DATA A ZNALOSTI & WIKT 2019

Zborník konferencie

Editori

Peter Butka
František Babič
Ján Paralič

Košice, KONGRES Hotel Roca
Slovensko
10. - 11. október 2019

Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach

Data a Znalosti & WIKT 2019

Editori

Peter Butka
Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie
Letná 9
042 00 Košice
Slovensko

František Babič
Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie
Letná 9
042 00 Košice
Slovensko

Ján Paralič
Technická univerzita v Košiciach
Fakulta elektrotechniky a informatiky
Katedra kybernetiky a umelej inteligencie
Letná 9
042 00 Košice
Slovensko

Partneri vydania
Česká společnost pro kybernetiku a informatiku
Slovak.AI

Autori sú uvedení v obsahu. Príspevky sú publikované v tvare dodanom autormi, bez významných zmien. Text neprešiel jazykovou úpravou. Každý príspevok bol recenzovaný minimálne dvoma až troma recenzentmi, ktorí sú členmi programového výboru konferencie.

Vydala
Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach, 2019

Elektronická verzia zborníka konferencie

1.vydanie, 188 strán

ISBN 978-80-553-3354-0

Predhovor

Stretnutia českej a slovenskej komunity venujúcej sa výskumu a vývoju aplikácií v oblastiach databázových technológií, informačných systémov, či dátového a znalostného inžinierstva, majú svoju dlhodobú tradíciu. Hlavnou pridanou hodnotou je práve ich komunitný charakter ako prostriedku pre vzájomnú informovanosť, predávanie skúseností a udržiavanie tradične výborných vzťahov medzi pracoviskami v Česku a na Slovensku. Okrem toho tieto stretnutia ponúkajú možnosť stretnutí a výmenu skúseností medzi odborníkmi z akadémie a priemyselnej praxe. Sme toho názoru, že tieto stretnutia majú zmysel aj v dnešnej dobe, napriek evidentnému tlaku na výskumníkov zúčastňovať sa najmä široko koncipovaných medzinárodných konferencií a kongresov. Aj preto nám je cťou, že sme dostali možnosť organizovať spoločné podujatie spájajúce opäť po roku konferenciu Data a znalosti a workshop WIKT (Workshop on Intelligent and Knowledge oriented Technologies). Tento ročník spoločného podujatia sa konal pod názvom **Data a znalosti & WIKT 2019** v dňoch 10. a 11. októbra 2019 v Košiciach v hoteli KONGRES Hotel Roca.

Data a znalosti je česko-slovenská odborná konferencia a súčasne komunitné stretnutie odborníkov zamerané na najlepšie postupy a vývojové trendy v oblasti dátového, informačného a znalostného inžinierstva, ako aj na využitie informačných technológií pri budovaní informačných systémov, vrátane výsledkov ich aplikácie v praxi. Tento rok sa konal 5. ročník konferencie, ktorá nadväzuje na dlhoročnú tradíciu dvoch prestížnych česko-slovenských konferencií: na konferenciu Datakon, ktorá existovala od roku 2000, kedy nadviazala na konferenciu s vtedy dvadsaťročnou tradíciou – Datasem, a na konferenciu Znalosti, ktorá existovala od roku 2001.

Workshop **WIKT** zaznamenal v rámci tohtoročného spoločného podujatia svoj 14. ročník. Jeho dlhodobým zameraním sú inteligentné a znalostne orientované technológie pre podporu posunu organizácií smerom k znalostnej ekonomike na základe výskumu a vývoja v tejto oblasti, ktorá sa stáva kľúčovým faktorom znalostnej ekonomiky. Cieľom je výmena informácií o prebiehajúcom výskume, diskusia o aktuálnych problémoch v oblasti inteligentných a znalostných technológií a možných spôsoboch ich riešenia. V neposlednom rade ide aj o výmenu skúsenosti s použitím relevantných pokročilých technológií a softvérových nástrojov, spôsobov ich využitia a nasadenia pri riešení úloh v praxi.

Medzi hlavné témy tohto ročníka spoločného podujatia Data a znalosti & WIKT 2019 patrili:

- Dátovo centrická bezpečnosť
- Technológia blockchain
- Multi-modelové databázové architektúry
- Získavanie, ukladanie, spracovanie a vizualizácia veľkých dát (Big Data)
- Tvorba, publikovanie a využívanie otvorených a prepojených dát (Linked Data)
- Indexovanie a vyhľadávanie textových a multimediálnych dát
- Procesy a roly v správe dát
- Architektúry podnikových systémov
- Strojové učenie, hlboké učenie, data mining (Data Science)
- Detekcia anomálií
- Anonymizácia a zachovanie súkromia pri dolovaní z dát
- Aplikácia strojového učenia v bioinformatike, v počítačovom videní a pri spracovaní reči

- Modelovanie používateľa, adaptívne a personalizované systémy
- Pokročilé používateľské rozhrania softvérových a informačných systémov
- Systémy pre správu znalostí v organizáciách
- Expertné, inteligentné a agentové systémy, výpočtová inteligencia
- Výpočtová lingvistika
- Ontologické a konceptuálne modely
- Automatické odvodzovanie a plánovanie
- Znalostné technológie a ich aplikácie
- Dolovanie v dátach a ich aplikácie
- Big data, možné prístupy, vhodné technológie
- Cloud, technické riešenia, aplikačné príklady
- Modelovanie informácií a znalostí, reprezentácia sémantiky
- Analýza a spracovanie informačných zdrojov
- Sociálny web a jeho aplikácie, analýza sociálnych sietí
- Personalizovaný web a jeho aplikácie, odporúčania
- Spracovanie informačných zdrojov v českom/slovenskom jazyku
- Sémanticky a servisne orientované architektúry
- Usudzovanie a odvodzovanie

V rámci konferenčného systému bolo prijatých 36 podaní, z nich 34 bolo akceptovaných pre prezentovanie na podujatí, z toho 24 formou prezentácie v klasických sekciách a 10 formou posterov. V rámci akceptovaných prác sme zaznamenali rôzne typy príspevkov, na základe ktorých je usporiadaný aj tento konferenčný zborník – výskumné príspevky (9), aplikačné príspevky (6), projektové príspevky (4), príspevky o prebiehajúcom výskume (7) a príspevky v rámci doktorandského sympózia (8). Všetky príspevky prešli recenzným konaním, za čo sa chceme týmto poďakovať členom programového výboru konferencie Data a znalosti & WIKT 2019. Samozrejme, vďaka patrí aj všetkým autorom príspevkov a účastníkom konferencie. V neposlednom rade by sme sa chceli poďakovať autorom pozvaných prednášok, ktorých abstrakty sú uvedené v úvodnej sekcii pred prijatými príspevkami. Súčasťou konferencie bola aj sekcia venovaná iniciatíve Slovak.AI.

Veľká vďaka patrí takisto partnerom konferencie, konkrétne spoločnosti ČSKI (Česká společnost pro kybernetiku a informatiku) a iniciatíve Slovak.AI. Na záver je na mieste sa poďakovať aj organizačnému výboru konferencie a hotelu, ktorí umožnili našej komunite opäť sa po roku stretnúť v tradične príjemnej atmosfére.

Košice, október 2019

Peter Butka, František Babič, Ján Paralič

Organizácia konferencie

Stály riadiaci výbor konferencie Data a znalosti

Predseda (Chair):	CHLAPEK, Dušan (VŠE Praha)
Členovia (Members):	BIELIKOVÁ, Mária (FIIT STU Bratislava)
	BURGET, Radek (FIT VUT Brno)
	HUJŇÁK, Petr (Per Partes Consulting Praha)
	KORDÍK, Pavel (FIT ČVUT Praha)
	MATIAŠKO, Karol (ŽU Žilina)
	PARALIČ, Ján (FEI TU Košice)
	POKORNÝ, Jaroslav (MFF UK Praha)
	POPELÍNSKÝ, Lubomír (FI MU Brno)
	RAUCH, Jan (VŠE Praha)
	RICHTA, Karel (FEL ČVUT Praha)
	SVÁTEK, Vojtěch (VŠE Praha)
	ŠALOUN, Petr (Univerzita Palackého v Olomouci)
	VALENTA, Michal (FIT ČVUT Praha)

Stály riadiaci výbor WIKT

Predseda (Chair):	BIELIKOVÁ, Mária (FIIT STU Bratislava)
Členovia (Members):	BUTKA, Peter (FEI TU Košice)
	HLUCHÝ, Ladislav (ÚI SAV Bratislava)
	HOMOLA, Martin (FMFI UK Bratislava)
	HORVÁTH, Tomáš (PrF UPJŠ Košice)
	CHUDÁ, Daniela (FIIT STU Bratislava)
	KRAJČI, Stanislav (PrF UPJŠ Košice)
	LACLAVÍK, Michal (Deloitte Digital)
	MACH, Marián (FEI TU Košice)
	MACHOVÁ, Kristína (FEI TU Košice)
	MATIAŠKO, Karol (FRI ŽU Žilina)
	NÁVRAT, Pavol (FIIT STU Bratislava)
	PARALIČ, Ján (FEI TU Košice)
	ROZINAJOVÁ, Viera (FIIT STU Bratislava)
	ŠALOUN, Petr (Univerzita Palackého v Olomouci)
	VOJTÁŠ, Peter (MFF UK Praha)
	ZENDULKA, Jaroslav (FIT VUT Brno)

Programový výbor konferencie Data a znalosti & WIKT 2019

Predseda (Chair):	BUTKA, Peter (FEI TU Košice)
Členovia (Members):	BABIČ, František (FEI TU Košice)
	BARLA, Michal (FIIT STU Bratislava)
	BARTÁK, Roman (MFF UK Praha)
	BARTÍK, Vladimír (FIT VUT Brno)
	BERKA, Petr (VŠE Praha)
	BIELIKOVÁ, Mária (FIIT STU Bratislava)
	BOHÁČIK, Ján (FRI ŽU Žilina)
	BURGET, Radek (FIT VUT Brno)
	DOSTAL, Martin (ZČU Plzeň)
	DROTÁR, Peter (FEI TU Košice)
	FARKAŠ, Igor (FMFI UK Bratislava)
	FIALA, Dalibor (ZČU Plzeň)
	GAZDA, Juraj (FEI TU Košice)
	GENČI, Ján (FEI TU Košice)
	GURSKÝ, Peter (PrF UPJŠ Košice)
	HAVLICE, Zdeněk (FEI TU Košice)
	HLUCHÝ, Ladislav (SAV Bratislava)
	HOLEŇA, Martin (ÚI AV ČR)
	HOLUBOVÁ, Irena (MFF UK Praha)
	HORVÁTH, Tomáš (PrF UPJŠ Košice)
	HRUŠKA, Tomáš (FIT VUT Brno)
	HUJŇÁK, Petr (Per Partes Consulting Praha)
	CHLAPEK, Dušan (VŠE Praha)
	CHUDÁ, Daniela (FIIT STU Bratislava)
	KLEČKOVÁ, Jana (ZČU Plzeň)
	KLÉMA, Jiří (FEL ČVUT Praha)
	KLÍMEK, Jakub (FIT ČVUT Praha)
	KOMPAN, Michal (FIIT STU Bratislava)
	KORDÍK, Pavel (FIT ČVUT Praha)
	KRAJČI, Stanislav (UPJŠ Košice)
	KRÁL, Pavel (ZČU Plzeň)
	KŘEMEN, Petr (FEL ČVUT Praha)
	KŘIVKA, Zbyněk (FIT VUT Brno)
	KUČERA, Petr (Komix)
	LABSKÝ, Martin (IBM TJW, Praha)
	LACKO, Peter (FIIT STU Bratislava)
	LACLAVÍK, Michal (Deloitte Digital)
	LEVASHENKO, Vitaly (FRI ŽU Žilina)
	MACH, Marián (FEI TU Košice)

MACHOVÁ, Kristína (FEI TU Košice)
MATIAŠKO, Karol (ŽU Žilina)
MIKULECKÝ, Peter (Univerzita Hradec Králové)
MOLHANEC, Martin (FEL ČVUT Praha)
MOUČEK, Roman (ZČU Plzeň)
NÁVRAT, Pavol (FIIT STU Bratislava)
NGUYEN, Giang (UI SAV Bratislava)
PARALIČ, Ján (FEI TU Košice)
PERGL, Robert (FIT ČVUT Praha)
PITNER, Tomáš (FI MU Brno)
POKORNÝ, Jaroslav (MFF UK Praha)
POPELÍNSKÝ, Lubomír (FI MU Brno)
RAUCH, Jan (VŠE Praha)
ŘEPA, Václav (FIS VŠE Praha)
RICHTA, Karel (FEL ČVUT Praha)
ROZINAJOVÁ, Viera (FIIT STU Bratislava)
RUDOVÁ, Hana (FI MU Brno)
SMRŽ, Pavel (FIT VUT Brno)
SRBA, Ivan (FIIT STU Bratislava)
STEINBERGER, Josef (ZČU Plzeň)
SVÁTEK, Vojtěch (VŠE Praha)
ŠALOUN, Petr (Univerzita Palackého v Olomouci)
TVAROŽEK, Jozef (FIIT STU Bratislava)
VALENTA, Michal (FIT ČVUT Praha)
VOJTÁŠ, Peter (MFF UK Praha)
ZAMAZAL, Ondřej (VŠE Praha)
ZENDULKA, Jaroslav (FIT VUT Brno)
ZÍMA, Martin (ZČU Plzeň)

Organizačný výbor konferencie Data a znalosti & WIKT 2019

Predseda (Chair):	BABIČ, František (FEI TU Košice)
Členovia (Members):	IVANČÁKOVÁ, Juliana (FEI TU Košice)
	MASLEJ KREŠŇÁKOVÁ, Viera (FEI TU Košice)
	PUSZTOVÁ, Ľudmila (FEI TU Košice)

Konferenciu organizuje

Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach

Partneri konferencie

Česká společnost pro kybernetiku a informatiku

Slovak.AI

Obsah

Aplikačné príspevky

Projektové príspevky

Príspevky o prebiehajúcom výskume

Doktorandské sympózium

Machine Learning ako akcelerátor StartUpu

Andrej Makovický

Kiwi.com
https://www.kiwi.com/

Pozvaná prednáška

Abstrakt. V Kiwi.com používame (skoro) plný rozsah analytických a ML metód. V praxi je dôležité vybrať tú správnu a najefektívnejšiu vzhľadom k biznisovým potrebám. Prezentácia vám ukáže ako to robíme u nás v Kiwi.com, na čom sme sa spálili a čo nám funguje dobre. Budú príklady aj dáta.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 13-13.

