumu. **Prezeranie (browsing)** je spôsob exploratívneho a intuitívneho vyhľadávania, založený na zvedavosti a náhodnom získavaní informácií na základe vizuálnych podnetov. V dátach bolo charakterizované typom udalosti view a list a tvorilo až 83% všetkých akcií.

V sledovanom elektronickom obchode teda výrazne prevažoval pasívny typ informačného správania (obr. 1). Väčšina akcií prehliadania prebehla bez ohľadu na kategóriu (view) a menšia časť po označení kategórie (list). Z hľadiska charakteru zľavového elektronického obchodu ide o pochopiteľné a očakávané informačné správanie, pretože predkladaná ponuka sa mení v závislosti od dostupných zliav a používateľa hľadajú pre nich zaujímavé ponuky. Ich zámer je v tomto prípade skôr informačný ako navigačný.



Obr. 1. Percento akcií vybraných používateľov v sledovanom elektronickom obchode

Krátky prieskum (charakterizovaný akciou search) značí priame vyhľadávanie prostredníctvom zadávania kľúčových slov dotazu do vyhľadávacieho okna. Takéto sedenia tvorilo 12% všetkých sedení. Používatelia využívajúci túto stratégiu majú predstavu o vyhľadávanom tovare, resp. službe a aj z údajov vyplýva, že ich zámer bol zrejmý už v prvých krokoch sedenia. Z celkového počtu takmer 1k sedení s použitím krátkého prieskumu bolo v 43,09% sedení použité vyhľadávanie (search) už v prvom alebo druhom kroku. Priamym zadaním dotazu v neskorších krokoch sa používatelia väčšinou vracali k zaujímavým ponukám nájdeným pasívnym prezeraním. V mnohých prípadoch na seba vyhľadávania počas sedení nadväzovali a používateľ sa rozhodol pre transakciu až po niekoľkých sedeniach.

Napriek tomu, že krátky prieskum sa vykonáva v menšej miere, venovať pozornosť používateľom s aktívnym informačným správaním a reflektovať ich informačné po-

treby sa vypláca, keďže pri tomto druhu informačného správania bola väčšia pravdepodobnosť uskutočnenia transakcie. V rámci nášho výskumu sme si za transakciu zvolili vloženie produktu do košíka (basket), pretože tieto transakcie reflektujú záujem používateľa a neboli ovplyvnené prípadnými problémami v posledných krokoch transakcie.

Z údajov v tabuľke 1 je vidieť, že intencionálne vyhľadávanie prostredníctvom krátkého prieskumu malo vplyv na vykonanie transakcie v sledovanom elektronickom obchode. Kým k transakcii viedlo 5 % sedení, zahŕňajúcich výhradne prezeranie (browsing), v prípade použitia krátkého prieskumu došlo k transakcii až pri 10 % sedení.

Tab. 1. Počet sedení v rámci prieskumových stratégií používateľov vzhľadom na transakcie

Prieskumová stratégia	% Všetkých sedení
Krátky prieskum s transakciou (košík)	1,2%
Krátky prieskum bez transakcie	10,7%
Prehliadanie (Browsing) s transakciou (košík)	4,1%
Prehliadanie (Browsing) bez transakcie	79%

Pri použití krátkého prieskumu je možné hlbšie analyzovať **informačné a navigačné zámery používateľov** a aspekty dotazov, použitých pri vyhľadávaní v elektronickom obchode. Sledovali sme kategórie, v rámci ktorých používatelia vyhľadávali. Navigačné zámery boli charakterizované názvom firmy alebo značky v dotaze, informačné zámery jeho prípadnými inými vlastnosťami, ako bolo napríklad čas, miesto, prípadne iný prívlastok v texte dotazu, ako bol najmä typ služby či produktu. Na rozdiel od internetového vyhľadávania, v internom vyhľadávaní skúmaného webového portálu výrazne neprevládajú informačné zámery vyhľadávania nad navigačnými, skôr sú rovnomerne rozložené. Ako bolo spomenuté vyššie, ak sa používateľ rozhodol využiť krátky prieskum zadáním dotazu, tak sa predpokladá, že má predstavu o hľadanej ponuke.

Kategórie s častými vyhľadávacími zámermi informačného charakteru v internom a internetovom vyhľadávaní sa však podobajú, ide o

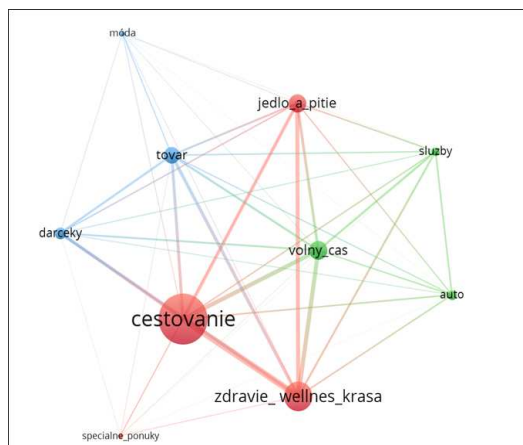
- zdravie, krásu, wellness a zábavu – dotazy so zadáním typu produktu, služby
- cestovanie – zadanie miesta (50% vyhľadávanií v kategórii cestovanie)
- tovar – zadanie typu produktu

Naopak navigačný zámer jednoznačne prevažuje v kategórii jedlo, pitie a reštaurácie, v ktorej až dve tretiny dotazov obsahovali názov reštaurácie. Vyhľadávanie reštaurácií podľa názvu indikuje dôležitosť budovania značky v tejto oblasti.

Najčastejšími dotazmi s transakčným zámerom (basket) sa stali dotazy z kategórií zdravie, krásu, wellness (1,4% dotazov) a tiež cestovanie (1,32% dotazov). Tieto kategórie vykazovali vo výskume Jansena a Booth [2] skôr stredné hodnoty, teda možno potvrdiť, že transakčný vyhľadávací zámer závisí aj od typu digitálneho prostredia, teda webovej alebo inej aplikácie, v ktorom sa vykonáva.

Z údajov vyplýva, že až 87% sedení používateľov je zameraných len na jednu kategóriu tovaru. Ak bolo v rámci jedného sedenia vyhľadávaných viac kategórií, naj-

častejšie išlo o kombináciu kategórií zdravie, krása, wellness a cestovanie ako vidieť podľa prepojenia uzlov na obrázku 3.



Obr. 3. Spoločný výskyt prezeraných kategórií počas sedení používateľov. Zdroj: vlastný obr.

4 Návrhy odporúčaní a diskusia

Vyššie uvedené zistenia možno využiť pre návrhy odporúčaní informácií pre ďalší nákup, prípadne pri predikovaní informačného správania používateľov. Konkrétne pri prezeraní jednej z uvedenej často vyhľadávanej kombinácie kategórií (napríklad cestovanie) je možné ponúknuť používateľovi aj podobné produkty z druhej kategórie (zdravie, krása, wellness). Problémom sú zistené nedostatky (absencie) automaticky priradených kategórií produktov a služieb na sledovanom portáli. Jedným z nedostatkov manuálneho kódovania, ktoré sme použili, je radenie dotazov výlučne do jednej kategórie, avšak vo výskume s využitím podobných metód [7] sa potvrdilo, že až 75% dotazov možno klasifikovať do jednej kategórie.

Alternatívou pre podporenie prieskumovej stratégie monitoringu je vytvorenie elektronického spravodaja, zameraného najmä na oblasť, ktorá bola najčastejšie vyhľadávaná jednotlivými používateľmi aj samostatne (zdravie, krása, wellness).

Pri analýze transakcií (basket vs. purchase) sme si všimli veľký nepomer medzi počtom produktov v košíku a reálne vykonaným nákupom. Pravdepodobnosť nákupu po uložení do košíka bola až 1:5,8. Príčina by mohla spočívať v problémoch v procese transakcie, ďalším dôvodom môže byť, že používatelia v prípade sledovaného e-obchodu využívajú košík ako úložisko produktov, ktoré ich zaujali. Aj tento fakt by mohol e-obchod využiť pre odporúčanie produktov pre individuálnych používateľov na vlastnom portáli alebo v rámci remarketingu.

Zistenia ohľadom typov dotazov pri prieskumových zámeroch majú dôležitú implikáciu pre úpravu rozhrania, konkrétne informačnú architektúru sledovaného webového sídla, najmä pre návrh jeho fazetových filtrov. To však platí iba v prípade väčšieho webového portálu s množstvom výsledkov vyhľadávania, ktoré by používatelia

potrebovali obmedziť. Kategóriu cestovanie by v tom prípade bolo možné doplniť o prehľadnejšie filtre na základe miesta, kategóriu tovar ale aj jedlo a pitie o fazetové filtre najčastejšie vyhľadávaných produktov (napr. burger) a v kategóriách krása, zábava, wellness a voľný čas by sa mohli uplatniť filtre podľa najčastejšie vyhľadávaných služieb (napr. masáž). Fazetové filtre nielen zlepšia použiteľnosť webového portálu, ale tiež podporia intencionálny krátky prieskum, ktorý má, ako sme zistili, vyšší transakčný potenciál.

Súvislosti zámeru používateľa a vykonania transakcie nemožno sledovať iba za pomoci analýzy transakčných logov, ktorú v príspevku využívame. Na vykonanie transakcie (konečný nákup) má vplyv najmä utilitárnska (pragmatická) motivácia používateľa, ovplyvnená pohodlnosťou pohybu na stránke, cenovými hodnotami, bezpečnosťou, dostupnosťou informácií a výberu. Rozhodnutie nakúpiť je ovplyvnené aj kontextuálnymi aspektami, s ktorými je potrebné rátať a ktoré sú doménou marketingu, ako je zákaznícky servis a vnímanie obchodu používateľom [8]. Sledovanie zámeru používateľov je však dôležité aj z toho dôvodu, že persuázia a príjemný zážitok pri nakupovaní nie je v elektronických obchodoch jediným predpokladom vykonania transakcie.

Získané dáta rozhodne nemožno zovšeobecniť, pretože reprezentujú interakciu používateľov iba s jedným elektronickým obchodom. Článok však môže slúžiť ako podnet pre ďalší výskum v oblasti zámerov používateľov pri vyhľadávaní v elektronických obchodoch, keďže problematika v súčasnosti nie je rozpracovaná aj napriek svojej aktuálnosti a dôležitosti najmä pre oblasť nájdateľnosti a optimalizácie elektronických obchodov. Sledovanie vyhľadávania na úzko špecializovanom webovom portáli však prináša aj výhodu eliminácie kontextu, najmä kontextu úlohy, situácie a domény, spomínaných Freundovou a Tomsovou [9]

Ďalšou výskumnou príležitosťou, ktorú plánujeme vykonať v budúcnosti, je prepojenie kvantitatívnych dát o interakcii používateľov s informáciami o osobnosti používateľov na základe modelu Big 5, získanými z dotazníkového prieskumu [5, 6, 10]. Zámery a stratégie vyhľadávania používateľov tak možno sledovať v kontexte typu osobnosti používateľa, čo umožní ešte cielenejšiu personalizáciu a odporúčanie informácií používateľom.

Literatúra

1. Steinerová, J., Grešková, M., Ilavská, J. *Informačné stratégie v elektronickom prostredí*. Univerzita Komenského, Bratislava, 2010.
2. Jansen, Bernard J., Booth, Danielle. Classifying web queries by topic and user intent. In: *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010. p. 4285-4290.
3. Broder, Andrei. A taxonomy of web search. In: *ACM Sigir forum*. ACM, 2002. p. 3-10. Dostupné na: <http://www.cis.upenn.edu/~nenkova/Courses/cis430/p3-broder.pdf>
4. Jansen, Bernard J. Search log analysis: What it is, what's been done, how to do it. *Library & information science research*, 2006, 28,3: 407-432.
5. Bielíková, M. a kol. Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore. In: *Proceedings in Informatics and Information Technologies: 11th Workshop on*

- Intelligent and Knowledge Oriented Technologies, 35th Conference on Data and Knowledge. Nakladatel'stvo STU, 2016
6. Bieliková, M. a kol. Rozsiahla dátová vzorka prepájajúca správanie používateľov e-obchodu s explicitne zistenými osobnostnými charakteristikami. In: *Proceedings in Informatics and Information Technologies: 12th Workshop on Intelligent and Knowledge Oriented Technologies, 36th Conference on Data and Knowledge*. Nakladatel'stvo STU, 2017
 7. Jansen, B.J., Booth, D., Spink, A. Determining the User Intent of Web Search Engine Queries. In: *Proceedings of the 16th international conference on World Wide Web*. NY: ACM, 2007. s. 1149-1150. Dostupné na: <https://dl.acm.org/citation.cfm?id=1242739>
 8. Shim, S. An online repurchase intentions model: The role of intention to search. In: *Journal of Retailing*, Volume 77, Issue 3, 2001, str. 397-416. Dostupné na: <https://www.sciencedirect.com/science/article/pii/S0022435901000513>
 9. Freund, L., Toms, E. Contextual Search: from Information Behaviour to Information Retrieval. In: *Proceedings of the Annual Conference of CAIS*. Canadian Association for Information Science, 2005.
 10. Al-Samarraie, H., Eldenfria, A., Dawoud, H. The impact of personality traits on users' information-seeking behavior. In: *Journal of Information Processing and Management: an International Journal*. Volume 53 Issue 1, 2017. Tarrytown: Pergamon Press, 2017. str. 237-247

Pod'akovanie: Tento príspevok bol spracovaný vďaka projektu HIBER, APVV 15-0508.

Annotation:

Search strategies and intentions in sales portals: Case study

Tracking intent and interactions with e-shops is especially important for the purpose of exploring and meeting the user needs. In this case study, transaction and search log analysis method is used to analyze the data. It turns out that browsing is the most common search strategy, nevertheless active searching using search query results in more transactions. User's intent (information, navigation, and transaction) depends to a large extent on the subject. Therefore, we also categorize users' intentions in the search for discounts that has many implications for information architecture and personalized recommendation of information in e-shops.

Influence of the individual's size on the island model architecture

Stepan Balcar

Charles University, Faculty of Mathematics and Physics
`Stepan.Balcar@mff.cuni.cz`

Abstract. Our motivation starts from recommendation systems on large data. We use hybrid evolutionary agent based system. Our system runs in parallel several optimization methods. Main challenge is the frequency of individuals migration. We describe technical restrictions which complicate parallelization techniques of helping each other computational methods. Size of data influences latency in communication and measurement. We experiment with architectural changes which enable to move the limits. At the same time we would like to give planners sufficient freedom to manage distributed computation.

Key words: recommendation systems · large data · evolution based island models · computation parallelization · agent based systems

1 Introduction

Our research focuses on recommending systems and utilizing evolving algorithms to predict user preferences. In this paper we would like to report on technological challenges and architectural solutions. We started with a system, originally developed for study of parallelization of standard stochastic and evolutionary optimization methods which was tested on problems: TSP, BP, vertex cover, ... ([3], implementation described in [4]). This system was based on the idea of islands model with migrants ([6]). When we tried to use this system for recommendation tasks, it turned out that main limit is the size of an individual.

Main idea of original [4]-system was to have a tool for experiments comparing single methods (SM), homogeneous island models (HoIM) and heterogeneous island models (HeIM). Homogeneous island models are known for a longer time. For a while, main research topic of HoIM was parameter selection, population size and ratio of mutation (see e.g. [7]). Heterogeneous approach was understood as diversity on the level of parameters (see [5]). M. Pilat and R. Neruda started a new approach to HeIM, where the heterogeneity is understood as diversity in methods [6].

Our original motivation in [3] was to create an experimental system sufficiently flexible to be able to show contribution of heterogeneity of diverse methods with migration and re-planning. For implementation we have selected the multi-agent framework Jade¹. Main advantage of agent framework is potential for

¹ <http://jade.tilab.com/>

obtaining an adaptive system which is able of inherent improvement. In this sense idea of our system was similar (but much more general) to the one of ([5]). In [3] it was shown that heterogeneity gives better results than homogeneity.

Recently we tried to use the potential of HeIM in the area of recommender systems [1]. In this domain we optimized model based solution in the form of latent factors predicting rating matrix. It turned out that individuals (models / recommendation) are several orders of magnitude larger than in previous experiments. This led to serious performance problems. In this paper we describe both the original system and architectural changes to be able to work with large individuals (and still getting acceptable results).

2 The architecture of the implemented heterogeneous islands model system

On Figure 1 we see basic idea of heterogeneous island models. Circles and colored points correspond to different computational resources and optimization methods. Methods help each other by sending migrants. To achieve a synergy, the orchestration of the whole is done by a planner.

From this concept arose the system architecture shown in Figure 2. Jade container correspond to island and computational agents (CA) correspond to optimization methods. Agents send each other ACL messages (FIPA). Agent "Central manager" manages distributed computation. Agent "Monitor" intercepts communication in the system and creates history of computation based on statistics. In each container there is a local master. One of the parameters of this model is the period of migration of individuals ("period of distribution" in Table 1). One has also to define neighboring relation, i.e. where to send individuals selected by sender.

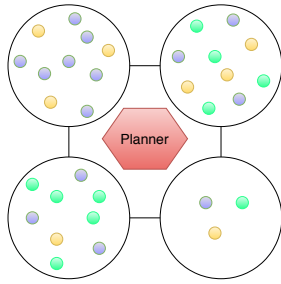


Fig. 1. System abstraction

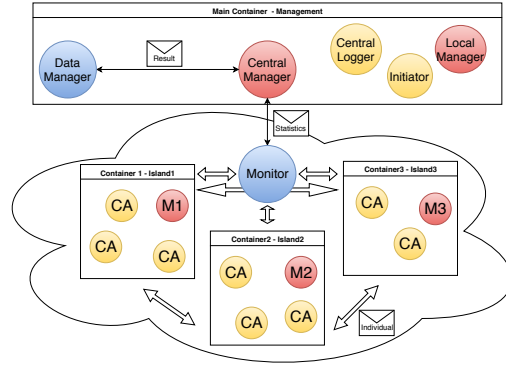


Fig. 2. System architecture

3 Planning and monitoring of the computation

The heart of heterogeneous system is the planner. Planner initializes and re-plans methods (commands local masters). The heart of the planner is the planner ideology. In Figure 3 we see the module "Planner ideology" (PI) which is exchangeable. In [2] (see also [3]) we can see several implementations of PI. Planner module gets on input informations on success and failures of running methods instances. Helpers consists of statistics about count of successful help of one method to another (see Figure 5). Pedigree are that of individuals (e.g. an individual in 50 iteration was influences (improved) by different methods). Statistics is a more detailed history on re-planning in different iterations.

On Figure 4 we see process of transformation of measurements done by agent Monitor to the history of computation. It is up to our choice how many information we store is statistics and history. Data stored by agent Data manager are helpful for analysis of system behavior. This is needed for human activity when creating a new planner ideology and tuning the system.

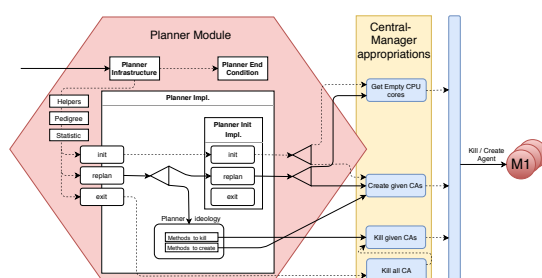


Fig. 3. Planner

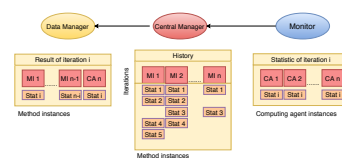


Fig. 4. Monitoring and creating history

Information sent by "Monitor" to "Central Manager" originally contained individuals created by method instances. This enabled to create different metrics, care about diversity and value of genetic material. In case of large individuals in recommendation domain this influenced middle-ware throughput. In the new version of the system results of statistics contain only hash of individuals, fitness, pedigree and some others. This of course limits possibilities of Planner ideology. It is left for future work to increase quality of recommendation under this restricted circumstances.

4 New architecture of computational methods

Further significant changes were done in architecture of optimization methods.

We experimented with 50 cycles of the planner. Inside each cycle of the planner (a time frame 1 minute) there were several cycles of optimization methods.

Nevertheless, in an heterogeneous system methods differ in number of generations (iterations) done in this 1 minute. Hence the system has to be designed asynchronous. Our first implementation of asynchronism was very liberal to all interruptions. In our new architecture (still asynchronous) we implemented selectors and inserters controlled by clocks. The clock was set to optimize simulated annealing performance. The main reason was, that simulated annealing performed very bad in [1] (with system version corresponding to second row in Table 1).

In recommender system domain our individual are latent factors (namely latent factors of users and latent factors of items) of big rating matrices. To compute fitness of an individual we have to calculate large part of matrix product and compare it to training examples. Each operation with an individual needs more time. That is why the choice of individuals from the queue of migrants is influencing convergence of the computation. Original architecture of the computational agent consisted of two queues of all incoming and outgoing individuals. In the new architecture (Figure 6) we understand data structures as a parameter of the system (much more general than queues).

From our experiments it follows importance of selectors and inserter choosing the individual to be included into the population or inserted to model of individuals waiting for departure. Sparing sending behavior is crucial for middle-ware throughput. In new system we pay more attention to sending behavior. Specifically, we implemented more options which enable effective distribution of genetic material and save communication load and memory.

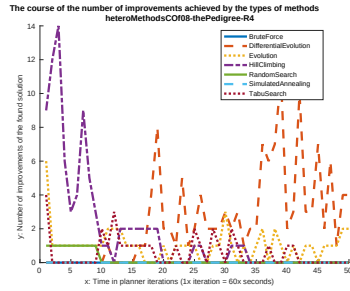


Fig. 5. Mutual help of methods

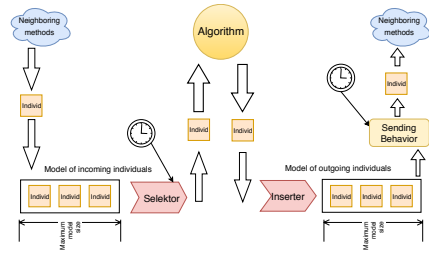


Fig. 6. New CA-architecture

5 Results

We run experiments on ml-100k² ml-1m³ and ml-10m⁴ datasets. Our new implementation is publicly available on Github⁵ (all the changes are backward compatible).

In order to get system running on large data and giving acceptable results we changed architecture and way of monitoring (as described above), parameters of the system: "periodOfDistribution" (i.e. frequency of migration), "distribution" (i.e. choice of neighboring relations), "indiv-transport" (from ontology to objects) and "models" (data models of incoming and outgoing individuals).

Table 1 describes phases of our system development. Original system was inapplicable. When we increased the period of distribution and bounded the number of neighbors to 3, system was able to process ml-100k dataset effective and in acceptable quality.

With new architecture we provided several experiments. This process was classical tuning of trade-off between quality of results, effectiveness and system through-put (schematically illustrated in Figure 7). We preserved distribution to 3 neighbors and frequency. All the changes of architecture and sending individuals as object (not ontology) gave sufficient a acceptable quality of results. Our next challenge was to decrease period (increase frequency) of distribution to 5s. One would expect the system be congested (we have kept 3 as the number of neighbors). Last change we display in Table 1 is change of data structures of incoming and outgoing individuals sufficiently outperformed danger of congestion. The system was still communicationally uncongested and the usage of memory decreased.

Architecture of system	ml-100k	ml-1M	ml-10M
Original system: periodOfDistribution=5s, distribution=all, indiv-transport=ontology	-	-	-
Original system periodOfDistribution=30s, distribution=3neighbors, indiv-transport=ontology	running	ineffective	-
Improved system: periodOfDistribution=30s, distribution=3neighbors, indiv-transport=object, models=(100k=queues, 1M=queues, 10M=queues)	running	running	running, on limits of memory
Improved system periodOfDistribution=5s, distribution=3neighbors, indiv-transport=object, models=(100k=queues, 1M=3last, 10M=1last)	running	running	running, less memory demanding

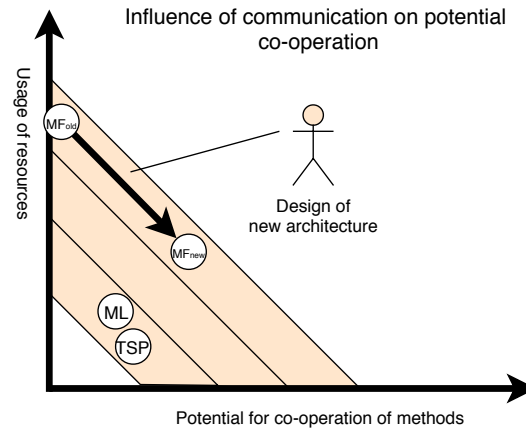
Tab. 1. Consequences of modifying system architecture

² <https://grouplens.org/datasets/movielens/100k/>

³ <https://grouplens.org/datasets/movielens/1m/>

⁴ <https://grouplens.org/datasets/movielens/10m/>

⁵ <https://github.com/sbalcar/distributedea/>

**Fig. 7.** Trade-off

6 Conclusions

We improved original system for classical evolutionary optimization to work with large data in recommender systems (more details in [1]). By several architectural changes we shifted the limits of communication congestion. In future work we would like to be competitive in quality of recommendation. Possible next architectural change are hierarchical heterogeneous island models.

References

1. Balcar, S.: Preference learning by matrix factorization on island models. In: ITAT'18, accepted, to appear in CEUR (2018)
2. Balcar, S., Pilat, M.: Online parallel portfolio selection with heterogeneous island model (accepted ictai'18). <http://arxiv.org/abs/1806.04528>
3. Balcar, S., Pilat, M.: Heterogeneous island model with re-planning of methods. In: Aguirre, H. (ed.) GECCO '18. ACM, New York, NY, USA (2018)
4. Balcar, S.: Heterogeneous island model with re-planning of methods (in czech). Master thesis, Martin Pilát supervisor (2017), <https://is.cuni.cz/webapps/zzp/detail/173747/>
5. Gong, Y., Fukunaga, A.: Distributed island-model genetic algorithms using heterogeneous parameter settings. In: 2011 IEEE CEC. pp. 820–827 (2011)
6. Pilát, M., Neruda, R.: Combining multiobjective and single-objective genetic algorithms in heterogeneous island model. In: CEC 2010. pp. 1–8 (2010). <https://doi.org/10.1109/CEC.2010.5586075>
7. Whitley, D., Rana, S., Heckendorn, R.B.: The island model genetic algorithm: On separability, population size and convergence. *J.Comp.Inf.Tech.* **7**, 33–47 (1998)

Acknowledgement. This work was supported by a Czech Ph.D grant SVV2018-260451.

Knowledge-Driven Manufacturing

Petr Kadera¹, Petr Novák¹, Jiří Vyskočil¹, and Václav Jirkovský¹

Czech Institute of Informatics, Robotics and Cybernetics Czech Technical University
in Prague

Jugoslávských Partyzánů 3, 160 00 Praha

{`petr.kadera`, `petr.novak`, `jiri.vyskocil`, `vaclav.jirkovsky`} @cvut.cz

Abstract We are currently witnessing an ongoing transformation of manufacturing industry from mass production to mass customization. Robotics plays in this process a key role. The former and mostly present robotic solutions were deployed to repeat a fixed sequence of predefined operations. This approach is not sufficient for customized production due to high costs for the reconfiguration of robotic systems. Thus, a knowledge-driven approach is proposed to overcome this issue by automatic creation of robotic control sequences. Both robotic system and final product are declaratively described. Consequently, the description is transformed to the Planning Domain Description Language (PDDL) format, which is then input for an arbitrary PDDL compliant planner (e.g., Fast-Downward). The planner finds (if such a plan exists) a sequence of robotic operations that lead to the final product. The approach is validated on a system composed of two KUKA robots and an automated conveying system that produces LEGO models.

Keywords: Flexible manufacturing · Robotics · Automatic Planning · PDDL

1 Introduction

Shorter time-to-market, high responsibility to customer needs and strong support for customization are expected properties of factories brought by the Fourth industrial revolution [3]. A key role in this process will play robotics as an enabler of the flexible production. However, current methods of programming robotic systems aims rather at repetitive than flexible production [1]. The behavior of robots is hard-coded and thus each modification of the robotic behavior requires change of the control code, which is not only expensive, but also error-prone. In this paper, we propose a method that build upon a formal specification of a particular robotic domain and a concrete product. These specifications are transformed in the form of Planning Domain Definition Language (PDDL), which is widespread format for describing planning problems [4].

2 Proposed method

The proposed method is a five-step process as follows: (i) Specification of a product, (ii) Specification of a production domain, (iii) Transformation of the

product specification to the PDDL, (iv) Computation of the production plan, (v) Interpretation of the plan by a Programmable Logical Controller (PLC) and execution by robots (see Fig. 1).

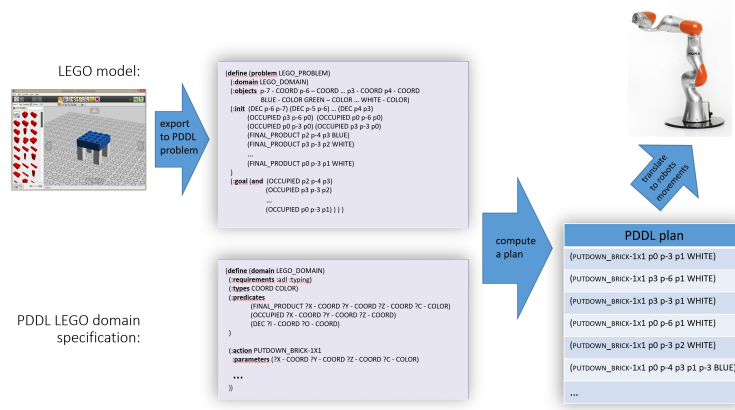


Figure 1. LEGO Building Flowchart.

The method does not prescribe a concrete format to describe the requested product. In this pilot application we use the freely available tool LEGO Digital Designer¹. The production domain is described directly in the PDDL form (see Fig. 2), whereas specification of the model is transformed to PDDL using a newly developed proprietary convertor (see Fig. 3). Since both production domain as well as the required product are formally specified, a PDDL compliant planner can be used. In this case, the Fast-Downward planner is used because it is recognized as a fast and reliable tool. The computed plan is transformed in a sequence of production steps that are communicated to a PLC.

3 Validation Study

The functionality of the proposed method is validated using laboratory facility Testbed for Industry 4.0² (see Fig. 4). The hardware setup of the facility includes a conveying transportation system Montratec, two KUKA Agilus robots and one KUKA LBR iiwa robot, PLC Siemens Simatic S7 -1512 and a PC workstation DELL Precision T1700.

¹ <https://www.lego.com/en-us/ldd>

² <https://www.ciirc.cvut.cz/testbed/>

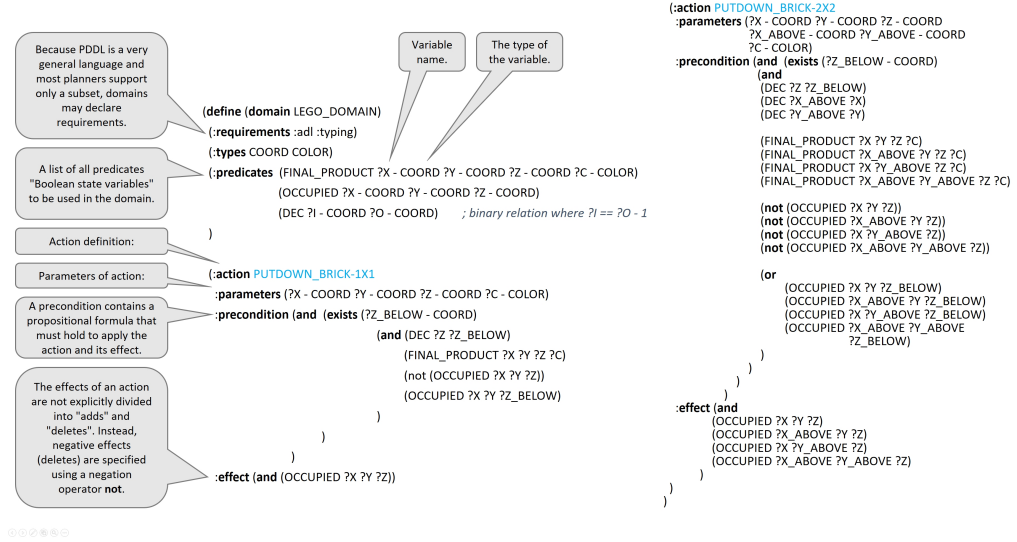


Figure 2. PDDL LEGO domain specification

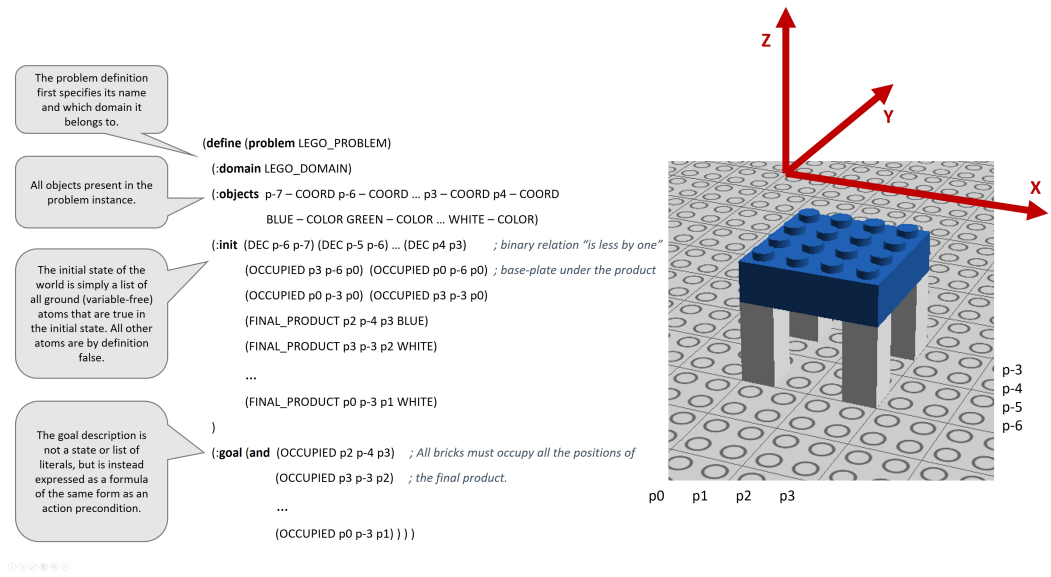


Figure 3. PDDL LEGO problem instance

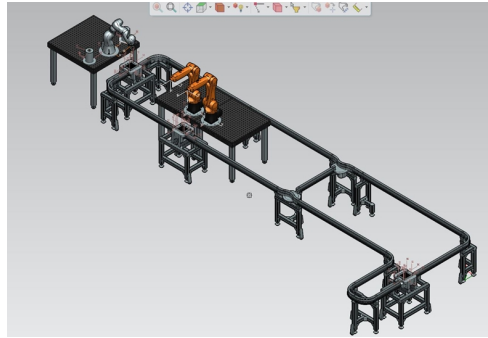


Figure 4. Testbed at CTU

4 Conclusions and Future Work

The presented method illustrates an example of a knowledge-driven robotic production which represents an alternative to traditional static robotic programs. Explicitly defined knowledge about the production domain as well as about the configuration of the required product is transformed into a planning problem that is solved using a universal planner. Consequently, the found plan is executed by available production resources. In the future, we plan to extend the scope of this method outside the LEGO domain. Therefore, Lego Digital Designer will not be suitable any more to represent the model of a product and it will be necessary to find an alternative. Web Ontology Language (OWL) [2] is a good candidate to describe the product structure as well as to handle the expert knowledge about the relevant production domain.

Acknowledgment

This research has been supported by the EU Horizon 2020 project DIGICOR and by the DAMiAS project funded by the Technological Agency of the Czech Republic.

References

1. Angerer, A., Vistein, M., Hoffmann, A., Reif, W., Krebs, F., Schönheits, M.: Towards multi-functional robot-based automation systems. In: Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics - Volume 2: ICINCO,. pp. 438–443. INSTICC, SciTePress (2015)
2. Bechhofer, S.: Owl: Web ontology language. In: Encyclopedia of database systems, pp. 2008–2009. Springer (2009)
3. Drath, R., Horch, A.: Industrie 4.0: Hit or hype?[industry forum]. IEEE industrial electronics magazine 8(2), 56–58 (2014)
4. McDermott, D., Ghallab, M., Howe, A., Knoblock, C., Ram, A., Veloso, M., Weld, D., Wilkins, D.: Pddl-the planning domain definition language (1998)

Představení projektu "Moderní zpřístupnění historických pramenů"

Pavel Král^{1,2}, Karel Halla³, Radek Široký⁴, Ladislav Lenc² a Jiří Martínek¹

¹ Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni
Univerzitní 8, 306 14 Plzeň

² Nové technologie pro informační společnost - NTIS, FAV ZČU v Plzni
Technická 14, 306 14 Plzeň
{pkral,llenc,jimar}@kiv.zcu.cz

³ Státní oblastní archiv v Plzni, Státní okresní archiv Cheb
Františkánské nám. 14, 350 02 Cheb
halla@soaplzen.cz

⁴ Západočeský institut pro ochranu a dokumentaci památek - ZIP o.p.s.
Tomanova 3-5, 301 00 Plzeň
rsiroky@zip-ops.cz

Abstrakt. Hlavním cílem projektu je zpřístupnění archivních pramenů z česko-bavorského příhraničí pomocí nejmodernějších informačních technologií. Bude umožněno vyhledávání informací podle lokalizačních údajů. Dále se zaměříme na přehlednou prezentaci a efektivní vyhledávání obrazových dokumentů. Také bude realizováno inteligentní full-textové zpřístupnění dokumentů v českém i německém jazyce. Informace budou k dispozici pomocí nadstavby existujícího portálu „Porta Fontium“.

Klíčová slova: česko-bavorská spolupráce · full-textové vyhledávání · Porta Fontium

1 Úvod

V současné době jsou digitalizované archivní prameny v česko-bavorském příhraničí k dispozici prostřednictvím portálu „Porta Fontium“. Tento systém vznikl především pro fyzickou ochranu cenných dokumentů, proto je vyhledávání informací stále závislé na ručním procházení a pročitání mnohostránkových rukopisů. Zároveň v aktuální verzi systému nejsou obsaženy historické mapy.