Rola UX v procese tvorby digitálnych produktov (pohľad z praxe)

Martin Krupa

ui42
https://www.ui42.sk/

Pozvaná prednáška

Abstrakt. Spoluzakladateľ Slovenskej UX asociácie (SUXA) porozpráva o svojich vyše 20-ročných skúsenostiach s tvorbou používateľských rozhraní pre softvér. Poskytne prehľad overených metód pri ich tvorbe. Odprezentuje aj svoje poznatky z oblasti UX vzdelávania.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 14-14.

Precízne poľnohospodárstvo - nápady a predstavy vs. realita

Tomáš Horváth

ELTE Eötvös Loránd University, Faculty of Informatics,
Department of Data Science and Engineering, Telekom Innovation Laboratories,
Pázmány Péter sétány 1/C, 1117, Budapest, Hungary
tomas.horvath@inf.elte.hu
http://t-labs.elte.hu/?page_id=151

Pozvaná prednáška

Abstrakt. Precízne poľnohospodárstvo (PP) priťahuje čoraz väčšiu pozornosť informatickej komunity. Avšak, ako informatici, pri plánovaní projektov v oblasti PP zvyčajne nerátame s faktormi, ktoré sú kritické pre úspešnosť celého projektu. V prednáške na reálnych príkladoch zhrniem naše znalosti a skúsenosti, ktoré je dobré vedieť predtým, než sa pustíme do akéhokoľvek projektu v oblasti PP.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 15-15.

Analýza významnosti faktorov vplývajúcich na závažnosť kardiovaskulárnych ochorení

Zuzana Pella[1], Oliver Lohaj[2], Ján Paralič[1], Dominik Pella[3]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
{zuzana.pella,jan.paralic}@tuke.sk
[2]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
oliver.lohaj@student.tuke.sk
[3]1. Kardiologická klinika UPJŠ LF a VÚSCH a.s.
Ondavská 8, 040 11 Košice
dominik.pella@gmail.com

Abstrakt. So stúpajúcou prevalenciou kardiovaskulárnych ochorení je potrebné detailnejšie skúmať vplyv rôznych faktorov, ktoré ich ovplyvňujú. Interdisciplinárnym spojením medicíny a dátovej analýzy sa pokúsime preukázať rôznu silu vplyvu jednotlivých faktorov na stupeň závažnosti kardiovaskulárnych ochorení. Údaje o pacientoch sme získali zo špecializovaného pracoviska 1. Kardiologickej kliniky UPJŠ LF a VÚSCH a.s., ktoré boli vo forme lekárskych správ postúpené na spracovanie pomocou vlastného softvéru PALS. Výsledkom tohto spracovania bol anonymizovaný dátový súbor obsahujúci medicínsku diagnostiku 820 pacientov v spektre 66 atribútov. Pomocou metódy postupného dopredného výberu sme na základe troch rôznych kritérií (*Adjusted RSq*, *Mallow's Cp*, *BIC*) získali tri skupiny atribútov. Takto vytvorené dátové súbory, spolu so súborom opierajúcim sa o vybrané atribúty na základe skúmania podobných výskumov a súborom obsahujúcim všetky atribúty, sme použili na vypracovanie troch typov modelov: rozhodovacie stromy, Naivný Bayesovský klasifikátor a kNN. Popisovaný výskum sa riadil metodológiou CRISP-DM.

Kľúčové slová: kardiovaskulárne ochorenia, CRISP-DM, datamining

1 Úvod

Ako uvádza Národné centrum zdravotníckych informácií, v roku 2016 boli kardiovaskulárne ochorenia najčastejšou príčinou úmrtnosti na Slovensku, predstavujúcou až 48,2% zo všetkých úmrtí [1]. Hoci do kategórií kardiovaskulárnych ochorení spadá viacero diagnóz, v našej práci sa zameriavame na aterosklerózu v koronárnych cievach a jej následky (progresia aterosklerotických plakov môže vyústiť až do totálneho uzáveru lumenu cievy [2]). Napredujúcu aterosklerózu je však možné včas podchytiť, čím sa znižuje aj riziko trvalých následkov. K tomu nám slúži niekoľko vyšetrení, najčastejšie koronárna angiografia (koronarografia). Vyšetrenie začína zavedením katétra femorálnym alebo radiálnym prístupom (*artéria femoralis* – stehenná tepna alebo *artéria radialis* – vretenná tepna), následne po zavedení vodiča sa urobí kontrolný nástrek koronárneho riečiska za pomoci kontrastnej látky. Pomocou RTG prístroja sú následne kontrolované priechody a prípadné zúženia koronárnych ciev [3]. Keďže ide o invazívny zákrok, okrem finančnej náročnosti so

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 16-19.

sebou prináša aj istú mieru rizika [4]. Z nášho pohľadu je preto dôležité snažiť sa v prvom rade o zníženie miery rizika a to spôsobom obídenia nutnosti podstúpiť koronarografiu a v druhom rade o zníženie finančnej náročnosti.

Popisovanú analýzu vlastného dátového súboru sme vykonali podľa metodológie CRISP-DM v jazyku R v prostredí R Studio.

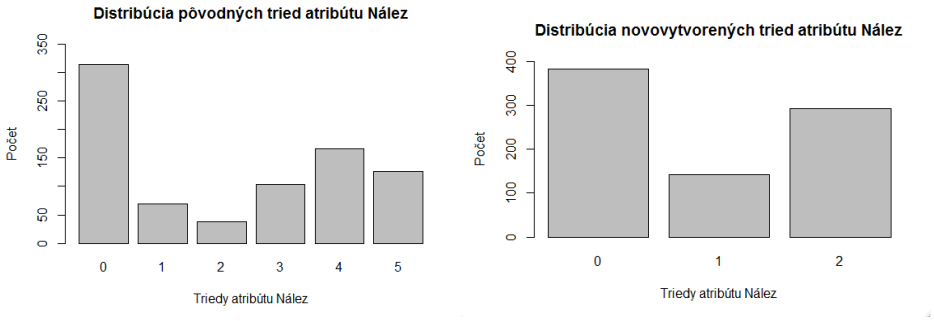
2 Dátový súbor pacientov a jeho príprava pre modelovanie

Náš výskum prebiehal na vlastnej dátovej vzorke, ktorú sme získali za pomoci špecialistov pôsobiacich na 1. Kardiologickej klinike UPJŠ LF a VÚSCH a.s. v Košiciach. Dátová vzorka zahŕňa pacientov hospitalizovaných vo VÚSCHu v období od júna 2017 do marca 2018 v podobe prepúšťacích správ vo formáte *.txt*. Tieto správy boli spracované softvérom PALS [5], ktorý však prešiel miernou úpravou. Súbor obsahujúci extrahované údaje pozostával z 820 pacientov a 66 atribútov popisujúcich osobnú a rodinnú anamnézu, laboratórne výsledky vyšetrení, výsledok EKG vyšetrenia a závery koronarografie [5]. Zastúpenie mužov a žien bolo v pomere 57:43 (467 mužov a 353 žien) vo veku 66.93 ± 9.37.

Jedným z dôležitých krokov prípravy dátového súboru bolo doplnenie chýbajúcich hodnôt. Zamerali sme sa na viacero spôsobov, no spomedzi skúmaných metód ako napr. priemerovania, metódy maximálneho očakávania, či metódy rough set sme zvolili nakoniec metódu k-najbližších susedov, nakoľko bola najúspešnejšia.

Ďalším krokom bola agregácia atribútu *Nalez*, z pôvodných 6 kategórií na 3:

- 0 – bez nálezu, prípadne 10% zúženie koronárnych ciev (pôvodne 0 a 1),
- 1 – zúženie 20 – 70%, povodie RIA (z lat. *Ramus Interventricularis Anterior* - vetva ľavej koronárnej artérie) (20 – 50% zúženie (pôvodne 2 a 3),
- 2 – zúženie nad 70%. RIA nad 50% (pôvodne 4 a 5).



Obr. 1. Porovnanie zmeny atribútu Nález

Výsledkom predspracovania bol súbor obsahujúci 819 záznamov v spektre 65 atribútov.

Nakoľko atribútov bolo väčšie množstvo, považovali sme za potrebné ich určitým spôsobom selektovať. K tomu nám pomohla metóda postupného dopredného výberu (z angl. *Forward Stepwise Selection* - FSS). Prvým kritériom výberu podmnožiny atribútov bol korigovaný koeficient determinácie – *Adjusted RSq*, ktorého maximálna hodnota bola pri počte 32 atribútov (množina atribútov FSS1). Druhým kritériom bola hodnota *Mallow's* C_p pre počet atribútov 23, pričom väčšina atribútov bola takmer totožná s predchádzajúcim výberom (množina atribútov FSS2). Posledným sledovaným kritériom bola hodnota

Bayesovho informačného kritéria *BIC* (množina atribútov FSS3). Požadovanú najnižšiu hodnotu sme získali pri počte atribútov 9.

Pre výber atribútov sme sa inšpirovali aj odbornými publikáciami venujúcimi sa podobnej problematike [6][7][8]. Takto získané skupiny atribútov sme postúpili na spracovanie za účelom vytvorenia modelov.

3 Modelovanie

Vo fáze modelovania sme sa zamerali na niekoľko typov algoritmov: rozhodovacie stromy, Bayesovskú klasifikáciu a metódu k-najbližších susedov.

Pre vytváranie rozhodovacích stromov (RS) sme použili *Conditional Inference Tree*, pričom rozdelenie dátového súboru na trénovaciu a testovaciu množinu bolo v pomere 80:20. Pri vytváraní modelu na základe Naivného Bayesovského klasifikátora (NBK) sme pracovali s rovnakými skupinami atribútov, pričom v niektorých prípadoch sme museli pristúpiť k výberu podskupiny numerických a logických atribútov. Podobne sme postupovali aj pri vytváraní modelov za použitia k-najbližších susedov (kNN). Vyhodnotenie úspešnosti modelov sme vykonali za pomoci kontingenčnej matice. Nasledujúca tabuľka uvádza presnosti vytvorených modelov pre rôzne skupiny atribútov.

Tab. 1. Prehľad úspešnosti vytvorených modelov

Výber atribútov	Úspešnosť klasifikácie		
	RS	NBK	k-NN
Celá množina	98.10%	69.11 %	45.34 %
Množina atribútov podľa [6]	52.91 %	52.87 %	41.86 %
Množina atribútov podľa [7]	52.91 %	50.79 %	40.11 %
Množina atribútov podľa [8]	53.00 %	37. 61 %	40.69 %
FSS1	68.61 %	57.14 %	42.44 %
FSS2	68.61 %	70.70 %	43.60 %
FSS3	70.35 %	71.92 %	69.77 %

4 Vyhodnotenie

Za najúspešnejší z vytvorených modelov z hľadiska miery presnosti považujeme RS pracujúci s celou množinou atribútov. Avšak použitie tohto modelu je veľmi obmedzené, nakoľko sa opiera aj o atribúty popisujúce výsledok nálezu koronarografie. Pre naše potreby však tento model nie je až tak vyhovujúci, nakoľko našou snahou je určiť potrebu absolvovania koronagrafie a teda reálne by sme nemali disponovať atribútmi popisujúcimi nález koronarografie. Zohľadnením týchto skutočností sa nám ako najvhodnejší z vytvorených modelov javí NBK pracujúci s množinou FSS3. Avšak jeho presnosť je veľmi nízka na to, aby sme ho mohli považovať za spoľahlivý a vhodný na určenie výsledku koronagrafie – vyšetrenia, ktoré môže pomôcť pri skvalitnení, prípadne až záchrane ľudského života. Metóda kNN nám poskytla najnižšiu presnosť v spektre použitých množín atribútov. Ako je z tabuľky Tab. 1. zrejmé, z pohľadu kritéria výberu atribútov metódou FSS najlepšie výsledky dosiahol zakaždým model (RS na celej množine nie je vhodným kandidátom) za použitia množiny FSS3, čo zodpovedá kritériu BIC.

5 Záver

V tomto článku popisujeme prácu s vlastným dátovým súborom, ktorý pochádza z 1. Kardiologickej kliniky UPJŠ LF a VÚSCH a.s. v Košiciach. Dátový súbor obsahuje informácie o 467 mužoch a 353 ženách vo veku 66.93 ± 9.37. V rámci prípravy súboru na modelovanie sme okrem doplnenia chýbajúcich hodnôt vytvorili viacero množín atribútov a v ďalšej fáze sme vytvorili viacero klasifikačných modelov. Získané výsledky však nemôžeme považovať za uspokojivé.

V ďalších fázach nášho výskumu sa zameriame na lepšie pochopenie dátového súboru v zmysle ošetrenie korelačných vzťahov, taktiež na vytvorenie nových podmožín atribútov a nových pomerov rozdelenia na trénovaciu a testovaciu množinu, rovnako na zameranie sa na n-násobnú krížovú validáciu, prípadne na ďalšie typy modelov.

Poďakovanie: Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-17-0550.