Hlavním cílem projektu je zpřístupnění archivních pramenů z česko-bavorského příhraničí širokému spektru uživatelů z řad odborné i laické veřejnosti pomocí nejmodernějších informačních technologií.

Prvním dílčím cílem je vůbec první spojení historických map a plánů ke společným česko-bavorským dějinám až do roku 1918 a jejich online prezentace. Dalším cílem bude umožnění vyhledávání informací podle lokalizačních údajů. Portál bude propojen s mapovými podklady. Bude tak možno vyhledávat různé typy dokumentů podle jejich místa na mapě. Zároveň bude umožněno zobrazení vybraných dokumentů na mapě. Dále se zaměříme na přehlednou prezentaci a efektivní vyhledávání obrazových dokumentů. Uživatel uvidí dokument ve

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 173-176.*

vrstvách (text, obrázky s popisem), což bude přehlednější. Dále bude nabídnuto vyhledávat v rukopisných kronikách podle extrahovaných vzorů. Čtvrtý dílčí cíl se zabývá inteligentním full-textovým zpřístupněním dokumentů. Uživatel bude moci provádět dotazy v přirozeném jazyce a odpověď systému bude více relevantní. Posledním cílem je prezentace výsledků projektu odborníkům i širší veřejnosti prostřednictvím workshopů a školení. Portál tak bude sloužit odborné i laické veřejnosti v mnohem větší míře, než dosud. Zároveň bude od uživatelů získána zpětná vazba, která přispěje k řešení výše uvedených dílčích cílů.

2 Řešení projektu

V prvním kroku provedou všechny zúčastněné archivy na bavorské a české straně ve svých elektronických a analogových pomůckách vyhledání a identifikaci pro příhraničí relevantních map. Přitom vedle podrobného popisu, který bude rozhodující pro konečné posouzení relevantnosti pro projekt, dojde také ke zhodnocení fyzického stavu mapových podkladů. Identifikované mapy a plány budou se získanými údaji obsahově zpracovány a indexovány (opatřeny metadaty), což umožní následné filtrování podle území, místa, časového rozmezí, osob atd. (dosaďná popis, pořízený především z analogových pomůcek a kartoték, proběhl pouze ve velmi hrubých obrysech a v žádném případě nesplňuje podmínky pro online využití v rámci internetové platformy).

V dalším kroku budou předem vytipované mapy a plány konzervovány, případně restaurovány, a připraveny k digitalizaci. Přitom by mělo dojít také k jejich odbornému adjustování (přebalení). Poté budou relevantní mapy a plány digitalizovány, georeferencovány a spojeny s metadaty pořízenými v rámci popisu a zveřejněny na internetu prostřednictvím portálu „Porta Fontium“. Paralelně budou probíhat překlady pořízených popisů do českého a německého jazyka.

V rámci portálu „Porta Fontium“ bude vytvořen mapový prohlížeč pro geografickou prezentaci veškerých dat včetně nově zpřístupněných historických map. Prohlížeč bude obsahovat různé mapové vrstvy. Mezi dokumenty a jejich lokalizací bude obousměrná vazba. Uživatel bude mít možnost vyhledávat různé typy dokumentů podle jejich místa na mapě s možností filtrace zobrazovaných typů dokumentu. Na druhé straně bude uživateli umožněno zobrazit vybrané dokumenty na mapě.

Dále bude vytvořen modul pro přehlednou prezentaci a prohledávání obrazových dokumentů, které tvoří velkou část zveřejněných historických pramenů. Budou analyzována obrazová data, bude provedena jejich segmentace (text, pozadí, grafická část, ...) a související anotace. V rámci portálu „Porta Fontium“ bude vytvořen prohlížeč obrazových dokumentů, který kromě jejich kvalitní prezentace umožní i vyhledávání v obrazových materiálech (například podle určitého typu architektury). Tento nový prohlížeč umožní i vyhledávání v rukopisných kronikách podle ručně extrahovaných částí psaného textu. Analogicky aplikace uživateli umožní najít fotografie (obrázky), které jsou podobné dané předloze (výřezu) obrázku.

Pro dosažení co nejlepších výsledků budou kombinovány tradiční zavedené postupy z oblastí automatického zpracování vzorů a strojového vidění spolu s novými metodami hlubokého učení (deep learning).

Jako další bude provedena segmentace tištěných dokumentů přístupných v databázích portálu "Porta Fontium". Každý dokument bude rozdělen na text, tabulky a obrázky. Text bude následně rozčleněn na sekce (nadpis, autoři, tělo, apod.). Dále bude textová vrstva (v podobě obrázku) převedena pomocí metod optického rozpoznávání znaků (OCR) do textové podoby. Takto upravené dokumenty budou zpracovány / analyzovány metodami automatického zpracování přirozeného jazyka (NLP). Bude provedena normalizace slov (převod slov do základního tvaru), sémantická analýza (určení významu textu), automatické určení pojmenovaných entit (důležitá slova v textu - např. osoby, organizace, adresy, časové údaje, apod.) , určení kategorií (politika, kultura, historie, apod.) a klíčových slov v dokumentu. V případě potřeby bude dále proveden automatický souhrn textů. Taktéž bude provedeno dvojjazyčné zpřístupnění dokumentů. Budou analyzovány dva možné přístupy řešení tohoto problému. První přístup bude pomocí strojového překladu (a případné následné ruční korekce). Druhý přístup spočívá v obecné (jazykově nezávislé) reprezentaci dokumentů. Na základě výsledků experimentů bude rozhodnuto o možnostech integrace do stávajícího systému.

Bude tak vytvořen moderní dvojjazyčný prohlížeč tištěných dokumentů, jehož cílem je zefektivnění vyhledávání informací v množině dokumentů. Uživatel bude moci provádět dotazy v přirozeném jazyce a odpověď systému bude více relevantní, než při hledání v podobě klíčových slov. (např. budou nalezeny všechny dokumenty, kde se vyskytuje slovo „les“ v libovolném tvaru, apod.).

Průběh celého projektu bude koordinován s širší odbornou veřejností zainteresovanou na využívání historických pramenů s cílem seznámení maximálního počtu potenciálních uživatelů portálu „Porta Fontium“ s jeho možnostmi a zároveň pro získání zpětné vazby a požadavků pro jeho efektivní provoz. Na počátku projektu bude probíhat diskuse o potřebách odborné veřejnosti, získané poznatky budou zohledněny při realizaci projektu. Výsledky projektu budou prezentovány na odborných konferencích a formou cílených školení. V rámci projektu bude také vytvořen modelový výchovně vzdělávací program pro prezentaci poznávacího potenciálu historických archivních pramenů určený pro studenty středních, případně pro žáky základních škol.

V současné době připravujeme odborný diskuzní workshop, který se uskuteční ve dnech 8. až 9. listopadu 2018 v Klášteře Teplá⁵.

3 Závěr

V rámci řešení projektu byla navázána ojedinělá spolupráce odborníků z příhraničí ze dvou sousedních států a několika výzkumných oblastí: historie, stavebně historický průzkum, archeologie a informatika, což umožní portál „Porta

⁵ Další informace k pořádanému semináři jsou k dispozici na <http://www.portafontium.eu/sites/default/files/workshop-2018-11-pozvanka.pdf>

Fontium“ rozšířit o unikátní funkčnost: moderní vizualizaci obrazových materiálů, efektivní vyhledávání požadovaných informací v obrazových materiálech (ručně psané texty a fotografie) a inteligentní dvojjazyčné full-textové vyhledávání. Tím se otevře cesta k možnostem rozsáhlého a tím i inovativního přeshraničního bádání pro historiky, regionální badatele a genealogy.

Poděkování: Tento článek vznikl díky podpoře projektu č. 211: Moderní zpřístupnění historických pramenů z programu přeshraniční spolupráce Česká republika – Svobodný stát Bavorsko Cíl EÚS 2014 -2020.

Annotation:

The main goal of this project is to make accessible archival resources from the Czech-Bavarian border region using state-of-the-art information technologies. It will be possible to search information based on geolocation. We also focus on a clear presentation and an effective search of the documents in a form of raster images. We further realize an intelligent full-text access to the printed documents in both Czech and German languages. The information will be available through an existing portal “Porta Fontium”.

Využitie dostupných otvorených dát na podporu výberu nehnuteľnosti v Slovenskej republike

Pavol Šatala¹, Tomáš Babič¹, Peter Butka¹

¹Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Slovenská republika
pavol.satala@tuke.sk, tomas.babic@student.tuke.sk,
peter.butka@tuke.sk

Abstrakt. Vlády a vládne organizácie zverejňujú veľké množstvo verejne dostupných dát, ktoré poskytujú informácie o ich fungovaní, finančných operáciách, o službách poskytovaných danou organizáciou a podobne. Tieto informácie sú povinne zverejňované formou štruktúrovaných súborov na internete. Článok poskytuje prehľad o objeme, kvalite a dostupnosti dát. Vybraná oblasť týchto dát je následne využitá pri tvorbe mobilnej aplikácie pre OS Android, ktorá poskytuje podporu pri výbere vhodnej nehnuteľnosti. Úlohou aplikácie je zobrazit' ponuku nehnuteľností s ohľadom na blízkosť služieb požadovaných kupujúcim. V závere článku je formou prieskumu vyhodnotená relevantnosť vybraných služieb pre výber nehnuteľností a taktiež vyhodnotený prínos funkcionality aplikácie pre zrýchlenie vyhľadania ponuky.

Kľúčové slova: otvorené dáta, mobilná aplikácia, nehnuteľnosti, kritéria výberu nehnuteľností.

Typ príspevku: Aplikčný príspevok

1 Úvod

Používatelia si pri výbere nehnuteľnosti kladú rôzne požiadavky. Po upresnení potrebnej veľkosti nehnuteľnosti v závislosti od jej požadovaného využitia, sa zvyčajne zákazníci zameriavajú na lokalitu nehnuteľnosti. Okrem subjektívnych preferencií niektorých lokalít používatelia sledujú dostupnosť služieb v požadovanom dosahu. Manuálne vyhľadanie takýchto lokalít je zdĺhavé a nepresné. V článku sa zaoberáme možnosťou využitia mobilnej aplikácie s využitím dostupných dát na uľahčenie výberu nehnuteľnosti podľa požadovaných kritérií.

2 Metódy

Na splnenie nášho cieľu sme na základe dostupných dát vytvorili aplikáciu umožňujúcu nájsť lokalitu a ponúkané nehnuteľnosti podľa zvolených kritérií. Hlavnou inšpiráciou boli podobné aplikácie zo Spojeného kráľovstva. Avšak naša aplikácia ponúka výrazne viacej parametrov na selekciu vhodnej lokality ako napríklad informácie o záplavových

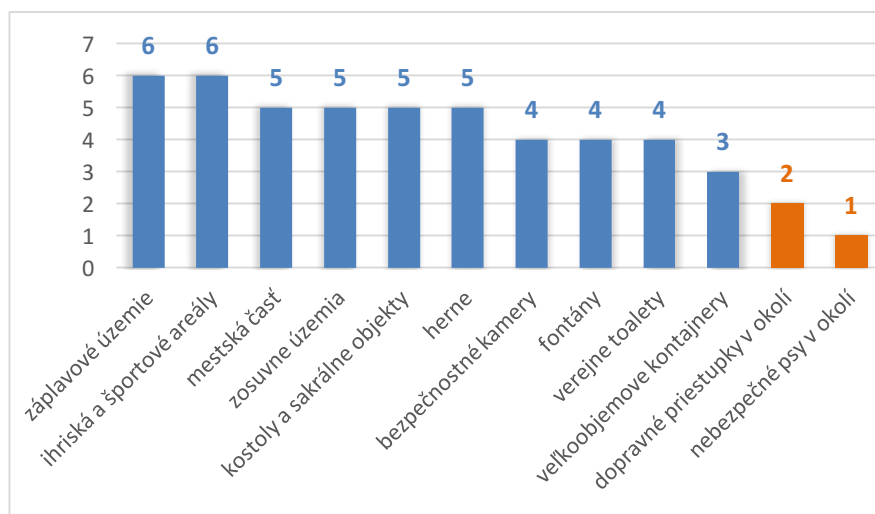
*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 177-181.*

územiach, ihriskách, športových zariadeniach, mestských častiach, zosuvných územiach, sakrálnych budovách, herniach, bezpečnostných kamerách, rozložení veľkoobjemových kontajnerov, prihlásených nebezpečných psoch a podobne.

Prvou úlohou pri vývoji aplikácie bolo určiť najdôležitejšie parametre nehnuteľnosti pre našu cieľovú skupinu.

2.1 Výber relevantných parametrov

Pre výber kritérií vhodných pre našu aplikáciu bol realizovaný prieskum formou dotazníka. Dotazník bol distribuovaný ľuďom v našej cieľovej skupine, ktorí predstavujú ľudí medzi 30. a 40. rokom života, hľadajúci si svoje nové bývanie. Úlohou respondentov bolo vybrať 5 najdôležitejších kritérií pri výbere nového bývania. Na výber mali medzi atribútmi poskytovanými dostupnými datasetmi. Dotazník vyplnilo 10 respondentov. Podľa tohto prieskumu päť najdôležitejších kritérií predstavujú: poloha v území ohrozenom povodňami, ihriská a športové areály, mestské časti, územia ohrozené zosuvmi pôdy, kostoly a sakrálné objekty. Tieto kritéria by mali byť zahrnuté do našej aplikácie. Detailné výsledky prieskumu sú znázornené na obr. 1.

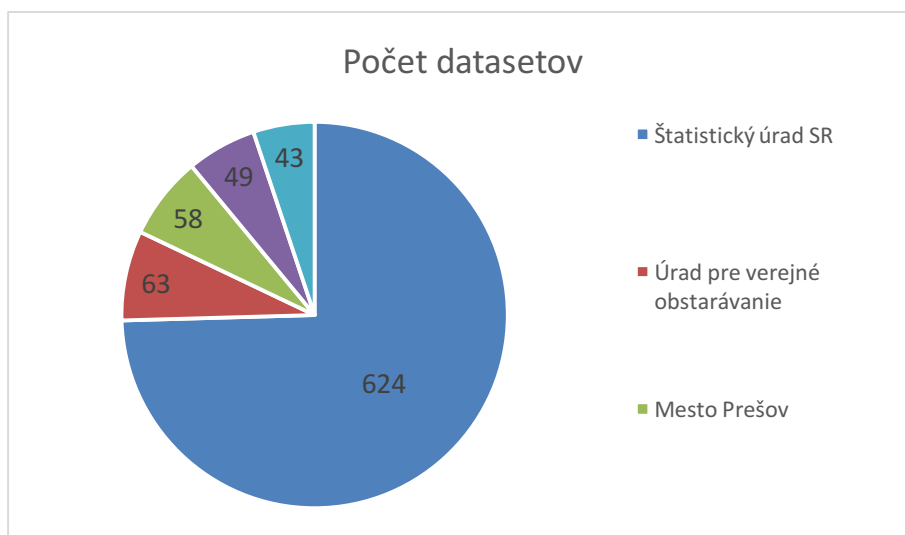


Obr. 1: Výsledky dotazníka

2.2 Dostupné dáta

Pre získanie informácií o pozícií objektoch sme sa zamerali na problematiku otvorených dát. Otvorené dáta sú definované ako verejne dostupné informácie, voľne využiteľné na prácu, ukladanie ako aj distribúciu[1]. Najfrekvencovanejším spôsobom distribúcie je publikovanie na internete. Podľa [2] hlavnými požiadavkami kladenými na otvorené dáta sú: komplexnosť, strojová čitateľnosť a verejná dostupnosť. Dáta by

mali byť poskytované v štruktúrovanom formáte (napr.: *.csv, *.xml, *.json textové súbory). Obrázky a *.pdf súbory nie sú považované za otvorené dáta, z dôvodu komplikovaného strojového spracovania. Komplexné celosvetové hodnotenie a porovnanie prístupu vlád k problematike otvorených dát si kladie za cieľ projekt The Open Data Barometer. Vyhodnocuje ako kvalitu zverejňovaných dát, tak aj ich dopad na sociálny život a inovácie v danej krajine. V roku 2016 sa na prvých piatich miestach umiestnili Spojené Kráľovstvo, Kanada, Francúzsko, Spojené štáty Americké a Južná Kórea. Slovenská republika sa umiestnila na 29. priečke[3]. Prvou iniciatívou Slovenska v tejto oblasti bolo spustenie portálu data.gov.sk v roku 2013. Tento portál poskytuje prístup k otvoreným dátam zo štátnej správy. Na základe zákona č. 264/2010z.z. všetky vládne organizácie, samosprávy a štátne podniky musia zverejňovať svoje zmluvy, faktúry a objednávky. Medzi najvýznamnejšie katalógy týchto otvorených dát v SR patria webové sídla slovensko.sk a data.gov.sk. Na obr. 2 je zobrazená početnosť dát zverejnených jednotlivými organizáciami na portáli data.gov.sk.



Obr. 2: Porovnanie počtu zverejnených datasetov jednotlivých organizácií na portáli data.gov.sk.

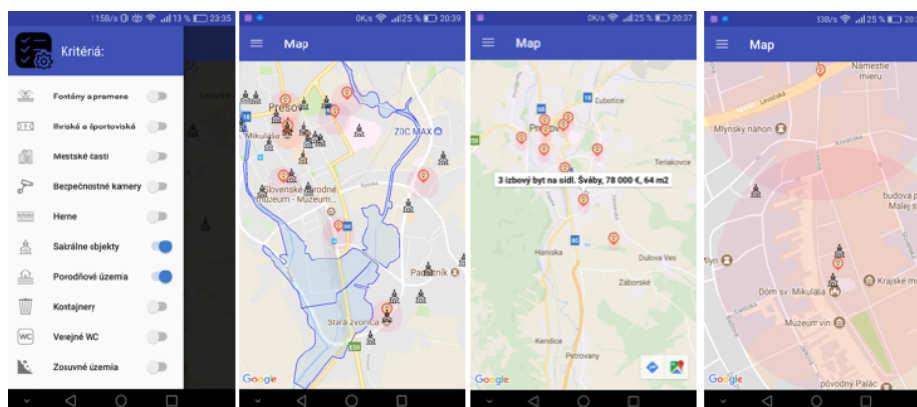
Medzi významných poskytovateľov otvorených dát patrí mesto Prešov. Mesto Prešov je krajské mesto ležiace na východe slovenskej republiky s približne 90 000 obyvateľmi. V našom riešení sme sa rozhodli prakticky realizovať jedno z riešení využívajúcich tieto dostupné dáta. Ako možnosť zvýšenia pridanej hodnoty dát, sme videli využitie aktualizovaných informáciách o objektoch v meste s cieľom poskytnúť podporu pri hľadaní správnej nehnuteľnosti.

3 Výsledné riešenie

Úlohou nášho riešenia bolo potvrdiť hypotézu, že mobilná aplikácia využívajúca otvorené dáta je schopná skrátiť čas potrebný pre nájdenie nehnuteľnosti spĺňajúcej kritéria zákazníka. Štatistika poštovej banky z 25. Januára 2018 ukazuje, že 76% mladých Slovákov žije so svojimi rodičmi. Naše riešenie je zamerané na ľudí vo veku okolo 30 rokov hľadajúcich nové bývanie. Aplikácia umožní používateľovi vybrať si požadované kritéria, ktoré sú následne vizualizované. Taktiež sú na mape vizualizované dostupné ponuky nehnuteľnosti.

3.1 Vizualizácia

Ako je možné vidieť na obr. 3 používateľom zvolené kritéria sú vizualizované formou bodu s význačným okruhom o požadovanej vzdialenosti (ihriská, sakrálné objekty, verejné toalety a pod.) alebo formou zvýraznenia polynómu (územia ohrozené povodňami a zosuvmi pôdy a pod). Dostupné ponuky sú vizualizované formou bodov.



Obr. 3: Vizualizácia zvolených kritérií a ponúk

4 Záver

Naše riešenie ukazuje spôsob využitia otvorených dát poskytovaných verejnou správou. Na potvrdenie našej hypotézy o možnosti skrátiť čas potrebný na vyhľadanie ponúk nehnuteľnosti sme realizovali testovanie medzi členmi našej cieľovej skupiny. Test prebiehal formou porovnania času potrebného na vyhľadanie ponuky prostredníctvom našej aplikácie a prostredníctvom ľubovoľného existujúceho riešenia (kombináciou existujúcich riešení). Testovania času sa zúčastnilo 5 subjektov. Priemerný čas potrebný na nájdenie vhodnej ponuky bez využitia našej aplikácie bol 1:56 minút. S využitím našej aplikácie to bolo 54 sekúnd, čo potvrdzuje našu hypotézu a dokazuje skrátenie času potrebného na nájdenie vhodnej ponuky s využitím našej

aplikácie. Celkové priemerné hodnotenie aplikácie desiatimi používateľmi bolo 4.5/5 (1-najhoršie, 5-najlepšie) [4].

Literatúra

1. Bošanská, D., Bárdy, J.: Strategická priorita Open Data. Ministerstvo financií Slovenskej republiky, online: www.informatizacia.sk/ext_dok-07_ea_sp_open_data_v02/22728c, 2018
2. Chlapek D., Kučera J., Nečaský M.: Metodika publikace otevřených dat veřejné správy ČR, Praha, 2012
3. OpenData Barometer online: https://opendatabarometer.org/?_year=2016&indicator=ODB;3/2018
4. Babič T.: Mobilná aplikácia využívajúca otvorené dáta z prostredia verejnej správy (Bakalárska práca), Technická univerzita v Košiciach

PodĎakovanie

Táto práca bola podporená Agentúrou na podporu výskumu a vývoja v rámci grantu č. APVV-16-0213. Taktiež tento príspevok vznikol s podporou VEGA 1/0493/16 a KEGA 005TUKE-4/2017 - Zavedenie praktickej výučby mobilných technológií prostredníctvom vývoja aplikácií pre inteligentné zariadenia

Annotation

Usage of government open data to simplify real estate selection in Slovak republic

Recently, the government and certain organizations publish a huge amount of data which often remain raw and unused data. The article is focused on open-data in Slovak republic. It shows amount and quality of data published by several government organizations. The selected data in the field of public relation is used to create mobile application which is able to provide this data for public to help to find new living. Mobile application is being developed on a base of Android which belongs among advanced platforms. The article also contains results of questionnaire choice of suitable data, relevant for real-estate domain. The conclusion shows that our application can reduce time to find required housing offer.

Semiautomatizované porovnávanie certifikačných schém cloudových služieb

Martin Labaj, Karol Rástočný, Daniela Chudá

Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
{meno.priezvisko}@stuba.sk

Abstrakt. Problematika bezpečnosti informačných technológií je v súčasnej dobe jedna z najvážnejších požiadaviek v súvislosti s využívaním cloudových služieb. Aby poskytovatelia cloudových služieb vedeli deklarovať potrebnú úroveň bezpečnosti spracovania ich údajov, podstupujú certifikácie nezávislými inštitúciami. Z dôvodu vypuklosti tejto problematiky ako aj z rôznorodých požiadaviek trhu však v súčasnosti existuje pomerne veľké množstvo rôznych certifikačných schém, štandardov a regulácií zameraných na bezpečnosť a spracovanie dát, čo vedie k neprimeranej záťaži poskytovateľov cloudových služieb ako aj neprehľadnosti trhu z hľadiska ich zákazníkov. V článku preto navrhujeme systém umožňujúci semiautomatizované spracovanie existujúcich certifikačných schém, štandardov a regulácií, ktorý ich pomocou metód spracovania prirodzeného jazyka formalizuje do ontológie. Takto formalizované schémy následne porovnáva, vďaka čomu pomáha používateľom a poskytovateľom cloudových služieb vyhodnotiť, splňanie požiadaviek na bezpečnosť dát cloudovými službami.

Kľúčové slová: certifikácia cloudových služieb, spracovanie prirodzeného jazyka, ontológia certifikačných schém

1 Problematika porovnávania certifikačných schém

Bezpečnosť informačných technológií je v súčasnosti ešte vypuklejšia vďaka častému používaniu cloudových služieb na rôznych úrovniach softvéru – od infraštruktúry, platforiem až po aplikácie. Existujú rôzne zaužívané certifikačné schémy, podľa ktorých poskytovatelia cloudových služieb podstupujú proces certifikácie a získaným certifikátom tak preukazujú vlastnosti svojich služieb pre svojich používateľov (zákazníkov cloudových služieb). Zároveň do tejto problémovej oblasti vstupujú aj rôzne štandardy a regulácie týkajúce sa rôznych odvetví (napríklad bankovníctvo), rôznych typov používateľov (typicky verejná správa) ako aj rôzne národné zákony v rôznych krajinách Európskej únie (EÚ). Vďaka tomu je situácia na trhu cloudových služieb neprehľadná najmä pri cezhraničnom poskytovaní, obstarávaní a používaní cloudových služieb.

Cezhraničný trh alebo pokus o jeho rozširovanie si tak vyžaduje značné úsilie zo strany poskytovateľov cloudových služieb ako aj zákazníkov. Poskytovatelia, pokiaľ chcú podchytiť zákazníkov, ktorí vyžadujú alebo musia vyžadovať konkrétne certifikáty, musia podstúpiť ďalšiu certifikáciu podľa viacerých schém, napriek tomu, že

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 183-186.*

rôzne štandardy a certifikačné schémy sa výrazne prekrývajú a poskytovateľ už mnoho požiadaviek z ďalšej pridávanej schémy splnil. Zákazník naopak nevie vyhodnotiť, či daný poskytovateľ certifikovaný podľa rôznych schém spĺňa jeho vlastné alebo jemu predpísané požiadavky. Jedným z riešení je vytvorenie ručného mapovania medzi schémami a externými požiadavkami ako aj medzi schémami navzájom – ak zákazník potrebuje schému A a poskytovateľ služieb certifikovaný podľa schémy B na základe vytvoreného mapovania pozná rozdiely medzi schémami a vie, ktoré konkrétne požiadavky zostali nepokryté, vďaka čomu vie určiť najefektívnejší postup doplnenia vlastností cloudovej služby a získať tak nového zákazníka.

Vytváranie ručného mapovania však naráža na značnú rozsiahlosť schém a požiadaviek ako aj ich počet. Ak uvažíme n dokumentov (ako *dokument* ďalej rozumieme akúkoľvek certifikačnú schému, zákon, reguláciu, zoznam požiadaviek), vytvorenie počiatočného mapovania predstavuje úsilie $O(n^2)$, porovnanie „každého s každým“. Navyše tieto dokumenty nie sú statické, ale v čase sa vyvíjajú a pri aktualizácii jediného dokumentu je potrebné zopakovať porovnanie so všetkými ostatnými s úsilím $O(n)$.

Preto navrhujeme systém umožňujúci semiautomatizované porovnávanie certifikácií cloudových služieb na základe formalizácie dokumentov pomocou ontológie. Práve prevodom voľne vyjadrených textových požiadaviek do ontologického zápisu získavame možnosť porovnávať všetky dokumenty navzájom po vynaložení počiatočného úsilia $O(n)$ na prevod n dokumentov, čo predstavuje značnú úsporu napriek tomu, že úsilie na formalizáciu jedného dokumentu môže byť vyššie než pri prostom vytváraní ručného mapovania a porovnávaní dvoch dokumentov. Takisto pri aktualizácii jedného dokumentu je potrebné aktualizovať iba formalizovaný opis toho konkrétneho dokumentu.

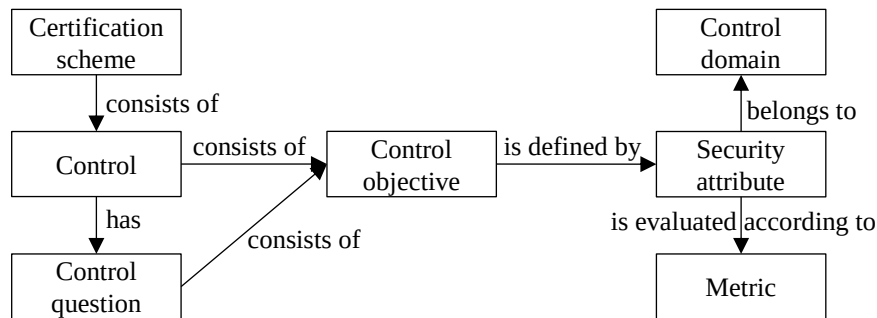
2 Použitie ontológie na formalizáciu certifikačnej schémy

Základom navrhovaného prístupu je nami navrhovaná ontológia bezpečnostných požiadaviek na cloudové služby vychádzajúca z formátu Linked USDL používaného na modelovanie dohôd o úrovni služieb [1]. Rôzne ontológie boli v minulosti použité na modelovanie zraniteľností, rizík, bezpečnostných politík [2, 3], avšak na vyššie uvedené účely je pre nás dôležité najmä zachytenie jednotlivých požiadaviek v tvare umožňujúcom jednoduché porovnávanie. Preto navrhujeme štruktúru v tvare *Control objective* (kontrolovaná vlastnosť) > *Security attribute* (bezpečnostný atribút) > *Metric* (metrika), ktorá umožňuje rozloženie požiadavky na časti, v rámci ktorých je možné vyhodnocovať prekrytia príbuzných požiadaviek vychádzajúcich z rôznych dokumentov (Obrázok 1). Napríklad teoretická požiadavka „Plot okolo dátového centra musí mať výšku minimálne 2 m.“ sa prevedie na „ ≥ 2 “ (*Control objective*) > „Výška plotu dátového centra“ (*Security attribute*) > „Výška“ (*Metric*).

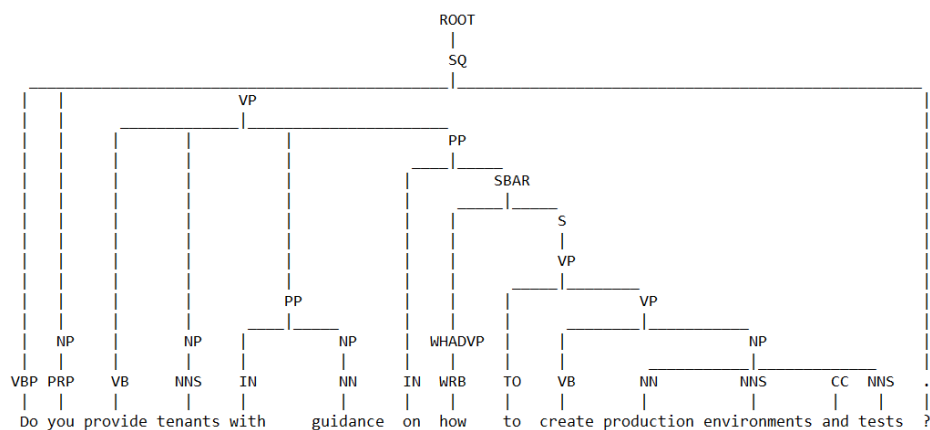
3 Automatizácia pomocou spracovania prirodzeného jazyka

Prevod požiadaviek, ktoré vo voľnom texte často predstavujú zložené rozvinuté súvetia (napr. „Do you provide tenants with guidance on how to create production

environments and tests?“, IVS-08.2 CCM v3.0.1), zachytávajúcích množstvo čiastkových jednotlivých detailov, môže predstavovať netriviálnu prácu. Na uľahčenie formalizácie navrhujeme použitie metód spracovania prirodzeného jazyka na semiautomatizáciu prepisu. Skladbu viet v požiadavkách môžeme dostupnými nástrojmi analyzovať, aplikovaním heuristik ich previesť na viaceré jednoduché vety a upraviť ich na tvar vyjadrený v požiadavke priamo prevoditeľnej do ontologického zápisu. Na Obrázku 2 je znázornený príklad anotácie jednoduchšej požiadavky z dotazníka certifikačnej schémy. Úplné spracovanie prirodzeného jazyka je v súčasnosti naďalej otvoreným výskumným problémom zahŕňajúcim napríklad tzv. *right node raising*, kde viaceré paralelné konštrukcie vety môžu zdieľať jeden predmet. Preto tento pomocný proces navrhujeme ako poloautomatizovaný, pričom pomocou spracovania prirodzeného jazyka môžeme navrhnúť prvotné rozdelenie rozvitej požiadavky na jednoduché vety prepisované do ontologickej formalizácie, ktoré schvaľuje a prípadne upravuje doménový expert.



Obr. 1. Logický model formalizácie pomocou ontologického prístupu.



Obr. 2. Príklad spracovania kontrolnej otázky IVS-08.2 z dotazníka certifikačnej schémy CCM v3.0.1 na samostatné požiadavky „Tenants must be provided with guidance on how to create production environments.“, „Tenants must be provided with guidance on how to create tests.“.

4 Záver

Navrhovaný systém sme experimentálne overili vytvorením prototypu vo forme webovej aplikácie implementovanej v jazyku Python pomocou rámca Django. Na spracovanie viet v prirodzenom jazyku sme použili knižnice Stanford CoreNLP a Nodebox English Linguistics. Pomocou prototypu systému sme prepísali vybrané časti schém CSA CCM a ISO 27001:2013 do ontologického formátu a vykonali vzájomné porovnanie.

Navrhovaný prístup nie je obmedzený len na doménu bezpečnosti cloudových služieb, vzhľadom na podobnosť s problematikou schém a požiadaviek sa ako veľmi príbuzná oblasť črtá problematika porovnania zákonov v rôznych krajinách – príkladom môžu byť zákony o cestnej premávke, ktoré sú naprieč krajinami EÚ veľmi podobné, no napriek tomu sa v mnohých detailoch výrazne líšia.

Literatúra

1. Garcia, J. M., Fernandez, P., Pedrinaci, C., Resinas, M., Cardoso, J., & Ruiz-Cortes, A.: Modeling Service Level Agreements with Linked USDL Agreement. *IEEE Transactions on Services Computing* 10(1) (2017), 52–65.
2. Takahashi, T., Kadobayashi, Y., & Fujiwara, H.: Ontological Approach Toward Cybersecurity in Cloud Computing. In: *Proc. of the 3rd int. conference on Security of information and networks - SIN '10*, ACM Press, New York (2010), 100–109.
3. Veloudis, S., & Paraskakis, I.: Ontological Templates for Modelling Security Policies in Cloud Environments. In: *Proc. of the 20th Pan-Hellenic Conference on Informatics - PCI '16*. ACM Press, New York (2016), 1–6.

PodĎakovanie: Tento článok vznikol vďaka čiastočnej podpore projektov EÚ 731845 – EU-SEC a APVV-15-0508.

Annotation:

Semiautomatic comparison of cloud service certification using natural language processing and ontologies

Security is currently one of the most important requirements for cloud service usage. Cloud service providers undergo various certifications through independent institutions and obtain certificates to signal security of their offered services. Currently, there are many various certification schemes, security standards and laws, and this complicates choosing the right service based on the regulations and requirements of the customer while it also places additional burden on providers wishing to enter other European markets. We propose a system allowing semi-automatic processing of existing certification schemes, standards and regulations using natural language processing to process them into ontological representation. The formalized representation allows automatic comparison of such documents and allows both cloud service users and providers to evaluate security requirements in cloud services.

Využitie kombinácie User experience a agilného prístupu v procese vývoja webovej aplikácie

Viera Pavlišková, František Babič, Ľudmila Pusztová

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Letná 9, 042 00 Košice

viera.pavlikova@student.tuke.sk, frantisek.babic@tuke.sk,
ludmila.pusztova.2@tuke.sk

Abstrakt. V súčasnosti je čoraz ťažšie vytvoriť webovú aplikáciu, ktorá by ponúkla používateľovi pozitívnu skúsenosť, motivovala ho k akcii alebo k opätovnej návšteve. V článku popisujeme proces vývoja webovej aplikácie pre reálneho zákazníka, v ktorom sme využili vhodné prvky UX dizajnu, agilnej metodiky a gamifikácie. User experience (UX) dizajn je dôležitý pre určité typy digitálnych produktov, keďže okrem šťastia, zážitku a pozitívnej skúsenosti vedie používateľa k efektívnej realizácii príslušných akcií, napr. úspešné dokončenie nákupu v e-shope. Agilný prístup oblasti softvérového vývoja je vhodný pre projekty, ktoré vyžadujú vyššiu mieru inovácií a upresňovanie v priebehu vývoja. Gamifikácia predstavuje aplikáciu herných prvkov a mechanizmov do neherných prostredí s cieľom podporiť motiváciu, angažovanosť a produktivitu cieľovej skupiny používateľov. Finálna responzívna webová aplikácia je výsledkom vlastnej kombinácie agilného prístupu a princípu HCD.

Kľúčové slova: webová aplikácia, UX, agilný model, HCD

1 Úvod a motivácia

V súčasnosti existuje veľké množstvo mobilných, či webových aplikácií. Nie je ťažké vytvoriť aplikáciu, avšak čoraz náročnejšie je presvedčiť a zaujať používateľa, prísť na trh s niečím novým, inovatívnym a pútavým. Jedným zo zaujímavých spôsobov, ako navrhnuť aplikáciu „na mieru“ pre používateľa, je zavedenie prvkov User Experience (UX) do procesu vývoja softvérového produktu.

J. Marsh vo svojej publikácii [4] spomína, že všetko, s čím sa človek stretáva a používa v ňom zanecháva používateľskú skúsenosť. Pri snahe zakomponovať UX dizajn do aplikácie by nemalo byť prioritou vytvoriť používateľskú skúsenosť, ale urobiť ju výnimočnou, dobrou a zaujímavou. Je to niečo, čo bude robiť používateľa šťastným a po použití produktu bude mať príjemný zážitok. Cieľom UX dizajnéra nie je len sprostredkovať šťastie, zážitok alebo skúsenosť, ale aj podporiť efektívnosť používateľa pri realizácii príslušných akcií. Veľa ľudí si pojem UX zamieňa za User Interface (UI), kde UI predstavuje používateľské rozhranie.

HCD (Human-centered Design) je jedna z metodík, pri ktorej sa zaoberáme dosiahnutím UX. Základný konceptuálny model HCD znázorňuje súvislosť medzi istou

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 187-192.*

úrovňou ľudskej potreby (Maslowova hierarchia potrieb) a dizajnovým vývojom (3 periódy) [8]:

- Perióda zameraná na funkčnosť, v rámci ktorej sa dizajn zameriaval na užitočné produkty a služby, ktoré spĺňajú základné ľudské potreby: fyziologické, potreba bezpečia a istoty.
- Perióda zameraná na spotrebiteľa, kedy sa dizajn presúva od užitočnosti k použiteľnosti; rozhodnutia dizajnérov sú motivované potrebou lásky, spolupatričnosti, uznania a úcty.
- Dizajn zameraný na človeka, ktorý sa snaží uspokojiť najvyššiu úroveň potrieb – sebarealizácie, ktorá zahŕňa kognitívne, estetické, sebarealizačné a sebavzdelávacie potreby.

Agilný model vytvárania digitálneho produktu predstavuje zdokonalený inkrementálny model, v rámci ktorého sú projektový tím a zákazník (koncový používateľ) v úzkom kontakte počas celého trvania procesu vývoja. Agilný model rozdeľuje očakávaný produkt na malé prírastky, ktoré sú poskytované v iteráciách. Scrum [6] predstavuje jednu z najpopulárnejších agilných metodík, o ktorej existuje množstvo dostupných materiálov, preto sa jej nebudeme venovať detailnejšie.

Gamifikácia patrí v súčasnosti medzi obľúbené termíny, ktoré sa často používajú pri rôznych príležitostiach v IKT doméne. Andrzej Marczewski v knihe [3] definuje gamifikáciu ako aplikovanie herných prvkov do reálnych životných úloh, ktoré ovplyvňujú správanie, zlepšenie motivácie a posilnenie angažovanosti. Ide o akési sústredenie sa na slová ako angažovanosť, motivácia, zmeny správania a produktivita.

Tento článok popisuje proces vývoja webovej aplikácie na základe identifikovaných požiadaviek a očakávaní reálnej cieľovej skupiny. Do fázy návrhu sme zakomponovali vybrané prvky UX dizajnu, aby sme zabezpečili čo najvyššiu spokojnosť cieľovej skupiny s finálnym produktom.

1.1 Prehľad existujúcich riešení

Autori Coatta a Gosper v článku [1] uvádzajú, že zakomponovanie rôznych prvkov ako UX dizajn, agilný prístup alebo interakcia človek-počítač (HCI) zefektívňujú vývoj softvérového produktu, napr. atď. Avšak práve UX dizajn je často zanedbávaný a do procesu vývoja je zavádzaný len v prípade, že je k dispozícii dostatočný rozpočet a časová dotácia. Dobre premyslený UX dizajn však môže byť rozhodujúcim faktorom pre úspech výsledného produktu.

Pri vývoji produktu je dôležité mať k dispozícii kvalitné používateľské príbehy, na ktorých sa dohodli členovia projektu; obsahujúce jasné definície súvisiacich pracovných postupov a prípadov použitia. Mnoho tímov sa obáva zavádzania agilnej metodiky pri realizácii UX dizajnu, pretože zahŕňa akceptovanie zmeny požiadaviek aj neskoro počas vývoja produktu. Po zavedení tohto prístupu softvérové tímy často zisťujú, že agilný prístup je efektívnejší na realizáciu UX dizajnu ako tradičný vodoпадový model [5].

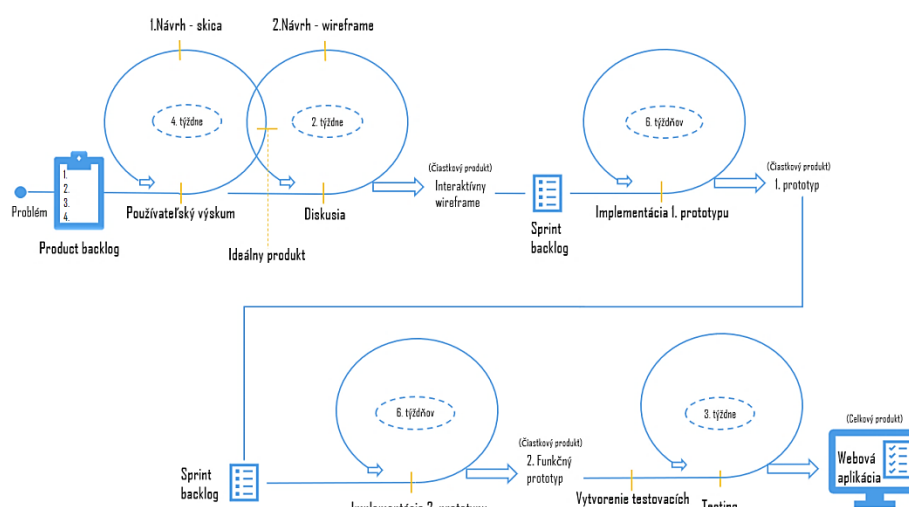
Zároveň sa výskumníci zaoberajú spôsobmi, ako efektívne riadiť túto kombináciu. Ferreira a kol. [7] vo svojej práci analyzovali interakcie medzi zabehaným Scrum tímom a UX dizajnérmi na spoločnom projekte. Zistili, že nebolo jednoduché nastaviť

pravidlá spolupráce najmä z pohľadu rozdielnych pracovných kultúr, spôsobu fungovania alebo vnímania UX ako vstupu z externého prostredia. Na základe štúdie realizovanej v softvérových spoločnostiach v Taliansku a Dánsku kolektív autorov na tento účel využil metódu s názvom „Cooperative Method Development“, aby zlepšil interakciu medzi všetkými účastníkmi vývoja a identifikovali kritické miesta [9].

2 Agilný model softvérového vývoja s prvkami UX

Podľa [2] je UX dizajn dôležitý pre určité typy digitálnych produktov: komplexné stránky alebo aplikácie, malo-obchodný predaj alebo on-line predaj, začínajúce biznisy, malé rozpočtové projekty a ďalšie. UX je kľúčom pre zložité stránky, pretože používatelia musia byť schopní dostať sa ľahko na stránku a porozumieť, ako s ňou pracovať a používať. Zanedbanie UX môže vyústiť do nedbalého webu, na ktorý sa ľudia opäť nevrátia. Zapamätateľné bohaté skúsenosti a pozitívna interakcia privedie používateľov na web opäť.

V našej práci sme použili kombináciu agilného modelu a metodiky HCD ako prostriedku pre dosiahnutie UX. Nami navrhnutou kombináciou (Obr.1) sme zabezpečili splnenie požiadaviek cieľových používateľov, aby výsledný produkt bol pochopiteľný, jednoducho použiteľný a skúsenosť s ním zapamätateľná, pozitívna a príjemná.



Obr. 1 Agilný model v kombinácii s HCD

2.1 Používateľský výskum

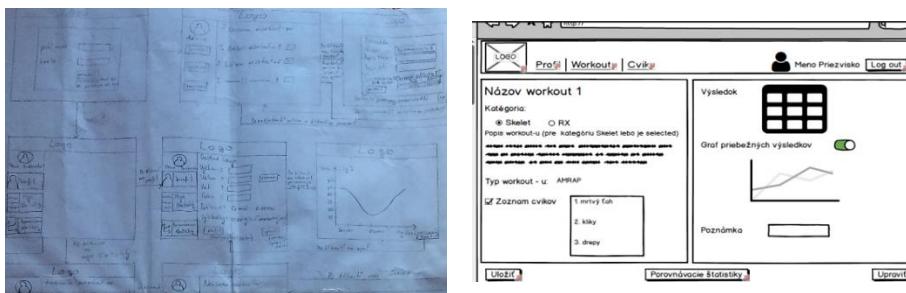
Vo fáze používateľského výskumu sme realizovali rôzne diskusie, rozhovory, skúmanie správania používateľov a individuálne pohovory. Cieľom bolo vytvoriť základnú množinu požiadaviek, funkčných aj nie-funkčných, ktoré by definovali očakávaný produkt. Dĺžka trvania používateľského výskumu bola jeden týždeň.

2.2 Ideálny produkt

Keďže cieľoví používatelia potrebovali aplikáciu, ktorá bude dostupná pre všetky platformy, rozhodli sme sa pre webovú responzívnu aplikáciu. Na základe vytvorenej množiny požiadaviek sme definovali tzv. „product backlog“.

2.3 Informačná architektúra

Informačná architektúra (IA) je dôležitou časťou UX dizajnu, ktorej výsledkom je vytvorenie logickej štruktúry a vzájomnej nadväznosti obsahu celej aplikácie. Aj napriek tomu, že je tento proces časovo náročnejší a jeho tvorba je namáhavejšia, do procesu návrhu riešenia sme ho zakomponovali, pretože je efektívnejší, výsledok je spoľahlivejší a navrhnutý podľa požiadaviek cieľovej skupiny. Na začiatku sme vytvorili sériu papierových prototypov - skíc. Skica slúži aj na definovanie funkcionality aplikácie, je možné ju upravovať až do momentu, pokiaľ je používateľ spokojný. Následne sme vytvorili sériu wireframes, ktoré slúžia nielen na zobrazenie obsahu ale aj ukážku fungovania interaktívnych prvkov.



Obr. 2 Ukážka papierovej skice a vytvorených wireframes.

2.4 Implementácia a testovanie

Dĺžku trvania jedného Sprintu sme odhadli na 5-6 týždňov; do Sprint backlogu sme postupne zaradovali požiadavky podľa priradenej priority. Prvý prototyp obsahoval základnú množinu funkcií, ktoré ešte nebolo možné prezentovať používateľom. Zároveň v Sprint backlogu zostalo ešte niekoľko nesplnených požiadaviek, implementáciou ktorých by sme dosiahli funkčný prototyp. Tie boli zaradené do Sprintu č.2 spolu s dodatočnými požiadavkami. Výsledkom bol funkčný prototyp responzívnej webovej aplikácie pripravený na testovanie.

Cieľom testovania bolo odhalenie skrytých chýb na základe vopred definovaných testovacích prípadov. Prvým krokom bolo testovanie samotným programátorom, ktoré odhalilo niekoľko chýb priamo v zdrojovom kóde, ktoré sa prejavovali nesprávnou alebo nefungujúcou akciou na používateľskom rozhraní. Druhým krokom bolo akceptačné testovanie pomocou vybranej skupiny koncových používateľov. Výsledkom bola pozitívna skúsenosť u väčšiny účastníkov testovania; aplikáciu hodnotili ako prehľadnú, jednoducho použiteľnú a vizuálnej príjemnú.

3 Záver

Tento článok popisuje proces vývoja webovej aplikácie prostredníctvom kombinácie agilného prístupu a HCD metodiky zastupujúcej spôsob, ako dosiahnuť pozitívnu používateľskú skúsenosť. Cieľom bolo zabezpečiť úzku a intenzívnu interakciu s cieľovou skupinou používateľov, efektívny manažment a zapracovanie používateľských požiadaviek.

Po úspešnom dokončení celého cyklu vývoja sme došli k záveru, že touto kombináciou dokážeme predísť viacerým chybám alebo nezhodám s cieľovým používateľmi. K tomu prispela aj edukatívna funkcia tohto prístupu, keď sme používateľa viedli k tomu, aby bola naša komunikácia čo najefektívnejšia, aby požiadavky definoval v kontexte použitia v praxi, aby si už vo fáze návrhu vyskúšal a zamyslel sa nad fungovaním čiastkových prototypov, či už v papierovej alebo elektronickej podobe.

Jednoduchosť, inovatívnosť, prehľadnosť a použiteľnosť sú známkami kvalitného UX dizajnu.

Literatúra

1. Coatta, T., Gosper, J.: UX design and Agile: a natural fit?. In: Communications of the ACM (2011) 54 60.
2. Cousins, C.: Why Does User Experience Matter? } (2017) Dostupné na internete: <<https://www.ceros.com/originals/why-does-user-experience-matter/>>.
3. Marczewski, A.: Gamification: A Simple Introduction. (2013).
4. Marsh, J.: UX for beginners. Kanada, O'Reilly Media (2015).
5. Plonka, L., et al.: UX design in agile: a DSDM case study. In: Cantone G., Marchesi M. (eds) Agile Processes in Software Engineering and Extreme Programming. Springer International Publishing (2014).
6. Sutherland, J., Schwaber, J.: Scrum Development Process. In: OOPSLA Business Object Design and Implementation Workshop, Object-Oriented Programming, Systems and Applications (OOPSLA) convention, Austin, Texas, USA (1995)
7. Ferreira, J., Sharp, H., Robinson, H.: User experience design and agile development: managing cooperation through articulation work. In: Journal of Software: Practice and Experience, 41(9), p. 963-974 (2011)
8. Zhang, T., Dong, H.: Human-centred design: an emergent conceptual model. Include 2009, Inclusive Design into Innovation: Transforming practice in design, research and business Royal College of Art, London (2009)
9. Ardito, C., Buono, P., Caivano, D., Costabile, M.F., Lanzilotti, R.: Investigating and promoting UX practice in industry: An experimental study. In: International Journal of Human-Computer Studies, 72(6), p. 542-551 (2014)

PodĎakovanie: Táto publikácia vznikla vďaka podpore APVV projektu č. APVV-16-0213 a KEGA projektu 005TUKE-4/2017.

Annotation:

Web application development process as a combination of User experience and Agile methodology

This article describes the process of developing a web application, utilizing appropriate elements of UX design, agile methodology, and gamification. User experience (UX) design is important for certain types of digital products, since, in addition to happiness, experience, and positive experience, the user can effectively implement the relevant actions, successful completion of e-shop purchases. The Agile methodology is suited to projects that require higher levels of innovation and refinement during the phases of design or development. The final responsive web application is the result of own combination of the Scrum and the HCD principle. Therefore, we were able to avoid multiple mistakes or disagreements with target users. Its educational function has also contributed to making the communication as effective as possible. The users were thinking about their processes when they have tested various partial prototypes.

Behaviorálna biometrická autentifikácia na mobilných zariadeniach

Kamil Burda, Daniela Chudá

Ústav informatiky, informačných systémov a softvérového inžinierstva,
FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
kamil.burda@stuba.sk, daniela.chuda@stuba.sk

Abstrakt. V príspevku sa zaoberáme autentifikáciou používateľa na základe jeho behaviorálnych biometrických charakteristík na mobilných zariadeniach. Príspevok poskytuje prehľad jednotlivých častí procesu modelovania používateľa pre jeho autentifikáciu od získavania surových dát cez extrakciu črt až po vykonanie samotnej autentifikácie. Autentifikáciu je možné reprezentovať ako problém binárnej klasifikácie, resp. detekcie anomálií, v závislosti od nasadenia riešenia v praxi. Na chybovosť autentifikácie, obzvlášť na mobilných zariadeniach, vplyva kontext používania zariadenia. Jedným z kontextov je rozdielnosť zariadení, čo spôsobuje značné rozdielnosti nielen v dôsledku odlišného správania používateľa, ale aj použitého hardvéru pre senzory na zariadeniach. Na základe experimentu vykonávania jednoduchých gest na troch rôznych zariadeniach možno skonštatovať, že rozdielne zariadenia majú veľmi negatívny vplyv na chybovosť autentifikácie a je potrebné prehodnotiť jednotlivé kroky modelovania používateľa na minimalizovanie vplyvu kontextov pre autentifikáciu používateľa.

Kľúčové slová: behaviorálna biometria · autentifikácia · mobilné zariadenia

1 Úvod

Vzhľadom na rozšírenosť mobilných zariadení (smartfóny a tablety) vzniká potreba ich zabezpečovania voči odcudzeniu. Tradičné mechanizmy na *autentifikáciu* (overenie identity) používateľa na mobilných zariadeniach, ako heslo, PIN alebo vzor, nemusia byť postačujúce vzhľadom na ich jednoduché odcudzenie – je možné ich uhádnuť alebo odpozorovať.

V súčasnosti skúmanou alternatívou je *behaviorálna biometria*, ktorá vyhodnocuje správanie používateľa pri interakcii s mobilným zariadením (napr. gestá na dotykovej obrazovke, pohyb zariadením). Behaviorálna biometria využíva dáta z bežne dostupných senzorov (dotyková obrazovka, akcelerometer), ťažšie sa odcudzuje (ťažko vedome napodobniť správanie iného človeka) a je možné ju vyhodnotiť aj počas používania zariadenia (t.j. autentifikácia prebieha na pozadí počas interakcie s aplikáciou). Vzhľadom na vyššiu chybovosť v porovnaní s fyziologickými charakteristikami nie je v súčasnosti možné využívať behaviorálnu biometriu ako samostatnú metódu autentifikácie používateľa.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 193-197.*

2 Budovanie modelu používateľa pre jeho autentifikáciu

Na vykonanie behaviorálnej biometrickej autentifikácie je potrebné, aby si aplikácia alebo systém vytvoril reprezentáciu používateľa, teda vybudovala tzv. *model používateľa*. Proces budovania modelu používateľa pre jeho autentifikáciu možno členiť do nasledovných krokov [6]:

- zaznamenávanie (logovanie) surových dát z ovládacích zariadení,
- predspracovanie surových dát,
- extrakcia črt zo surových dát,
- porovnanie modelu používateľa s novozískanými vzorkami a určenie, či ide o oprávneného používateľa (vlastníka zariadenia) alebo narušiteľa (cudzej osoby).

2.1 Zaznamenávanie dát zo senzorov mobilného zariadenia

Vzhľadom na predpokladanú jedinečnosť správania používateľov dokážeme na autentifikáciu používateľa zúžitkovať dáta z nasledovných senzorov mobilných zariadení:

- dotyková obrazovka – poskytuje x a y súradnice pohybu po obrazovke, plochu dotyku a časovú pečiatku,
- akcelerometer, gyroskop, magnetometer – poskytuje hodnoty na osiach x , y a z spolu s časovou pečiatkou.

Ďalšie dostupné senzory ako napr. senzor blízkosti alebo svetla buď poskytujú nedostatočné množstvo dát alebo sú irelevantné pri vyhodnotení správania používateľa.

2.2 Predspracovanie dát

Na autentifikáciu nie je vhodné používať priamo surové dáta, keďže je dôležité, aby biometrické charakteristiky boli jedinečné a zároveň konzistentné pre každého používateľa. Preto je potrebné surové dáta zoskupiť do *segmentov* a zo segmentov extrahovať vhodné črty.

Surové dáta môžu byť rozdelené do segmentov podľa jedného alebo viac z nasledovných kritérií:

- explicitné udalosti, napr. začiatok a koniec gesta na dotykovej obrazovke [4,5],
- pravidelné intervaly s posuvným oknom (používané najmä pre dáta z akcelerometra),
- vzory (opakujúce sa motívy) v dátach [9].

Senzory ako akcelerometer, gyroskop alebo magnetometer vzhľadom na nepresnosť merania obsahujú šum, ktorý skresľuje zaznamenané dáta. Pri predspracovaní sa teda často používa aj filtrovanie šumu, resp. vyhladenie dát, napr. pomocou pohyblivého priemeru alebo Kalmanovho filtra.

2.3 Extrakcia črt

Z každého segmentu sa extrahujú (vypočítajú) črty, ktoré sú potenciálne jedinečné a konzistentné pre každého používateľa. Príkladmi črt z gesta zvislého ťahu na dotykovej obrazovke sú začiatočná a koncová pozícia gesta, priemerná rýchlosť vykonania gesta, štandardná odchýlka rýchlosti vykonania gesta, čas trvania gesta.

Črty možno extrahovať systematicky kombinovaním viacerých kritérií – použité štatistiky (napr. priemer, štandardná odchýlka) alebo použitá časť segmentu (celý segment, začiatok, koniec) [2]. Vzhľadom na potenciálne vysoký počet extrahovaných črt je potrebné odstrániť redundantné (navzájom korelujúce) a irelevantné (nevhodné pre autentifikáciu) črty pomocou metód výberu črt typu filter (napr. informačný zisk) alebo obálka (napr. prírodou inšpirované algoritmy).

Črty extrahované z jedného gesta tvoria vektor – *vzorku*. Množina vzoriek získaných zo segmentov teda potom reprezentuje používateľa.

2.4 Autentifikácia

Po získaní vzoriek zo surových dát je potrebné vybudovať model používateľa pre jeho autentifikáciu. Pre autentifikáciu sú využívané predovšetkým metódy strojového učenia, ako napr. k -najbližších susedov, mechanizmus podporných vektorov alebo náhodný les.

Keďže výsledkom autentifikácie je rozhodnutie, či daná vzorka patrí do modelu oprávneného používateľa alebo nie, je možné autentifikáciu realizovať dvoma prístupmi – binárna klasifikácia alebo detekcia anomálií.

Pri binárnej klasifikácii sa vybraná metóda natrénuje na vzorkách od oprávneného používateľa a narušiteľov. V prípade detekcie anomálií sa metóda natrénuje iba na vzorkách od oprávneného používateľa.

Binárna klasifikácia je vhodným riešením pre dodatočnú autentifikáciu vo webovej službe (napr. internet banking) vďaka možnosti nenáročného získavania dát od viacerých používateľov. V takomto prípade môže mať každý používateľ priradený svoj vlastný model, kde trieda narušiteľa reprezentuje podmnožinu vzoriek od ostatných používateľov [8].

Pre zabezpečenie prístupu do mobilného zariadenia lokálne je vhodnejšia voľba detekcia anomálií, keďže získavanie vzoriek od cudzích používateľov je v tomto prípade nepraktické a potenciálne nebezpečné. Detekcia anomálií však v súčasnosti dosahuje horšie výsledky než binárna klasifikácia [3], hoci kombinácia výsledkov viacerých detektorov anomálií môže viesť k zlepšeniu [7].

Na vyhodnotenie úspešnosti, resp. chybovosti biometrického systému sa zvyčajne používajú: miera falošného prijatia (FAR), miera falošného odmietnutia (FRR), priemer FAR a FRR (HTER) a vyrovnaná chybovosť (EER). Chybovosť EER sa v súčasných prácach pohybuje zvyčajne do 20%, čo indikuje, že zatiaľ nie je možné využívať behaviorálnu biometriu ako samostatnú metódu autentifikácie.

Chybovosť možno znížiť autentifikáciou na základe nielen jednej vzorky, ale sekvencie vzoriek [1], vytvorením jednej vzorky z črt extrahovaných z viacerých senzorov [4], alebo trénovaním klasifikátorov pre každý senzor zvlášť a následnou

kombináciou výsledkov, napr. váhovaným súčtom. Vysoká chybovosť môže byť zapríčinená aj meniacim sa prostredím, teda *kontextom*, v ktorom používateľ používa mobilné zariadenie, napr. telesná poloha alebo rozdielny hardvér zariadenia.

3 Vplyv rozdielneho hardvéru na autentifikáciu používateľa

Jedným z kontextov je používané zariadenie. Ak používateľ disponuje viacerými zariadeniami, extrahované črty sa líšia naprieč zariadeniami vzhľadom na odlišnú veľkosť dotykovej obrazovky, hmotnosť zariadenia, alebo hardvér senzorov s odlišnou presnosťou meraní.

Z experimentu s 18 účastníkmi, v ktorých účastníci kreslili a znázorňovali kruh na dotykovej obrazovke a následne pohybom zariadenia, sme vykonali autentifikáciu používateľa na troch rôznych zariadeniach (dva smartfóny a jeden tablet), pričom aj v trénovacej aj v testovacej množine sme použili vzorky zo všetkých zariadení. Pre každého používateľa sme natrénovali samostatný model používateľa.

Tabuľka 1 ukazuje výsledky autentifikácie používateľa pomocou náhodného lesu pri rôznom percentuálnom podiele počtu vzoriek od vlastníka. Výsledky naznačujú, že kvôli neakceptovateľne vysokej miere FRR je potrebné sa ďalej zaoberať znižovaním závislosti od konkrétneho zariadenia.

Tab. 1. Autentifikácia používateľa na rôznych mobilných zariadeniach.

Pomer vzoriek od vlastníka	FAR	FRR
35%	7%	86%
50%	7%	85%
85%	15%	70%

4 Zhrnutie

Tento príspevok predstavuje stručný prehľad možností biometrickej autentifikácie na mobilných zariadeniach. V závislosti od prípadu použitia je potrebné sa na problém behaviorálnej biometrickej autentifikácie pozeriť ako binárna klasifikácia alebo detekcia anomálií.

Keďže kontext používania mobilného zariadenia značne ovplyvňuje chybovosť autentifikácie, v budúcnosti sa budeme zaoberať minimalizovaním vplyvu kontextu na autentifikáciu používateľa.

Literatúra

1. Ben Kimon, L., Mirsky, Y., Rokach, L., Shapira, B.: User Verification on Mobile Devices Using Sequences of Touch Gestures. In: Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 365–366. UMAP '17, ACM, New York, NY, USA (2017)
2. Burda, K., Chuda, D.: Influence of Body Postures on Biometric User Authentication on Mobile Devices. Krems an der Donau, Austria (Jan 2018)
3. Buschek, D., De Luca, A., Alt, F.: Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1393–1402. CHI '15, ACM, New York, NY, USA (2015)
4. De Luca, A., Hang, A., Brudy, F., Lindner, C., Hussmann, H.: Touch Me Once and I Know It's You!: Implicit Authentication Based on Touch Screen Patterns. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 987–996. CHI '12, ACM, New York, NY, USA (2012)
5. Frank, M., Biedert, R., Ma, E., Martinovic, I., Song, D.: Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. IEEE Transactions on Information Forensics and Security **8**(1), 136–148 (Jan 2013)
6. Jain, A.K., Ross, A.A., Nandakumar, K.: Introduction to Biometrics. Springer US, Boston, MA (2011)
7. Kumar, R., Kundu, P.P., Phoha, V.V.: Continuous Authentication Using One-class Classifiers and their Fusion. arXiv:1710.11075 [cs] (Oct 2017)
8. Lee, W.H., Lee, R.B.: Implicit Smartphone User Authentication with Sensors and Contextual Machine Learning. arXiv:1708.09754 [cs] (Aug 2017)
9. Li, Y., Xie, M., Bian, J.: SEGAUTH: A segment-based approach to behavioral biometric authentication. In: 2016 IEEE Conference on Communications and Network Security (CNS). pp. 1–9 (Oct 2016)

Pod'akovanie: Tento článok vznikol vďaka čiastočnej podpore projektu APVV-15-0508 Informačné správanie sa človeka v digitálnom priestore.

Annotation:

Behavioral Biometric Authentication on Mobile Devices

In this paper we examine the process of user authentication on mobile devices based on behavioral biometric characteristics. This paper presents an overview of the process of user modeling for authentication, namely raw data logging, feature extraction and user model comparison. Authentication can be represented as a binary classification or novelty detection problem depending on the real-world scenario. The context in which mobile devices are used by the user may have a significant impact on authentication error rate. One of the contexts is different hardware that is not only influenced by user behavior (e.g. due to a different screen resolution or weight of each device), but also by different sensors built into the devices. Based on our experiment involving performing simple gestures on three mobile devices, we conclude that different devices have a significant negative impact on authentication error rate. Therefore, to develop a user model as context-independent as possible, alternate approaches to each stage of user modeling must be considered.

Personalizované odporúčanie krátkodobých ponúk

Peter Gaspar, Ivan Srba, Michal Barla, Michal Kompan a Maria Bielikova

Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2, 841 04 Bratislava
{peter_gaspar, ivan.srba, michal.barla, michal.kompan,
maria.bielikova}@stuba.sk

Abstrakt. Personalizované odporúčanie je už neoddeliteľnou súčasťou moderných webových aplikácií. V závislosti od špecifickej domény sa pri navrhovaní konkrétnych prístupov musíme vysporiadať s vlastnosťami daného sídla. V tomto článku prezentujeme opis reálneho odporúčacieho systému aplikovaného v doméne elektronického obchodu. Konkrétna inštancia elektronického obchodu je typická krátkym cyklom aktuálnosti predávaných produktov, čo kladie špecifické nároky pri návrhu a realizácii daného riešenia.

Kľúčové slová: personalizované odporúčanie · model používateľa · elektronický obchod

1 Úvod

S personalizovaným odporúčaním sa v súčasnosti stretávame v mnohých doménach (napr. filmy, hudba, online nakupovanie). Výskumníci sa pritom zameriavajú najmä na skvalitnenie výstupu odporúčania (z pohľadu relevancie pre používateľa). Vo väčšine domén vidíme potrebu zamerať návrh odporúčacieho systému na špecifiká webového sídla.

Príkladom takejto domény je e-obchod. Odporúčanie v doméne e-obchodu je užitočné z troch hľadísk [2]:

- *Od prehliadania k nákupu:* používatelia často prehliadajú e-obchody bez toho, aby si niečo zakúpili. Odporúčače im môžu pomôcť v tomto procese nájsť zaujímavé produkty.
- *Křížový predaj* (*angl. cross-sell*): Odporúčanie ďalších produktov môže navýšiť obsah nákupného košíka (napr. odporúčaním ďalších/doplňkových produktov k produktom, ktoré má zákazník vložené v košíku).
- *Lojálnosť:* odporúčania prinášajú personalizovaný zážitok pri návšteve stránok. V prípade, že je stránka šitá na mieru používateľa, má tendenciu byť lojálnejší a opätovne sa na ňu vracáť.

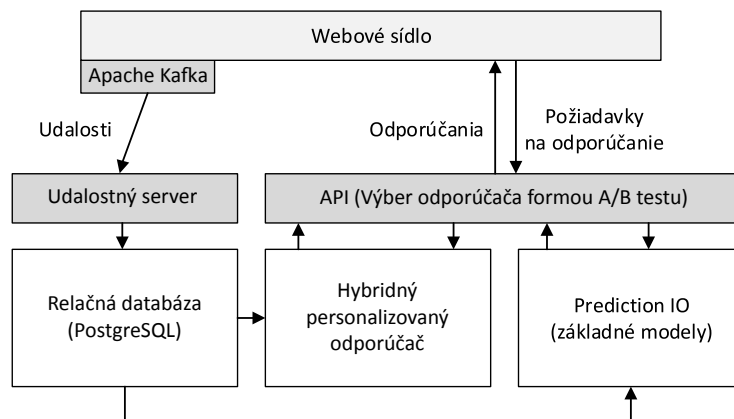
Aby mohli byť tieto hľadiská splnené, často sa vyžaduje prispôsobenie návrhu odporúčania na špecifiká tejto domény. Jeden z pohľadov je zameranie na predávané

produkty, kde sa odporúčanie niekedy musí vysporiadať nielen s generickým tovarom, ale napríklad aj so službami. Vo viacerých doménach sú ponúkané produkty s obmedzenou dobou platnosti. Ďalším úskalím, s ktorým sa môžeme stretnúť, je škálovateľnosť odporúčania, a to na viacerých úrovniach: predspracovanie dát, tréningovanie modelu, generovanie odporúčania a vyhodnotenie.

V našej práci sme navrhli odporúčací systém nasadený v doméne elektronického obchodu, ktorý patrí do špecifickej skupiny zľavových portálov. Realizácia berie do úvahy špecifiká samotného obsahu - krátka aktuálnosť predávaných zliav, ale aj nároky súvisiace s nasadením riešenia - výpočtová zložitosť, rýchlosť odozvy a pod.

2 Navrhnuté riešenie

Navrhnutý odporúčací systém sa skladá z viacerých samostatných komponentov, ktoré navzájom medzi sebou komunikujú (Obrázok 1). Z webového sídla e-obchodu získavame v reálnom čase udalosti, ktoré zahŕňajú aktivitu používateľov v e-obchode (zobrazovanie produktov, vkladanie do košíka, nákup, hodnotenie, vyhľadávanie, a pod.) a metadáta o dostupných produktoch. Tieto udalosti následne ukladáme do dátového úložiska (PostgreSQL databáza).



Obr. 1. Architektúra navrhnutého personalizovaného odporúčacieho systému.

Získané udalosti z dátového úložiska vstupujú do odporúčačov. V našej práci využívame 2 základné systémy odporúčania: generické odporúčanie poskytované predikčným systémom Apache PredictionIO¹ a špecifické nami vyvíjané hybridné modely odporúčania. Apache PredictionIO je platforma s otvoreným zdrojovým kódom zaradená do Apache Foundation projektu. V rámci nej využívame šablónu tzv. univerzálneho odporúčača², ktorý je založený na kolaboratívnom

¹ <https://predictionio.apache.org>

² <https://github.com/actionml/universal-recommender>

filtrovaní a využíva algoritmus *Correlated Cross-Occurrence*³. Ten umožňuje jednoducho realizovať scenáre typické pre e-obchody: odporúčanie podobných produktov, odporúčanie v nákupnom košíku, populárne produkty, a pod. Apache PredictionIO využíva pre uloženie a vyhľadávanie v modeli odporúčania ElasticSearch⁴, škálovateľnosť riešenia zabezpečuje knižnica Spark⁵.

Hybridné modely odporúčania sa zameriavajú na lepšiu adaptáciu na špecifickú doménu. Príkladom je, že každý produkt má obmedzenú dobu zobrazenia a obmedzenú dobu platnosti (počas ktorej ho možno využiť). Taktiež sú tieto modely schopné automatizovane identifikovať produkty, ktoré sa po čase opäť objavajú v ponuke e-obchodu (táto informácia nie je explicitne dostupná v dátach). Tieto skutočnosti je možné zohľadniť aj v rámci Apache PredictionIO, avšak naše experimenty ukázali, že toto riešenie nebolo postačujúce.

V hybridnom modeli odporúčania vychádzame z hybridného prístupu k odporúčaniu založenom na maticovej faktorizácii, ktorá využíva aj vlastnosti používateľa a položiek. Pre realizáciu riešenia využívame knižnicu LightFM [1], pre naučenie modelu odporúčania stratovú funkciu WARP (Weighted Approximate-Rank Pairwise) [3]. Okrem kolaboratívneho prístupu je súčasťou hybridného modelu aj obsahový odporúčač. Ten umožňuje nachádzať podobné produkty na základe príslušnosti ku kategóriám, textového opisu (využívame základné spracovanie textu, ako tokenizácia, lematizácia a prevod na model TF-IDF) a na základe latentnej reprezentácie produktov, ktorá je výstupom tréningu modelu.

Výhodou nášho riešenia v porovnaní s Apache PredictionIO sú okrem prispôbitelnosti na špecifickú doménu aj menšie nároky na hardvérové prostriedky (najmä pamäť) a rýchlejší čas tréningu modelu.

V závislosti od umiestnenia výsledkov odporúčania v rámci webového sídla navyše rozlišujeme dva prístupy k odporúčaniu:

- alternatívne odporúčanie – umiestnené v detaile produktu; snažíme sa odporúčať položky podobné k tej, ktorú používateľ aktuálne zobrazuje,
- komplementárne odporúčanie – umiestnené v nákupnom košíku; snažíme sa odporúčať doplnkové položky k tým, ktoré má používateľ v nákupnom košíku.

V prípade PredictionIO realizujeme oba prístupy pomocou kontextuálneho odporúčania, ktoré považuje za kontext zobrazený produkt resp. produkty v košíku. Pri hybridnom alternatívnom modeli počítame kosínusovú podobnosť medzi vektorom reprezentujúcim zobrazený produkt a vektormi ostatných produktov (vo webovom sídle) na základe obsahu (využívame opis produktu). Pri hybridnom komplementárnom modeli počítame kosínusovú podobnosť medzi položkou v košíku a ostatnými produktami e-obchodu na základe pravdepodobnosti, že sa spoločne nakupujú (využívame latentné črty produktov, ktoré vytvára pri tréningu modelu knižnica LightFM).

³ <https://mahout.apache.org/users/algorithms/intro-cooccurrence-spark.html>

⁴ <https://elastic.co>

⁵ <https://spark.apache.org>

Oba modely (PredictionIO a hybridný model) umožňujú personalizované i nepersonalizované odporúčanie, pričom personalizované odporúčanie podporujú už od prvej interakcie používateľa s produktom webového sídla. V prípade, že modely nemajú k dispozícii interakcie používateľov, automaticky poskytujú populárne produkty ako výsledok odporúčania.

Samotné odporúčanie produktov je realizované v aplikačnom rámci Flask⁶ založenom na jazyku Python. Ten zároveň využívame aj na prijímanie udalostí z webového sídla, a to vrátane spätnej väzby na prezentované odporúčanie. Pri prichádzajúcej požiadavke na odporúčanie realizujeme výber konkrétneho systému odporúčania (Apache PredictionIO, hybridný model).

3 Overenie

Pri overení riešenia sme porovnávali úspešnosť odporúčaní PredictionIO a vlastného hybridného modelu. Zamerali sme sa na metriky relevantné pre doménu e-obchodu: miera prekliku (*angl.* click-through-rate, CTR) a miera obratu (*angl.* conversion rate). Mieru prekliku sme vypočítali ako podiel počtu kliknutí na odporúčanie z počtu zobrazení odporúčania. Mieru obratu sme vypočítali ako podiel počtu vytvorených objednávok (obsahujúcich produkty z odporúčania) a počtu zobrazení odporúčania, resp. kliknutí na odporúčanie.

Tab. 1. Porovnanie úspešnosti odporúčaní PredictionIO a hybridného odporúčania pre alternatívny (Alternat.) a komplementárny (Komplement.) prístup.

Metrika	PredictionIO		Hybridný model	
	Alternat.	Komplement.	Alternat.	Komplement.
Miera prekliku	0,1442	0,0211	0,2226	0,0164
Miera obratu (zo zobrazenia)	0,0013	0,0006	0,0015	0,0001
Miera obratu (z kliku)	0,0087	0,0303	0,0068	0,0069

Overenie sme realizovali online, a to v prostredí slovenského zľavového portálu. Používateľom sme počas sledovaného obdobia (2 týždne) prideliťovali v pomere 50:50 odporúčací systém, ktorý im generoval odporúčania (PredictionIO, Hybridný model). Toto pridelenie sa pritom pre používateľa nemenilo, čím bola zachovaná konzistentnosť jeho správania (alternatívne aj komplementárne odporúčania mu boli generované vždy rovnakým systémom).

Ako môžeme vidieť z výsledkov (Tabuľka 1), pri alternatívnom prístupe dosahuje vo všeobecnosti lepšie výsledky hybridný model, pri komplementárnom prístupe PredictionIO. Rozpor medzi mierou obratu zo zobrazenia a kliku vidíme najmä v správaní používateľov, kedy používatelia mali v prípade hybridného modelu tendenciu častejšie klikať na výsledky odporúčania (vidíme z miery prekliku), čo sa prejavilo na tom, že menovateľ v miere obratu z kliku dosahoval

⁶ <https://flask.pocoo.org>

vyššie hodnoty. V celkovom počte nákupov (číslo neuvádzame) dosahoval hybridný model lepšie výsledky ako PredictionIO.

Škálovateľnosť riešenia preukazuje viac ako 224 tisíc udalostí o správaní používateľov prijatých z elektronického obchodu denne ako aj viac ako 105 tisíc vygenerovaných odporúčaní. Vďaka použitiu distribuovaných technológií sa odporúčania generujú v priemere za 50 ms.