Literatúra

1. Národné centrum zdravotníckych informácií: Zdravotníctvo Slovenskej republiky v číslach 2016, Január 2018 (Online: http://www.nczisk.sk/Documents/publikacie/analyticke/zdravotnictvo_slovenskej_republiky_v_cislach_2016.pdf).
2. American Heart Association: Atherosclerosis, Apríl 2017 (Online: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/atherosclerosis).
3. Kardiocentrum Nitra: Koronarografia - Koronárna angiografia - Katetrizácia srdca (Onli-ne: http://www.kcnr.sk/pre-pacientov/edukacne-texty/koronarografia.html)
4. Stredoslovenský ústav srdcových a cievnych chorôb: Koronarografické vyšetrenie. (On-line: https://www.suscch.eu/page.php?63)
5. Pella, Z., Milkovič, P., Paralič, J.: Application for Text Processing of Cardiology Medical Records. In: DISA 2018: IEEE World Symposium on Digital Intelligence for Systems and Machines: proceedings, Danver (USA): Institute of Electrical and Electronics En-gineers (2018), 169-174.
6. Hongzeng, X. et al.: Development Of A Diagnosis Model For Coronary Artery Disease. Indian Heart Journal, vol 69, no. 5, Elsevier BV (2017), 634-639, doi: 10.1016/j.ihj.2017.02.022.
7. Verma, L., Srivastavas, S.: A Data Mining Model for Coronary Artery Disease Detection Using Noninvasive Clinical Parameters. Indian Journal of Science and Technology, vol. 9, no. 48 (2016) doi: 10.17485/ijst/2016/v9i48/105707.
8. Tayefi, M. et al.: Hs-CRP Is Strongly Associated With Coronary Heart Disease (CHD): A Data Mining Approach Using Decision Tree Algorithm. Computer Methods And Pro-grams In Biomedicine, vol 141, Elsevier BV (2017), 105-109, doi: 10.1016/j.cmpb.2017.02.001.

Content-aware Collaborative Filtering in Point-of-Interest Recommendation Systems

Samigullina Guzel[1], Jaroslav Kuchář[1]

[1]Faculty of Information Technology, Czech Technical University in Prague
Thákurova 9, 160 00 Prague
{samigguz,jaroslav.kuchar}@fit.cvut.cz

Abstract. With the availability of the vast amount of users and Location-based social networks, the problem of POI recommendations has been widely studied and received significant research attention in the last years. While previous works of POI recommendation mostly focused on investigating the spatial, temporal, and social influence, the use of additional content information has not been directionally studied. In this paper, we propose the content-aware matrix factorization method based on incorporating POI attributes and categories information. We propose two variants of the algorithm that can work with an explicit and implicit feedback. Experimental results show that the proposed method improves the quality of recommendation and outperforms most state-of-the-art collaborative filtering algorithms.

Keywords: POI Recommendation System, matrix factorization, implicit and explicit feedback

1 Introduction

Location-based social networks (LBSNs) have become very popular and attracted lots of attention from internet users, business and academia with the increasing popularity of GPS-enabled mobile devices. Typical location-based social networks include Foursquare, Yelp, Facebook Place, GeoLife, etc. The number of users in those networks is huge, for example, Foursquare had more than 50 million monthly active users on October 2018, and Yelp had about 33 million unique visitors by the end of 2018.

With the availability of the vast amount of users' visiting history, the problem of POI recommendations has been widely studied and received significant research attention in the last years. While previous works of POI recommendation mostly focused on investigating the spatial, temporal, and social influence, the use of additional content information has not been directionally studied. Such additional information can not only improve the performance of the recommendation but also help to overcome the so-called cold start problem.

In this paper, we propose the content-aware matrix factorization method based on incorporating POI attributes and categories information to overcome the cold start item problem, and consequently improve the quality of recommendation. We propose two variants of the algorithm that can work with an explicit and implicit feedback.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 20-23.

2 Algorithm description

As a baseline approach, we use the state-of-the-art matrix factorization method (described in the work [4]), where the regularized objective function is computed only over the observed entries in rating $m \times n$ matrix R. The proposed step is to utilize content information to regularize the matrix factorization. In the work [6], Yu, Wang, and Gao propose to add the following item relationship regularization term based on item attribute information to constrain the baseline matrix factorization framework:

$$\frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}S_{ij}||v_i - v_j||^2$$

where β is the regularization parameter to control the impact from the item attribute information, S is similarity matrix where $S(i,j)$ represents the similarity between items i and j based on their item attribute information, v_i is the ith row of the unknown matrix V that is referred to as an *item factor*.

By adding the item relationship regularization term into Equation (1), the standard metric factorization objective function J changes to:

$$J^* = \min_{U,V}\frac{1}{2}||R - UV^T||^2 + \frac{\lambda}{2}||U||^2 + \frac{\lambda}{2}||V||^2 + \frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}S_{ij}||v_i - v_j||^2$$

where $m \times k$ matrix U and $n \times k$ matrix V are the unknown matrices, which need to be learned to minimize the objective function, $\lambda > 0$ is the regularization parameter that controls the weight of the regularization term.

In datasets each POI is usually represented by a collection of attributes (i.e., WiFi - true/false, price range, parking - true/false, noise level and so on) and categories (i.e., bakery, restaurant, coffee shop), so we will use attributes A and categories C to compute similarity between each pair of POIs. To compute similarity matrix S we construct the following similarity measure:

$$Sim(i,j) = \frac{\sum_{k=1}^{D}\delta(a_{i,k},a_{j,k})}{D} + \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

where D is the number of attributes and $\delta(a_{i,k}, a_{j,k})$ returns 1 if attributes$a_{i,k} = a_{j,k}$ and 0 otherwise. The most popular and effective ways to solve this optimization problem is to use the stochastic gradient descent approach (SGD) that applied to find a local minimum solution of the objective function.

The discussed matrix factorization approach is designed for rating predictions in a typical scenario of recommendation system. In POI recommendation systems user preferences are usually expressed in the form of check-ins, a higher visit frequency corresponds to larger confidence of preference for the location. Due to the unique characteristics of implicit feedback (e.g., lack of negative feedback), it is necessary to design adjusted algorithm. In the work [1], Yifan Hu et al. propose to consider all non-visited locations as negative examples when the weights to all negative examples are assigned the same value, i.e., 1. Thus, the weight matrix W can be defined as follows:

$$w_{ui} = \begin{cases} \mu(c_{ui}^*) + 1, & if c_{ui}^* > 0 \\ 1, & otherwise \end{cases}$$

where $\mu(c_{ui}) > 0$ is a monotonically increasing function and $c_{ui}^* \in \{0,1\}$ is the entry of matrix C^* that indicates whether a user u has visited a POI i. Based on this weighted matrix W, the objective function for the implicit feedback is represented as follows:

$$\min_{U,V} \frac{1}{2}||W \odot (C^* - UV^T)||^2 + \frac{\lambda}{2}||U||^2 + \frac{\lambda}{2}||V||^2 + \frac{\beta}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} S_{ij}||v_i - v_j||^2$$

where $\odot$ is the Hadamard product operator, C^* is 0/1 matrix described above.

This optimization problem can also be solved using the stochastic gradient descent approach. However, due to the weight setting, the approximation error is summed over all entries in the user-POI matrix and stochastic gradient descent algorithm becomes too expensive. Fortunately, the approximate error can be efficiently reduced via Alternative Least Squares algorithm (ALS) and its time complexity for each iteration is in proportion to the total number of visited locations.

3 Datasets

In our experiments we use the public Foursquare[1] and Yelp[2] datasets. Because the Foursquare dataset does not contain attribute information about POIs that is required by our algorithm, we propose to conflate the POIs based on multiple attributes from the Foursquare dataset using the Yelp API. The Foursquare dataset and Yelp API have several overlapping attributes that we used as input to match POI - geographic location, venue name and categories. For all returned items from Yelp API we compute distance, similarity of name and categories using the following weighted model:

$$Sim(a,b) = \lambda_1 \cdot Dist(a,b) + \lambda_2 \cdot Lev_{name}(a,b) + \lambda_3 \cdot Cat(a,b)$$

where a is the target venue in the Foursquare dataset, b is the returned venue from Yelp API, $\lambda_1, \lambda_2, \lambda_3$ are regularization parameters to control the impact of each similarity measure. If venue with the highest similarity value satisfies a pre-specified matching threshold, a match is found (more information can be found in [5]).

The Yelp data containing ratings for POIs will be used by our first content-aware SGD algorithm for the explicit feedback and the extended Foursquare data containing check-in data will be used by the second ALS algorithm for the implicit feedback.

4 Experiments

For each user, we randomly select 30% of her visiting locations as testing data (also referred to as ground truth) to evaluate the performance of different algorithms. The remaining portions from each user constitute a training set for learning the parameters of the proposed model. Based on the training sets, we construct a user-POI rating matrix R and check-in frequency matrix C for Yelp and Foursquare respectively.

[1] Available at https://sites.google.com/site/yangdingqi/home/foursquare-dataset

[2] Yelp dataset challenge Round 13 (access date: March 2019), https://www .yelp.com/dataset/challenge

Also, we explored the impact of control parameters of our algorithms to quality of recommendation and tune them based on the training set to find the optimal values (more information can be found in [5]). Subsequently, we used the best values of parameters during comparison with other approaches. For all involved recommendation algorithms, we set λ= 0.1 and the control parameter β is set to 0.1.

4.1 Comparison with other approaches

To evaluate the performance of the proposed method, we choose the following state-of-the-art collaborative filtering approaches for comparison - RSVD (Regularized SVD approach), WMF (Weighted matrix factorization) [1], SVD++ (an extension of SVD) [2], UCF (User-based Nearest Neighbor algorithm) and GeoMF (state-of-the-art method for POI recommendation) [3]. The results of comparison the above selected recommendation algorithms with our two algorithms for explicit and implicit feedback are plotted in the following figures.

For the first algorithm for explicit feedback, where a user specifies preference using ratings, we use Mean Absolute Error (MAE) to measure the prediction accuracy. In implicit-feedback recommendation methods, the learned model is assessed by its capacity of finding the ground truth locations for each user among the top k ranked locations. So it makes sense to look into rank-based metrics, such as widely used rank-based metrics Precision@k and nDCG@k.

In Fig. 1 you can see, that in terms of MAE, the first algorithm outperforms UCF, RSVD, SVD++, but GeoMF gives the better MAE values on 7.5% than our algorithm. In term of Precision@5, only GeoMF outperforms the second algorithm of implicit feedback by around 7%. As for Precision@10 (Fig. 2), only GeoMF and SVD++ are superior to our second algorithm of implicit feedback by about 12%.

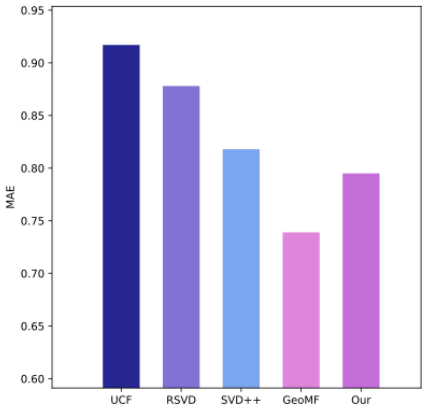


Fig. 1. MAE comparison with the first algorithm

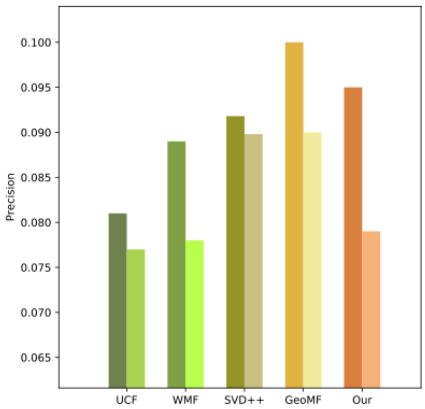


Fig. 2. Comparison of Precision@k with the second algorithm (k=5, 10)

This observation confirms the assumption that the use of POI content information can improve the quality of recommendations. However, almost always GeoMF algorithm is superior to our proposed methods.

4.2 Performance on Cold Start Items

One of disadvantage of the collaborative filtering approach is the so-called cold start problem, which refers to the general difficulty in performing collaborative filtering for users and items that are relatively new. Because we use additional content information for construct similarity matrix between POIs, it can help to deal with the cold start item issue in recommender systems. To evaluate it, we group POIs according to the number of observed ratings and check-ins on POIs in the training set, and then compare the values of metrics of different POIs groups with other selected recommendation algorithms.

The results of the comparison of selected recommendation approaches with our first algorithm are plotted in Fig. 3. You can see, that the first algorithm is able to generate better recommendations than other algorithms when the POIs have few observed ratings (11-30). As more observed ratings are given (>30), the improvement of our proposed approach gradually reduces.

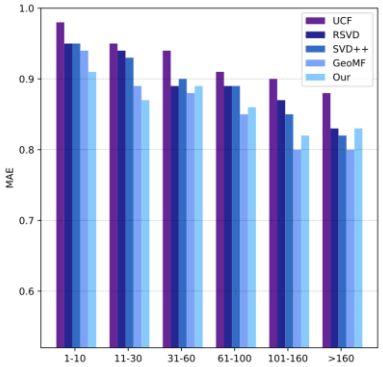


Fig. 3. Comparison of MAE with the first algorithm

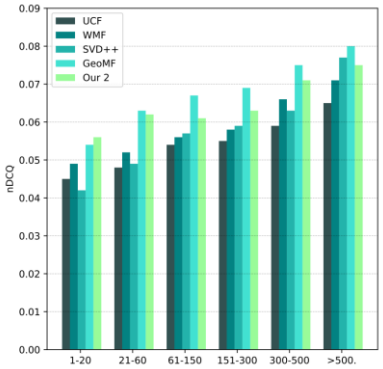


Fig. 4. Comparison of nDCG@5 with the second algorithm

The results of of the comparison of selected algorithms with our second algorithm for implicit feedback are plotted in Fig. 4. The figure shows similar change trends as for the first algorithm.

These observations indicate that our proposed algorithms can cope with cold start item problem and the improvement is directly related to the use of additional content information, such as POI attribute and categories information.

5 Conclusion

Experimental results show that our method improves the quality of recommendation and can effectively cope with the so-called problem. The proposed method outperforms most state-of-the-art collaborative filtering algorithms, only GeoMF algorithm has surpassed our approach. The GeoMF algorithm integrates geographical influence by modeling users' activity regions and the influence propagation on geographical space. Therefore, to improve the recommendation quality it makes sense to consider using the geographical influence on users' check-in or rating behaviors based on the assumption that users tend to visit nearby locations within a radius of activity regions. Furthermore, it would be interesting to investigate the

recommendation effect of content information compared to other information, such as temporal or social information. Also, in our work, we use the simple similarity measure between attributes. Because some of the attributes are in categorical structure, it would be fine to consider some similarity measure, that reflects the relationship between categorical data.

Acknowledgements: This research was supported by Faculty of Informatics, Czech Technical University in Prague.