4 Záver

V našom projekte sme sa zamerali na vytvorenie odporúčacieho systému s dôrazom na špecifiká zvolenej domény a požiadavky na produkčné riešenie. Prezentovaný systém umožňuje v reálnom čase zaznamenávať aktivitu používateľov a v pravidelných intervaloch pretrénovať model odporúčania, čo zaručuje reflektovanie zmien v správaní používateľov. Riešenie pozostáva zo sady odporúčacích algoritmov cielených na rôzne hľadiská - napr. zvýšenie spokojnosti používateľov, odporúčanie doplnkových produktov. V ďalšej práci sa sústreďujeme na ďalšie úlohy, v ktorých prináša personalizované odporúčanie pozitívne výsledky - tvorba personalizovaných noviniek, adaptívnu úvodnú stránku a pod.

Literatúra

1. Kula, M.: Metadata embeddings for user and item cold-start recommendations. In: Bogers, T., Koolen, M. (eds.) Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20 2015. CEUR-WS.org (2015), CEUR Workshop Proceedings, vol. 1448, pp. 14–212.
2. Schafer, JB., Konstan, J., Riedl, J.: Recommender systems in e-commerce. In: Proceedings of the 1st ACM Conference on Electronic Commerce, ACM, New York, NY, USA (1999), EC '99, pp. 158–1663.
3. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI (2011)

PodĎakovanie: Tento článok vznikol vďaka čiastočnej podpore projektov APVV-15-0508, VG 1/0667/18, a KEGA 028STU-4/2017.

Annotation:

Personalized recommendation of short-term offers

Personalized recommendation is an essential part of modern Web applications. It is necessary to consider and reflect specific domain characteristics in the recommender design. In this paper we present a short overview of the real-world recommender designed for the e-commerce domain. Proposed recommender is designed to deal with the short-term items availability. It addresses the scalability and high availability.

We use several recommender algorithms to address various goals, e.g., cross-sale. For this purpose a set of default and custom recommenders were proposed.

Benchmark of LFS and other Item-based Recommendation Algorithms on Public Dataset

Michal Laclavik¹ and Marek Ciglan¹

Magnetic Media Online, www.magnetic.com, laclavik@magnetic.com

Abstract. In this paper, we propose a simple non-personalized product recommendation algorithm for e-commerce websites. Algorithm uses purchase history as a training data and outputs product-to-product similarities, which are used for product recommendations on product detail pages. Product similarities are computed based on latent product features computed by Matrix Factorization on user-product matrix or session-product matrix. The recommendations include replacement for viewed products as well as complementary products to this product. Item-based algorithms fall into category of non-personalized product recommendation algorithm, but still are widely used for recommendations on any e-commerce websites. We compare the algorithm with item-KNN, Apriori and POP (top seller) algorithms on publicly available dataset and benchmark them on several important metrics, which need to be taken into account on any e-commerce site to provide useful recommendations.

1 Introduction

Product recommendation is now days essential part of any e-commerce site. Since the Netflix prize 2009¹, personalized recommendation algorithms became popular and delivered value in many areas, where personalized context is a key driver, for example in movie recommendations. The reality of many e-commerce sites is that personalized algorithms like matrix factorization [6] does not work that well for 2 reasons, which we observed also at Magnetic:

- The first is that current session context like product(s) viewed is much more stronger than user history, although both can be combined.
- The second problem is that many e-commerce sites have high number of traffic for unknown users (cold start), and you still want to serve relevant recommendations, so it is more important to focus on current product page context or context of the session.

That is why, there is still need for non-personalized item-based recommendation algorithms, which can deliver relevant recommendations just based on product context. Such algorithms can be used on checkout or add-to-cart page, but mainly on product-detail page, which serves more than 90% of recommendations on typical e-commerce site in user session. At the end, these algorithms

¹ <http://netflixprize.com/index.html>

are usually combined by machine learning algorithms like [8], in order to derive appropriate context and appropriate recommendation for current user session including also personalized user context, but good quality non-personalized recommendations are important input into ML models.

Typical item-based algorithms are: shopping basket analysis algorithms like apriori [1] [3], mahout item similarity², KNN [9] or content based algorithms. Also simple algorithms like new arrivals, top seller in category or random in the category can serve well in a product context, where not enough of purchase or click data is available yet.

The main contributions of the paper are the following:

- We propose a new item-based algorithm, which can be trained on purchase data, and it is able to deliver recommendations in much wider context than typical basket-analysis algorithms or item-KNN.
- We implemented three basic existing algorithms: item-KNN, Apriori and POP in the form of IPython notebook, which is available for further experiments and extension
- We evaluate the algorithms on publicly available data set³, which makes experiments reproducible
- implementation of algorithms and evaluation is available on-line⁴

2 LFS Algorithm

Latent Features Similarity Algorithm (LFS) is based on Matrix Factorization [6], concretely its MLlib implementation⁵, based on ALS for implicit feedback [5]. Similar algorithm to LFS was proposed in [7], however it combine both user and product content, while here we focus on non-personalized algorithms and their benchmark.

In LFS, we compute product latent features by factorizing user-product or session-product matrix. Then we ignore the user (or session) latent vectors and use only product vectors to compute product-to-product similarity by simple dot product on latent vectors. These results into product to product similarity, where we take top-N similar products for each product and serve them as recommendations.

For product-to-product similarity we have used dot product, but we also tried products vectors cosine similarity in the latent space, however results with simple dot product were much better in offline evaluation.

In the paper we trained algorithm only on session-product matrix since user information is not available in the dataset. Anyhow, we tested both user-product and session-product training data on 4 e-commerce websites and in all cases models trained on session-product matrix delivered same or better results.

² <https://mahout.apache.org/users/algorithms/intro-cooccurrence-spark.html>

³ <http://2015.recsyschallenge.com/challenge.html>

⁴ <https://github.com/misos/item-based-recommendation-algorithms-eval>

⁵ <http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html>

We are using 128 latent features. In most of the experiments, more features contributed to better quality performance, but going for more than 100 features gain us just small quality improvement, when compared with the size of the model and time to train the model. So we just did arbitrary decision to use 128 features based on the application needs.

It was interesting to see, how well latent features represented the reality. For example on e-commerce site selling pets related products, it nicely recommended products of different kind grouped together for kitten, puppy or reptile although training data had no information about category. On the home accessories e-commerce site it recommended similar products, where LFS algorithm detected features like product shape, pattern, product material (metal, glass) or also other features like seasoning, where it recommended other fall related products having artifacts like leaves or pumpkin to product with falling leaves artifacts.

3 Evaluation

We evaluate following algorithms:

- proposed LFS algorithm,
- a simple popularity algorithm based on recommending top seller products,
- classical Apriori [1] algorithm and
- item-KNN [9] algorithm.

LFS is implemented as defined in section 2. For the item-KNN algorithm, we are using cosine similarity, since we have just implicit user feedback, browsed products, no ratings. For Apriori, we tuned the algorithm to return best results on the given dataset, where we need to have at least 5 relations between products and the score is higher than 10%. Same score threshold was applied for recommendations from item-KNN.

Data set: we picked publicly available dataset of RecSys Challenge 2015 [2] for evaluation in order to make experiments reproducible. We are also providing ipython notebooks with all the code of creating models and evaluation.

Evaluation protocol: In the RecSys Challenge 2015 [2], algorithms were evaluated on two tasks: predicting of purchase sessions; and predicting purchased items. In our evaluation, we focus on predicting purchased items, but in a bit slightly different way, matching the business needs and metrics of e-commerce clients. In real e-commerce website, you always want to fill recommendations zones with useful recommendations. So if algorithm is not able to recommend anything or not confident enough, you still want to recommend at least top sellers in category, new arrivals or promoted products. This is why, we are also evaluating percentage of served recommendations and server sessions by each algorithm. We evaluated all algorithms with 1, 3, 5 and 10 recommendations per page. In the paper, we provide just results for 5 recommendations per page, because other numbers were directionally very similar.

We have used 90 days of data to train the models and next 14 days to evaluate the model. So we have generated 5 recommendations per viewed product page

for browsing history of next 14 days after training period, and then compared recommended (session, product) pairs with purchased (session, product) pairs in same period. This is obviously a bit simplified evaluation protocol, but evaluating and training data are clearly independent also in the terms of out of time sample.

In Table 1, we can see the results of evaluation. As already mentioned we have compared LFS with other 3 standard algorithms: TS (Top Sellers), Apriori and Item-KNN. TS was able to serve always, but with poor results. It could be a bit better if serving Top Seller in category instead of overall top seller products. Apriori and KNN had lower session conversion rate than LFS, but on the other hand higher recommendation conversion rate (recommended purchases/all recommendations). Still Apriori was able to serve only 9% of all requested recommendations and KNN 38% while LFS could serve 96% of all requested recommendations (5 per visited product detail page). In addition with 5 recommendation on page, LFS was able to recommend 49% of all purchases.

Table 1. Evaluation results.

Algo	CR (sessions)	CR (rec)	served rec	served sessions	rec purchases
LFS	6.70%	0.38%	96%	98%	49%
KNN	4.76%	0.54%	38%	77%	27%
Apriori	2.01%	0.61%	9%	50%	8%
TS	0.18%	0.01%	100%	100%	1%

As we can see, each of evaluated algorithm has advantage at least in one of the used metrics. TS can always fill in requested recommendation slots. KNN can identify decent portion of future purchases as well as it has higher conversion rate per recommendation than LFS. Classical Apriori algorithm on the other hand can serve best recommendation (highest purchase rate per provided recommendations) but only in limited context, but LFS combined most of good properties with winning in conversion rate per session and in recommending most of purchased products.

We have also used LFS similarity score as a feature for Machine Learning model similar to [8], which combines various algorithms and reorder final recommendations to be served. The similarity score between displayed product and to-be-served-recommended-product computed by LFS was one of top 3 features for machine learning model based on Gradient Boosting Decision Tree [4]

Scalability: To conclude the evaluation, we should also mention the scalability of the solution. LFS is trained on purchased history on ApacheSpark MLlib implementation of ALS [5] which is scalable. It scales well on Apache Spark. All the steps of algorithm can be scaled and computed in Apache Spark cluster.

4 Conclusion

Item-based recommendation algorithms provide basic non-personalized product recommendations on e-commerce sites. Nowadays more advanced machine learning algorithms exist, which can be applied and tuned well in many recommendation contexts, but item-based non-personalized algorithms are still valuable input to build better models for typical e-commerce site.

In this paper we implemented and evaluated simple LFS algorithm, which allows to recommend replacement or complementary products for product detail pages of e-commerce sites. Its advantage is that it can serve recommendations to almost all users with good quality of relevant recommendations. The model is trained on purchased data, so the only requirement is to have a purchased history of the product for which complementary or replacement recommendations are served. The LFS performs better than other state-of-the-art item-based algorithms like item-KNN or Apriori.

In this paper we also implemented and evaluated three basic item-based algorithms on publicly available dataset. We hope that provided implementation and code can serve as a benchmark for further evaluation of additional item-based recommendation algorithms

References

1. R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.
2. D. Ben-Shimon, A. Tsikinovsky, M. Friedmann, B. Shapira, L. Rokach, and J. Horerle. Recsys challenge 2015 and the yoochoose dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys '15*, pages 357–358, New York, NY, USA, 2015. ACM.
3. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *ACM SIGMOD Record*, volume 26, pages 255–264. ACM, 1997.
4. J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.
5. Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. IEEE, 2008.
6. Y. Koren, R. Bell, C. Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
7. M. Kula. Metadata embeddings for user and item cold-start recommendations. In *2nd Workshop on New Trends on Content-Based Recommender Systems co-located with RecSys 2015, Vienna, Austria*, volume 1448 of *CEUR Workshop Proceedings*, pages 14–21. CEUR-WS.org, 2015.
8. P. Romov and E. Sokolov. Recsys challenge 2015: Ensemble learning with categorical features. In *Proceedings of the 2015 International ACM Recommender Systems Challenge, RecSys '15 Challenge*, pages 1:1–1:4, New York, NY, USA, 2015. ACM.
9. B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web, WWW '01*, pages 285–295, New York, NY, USA, 2001. ACM.

Project Thoth – A Recommendation Engine for Machine Learning Applications

Fridolín Pokorný, Christoph Görn

Red Hat Czech s.r.o., Purkyňova 99, 612 00 Brno
{fridolin,goern}@redhat.com

Abstract. There has been seen a hype over the past years in AI and machine learning. Machine learning and AI applications are used in production systems which require significant effort to ensure applications behave correctly and are fully operational. In this paper we describe a recommendation system for AI and machine learning applications which use popular open source machine learning libraries. The described system, Thoth, is capable of storing observations based on which it predicts possible misbehavior in application, application assembling errors or issues when integrating application with other components in a deployment.

Keywords: machine learning · artificial intelligence · big data · graph database · data processing · distributed systems · Python

1 Introduction

In this paper we present the designed platform that is a recommendation engine for applications and their dependency stacks. The main purpose of the designed system is to predict misbehavior in applications as well as how to prevent from them. The problematic of an application stack is described with a listing of possible issues that can arise when relying directly or indirectly on a logic that, in the end, create the resulting runnable application compound by libraries.

2 Application Stack

To build complex applications, there is often a need to create small building blocks that encapsulate required logic into an easy to access interface. These small units are often reusable and thus create a separate libraries used by libraries themselves as well as by the created application. The described dependency chaining creates direct or transitive requirements (also know as an application stack) for an application that is handled by a dependency management system in the given language ecosystem. Dependency management system helps to resolve dependencies given the version range specified to meet required functionality offered by required dependencies.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 211-214.*

Given the complexity in applications and their dependencies answering a question which dependencies in which versions are the best possible fit for available hardware in the given runtime environment is not trivial and requires knowledge about the language ecosystem used and behavioral characteristics of the successfully assembled application.

As applications evolve, it is not rare to see a broken application that does not assemble due to changes in libraries used, or simply breaks the expected behavior even without any visible notices. Having properly locked versions of libraries helps to solve issues in most cases, but this approach does not help to keep application always up-to-date with recent fixes in libraries used. An automated system Thoth, presented in the upcoming sections, answers and predicts which libraries in which versions are the most suitable ones for the given requirements saving computational and developer time and, moreover, learns from errors that create patterns already seen in other application stacks.

2.1 Modelling Application Stacks in a Graph Database

Considering dependencies, their relations between each other, the whole application can be modeled in a graph database. We chose JanusGraph graph database to be our data store for storing data in a graph form.

We model two main distinguishable information:

- *Ecosystem data itself* – how the given ecosystem looks like based on a declarative version range specification, what are dependencies between packages and observations like whether the given package is resolvable.
- *Observations* – observations we gathered about packages and application stacks during continuous integration system (CI) or continuous delivery system (CD).

This data model connects data needed for creating recommendations for a user.

2.2 Observations and Predictions

To give recommendations to a user there is a need to aggregate data. Thoth for this purpose aggregates raw data mentioned in the section 2.1. These data create a base for two core types of observations driving recommendations:

- *Package level data* – information gathered solely on package level. An example of such information can be a fact that the given package is installable into the requested runtime environment.
- *Application stack level data* – information about application stack assembling and errors that were found during application assembling in CI or negative feedback in the production environment on deployment or runtime (CD or directly user feedback).

These two fundamental observations are inputs for the scoring function which outputs the best possible application stack with fully locked down versions for deployment or an experimental application stack for gathering additional observations respecting new package releases or new environments.

3 Thoth Architecture and Development

Project Thoth runs on OpenShift to guarantee high horizontal scalability and high availability in case of failures. In some cases, there is a need to directly execute untrusted and potentially malicious code shipped within analyzed Python packages. For this purpose, there is done a restriction using OpenShift's Network Policy mechanism to restrict networking access. Also, OpenShift's namespaces help us with assigning resource quotas for application to restrict from DoS attacks and serve user requests asynchronously, if needed.

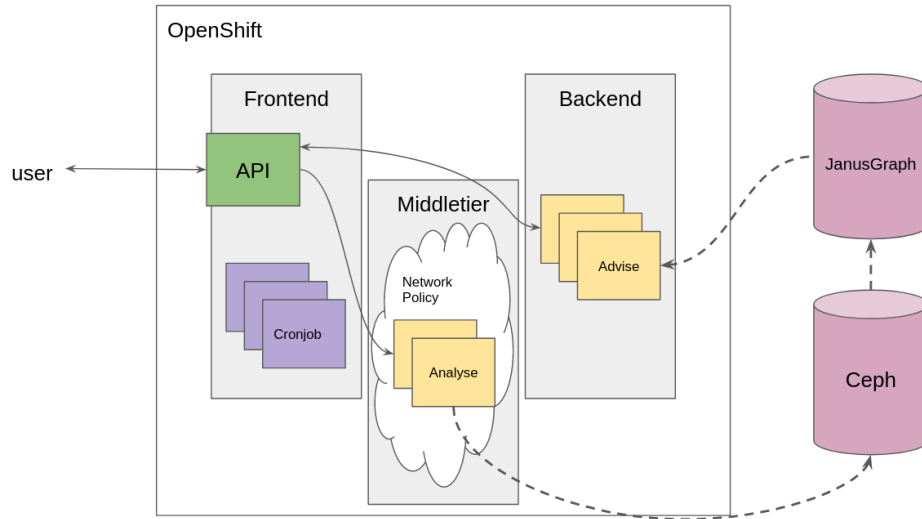


Fig. 1: An architecture of Thoth.

The deployment is separated into three namespaces. In the *Frontend* namespace there are available cronjobs, that perform periodic tasks ensuring the whole system is up-to-date, and a scalable API service that is the main interaction point for a user.

The *Middletier* namespace is the restricted environment where all analyses are performed. Raw results are serialized into a JSON format and stored on an object storage Ceph [3]. One of the cronjobs in the *Frontend* namespace ensures all results are stored properly in the graph database. The database separation is needed due to security (possibly malicious software does not access production

data in the JanusGraph) and to guarantee fully recoverable production graph database.

The last namespace called *Backend* is used to guarantee SLA of the application for serving user requests. There is allocated a resource pool that is assigned for the recommendation engine itself.

3.1 Lab

As discovering new fields and ideas require experiments, project Thoth is designed of loosely coupled packages that, when assembled together, create a ready to deploy Thoth but can also assemble an interactive experimental development environment using JupyterHub and its Jupyter Notebooks.

All experiments can be done in an interactive form that is sharable in the team using Jupyter Notebooks. This allows experiments to be easily accommodated in production code.

3.2 Bots – Junior Developers

To automate every-days tasks, project Thoth comes with bots that ensure the project dependencies and Thoth's components are always up-to-date. If a component in the Thoth's architecture gets updated, a bot automatically updates all the relevant components. Moreover, if there is given a human review, another bot ensures the pull request is merged only if the change does not break existing parts of Thoth (CI/CD checks).

4 Conclusion

Project Thoth is open source and available on GitHub [1]. It is written in the Python programming language to support popular Python machine learning and AI packages out of the box. The deployment is orchestrated by OpenShift [2] to guarantee high horizontal scalability to serve high load of user requests as well as a separation needed for security. The whole system is exposed via an API endpoint where users can get recommendations for their machine learning and AI applications.

References

1. Project Thoth on GitHub, <https://github.com/thoth-station/>. Last accessed 28 May 2018
2. OpenShift: Container Application Platform by Red Hat, <https://www.openshift.com>. Last accessed 28 May 2018
3. Ceph: Homepage <https://www.ceph.com>. Last accessed 28 May 2018

Autoencoders vs. others for anomaly detection

Kristína Miklášová, Ondřej Lomič, Lukáš Císar, Luboš Popelínský, and
Veronika Krejčířová

KD Lab FI MU Brno
{433383, lamanchy, 475964}@mail.muni.cz, popel@fi.muni.cz

Abstract. The paper deals with a task of finding anomalies in the set of pictures using *autoencoders* and comparison of results with other methods searching for outliers, namely *LOF* and *z-score*. Outliers found by these methods are compared to outliers found by project team members. Process consists of preprocessing of pictures using pretrained *deep neural nets* (one at a time), reducing dimension using *PCA*, normalization of features and applying methods on pictures, either on a whole set or subsets divided by classes (dividing the pictures to groups by objects of interest that can be found in them). Output of methods with different attribute settings was compared to outliers found by team members using *confusion matrix* and *F1-score*. The results were not very positive, no significant relationships were found between anomalies found by team members and by anomalies found by individual methods. Possible reasons for this are discussed.

Keywords: Autoencoders · Local Outlier Factor (LOF) · z-score · anomalies · PascalVOC image dataset · feature extraction

1 Introduction

In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset [1]. Anomaly detection is very common problem that occurs in many fields like bank fraud, biomedicine, text processing etc. It can be for example decreased cardiac activity in ECG or a lot of bank transfers per day.

There are many methods to solve this task like density-based techniques (k-nearest neighbors, local outlier factor [3]), one-class support vector machines, cluster analysis, neural networks, Hidden Markov models...

In this paper, we focus on three selected techniques - *autoencoders* [2], *local outlier factor (LOF)* and statistical method *z-score* [4]. Our dataset is subset of PascalVOC image dataset [6] that consists of 4 classes - "Planes", "People", "Planes and People" and "Others". We tried to detect anomalies across whole dataset as well as across classes without "Others" class.

2 Used methods

Autoencoders. An *autoencoder* is an artificial neural network used for unsupervised learning of efficient codings. The aim of an autoencoder is to learn a

representation (encoding) for a set of data, typically for the purpose of dimensionality reduction.

Architecturally, the simplest form of an autoencoder is a feedforward, non-recurrent neural network having an input layer, one or more hidden layers and an output layer that has the same number of nodes as the input layer. The purpose of this type of network is to reconstruct its own inputs (instead of predicting the target value Y given inputs X).

Local Outlier Factor (LOF). *The local outlier factor* is based on a concept of a local density, where locality is given by k -nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers.

The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in *LOF* is an additional measure to produce more stable results within clusters.

Z-score. It is a measure of how many standard deviations below or above the population mean a raw score is. A *z-score* is also known as a standard score and it can be placed on a normal distribution curve. In order to use a *z-score*, you need to know the mean μ and also the population standard deviation σ .

3 Workflow of anomaly detection

3.1 Introduction

At the beginning, we were given a subset of 394 pictures from PascalVOC image dataset, mainly showing people, planes and few other topics like animals, bottles etc. The initial task to perform was to find outliers in the set. First, a criterion was to find outliers on a set as a whole. Next step was to look for outliers in a subsets of original set - in all classes separately.

After each member individually made his findings on anomalies, our results were compared - only 4 pictures were marked by all three members, 16 were marked by two and the majority of anomalies was marked by only one person. This result may be explained by the fact that each member had his/her own criteria for marking a picture anomalous.

3.2 Preprocessing

For this purpose, *Keras* [5] platform using *TensorFlow* backend was used. It is a *Python* library which makes it easy for deep neural networks to be up and running. The pretrained neural network [7] - *vgg16*, *vgg19*, *resnet* and *xception* - was used to extract a representation of original data. An input was the vector of

values where each value represents the color of each pixel in the original image. This input was transformed to the feature vector of lower dimensions which is an output of latter layers of the neural network. This representation gives us more abstract qualities of pictures than only colors of pixels. In case of *vgg16* and *vgg19*, the feature vector has size 512 and in case of two other pretrained nets, the size of feature vector is 2048.

These vectors were transformed by *PCA* [8] into lower dimensions, values 512, 128 or 32 were tried out. Additionally, vectors' components have been normalized into values between 0 and 1. These normalized data as well as unnormalized data were used as an input for our anomalies detection methods with different set ups.

One of the attributes was also a set upon which outliers were being searched for - *classless*, *classbased* or *other*. *Classless* means that dataset for learning consists of all classes - "People", "Planes", "People and Planes", "Others". *Class-based* setting contains only three classes without "Others". In case of *other* option, we mix all images from "Planes" with some pictures from class "Others" (we can decide how many images in percentage we want from others).

3.3 Experimentation with attributes

Preprocessing

To load images, extract feature vector and prepare them as input into algorithms, we could set these parameters:

- *feature_model_name* - model of pretrained neural network,
- *pca.dimensions* - size of vector after *PCA* reduction,
- *normalize* - we could decide whether input will be normalized or not.

4 Evaluation

These results were obtained by comparison with anomalies detected by us, image was marked as an anomaly when at least two people (out of three) marked it as an anomaly. We annotated images as *classbased* data - looking for an anomalies within each class and *classless* data - searching anomalies in whole dataset.

During the research, we have tried hundreds of combinations of parameters in order to understand the results we were receiving. After each run in a specific combination, graphic output was generated presenting the findings.

Combinations of parameters we tried:

- 4 neural networks for feature extraction - *vgg16*, *vgg19*, *xception* **or** *resnet50*
- normalization - *True* **or** *False*
- *PCA* dimensions - 32, 128, 512
- type of anomaly detection made by us - intersection **or** union
- type of anomalies - *classbased*, *classless* **or** *others*
- number of nearest neighbours - 3, 5, 7

- architecture of autoencoder - 1 hidden layer with 3/4 neurons w.r.t. input layer and sigmoid activation function **or** 2 hidden layer (the first layer was the same as in previous case and in the second layer, there were 3/5 neurons w.r.t. input layer and relu function) **or** 3 hidden layers (first two layers as previous ones and the third layer had 2/5 neurons w.r.t. input layer and relu function)
- batch size - 32, 16, 8

We summarized received results in Table 1, where we show values for all compared methods across each model as well as all models together. These results were obtained by comparison with our labeled anomalies. In addition, we compared *autoencoders* results with results obtained from *LOF* method in Table 2.

Results (in %)						
Metric	Method	vgg16	vgg19	resnet	exception	all models
Max accuracy	<i>autoencoders</i>	93,68	95	94	96	96
	<i>LOF</i>	96	94,29	97	94,74	97
	<i>z-score</i>	96	94,74	97	95,24	97
Average accuracy	<i>autoencoders</i>	90,62	90,76	90,51	90,6	90,62
	<i>LOF</i>	89,67	89,82	90,02	89,9	89,85
	<i>z-score</i>	93,62	92,55	90,82	92,55	92,39
Max recall	<i>autoencoders</i>	32	21,33	30	14,67	32
	<i>LOF</i>	70	50	43,33	43,33	70
	<i>z-score</i>	50	50	40	60	60
Average recall	<i>autoencoders</i>	4,87	2,62	6,77	2,16	4,1
	<i>LOF</i>	7,48	5,07	10,76	7,1	7,6
	<i>z-score</i>	4,44	3,61	5,6	8,73	5,6
Max F1 score	<i>autoencoders</i>	0,267	0,221	0,333	0,168	0,333
	<i>LOF</i>	0,667	0,533	0,433	0,433	0,667
	<i>z-score</i>	0,533	0,413	0,333	0,567	0,567
Average F1 score	<i>autoencoders</i>	0,049	0,027	0,062	0,026	0,041
	<i>LOF</i>	0,065	0,055	0,093	0,07	0,07
	<i>z-score</i>	0,05	0,03	0,05	0,084	0,054

Table 1. Results across all models and methods.

Autoencoders vs. LOF method (in %)					
Max accuracy	Avg accuracy	Max recall	Avg recall	Max F1 score	Avg F1 score
100	93,28	32,37	100	1	0,33

Table 2. Autoencoders results compared with results from LOF method

4.1 The most detected images

We grouped detected images by method that were used for anomaly detection. We show only most detected stuff - the count of images is different for each method.

There are some images that all methods detected as an anomaly - airplanes in the sky - as well as pictures that are unique for each method (e.g. a pole vaulter, sheep or a man on the black-and-white photo). It's very interesting that our labeled anomalies has no intersection with results obtained by methods.

4.2 The best configuration

We tried several combinations of parameters and determined best configurations in term of *accuracy* and *F1 score*.

The best configuration for autoencoders				
Metrics	Batch size	Model	PCA dimen.	Layers
Max acc	8	xception	128	[0,75; 'sigmoid'], [0,6; 'relu'], [0,4; 'relu']
Max F1	8	resnet	32	[0,75; 'sigmoid'], [0,6; 'relu'], [0,4; 'relu']

Table 3. The best configurations with max accuracy and F1 score for autoencoder

The best configuration for LOF method				
Metrics	Batch size	Model	PCA dimen.	Number of neighbours
Max accuracy	16	resnet	32	5
Max F1 score	8	vgg16	32	5

Table 4. The best configuration for LOF method

The best configuration for z-score			
Metrics	Batch size	Model	PCA dimen.
Max acc	16	resnet	128
Max F1	16	xception	32

Table 5. The best setting of parameters for z-score

4.3 Classless and classbased results

Now, we try to present significant configurations of parameters with a brief discussion about the results. The discussion is grouped by a type of pictures as input - starting with *classless* (a whole set), followed by *classbased*.

Two example configurations are presented here, difference is in a number of *PCA* dimensions and a set of humans found outliers results were compared to.

First, normalized data with 512 dimensions *PCA* on a model *vgg16* with intersection of people found anomalies configuration and a batch size of 16 is presented. Here, as a rule in *classless* settings, no (or 1 to 4) anomalies were agreed on by people and selected methods. But interestingly, anomalies which were agreed on by 2 or more methods visually all come to the same category of planes with a simple, mostly single color background.

A second example of *classless* configuration has a batch size of 16, *PCA* dimension is 128 and a set of human found anomalies consists of their union.

To continue, *classbased* outliers are going to be inspected. In this section, let us start with a class "Planes". Here, two configurations will be mentioned. In the first one, non-normalized data with 32 dimensions *PCA* on a model *vgg16* with intersection of people found anomalies configuration and a batch size of 16 is was used.

What we find interesting is a difference in found anomalies in these different settings. But as in a great majority of cases, no common outliers were found by humans and the methods.

To move on, let us see a *classbased* anomalies search in classes "Planes", "People" and both. Here, one output in a non-normalized data with 32 dimensions *PCA* on a model *vgg16* with intersection of people found anomalies configuration and a batch size of 16 configuration is presented. Of all the outliers, only one was agreed on both by humans and the methods. Interestingly, the same configuration on a normalized data returned methods-found outliers consisting purely by planes on a single color background.

5 Conclusion

The lack of common anomalies seems to lie in a different qualities of pictures team members and methods looked at. A proposed topic for further research would be an iterative model where methods would ask if anomalies found by them are really anomalies and adjust their results accordingly. Also, a possible reason for a bad performance may be in a relative small number of training examples.

References

1. Anomaly detection, https://en.wikipedia.org/wiki/Anomaly_detection.
2. Autoencoders, <http://ufldl.stanford.edu/tutorial/unsupervised/Autoencoders/>.
3. Local Outlier Factor, <http://www.dbs.ifi.lmu.de/Publicationen/Papers/LOF.pdf>.
4. z-score, <http://www.statisticshowto.com/probability-and-statistics/z-score/>.
5. Keras documentation, <https://keras.io/>.
6. PascalVOC image dataset, <http://host.robots.ox.ac.uk/pascal/VOC/>.
7. Pretrained neural networks, <https://github.com/tensorflow/models/tree/master/research/slim#Pretrained>.
8. PCA, <ftp://statgen.ncsu.edu/pub/thorne/molevovclass/AtchleyOct19.pdf>.

Acknowledgements. This work has been supported by Faculty of Informatics, Masaryk University Brno.

FireProt: web server for automated design of thermostable proteins

Miloš Musil¹²³, Jan Štourač¹², Jaroslav Bendl¹²³, Jan Brezovský¹², Zbyněk Prokop¹², Jaroslav Zendulka³, Tomáš Martínek³, David Bednář¹², and Jiří Damborský¹²

¹ Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University

Kamenice 5, Bld. A13, 625 00 Brno, Czech Republic

imusilm@fit.vutbr.cz

² International Centre for Clinical Research, St. Anne's University Hospital Brno Pekařská 53, 656 91 Brno, Czech republic

³ Department of Information Systems, Faculty of Information Technology, Brno University of Technology Božetěchova 2, 612,66 Brno

Abstract. There is a continuous interest in increasing proteins stability to enhance their usability in numerous biomedical and biotechnological applications. Only a single-point mutations with a small impact on protein stability are typically predicted with the existing tools, while more complex multiple-point mutants are necessary for useful stabilization. Here, we present FireProt, a web server for automated design of multiple-point thermostable mutants that combines structural and evolutionary information in its calculation core. The server is complemented with interactive, easy-to-use interface that allows users to directly analyze and optionally modify designed thermostable mutants.

Keywords: protein stability · thermostable proteins · hybrid method · protein engineering

1 Introduction

Proteins are widely used in numerous biomedical and biotechnological applications. However, naturally occurring proteins cannot usually withstand the harsh industrial environment, since they are mostly evolved to function at mild conditions. Therefore, there is a continuous interest in increasing proteins stability to enhance their industrial usability. A number of in silico tools for the prediction of the effect of mutations on proteins have been recently developed to improve their stability, thus surpassing their natural limitations. However, due to the potentially antagonistic effect of mutations, only single-point mutations are usually predicted in silico and have to be followed by laborious and costly protein expression, purification and characterization. Single-point mutations typically

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 221-224.*

enhance the melting temperature of target proteins only by units of degree. A much higher degree of stabilization can be achieved by constructing multiple-point mutants. Unfortunately, the automated design of multiple-point mutants is uneasy, due to the existence of the potential antagonistic effects of individual mutations. To address these issues, FireProt web was developed [1].

2 Dataset

One of the most crucial properties of every predictive tool is the construction of the reliable training dataset. Our dataset was mainly obtained from the ProTherm database [2]. ProTherm is the most extensive collection of thermodynamic parameters such as Gibbs free energy, heat capacity or enthalpy, measured as a difference between wild-type and mutant protein.

ProTherm was lastly updated in February 2013 and up to date it contains nearly 26,000 entries of both single- and multiple-point mutants designed over 740 unique proteins. However, even though ProTherm database is the most common source of data used for the training of the existing tools, in its current state it suffers from several serious issues such as missing values, opposite signs or various inconsistencies in measures. To deal with this issues, ProTherm had to be manually curated and all the redundancies were discarded during the process of the final dataset construction. Furthermore, we have also omitted mutations with the measured change of Gibbs free energy from the interval of $(-0.5, 0.5)$ since the average experimental error of measurement is about 0.48 kcal/mol [3].

Our training dataset was later expanded with experimental data obtained from the HotMuSiC dataset [4] and the final set consists of 1,573 mutations (349 stabilizing, 1,224 destabilizing) designed over 103 unique proteins. Included proteins were completed with corresponding 3D structure in the highest quality available (based on resolution and R-factor).

3 FireProt method

The web server integrates sixteen computational tools and utilizes both sequence and structural information in the process. At first, potentially stabilizing single-point mutations are identified via two separate branches: one relying on the estimation of the change of free energy upon mutation and second utilizing back-to-consensus approach. The energy-based approach is employing FoldX [5] and Rosetta [6] tools to estimate free energy via the statistical and physical force fields. Furthermore, preceding filters accelerate the calculation by omitting the high risk residues – ones that are located in the highly conserved or correlated regions of the protein. It was observed that functional and structural constraints in proteins generally lead to the conservation of amino acid residues and similarly, correlated residues ordinarily help to maintain protein function, folding and stability. Mutations conducted on these positions are therefore considered unsafe and excluded from the design of stable proteins. The second approach is based on the information obtained from multiple sequence alignment. The

most common amino acid in each position of protein sequence often provides a non-negligible effect on protein stability. Thus, FireProt implements majority and frequency ratio approach to identify mutations at positions where wild-type amino acid differs from the most prevalent one.

In total, three protein designs are provided by FireProt strategy. The first design includes only the mutations from energy-based approach, the second contains the mutations suggested by the evolution-based approach and the third is the combination of both. Naturally, because of potentially antagonistic effect between individual mutations, these mutations cannot be combined blindly. To avoid possible clashes, FireProt strategy is trying to minimize antagonistic effects by utilizing Rosetta. The physical force field is used to evaluate all pairs of single-point mutations within the range of 10 Å and subsequently the mutations are introduced into the stable protein via the edge- based graph algorithm. In this way, the colliding mutations are excluded from protein and algorithm stops once there are no mutations left or the stabilizing effect of analyzed pair of mutations drops below defined threshold. Detail workflow of the FireProt web can be found in the Figure

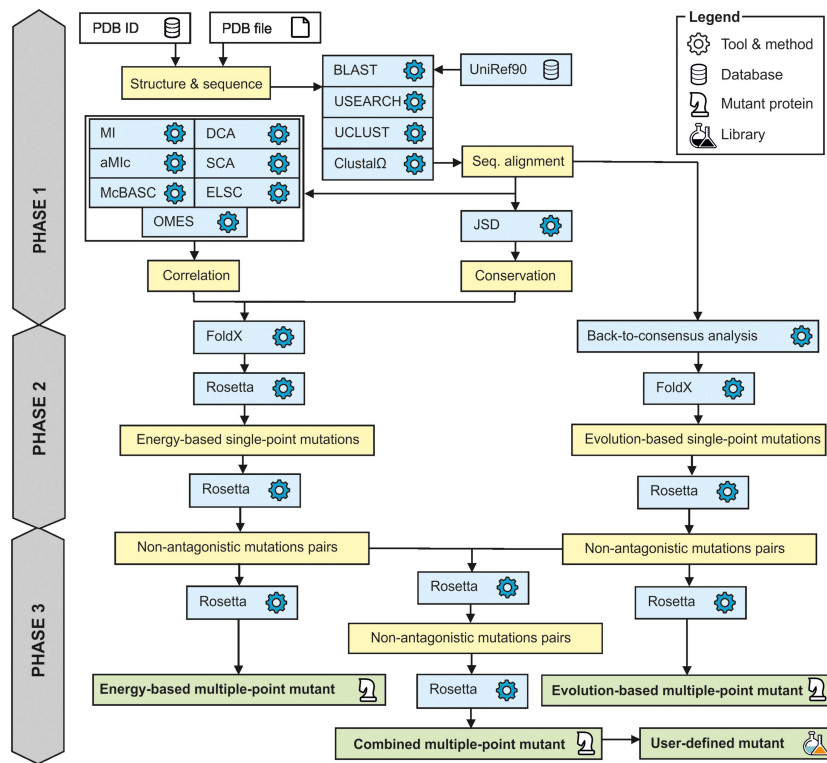


Fig. 1. Complete workflow of the FireProt web.