References

1. Hu, Y., Koren, Y., et al.: Collaborative Filtering for Implicit Feedback Datasets. In: Eighth IEEE International Conference on Data Mining (2008), 263–272.
2. Koren, Y..: Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (2008), 426–434.
3. Lian, D., Zhao, C., et al.: Geomf: Joint geographical modeling and matrix factorization for point-of-interest recommendation. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (2014), 831-840.
4. Paterek, A.: Improving regularized singular value decomposition for collaborative filtering. In: Proc. of KDD cup and workshop (2007), 5–8.
5. Samigullina, G: Algorithms for collaborative filtering in Point-of-Interest Recommendation Systems. Prague, 2019. Diploma Thesis. Czech Technical University in Prague, Faculty of Information Technology. Available from: https://dspace.cvut.cz/bitstream/handle/10467/82612/F8-DP-2019-Samigullina-Guzel-thesis.pdf
6. Yu, Y., Wang, C., et al.: Attributes Coupling based Item Enhanced Matrix Factorization Technique for Recommender Systems. IEEE Transactions on knowledge and data engineering , 2014.

Návrh znalostného modelu v doméne prevádzky IT pre predikciu bezpečnostných incidentov

Martin Sarnovský[1], Pavol Halás[1]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
matin.sarnovsky@tuke.sk, pavol.halas@student.tuke.sk

Abstrakt. Článok sa zameriava na použitie znalostných modelov v prediktívnych úloh v oblasti prevádzky IT. Cieľom bolo navrhnúť model, ktorý by zachytával externé doménovo-špecifické znalosti modelovanej oblasti. Takéto znalosti potom môžu byť z modelu extrahované pri riešení predikcie typu sieťových útokov a môžu napomôcť zlepšovaniu výsledkov analytických modelov používaných v týchto úlohách. Navrhnutý znalostný model pokrýva základné doménové koncepty popisujúce bezpečnostné incidenty a ich charakteristiky vrátane taxonómie sieťových útokov. Model bol implementovaný v štandarde OWL a vyhodnotený pomocou množiny kompetenčných otázok.

Kľúčové slová: ontológie, znalostné modely, prediktívne modelovanie

1 Úvod a motivácia

Ontológia je termín, ktorý sa používa na označenie zdieľaných znalostí v určitej oblasti záujmu, ktorý môže byť použitý ako jednotný rámec pre riešenie mnohých problémov [1]. Pomáha zlepšovať komunikáciu medzi ľuďmi, organizáciami a systémami pomocou definovania jednotných pojmov a ich vzájomných vzťahov. Medzi hlavné výhody ontológií patrí zdieľanie znalostí a možnosť opätovného použitia nadobudnutých znalostí. Poskytuje určitý pohľad na danú doménu pomocou skupiny konceptov (napr. entity, atribúty, procesy), ich definícií a vnútorných vzťahov.
Oblasť detekcie sieťových útokov je čoraz dôležitejšia pre bezpečnosť webových aplikácií, pretože chráni komunikáciu a citlivé údaje miliónov užívateľov. Tradičné bezpečnostné riešenia poskytujú prvú líniu obrany proti útokom a detegujú známe bezpečnostné nedostatky. Chýba im sémantika, preto nie sú schopné rozoznať nové a kritické chyby [2]. Systémy pre detekciu sieťových útokov by mali obsahovať aj sémantickú časť, ktorá by vedela rozoznať kontextový charakter útoku.
Kľúčovou úlohou práce popísanej v tomto článku bolo vytvorenie znalostného modelu domény detekcie sieťových útokov, ktorý by mohol byť využívaný v prediktívnych analytických úlohách na rôzne účely, napr. pre generalizovanie alebo špecifikáciu cieľového atribútu alebo pre zlepšenie prediktívnych modelov pomocou odvodenia doménovo-špecifických znalostí, ktoré nemusia byť explicitne obsiahnuté v modelovaných dátach. V článku sa zameriame na predstavenie štruktúry takéhoto

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 26-29.

znalostného modelu a predstavíme si niektoré z možnosti jeho použitia pri úlohách predikcie typu sieťového útoku v dátach z prevádzky IT infraštruktúry.

2 Využitie ontológií pri detekcii bezpečnostných incidentov

Využitiu sémantických modelov v prediktívnych úlohách v oblasti detekcie sieťových útokov sa venovalo v minulosti viacero rôznych prác. V [3] sa autori zamerali na vytvorenie ontológie pomocou jazyka DAML+OIL, ktorá charakterizovala DoS (Denial of Service) typ útokov. Ontológia slúžila ako základ systému detekcie útokov, pričom využívali rámec Jena pre interakciu ontológie a agentov. Týmto spôsobom mohli extrahovať sémantický vzťah všetkých útokov z ontológie. Výsledky pri predikcii DoS útokov dokazujú, že použitím ontológie sa znížil počet falošne pozitívnych a falošne negatívnych nahlásení útoku oproti ostatným systémom pre detekciu útokov, ktoré využívali rôzne klasifikátory (napr. boosting a iné). Autori v práci [4] použili rovnako jazyk DAML+OIL na vytvorenie ontológie popisujúcej sieťové útoky zaznamenaná v datasete KDD99 [5]. Výsledky modelu založenom na ontológii boli vo všetkých kategóriách (presnosť modelu, miera falošných nahlásení útoku atď.) lepšie ako ostatné, štandardné klasifikačné modely. V [6] je popísaná ontológia implementovaná v jazyku OWL (Web Ontology Language), ktorá charakterizuje 3 vybrané útoky z rovnakého datasetu. Podobne v predošlých dvoch prístupoch, svoje výsledky systému založeného na ontológii porovnával vo viacerých kategóriách s modelmi, ktoré využívali algoritmy strojového učenia a dolovania dát. Spomínané publikácie demonštrujú niekoľko z možností využitia sémantických modelov v systémoch pre detekciu počítačových útokov.

3 Návrh znalostného modelu

Znalostný model popísaný v tomto článku je ontológia, ktorú sme vytvárali podľa metodológie Grüninger a Fox [7] a z časti aj podľa metodológie Methontology [8]. Ako formalizmus pre našu ontológiu sme si zvolili jazyk OWL, pretože je jedným zo súčasných popredných jazykov pre tvorbu ontológií, ktorý ponúka široké možnosti v oblasti sémantiky. V nasledujúcich podkapitolách si predstavíme hlavné stavebné bloky navrhovanej ontológie.

3.1 Štruktúra modelu

Pri vytváraní znalostného modelu sme sa snažili vytvoriť takú štruktúru tried, ktorá by podchytila všetky podstatné javy domény systémov pre detekciu sieťových útokov. Medzi hlavné triedy ontológie patrí:

- *Connections* - trieda reprezentuje stav jednotlivých záznamov pripojení a pozostáva z viacerých podtried, ktoré vytvárajú hierarchiu. Trieda sa ďalej delí na triedu *Attacks* reprezentujúcu útoky a triedu *Normal*, ktorá hovorí, že dané pripojenie je v bezpečí. Trieda *Attacks* sa delí na 4 podtriedy, ktoré reprezentujú

hlavné druhy útokov (napr. DoS útoky). Hierarchia bola vytváraná podľa popisu úloh datasetu KDD99.

- *Effects* - trieda obsahuje podtriedy, ktoré zastupujú všetky možné následky útokov (napr. spomalenie odozvy servera, vykonávanie príkazov pod root prístupom atď.)
- *Mechanisms* - podtriedy zastupujú všetky možné príčiny jednotlivých útokov ontológie (slabá údržba prostredia, zlá konfigurácia atď.)
- *Flags* - zastupuje normálne alebo chybové stavy jednotlivých pripojení (neodpovedanie služby, pokus o pripojenie do siete bol zamietnutý atď.)
- *Protocols* - reprezentuje typy protokolov na ktorých beží pripojenie (TCP, UDP a ICMP)
- *Services* - reprezentuje jednotlivé typy služieb pripojenia (http, telnet atď..)
- *Severities* - reprezentuje závažnosť daného útoku, jej podtriedy zastupujú úroveň závažnosti (slabá, stredná a vysoká).
- *Targets* - reprezentuje možné ciele daného druhu útoku (používateľ, sieť).
- *Models* - reprezentuje prediktívne modely v danej doméne

V navrhovanom modeli sú vytvorené dva typy vzťahov: objektové vlastnosti a vlastnosti dátového typu. V nasledujúcej tabuľke (Tab.1) sú znázornené všetky objektové vlastnosti s ich doménami a rozsahom.

Tab. 1. Objektové vlastnosti sémantického modelu

Názov vlastnosti	Doména	Rozsah
hasSeverity	Connections	Severity
hasProtocol	Connections	Protocols
hasMechanism	Connections	Mechanisms
hasFlag	Connections	Flags
hasService	Connections	Services
hasEffect	Connections	Effects
hasTarget	Connections	Targets
hasGranularityLevel	Models	Connections
providesService	Protocols	Services

3.2 Vyhodnotenie modelu

Pred samotným vytvorením ontológie sme definovali zoznam kompetenčných otázok podľa metodológie Grüninger a Fox, ktoré tak definovali požiadavky na znalostný model. Následne sme tieto kompetenčné otázky preformulovali z neformálnej podoby do dopytovacieho jazyka. Cieľom bolo demonštrovať, že navrhnutá ontológia je vhodná na reprezentáciu požadovaných znalostí a tieto je možné z nej pomocou dopytovacieho jazyka vyhľadať. Nasledujúci príklad (Tab. 2) demonštruje ukážku kompetenčnej otázky a príslušného dopytu do ontológie v jazyku SPARQL. Celkovo sme sformulovali 10 kompetenčných otázok, ktoré sme dokázali vyjadriť pomocou SPARQL dopytu.

Tab. 2. Ukážka vyhodnotenia modelu pomocou kompetenčných otázok a SPARQL dopytov

Kompetenčná otázka	SPARQL dopyt
Na akom protokole sa vyskytujú pripojenia s útokom *Land* a aké sú jeho príčiny, následky a závažnosť ?	SELECT ?object ?subject WHERE { on:Land rdfs:subClassOf ?obj. ?obj owl:onProperty ?object; owl:someValuesFrom ?subject.}

4 Záver

V článku sme predstavili návrh znalostného modelu pre doménu sieťových útokov. Cieľom bolo navrhnúť model pokrývajúci koncepty domény detekcie sieťových útokov vrátane aspektov použiteľných pri prediktívnych úlohách ich detekcie. Navrhnutý model je možné použiť vo viacerých scenároch analytických úloh zaoberajúcich sa predikciou bezpečnostných útokov. Napríklad je možné využiť taxonómiu typov útokov pri predikcii konkrétnych druhov útokov, či jej využitie pre predikciu typu útoku na základe podobnosti existujúcich inštancií. V budúcnosti vidíme možnosti rozšírenia modelu o koncepty popisujúce preventívne opatrenia pre jednotlivé sieťové útoky, prípadne využitie ontológie v úlohách hľadania sekvencií a častých vzorov v dátach prevádzky IT prostredia.

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

Literatúra

1. Uschold, M., Gruninger, M.: Ontologies: principles, methods and applications. The Knowledge Engineering Review. 11, 93–136 (1996).
2. Godse, M.,et.al: Semantic Host Based Intrusion Detection. International Journal of Software Engineering and its Applications. 9, 167–173 (2015).
3. Abdoli, F., Kahani, M.: Ontology-based distributed intrusion detection system. In: 2009 14th International CSI Computer Conference, CSICC 2009. pp. 65–70 (2009).
4. Hung, S.-S., Liu, D.S.-M.: A User-centric Intrusion Detection System by Using Ontology Approach. Presented at the (2007).
5. Georges, J., Milley, A.H.: KDD'99 competition. ACM SIGKDD Explorations Newsletter. 1, 79 (2007).
6. Khairkar, A.D., Kshirsagar, D.D., Kumar, S.: Ontology for detection of web attacks. In: Proceedings - 2013 International Conference on Communication Systems and Network Technologies, CSNT 2013. pp. 612–615 (2013).
7. Gruninger, M., Fox, M.S.: Methodology for the Design and Evaluation of Ontologies. In: Industrial Engineering. pp. 1–10 (1995).
8. Mariano Ferndndez, Asuncion Gomez-Perez, N.J.: METHONTOLOGY: From Ontological Art Towards Ontological Engineering. In: Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series. pp. 115–122 (1997).

Analýza a syntéza dát pre určenie rizika vzniku kardiovaskulárneho ochorenia

Zuzana Pella[1], Karin Jana Szilárdy[2], Ján Paralič[1], Dominik Pella[3]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
{zuzana.pella,jan.paralic}@tuke.sk
[2]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
karin.jana.szilardy@student.tuke.sk
[3]1. Kardiologická klinika UPJŠ LF a VÚSCH a.s.
Ondavská 8, 040 11 Košice
dominik.pella@gmail.com

Abstrakt. Vďaka svojej závažnosti a stúpajúcej miere úmrtí následkom kardiovaskulárnych ochorení sú tieto stále vo väčšej miere objektom záujmu mnohých výskumníkov. Opakovane sa vykonávajú analýzy na voľne dostupných dátových súboroch, pričom autori týchto výskumov sa snažia o nový príspevok z hľadiska preukázania kauzality vzťahov medzi rôznymi atribútmi vedúcimi k nepriaznivému zdravotnému stavu pacientov, či o úpravu rôznych dataminingových algoritmov so snahou vylepšenia predchádzajúcich výsledkov. Naša práca sa zameriava na zistenie vplyvu syntézy dvoch často používaných voľne dostupných dátových súborov na základe ich podobnosti, predpokladajúc že rozšírenie dátovej vzorky by mohlo viesť k lepším výsledkom analýzy týchto dát. Celá naša práca sa riadi metodológiou CRISP-DM, kde v odpovedajúcich fázach vytvárame spojený dátový súbor, následne okrem úplnej množiny dostupných atribútov hľadáme aj ich vhodné redukcie. Spojený dataset s niekoľkými výbermi atribútov sú následne použité vo fáze modelovania za účelom vytvorenia modelov pomocou algoritmov rozhodovacieho stromu *Conditional Inference Trees* (ctree), Naivného Bayesa (NB), metódy *Support Vector Machines* (SVM) a k-najbližších susedov (kNN). Naše očakávania o zlepšení kvality výsledných klasifikačných modelov na spojenom dátovom súbore sa však nepotvrdili.