A high time demands of Rosetta analysis were one of the most excruciating issues with the FireProt protocol. Even with the application of preceding filters over 100 mutations is usually left for precise, but slow, Rosetta calculations. For this reason, several force fields and protocols were evaluated on the dataset of 1,573 mutations. Based on the results of these evaluations, the best trade-off between the time requirements and precision was selected. Moreover, the FireProt time demands were drastically reduced by massive parallelization via the use of MetaCentrum. These measures have lowered the calculation time from several weeks to a single day.

4 Conclusions

FireProt is a web server that provides users with a one-stop-shop solution for the design of thermostable multiple-point mutants. It combines energy- and evolution-based information in its calculation core and all the protocols were highly optimized to increase the calculation speed, while maintaining the prediction accuracy. The server is complemented by an easy-to-use graphical interface that allows users to interactively analyze individual mutations selected as a part of energy- of evolution-based approach together with the ability to design their own multiple-point mutants on top of our robust strategy. The web server is freely available at <https://loschmidt.chemi.muni.cz/fireprot/>.

References

1. Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., Martinek, T., Bednar, D., Damborsky, J.: FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Research* **45**(W1), W393–399 (2017)
2. Bava, K.A., Gromiha, M.M., Uedaira, H., Kitajima, K., Sarai, A.: ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Research* **32**, D120–121 (2004)
3. Khatum, J., Khare, S.D., Dokholyan, N.V.: Can contact potentials reliably predict stability of proteins? *Journal of Molecular Biology* **366**(5), 1223–1238 (2004)
4. Pucci, F., Bourgeas, R., Rooman, M.: Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Scientific Reports* **6**, 23257 (2016)
5. Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *Journal of Molecular Biology* **320**, 369–387 (2002)
6. Kellogg, E.H., Leaver-Fay, A., Baker, D.: Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins* **79**, 830–838 (2011)

Inteligentná optimalizácia pokrytia a efektívnosti heterogénnych sietí využitím strojového učenia bez učiteľa

Eugen Šlapak¹, Marcel Vološin¹ a Juraj Gazda¹

Katedra počítačov a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice
`eugen.slapak@student.tuke.sk`

Abstrakt. V tejto práci prezentujeme rámec pre zlepšenie kvality topológie Heterogénnej siete (HetNet), pomocou online mini-batch K-Means algoritmu, pre zhlukovanie, ktorý patrí do skupiny algoritmov strojového učenia bez učiteľa.

Modelovanie topológie HetNet-u sa často spolieha na bodové procesy stochastickej geometrie, ako napríklad Poisson point process (PPP) alebo Binomial point process (BPP) bez zahrnutia a priori znalostí o lokáciách používateľov. Náš navrhovaný prístup naopak využíva dostupnú informáciu o vzoroch mobility používateľov. Za účelom napodobenia komplexnej mobility dynamických používateľov, charakterizujeme mobilitu používateľov pomocou dvoch odlišných vzorov: deterministickým modelom založeným na Bézierových krivkách a stochastickým Lévy flight modelom mobility. Výsledky simulácií potvrdili naše očakávania a ukázali, že topológia HetNet-u vytvorená navrhovanou metódou poskytuje vyššiu priepustnosť v porovnaní s BPP prístupom.

Kľúčové slová: HetNet · dizajn topológie · mini-batch K-Means · strojové učenie

1 Úvod

Heterogénne siete (HetNet-y) predstavujú sľubný prístup, pre zlepšenie ukazovateľov kvality siete, v porovnaní s konvenčnými jednostupňovými bezdrôtovými celulárnymi sieťami. HetNet-y sú však citlivé na interferenciu, ktorá by sa mohla výrazne zvýšiť za predpokladu, že by operátori a jednotlivci rozmiestňovali bázové stanice s malým rozsahom (BSMR) nekoordinovaným spôsobom. Problém interferencie vyplývajúcej z rozmiestnenia veľkého množstva BSMR zostáva dôležitým výskumným problémom.

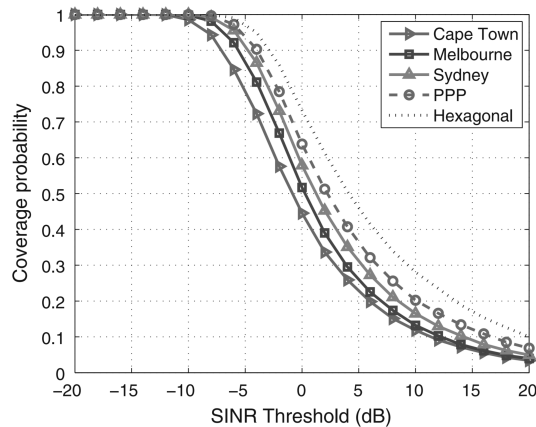
Prístup bežne používaný pre modelovanie HetNet-ov sa spolieha hlavne na aplikáciu stochastickej geometrie. Stochastický bodový proces bežne používaný pre modelovanie rozmiestnenia buniek siete - Poissonov bodový proces (Poisson point process, PPP), poskytuje adekvátnu až mierne nadhodnotenú aproximáciu pravdepodobnosti pokrytia pre rozmiestnenie buniek v rôznych mestách, táto skutočnosť je ilustrovaná na Obr. 1.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 225-230.*

Klasický hexagonálny model poskytuje ešte výraznejšie nadhodnotenú aproximáciu, ktorá vo väčšine prípadov porovnania s charakteristikami skutočných bezdrôtových celulárnych sietí predstavuje hornú hranicu pravdepodobnosti pokrytia (Obr. 1).

Stochastická geometria vo všeobecnosti neodráža hustotu rozmiestnenia jednotlivých koncových zariadení (KZ). Naším zámerom je využiť vzory lokácií a mobility používateľov, pre vylepšenie kvality poskytovaných telekomunikačných služieb rozmiestnením BSMR do lokácií s vyššou pravdepodobnosťou výskytu KZ. Intuitívnym riešením, ktoré nám umožní zahrnúť túto informáciu do algoritmu rozmiestnenia prvkov celulárnej siete je metóda mini-batch K-Means zhľukovania.

V tejto práci posúdime kvalitu výsledkov mini-batch K-Means zhľukovacieho algoritmu, ktorý bude použitý pre optimalizáciu rozmiestnenia BSMR na základe dát o vzoroch mobility používateľov. Ako metódu s ktorou budeme takéto rozmiestňovanie pomocou zhľukovania porovnávať, budeme uvažovať bežne používaný BPP prístup zo stochastickej geometrie. Môžeme konštatovať, že navrhovaná metóda prekonáva BPP prístup naprieč všetkými simulačnými parametrami a poskytuje výrazné zlepšenie kvalitatívnych metrik bezdrôtovej siete.



Obr. 1. Obr. 1. Pravdepodobnosti pokrytia merané v rôznych mestách porovnané s pravdepodobnosťami pokrytia pre PPP a Hexagonálny model rozmiestňovania básových staníc.[1] .

2 Systémový model

HetNet je modelovaný ako dvojstupňová sieť s básovými stanicami veľkého rozsahu BSVR ktoré majú predurčenú lokáciu, ktorá sa počas behu simulácie

nemení, keďže sa predpokladá, že operátor sa bude chcieť vyhnúť nákladom spojeným so zmenou lokácií BSVR a ich sieťovej infraštruktúry. BSMR sú zo začiatku rozmiestnené stochasticky pomocou BPP, neskôr sú z tohto počiatočného stavu rozmiestňované pomocou K-Means algoritmu. Všetky antény báзовých staníc ($BS = BSVR + BSMR$) sú v simulácii považované za izotropické, teda ich direktivita je rovná jednej.

Pre výpočet útlmu signálu jeho šírením bol použitý Hata-Okumura model. Interferencia, ktorá je v simulácii braná do úvahy, zahŕňa len interferenciu medzi BS pracujúcimi na rovnakej frekvencii, zatiaľ čo frekvencie používané BSVR a BSMR sa líšia. Termálny šum, ktorý je počítaný ako $k \cdot T \cdot B_{FB} = N_0 \cdot B_{FB}$ je tiež v rámci simulácie braný do úvahy, kde k je Boltzmannova konštanta, $N_0 = k \cdot T$ reprezentuje spektrálnu hustotu šumu, B_{FB} je šírka pásma jedného frekvenčného bloku (FB) (v našej simulácii berieme do úvahy šírku pásma 180kHz) a T je termodynamická teplota.

Každá BS má limitovaný počet FB, ktoré sú distribuované každému KZ v každom simulačnom kroku, keďže v kratších časových úsekoch sa prirodzene nevyskytujú žiadne udalosti, ktoré by si vyžadovali prerozdelenie FB medzi používateľmi. Distribúcia FB jednotlivým KZ je riadená round-robin algoritmom, čím je zabezpečené relatívne férové rozdelenie FB medzi nich.

Tiež v simulácii uvažujeme typický downlink scenár LTE systému, bez agregácie bezdrôtového pripojenia. Taktiež, každý koncový používateľ používa v každom okamihu maximálne jedno bezdrôtové zariadenie.

Všetky parametre simulácie sú uvedené v tabuľke Tab. 1.

3 Navrhovaný prístup

Ako vhodná metóda pre úlohu rozmiestnenia BSMR bol zvolený mini-batch K-Means algoritmus, keďže minimalizuje sumu štvorcov vzdialeností všetkých vzoriek, k centroidom, ktoré sa k nim nachádzajú najbližšie. Formálnejšie povedané, minimalizuje hodnotu funkcie zobrazenej v rovnici (1) [2].

Kroky K-Means algoritmu:

1. Priradenie každej vzorky práve jednému zhľuku, s minimálnym štvorcom Euklidovskej vzdialenosti k stredu zhľuku.
2. Výpočet nových stredov zhľukov, ktoré sa stanú centroidmi týchto zhľukov. Stredy zhľukov mohli, ale nemuseli zmeniť svoje lokácie v predchádzajúcom kroku algoritmu, keďže je možné, že žiadna vzorka nebola pridaná ku zhľuku alebo odstránená zo zhľuku.

Centroidy postupne konvergujú k fixným lokáciám:

$$J(C) = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - \mu_k\|^2, \quad (1)$$

kde $C = \{c_k\}$, $k = 1..K$ je množina všetkých zhľukov, $X = \{x_i\}$, $i = 1..n$ je množina všetkých tréningových vzoriek a μ_k je stred zhľuku c_k .

Tab. 1. Parametre systémového modelu

Parameter	Value
Metóda simulácie	Monte Carlo po 100 pokusov na jeden počet BS
Dĺžka jednej simulácie	800 hours
Simulovaná plocha	1km^2
Vysielací výkon BS	BSVR: 40.3 W BSMR: 6.3 W
Výška umiestnenia antény BS	BSVR: 10 m BSMR: 2 m
Typ antény	0 dBi (izotropická)
Nosná frekvencia	BSVR: 2.1 GHz BSMR: 3.5 GHz
Počet FB	každý BSVR: 100 FB každý BSMR: 20 FB
Šírka pásma FB (B_{FB})	180 kHz
Systémová šírka pásma	BSVR: 18 MHz BSMR: 3.6 MHz
Model útlmu signálu	Hata
Termodynamická teplota	290 K
Počet Lévy Flight KZ	100
Počet HS-ov	10, každý tvorený 10-timi KZ,
Celkový počet KZ	200

V našom prípade vzorky reprezentujú lokácie jednotlivých KZ, v jednotlivých krokoch simulácie, a lokácie BSMR sú určené lokáciami jednotlivých centroidov. Priemerná vzdialenosť používateľa k ľubovoľnej BSMR dodatočne priemerovaná cez čas, je tak minimalizovaná. Minimalizácia tejto priemernej vzdialenosti používateľov k BSMR, prirodzene pomáha pri zvýšení priemernej priepustnosti poskytovanej KZ. Mini-batch varianta tohto K-Means zhlučovacieho algoritmu [3] bola pre našu implementáciu zvolená z toho dôvodu, že prináša výpočtové zrýchlenie voči klasickému K-Means pre veľké dátové sady a umožňuje online pridávanie tréningových vzoriek. V našej implementácii je v každom kroku simulácie K-Means zhlučovaciemu algoritmu dodávaná jedna časť údajov o lokáciách používateľov (pre daný simulovaný čas, reprezentovaný, daným simulačným krokom).

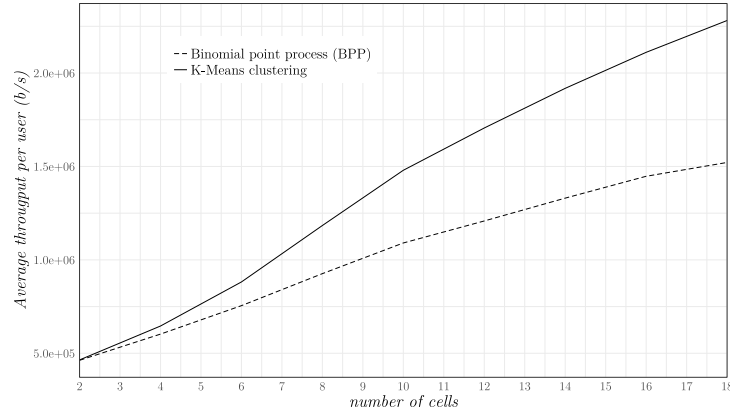
4 Výsledky

Kvalitatívne výsledky rozmiestnenia BSMR pomocou BPP, aj pomocou mini-batch K-Means algoritmu trénovaného na 800 simulovaných hodinách bežiacej simulácie sú kvantifikované priemernou priepustnosťou pre jedno KZ. Maximálna teoretická priepustnosť pre každé KZ je vypočítaná pomocou Shannon-Hartleyho teórie ako:

$$C = B_{FB} \log_2 \left(1 + \frac{S}{I+N} \right), \quad (2)$$

kde B_{FB} označuje šírku pásma jedného FB a $\frac{S}{I+N}$ pomer signálu k sume interferencie a šumu.

Výsledky merania priemernej priepustnosti pre rozličné počty BS sú vizualizované na Obr. 2.



Obr. 2. Porovnanie priemernej priepustnosti na KZ merané pre rôzne počty BS.

Priepustnosť, pre dve BS pre BPP aj mini-batch K-Means scenár rozmiestnenia sa nelíši, pretože sú v tomto prípade rozmiestňované len dve BSVR na vopred určené statické lokácie, bez toho, aby boli v jednom, alebo druhom scenári prítomné BSMR. S rastúcim počtom BSMR postupne rastie rozdiel v priepustnosti medzi BPP a mini-batch K-Means zhľukovaním na hodnoty až 0.75 mbits, tento rozdiel teda dosahuje 66 % (v prípade rozmiestňovania 18-tich BS).

5 Záver

Mini-batch K-Means zhľukovací algoritmus poskytoval vyššie priemerné priepustnosti v porovnaní s BPP rozmiestňovaním pre všetky skúmané parametre.

tre simulácie. Aj napriek tomu, že mini-batch K-Means algoritmus negaran-
tuje optimálnosť nájdeného rozmiestnenia BSMR, je tento nedostatok kom-
penzovaný vyššou výpočtovou efektívnosťou v porovnaní s plne vyčerpávajúcím
prehľadávaním celého priestoru možných riešení, pre ktoré rastie výpočtová
zložitosť prudko exponenciálne s lineárne narastajúcou veľkosťou problému. V na-
sledujúcom výskume, sa budeme zameriavať jednak na optimalizáciu rozmiestne-
nia BSMR reflektujúceho vzory mobility pre KZ a zároveň kontrolu vysielacieho
výkonu, ktorá vytvára zákaly tzv. *zelených komunikačných sietí*.

Odkazy

- [1] C.-H. Lee, C.-Y. Shih a Y.-S. Chen, “Stochastic Geometry Based Models for Modeling Cellular Networks in Urban Areas”, *Wirel. Netw.*, roč. 19, č. 6, s. 1063–1072, srp. 2013, ISSN: 1022-0038. DOI: 10.1007/s11276-012-0518-0. WWW: <http://dx.doi.org/10.1007/s11276-012-0518-0>.
- [2] A. K. Jain, “Data Clustering: 50 Years Beyond K-means”, *Machine Learning and Knowledge Discovery in Databases Lecture Notes in Computer Science*, s. 3–4, DOI: 10.1007/978-3-540-87479-9_3.
- [3] D. Sculley, “Web-scale K-means Clustering”, in *Proceedings of the 19th International Conference on World Wide Web*, ř. WWW ’10, Raleigh, North Carolina, USA: ACM, 2010, s. 1177–1178, ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772862. WWW: <http://doi.acm.org/10.1145/1772690.1772862>.

PodĎakovanie: Táto práca bola podporená slovenskou Agentúrou na podporu výskumu a vývoja, číslo projektu APVV-15-0055.

Annotation:

Intelligent optimization of coverage and efficiency of heterogeneous networks using unsupervised machine learning

Method of intelligent small cell placement enabling cellular network providers to place small cells without human supervision of an expert, to achieve improvements of qualitative characteristics of the network, is presented in this work.

For the purpose of testing various approaches to base station placement, simulation of HetNet and user movement in 2D space was created. Deterministic trajectories are placed in an intentionally simple symmetric pattern, in order for less difficult interpretation of placement patterns provided by machine learning algorithm.

Presented placement optimisation method allows non exhaustive search of candidates for base station placement, reducing computational complexity in exchange for possibility that suboptimal solution (but one that may be still satisfying operator’s requirements), instead of globally optimal one, is found.

Dynamic Social Networks and Their Analysis

Iveta Mrázová¹ and Peter Zvirinský²

¹ Faculty of Mathematics and Physics,
Department of Theoretical Computer Science and Mathematical Logic
Charles University, Prague, Czech Republic
`iveta.mrazova@mff.cuni.cz`

² Faculty of Mathematics and Physics,
Department of Theoretical Computer Science and Mathematical Logic
Charles University, Prague, Czech Republic
`peter.zvirinsky@gmail.com`

Abstract. Social networks are usually represented as (static) graphs; their vertices correspond to individuals and edges represent their mutual relationships. With the aim to study effectively also their dynamic versions, we will discuss here various tools developed for social network analysis based on their performance on big graph data evolving in time, support for parallel computations, algorithms they provide and their flexibility for the user.

Keywords: social network analysis, big data, graph data mining

1 Introduction

Complex relationships among individuals and/or their groups in popular large-scale social networks like Facebook, Twitter, or LinkedIn can be described as graphs — the vertices correspond to the individuals and the edges represent their mutual relationships. Most social network analyses performed so far were, however, restricted to static social networks that do not evolve over time.

This assumption is often not met for many social networks studied nowadays with the aim to understand various social phenomena, their emergence and impact in a time-evolving setting. The goal of this work is thus to specify the notion of a dynamic social network and to summarize requirements to be put on its analysis. Based on the review of existing graph mining tools, we will design the main structure of a new system for the analysis of dynamic social networks.

2 Related Works

Social networks are used to describe complex relationships between the considered subjects (people, companies, or other entities in the network). Recent rise in the availability of network-structured data, mostly from social networks (e.g., Facebook, Twitter, LinkedIn), but also from other sources, has created an increased demand for processing large graphs. Popular social networks furthermore tend to change their character, size and structure over time.

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 231-236.*

Most of the existing graph mining tools applicable to social network analysis are oriented towards static graphs and can be divided into two categories: user friendly libraries designed for single machine computation and large-scale distributed graph processing systems. The single machine systems usually provide a large collection of common network analysis algorithms – cf., e.g., igraph [2], NetworkX [4], or SNAP [8]. Unfortunately, they often do not scale well to large graphs or the chosen graph abstraction does not directly support parallel computations and each algorithm must be parallelized separately [8].

Distributed systems like Pregel [9], PowerGraph [3], or GraphX [10], can scale to extremely large graphs easily. Yet, they provide only a handful of common graph algorithms and therefore cannot be used for most network analyses usually performed. The main reason for this limit is the rather restrictive idea of the gather-apply-scatter (GAS) computational model they use. Although the GAS model is well suited for some algorithms such as PageRank, HITS or connected components, its usability for other types of algorithms remains limited [7].

All of the above-discussed tools understand the evolution of a dynamic social network as a series of snapshots and perform the analysis of each snapshot separately. This approach is obviously tedious and requires to implement the necessary time dependent logic over and over again from scratch. To overcome these disadvantages, a few frameworks oriented towards time-evolving graphs have been proposed recently, e.g., Chronos [5], Kineograph [1], or GraphTau [6]. Their main focus is, however, put on the distributed setting and high performance rather than flexibility and implementation of a sufficiently large collection of network mining algorithms.

3 Dynamic Social Networks

Informally, dynamic social networks can be understood as graphs that can change their structure over time, i.e., edges and vertices can be added and/or removed from the graph at different times. Our point of view is edge-centric, meaning that the time information is attached only to the edges (as a list of activity relevant time spans). At a given time point, the vertices thus exist only if they are connected to an active edge. Internally the input graphs correspond to a (static) graph with specific attributes assigned to the edges that indicate when the given edge was actively in use.

Definition 1. A (static) social network is a directed graph $G = (V, E)$, where V is the set of vertices and $E = \{(v_k, v_l) | v_k, v_l \in V\}$ is the set of directed edges.

Further, we will assume to see social networks across a series of discrete time points $t_1, \dots, t_{n-1}, t_n$. At each time point t_i ; $1 \leq i \leq n$, we observe an instance of a (static) social network $G_i = (V_i, E_i)$. Obvious changes that may occur between two time-points t_{i-1} and t_i ; $1 < i \leq n$, thus include the addition or deletion of edges, and the appearance or disappearance of vertices.

Definition 2. Let n be the number of the considered time points and let us observe an instance of the (static) social network $G_i = (V_i, E_i)$ at each i . A

dynamic social network is a directed graph $G_D^n = (V_D^n, E_D^n)$, where

- $V_D^n = \bigcup_{i=1}^n V_i$ is a set of vertices that have appeared at any of the considered time points $t_i; 1 \leq i \leq n$, and
- $E_D^n = \bigcup_{i=1}^n E_i$ is a set of edges that have appeared at any of the considered time points $t_i; 1 \leq i \leq n$, and all such edges are labeled by a sequence of ordered pairs $[i, j]$ that specify the time points at that the edge existed in the network, i.e., from the time point t_i to the time point t_j in this case. The time interval $[t_i, t_i]$ can be abbreviated as $[t_i]$.

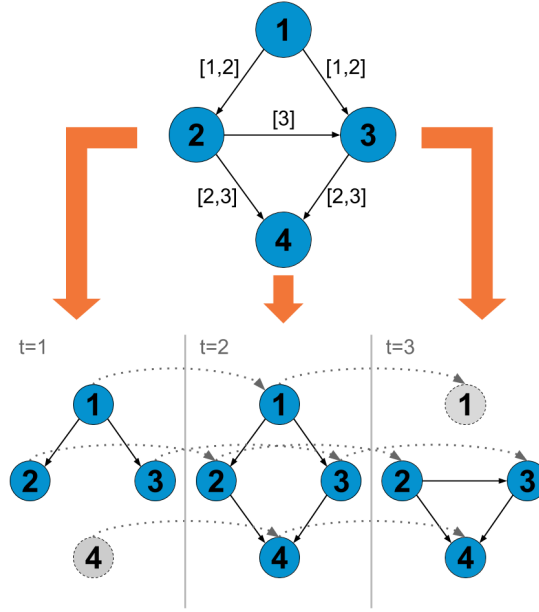


Figure 1. Unfolding of a dynamic social network into an expanded graph consisting of 3 time slices t_1, t_2 and t_3 . A copy of each active vertex and edge is created for each time slice. Additional meta edges connect the copies of the same vertices across subsequent time slices to support information flow both within the network and across time.

To address the network dynamics across time, this graph can be unfolded into the so-called expanded graph. The expanded graph shall provide a transparent view of the network at each of the considered time points. Further, we will denote it as a slice. Expanded graphs thus incorporate multiple graphs that are mutually inter-connected based on the user-defined slicing function (e.g., hourly, daily, etc.). Each slice contains a copy of active edges and vertices from the original graph. To support information flow also between the vertices/edges from different time slices, new "meta" level edges are added between the vertices from different time periods. Fig. 1 illustrates the unfolding process.

Definition 3. An expanded graph of a dynamic social network $G_D^n = (V_D^n, E_D^n)$ is a directed graph $G_E^n = (V_E^n, E_E^n)$, where

- $V_E^n = \{v_k^i | v_k \in V_D^n \wedge 1 \leq i \leq n\}$ is the set of vertices and
- $E_E^n = \{(v_k^{i'}, v_l^{i'}) | i \leq i' \leq j \text{ for any time interval } [i, j] \text{ labeling } (v_k, v_l) \in E_D^n\} \cup \{(v_k^i, v_k^{i+1}) | v_k^i \in V_D^n \wedge 1 \leq i < n\}$ is the set of directed edges.

4 GraphSlices — Requirements and System Architecture

The concept of dynamic social networks obviously places new requirements on the graph mining tools. In particular, these requirements include extensibility and ease of use, support for transparent time-dependent operations, a variety of algorithms provided for the analysis of social networks and rich support for high performance computing solutions. Below, we will outline the overall structure of our new framework GraphSlices designed with the aim to meet the aforementioned requirements. Yet to support network analysis both across time and various other data dimensions that might be considered by the user later on, the number of dimensions remains free. The actual implementation of the proposed framework consists of the following main components (cf. Fig. 2):

- **Graph** — this component provides the graph interface with two basic implementations, a serial and a parallel one. The interface provides access to high level operations over the graphs including the attributes associated with the vertices or edges. Each vertex and edge is identified by a unique ID and is wrapped into a corresponding **Vertex** or **Edge** class. Graph computations are performed by chaining multiple operations on the graph and every operation returns a new transformed version of the original graph.
- **Algorithms** — a module containing various network analysis algorithms. So far, PageRank, HITS, coloring and label propagation have been implemented. Other techniques are foreseen, e.g., community detection and information flow algorithms as well as new approaches based on advanced AI and machine learning principles capable of revealing emerging structures and their evolution within the network. An example of a simple end2end use-case for calculating the evolving PageRank is shown in Listing 1.1
- **Builders** — a component with the utilities to load/save read-only copies of the graphs from/to the disk
- **Generators** — a module containing functions to generate different classes of random graphs to study and test various graph properties

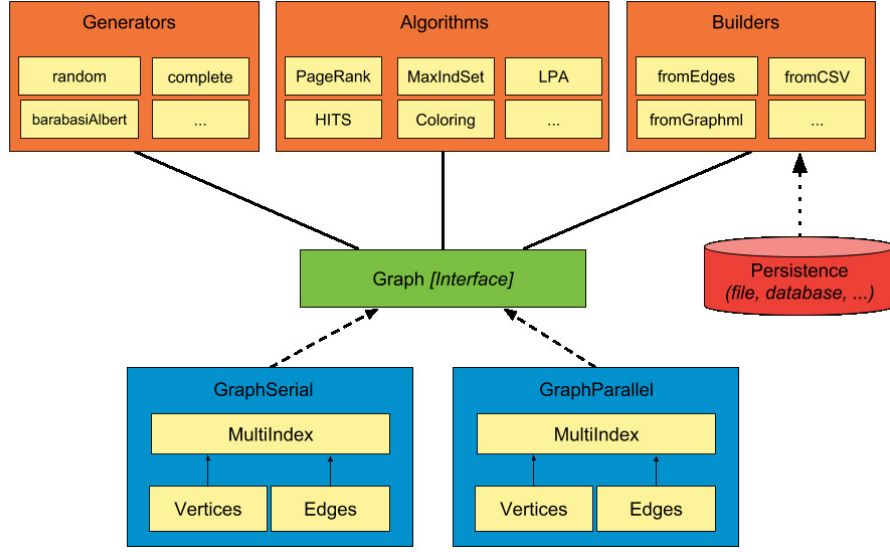


Figure 2. Architecture of the proposed framework GraphSlices: At the center lies the Graph interface which is responsible for all graph manipulation operations. The two basic implementations supported by this interface are the serial single threaded implementation and the parallel multi-threaded implementation. The rest of the code is organized into 3 modules: Generators with various methods to construct synthetic graphs, Algorithms that contain a suite of techniques implemented for social network analysis, and Builders that load the graphs to/from the memory.

```

val graph = Builders.loadFromFile("...")

val slicedGraph = graph.pushDimension(
  edge => edge.data.map(time => (time, 1.0))
)

val prGraph = Algorithms.evolvePageRank(slicedGraph)

```

Listing 1.1. An example for the dynamic PageRank: Using the example from Fig. 1, we first load the (dynamic) graph structure, vertex and edge attributes from a file. Afterwards, assuming the time indicators have the form of a list, we create an expanded form of the network graph and compute the evolving PageRank.

The proposed framework follows functional programming paradigms, therefore we have decided to implement it in Scala. Scala data structures are optimized for sequential processing and all data resides in memory. In addition to GAS, our abstraction of graph data thus provides also access to vertex/edge iterators that allow very fast sequential processing and support the implementation of a wide variety of graph algorithms. The dynamics of the graphs is handled by means of a lazily initialized index that allows the graphs to be effectively “sliced” and

processed based on time or any other dimension. Scala further supports efficient graph manipulation and powerful parallelization options. GraphSlices has been developed for single big-memory multiple-cores machines, and as such provides a useful trade-off between high performance, flexibility and ease of use.

5 Conclusions

The concept of dynamic social networks represents a novel approach to the analysis of network structured data that evolves over time. This paper provides a survey of latest graph mining tools, together with an assessment of their advantages and disadvantages. To facilitate the otherwise cumbersome analysis of evolving network models, we have specified the notion of a dynamic social network and summarized a set of requirements to be addressed by the supporting software frameworks.

Using the insight obtained while analyzing previous approaches to social network analysis, we have designed the main structure and functionality of a new graph processing framework called GraphSlices. Its goal is to support high-level operations designed to manipulate and analyze dynamic social networks. It shall support multiple computational models, including GAS, and it shall be equipped with a sufficiently large collection of algorithms applicable to the solution of various practical network analysis tasks. Our further research shall include also thorough performance tests of the implemented software.

Acknowledgements: This work was supported by project GA UK No. 120616.

References

1. Cheng, R. et al.: Kineograph: Taking the pulse of a fast-changing and connected world. In: Proc. of EuroSys'12 (2012) 85–98.
2. Csardi, G. and Nepusz, T.: The igraph software package for complex network research. *Int. Journal, Complex Systems*, 1695 (2006).
3. Gonzalez, J.E. et al.: Powergraph: Distributed graphparallel computation on natural graphs. In: Proc. of the 10th USENIX Symposium of OSDI'12 (2012) 17–30.
4. Hagberg, A.A., Schult, D.A. and Swart, P. J.: Exploring network structure, dynamics, and function using NetworkX. In: Proc. of SciPy'2008 (2008) 11–15.
5. Han, W. et al.: Chronos: A graph engine for temporal graph analysis. In: Proc. of EuroSys'14 (2014) 1:1–1:14.
6. Iyer, A.P., Li, L.E., Das, T. and Stoica, I.: Time-evolving graph processing at scale. In: Proc. of GRADES '16 (2016) 5:1–5:6.
7. Kalavri, V., Vlassov, V. and Haridi, S.: High-level programming abstractions for distributed graph processing. *IEEE TKDE* 30(2) (2016) 305–324.
8. Leskovec, J. and Sosič, R.: Snap: A general-purpose network analysis and graph-mining library. *ACM Trans. Intell. Syst. Technol.* 8(1) (2016) 1:1–1:20.
9. Malewicz, G. et al.: Pregel: A system for large-scale graph processing. In: Proc. of SIGMOD'10 (2010) 135–146.
10. Xin, R.S., Gonzalez, J.E., Franklin, M.J. and Stoica, I.: Graphx: A resilient distributed graph system on spark. In: Proc. of GRADES'13 (2013) 2:1–2:6.

Siamese Convolutional Neural Networks for Recognizing Partial Entailment

Martin Vít¹

NLP Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 60200 Brno
info@martinvita.eu

Abstract. Recognizing textual entailment (RTE), i. e., a decision problem whether a sentence (called hypothesis) can be inferred from a given text, became a well established and widely studied task. As a consequence of the traditional binary (or ternary) class formulation, it is not possible to express the fact that a fragment of the hypothesis is entailed by the text, even though the “whole” entailment of the hypothesis from the text does not hold. The notions of partial textual entailment – and faceted entailment in particular – address this problem. In this paper, we introduce a siamese CNN architecture with a static attention mechanism together with a sentence compression and provide an evaluation over modified SemEval 2013 Task 8 dataset.

Keywords: Partial Textual Entailment · Convolutional Neural Networks · Siamese Architectures

1 Introduction

Recognizing textual entailment (RTE) task attracts attention of NLP community for years. It is usually formalized as a classification problem – to classify “text-hypothesis pairs” into two (Entailment/Non Entailment) or three (Entailment/Non Entailment/Neutral) classes. Nevertheless, these task formulations cannot properly model a situation when only a fragment of hypothesis H is entailed by the given text T (but the whole H not) and/or a fact that H is *almost entailed* by T . In real-world applications, there is a demand for such a graded approach. Recognizing partial entailment has many potential applications, including computing semantic textual similarity [5], detecting plagiarism, multi-document summarization [2] etc. Recently, there exist many approaches to RTE problem but, in contrast, the task of recognizing faceted textual entailment is highly neglected.

The ideas of partial textual entailment and faceted entailment were introduced in [6] in order to provide an automatic evaluation functionality for intelligent tutoring systems – roughly said, the goal is to recognize whether the student’s answer is at least partially correct, i. e., the correct answer to a given question is at least partially entailed by the student’s answer. This task was later elaborated by Omer Levy et al. [4]. According to them, we say that an ordered

pair $(T; H)$ forms a *partial textual entailment* if a fragment (component) of the hypothesis H is entailed by T . In [6], the model of *facets* has been defined: a *facet* is an ordered pair of words contained in the hypothesis accompanied by a direct semantic relation between these two words (or multi-word expressions). In RTE-8 challenge [2] as well as in this paper, a simplified facet model is used: the explicitly expressed semantic relation is omitted, hence the facet consists of only a pair of words (w_1, w_2) from the hypothesis. For instance, if the hypothesis is a sentence “*The main job of muscles is to move bones.*”, then $(\text{muscles}, \text{move})$ is an example of one possible facet. The *agent role* of “muscles” in this case is not provided.

Recognizing faceted entailment – task definition: Given (i) a text T , (ii) a hypothesis H , and (iii) a facet (w_1, w_2) , i.e., an ordered pair of terms in H , the task consists of determining whether T expresses, either directly or by implication, the same relationship between the facet words w_1, w_2 as in H . If so, the output label is **Expressed**, otherwise **Unaddressed**, [2].

Let us provide an example used also in [4]: **T:** Muscles generate movement in the body. **H:** The main job of muscles is to move bones. The facet (muscles, move) in context of H is expressed by T , but, in contrast the facet (move, bones) in context of H is unaddressed by T .

Currently, there is only one known system for recognizing faceted entailment – the system of Levy et al., [4] that is based on adaptations of three different RTE approaches/modules (none of them uses neural networks).

2 The Core of Our Approach – Siamese CNNs

Deep learning has meant a revolution in various fields of NLP. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs, especially LSTMs) are the main architectures used in various NLP tasks including machine translation, answer selection or textual entailment.

Recognizing faceted textual entailment task deals with three entities as inputs: *text*, *hypothesis*, *facet*. The underlying idea of our approach can be briefly described as follows: both text and hypothesis are represented using convolutional neural networks (input $\rightarrow$ convolutions with different filter widths $\rightarrow$ pooling layers) that share corresponding weights, i. e. they are two instances of the same network: one for representation of T , the second one for representation of H . These two vector representations are then concatenated together with the representation of the facet and over this representations, several densely connected layers are applied. The last (output) layer is then a single sigmoid neuron, since we have a binary decision problem. (An architecture of this kind for “classical” RTE is depicted in [9] as BCNN.)

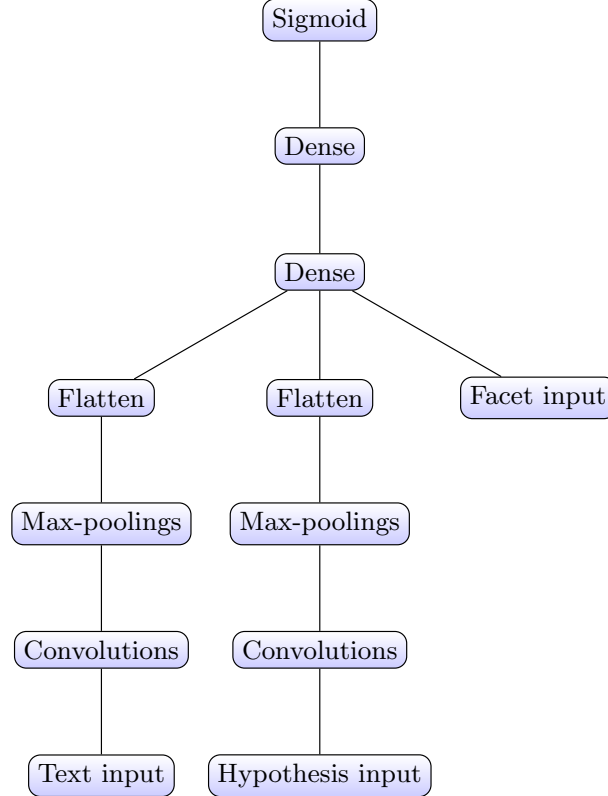
Sentences (text, hypothesis) are sequences of words – let us assume they have the same length l . Each word w is represented by its embedding $e(w)$ in the d_0 -dimensional real space. For the purpose of this work, we use GloVe embeddings [7] with dimension $d_0 = 50$. This leads to a real matrix representation of dimension $(d_0 \times l)$ for each sentence – two main input layers. We have also an

auxiliary input layer for facets. These are represented by d_0 dimensional vectors that are *centroids* of facet word embeddings. Since the sentences have different lengths, l is the maximal sentence length and sentences of length lower than l are padded with zero vectors.

After the input layers, there are simultaneously several convolution layers: 6 with filter width 2, 6 with filter width 3, and 2 with filter width 4, these are connected to max pooling layers with widths 2, 3, 4 respectively. The number of filters and their widths were determined by exhaustive search. To avoid overfitting, L_2 regularization was used.

As already mentioned, these two instances of the same CNNs (i. e., with shared weights and other parameters) provide text/hypothesis vector representations. This architecture is typically called a “siamese architecture”. These two representations as well as the facet representation become an input to two densely connected layers both with 4 ReLU units. Final decision about faceted entailment is done by a single sigmoid unit, see Figure 1. We have ~ 3000 trainable parameters in total. The training was performed in 128 epochs using RMSprop.

Fig. 1. Overall architecture



2.1 Improvements: static attention and sentence compression

Our improvements are based on the assumption that different words in the text/hypothesis have different importance from the faceted entailment viewpoint: words in the text/hypothesis that are similar to those in the facet, should be adjusted in the sentence representation and, moreover, words with low similarity should be removed.

The similarity on the word level is computed using function:

$$\rho(x, y) = \frac{1}{1 + |x - y|},$$

where $|\cdot|$ stands for the standard Euclidean distance, x and y are vectors of dimension d_0 , i. e. word vectors. We use Euclidean distance instead of a “more standard” cosine similarity according to investigations presented in [10].

For each word w_i in the text/hypothesis ($1 \leq i \leq l$), let λ_i be its weight w.r.t. the facet (v_1, v_2) defined by:

$$\lambda_i(w_i) = \max\{\rho(e(w_i), e(v_j)) \mid j = 1, 2\}.$$

Note that words v_1, v_2 in the hypothesis forming the facet, have their corresponding values of

$$\lambda(v_1) = \lambda(v_2) = 1.$$

Static attention-based CNN for recognizing faceted entailment will have formally the same architecture as in the previous case, but instead of feeding the input layer by the concatenated column vectors $e(w_1), e(w_2), \dots, e(w_l)$ for the sentence $w_1, w_2, \dots, w_l$, we will use the concatenation of columns

$$\lambda_1(w_1) \cdot e(w_1), \lambda_2(w_2) \cdot e(w_2), \dots, \lambda_l(w_l) \cdot e(w_l),$$

i. e., embedding vectors are multiplied by scalars – the weights with respect to a given facet (in the implementation, this is done by a point-wise tensor multiplication layer).

We use rather the word “static” together with “attention”, to distinguish the approaches like [10] (and attention mechanisms used in the field of image processing), since there are no weights to be learned during the training in order to specify which words/input areas are more important than the others – this is done before the learning process in a fixed way.

The *static attention* is modeled by multiplying word embedding vectors in the text/hypothesis representation by scalars (weights expressing the word similarity “facet word–word in the text/hypothesis”). *Sentence simplification* is a process of altering sentence such that it becomes simpler according to syntactic or lexical criteria [3]. Sentence simplification can be used in a preprocessing pipeline in a variety of tasks including text summarization, sentence fusion or machine translation. One of the prior approaches to sentence simplification is based on sentence compression – i. e., word deletion in particular. In our approach, we drop out words that have a low similarity to those contained in facets. This process removes some n -grams, but also introduces new ones that were not present

in the original sentence and that were obtained by skipping a deleted word (that will be later covered by a convolutional filter of width n). In parallel, some n -grams that contained irrelevant words were dismissed.

3 Evaluation and Results

Nowadays, there is only a limited amount of training/test data for recognizing faceted entailment task. The only one “official” annotated corpus for this purpose is known, it is a subset of SciEntsBank that was used in SemEval 2013 RTE-8 challenge [1]. For our purposes we have modified this corpus by adding a part of test data to the training dataset in order to have a bigger training set – since the original training set contained less items than the test one (!): 13145 (train) vs. 16263 (test). This was done to increase the amount of training data that deep learning models usually require).

Our training set includes 27408 items, the test set 2000 (randomly selected from the original SemEval test set).

We achieved the precision **0.845** and recall 0.832 when employing mentioned improvements, and precision **0.786** and recall 0.779 without them.

4 Conclusion and Further Work

We have proposed a siamese attention-based convolutional CNNs that works over compressed sentences (text, hypothesis). As far we know, it is a first deep learning approach applied to this task. We have demonstrated that the static attention mechanism provides a significant added value in the task of recognizing faceted entailment. Proposing advanced architectures for this task belongs to further work as well as their sound evaluation: the further work will include experiments and evaluations of other deep neural network architectures, mainly (bi-)LSTM based ones, [8]. The other way of investigations will be focused on the impact of different initial sentence representations, such that dimensionality of word embedding used, different types of them, e.g. GloVe, word2vec, freetext etc., as well as the impact of different sentence length in the sentence compression step. The results will be covered by a paper in preparation.

References

1. Bentivogli, L., Dagan, I., Magnini, B.: The recognizing textual entailment challenges: Datasets and methodologies. In: Handbook of Linguistic Annotation, pp. 1119–1147. Springer (2017)
2. Dzikovska, M.O., Nielsen, R.D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., Clark, P., Dagan, I., Dang, H.T.: Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Tech. rep., NORTH TEXAS STATE UNIV DENTON (2013)
3. Hasler, E., de Gispert, A., Stahlberg, F., Waite, A., Byrne, B.: Source sentence simplification for statistical machine translation. *Computer Speech & Language* **45**, 221–235 (2017)

4. Levy, O., Zesch, T., Dagan, I., Gurevych, I.: Recognizing partial textual entailment. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 451–455 (2013)
5. Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., Agirre, E.: Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems* **119**, 186–199 (2017)
6. Nielsen, R.D., Ward, W., Martin, J.H.: Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering* **15**(4), 479–501 (2009)
7. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
8. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of cnn and rnn for natural language processing. arXiv preprint arXiv:1702.01923 (2017)
9. Yin, W., Schütze, H., Xiang, B., Zhou, B.: Abcnn: Attention-based convolutional neural network for modeling sentence pairs. arXiv preprint arXiv:1512.05193 (2015)
10. Yin, W., Schütze, H., Xiang, B., Zhou, B.: Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association of Computational Linguistics* **4**(1), 259–272 (2016)

Datová sada pro detekci dezinformačního obsahu - případová studie Novičok v Česku

Martin Řimnáč

Ústav informatiky AV ČR, v.v.i
Pod Vodárenskou věží 2, 182 07 Praha 8
`rimnacm@cs.cas.cz`

Abstrakt. Publikování dezinformací na webu hraje stále větší roli, proto vyvstává otázka, jak takovému obsahu čelit a nebo na jeho potenciální závadnost alespoň upozornit. Propaganda využívá dezinformací k relativizaci skutečností, jejichž popis se snaží většinou nepřímou zpochybnit. Příspěvek formou případové studie v konkrétní kauze formálně popisuje výroky prezentované v článcích publikovaných na webu a to včetně účelu jejich publikování; všímá si některých zajímavých aspektů prezentovaných dezinformací a hledá model pro jejich popis. Cílem příspěvku je informovat o vznikající datové sadě a ilustrovat základní použité dezinformační techniky včetně důsledků jejich použití.

Klíčová slova: Dezinformace · Web · Pravděpodobnost · Entropie

1 Úvod

Dnešní web umožňuje publikovat příspěvek, ať již formou diskuzí, blogů či jiných médií téměř každému. Se zvyšující se jednoduchostí publikace názoru a značnou mírou i rychlosti dosahu je možné takových příspěvků využívat k šíření propagandy, neboli *k působení na publikum formou úmyslných formulací myšlenek, postojů chování s cílem dosáhnout reakce v souladu s úmysly a potřebami propagandisty* [4]. Propaganda se snaží relativizovat skutečnost generováním (souboru) hypotéz zamlžující realitu, u některých typů dat i algoritmicky [6] pomocí specializovaných podobnostních měr.

Propaganda je šířena prostřednictvím dezinformací, *nepravdivých nebo zavádějících informací, jejichž publikací autor ovlivňuje nebo manipuluje příjemce informace* [4,5]. Naopak *nepravdivé informace, které nemají cíl manipulovat a vznikají pouze z nepochopení podstaty informace*, označujeme jako fámy, popřípadě misinformace. Obsah, který obsahuje dezinformace, pak nazýváme *fake news* [4], je-li obsahem celý článek, pak *hoax* [5]. Úspěšným označíme takový hoax, který zůstane zveřejněn po dlouhou dobu (provádí-li se korekce), dostane se mu mediální pozornosti a přiláká mnoho uživatelů [5].

Konvenční média se snaží eliminovat prezentování dezinformací a hoaxů, ať již formou přímé lingvistické analýzy obsahu [3] nebo nepřímo na základě vlastností obsahu jako je délka textu, formátování či počet externích odkazů [5]. Eliminace (zavádějícího, ale zákonem nezapovězeného) obsahu však naráží na etické problémy spojené se svobodou slova.

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 243-248.*

2 Formalismus pro analýzu prezentovaných výroků

V rámci navrženého formalismu výroky mohou podporovat, oslabovat či rozporovat jiné výroky, mohou však i napadat či velebit aktéry. Jak výroky, tak aktéři mohou mít hierarchické uspořádání. Pro lepší srozumitelnost formalismus umožňuje stanovit, že daný výrok je reakcí na některý z předchozích výroků nebo že pro svou platnost využívá jiného výroku; rovněž zaznamenává vágnost výroku (nejpravděpodobnější, prý) nebo použití nějaké techniky (například ironie).

2.1 Datová sada Novičok v Česku

Jako pilotní projekt pro analýzu publikovaných výroků byla vybrána kauza *otravy dvojitého agenta Sergeje Skripala a jeho dcery Julie armádním nervově paralytickým jedem A234, který patří do skupiny látek neoficiálně označovaných jako novičok vyvinutých v 70. až 80. letech v Sovětském svazu v rámci programu Foliant*, mezi analyzovaná mainstreamová média byly vybrány portály *aktualne.cz*, *e15.cz*, *echo24.cz*, *idnes.cz* a *lidovky.cz*; z alternativních (podle [1]) médií pak *ac24.cz*, *aeronet.cz*, *cz.sputniknews.com*, *parlamentnilisty.cz* a *prvnizpravy.cz*. Relevance publikovaných článků byla každodenně aktualizována pomocí služby Google Common Search. Relevantní články byly manuálně analyzovány a jejich obsah extrahován do výroků, které byly v rámci formalismu otagovány.

V datové sadě se mainstreamová média snaží o vyvážený přístup včetně citace postojů jednotlivých stran, typicky *Velká Británie z útoku na Skripalovi viní Moskvu, která to ale popírá*¹, ke konci článků bývá uváděno shrnutí stávajícího stavu. Tento perex je uváděn i u článků typu rozhovor, kdy článek prezentuje stanovisko jediného aktéra. Počet prezentovaných výroků bývá řádově nižší desítky. Naopak alternativní média v datové sadě velmi často přebírají části z jiných médií a podstatně častěji prezentují postoj jedné strany; vykazují podstatně vyšší podíl formulací, které se snaží relativizovat výroky. Alternativní média, podobně jako diskuzní fóra, vedle relativizace používají i ironie a osobních útoků, počet výroků bývá oproti mainstreamovým podstatně nižší (běžně menší než 5) a nebo naopak velmi vysoký.

2.2 Ukázky relativizace výroků z datové sad

Jako příklad relativizace uveďme výrok Velké Británie, že *původ látky použité k otravě Skripalových je v Rusku*² podpořený *místem vývoje novičoků v bývalém*

¹ například https://www.lidovky.cz/otravne-latky-se-v-cesku-vyrabeji-jen-kvuli-testovani-vlastnosti-tvrdi-odbornik-g13-/zpravy-domov.aspx?c=A180317_185626_ln_domov_ele

² například https://www.lidovky.cz/novicok-pribeh-desiveho-smrticiho-jedu-o-nemz-rusko-od-pocatku-lhalo-liw-/zpravy-svet.aspx?c=A180318_144805_ln_zahranici_mpr nebo https://technet.idnes.cz/sergej-skripal-novicok-fsb-putin-kreml-mi6-oleg-penkovskij-gru-pbl-/vojenstvi.aspx?c=A180331_103433_vojenstvi_kuz

SSSR³, detekcí otravné látky u Skripalových jako novičok A234⁴ a současnou vnitropolitickou situací v Rusku; výrok podporuje odpovědnost Ruska za útok. Rusko zareagovalo sadou postupně uvolňovaných výroků. Nejprve oznámilo, že *nejpravděpodobnějším zdrojem smrtící látky typu Novičok je podle ruské diplomacie Česká republika, Slovensko, Británie nebo Švédsko. Jmenované země podle mluvčí od konce 90. let provádějí intenzivní výzkum látek z projektu Novičok*⁵, posléze dementovalo *Na území Sovětského svazu či v dobách jeho existence ani v éře Ruské federace nebo na jejím území se nikdy neprováděly výzkumy, jejichž skutečné nebo kódovací jméno bylo Novičok*.⁶ Nelze si nepovšimnout rozporu mezi výroky (prezentovanými též den ministerstvem zahraničí Ruské Federace), zatímco druhý popírá existenci novičoku (avšak jako programu, nikoliv jako látky), první udává zdroj látky novičok (ve smyslu látky i programu) mimo Rusko s apelem na vývoj (dle dalších publikovaných výroků se *intenzivní výzkum látek provádí za účelem protichemické obrany*⁷, na čemž se shodují obě strany).

Mainstreamová média publikují výrok o *mikrosyntéze, která není podle mezinárodní úmluvy považována za výrobu*⁸, se kterým však nejen alternativní média, ale i rozhovory v mainstreamových médiích pak laborují ve smyslu *Rusové nás poměrně od samého začátku dávali do spojitosti jenom jako zemi, která umí Novičok vyrobit, o čemž není důvod pochybovat, že umí*⁹ stírající rozdíl mezi mikrosyntézou a výrobou pro bojové použití. Po eskalování diplomatického sporu pak Ruská strana doplňuje, že *Prahu nikdo v ničem neobviňoval a neobviňuje. Na rozdíl od britských kolegů žádná obvinění nevznášíme. Mluvíme jen o tom, že dokonce i v mediálním prostoru – a není to propaganda, nejsou to ruská, ale česká média – existuje velké množství materiálů, potvrzujících potenciál Česka v oblasti chemických výzkumů. To je prostě třeba brát v úvahu*¹⁰. Ukázka ilustruje posun sémantiky, který relativizuje připouštěním hypotetické možnost schopnosti, být v jiných ohledech rozporuplné.

³ například https://www.lidovky.cz/jed-novicok-byl-nejmocnejsi-chemickou-zbrani-sovetsti-vedci-za-nej-dostali-leninovu-cenu-g74-/zpravy-svet.aspx?c=A180313_084853_ln_zahranici_ele

⁴ například https://www.lidovky.cz/novicok-pribeh-desiveho-smrticiho-jedu-o-nemz-rusko-od-pocatku-lhalo-liw-/zpravy-svet.aspx?c=A180318_144805_ln_zahranici_mpr

⁵ například https://www.lidovky.cz/kreml-reaguje-na-zemanovotvrzeni-o-novicoku-je-to-pry-dukaz-ze-mayova-lze-15v-/zpravy-domov.aspx?c=A180504_082512_ln_domov_lar

⁶ https://www.lidovky.cz/cesko-slovensko-britanie-nebo-svedsko-jsou-moznymi-puvodci-latky-novicok-tvrdi-moskva-gta-/zpravy-svet.aspx?c=A180317_151552_ln_zahranici_ele

⁷ například https://zpravy.idnes.cz/ministr-zahranici-martin-stropnický-komentar-velvyslanec-novicok-sergej-skripal-g2m-/domaci.aspx?c=A180321_125640_domaci_jn

⁸ https://zpravy.idnes.cz/mimoradna-schuze-snemovny-novicok-dtb-/domaci.aspx?c=A180530_212939_domaci_kop

⁹ například subportál u lidovky.cz, http://neviditelnypes.lidovky.cz/rozhovor-rusko-skripala-neotravilo-d78-/p_zahranici.aspx?c=A180527_172153_p_zahranici_wag

¹⁰ například <https://echo24.cz/a/SduKV/moskva-opet-ukazala-na-prahu-novicok-ma-vyznamne-misto-v-ceskem-vyzkumu>

Sémantický posun lze v datové sadě ilustrovat i u mainstreamových médií nebo důvěryhodných organizací; zatímco původně látka A230 (která byla v Česku (mikro)syntetizována, odlišná od látky A234 použité k otravě Skripalových) nebyla označena jako novičok (údajně zpráva BIS, SÚJB), postupně si toto označení získávala (Vojenský výzkumný ústav, později Vojenské zpravodajství, prezidentská kancelář, následně většina médií).

Podobně je to s množstvím syntetizované látky a jejím limitem dle úmluvy o nešíření chemických zbraní, některá alternativní média mluví o zákazu jakéhokoliv množství¹¹, řádově mikrogramech (u popisu mikrosyntézy; např. ministerstvo obrany)¹², řádově o gramy¹³, do 10 kg¹⁴ až po 1 tunu (alternativní média¹⁵).

Cíl relativizace je většinou omezen na kontext článku nebo časové okno, zatímco koncem března platilo *Prezident republiky Miloš Zeman dosáhl dalšího velkého vítězství na poli diplomacie. ... bylo dohodnuto, že ... ČR nevyhostí 30 ruských diplomatů, ale jenom 3 ruské "pracovníky" z ruské ambasády*¹⁶, v polovině května stejný autor na stejném portálu předmětné vítězství označuje jako *Další útok proti Miloši Zemanovi přišel v kauze Skripal, kdy ministr zahraničí za Hnutí ANO Martin Stropnický vyhostil z ČR tři ruské diplomaty*¹⁷. Zatímco první výrok lichotí prezidentovi, pozdější útočí na představitele vlády a lichotí proruskému postoji. Podobné chování vykazuje výrok *Rusom, ktorí tvrdia, že by na základe analýzy dokázali nájsť výrobcu*¹⁸ reflektující Britové odmítají Rusku

¹¹ <https://aeronet.cz/news/novicok-nikdy-neexistoval-ve-skutecnosti-slo-o-noviczech-vtipkuj-ruska-media-a-nejsou-daleko-od-pravdy-base-line-npl-a-230-se-totiz-pry-v-cesku-nevyrabela-ale-jenom-testovala/>

¹² například <https://echo24.cz/a/SAkpU/novicok-se-v-cr-vyrabel-v-mikroskopickem-mnozstvi-uvadi-ministerstvo> nebo https://www.lidovky.cz/drabova-latky-zvane-novicok-se-v-cesku-nevyrabely-syntetizoval-se-mohly-mikrogramy-a-230-nebyla-v-i2x-/zpravy-domov.aspx?c=A180504_103412_ln_domov_gib

¹³ například <https://nazory.aktualne.cz/komentare/pr-bublina-slechtova-sejmula-pana-novicoka/r7e0cbf325a8011e8b8310cc47ab5f122/>

¹⁴ <https://aeronet.cz/news/novicok-nikdy-neexistoval-ve-skutecnosti-slo-o-noviczech-vtipkuj-ruska-media-a-nejsou-daleko-od-pravdy-base-line-npl-a-230-se-totiz-pry-v-cesku-nevyrabela-ale-jenom-testovala/>, ruský velvyslanec podle <https://echo24.cz/a/Svnev/rusove-couvaji-cesko-jsme-nikdy-neobvinili-jen-jsme-konstatovali-fakta>

¹⁵ <https://www.parlamentnilisty.cz/arena/monitor/Docent-ke-kauze-novicok-Tutlalo-se-neco-hodne-svinskeho-Proto-informovane-osoby-zacaly-tak-radit-proti-Zemanovi-534799>

¹⁶ <https://aeronet.cz/news/zdroj-z-okoli-hradu-chteli-po-nas-vyhostit-30-ruskych-diplomatu-nakonec-ale-tu-nulu-za-trojkou-museli-skrtnout-dalsi-politicke-vitezstvi-prezidenta-milose-zemana-neocon-se-nazral/>

¹⁷ <https://aeronet.cz/news/uz-vime-proc-cssd-pozadovala-kreslo-ministra-vnitro-na-vnitru-bude-sedet-bakaluv-clovek-z-americkeho-aspen-institute-a-pomalou-se-zacina-rysovat-proc-babis-neposlechl-milose-zemana/>

¹⁸ <http://www.parlamentnilisty.sk/arena/politologovia/Tereza-Spencerova-Novick-Pomaly-bezna-vec-ma-to-vslikto-Putin-po-veceroch-nad-cmudiacimi-kadickami-vyvijsa-nove-zbrane-Tie-dristy-a-taraniny-o-Rusku-299029>

poskytnout tento vzorek¹⁹, jenž přestal být používán po té, co *Podle britských i ruských expertů výrobce ani není možné zjistit*²⁰.

Relativizovat lze i všeobecně přijímaný výrok; zatímco informaci, že *Julija Skripalová v pondělí opustila nemocnici*²¹ je publikována všemi sledovanými zdroji, o den později je výrok upřesněn, že se tak tomu stalo *odpoledne*²², za cca měsíc později pak, že *byla minulý měsíc potají a uprostřed noci propuštěna z léčby*²³. Rozporuplnost v takovém případě přináší pouze část výroku.

2.3 Budoucí strojové zpracování

Datová sada je předurčena pro další výzkum v následujících oblastech

- Pravděpodobnostní přístup - Výše uvedené scénáře mají společný efekt, kterým je zvýšení binární entropie [2] platnosti výroku. Publikovaná relativizace (jak je patrné většinou nepřímou) oslabuje původní výrok, zvyšuje tak v člancích relativní četnost výroků opačného významu a tím zvyšuje entropii. Cíle relativizace jsou většinou krátkodobé, zatímco cíle propagandy dlouhodobější, neboť bývají spojeny s postojem autora k aktérům spojených s výrokiem. I tuto vazbu opět lze kvantifikovat pomocí relativních četností.
- Hledání vzorů scénářů - Expresivita datové sady umožňuje definování vzorů scénářů a jejich následné automatické hledání. Tímto způsobem lze detekovat nejen protistrany a jejich vzájemné konfrontace, ale i rozpory napříč jednotlivými články včetně časových konsekvencí. Cílem detekce není vybrání jediné správné odpovědi (faktu), ale upozornit na fakta, která si protirečí, která nejsou potvrzena jinými zdroji nebo autory.
- Ontologie vlastností - Formalismus je definován pomocí jednoduchého ontologického popisu. Ten může být oproti datové sadě rozšířen o další ať již strojově odhadnuté či uživatelem definované vazby, které poukazují na podobné vlastnosti výroků (například výroky z pokračující konfrontace dvou stran).

3 Závěr

Cílem příspěvku je informovat o vznikající datové sadě popisující konkrétní kauzu Novicok v České republice, ve které se střetávají Velká Británie, Rusko a Česká republika, podle [4] historicky propagandistické mocnosti. Příspěvek

¹⁹ například <http://www.ac24.cz/zpravy-ze-sveta/12410-jake-staty-dodat-novicok-cesko>

²⁰ například <https://echo24.cz/a/SMNNY/britove-pry-nasli-tajnou-ruskou-laborator-odkud-pochazi-jed-novicok>

²¹ například <https://www.echo24.cz/a/SJ4wm/nehci-vasi-pomoc-odmitla-rusy-dcera-otraveného-agenta-skripala>

²² <http://zpravy.e15.cz/zahranicni/udalosti/skripalova-byla-propustena-z-nemocnice-stav-jejeho-otce-se-lepsi-1345466>

²³ <https://aeronet.cz/news/novicok-nikdy-neexistoval-ve-skutecnosti-slo-o-noviczech-vtipkují-ruska-media-a-nejsou-daleko-od-pravdy-base-line-npl-a-230-se-totiz-pry-v-cesku-nevyrabela-ale-jenom-testovala/>

formou případové studie popisuje z datové sady konkrétní scénáře kombinující výroky a dává nahlédnout do možnosti pomocí pravděpodobnostního přístupu popsat důvody, proč k relativizaci daného výroku dochází. Stejný přístup pak umožňuje detekovat výroky, které mají relativizující účinek. Žádná infromatická míra však nenahradí lidský úsudek, avšak může upozornit na klíčové rozdíly v prezentovaných výroci. Jejich agregace může identifikovat autory a stranu, kterou svoji propagandou upřednostňují.

Literatura

1. Přehled dezinformačních a manipulativních webu (Think-tank Evropské hodnoty [online]), http://www.evropskehodnoty.cz/fungovani-ceskych-dezinformacnich-webu/weby_list/
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience (July 2006)
3. Galitsky, B.: Detecting rumor and disinformation by web mining. In: AAAI Spring Symposium Series (2015), <https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10251/10094>
4. Gregor, M., Vejvodová, P.: Nejlepší kniha o fake news. CPress, Brno (2018)
5. Kumar, S., West, R., Leskovec, J.: Disinformation on the web. In: Proceedings of the 25th International Conference on World Wide Web - WWW16. Association for Computing Machinery (ACM) (2016). <https://doi.org/10.1145/2872427.2883085>, <https://doi.org/10.1145/872427.2883085>
6. Whang, S.E., Garcia-Molina, H.: Disinformation techniques for entity resolution. In: Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management. pp. 715–720. CIKM '13, ACM, New York, NY, USA (2013). <https://doi.org/10.1145/2505515.2505636>, <http://doi.acm.org/10.1145/2505515.2505636>

Poděkování Tento článek vznikl s podporou na dlouhodobý koncepční rozvoj výzkumné organizace RVO:67985807.

Annotation

Web Disinformation Detection - Case Study - Novicok in Czechia

The paper presents a case study of the propaganda usage on a real cause of double agent Sergei Skripal. The formal model describing statements published in web articles is announced and particular interesting aspects of used disinformation are provided together with the reasons, why the disinformation is published. The paper is aimed at the presentation of the data collection to have been created and provides a brief discussion on the used propaganda techniques.

Self-Organizing Neural Networks for the Analysis of Country Development

Ildefonso Padrón Peña¹, Oto Mráz²

¹Department of Theoretical Computer Science and Mathematical Logic,
Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic
Malostranské náměstí 25, 118 00 Prague 1
ildepadronhosna@gmail.com

²Department of Informatics, Faculty of Natural & Mathematical Sciences,
King's College, London, UK
Bush House, 30 Aldwych, WC2B 4BG London
oto.mraz@gmail.com

Abstract. National economies can be influenced in various ways. Our research is focused on post-Soviet and Spanish speaking countries. To understand the actual state of the countries and to assess possible trends for their development, we use the SOM-networks. Based on development indicators provided by the World Bank, our approach is shown to explain both major trends as well as changes in the development of the countries. Stable economies are characterized by minor changes and sustained development. Crises, on the other hand, tend to manifest themselves rather by turbulent movements of the economy.

This paper demonstrates various changes in the country's economy, such as economic and moral crises, or socio-political turmoils, but also positive events like joining economical unions or organizing international sports events. Further, we will provide possible reasons for the causes of these trends and crises.

Key words: self-organizing maps, macroeconomic analysis with neural networks, data analysis, data visualization.

1 Introduction

Crises may take different forms such as war, economic downturn, natural disasters, etc. Conversely, events like hosting the football championships can bring tangible improvements, in Russia, it increased the Gross Domestic Product (GDP) by an extra 0.2% [7]. For different countries, however, their impact on society may differ. The main goal of our work is therefore to propose a new methodology capable of detecting regularities or anomalies in country development represented in the form of time-series data. The amount and character of the found anomalies might, namely, help to assess and visualise the actual state of the investigated country. This paper is focused on the evolution of post-Soviet and Spanish speaking countries based on data from the World Bank. We will consider a large range of indicators, i.e. GDP per capita, inflation or literacy rates, processed using self-organizing Kohonen maps (SOMs).

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 249-254.*

2 Related Work

Various research papers demonstrate applicability of neural networks and other machine learning techniques to the analysis of economic data. Recent models for time-series forecasting [2] and data clustering methods [1], [3], [5], [8] have been studied with emphasis on different aspects, e.g., time and space complexity, dependencies between variables, or correlation between data size and the model's architecture. Unfortunately, most of the known clustering approaches do not adequately reflect the topological structure of the data, which might provide better insight into the details of the data set. The capability of topology-preserving mappings represents, however an advantage of SOMs [4] with a huge potential impact on the result interpretation.

Kohonen maps provide a powerful tool allowing to process high dimensional data by reducing the dimensionality of the input space. The arrangement of the neurons on a 2D-grid should reflect the topology of the input structure. After a large number of training iterations, the neurons will be located accordingly on the map.

3 Data and its Pre-Processing

Our data was obtained from the World Bank database [9]. Altogether, 19 different indicators showing economic and social aspects from 2000 until 2012 were used, each carrying equal weight (**Tab. 1**). In our research we investigated the development of 16 post-Soviet states and 17 Spanish speaking countries in 2000-2012¹. In total, 14.66% and 13.41% of the data was missing from the post-Soviet and the Spanish speaking countries, respectively. For the approximation, the missing ends were substituted by the closest known value. Missing intervals with known bounding values were linearly approximated. If for a given indicator, a country had no data at all, the mean of the indicator was used. Then, the values of the i -th indicator, denoted val_i^{old} , were transformed into val_i^{new} , such that the transformed values retained the original distribution and remained in the range [0,1]. Equation (1) was used for pre-processing:

$$val_i^{new} = \frac{val_i^{old} - val_i^{min}}{val_i^{max} - val_i^{min}} \quad (1)$$

val_i^{min} and val_i^{max} correspond to the min and max values of the original val_i^{old} .

4 Supporting Experiments

For the prepared data, two SOMs with the same initial parameters were trained (one for the post-Soviet and one for the Spanish speaking countries). The SOM model used had a 2D hexagonal grid with 20×30 neurons in the Kohonen layer as this provides

¹ The post-Soviet countries are Russia, Estonia, Latvia, Lithuania, Belarus, Ukraine, Moldova, Georgia, Armenia, Azerbaijan, Kazakhstan, Uzbekistan, Turkmenistan, Tajikistan and Kyrgyzstan, Czech Republic was added for reference. The Spanish speaking states include Argentina, Bolivia, Chile, Colombia, Cuba, Dominican Republic, Ecuador, El Salvador, Honduras, Mexico, Panama, Paraguay, Peru, Puerto Rico, Spain, Uruguay and Venezuela.

a reasonably large space to obtain convincing results given the data size. Both SOMs were trained for 50000 iterations. Next, the data for each country was projected onto the respective SOM. The evolution of the economy over time is shown by a line connecting data points representing two consecutive years. Lastly, the initial (2000) and final (2012) data points were marked in green and red, respectively.

Among the post-Soviet countries two clusters were formed. The larger one on the right contains mainly the poorer regions, like Central Asia or the Caucasus. The smaller one on the left includes rather richer countries, e.g., the Baltic States. In the case of the Spanish speaking countries, geographically close regions tend to be adjacent to each other on the SOM. The main cluster on the right contains most Latin American countries. The two smaller clusters on the top left correspond to Spain and Puerto Rico (due to its political relationship with the US, an unincorporated territory).

Tab. 1. The used macroeconomic indicators

Development indicator	Units
1. GDP growth	annual %
2. GDP per capita	current US\$
3. GNI per capita, Atlas method	current US\$
4. GNI per capita, PPP	current international \$
5. Exports of goods and services	% of GDP
6. Imports of goods and services	% of GDP
7. GINI index	World Bank estimate
8. Foreign direct investment per capita	Balance of Payments, current US\$
9. Unemployment rate	% of total labour force
10. Poverty headcount ratio at national poverty lines	% of population
11. Inflation, consumer prices	annual %
12. Inflation, GDP deflator	annual %
13. Internet users	per 100 people
14. Life expectancy at birth, total	years
15. Adult literacy rate, population 15+ years, both sexes	%
16. Agriculture, value added	% of GDP
17. CO ₂ emissions	metric tons per capita
18. Central government debt, total	% of GDP
19. Population density	per km ²

5 Analysis of Country Development

To find regularities or anomalies, we used the Manhattan distance, $d_M(\mathbf{c}^t, \mathbf{c}^{t+1})$:

$$d_M(\mathbf{c}^t, \mathbf{c}^{t+1}) = |c_i^t - c_i^{t+1}| + |c_j^t - c_j^{t+1}|$$

of Kohonen neurons $\mathbf{c}^t$ with the coordinates $\mathbf{c}^t = (c_i^t, c_j^t)$ and $\mathbf{c}^{t+1}$ with the coordinates $\mathbf{c}^{t+1} = (c_i^{t+1}, c_j^{t+1})$ representing the considered country in the respective years t

and $t + 1$. Then we compared the obtained values $d_M(\mathbf{c}^t, \mathbf{c}^{t+1})$ with a threshold distance experimentally chosen to be 10% of the biggest possible displacement in the map. Since we used a 20 by 30 grid of neurons, the max Manhattan distance between any two neurons is $|30 - 1| + |20 - 1| = 29 + 19 = 48$ giving a threshold value of 4.8. If the obtained distance was higher, that year was marked as a significant change. This metric proved great in identifying the most important changes, which correlated very well with the socio-economic development in those years. **Tab. 2** contains the most interesting countries, with the significant changes in highlighted in grey.

Tab. 2. Positions of 6 example countries on the Kohonen maps

Country	2000	01	02	03	04	05	06	07	08	09	10	11	12
Spain	24	24	24	2	2	2	1	1	0	0	0	0	0
	0	0	0	10	10	10	9	9	10	7	7	7	8
Estonia	16	16	2	2	3	3	3	3	4	2	3	3	4
	16	16	5	5	3	3	5	5	4	4	4	4	5
Azerbaijan	8	8	3	3	4	6	3	5	5	5	15	12	14
	24	24	19	20	19	19	17	18	17	17	23	24	24
Argentina	17	17	14	16	16	24	24	24	24	22	27	27	23
	0	0	5	9	9	2	2	2	2	0	2	2	1
Armenia	3	4	4	4	4	4	3	3	4	4	5	5	5
	15	14	14	14	14	14	14	14	15	16	14	15	16
Bolivia	19	20	20	20	19	21	21	21	22	21	20	20	21
	6	5	5	5	5	7	8	8	8	11	11	11	10

Considering post-Soviet countries, Estonia, e.g., started in the main cluster, but moved to the smaller one in 2002, most likely due to its EU accession. Conversely, Azerbaijan changed its position quite frequently but lacked sustained progress. The first change in 2002 could be due to the constitutional amendments. In 2006 GDP growth peaked at 41.5%. Lastly in 2010 the price of oil increased heavily, which Azerbaijan relies on unlike its neighbours Georgia and Armenia.

Looking at the Spanish speaking countries, the isolation of Spain in a separate cluster since 2003 could be explained by the Euro adoption in 2002. Considering Argentina, big position changes happened in 2002, 2003, 2005, 2010, and 2012. The first one can be related to the Argentine Great Depression (1998-2002, the economy shrank by 28%). The economic recovery started with 2003 presidential elections. In 2005 the GDP exceeded pre-crisis levels. In 2009 Argentina was affected by the global recession. First GDP growth decreased to nearly zero, and then expanded to around 9% in 2010 falling back to 1,9% in 2012, one the lowest in Latin America.

Analysing the evolution of the countries in the Kohonen maps, we can thus distinguish the following three main types of countries (a complete set of the Kohonen maps as well as the data used can be accessed online [6]):

1. Countries with a very local movement on the grid. This can be understood as steady economic development, with no considerable socio-economic changes, e.g. Armenia (**Fig.1**), Tajikistan, Bolivia or El Salvador.

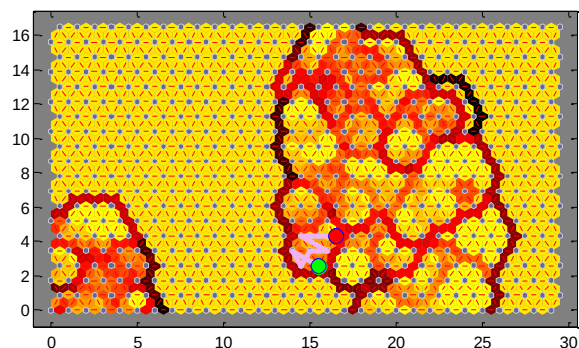


Fig.1.Projection of data from Armenia

2. Regions which underwent a notable shift such as Spain (**Fig.2**), Estonia, Latvia or Lithuania. Major changes in the country's position in the lattice often relate to significant social, political, or economic changes.

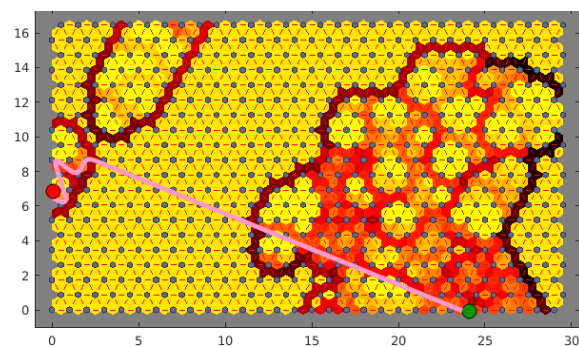


Fig.2. Projection of data from Spain

3. Countries behaving in an unstable manner, making several significant movements back and forth. These are often unstable economies with high levels of corruption and social insecurity, e.g., Azerbaijan (**Fig.3**), Moldova, Argentina or Colombia.

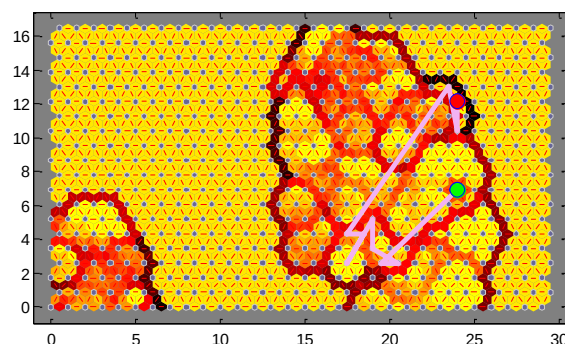


Fig.3. Projection of data from Azerbaijan

6 Conclusion

Our research has shown the usefulness of Kohonen maps for the analysis of high-dimensional macroeconomical data. The processed 19 indicators for post-Soviet and Spanish speaking countries were projected onto a 2D SOM-map, which helped us group together countries with similar development. Our original contribution to the area consists in introducing a new methodology based on the analysis of pre-trained Kohonen maps to identify anomalies in the development of time-series data.

Major shifts in the positions of the countries on the Kohonen maps were often accompanied by socio-political or economical reforms, like EU accession. Conversely, countries rich in oil shift a lot, yet without any steady improvements. Other reasons for instability may include also significant economic downturns, natural disasters or wars. Within the framework of our further research we would like extend our experiments to more countries from different parts of the world and to investigate additional indicators, identifying the crucial ones for a country's stable development and economic growth. This knowledge would help us predict which country will develop stably, and which one is likely to undergo a considerable change in the near future.