Kľúčové slová: Clevelenad dataset, Z-Alizadeh Sani dataset, CRISP-DM, modelovanie

1 Úvod

V pomyselnom rebríčku celosvetových prvenstiev príčin úmrtí zaberajú kardiovaskulárne ochorenie (KVO) prvé miesto [1]. Premietnutím do čísel, ročne

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 30-34.

zomrie na KVO 610 tisíc ľudí v Amerike [2]. Na Slovensku to bolo v roku 2016 až 48,2% zo všetkých úmrtí [3].

Potvrdenie alebo vyvrátenie prítomnosti KVO sa uskutočňuje pomocou dvoch typov vyšetrovacích metód: invazívnych a neinvazívnych [4]. Najpresnejšou metódou je koronárna angiografia, typ invazívneho vyšetrenia, ktorý však so sebou prináša viacero rizík [5].

Alarmujúce čísla úmrtí následkom KVO a rizikovosť vyšetrenia sú dôvodom, prečo sa tak mnoho výskumníkov z oblasti zdravotníctva či informatických vied zaoberá touto problematikou [6,7,8,9,10,11]. Zameriavajú sa na 2 najrozšírenejšie dátové súbory: Cleveland [12] a Z-Alizadeh Sani [13]. Na tieto sa zameriavame aj my, ale z iného uhla. Našou snahou bolo vytvorenie jedného súboru na základe podobných atribútov a následne ho podrobiť dátovej analýze, pričom sme sa riadili metodológiou CRISP-DM. Očakávame, že spojenie dátových súborov môže okrem zvýšenia presnosti predikcie prítomnosti ischemickej choroby srdca taktiež prinajmenšom nový pohľad na výber atribútov, ktoré participujú na prítomnosti spomínaného ochorenia.

2 Práca s dátovými súbormi Cleveland a Z-Alizadeh Sani

Cieľom analýz nad dátovými súbormi Cleveland a Z-Alizadeh Sani bolo predikovať s čo najvyššou presnosťou, či pacient trpí alebo netrpí ischemickou chorobou srdca (ICHS, jeden z typov KVO).

Údaje sme získali z portálu UCI Machine Learning Repository. Dátový súbor Cleveland zahŕňa údaje o 303 pacientoch v spektre 14 atribútov, pričom cieľový atribút je kategorický a odzrkadľuje závažnosť ICHS (hodnoty 0-4). Z-Alizadeh Sani dataset obsahuje zhodne 303 záznamov, avšak atribútov je až 54 a binárny cieľový atribút definuje pacienta z pohľadu prítomnosti alebo absencie ICHS. Porovnaním dátových súborov, vo fáze pochopenia dát, sme zistili, že majú 4 totožné atribúty a ďalších 9 atribútov, vrátane cieľového, bolo podobných. Z podobných atribútov sme vytvorili nové, vhodným zjednotením hodnôt atribútov. Jednotlivé atribúty a ich úpravy boli konzultované s medicínskym expertom. Ako príklad uvádzame zjednotenie atribútu popisujúceho typickú bolesť hrudníka. V súbore Cleveland bol tento atribút reprezentovaný hodnotami 1-4, pričom hodnota 1 značila prítomnosť symptómu a ostatné hodnoty absenciu symptómu. V súbore Z-Alizadeh Sani bol tento atribút binárneho charakteru, teda hodnota 0 značila absenciu a hodnota 1 prítomnosť symptómu. Výsledkom zjednotenia bol binárny atribút odpovedajúci použitiu v súbore Z-Alizadeh Sani.

Dátový súbor vytvorený za pomoci syntézy obsahoval 606 záznamov, ktoré zobrazoval pomocou 13 atribútov. Pre možnosť väčšej variability množín dát použitých vo fáze modelovania sme sa rozhodli na náš súbor aplikovať metódu postupného dopredného výberu (z ang. *Forward Stepwise Selection*), čím sme získali spolu s cieľovým atribútom redukovanú množinu 10 atribútov.

Vo fáze modelovania sme sa zamerali na použitie štyroch rôznych algoritmov, ktorých výber sa zakladal na analýze najčastejšie používaných algoritmov pre analýzu medicínskych dát. Tieto algoritmy boli použité na dátové súbory Cleveland a Z-Alizadeh Sani, syntézou vytvorený dátový súbor pozostávajúci zo všetkých atribútov a syntézou vytvorený dátový súbor s redukovanou množinou atribútov. Prehľad

výsledkov použitých algoritmov na týchto dátových súboroch ponúkame v nasledujúcej tabuľke (Tab. 1).

V tabuľke uvádzame pre každú množinu najvyššiu hodnotu presnosti, špecificity a citlivosti tučným písmom, naopak najnižšiu hodnotu uvádzame kurzívou. V tabuľke sa taktiež nachádzajú podčiarknuté hodnoty, ktoré poukazujú na najvyššie dosiahnuté hodnoty presnosti, špecificity a citlivosti na spojenom dátovom súbore, či už pre celú alebo redukovanú množinu atribútov.

Tab. 1. Prehľad úspešnosti modelov nad dátovými súbormi

Modely	Cleveland Heart Disease	Z-Alizadeh Sani	Syntézou vytvorený dátový súbor (všetky atribúty)	Syntézou vytvorený dátový súbor (redukovaná množina atribútov)
Presnosť modelov na testovacej množine				
Naivný Bayes	*83,83%*	**88,45%**	72,61%	74,92%
SVM	79,03%	83,87%	76,72%	73,28%
ctree	75,81%	82,25%	68,97%	73,28%
3-NN	65,93%	54,95%	64,57%	73,71%
5-kNN	65,93%	61,54%	66,29%	74,29%
10-kNN	59,34%	63,74%	63,43%	73,71%
Špecificita modelov na testovacej množine				
Naivný Bayes	87,80%	88,43%	68,92%	71,71%
SVM	*89,66%*	**92,68%**	70,83%	66,67%
ctree	*89,66%*	87,80%	56,25%	66.67%
3-NN	70,21%	16,67%	55,56%	63,89%
5-kNN	68,09%	10,00%	51,39%	63,89%
10-kNN	65,96%	10,00%	50,00%	63,89%
Citlivosť modelov na testovacej množine				
Naivný Bayes	79,14%	88,51%	75,21%	77,18%
SVM	69,70%	66,67%	80,88%	77,94%
ctree	63,64%	71,43%	80,00%	77,94%
3-NN	61,36%	73,77%	70,87%	80,58%
5-kNN	63,64%	86,89%	76,70%	81,55%
10-kNN	*52,27%*	**90,16%**	72,82%	80,58%

Najlepšie výsledky v podobe priemerných hodnôt nad rôznymi dátovými súbormi a množinami dosahoval algoritmus Naivného Bayesa s priemernou presnosťou 79,95%. Navyše rozdiel medzi citlivosťou a špecificitou bol pri modeloch na báze tohto algoritmu najmenší zo všetkých porovnávaných algoritmov.

Podobným spôsobom sme porovnali aj úspešnosť modelov nad jednotlivými dátovými súbormi a množinami atribútov. Musíme však konštatovať, že spojenie dátových súborov nám neprinieslo zlepšenie presnosti modelov oproti modelom nad pôvodnými dátovými súbormi samostatne, čo je pravdepodobne dané veľkým rozdielom v počte atribútov u spájaných dátových súborov (14 vs. 54), čo umožnilo

využiť iba menšiu časť informácií dostupných v prípade datasetu Z-Alizadeh Sani. Ak sa však pozrieme už iba na efekt použitia metódy FSS na redukciu počtu atribútov, ten bol vždy pozitívny (t.j. viedol k zvýšeniu všetkých sledovaných parametrov) v prípade všetkých testovaných algoritmov okrem SVM. Najvýraznejšie zlepšenia sme pozorovali v prípade kNN.

3 Záver

Predkladaný príspevok popisuje prácu nad voľne dostupnými dátovými súbormi Cleveland a Z-Alizadeh Sani. Na tieto a taktiež na syntézou z nich vytvorený dátový súbor sme aplikovali 4 algoritmy dolovania dát s cieľom identifikovať prítomnosť ischemickej choroby srdca u pozorovaných pacientov. Získané výsledky nenaplnili naše očakávania ohľadom zvýšenia úspešnosti modelov na spojenom dátovom súbore.

Poďakovanie: Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-17-0550.

Literatúra

1. World Health Organiyation: Cardiovascular diseases (CVDs), Máj 2017 (Online: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
2. Centers for Disease Control and Prevention: Heart Disease – Heart Disease Statistics and Maps, November 2017 (Online: https://www.cdc.gov/heartdisease/facts.htm).
3. Národné centrum zdravotníckych informácií: Zdravotníctvo Slovenskej republiky v číslach 2016, Január 2018 (Online: http://www.nczisk.sk/Documents/publikacie/analyticke/zdravotnictvo_slovenskej_republiky_v_cislach_2016.pdf).
4. Vargová, V., Fedačko, J.: Ischemická choroba srdca. Národný portál zdravia (2016) (Online: https://www.npz.sk/sites/npz/Stranky/NpzArticles/2013_06/Ischemicka_choroba_srdca.aspx?did=4&sdid=31&tuid=0&page=4&).
5. Mayo Clinic: Coronary angiogram (Online: https://www.mayoclinic.org/tests-procedures/coronary-angiogram/about/pac-20384904).
6. Yadav, C., Lade, S., K Suman, M.: Predictive Analysis for the Diagnosis of Coronary Artery Disease using Association Rule Mining. International Journal of Computer Applications 87(4) (2014), 9-13.
7. Verma, L., Srivastava, S., Negi, P.C.: A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. Journal of Medical Systems 40 (7) (2016).
8. El-Bialy, R. et al.: Feature Analysis of Coronary Artery Heart Disease Data Sets. Procedia Computer Science 65 (2015), 459-468.
9. Sani, Z. et al.: Diagnosing coronary artery disease via data mining algorithms by considering laboratory and echocardiography features. Research in Cardiovascular Medicine 2 (3) (2013).

10. Purushottam, Saxena, K., Sharma, R.: Efficient Heart Disease Prediction System. Procedia Computer Science 85 (2016), 962 – 969.
11. Rajeswari, K., Vaithiyanathan, V., Neelakantan, T.R.: Feature Selection in Ischemic Heart Disease Identification using Feed Forward Neural Networks. Procedia Engineering 41 (2012), 1818-1823.
12. UCI Machine Learning Repository: Heart Disease Data Set. Archive.ics.uci.edu (Online: https://archive.ics.uci.edu/ml/datasets/heart+Disease).
13. UCI Machine Learning Repository: Z-Alizadeh Sani Data Set. Archive.ics.uci.edu (Online: https://archive.ics.uci.edu/ml/datasets/Z-Alizadeh+Sani).

Klasifikácia rádiových galaxií metódami hlbokého učenia

Viera Malej Krešňáková[1], Kristián Hai Le Thanh[2], Eduard Pizur[2], Peter Butka[1]

[1,2] Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
[1]{viera.maslej.kresnakova, peter.butka}@tuke.sk
[2]{hai.kristian.le.thanh, eduard.pizur}@student.tuke.sk

Abstrakt. V tomto príspevku prinášame klasifikáciu kompaktných a rozšírených rádiových galaxií. Klasifikujeme tak galaxie podľa morfológie do 4 tried. Na klasifikáciu sme použili hlboké učenie, konkrétne konvolučné neurónové siete. Najvyššou dosiahnutou presnosťou (98 %) sme prekonali doteraz publikované výsledky.

Kľúčové slová: astronómia, rádiové galaxie, hlboké učenie, konvolučné neurónové siete

1 Úvod

Rádioastronómia vznikla v roku 1932, zásluhou Američana českého pôvodu Karla Janskeho [1], ktorý ako prvý objavil rádiové žiarenie prichádzajúce z kozmu. Ďalej nasledoval objav za objavom ako napríklad žiarenie rozsiahlych vodíkových mračien na vlnovej dĺžke 21 cm, organickej molekuly hydroxilu na 18 cm, objav pulzarov či jeden z najdôležitejších objavov - fluktuácie reliktového žiarenia. Za necelé trištvrte storočia sa rádioastronómia stala jedným z najdôležitejších spôsobov poznávania vesmíru. Atmosféra rádiové lúče prepúšťa, takže je možné stavať obrovské rádiové detektory priamo na zemskom povrchu. S príchodom nových observatórií ako **Square Kilometer Array** [2], získavame veľké množstvo dát obsahujúcich rádiové galaxie. Klasifikácia galaxií podľa tried je dôležitý krok, pretože rôzne triedy rádiových galaxií nám umožňujú pochopiť vznik a vývoj galaxií, ich pod-zložiek ako svetelnosť, hmotnosť či rýchlosť tvorby hviezd alebo môžu byť použité ako indikátory kozmického prostredia. Tradične sú tieto galaxie klasifikované ako Fanaroff- Riley (FR) typu I alebo II a BENT [3]. Tieto typy nazývame rozšírené (extended radio galaxies). V tomto príspevku prinášame aj klasifikáciu kompaktných rádiových galaxií. Klasifikujeme tak galaxie podľa morfológie do 4 tried a to COMP (kompaktné) alebo rozšírené (FRI, FRII, BENT).

Tradične sa triedy FR identifikovali vizuálnou kontrolou. Vznikli katalógy ako NRAO VLA Sky Survey (NVSS) [4], Sydney University Molonglo Sky Survey (SUMSS) [5] a the Faint Images of the Radio Sky at Twenty-Centimeters (FIRST) [6], ktoré obsahujú viac ako 2 milióny objektov. Veľkú zásluhu na anotovaní má taktiež

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 35-38.

občiansky projekt Radio Galaxy Zoo [7]. Takýto objem anotovaných dát inšpiroval vedcov k vytvoreniu automatickej klasifikácie objektov.