Literature

1. Collan, M., Eklund, T., Back, B.: Using the self-organizing map to visualize and explore socio-economic development. *EBS Review* 22 (2007) 6–15.
2. Cook, T.R., Hall, A.S.: Macroeconomic Indicator Forecasting with Deep Neural Networks. Research Working Paper No. 17-11. Federal Reserve Bank of Kansas City, USA (2017) 1–38, <https://doi.org/10.18651/RWP2017-11> [online accessed on 17. 6. 2018].
3. Focardi, S.M.: Clustering economic and financial time series: Exploring the existence of stable correlation conditions. Discussion Paper No. 04, TheIntertek Group, Paris, France (2001) 1–15, https://www.researchgate.net/profile/Sergio_Focardi2/publication/2547390_Clustering_economic_and_financial_time_series_Exploring_the_existence_of_stable_correlation_conditions/links/56b379c608ae5deb2657d5ba.pdf [online accessed on 17. 6. 2018].
4. Kohonen, T.: Self-organizing maps. Springer-Verlag, Berlin, 2001.
5. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *JMLR* 9 (2008) 2579–2605.
6. Mráz, O., Peña, I. P.: Kohonen map research (2018), <https://otomraz.wixsite.com/kohonenmapresearch>
7. Palata, L.: Russia was different. For a while. (Rusko bylo jiné. Na chvíli.) *Mladá Fronta Dnes*, Praha, 18. 7. 2018 (2018) p.7 (in Czech).
8. Řezanková, H.: Cluster Analysis of Economic Data. *Statistika: Statistics and Economy Journal* 94 (2014) 73–86.
9. World Bank: World Bank Open Data. World Bank, Washington, USA (2018), <https://data.worldbank.org> [Online accessed on 17. 6. 2018]

Predikcia úpadku s.r.o. využitím metód pre rozpoznanie odľahlých bodov na nevyvážených dátach

Peter Drotár¹, Martin Zorinčák² a Peter Gnip¹

¹ Katedra počítačov a informatiky, Technická univerzita v Košiciach

² Katedra financií, Technická univerzita v Košiciach

Abstrakt. Tento príspevok sa zaoberá problematikou predikcie bankrotu spoločností s ručením obmedzeným na Slovensku. Cieľom bolo vytvoriť vhodný klasifikačný model na základe dátovej sady výrazne nevyváženého charakteru, kde pomer nebankrotujúcich spoločností výrazne prevyšoval počet bankrotujúcich podnikov. Pri riešení tohoto problému sme použili tri rozdielne metódy strojového učenia. Najlepšiu úspešnosť predikcie sme dosiahli prostredníctvom klasifikátora LSAD, kde bola zaznamenaná hodnota geometrického priemeru vo výške 92% pre obdobie jedného roka pred rokom evaluácie. Výsledky boli overené na dátových záznamoch spoločností z odvetvia priemyslu a stavebníctva.

Kľúčové slová: Strojové učenie, nevyvážená dátová sada, metóda podporných vektorov, detekcia odľahlých bodov, bankrot spoločností

1 Úvod

Predikcia bankrotu spoločností je problematika, ktorá je diskutovaná už niekoľko desaťročí, avšak stále nie je dostatočne vyriešená. Schopnosť včasne a úspešne predpovedať, že fungovanie spoločnosti a jej finančné zdravie sa dostáva do ťažkostí, môže byť veľmi užitočným nástrojom v rukách vedenia podniku, ktoré môže včas reagovať na blížiacu sa nepriaznivú situáciu a tým predísť možným ťažkostiam, ktoré môžu ohroziť chod, ba až existenciu spoločnosti [1].

Prvé pokusy predikcie bankrotu spoločností boli založené na porovnávaní hodnôt pomerových ukazovateľov bankrotujúcich a nebankrotujúcich spoločností. Rozvoj štatistických metód zaznamenal výrazný úspech pri vytvorení Altmanovho Z-score [2]. Táto metóda bola postupne nahradená inými štatistickými metódami (logit a probit). Postupom času sa začali štatistické metódy nahrádzať sofistikovanejšími metódami, medzi ktoré môžeme zaradiť metódu podporných vektorov, neurónové siete, generické algoritmy [4] a metódu rozhodovania na základe predchádzajúcich skúseností [3]. Aktuálny prehľad existujúcich riešení a postupov je uvedený v prehľadovom článku [5]. Podstatou týchto je riešenie binárneho klasifikačného problému, ktorého princípom je zaradenie spoločnosti do bankrotujúcej alebo nebankrotujúcej skupiny. Vzhľadom na charakter nastolenej problematiky, často dochádza k problému nevyváženej dátovej sady,

ktorá predstavuje reálnu situáciu z praxe, keďže počet spoločností v úpadku predstavuje len zlomok z celkového počtu spoločností. Preto je potrebné k tejto problematike pristupovať ako k problému nevyvázenej dátovej sady.

2 Charakteristika dátovej sady

V tomto príspevku sme sa zamerali na analýzu a spracovanie dátovej sady výrazne nevyváženého charakteru (*imbalanced dataset*), ktorú tvorili záznamy spoločností s ručením obmedzením (s.r.o.) pôsobiace v rôznych odvetviach hospodárstva na Slovensku počas obdobia rokov 2010-2016. Spracovávané podniky boli charakterizované dátami z účtovných závierok za obdobie troch rokov pred rokom evaluácie R , ktorý vyjadruje rok, v ktorom bola spoločnosť vyhodnotená ako bankrotujúca/nebankrotujúca. Za bankrotujúcu spoločnosť považujeme tú, na ktorú bolo v danom roku evaluácie vyhlásené konkurzné konanie. V opačnom prípade bola spoločnosť označená ako nebankrotujúca. V tejto štúdii sme sa zamerali na vytvorenie klasifikačného modelu zameraného na predikciu bankrotu spoločností výhradne pre odvetvie priemyslu a stavebníctva. Prehľad konkrétnych pomerov bankrotujúcich a nebankrotujúcich spoločností pre jednotlivé roky evaluácie je znázornený v tabuľke 1.

Tab. 1. Pomer bankrotujúcich a nebankrotujúcich firiem pre jednotlivé roky evaluácie

	2013	2014	2015	2016
Priemysel	30/4077	30/4450	26/5019	14/5840
Stavebníctvo	25/1205	30/1418	20/1749	14/2174

Dátové záznamy boli spracované v programovacom jazyku *Python* s využitím metód obsiahnutých v knižnici strojového učenia *Scikit-learn* [6]. Každá účtovná závierka obsahovala 20 štandardných ekonomických ukazovateľov, pričom stručný prehľad základných pomerových ukazovateľov je uvedený v tabuľke 2. Tie boli doplnené ďalšími pomerovými ukazovateľmi (návratnosť investícií, mzdy/tržby a mzdy/pridaná hodnota).

Tab. 2. Prehľad použitých finančných pomerových ukazovateľov

zadlženosť	rentabilita	aktivita	likvidita
celková	aktív	doba obratu aktív	okamžitá
finančná páka	vlastného imania	doba obratu zásob	pohotová
vlastného imania	tržieb	doba obratu kd. pohľadávok	bežná
krytie cudzích zdrojov		doba obratu kd. zásob	
voči bankám			
finančné krytie cudzích zdrojov			
krytie stálych aktív			

3 Použitá metodológia

V tejto štúdií sme použili metódu podporných vektorov (*SVC*) [7] s natrénovaním na dvoch triedach a dve metódy na detekciu odľahlých bodov: metódu podporných vektorov s natrénovaním na jednej triede (*OCSVM*) [8] a metódu na detekciu anomálií využitím metódy najmenších štvorcov (*LSAD*) [9]. V porovnaní s našou predchádzajúcou prácou [12] sme pridali metódy *SVC* a *LSAD*, ktoré ukázali sľubný potenciál.

Chýbajúce hodnoty dátových záznamov boli imputované strednou hodnotou príslušného pomerového ukazovateľa. Ukazovatele boli taktiež škálované na nulovú strednú hodnotu a jednotkový rozptyl. Na natrénovanie klasifikátorov bolo použitých 80% vzoriek dátových záznamov. Trénovacia vzorka bola v prípade klasifikátora *OCSVM* a *LSAD* tvorená len dátami z majoritnej triedy. Testovacia vzorka pozostávala zo zvyšných 20% nebankrotujúcich podnikov a celej vzorky bankrotujúcich podnikov. V prípade klasifikátora *SVC* boli dátové záznamy bankrotujúcich spoločností zastúpené v testovacej aj trénovacej vzorke.

Experimenty boli opakované 1000-krát s náhodným výberom trénovacích dát a použitím všetkých kombinácií zvolených hodnôt hyperparametrov, ktorých prehľad je uvedený v tabuľke 2. Úspešnosť natrénovaných modelov bola hodnotená využitím metriky geometrického priemeru (*G-mean*) [10], ktorá zohľadňuje úspešnosť predikcie oboch tried binárneho klasifikačného problému, pričom dbá na ich vyváženosť. V prípade, že niektorá z tried dátovej sady dosahuje nízku úspešnosť predikcie, ovplyvní to výslednú hodnotu úspešnosti natrénovaného modelu aj napriek tomu, že presnosť druhej triedy je enormne vysoká [11]. Jednotlivé výsledky boli priemerné.

Tab. 3. Prehľad použitých hodnôt hyperparametrov pri jednotlivých metódach

Klasifikátor	Parameter	Základná množina hyperparametrov
SVC	class_weight c	[50; 100; 200; 300; 400; 500; 600] [0,0001; 0,001; 0,01; 0,1; 1; 10; 100; 1000]
OCSVM	nu	[0,01; 0,1; 1; 2; 3; 5; 10]
	gamma	[0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9; 1; 2; 5; 10]
	degree	[0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9; 1; 2; 5; 10]
LSAD	rho	[0,01; 0,1; 1; 2; 3; 5; 10]
	sigma	[0,1; 0,15; 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,65; 0,7; 0,75; 0,8; 0,85; 0,9; 1; 2; 5; 10]

4 Dosiahnuté výsledky

Hroziaci bankrot spoločností sme sa v prvej fáze tejto štúdie snažili predikovať na základe dostupných záznamov pomerových finančných ukazovateľov z obdo-

bia predchádzajúcich troch rokov jednotlivo, t.z. pre s.r.o. bankrotujúce v evaluačnom roku 2015 boli na predikciu bankrotu použité dáta z roku 2014 (R-1), 2013 (R-2) a taktiež 2012 (R-3) samostatne. Následne boli na predikciu použité dáta z dvoch rokov idúcich po sebe, t.z. pre podniky bankrotujúce v evaluačnom roku 2015 boli použité dáta z obdobia rokov 2014 a 2013 (R-1&R-2) a taktiež 2013 a 2012 (R-2&R-3), čím sme docielili použitie viacerých atribútov naraz. Na predikciu sme taktiež použili všetky dátové záznamy dostupné pre konkrétny rok evaluácie, t.j. pre rok 2015 boli použité dáta z rokov 2014 a 2013 a 2012 (R-ALL) súčasne. Rovnaký postup bol uplatnený pre všetky evaluačné obdobia pri oboch odvetviach hospodárstva. Dosiahnuté hodnoty úspešnosti natrénovaných modelov sú uvedené v tabuľke 4.

Tab. 4. Úspešnosť natrénovaných modelov vyjadrená geometrickým priemerom (v %)

	Data	Odvetvie priemyslu				Odvetvie stavebníctva			
		2013	2014	2015	2016	2013	2014	2015	2016
SVC	R-1	73,11±9,46	73,16±8,83	83,61±7,43	73,66±22,37	67,71±15,78	84,96±5,06	80,99±11,80	86,13±15,01
	R-2	71,98±10,82	65,21±9,19	71,12±12,84	54,12±27,17	63,32±8,49	63,83±12,39	73,60±13,12	80,01±14,54
	R-3	66,87±10,15	58,58±9,76	75,50±9,57	62,96±12,97	63,08±13,43	53,64±15,22	62,45±16,67	71,33±13,61
	R-1&R-2	73,97±10,95	69,67±12,87	79,29±8,84	72,40±24,33	68,24±11,62	80,06±9,81	80,60±13,06	86,76±9,53
	R-2&R-3	69,73±10,06	62,17±8,47	70,16±11,60	58,83±25,35	63,67±9,83	59,64±11,06	70,22±10,31	74,49±15,48
	R-ALL	71,27±10,75	66,61±12,83	83,53±9,20	70,52±24,35	67,19±12,80	77,88±10,28	79,50±12,75	84,63±10,35
OCSVM	R-1	61,47±2,84	73,12±1,96	72,50±2,35	82,37±1,22	69,79±3,41	74,14±2,09	84,28±2,17	80,13±1,48
	R-2	69,47±2,63	61,45±2,32	70,71±1,28	57,61±5,20	59,60±2,87	59,52±4,55	66,59±5,48	67,77±3,53
	R-3	65,01±2,94	60,91±2,50	66,66±1,85	58,53±5,44	51,09±2,08	60,55±3,27	62,59±2,53	64,94±7,33
	R-1&R-2	68,80±3,25	65,76±1,41	72,71±3,01	79,24±0,73	69,86±3,32	66,03±3,69	84,64±1,39	77,23±2,10
	R-2&R-3	67,61±2,55	59,03±1,75	67,06±1,72	63,19±4,11	53,32±2,73	56,92±2,69	59,23±2,76	65,37±2,89
	R-ALL	68,68±3,18	62,48±1,85	73,74±1,57	79,13±1,89	66,36±2,86	61,85±3,76	84,08±2,04	73,19±2,36
LSAD	R-1	76,67±2,27	76,73±2,01	75,57±3,61	87,86±1,32	77,86±2,36	82,86±3,3	86,62±1,73	91,54±1,74
	R-2	72,05±3,06	70,48±3,21	71,38±2,89	71,32±2,97	59,98±2,19	65,22±3,16	71,18±2,62	76,33±1,93
	R-3	60,43±3,57	62,63±3,31	65,22±3,27	53,66±3,27	55,56±3,45	56,53±2,67	63,01±4,36	71,57±4,41
	R-1&2&3	75,12±2,27	76,71±1,84	73,35±2,56	82,84±1,74	71,06±2,80	75,22±1,86	83,62±1,78	83,83±2,46
	R-2&R-3	69,11±2,14	65,89±2,57	66,49±2,41	68,22±2,43	56,51±2,74	64,29±4,01	65,12±2,68	74,73±2,63
	R-ALL	74,46±2,13	73,84±2,25	72,35±2,07	79,59±1,94	61,29±4,01	72,27±2,10	74,21±3,00	76,92±2,85

Zo získaných hodnôt úspešnosti predikcie natrénovaných klasifikačných modelov môžeme skonštatovať, že dáta z obdobia rokov bližších k roku evaluácie (R-1) majú v prípade všetkých použitých metód výpovednejšiu hodnotu. Najvyššie hodnoty úspešnosti boli zaznamenané pre klasifikátor LSAD s hodnotou geometrického priemeru od 75.57% - 87.86% (priemysel) a od 77.86% - 91.54% (stavebníctvo). Dosiahnuté hodnoty úspešnosti sa taktiež vyznačovali pomerne nízkou hodnotou štandardnej odchýlky nepresahujúcou hodnotu 3.3%. Pomerne vysoké hodnoty úspešnosti pri modeloch SVC nemôžeme považovať za spoľahlivé vzhľadom na vysoké hodnoty štandardnej odchýlky.

Pridanie nových atribútov (použitie 2 alebo 3 rokov súčasne) neprinieslo pri žiadnom klasifikátore výrazne zlepšenie. Vo väčšine prípadov sme mohli pozorovať zníženie presnosti natrénovaného modelu. Dôvodom môžu byť samotné dáta, pretože pridaním ukazovateľov z ďalšieho roku nie je pridaná žiadna nová informácia a taktiež neschopnosť klasifikátora dostatočne zmapovať spracovávané dáta.

5 Záver a ďalšie smerovanie výskumu

V tomto príspevku sme uviedli prvotné výsledky predikcie bankrotu spoločností s ručením obmedzeným dosiahnuté pomocou metódy SVC a dvoch metód na detekciu odlahlých bodov OCSVM a LSAD. Najlepšie hodnoty úspešnosti sme dosiahli pri použití klasifikátora LSAD s úspešnosťou od 75.57% - 87.86% (priemysel) a od 77.86% - 91.54% (stavebníctvo). Tieto modely sa taktiež vyznačovali pomerne nízkymi hodnotami štandardnej odchýlky, čo upevňuje doveryhodnosť dosiahnutých výsledkov. Na základe dosiahnutých výsledkov môžeme skonštatovať, že úpadok spoločností je pri použití výrazne nevyváženej dátovej sady predikovateľný.

V ďalšej štúdií sa pokúsime zamerať na detailnejšiu analýzu parametrov vplývajúcich na bankrot spoločností. Taktiež sa pokúsime o použitie iných metód strojového učenia (metódy pre klasifikáciu s učiteľom, podvzorkovanie a prevzorovanie). Jednou z možností je taktiež návrh nových atribútov, ktoré dokážu efektívnejšie reprezentovať finančnú situáciu v spoločnosti a tým umožniť klasifikačným algoritmom jednoduchšie spracovanie údajov.

Literatúra

1. Bellovary, J. L., Giacomino, D. E., Akers, M. D.: 2007. *A review of bankruptcy prediction studies: 1930 to present*. Journal of Financial Education, s. 1-42.
2. Altman, E. I.: 1968. *Financial ratios, discriminant analysis and the prediction of corporate bankruptcy*. The Journal of Finance, 23(4), s. 589-609.
3. Aamodt, A.: 1995. *Case-based reasoning - an introduction*. A newforce in advanced systems development, Unicom Seminars, s. 9-23.
4. Shin, K. S., Lee, Y. J.: 2002. *A generic algorithm application in bankruptcy prediction modelling*. Expert Systems with Applications, 23(3), s. 321-328.
5. Alaka, H. A., et al.: 2018. *Systematic review of bankruptcy prediction models: Towards a framework for tool selection*. Expert systems with Applications, 94. s. 164-184.
6. Pedregosa, F., et al. 2011. *Scikit-learn: Machine Learning in Python*. JMLR 12, s. 2825-2830.
7. Raschka, S.: 2015. *Python Machine Learning*. Birmingham: Packt Publishing Ltd., 2015. ISBN 978-1-78355-513-0.
8. Manevitz, L., Yousef, M.: *One-Class SVMs for Document Classification*. Journal of Machine Learning Research 2, 139-154 (2001).
9. Quinn, J., Sugiyama, M.: 2014. *A least-squares approach to anomaly detection in static and sequential data*. Pattern Recognition Letters, (44), s. 36-40.
10. Kubat, M., Matwin, S.: 1997. *Addressing the curse of imbalanced training sets: one-sided selection*. ICML (Vol. 97, s. 179-186).
11. Hido, S., et al.: 2009. *Roughly balanced bagging for imbalanced data*. In Statistical Analysis and Data Mining: The ASA Data Science Journal, (2), s. 412-426.
12. Gnyp, P., Zoricak, M., Drotar, P.: 2017. *Predikcia úpadku spoločností s ručením obmedzeným využitím metód pre rozpoznanie odlahlých bodov*. Data a Znalosti 2017, Plzen, CR.

PodĎakovanie: Táto práca bola podporená Agentúrou na podporu výskumu a vývoja projektom číslo APVV-15-0358.

Annotation: In this paper, we present the bankruptcy prediction model for the limited liability companies in Slovakia. The model is based on data from the annual reports, where the main challenge is the imbalanced distribution of classes. The number of non-bankrupt companies is significantly higher than the number of the bankrupt companies. We compared three different methods. The methods achieving the highest prediction performance is LSAD yielding as high as 92% score measured by the geometric mean. The results were validated on two datasets: one from the manufacturing industry and one from the building industry.

SoluProt: Prediction of Protein Solubility

Jiří Hon^{1,2,3}, Martin Marušiak³, Tomáš Martínek³, Jaroslav Zendulka³,
David Bednář^{1,2}, Jiří Damborský^{1,2}

¹ Loschmidt Laboratories, Centre for Toxic Compounds in the Environment
RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk
University, 625 00 Brno

² International Clinical Research Center, St. Annes's University Hospital Brno,
656 91 Brno

³ IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno
University of Technology, 612 66 Brno
`ihon@fit.vutbr.cz`

Abstract. Protein solubility poses a major bottleneck in production of many therapeutic and industrially attractive proteins. Experimental solubilization attempts are plagued by relatively low success rates and often lead to the loss of biological activity. Therefore, any advance in computational prediction of protein solubility may reduce the cost of experimental studies significantly. Here, we propose a novel software tool SoluProt for prediction of solubility from protein sequence based on machine learning and TargetTrack database. SoluProt achieved the best accuracy 58.2% and AUC 0.61 of all available tools at an independent balanced test set derived from NESG database. While the absolute prediction performance is rather low, SoluProt can still help to reduce costs of experimental studies significantly by efficient prioritization of protein sequences. The main SoluProt contribution lies in improved preprocessing of noisy training data and sensible selection of sequence features included in the prediction model.

Keywords: protein · solubility · prediction · machine-learning

1 Introduction

The key biochemical process limited by protein solubility is a heterologous expression – a manipulation of a living organism, usually *Escherichia. coli*, to produce a target protein of interest. Unfortunately, many proteins heterogously expressed are not sufficiently soluble. Although various biochemical approaches aimed at improving protein solubility were developed, solubilization attempts are plagued by relatively low success rates and often lead to the loss of biological activity.

Recently, various software tools were developed for prediction of solubility based on protein sequence – from the first reasonable model proposed by Wilkinson and Harrison [5] (mWH) to a battery of tools using more advanced machine-learning techniques, namely SOLpro [9], PROSO II [13], ccSOL omics

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 261-265.*

[1], Protein-Sol [6] and DeepSol [7]. The accuracy of available tools is, however, still limited and the problem remains unsolved. Moreover, the performance of the tools often turns up to be highly overestimated when later validated on larger test sets.

Here, we propose a novel software tool SoluProt that is more accurate than existing methods for prediction of solubility from protein sequence. The main effort was focused on dataset preprocessing, feature selection and unbiased comparison against existing methods.

2 SoluProt datasets

The main source of SoluProt training dataset is TargetTrack (TT) database [2]. As TT does not provide solubility information directly but only store reports from protein crystallization experiments, we inferred solubility computationally. All proteins that achieved soluble status or any subsequent stages requiring soluble expression, were considered soluble. When expression or purification failure was reported in the stop status, the protein was considered insoluble.

We did several additional steps to improve dataset quality. First of all, we performed expert-assisted keyword matching combined with manual checking of TT annotations to keep only those proteins expressed in the most common expression organism *E. coli*. This step significantly reduced the noise in the training dataset as it is known that a protein soluble in one expression system might be insoluble in another one. We also removed short proteins, sequences with undefined residues and transmembrane proteins predicted by *TMHMM* [8] tool. Next, we removed insoluble sequences that now have protein structure solved in PDB database. This step reflects the current advance in molecular biology as some older TT proteins can now be produced and crystallized using current technology.

In the last step, we reduced sequence redundancy in the training set by clustering to 25% identity using *MMseqs2* [14]. This was done separately for positive and negative samples to avoid simplification of the prediction problem. We also balanced class and sequence length distribution of soluble and insoluble samples such that the length alone was not decisive in the prediction and both classes were represented equally. Finally, the SoluProt training set consists of 10,912 protein sequences, having equal number of soluble and insoluble samples.

To build SoluProt test set, we used dataset of 9,703 proteins expressed in *E. coli* using unified production pipeline of North East Structural Consortium [12] (NESG) that measures solubility at integer scale from 0 to 5. The NESG dataset was processed using similar procedure as the training set, only the computational solubility derivation and expression system filtration steps were skipped as these were pointless in this case. Finally, 3,788 sequences remained in the balanced SoluProt test set having no overlap with the training set.

3 SoluProt predictor

SoluProt predictor is implemented in Python using *scikit-learn* [10], *Biopython*, *pandas* and *NumPy* libraries. Random forests regression [3] (RF) model was used for both feature pruning and the final prediction.

Prediction features were selected from a set of 253 values in eight subgroups: i) single amino acid content, ii) amino acid dimer content, iii), average physico-chemical features, iv) average flexibility by *DynaMine* [4], v) secondary structure content by *FESS* [11], vi) average disorder by *ESPRITZ* [15], vii) amino acids in transmembrane helices by *TMHMM* and viii) maximal identity to sequences in PDB.

Finally, 36 features were used in the prediction model. The five most dominant features are: i) maximal identity to PDB, ii) lysine to arginine ratio, iii) amino acid content, iv) isoelectric point and v) amino acid dimer content.

Optimization of random forests hyper-parameters was performed using 5-fold cross-validation and grid search on the SoluProt training set. The over-training was quantified by a difference between the accuracy on the training and validation part while the lower difference was preferred. The final model had average cross-validation accuracy 69% on training part and 65.3% on validation part.

SoluProt achieved the best accuracy (58.2% at threshold 0.53) and AUC (0.61) on the SoluProt test set from all available tools (Table 1 and Figure 1). With the test set, SoluProt reports 969 true positives, 1,237 true negatives, 657 false positives and 925 false negatives. This indicates that proper prediction of positive samples is the most problematic part of the solubility prediction problem.

Table 1. Performance of sequence-based solubility prediction tools on SoluProt test set. The best possible accuracy and corresponding threshold is presented.

Tool	AUC	Accuracy	Threshold
SoluProt	0.607	0.582	0.530
PROSOII	0.596	0.572	0.610
ESPRESSO	0.571	0.555	0.561
DeepSol	0.552	0.541	0.443
mWH	0.544	0.539	0.561
Protein-Sol	0.545	0.536	0.617
SOLpro	0.521	0.523	0.376
ccSOL omics	0.512	0.517	0.750

Although SoluProt is one of the best performing tool, its absolute accuracy is rather low. However, SoluProt can still be very useful for protein sequence prioritization (Figure 1). Using SoluProt prediction score on SoluProt test set to leave only 10% best proteins will increase the number of soluble proteins by

nearly 50% in comparison to what would be expected when selecting the same amount of proteins from the test set randomly.

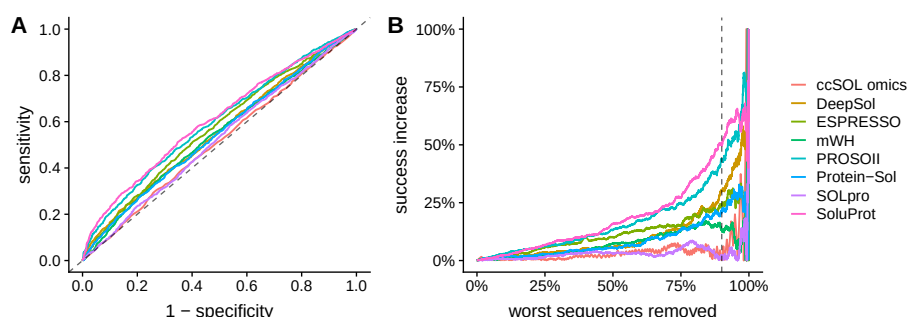


Fig. 1. Performance comparison. (A) ROC curves on SoluProt test set. (B) Increase of soluble sequences in prioritization task on the SoluProt test set in comparison to random subsampling. For SoluProt, removing 90% of worst scoring sequences (dashed line) will increase the number of soluble proteins by nearly 50%.

4 Conclusions

SoluProt tool moves the computational solubility prediction one step forward. The main SoluProt contribution lies in improved preprocessing of noisy TargetTrack training data and sensible selection of sequence features included in the prediction model. SoluProt achieved the best accuracy 58.2% and AUC 0.61 while being also the best tool in the task of protein prioritization.

However, there is still some space for improvement. Currently, the protein sequence is usually translated to numeric features that do not consider amino acid order. Consequently a protein sequence and its shuffled variant can have the same feature value while their solubility may differ significantly. This could be solved by embedding protein sequence into a vector space that would preserve sequence similarity. Another option might be the application of current advance in computational modelling of protein structures to build a spatial model from the protein sequence and subsequently derive more descriptive features that reflect the protein structure.

References

1. Agostini, F., Vendruscolo, M., Tartaglia, G.G.: Sequence-based prediction of protein solubility. *Journal of Molecular Biology* **421**(2-3), 237–241 (2012)
2. Berman, H.M., Gabanyi, M.J., Protein Structure Initiative Network of Investigators: Protein Structure Initiative – TargetTrack 2000–2017 – all data files (2017)

3. Breiman, L.: Random forests. *Machine Learning* **45**(1), 5–32 (2001)
4. Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., Vranken, W.F.: The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Research* **42**(Web Server issue), W264–270 (2014)
5. Davis, G.D., Elisee, C., Newham, D.M., Harrison, R.G.: New fusion protein systems designed to give soluble expression in *Escherichia coli*. *Biotechnology and Bioengineering* **65**(4), 382–388 (1999)
6. Hebditch, M., Carballo-Amador, M.A., Charonis, S., Curtis, R., Warwicker, J.: Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**(19), 3098–3100 (2017)
7. Khurana, S., Rawi, R., Kunji, K., Chuang, G.Y., Bensmail, H., Mall, R.: DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**(15), 2605–2613 (2018)
8. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L.: Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**(3), 567–580 (2001)
9. Magnan, C.N., Randall, A., Baldi, P.: SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* **25**(17), 2200–2207 (2009)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011)
11. Piovesan, D., Walsh, I., Minervini, G., Tosatto, S.C.E.: FELLs: fast estimator of latent local structure. *Bioinformatics* **33**(12), 1889–1891 (2017)
12. Price, W.N., et al.: Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microbial Informatics and Experimentation* **1**(1) (2011)
13. Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., Frishman, D.: PROSO II—a new method for protein solubility prediction. *FEBS Journal* **279**(12), 2192–2200 (2012)
14. Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**(11), 1026–1028 (2017)
15. Walsh, I., Martin, A.J.M., Di Domenico, T., Tosatto, S.C.E.: ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* **28**(4), 503–509 (2012)

Acknowledgement: This work was supported by ELIXIR CZ research infrastructure project [MEYS Grant No: LM2015047] including access to computing and storage facilities, IT4Innovations excellence in science project of the Ministry of Education, Youth and Sports of the Czech Republic [IT4I XS – LQ1602] and ICT tools, methods and technologies for smart cities project of the Brno University of Technology [FIT-S-17-3964]. Computational resources were supplied by the CESNET [LM2015042] and the CERIT Scientific Cloud [LM2015085], provided under the programme “Projects of Large Research, Development, and Innovations Infrastructures”.

digIS: An automated pipeline for detecting distant and novel insertion sequence elements in prokaryotes

Janka Puterová¹ and Tomáš Martínek²

¹ Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Božetěchova 2, 612,66 Brno, Czechia

`iputerova@fit.vutbr.cz`

² Department of Computer Systems, Faculty of Information Technology, Brno University of Technology, Božetěchova 2, 612,66 Brno, Czechia

`martinto@fit.vutbr.cz`

Abstract. Insertion sequences (ISs) are short DNA elements (typically less than 5 kb) representing the simplest transposable elements. They are widespread in prokaryotes and may occur in high numbers in their genomes. ISs play important role in prokaryotic genome structure and evolution, for example, they can cause large rearrangements in the host genome or gene inactivation. Current ISs annotation tools depend either on existing genome annotations or databases of known IS elements that results in an inability to detect novel IS elements or distant members of individual IS families. Here we introduce an approach for detection of distant members of IS families or novel IS elements based on manually curated profile hidden Markov models of transposases' catalytic domains which tackles the aforementioned shortcomings.

Key words: insertion sequence, transposase, hidden Markov model

1 Introduction

Insertion sequences (ISs) are the smallest and most abundant autonomous transposable elements in prokaryotic genomes, usually ranging from 700 bp to 3 kb in length. Majority of ISs possess one or two ORFs that encode a transposase (Tpase) or helper proteins. Tpase contains a conserved catalytic domain which catalyzes DNA cleavage.

Today, high-throughput sequencing techniques provide an enormous amount of data and availability of thousands of partially assembled or complete genomes requiring enhanced methods dISs detection and annotation. Although IS elements are relatively simple, their detection is challenging because of their noteworthy sequence and structural diversity.

At present, there are several available tools for detection of IS elements in genomes which are based on homology approach. They can be divided into two main categories: *i*) BLAST-based methods against reference libraries and *ii*) library-based methods using profile hidden Markov Model (pHMM) searches.

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 267-272.*

The first category is represented by tools IScan [16], ISSaga [15], OASIS [11] and ISQuest [1] which use ISFinder database [14] as a reference library in the combination with BLAST software. They are able to annotate IS elements in completely assembled and annotated genomes (IScan, ISSaga, OASIS) or in raw fragment data or partially assembled genomes (ISQuest). Their major weaknesses include a dependency on existing annotations and databases and an inability to find distant/novel members of IS families.

The latter category includes TnpPred [9] and ISEScan [18]. These tools take advantage of pHMMs which provide more sensitive searching and are able to detect remote homologs of known IS elements. TnpPred is able to detect only transposases, does not report full-length IS elements and provides pHMMs for 19 IS families only, although ISs are now classified into 28 families. ISEScan uses 621 pHMMs built from Tpsases contained in ACLAME database [6] but 355 of these models are made up from one sequence only.

In this paper, we present an approach for detection of distant members of known IS families or new IS elements utilizing manually curated pHMMs for Tpsases' catalytic domains which eliminates the shortcomings of existing IS-related tools. Due to the fact that Tpsase is composed of several often variable domains, the profile-based search is quite specific for 28 known families and is unable to detect new IS families. The fundamental idea is to search for the most conserved part of Tpsase (the catalytic domain) and to perform efficient filtering of false positives.

2 Materials and Methods

digIS pipeline is a complex tool implemented in Python whose development consisted of various steps which will be described in the following sections.

2.1 Datasets

To build pHMMs for Tpsases' catalytic domains, Tpsase sequences were downloaded from ISFinder database [14]. For filtering purposes, a positive and negative database was constructed as follows: reviewed protein sequences from various sources (SwissProt [2], RefSeq [8] and Protein Clusters database [12] for Bacteria and Archaea) were downloaded and searched by *hmmsearch* from HMMER3 package [7] using our pHMMs. Resulting hits which have sequence identity with ISFinder database greater than 80% or contain keyword 'transposase', 'insertion element' or 'transposon' are included in the positive database, otherwise in the negative database.

For evaluation purposes, a set of 58 prokaryotic genomes was downloaded from ISBrowser [5]. Annotations of IS elements together with their genomic locations in these genomes were downloaded from ISBrowser database [5] which were later used for performance comparison of our pipeline with existing tools TnpPred [9] and ISEScan [18].

2.2 Profile HMMs Construction

Multiple sequence alignment for sets of Tpsases belonging to the same family was performed by ClustalO [13] and visualized using Jalview Version 2 [17]. Alignments were manually refined and cut with respect to positions of catalytic residues (usually DDE) and a secondary structure of particular catalytic domain predicted by JPred4 [4]. Such processed sequences were used to construct resultant profile HMMs using *hmmbuild* from HMMER3 package [7].

2.3 digIS Pipeline Workflow

The *digIS* pipeline works as follows: Firstly, the input DNA sequences in FASTA format are translated into corresponding protein sequences in all six frames using *transeq* tool, which is included in EMBOSS package [10]. Next, our manually curated pHMMs are searched by *hmmsearch* from HMMER3 against the protein database assembled from translated input sequences with the e-value set to 10^{-3} . Overlapping hits and hits of the same query and at the distance no greater than *merge_distance* (default=300) are merged together. The resulting hits are extended in size defined by parameter *context_size* (default=1000). These are compared with sequences in the positive and negative database to filter out false positive hits.

2.4 Evaluation

Putative false positives (hits which are not in ISBrowser database) detected by *digIS* pipeline, which represent possible distant or novel IS elements, were expanded to both sides (max 5kb in length) and verified if they are distant/novel IS elements on multiple levels: *i*) DNA sequence level - identity with DNA sequences of all IS elements in ISFinder database, *ii*) protein sequence level - identity with protein sequences of ORFs in ISFinder database, *iii*) annotation level - considering GenBank [3] annotation. To determine, if FPs are distant/novel IS elements or not, we employed following scoring function:

$$fp_score = dna_iden + prot_iden + gb_annot$$

where *dna_iden* $\in [0, 100]$ is DNA sequence identity, *prot_iden* $\in [0, 100]$ is protein sequence identity and *gb_annot* acquires value 50 if the GenBank annotation is of a type 'mobile.element' or contains keywords 'transposase' or 'insertion element protein' in the protein product, value 25 if GenBank annotation has value 'hypothetical protein' in the protein product, value 0 if there is no GenBank annotation and value -25 if the GenBank annotation contains keyword other than 'transposase', 'insertion element protein' and 'hypothetical protein'. The *fp_score* can acquire value in the range -50 to 250. Based on the *fp_score*, the FPs were divided into three categories: *i*) true positives, *fp_score* $\in [250, 150)$, *ii*) putative novel elements, *fp_score* $\in [150, 50)$, *iii*) false positives, *fp_score* $\in [50, -50]$.

3 Results

The performance of the *digIS* pipeline was evaluated on 58 annotated prokaryotic genomes from ISBrowser database and compared with TnpPred and ISEScan tools based on their false discovery rate (FDR). The results are displayed in Table 1. *digIS*(Tnp-free) means that we used ISBrowser dataset in which elements without Tpnase were excluded. *digIS* pipeline found significantly more distant/novel elements (209) in comparison with the other tested tools (ISEScan - 31, TnpPred - 88) while having comparable FDR with ISEScan (0.037 vs 0.020). In contrast, TnpPred performed poorly with regard to FDR (0.163).

Tab. 1. Performance of ISEScan, digIS and TnpPred tools.

Tool	N	TP	FP_Novel			all_TP	real_FP	FDR
			TP	distant/novel	FP			
ISEScan	1207	941	155	31	22	1096	22	0.020
digIS	1207	846	280	209	43	1126	43	0.037
digIS (Tnp-free)	1058	846	280	209	43	1126	43	0.037
TnpPred	1207	952	135	88	211	1087	211	0.163

* *N* - Number of IS elements, *TP* - True Positives, *FP* - False Positives, *FP_Novel* - False Positives before filtering, *all_TP* - TP + FP_Novel(TP), *real_FP* - FP_Novel(FP), *FDR* - False Discovery Rate.

4 Discussion

Performance test of digIS with other two tools, TnpPred and ISEScan, indicates a significant improvement on an ability to detect distant or novel IS element while preserving reasonable FDR. We can conclude that digIS is able to find distant/novel IS elements which have yet been annotated and in this respect outperforms currently available IS-related tools.

Bibliography

1. Biswas, A., Gauthier, D.T., Ranjan, D., Zubair, M.: ISQuest: Finding insertion sequences in prokaryotic sequence fragment data. *Bioinformatics* **31**(21), 3406–3412 (2015). <https://doi.org/10.1093/bioinformatics/btv388>
2. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A.: UniProtKB/Swiss-Prot. *Methods in molecular biology* (Clifton, N.J.) **406**, 89–112 (2007)
3. Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W.: GenBank. *Nucleic acids research* **44**(D1), D67–72 (jan 2016). <https://doi.org/10.1093/nar/gkv1276>

4. Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J.: JPred4: a protein secondary structure prediction server. *Nucleic acids research* **43**(W1), W389–94 (jul 2015). <https://doi.org/10.1093/nar/gkv332>
5. Kichenaradja, P., Signier, P., Pérochon, J., Chandler, M.: ISbrowser: An extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids Research* **38**(SUPPL.1), 62–68 (2009). <https://doi.org/10.1093/nar/gkp947>
6. Leplae, R., Lima-Mendez, G., Toussaint, A.: ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research* **38**(suppl.1), D57–D61 (jan 2010). <https://doi.org/10.1093/nar/gkp938>
7. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M.: Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **41**(12), e121–e121 (jul 2013). <https://doi.org/10.1093/nar/gkt263>
8. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badredin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O’Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D.: Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* **44**(D1), D733–D745 (2016). <https://doi.org/10.1093/nar/gkv1189>, <http://dx.doi.org/10.1093/nar/gkv1189>
9. Riadi, G., Medina-Moenne, C., Holmes, D.S.: TnpPred: A Web Service for the Robust Prediction of Prokaryotic Transposases. *Comparative and functional genomics* **2012**, 678761 (nov 2012). <https://doi.org/10.1155/2012/678761>
10. Rice, P., Longden, I., Bleasby, A.: Emboss: the european molecular biology open software suite. *Trends in genetics* **16**(6), 276–277 (2000)
11. Robinson, D.G., Lee, M.C., Marx, C.J.: OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Research* **40**(22), e174–e174 (dec 2012). <https://doi.org/10.1093/nar/gks778>
12. Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E., Ye, J.: Database resources of the national center for biotechnology information. *Nucleic Acids Research* **37**(suppl.1), D5–D15 (2009). <https://doi.org/10.1093/nar/gkn741>, <http://dx.doi.org/10.1093/nar/gkn741>
13. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**(539) (2011). <https://doi.org/10.1038/msb.2011.75>
14. Signier, P.: ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research* **34**(90001), D32–D36 (2006). <https://doi.org/10.1093/nar/gkj014>
15. Varani, A.M., Signier, P., Goubeyre, E., Charneau, V., Chandler, M.: ISSaga is an ensemble of web-based methods for high throughput identification and semi-

- automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology* **12**(3), R30 (mar 2011). <https://doi.org/10.1186/gb-2011-12-3-r30>
16. Wagner, A., Lewis, C., Bichsel, M.: A survey of bacterial insertion sequences using IScan. *Nucleic Acids Research* **35**(16), 5284–5293 (2007). <https://doi.org/10.1093/nar/gkm597>
 17. Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., Barton, G.J.: Jalview Version 2 — a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9), 1189–1191 (2009). <https://doi.org/10.1093/bioinformatics/btp033>
 18. Xie, Z., Tang, H.: ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics* **33**(21), 3340–3347 (2017). <https://doi.org/10.1093/bioinformatics/btx433>, <http://dx.doi.org/10.1093/bioinformatics/btx433>

Acknowledgement: Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme ”Projects of Large Infrastructure for Research, Development, and Innovations”(LM2010005), is greatly appreciated. This work was supported by grant of the Brno University of Technology [FIT-S-17-3964].

Cytometry analysis and its problems

Tomáš Jarábek¹ and Mária Lucká²

¹ Fakulta informatiky a informačných technológií STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
`tomas.jarabek@stuba.sk`

² Fakulta informatiky a informačných technológií STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
`maria.lucka@stuba.sk`

Abstrakt Mass cytometry or CyTOF (from Cytometry by time-of-flight mass spectrometry) and the analysis of its data is a developing field in machine learning. Historically cytometry data was analyzed by manual clustering, but with the growing number of parameters obtainable from mass cytometry, the need for automated analysis by machine learning algorithms became clearer. There are many problems that algorithms analyzing the data must address. Some of these problems are the approach to missing or corrupted data, the need of recalibration to compare samples from different experiments and the most addressed problem, which is the identification of different cell type subpopulations in the data, which can then be used for further analysis. We show approaches how to address these problems as well as the downsides of modern algorithms.

Keywords: Cytometry · CyTOF · Data analysis

1 Introduction

Knowledge of the cellular structure acquires the usage of flow cytometry (FCM) or more recently mass cytometry (CyTOF) for single-cell analysis enabling the study of cellular and phenotypic diversity of complex cellular systems, in particular the immune system. Results are mainly used in clinical immunology [8].

Results of this analysis provide data on hundreds-of-thousands to millions of cells with up to 40 measured parameters per cell. Study of the data uses many machine learning algorithms to observe for expressions of individual cell types to identify differences between experimental groups.

The resulting data suffers from various deficiencies [3]. It usually contains debris and broken cells in the data samples which have to be filtered out. Data normalization is usually done by adding standardized polystyrene beads embedded with metal lanthanides into the studied sample which helps with the correction of both short-term and long-term signal fluctuations. This data normalization has been shown to not remove all problems which are being resolved by this normalization.

There are two main problems which cytometry analysis tries to address:

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 273-277.*

1. The identification of cell subpopulations in data.
2. The identification of the symptoms that are relevant for the diagnosis of diseases.

Most of the algorithms which analyze cytometry data are meant to address the first problems. By doing so they try to replace the step called "gating" [10] which is a technique of manually clustering the data. A step of manual gating consists of plotting the data in 2D plots with axes corresponding to two chosen parameters and manually clustering them for further analysis. This way it is repeated until it can reliably identify individual cell types. This technique has been used when the number of parameters was much smaller but is not scalable with increasing number of parameters.

2 Existing algorithms

There are many algorithms that analyze cytometry data. We will outline some of them to show major drawbacks of using them in automated cytometry analysis.

2.1 SPADE

SPADE (Spanning-tree progression analysis of density-normalized events) [9] is a tool, that is used to computationally analyze and visualize high-dimensional mass cytometry data. Spade displays cellular subpopulations in a branched tree structure, thus visualizing high-dimensional data in a 2D manner. The SPADE tree consists of connected nodes which represent clusters of cells, also providing information about the relation between cell types.

The SPADE algorithm can be summarized as:

1. Density-dependent down-sampling of cytometry data.
2. Agglomerative clustering of the down-sampled data.
3. Construction of the minimum spanning tree.
4. Up-sampling from the constructed SPADE tree.

The downside of this algorithm is that it depends on the interpretation of the results by biologists thus introducing bias to the analysis.

2.2 t-SNE

The nonlinear dimensionality reduction through t-Distributed Stochastic Neighbor Embedding (or t-SNE) [7] is used for the visualization of multidimensional cytometry data. t-SNE visualizes high-dimensional similarities between cells in an understandable 2D or 3D scatter plot. The distance between cells in a t-SNE map reflects their distance in the high-dimensional space.

The computational steps of t-SNE are:

1. Calculation of a pairwise similarity matrix for all cells based on their high-dimensional distances.

2. A low-dimensional similarity matrix is computed based on a initially random locations for each cell in the two (or three) t-SNE dimensions.
3. In an iterative process, the algorithm tries to minimize the difference between the high and the low dimensional matrices. Thereby it adjusts the position of all cells in the 2D or 3D space.

2.3 ACCENSE

A variation on the t-SNE algorithm with density-based partitioning is the algorithm ACCENSE (Automatic Classification of Cellular Expression by Nonlinear Stochastic Embeddings)[11].

The first part ACCENSE works in the same way as t-SNE, that is, it extracts a 2D representation of the data in a nonlinear way. Next it uses a kernel-based estimate of the 2D probability density of cells and its local maxima which correspond to cell subpopulations.

We can summarize the outline for the ACCENSE algorithm as:

1. Dimensionality reduction using t-SNE.
2. Peak detection using a peak-finding algorithm.
3. Population classification.

This approach, although heuristic, allows to approximately identify clusters of cells without having to specify the number of subpopulations.

2.4 Summary

There are many other algorithms for analyzing cytometry data like PhenoGraph [4], Wanderlust [1], Citrus [2], FlowSOM [12] or DeepCyTOF ([6], [5]). In the following part we will analyze the drawbacks of existing algorithms in addressing the main problems of cytometry analysis.

3 The drawbacks of current algorithms

To summarize the problems of existing algorithms the following observations can be made:

- The algorithms need to address missing or corrupted data which are present at the beginning of the analysis.
- The algorithms may need to recalibrate the data samples to compare different experiments.
- It is important to approach the problem of identifying different cell subpopulations by doing unsupervised clustering of data in such a way that new cell types may possibly be discovered.
- Most of the algorithms only address the problem of identifying different cell subpopulations but leave the interpretation of the results to biologists for the identification of the symptoms that are relevant for the diagnosis, thus introducing bias to the analysis.

It is also important to take the available computing resources it has for training into account, since the amount of data in cytometry analysis is relatively large. A sample of data corresponding to one person usually has from tens of thousands to hundreds thousands of cells each with the analyzed parameters (up to 40 parameters).

4 Conclusion

It is important to address the problems which were mentioned when designing new and better algorithms for cytometry analysis. Since the analysis works with a large amount of data, approximately 10^5 to 10^6 cells, the usage of deep learning for this task is very promising. There are many neural network architectures which can be used to address the problems. In the future we want to propose a deep learning architecture which will address the following tasks. The network will have the capability to remove or correct missing or corrupted data as well as the ability to recalibrate samples if needed. The network should also have the ability to identify important cell subpopulations in such a way that it will identify those cell types that it needs for the task but still takes into account existing cell type divisions.

5 Acknowledgment

This work was partially supported by the Scientific Grant Agency of The Slovak Republic, Grant No. VG 1/0458/18 and APVV-16-0484.

Reference

1. Bendall, S.C., et al.: Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell* **157**(3), 714 – 725 (2014)
2. Bruggner, R.V., et al.: Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences* **111**(26), E2770–E2777 (2014)
3. Florian, M., et al.: The end of gating? an introduction to automated analysis of high dimensional cytometry data. *European Journal of Immunology* **46**(1), 34–43
4. Levine, J.H., et al.: Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**(1), 184–197 (jul 2015)
5. Li, H., et al.: Deepcytof: Automated cell classification of mass cytometry data by deep learning and domain adaptation. *bioRxiv* (2016)
6. Li, H., et al.: Gating mass cytometry data by deep learning. *Bioinformatics* **33**(21), 3423–3430 (2017)
7. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008)
8. Maecker, H.T., McCoy, J.P., Nussenblatt, R.: Standardizing immunophenotyping for the human immunology project. *Nature Reviews Immunology* **12**, 191–200 (2012)

9. Qiu, P., et al.: Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nat Biotechnol* **29**(10), 886–891 (Oct 2011)
10. Shapiro, H.: *Practical Flow Cytometry*. Wiley (2005)
11. Shekhar, K., Brodin, P., Davis, M.M., Chakraborty, A.K.: Automatic classification of cellular expression by nonlinear stochastic embedding (accense). *Proceedings of the National Academy of Sciences* **111**(1), 202–207 (2014)
12. Sofie, V.G., et al.: Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A* **87**(7), 636–645

Webová aplikace statisticke_nastroje podporující správný výběr a realizaci zvolené statistické metody

Lubomír Štěpánek^{1,3}, Čestmír Štuka¹ a Martin Vejražka²

¹ Ústav biofyziky a informatiky

² Ústav lékařské biochemie a laboratorní diagnostiky

1. lékařská fakulta

Univerzita Karlova

{lubomir.stepanek, cestmir.stuka, martin.vejrazka}@lf1.cuni.cz

³ Katedra biomedicínské informatiky

Fakulta biomedicínského inženýrství

České vysoké učení technické v Praze

lubomir.stepanek@fbmi.cvut.cz

Abstrakt. Výzkumníci i studenti často čelí problému správného výběru statistické metody, která by měla být aplikována a vhodně tak analyzovala jejich data. I přes řadu dostupných komerčních a otevřených statistických balíčků existuje prostor pro nástroj kombinující výhody obou typů distribucí. Naší snahou je zjednodušit proces výběru a aplikace statistické metody pomocí vytvoření webové aplikace, která je volně k dispozici online. Aplikace `statisticke_nastroje` je vytvořena pomocí jazyka R a jednoho z jeho balíčků, `Shiny`. R je otevřený programovací jazyk a prostředí věnované zejména statistickým výpočtům a následným grafickým náhledům. Aplikace je vystavena na R-serveru 1. lékařské fakulty Univerzity Karlovy na http://shiny.statest.cz:3838/statisticke_nastroje/

a je založena na vývojovém diagramu navigujícím k použitelné statistické metodě – diagram umožňuje výzkumníkovi vybrat nejvhodnější statistickou metodu ve smyslu průchodu diagramem krok po kroku dle toho, jak odpovídá na dotazy „kladné“ diagramem. Po určení nejvhodnější metody je výzkumník odkazem přesměrován na záložku aplikace, která nabízí provedení dané metody nad daty, která byla do aplikace nahrána. Nabídku uživatelsky přátelské aplikace a vývojového diagramu mapujícího rozhodovací proces, který pokrývá aplikovatelné statistické metody a ukazuje na tu nejvhodnější, považujeme za cestu, jak celý tento proces zjednodušit. Přímé spojení (pomocí hyperlinku) mezi návrhem metody a záložkou s její realizací může zlepšit uživatelský komfort.

Klíčová slova: statistické výpočty · jazyk R · Shiny · webová aplikace · biostatistika · statistická podpora · statistické rozhodování · vývojový diagram

*J. Zendulka, M. Bieliková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 279-284.*

1 Úvod

Výběr správné statistické metody vhodné pro analýzu dat výzkumníka je důležitou součástí přípravy a realizace vědecké práce. Správnost této volby ovlivňuje i to, zda bude odborné sdělení nakonec redakčním řízením přijato, či nepřijato k publikování ve vědeckém časopisu. Ve snaze zjednodušit celý proces jsme navrhli a implementovali webovou aplikaci, jež pomáhá badateli čelícímu otázce výběru a vypracování statistického modelu ve své analýze. Grafické rozhraní aplikace je založeno zejména na využití vývojového diagramu, který uživatele naviguje k výběru a použití vhodné statistické metody.

Nástrojů a knihoven pro statistickou analýzu dat je na trhu dostupná velká řada [1]. Jde buďto o „klikací“ a mnohdy komerční software typu IBM SPSS [2], Dell Statistica [3], SAS [4], Minitab [5] či do jisté míry i tabulkový procesor MS Excel [6] a mnohé další, anebo o pokročilejší a speciálnější nástroje umožňující vlastní skriptování, programování a řešení rozsáhlých a komplexních úloh – zde se jedná například o prostředí Matlab [7] a Stata [8] či jazyk R se svými základními knihovnami [9] a jazyk Python s knihovnami SciPy, NumPy či Pandas [10], případně mnohé další.

Nespornou výhodou „klikacích“ nástrojů je jejich uživatelská přívětivost, nevýhodou pak relativní uzavřenost daná konečným počtem nabízených metod a implementovaných postupů, a tedy nemožnost řešit komplexnější a individuální úlohy. Limitující jsou zpoplatněné licence komerčních balíčků, a to hlavně mezi uživateli-jednotlivci, kam patří i studenti vysokých škol, kteří analyzují data ve svých závěrečných pracích.

Snaha o integrování konkrétnější podoby podpory správného výběru vhodné statistické metody, která se nejlépe hodí pro badatelova data a hypotézy, je naopak relativně vzácná. Produkty obvykle nabízejí pouze tištěný či online manuál k ovládání uživatelského rozhraní.

Cílem snažení autorů je poskytnout dobře dostupný, nekomerční nástroj pro co možná nejsnazší výběr a následnou aplikaci statistické metody, která je vhodná pro výzkumný záměr, hypotézy a data badatele. Cestou k takovému nástroji je dle autorů aplikace dostupná online a nevyžadující instalaci. Baterií implementovaných a nabízených metod je aplikace zaměřena zejména (ale nejen) na oblast biomedicíny. Práce s aplikací by měla být intuitivní a neměla by vyžadovat nutnost studia dlouhých manuálů. Těmto ideám se snaží aktuální verze vyvíjené a nabízené aplikace `statisticke_nastroje` co nejvíce přiblížit.

2 Materiál a metody

Aplikace `statisticke_nastroje` je volně a zdarma dostupná online. Jádru aplikace je napsáno v jazyce R. Jazyk R je „free-as-in-beer“ a „free-as-in-speech“, tj. je zdarma ke stažení a je možné svobodně upravovat jeho zdrojový kód. R je programovací jazyk a prostředí pro statistické výpočty a grafiku a je široce využíváno nejen statistiky. Užití jazyka R zajišťuje reliabilitu výpočtů prováděných zmíněnou aplikací.

Offline napsaný kód v jazyce R byl s použitím R-kového balíčku Shiny, který poskytuje online framework pro R-kové skripty, nahrán online na R-server 1. lékařské fakulty Univerzity Karlovy, kde je aplikace dostupná na adrese

http://shiny.statest.cz:3838/statisticke_nastroje/.

Jazykem volby aplikace je čeština, neboť v tuto chvíli cílí na česky mluvící výzkumníky a studenty.

Vývojový diagram integrovaný do aplikace pomáhá učinit správná statistická rozhodnutí v závislosti na použitých datech a zamýšlených statistických hypotézách; design diagramu vychází ze zásad „dobré praxe“ problematiky a z autorových zkušeností v oblasti; korektnost závěrů vývojového diagramu lze porovnat s doporučeními, která pro daný typ dat a hypotéz navrhuje např. [11].

Statistické metody, které jsou implementovány, pokrývají jak parametrické testy (t -testy, F -testy, ANOVA a další), tak neparametrické metody (Wilcoxonovy testy, Kruskal-Wallisův test, Friedmanův test a další), viz tabulku 1. Metody jsou přímo linkovány z koncových bodů diagramu k záložkám, které pak výpočet dané metody na vložených datech provádí.

Tab. 1. Implementované statistické metody

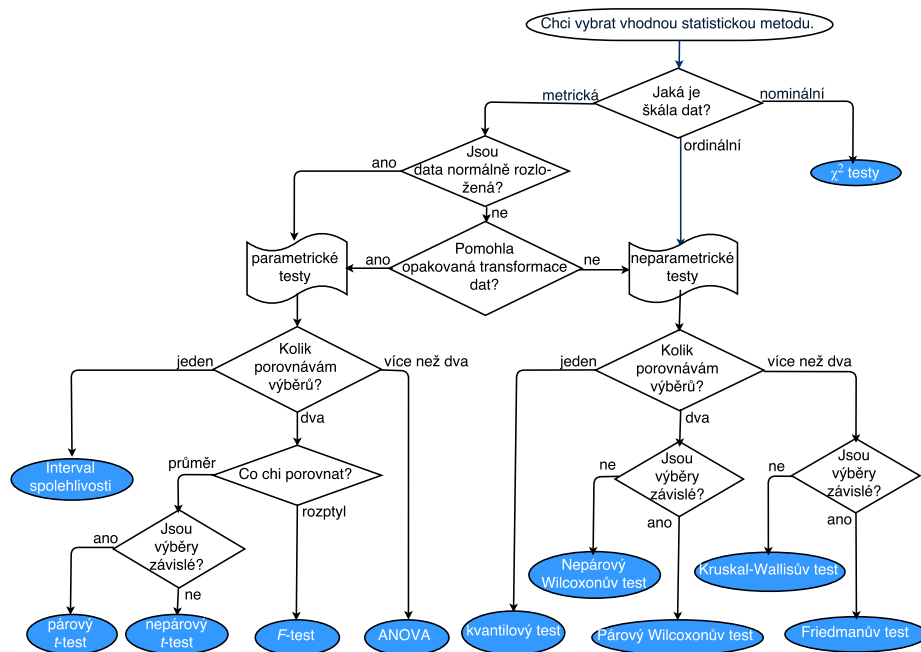
parametrické metody	neparametrické metody
t -testy	Wilcoxonovy testy
F -testy	Kruskal-Wallisův test
χ^2 -testy	Friedmanův test
AN(C)OVA	...
Shapiro-Wilkův test	

3 Výsledky

Klíčovou komponentou aplikace je zmíněný vývojový diagram, viz obrázek 1. Umožňuje výzkumníkovi, který již nasbíral data a formuloval výzkumnou hypotézu, vybrat pokud možno nejvíce odpovídající statistickou metodu, a to tak, že výzkumník projde diagramem v naznačeném směru (shora dolů) na základě odpovědí na pokládané otázky. Ve chvíli, kdy se výzkumník dostane postupným odpovídáním na otázky až ke koncovému bodu diagramu, je zde připraven odkaz na záložku, která vede k dané statistické metodě umožňující provést celou analýzu.

Jsou-li předtím do aplikace online nahrána adekvátní data a uživatel nastaví parametry analýzy, ta proběhne a uživateli nabídne možnost stáhnout offline výstupy vyprodukované analýzou – tabulky se sumáři, diagramy s infografikou.

Hlavní rozhodovací uzly vývojového diagramu vyžadují zodpovězení otázek týkajících se typů a struktur nasbíraných dat, empirického rozložení dat, počtu



Obr. 1. Vývojový diagram

porovnávaných výběrů a toho, zda by analýza měla porovnávat střední hodnoty nebo míry variability dat.

Uživatelské rozhraní aplikace odpovídá současným nárokům na webové utility. Roli back-endu hraje v aplikaci již zmíněné R, roli front-endu pak R-kový balíček *Shiny*, dále HTML, CSS a *javascript*, díky kterým je aplikace uživatelsky komfortnější, obrázek 2. Posledně jmenovaný jazyk umožňuje mít v aplikaci *busy indikátor* (symbol přesýpacích hodin při zaneprázdnění aplikace), thumbnaily (náhledy popisek při hover-over navigování nad odkaz), message hlášky a další funkcionality.

4 Diskuze a závěr

Přínos aplikace je momentálně předpokládán zejména pro uživatele z řad bakalářských a magisterských studentů různých oborů (bio)medicíny, kteří hledají relativně jednoduchý, přesto použitelný nástroj pro vypracování statistické analýzy do závěrečné práce.

Výhodou pro ně může být, že je aplikace zcela volně přístupná, není omezena žádnou licencí. Dalším plusem může být relativně velký prostor pro rychlé doplnění (doimplementování) chybějící, nicméně žádané metody.

Oproti studenty často používanému MS Excelu[®] nabízí aplikace možnost snadného vykreslení (a stažení) krabicového diagramu. Oproti jiným statistick-



Kvalita statistického aparátu použitého v odborném sdělení má jistě vliv na celkový dojem z publikace. Prvním krokem každé statistické analýzy je volba správného statistického modelu. Tento krok je však trikový a mnohdy netriviální i pro zkušeného badatele. Uživatelsky přátelská aplikace, která byla vyvinuta, se snaží tento krok zjednodušit. Propojení mezi doporučením konkrétní statistické metody, vyplývající z intuitivního vývojového diagramu, a záložkou, která analýzu danou metodou provede, má za cíl zlepšit uživatelský komfort.

[1] Zulfikar Ali a SBala Bhaskar. „Basic statistical tools in research and data analysis“. In: *Indian Journal of Anaesthesia* 60.9 (2016), s. 662. DOI: 10.4103/0019-5049.190623. URL: <https://doi.org/10.4103/0019-5049.190623>.

- [2] Corp. IBM. *IBM SPSS Statistics for Windows, Version 22.0*. Armonk, NY: IBM Corp. 2013.
- [3] Software Dell. *Statistica, Version 13.0*. 2015.
- [4] SAS Institute Inc. *SAS 9.4 Language Reference: Concepts*. Cary, NC, USA: SAS Institute Inc., 2013. ISBN: 1612905641, 9781612905648.
- [5] Inc. Minitab. *Minitab 17 Statistical Software*. State College, PA. 2010.
- [6] Microsoft. *MS Office Excel, version 16*. 2016.
- [7] C. M. Thompson a L. Shure. *Global Optimization Toolbox: User's Guide*. 2011.
- [8] StataCorp. *Stata Statistical Software: Release 14*. 2015. DOI: 10.2307/2234838.
- [9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL: <https://www.R-project.org/>.
- [10] Python Software Foundation. *Python Language Reference, version 2.7*. 2017. URL: <http://www.python.org>.
- [11] Jan Hendl. *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál, 2015. ISBN: 978-80-262-0981-2.

Annotation:

*Web-based application **statisticke_nastroje** enabling correct statistical test selecting and its execution*

Both researchers and students often face the problem of correct choice of the right statistical method which should be applied to analyse their data properly. Although there are plenty of commercial and open-source statistical packages, a space for a tool combining advantages of both such distributions still does exist. We tried to simplify the process of correct statistical test selecting and execution using a development of an online freely available web application. The application **statisticke_nastroje** is built with R language and one of its packages, Shiny. The R is an open-source programming language and environment designed for statistical computing and graphics. The application is available at R-server of the First Faculty of Medicine, Charles University, at http://shiny.statest.cz:3838/statisticke_nastroje/ and is based on flowchart diagram navigating to an applicable statistical method. The integrated flowchart enables a researcher to pick the most appropriate statistical method by the means of going through the flowchart step-by-step as she successively makes the decisions by answering the ordered questions offered by the chart, and, after that, she is linked to a tab of the application which allows using the suggested method and which provides an interface for uploading researcher's data and performing an analysis. We perceive that providing a user-friendly online application and flowchart mapping the decision-making process of possible statistical test choosing and pointing out the most appropriate one is the way how to make the described process more easy-to-follow. The direct connection between best possible statistical method and graphical interface performing the method computations could increase a level of user comfort.

PhD Odyssey

Interweaving PhD stories into the Semantic Web

Viet Bach Nguyen

Faculty of Informatics and Statistics
University of Economics, Prague
W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic
`nvbach91@outlook.com|nguv03@vse.cz`

Abstract. In this paper, I present a research project proposal for my upcoming PhD study in Applied Informatics. The main course of this project is ontology development for the academic domain, specifically about research activities of PhD students throughout their academic pilgrimages. The next goal is to create a game-like reference application featuring PhD story sharing, which would validate these ontologies. I also provide a pilot version of the project's specification, anticipated issues, outline and approach of my future research activities.

Keywords: academic domain · linked data · ontology engineering · PhD story · qualitative research · semantic web

1 Introduction

Ontology engineering can be perceived as a set of activities revolving around the creation of sophisticated domain-specific meta-modeling schemas. Each ontology model represents a formal definition of some domain-specific terms, concepts, and relations amid to describe an area of knowledge [9].

To build such a thing, the ontology practitioner must go through various steps to reach his goals [7]. Guidelines, frameworks, and best practices in the process of ontology development are still in an early stage of maturity. It is, however, an up-to-date research topic [4] thanks to the potential usage of ontologies in making knowledge explicit for utilization in knowledge-based systems and many applied areas such as artificial intelligence, life science, education and more [5].

2 Specification of research subject

As a PhD student, I would like to contribute to the scientific community by learning and bringing my own experience in ontology development. To do so, I propose a major subject in which ontology engineering activities are involved. This project will be focused on the analysis and creation of ontologies for the academic domain, specifically about PhD journeys in Applied Informatics, which would allow on-going and former students to structure their story experiences

*J. Zendulka, M. Bielíková, R. Burget, Z. Křivka (eds.),
Data a znalosti & WIKT 2018, Brno 11.-12.10.2018, pp. 285-288.*

and share them on the web. PhD experience may vary from person to person, but all of them must have something in common and I aim to formally generalize them using ontologies.

Many new PhD students often find themselves in rather difficult situations throughout the process of completing their program due to the insufficiency of important information and guidelines. The final objective of my dissertation would be to develop an application to support these students in the future. The main use case of this application would be to educate new students who just entered the scientific research world using an interactive and progressive story-telling game.

My dissertation activities will essentially revolve around studying existing PhD experiences. Research will include studying existing ontologies about research methods and publications involving ontologies and the Semantic Web. Other resources of interest include published PhD stories of computer science students around the world, either in written or audiovisual form. All this knowledge would be used for an interactive guide in creating and sharing PhD stories.

Presumably, this type of research is quite extensive. For that reason, boundaries and goals should be determined during the earliest phase of the project. As mentioned above, the activities and output of my dissertation would be bounded by Applied Informatics and might be limited to a selected geographical territory. Moreover, an early form of the prototyped version of the reference application would only feature a simple tagging model for story creation.

3 Related questions and issues

Abstraction and localization. PhD stories may come from many different countries and presented in many different styles. How to best comprehend the diversity of verbalized expressions and generalize them with abstractions? Is there a common pattern to look for?

Quality of resources. Throughout the process of studying existing PhD stories and memoirs, one might come to a problem where stories are incomplete or distributed unevenly in terms of continuity which might cause distortion of facts or misinterpretation of the author's original thought.

Regulations. The proposed application system will definitely work with personal data and other sensitive information. This poses a real challenge for the service provider to organize and protect data in a compliant and secure way. PhD stories would be accessible by the public and should not violate any law or regulations. One way to deal with this kind of problems is to inform users and obtain consent regarding their personal data. A more proactive approach is to use data anonymization techniques.

Interconnected stories. The question is how to deal with related stories as there might be a period of time when two or more stories can have a crossover. A proposed approach is to let the stories be viewed from different perspectives of each participant, respectively.

4 Anticipated strategy and research approach

To partially visualize the layout and scope of my future work, I have defined some of the tasks and goals as follows:

- Bibliography research including scientific conferences and journals.
- Research of ontology-related subjects including frameworks, best practices, guidelines and design approaches for utilization in real applications.
- Inspiration from old and contemporary PhD stories which are available online or documented as biographies.¹ Here lies a possibility for application of *thematic analysis* [6] involving practical methods such as examination, sampling, induction, deduction, inference and pattern recording.
- Interpretative ethnographic research in forms of participated observation or indirect surveillance. Resources of interest also include various types of documented or live materials.
- Creation of ontology models including re-usage of other ontologies.
- Creation of a domain-specific knowledge graph containing connected facts, figures and interrelations between concepts. This would acquire and integrate relevant information into an ontology and possibly derive new knowledge (similarly to DBpedia, Wikidata or Google Knowledge Graph, from which there would be some partial re-usage of concepts).
- Design and implementation of an interactive web application, which would presumably leverage ontology-driven software engineering [2].
- Partial gamification of that application with a simple but pleasant graphical design, inspired by the game of the Accountant².
- Analysis of personal data usage – compliance with GDPR.
- Possible application of story pattern mining during user input.

Since the main feature of my experimental activities is Design Research, the anticipated results of my work would be the following artifacts, all of which must then be practically validated by means of documented case studies:

- A **set of ontologies** describing the academic domain, providing knowledge of scientific research methods and methodologies.
- A **knowledge graph** instantiating the ontologies.
- A **software** showcasing the usefulness of the mentioned artifacts.

5 Related works

There are several related projects whose tangible outputs are ontologies that are devoted to the academic domain and research. The following ones have some features of interest that could be closely related to my dissertation.

¹ Many modern PhD students share their study experience online in forms of written blogs or audiovisual documentations aka vlogs.

² The Accountant game – <https://jplusplus.github.io/the-accountant/>

OLOUD [3] is an ontology that provides a high level model covering multiple education related use cases involving Linked Open University Datasets.

VIVO [1] is an internationally adopted ontology which enhances the discovery of researchers and collaborators across disciplines and organizations.

SWRC [8] is an ontology, in which research communities and relevant related concepts are modeled. It also emphasizes the importance of ontology re-usage.

6 Conclusion

The final objective of my dissertation project is to build an easily accessible, attractive knowledge-based system, providing users with an enjoyable process of learning about research processes and methods from a first-person-like perspective. This system can then be conceived as a common sharing platform among the scientific domain between academics, who are willing to share their experience in a less formal but fun way.

I believe my future work will help open the doors and create opportunities for new and on-going PhD students to do a better job upon their pilgrimage, pushing the scientific research community forward.

Acknowledgments. I would like express my sincerest gratitude towards Prof. Vojtěch Svátek for supporting me in my future study as my PhD advisor. The research has been partially supported by the CSF under 18-23964S.

References

1. Börner, K., Conlon, M., Corson-Rikert, J., Ding, Y.: *VIVO: A Semantic Approach to Scholarly Networking and Discovery*. Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers (2012)
2. de Cesare, S., Geerts, G.L., Holland, G., Lycett, M., Partridge, C.: *Ontology-driven Software Engineering*. In: *OOPSLA'09*. pp. 723–724. ACM (2009)
3. Fleiner, R., Szász, B., Micsik, A.: *OLOUD – An Ontology for Linked Open University Data*. *Acta Polytechnica Hungarica* **14**, 63–82 (Jan 2017)
4. Gangemi, A., Peroni, S., Shotton, D., Vitali, F.: *The Publishing Workflow Ontology (PWO)*. *Semantic Web* **8**(5), 703–718 (2017)
5. Gasevic, D., Djuric, D., Devedzic, V.: *Model Driven Engineering and Ontology Development* (2. ed.). Springer (2009)
6. Guest, G., MacQueen, K.M., Namey, E.E.: *Applied Thematic Analysis*. Thousand Oaks, CA: SAGE Publications Inc. (2012)
7. Noy, N.F., McGuinness, D.L., et al.: *Ontology development 101: A Guide to Creating your First Ontology* (2001)
8. Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D.: *The SWRC Ontology – Semantic Web for Research Communities*. In: *Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence*. pp. 218–231. *EPIA'05*, Springer-Verlag, Berlin, Heidelberg (2005)
9. Troullinou, G., Kondylakis, H., Daskalaki, E., Plexousakis, D.: *Ontology Understanding without Tears: The Summarization Approach*. *Semantic Web* **8**(6), 797–815 (2017)

Rejstřík autorů

Andrejčíková, Nadežda	139	Jarábek, Tomáš	273
Andrešič, David	89	Jirkovský, Václav	115, 127, 169
Babič, František	59, 187	Kadera, Petr	127, 169
Babič, Tomáš	177	Kompan, Michal	65, 199
Balázová, Veronika	101	Kostová, Sandra	95, 155
Balcar, Štěpán	163	Kováč, Jozef	149
Barla, Michal	199	Král, Pavel	173
Bednár, Peter	59, 65	Krejčířová, Veronika	215
Bednár, David	261	Kučera, Jan	49
Bieliková, Mária	65, 95, 101, 199	Kuchař, Jaroslav	111
Blaho, Radoslav	65	Labaj, Martin	183
Blšták, Miroslav	139	Lacko, Peter	65
Bončo, Tomáš	77	Laclavík, Michal	205
Bou Ezzeddine, Anna	139	Lenc, Ladislav	173
Burda, Kamil	193	Lóderer, Marek	139
Butka, Peter	59, 65, 107, 177	Loebl, Jaroslav	139
Ciglan, Marek	205	Lomič, Ondřej	215
Císar, Lukáš	215	Lucká, Mária	139, 273
Damborský, Jiří	261	Machová, Kristína	59, 65, 149
Drotár, Peter	255	Malkin Ondik, Irina	65
Dudáš, Marek	49	Marcinko, Ján	143
Gašpar, Peter	199	Maršáleková, Dominika	107
Gašpar, Vladimír	59	Martínek, Jiří	173
Gazda, Juraj	225	Martínek, Tomáš	261, 267
Genči, Ján	71	Martonová, Alena	65
Gnip, Peter	255	Marušiak, Martin	261
Görn, Christoph	211	Mesárošová, Barbora	65
Halla, Karel	173	Miklášová, Kristyna	215
Hon, Jiří	261	Móro, Róbert	65, 101
Hrčková, Andrea	155	Mráz, Oto	249
Chmelař, Petr	111	Mrázová, Iveta	231
Chudá, Daniela	65, 183, 193	Musil, Miloš	221

Mynarz, Jindřich	49	Sarnovský, Martin	59, 65, 143
Námešný, Matúš	77	Smatana, Miroslav	107
Návrat, Pavol	65	Srba, Ivan	65, 199
Nguyen, Viet Bach	285	Suchánek, Marek	121
Novák, Petr	169	Svátek, Vojtěch	49
Obitko, Marek	127	Šaloun, Petr	89
Ondřišová, Miriam	155	Šátala, Pavol	177
Oravcová, Katarína	143	Šebek, Ondřej	115
Padrón Peña, Ildefonso	249	Šimko, Jakub	65, 95, 155
Paralič, Ján	59, 65, 83	Šimko, Marián	65
Pavlišková, Viera	187	Široký, Radek	173
Pečíková, Bronislava	89	Šlapak, Eugen	225
Pergl, Robert	121	Štěpánek, Lubomír	279
Petliak, Nataliia	77	Štuka, Čestmír	279
Pokorný, Fridolín	53, 211	Tmenova, Oleksandra	77
Pokorný, Jaroslav	43	Vadovský, Michal	83
Popelínský, Luboš	77, 215	Vejražka, Martin	279
Pusztová, Ľudmila	187	Víta, Martin	237
Puterová, Janka	267	Vološin, Marcel	225
Rástočný, Karol	183	Vrablecová, Petra	139
Rozinajová, Viera	139	Vyskočil, Jiří	169
Rychtycký, Nestor	115	Zendulka, Jaroslav	261
Řepa, Václav	133	Zorinčák, Martin	255
Řimnáč, Martin	243	Zvirinský, Peter	231

Jaroslav Zendulka, Mária Bieliková, Radek Burget, Zbyněk Křivka
(Editoři)

DATA A ZNALOSTI & WIKT 2018, sborník konference

Vydalo Vysoké učení technické v Brně roku 2018
Určeno pro účastníky konference DATA A ZNALOSTI & WIKT 2018
1. vydání, 2018

Elektronické vydání.
291 stran

ISBN 978-80-214-5679-2