2 Klasifikácia rádiových galaxií vs. hlboké učenie

Konvolučné neurónové siete [8] sú dopredné neurónové siete, ktoré sa úspešné používajú na analýzu obrazu a textu. Pozostávajú z konvolučných a vzorkovacích vrstiev. Konvolučná vrstva obsahuje množinu filtrov, ktoré sa postupne aplikujú na obrázok a vzorkovacia vrstva sa používa na nelineárne zmenšenie výstupu konvolúcie. V praxi to znamená, že počas trénovania si sieť na jednotlivých vrstvách vytvára rôzne stupne abstrakcie vstupu. Nižšie vrstvy odhaľujú základné črty, pričom ďalšie vrstvy sa viac špecializujú a vytvárajú zložité koncepty. V článku [9] sme poukázali na použitie metód hlbokého učenia na analýzu dát v astrofyzike. Pomocou hlbokého učenie a veľkého množstva anotovaných obrázkov sa podarilo úspešne klasifikovať aj rádiové galaxie [10] [11] [12].

V príspevku [10] autori uvádzajú použitie metódy transferového učenia (transfer learning*)* na klasifikáciu rádiových galaxií. Autori tu využili konkrétne 13 vrstvovú konvolučnú neurónovú sieť, kde dosiahli 89% presnosť klasifikácie pri testovaní dát z katalógu FIRST. V tejto štúdií klasifikovali len galaxie typu FRI a FRII z katalógu FIRST a NVSS. Klasifikátor pre 4 triedy (COMP, FRI, FRII a BENT) použili prvý krát až v štúdií [11], kde pomocou jednoduchej neurónovej siete dosiahli priemernú presnosť 97 %. V práci [12] klasifikovali rádiové galaxie do troch tried (FRI, FRII a BENT), kde autori na základe zloženého klasifikátora (finálna klasifikácia je výsledok troch binárnych klasifikátorov) dosiahli priemernú presnosť modelu 88 %. Všetky výsledky vyjadrené ako presnosť, úspešnosť, návratnosť a F1 skóre sú uvedené v Tab. .

3 Experimenty

Náš príspevok prináša klasifikáciu rádiových galaxií podľa morfológie na kompaktné (COMPT) a rozšírené galaxie (FRI, FRII a BENT). Na trénovanie sme použili novú dátovú množinu z katalógu FIRST, ktorá pozostávala dohromady z 4 576 obrázkov, ktoré boli rozdelené do štyroch tried. Trieda BENT obsahovala 1 247 obrázkov, COMPT (1 115) , FRI (1 109) a FRII (1 105).

Prvé experimenty viedli k použitiu *transferového učenia.* Na základe výsledkov detekcie spájania galaxií [12], sme sa rozhodli na klasifikáciu rádiových galaxií použiť najúspešnejšiu sieť – *ResNet18* [13]. Táto sieť bola pôvodne vytvorená na klasifikáciu obrázkov z bežného sveta na dátovej množine *ImageNet* [14], kde sa klasifikovali obrázky do 1000 tried. Najvyššiu presnosť pri použití *ResNet18* sme na testovacích dátach (457 obrázkov) získali cez 98 %. Podrobnejšie výsledky sú uvedené v tabuľke Tab.1.

Ďalšia časť experimentov sa zameriava na učenie od začiatku. Vytvorili sme jednoduchšiu konvolučnú neurónovú sieť, ktorá na extrakciu príznakov využíva tri

konvolučné vrstvy s aktivačnou funkciou *ReLu*, *Max Pooling* a *Dropout*, Klasifikáciu do 4 tried zabezpečuje *Softmax* funkcia, pričom výsledná priemerná presnosť klasifikácie na testovacích dátach je 96 %.

Tab. 1. Porovnanie výsledkov klasifikácie rádiových galaxií

	presnosť	**návratnosť**	**F1 skóre**	**počet**
Výsledky našej metódy – transferové učenie				
COMPT	1,00	1,00	1,00	111
BENT	0,98	0,98	0,98	125
FRI	0,99	0,97	0,96	111
FRII	0,96	0,97	0,96	110
Priemer	**0,98**	**0,98**	**0,98**	457
Výsledky našej metódy – učenie od začiatku				
COMPT	0,97	0,99	0,98	112
BENT	0,93	0,96	0,95	126
FRI	0,99	0,92	0,95	112
FRII	0,94	0,95	0,95	111
Priemer	0,96	0,96	0,96	461
Výsledky dosiahnuté v [10] – učenie od začiatku				
COMPT	0,98	0,98	0,98	1000
BENT	0,96	0,98	0,97	1000
FRI	0,98	1,00	0,99	1000
FRII	0,96	0,93	0,95	1000
Priemer	0,97	0,97	0,97	4000
Výsledky dosiahnuté v [9] – transferové učenie				
FRI	0,95	0,85	0,90	80
FRII	0,83	0,94	0,88	117
Priemer	0,89	0,89	0,89	197
Výsledky dosiahnuté v [11] – učenie od začiatku				
BENT	0,95	0,79	0,87	77
FRI	0,91	0,91	0,91	53
FRII	0,75	0,91	0,83	57
Priemer	0,88	0,86	0,86	187

4 Záver

Podarilo sa nám klasifikovať rádiové galaxie do 4 tried s 98% presnosťou, čím sme prekonali doteraz publikované výsledky. Tento model bol vytvorený metódou

transferového učenia, kde sme modifikovali sieť *ResNet18*. Pri vytváraní modelu od začiatku, sme získali priemernú presnosť 96 %. Tieto výsledky sú porovnateľné s doteraz publikovanými prácami v tejto oblasti.

Poďakovanie: Táto práca bola podporená VEGA grantom č. 1/0493/16 a APVV grantom APVV-16-0213.

Literatúra

1. F.Ghigo, „Karl Jansky and the Discovery of Cosmic Radio Waves," 2008. [Online]. Available: www.nrao.edu.
2. „The SKA project," 20 06 2019. [Online]. Available: www.skatelescope.org.
3. B. L. Fanaroff J. M. Riley, „The Morphology of Extragalactic Radio Sources of High and Low Luminosity," MNRAS, %1. vyd.167, pp. 31-32, 1974.
4. J. J. Condon, W. D. Cotton, et„The NRAO VLA Sky Survey," The American Astronomical Society, zv. 115, %1. vyd.5, pp. 1693-1716, 1998.
5. D. C.-J. Bock, M. I. Large, Elaine M. Sadler, „SUA Wide-Field Radio Imaging Survey of the Southern Sky. I. Science goals, survey design and instrumentation," The Astronomical Journal), %1. vyd.117, p. 1578, 1999.
6. Becker, Roberth H.; White,Richard L.; Helfand, David J., „The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters," Astrophysical Journal, %1. vyd.450, p. 559, 1995.
7. J. K. Banfield O. I. Wong K. W. Willett R. et al., „Radio Galaxy Zoo: host galaxies and radio morphologies derived from visual inspection," MNRAS, zv. 3, %1. vyd.453, pp. 2326-2340, 2015.
8. Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.
9. V. M. Krešňáková, „Overview of deep learning methods applied for data analysis in astrophysics," SCYR, pp. 127-130, 2019.
10. H. Tang; A. M. M. Scaife; J. P. Leah, „Transfer learning for radio galaxy classification," MNRAS, 2019.
11. Wathela Alhassan A R Taylor Mattia Vaccari, „The FIRST Classifier: compact and extended radio galaxy classification using deep Convolutional Neural Networks," MNRAS, zv. 2, %1. vyd.480, pp. 2085-2093, 2018.
12. Arun Aniyan, Kshitij Thorat, „Classifying Radio Galaxies with Convolutional Neural Network," The asttrophysucal journal, p. 15, 2017.
13. Eduard Pizur, Viera Maslej Krešňáková, Peter Butka, „Detekcia spájania galaxií pomocou metód hlbokého učenia," Electrical Engineering and Informatics 10, 2019.
14. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, „Deep Residual Learning for Image Recognition," Computer Vision and Pattern Recognition, 2015.
15. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, „ImageNet: A Large-Scale Hierarchical Image Database," IEEE Computer Vision and Pattern Recognition (CVPR), 2009.

Modelovanie témy nad dátami s multimodálnym obsahom

Miroslav Smatana[1], Peter Butka[1], Patrícia Kočiščáková[1]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
{miroslav.smatana, peter.butka}@tuke.sk,
patricia.kociscakova@student.tuke.sk

Abstrakt. V súčasnosti s rozširovaním využitia výpočtovej techniky rastie aj množstvo dát v digitálnej forme. S nástupom sociálnych médií sa rozširuje škála rôznych typov dát. Ide o multimodálny obsah, ktorý je tvorený nie len textami, ale aj obrázkami, videami alebo zvukovými nahrávkami. Tieto dáta obsahujú množstvo užitočných informácií, ktoré môžu byť potencionálne užitočné napr. v procese rozhodovania spoločnosti. Existuje široká škála metód analýzy takýchto dát za účelom extrakcie užitočných informácii. V tejto práci sme sa zamerali na analýzu textových a obrázkových dát pomocou metódy modelovania tém, ktorá bola implementovaná pomocou neurónových sieti.

Kľúčové slová: multimodálny obsah, neurónové siete, modelovanie tém

1 Úvod

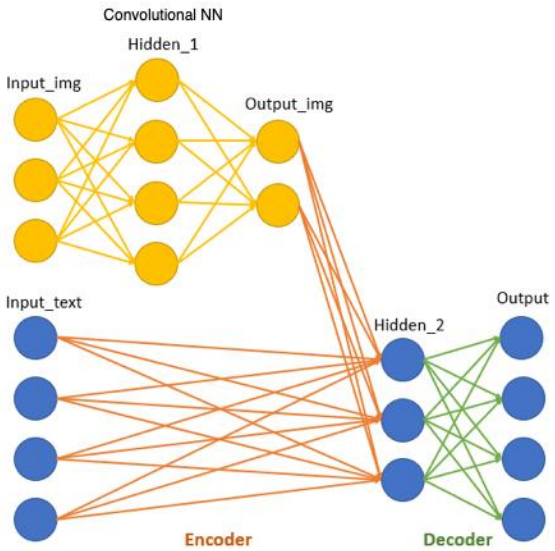
V dnešnej dobe počítačov, s narastajúcim počtom sociálnych médií, narastá počet dát v digitálnej podobe. Tieto dáta sú reprezentované príspevkami tvorenými prevažne multimodálnym obsahom na týchto sociálnych médiach (sociálne siete, diskusné fóra, e-noviny a pod.). Môžu to byť napríklad texty, obrázky, videá či zvukové nahrávky. Tieto dáta môžu obsahovať užitočné informácie z pohľadu firiem (napr. ako ľudia reagujú na niektorý z produktov) alebo aj z pohľadu samotných používateľov (napr. čo sa deje v mojom okolí). Problémom však je množstvo týchto dát, napr. na sociálnej sieti Facebook sa denne vygeneruje okolo 100TB dát a na sociálnej sieti Twitter okolo 175 miliónov príspevkov. Preto nastáva potreba automatického spracovania a analýza takýchto dát zo sociálnych médií za účelom extrakcie užitočných informácii. Existuje niekoľko metód určených na riešenie tejto úlohy, my sme sa rozhodli využiť metódy modelovania tém, ktorého cieľom je vytváranie skupín príbuzných slov (tém) zo vstupnej kolekcie dát. Modelovanie tém nám ukázalo nové možnosti prehľadávania, vyhľadávania a sumarizácie vstupných dát. Doposiaľ bolo vyvinutých niekoľko metód určených pre extrakciu tém ako Latentná Dirichletová alokácia (LDA) [1], hierarchický Dirichletov proces (HDP) [2], gamma poissonov model (GAP) [3] alebo lognormálový gamma poissonov model (LNGAP) [4].

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 39-42.

Problémom týchto metód je, že ich je možné aplikovať len na dáta jedného typu tj. len textové správy alebo len obrázky a pod. Preto sme sa v tejto práci zamerali na navrhnutie metódy založenej na hlbokom učení schopnej spracovanie dát, ktoré môžu obsahovať zmiešaný obsah a to text v spojení s obrázkami.

2 Navrhovaná metóda

Cieľom navrhovanej metódy je extrakcia tém nad príspevkami, ktoré môžu pozostávať zo zmiešaného obsahu textov a obrázkov. Celá táto metóda je založená na prístupoch hlbokého učenia [5] a jej architektúra je zobrazená na Obr.1.



Obr. 1 Architektúra navrhovanej metódy

Navrhovaný model sa skladá z 2 hlavných častí – enkóder (Encoder) a dekóder (Decoder), ktoré sú zvýraznené aj odlišnými farbami. Enkóder je reprezentovaný oranžovými prepojeniami a dekóder je zobrazený zelenými prepojeniami. Enkóder je tvorený dvomi rôznymi vstupmi. Jeden je klasický textový vstup - Input_text (neuróny na tejto vrstve predstavujú jednotlivé slová a ak je ich hodnota 1 tak sa dané slovo v texte nachádza ak 0 tak sa nenachádza) a druhý obrázkový vstup je výstupom (Output_img) konvolučnej neurónovej siete, ktorej úlohou je extrakcia užitočných informácií zo vstupného obrázka. Neuróny na vrstve Hidden_2 reprezentujú vytvorené témy a na aktiváciu neurónov využíva sigmoidálnu aktivačnú funkciu. Posledná vrstva (Output) je výstupom celého modelu a cieľom je rekonštrukcia textového vstupu na tejto vrstve (počet neurónov na tejto vrstve je zhodný s počtom neurónov na vrstve Input_text).

Ak chceme z uvedeného modelu extrahovať témy pre jednotlivé príspevky tak pre jednotlivé vstupy sledujeme aktiváciu neurónov na vrstve Hidden_2, kde čím viac sa

aktivácia neurónu blíži k hodnote 1 tým viac daný príspevok pojednáva o tejto téme. Naopak ak chceme extrahovať kľúčové slová pre dané témy tak na vrstve Hidden_2 nastavíme hodnotu neurónu pre danú tému na hodnotu 1 a hodnoty ostatných neurónov na hodnotu 0 a následne sledujeme aktivácie neurónov (slov) na výstupnej vrstve. Taktiež je možné z tohto modelu extrahovať aj kľúčové slová pre vstupné obrázky z príspevkov a to tak, že hodnoty textového vstupu nastavíme na hodnotu 0 a sledujeme aktivácie výstupu len pre obrázok na vstupe.
Aby sme dosiahli distribúciu tém pre jednotlivé príspevky ako pri klasických metódach modelovania tém upravili sme chybu učenia na skrytej vrstve Hidden_2 pridaním penalizácie pre témy:

$$J_{topic}(t) = J(t) + \Omega(t) \tag{1}$$

$$\Omega(t) = \alpha \sum_{i=1}^{m} KL(\rho||\rho_i') + \beta \sum_{i=1}^{m} KL(\varsigma||\varsigma_i') + \gamma \sum_{i=1}^{h} KL(\sigma||\sigma_i') \tag{2}$$

‚kde $J(t)$ predstavuje pôvodnú chybovú funkciu (v našom prípade „binary cross-entrophy“) a $\Omega(t)$ představuje penalizáciu pre témy, KL predstavuje Kullback-Leibler divergenciu [6]. α, β, γ predstavujú váhu jednotlivých časti funkcie na výslednú chybu, $\rho\ \varsigma\ \sigma$ sú parametre pre penalizáciu tém (konštanty blízke nule, napr. 0,05. Pričom musí platiť $\varsigma < \rho$), $\rho_i^{'}$ je priemerná aktivácia skrytých neurónov pre i-tý trénovací príklad, $\varsigma_i^{'}$ je medián aktivácie skrytých neurónov pre i-tý trénovací príklad, $\sigma_i^{'}$ je priemerná aktivácia skrytých neurónov nad celou trénovacou množinou.

3 Experimenty

Navrhnutý model sme testovali z pohľadu kvality extrahovaných kľúčových slov pre príspevky, ktoré obsahujú len obrázky bez textu, kde sme najskôr model naučili na množine 100 príspevkov, ktorá obsahovala 50 príspevkov o futbale a 50 príspevkov o tenise. Následne sme skúšali extrahovať kľúčové slová pre nové príspevky s obrázkami. Výsledok pre jeden príspevok je zobrazený na Obr.2.

['match', 'world', 'number', 'open', 'second', 'playing', 'champion', 'three', 'australian', 'seed']

Obr. 2 Extrahované kľúčové slová pre obrázok o tenise

Ako je možné vidieť tak naučený model bol schopný extrahovať zmysluplné kľúčové slová pre daný obrázok.

4 Záver

V práci bol prezentovaný model, založený na hlbokom učení, určený na extrakciu tém z multimodálneho obsahu. Ako bolo z experimentov možné vidieť, tak tento model bol schopný extrahovať zmysluplné kľúčové slová pre nové, doposiaľ ešte nevidené obrázky. Avšak v budúcnosti je potrebné daný model otestovať najmä z pohľadu kvality extrakcie tém pre vstupné príspevky.

Poďakovanie: Tento príspevok vznikol s podporou projektov APVV-17-0267, APVV-16-0213 a VEGA č. 1/0493/16.

Literatúra

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022
2. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2012). Hierarchical dirichlet processes. Journal of the american statistical association
3. Ogura H., Amano H., Kondo M. (2013). Gamma-Poisson Distribution Model for Text Categorization
4. Valverde J. A. (2013). Full Bayes Poisson gamma, Poisson lognormal and zero inflated random effects models: Comparing the precision of crash frequency estimates
5. Goodfellow I. A kol. (2016). Deep Learning, MIT Press.
6. Kullback S. (1968). Information Theory and Statistics.

User identification with keyboard and mouse movement dynamics

Vladimír Jančok, Daniela Chudá

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovicova 2, 842 16 Bratislava, Slovakia
vladimir.jancok@stuba.sk

Abstract. In this paper we investigate the possibility of identifying users based on the way of keyboard writing and mouse movement dynamics. This method can enhance or replace the traditional login-and-password approach when attempting to gain access to a system. In our experimental setup, we use the standard mouse and keyboard to gather the user data. When performing the experiments, we extract the appropriate behavioral features from the user actions. The resulting user model should be able to identify users with higher accuracy when performing a single web site login attempt.

Key words: user identification, keyboard and mouse dynamics, biometrics

1 Introduction and Related Work

Computers frequently process and store sensitive data such as payments, private emails or social networks accounts. The theft of device or user account can have serious consequences. Even the basic authentication mechanism (e.g. strong password) can be enhanced by behavioral biometrics with the possibility of background authentication without additional user experience degradation or privacy concerns that might be perceived by more complex biometric authentication methods.

Biometric characteristics such as dynamics of keyboard writing, mouse movement dynamics, activities performed in computer system are characteristic for each user. We can use these features when modeling user for identification, recognition of emotional status, detection of computer experiences [7].

Computer mouse is a valuable source of data, generating events such as mouse movement, mouse-down (button press), mouse-up (button release) and wheel scrolling. We collectively refer to events as raw data. Each event is associated with coordinates of the cursor and the time of occurrence of the event. Afterwards the collected data is preprocessed and the events are grouped and associated with corresponding actions performed by the user. Various universal metrics can be calculated from the mouse actions, such as curvature or velocity. Each metric is further processed to compute a single value called feature. A vector of features computed from metrics is referred to as instance and is added to a *biometric profile* of the user [2].

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 43-48.

Our research focus on the area of building profiles with computer system for identification and increased accuracy of user profiles for different types of users. Naturally, in a web-based environment, users work with the computer mouse much more than with the keyboard [5]. User modeling heavily depends on information provided by the user. The simplest way of building a user model is explicit user modeling, when the user is asked to provide relevant information directly. However, a big challenge of research is to minimize obtrusion and therefore implicit user modeling is trending. This approach relies on feedback from the user in form of actions [5].

A number of studies have been devoted to user authentication using standard input devices [9, 10].

Like all biometric authentication systems, mouse dynamics authentication (MDA) systems involve an enrollment phase and a verification phase. There is still a reasonable potential in increasing the accuracy, speed and cost of biometric authentication in web environment that have become the primary focus of our research.

MDA systems can be classified according to their mode of verification [10].

1. *Static approach* - collect and verify a user's mouse data at specific times (e.g. at login time). Features computed from the user's movement between each pair of dots comprise the enrollment signature. Authenticating involves the same series of dot-to-dot movements, which are compared against the enrollment signature. Mouse movements can be recorded through JavaScript embedded in the web page and sent to a server for processing.
2. *Continuous approach* - collect and verify the user's mouse data repeatedly throughout the entire session. Mouse events are aggregated as higher-level actions such as point-and-clicks or drag-and-drops, characterized by action type, distance, duration and direction. Consecutive actions over some time frame are grouped into sessions.

Behavioral biometrics can provide user authentication directly while using the computer or device. Each person uses their device differently, whether they interact with a mobile device or a computer. Based on this theory and user's unique behavior, we are able to determine whether it is an authorized person and to detect the unauthorized access by an imposter or attacker [3]. Recent research [1] focused on the application of deep learning on datasets of the keystroke dynamics by using convolutional neural networks (CNN) and Gaussian data augmentation technique managed to achieve 10% higher accuracy and 7.3% lower equal error rate (EER) than existing methods. Results were achieved on three publicly available datasets collected from 83 web users entering the same phrase at various sessions and several additional users in controlled environment.

2 User Model for Identification

We propose a user model adapted for static behavioral biometric identification. The proposed solution presented in this paper is a further development of previous research introduced in [4].

We can describe a user model as a reference representation of a user built in the system or application. The process of model creation of user identification or further authentication consists of logging the raw data from input devices, pre-processing of data (e.g. normalization), relevant features extraction and the comparison of user model with collected samples [5]. In our work we examine the setup of training and testing modules in order to simplify the logging system implementation as an effective research tool.

Our project utilizes a functional logger for events from a computer mouse and also for events from a mobile device. This data is stored in a database on the server. It provides data pre-processing and performs a simple classification.

3 Evaluation of Proposed User Model

In order to fully verify the updated user model, several experiments will need to be conducted. In our conditions up to 30 participants can be considered as a sufficient sample.

Once the samples are extracted from the data, we perform the classification (user identification) for each type of event separately. Since the number of samples can vary greatly across users, it can be temporarily reduced also up to 30 for each type.

This paper presents a detection method, which consists of tracking cursor movement and scrolling, extracting features from data and activity type classification.

The experimental results for user identification are shown in Figure 1.

3.1 Datasets and logged data processing

Experimental dataset may contain selected mouse events for a particular session for each user performing the defined tasks on a website.

Tab. 1. *Mouse event dataset structure sample*

time_in_millis	x_hscr	y_vscr	event type	x_hwin	y_vwin	element_bellow
1463658305989	928	756	mousemove	600	671	img
1463658306013	927	757	mousemove	600	672	img
1463658306134	927	757	mousemove	600	672	img
1463658306212	927	757	mousemove	600	672	img
1463658306212	927	672	click	600	672	img

Machine learning module is responsible for data preprocessing, feature extraction and user authentication. It is written in Python 3 language, using the following libraries: *sklearn, pandas, numpy* and *bokeh*. Machine learning module can also provide data visualization functionality.

Once the most important features from the dataset are chosen, they can be used to train the user model. Model for particular user contains the samples that belongs to him

or her. The user model distances can be clustered into two groups: samples from a user and samples from attackers.

Normally the samples from the user have on average shorter distance to model that samples from attackers. In addition the attacker samples are more diverse. Correct limit definition is fundamental, as low limit would cause higher False Rejection Rate (FRR) and high limit could classify the attacker as legitimate user and thus resulting in higher False Acceptance Rate (FAR).

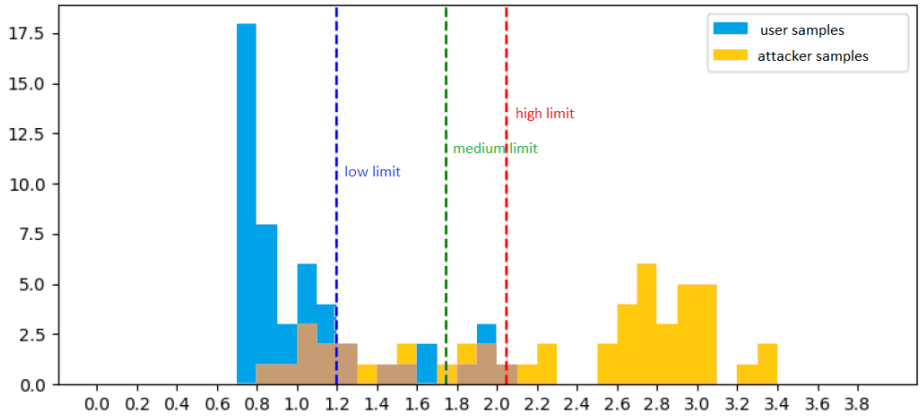


Figure 1. *Graph of sample distances from the user model (low limit – high FRR, users are incorrectly classified; high limit – high FAR, attackers remain undetected)*

3.2 Source code sample for mouse event data gathering

In order to gather user data from mouse actions on a particular web site a JavaScript can be called in index.html body and defined payload can be sent to the logging server.

```
getMouseEventData(event) {
    const rawData = {
      eventType: enumEvents[event.type],
      time: event.timeStamp,
      payload: {
        positionX: event.screenX,
        positionY: event.screenY,
        mouseButton: this.getButton(event),
      },
    };

    if (event.type === 'wheel') {
      rawData.payload['scrollDeltaX'] = event.deltaX;
      rawData.payload['scrollDeltaY'] = event.deltaY;
    }
    return rawData;
  }
```

```
  getButton(event) {
    if (event.type === 'mousedown' || event.type ===
'mouseup') {
      return enumButtonsClick[event.button];
    }
```

Latest version of the complete script and documentation can be found on: *https://gitlab.com/tp-fastar/logger-web/blob/develop/src/mouse-logger.js*.

4 Conclusions and Future Work

In this paper we introduced and partially elaborated a user model for biometric identification based on standard input devices such as mouse and keyboard applicable in general use web environments. In comparison with existing user identification or authentication approaches the current state of our proposed approach to user identification should be scaled to more users in order to confirm the experimental results.

Future work will include improving the identification accuracy and speed by selecting the most relevant and least redundant individual features rather than considering entire groups of features or choosing their most commonly used subset. Dynamic user identification should be also included.

Acknowledgments: This work was partially supported by the Slovak Research and Development Agency under the contract No. APVV-17-0267 - Automated Recognition of Antisocial Behaviour in Online Communities and project Modelling, prediction and evaluation of user behavior based on the web interaction for adaptation and personalization the Scientific Grant Agency of the Slovak Republic, grant No. VG 1/0667/18.

References

1. H. Ceker and S. Upadhyaya: “Sensitivity analysis in keystroke dynamics using convolutional neural networks,” 2017 IEEE Work. Inf. Forensics Secur. WIFS 2017, vol. 2018-January, pp. 1–6, 2018
2. D. Chudá, D. Krátky, K. Burda: Biometric Properties of Mouse Interaction Features on the Web Interacting with Computers, iwy015, https://doi.org/10.1093/iwc/iwy015, 2018
3. D. Chudá and P. Krátky: “International Conference on Computer Systems and Technologies-CompSysTech’14 Usage of computer mouse characteristics for identification in web browsing.”
4. A.K. Jain, A.A. Ross, K. Nandakumar: Introduction to Biometrics. Springer US, Boston, MA (2011).
5. P. Krátky and D. Chudá, “Recognition of web users with the aid of biometric user model,” J. Intell. Inf. Syst., vol. 51, no. 3, pp. 621–646, Dec. 2018.

6. P. Krátky, T. Repiský, and D. Chudá: “Is the Visitor Reading or Navigating?” in Proceedings of the 18th International Conference on Computer Systems and Technologies - CompSysTech’17, 2017, pp. 80–87.
7. P. Krátky and D. Chudá: “Estimating Gender and Age of Web Page Visitors from the Way They Use Their Mouse,” in Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion, 2016, pp. 61–62.
8. S. Mondal and P. Bours, “A study on continuous authentication using a combination of keystroke and mouse biometrics,” Neurocomputing, vol. 230, no. November, pp. 1–22, 2017.
9. J. Shelton, J. Adams, D. Leflore, and G. Dozier, “Mouse tracking, behavioral biometrics, and GEFE,” Conf. Proc. - IEEE SOUTHEASTCON, 2013.
10. C. Shen, Z. Cai, X. Guan, and J. Wang, “On the effectiveness and applicability of mouse dynamics biometric for static authentication: A benchmark study,” Proc. - 2012 5th IAPR Int. Conf. Biometrics, ICB 2012, pp. 378–383, 2012.
11. N. Zheng, A. Paloski, and H. Wang, “An Efficient User Verification System Using Angle-Based Mouse Movement Biometrics,” ACM Trans. Inf. Syst. Secur., vol. 18, no. 3, pp. 1–27, Apr. 2016.

Vysvetľovanie rozhodnutí neurónových sietí s využitím narušenia vstupu so zahrnutím interakcií

Branislav Pecher[1], Jakub Ševcech[1]

[1]Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
{branislav.pecher, jakub.ševcech}@stuba.sk

Abstrakt. Modely hlbokých neurónových sietí sú vo všeobecnosti považované za čierne skrinky, ktorým chýba transparentnosť, čo bráni ich prijatiu v mnohých oblastiach. Byť si istý, že náš model sa správa tak ako by mal je pri takýchto modeloch veľmi cenné. Preto sa metódy, ktoré poskytujú vysvetlenia rozhodnutí modelov stávajú veľmi populárne. V práci navrhujeme novú metódu na vysvetľovanie individuálnych rozhodnutí neurónových sietí za použitia atribučného prístupu založenom na narušení, ktorý vie pri práci zohľadniť interakcie v dátach. Túto metódu overujeme pomocou nových prístupov, ktoré si vedia lepšie poradiť s ťažkosťou overovania atribučných metód. Návrh a overenie metódy je realizovaný nad textovými dátami s vektorovou reprezentáciou.

Kľúčové slová: neurónové siete, interpretovateľnosť, vysvetlenie rozhodnutí

1 Úvod

Hlboké neurónové siete sú vo všeobecnosti jedným z najpresnejších modelov strojového učenia. Aj keď sa v teórií javia ako modely použiteľné na každú úlohu, ich použitiu v praxi bráni to, že sú vnímané ako čierne skrinky, ktoré nemožno dobre vysvetliť.

Vysvetliteľnosť modelov je dôležitá [1]. Klásť dôraz iba na to ako sa modelu darí na danom probléme môže byť zavádzajúce. Keďže dáta ktoré používame na trénovanie a taktiež trénovací proces sú pripravené ľuďmi, ľahko sa môže stať, že problémy v nich sú prehliadnuté, alebo dokonca neúmyselne do nich zavedené [1]. Takéto problémy následne vedú k nesprávnemu modelu, ktorý má tendenciu robiť chyby, ktoré sú neakceptovateľné hlavne v doménach s vysokou cenou za chybu. Iba pomocou vysvetľovania rozhodnutí je možné modely kontrolovať. Vysvetlenie rozhodnutí nám pomáha lepšie pochopiť model, dôverovať mu, alebo ho upraviť tak aby sa správal tak ako má.

V našej práci navrhujeme novú atribučnú metódu na vysvetľovanie individuálnych rozhodnutí neurónových sietí, ktorá vie zobrať do úvahy interakcie v dátach. Vysvetlenia sú reprezentované vo forme dôležitosti jednotlivých vstupných atribútov pre dané rozhodnutie. V práci sa zaoberáme textovými dátami s vektorovou reprezentáciou.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 49-53.

2 Súvisiace práce

Problém vysvetliteľnosti neurónových sietí sa rieši už od 80. rokov, kedy typickým prístupom bola *extrakcia pravidiel* [2, 3]. So zvyšovaním počtu neurónov a skrytých vrstiev a nástupom hlbokých architektúr sa tento problém ešte zvýraznil a v súčasnosti sa štandardne využívajú prístupy použitia *zástupného* modelu, alebo *atribučné* metódy.

Pri *extrakcii pravidiel* sa pomocou správania natrénovanej neurónovej siete vytvoria jednoduché pravidlá, ktoré jej správanie popisujú. Významným reprezentantom tohto prístupu je *DeepRED* [4], ktorý je určený na použitie nad hlbokými neurónovými sieťami, s ktorými si iné prístupy extrakcia pravidiel nevedia dobre poradiť. Tento prístup rozkladá neurónovú sieť na pravidlá pre rozhodovací strom postupne po vrstvách. Výsledný model veľmi blízko aproximuje správanie siete. Aj napriek použitiu rôznych orezávaní je výsledný strom veľmi mohutný a teda nie veľmi interpretovateľný.

Pri použití *zástupného* modelu sa natrénuje jednoduchší model, väčšinou na obmedzenej podmnožine pozorovaní, ktorý je ľahšie interpretovateľný a vysvetlenie pre tento jednoduchý model sa použije na vysvetlenie pôvodného modelu. Jedným reprezentantom tohto prístupu je metóda nazvaná *Locally Interpretable Model-agnostic Explanations* (LIME**)** [5]. Pri tomto prístupe sa vysvetľujú individuálne rozhodnutia pre pozorovania tak, že sa preskúma okolie daného pozorovania a z neho sa vygenerujú nové pozorovania. Tieto pozorovania sa ováhujú na základe vzdialenosti od pôvodného pozorovania a následne sa použijú na natrénovanie jednoduchšieho modelu, ktorým je častokrát lineárny model, alebo rozhodovací strom. Vysvetlenia pre tento model sa následne využijú na vysvetlenie pôvodného rozhodnutia. Tento proces je potrebné zopakovať pre každé pozorovanie zvlášť čo zapríčiňuje výrazné spomalenie vysvetľovania väčšieho počtu rozhodnutí.

Pri *atribučných* metódach sa každému vstupnému atribútu priradí skóre, ktoré vyjadruje jeho dôležitosť pre dané rozhodnutie. Jedným reprezentantom takýchto metód je *Layer-wise Relevance Propagation* (LRP) [6]. V tomto prístupe sa najprv na výstupnej vrstve definuje skóre. Toto skóre je následne šírenie naprieč sieťou na vstupnú vrstvu za pomoci posielania správ, ktoré blízko korešponduje s ováhovanými prepojeniami v sieti. Keďže tento prístup využíva gradient chyby v sieti, radí sa medzi gradientové metódy, ktoré sú závislé na architektúre. Druhým typom sú prístupy založené na narušení, ktoré rátajú skóre priamo, zavedením narušenia do vstupu a pozorovaním zmeny vo výstupe. Použitie tohto prístupu na obrázky za pomoci posuvného okna predstavil Zintgraf a iní [7]. Hoci sú prístupy založené na narušení nezávislé od architektúry, pri veľkom množstve atribútov môže vysvetlenie jedného rozhodnutia trvať aj niekoľko hodín a nevedia sa dobre vysporiadať s interakciami v dátach.

3 Narušenie vstupu so zohľadnením interakcií

V našej práci navrhujeme novú atribučnú metódu založenú na narušení vstupu, ktorá dokáže zobrať do úvahy interakcie v dátach a tak vytvárať presnejšie vysvetlenia. Každému vstupnému atribútu priradíme skóre, ktoré vyjadruje dôležitosť tohto atribútu

pre dané konkrétne rozhodnutie. Vysvetlenie jedného rozhodnutia prebieha tak, že toto skóre sa určí pre každý vstupný atribút daného pozorovania a tým sa určí, ktoré atribúty boli najdôležitejšie pre dané rozhodnutie. Týmto prístupom vieme identifikovať tie atribúty, ktoré podporujú, alebo potláčajú rozhodnutie. Predpokladáme, že vstupné textové dáta sú vo vektorovej reprezentácii. To nám zaručuje zachovanie informácie o poradí slov a tým aj jednotlivé interakcie, pričom ich vieme jednoducho identifikovať.

Vysvetlenie jedného rozhodnutia prebieha v nasledujúcich 4 krokoch. V prvom kroku vytvoríme ***referenčný výstup*** s ktorým sa budeme porovnávať. Ten vytvoríme tak, že predikujeme výstup pre dané pozorovanie pomocou natrénovanej neurónovej siete a zoberieme pravdepodobnosť najpravdepodobnejšej triedy ako našu referenciu.

V ďalšom kroku určíme pre každé vstupné slovo ich ***interakcie***. Tie je možné určiť 3 rôznymi prístupmi. Prvá možnosť je jednoducho zobrať pevný počet predchádzajúcich a nasledujúcich slov. Tým dostávame množinu interagujúcich slov vždy s rovnakou veľkosťou, avšak pre niektoré slová môže byť ich dôležitosť precenená. Druhou možnosťou je vyrátať podobnosť (napríklad kosínovú) medzi práve vyšetrovaným slovom a všetkými ostatnými slovami a všetky s podobnosťou vyššou ako určitý prah zobrať ako korelované. Tento prístup produkuje presnejšie výsledky za cenu rýchlosti a rozdielnej veľkosti množín slov, s ktorou sa treba neskôr vysporiadať. Poslednou možnosťou je zobrať kombináciu predchádzajúcich dvoch prístupov.

Po identifikovaní interakcií sa do pozorovania zavedie ***narušenie***, konkrétne sa naruší vyšetrované slovo spolu so slovami s ktorými interaguje. Narušenie ktoré využívame je úplne odstránenie daných slov zo vstupu tým, že sa namiesto nich použije špeciálny nulový vektor, ktorý reprezentuje absenciu signálu.

Keď už máme takto narušený vstup, môžeme ho použiť na určenie ***finálneho skóre.*** Pre takto narušený vstup sa určí výstup siete, zoberie sa pravdepodobnosť triedy určenej v prvom kroku a porovná sa s našou referenciou. Hodnotu finálneho skóre pre atribút určujeme za pomoci nasledujúceho princípu: ak nastal výrazný pokles pravdepodobnosti po zavedení narušenia, môžeme povedať že dané slovo bolo dôležité pre podporu rozhodnutia. Na druhú stranu ak sa pravdepodobnosť zvýšila tak ide o dôležité slovo, ktoré však rozhodnutie potláča. Ak nedošlo k žiadnej zmene, alebo iba k minimálnej, išlo o slovo bez významu pre dané rozhodnutie. Na výpočet preto používame jednoduchý rozdiel. Tento rozdiel je normalizovaný počtom odstránených slov a pripočítaný k dôležitosti každého odstráneného atribútu. Po prejdení všetkých atribútov sa ešte jednotlivé dôležitosti normalizujú počtom týchto pripočítaní.

4 Experiment: porovnanie s inými atribučnými metódami

Najlepším spôsobom ako overiť našu atribučnú metódu je porovnať jej výsledky s výsledkami iných atribučných metód. Na uskutočnenie tohto porovnania však potrebujeme nejaký spôsob ako povedať, ktorý výsledok je lepší, nejaký zlatý štandard voči ktorému sa porovnať. Ten však neexistuje. Prvou časťou tohto experimentu je teda vytvoriť takýto zlatý štandard s využitím spätnej väzby od ľudí.

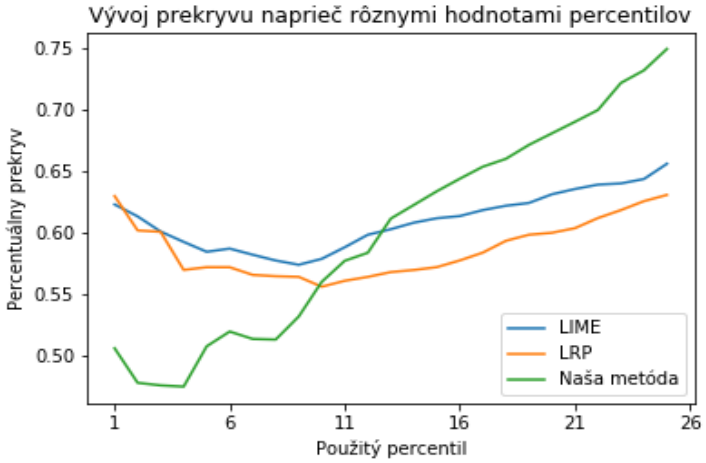
Na vytvorenie tohto zlatého štandardu, a aj následné porovnanie medzi atribučnými metódami, sme využili problém analýzy sentimentu nad textovými hodnoteniami filmov z IMDB. Tento problém sme si vybrali preto, lebo je veľmi intuitívny pre ľudí

a jeho výsledkom je množina dôležitých slov podporujúcich a potláčajúcich rozhodnutie. Vytvorenie zlatého štandardu teda prebieha tak, že sa každému používateľovi zobrazí niekoľko hodnotení, ktoré sa medzi rôznymi ľuďmi aj opakujú, a sú požiadaní aby vyznačili slová, ktoré pre nich vyjadrujú pozitívny a negatívny sentiment, pričom počet možných vyznačených slov neobmedzujeme. Takto dostávame množinu dôležitých slov od ľudí, voči ktorej je možné sa porovnať.

Druhou časťou experimentu je už priamo porovnanie s atribučnými metódami, konkrétne s metódami LIME [3] a LRP [4] opísanými v kapitole 2. Každá z týchto metód dokáže ako výsledok vrátiť usporiadaný zoznam slov, extrahovaných z vysvetľovanej neurónovej siete pre jedno konkrétne pozorovanie, spolu s dôležitosťou týchto slov. Aby sme overili správanie v rôznych nastaveniach, dôležité slová z metód určujeme na základe variabilného percentilu dôležitosti, pričom nadobúda hodnotu medzi 1 a 25. Ak teda máme percentil s hodnotou 5, z jednotlivých metód používame prvých 5% najdôležitejších slov s pozitívnym a negatívnym sentimentom zvlášť. Na určenie interakcií v našej metóde používame kombináciu kosínovej podobnosti s hranicou podobnosti 70% a pevného počtu 2 predchádzajúcich a nasledujúcich slov.

Určené slová z jednotlivých metód porovnávame so zlatým štandardom pomocou percentuálneho prekryvu, pričom porovnávame zvlášť slová s pozitívnym a negatívnym sentimentom. Výsledky následne porovnávame naprieč metódami.

Výsledky tohto porovnania možno vidieť na obrázku 1. Vidíme, že pri použití väčšieho množstva dôležitých slov dosahuje naša metóda lepšie výsledky ako zvyšné dve, čo značí, že dokáže lepšie identifikovať väčšie množstvo slov dôležitých pre rozhodnutie, ktoré by inak boli skryté kvôli interakciám. Avšak pri malom počte slov dáva horšie výsledky čo značí, že nedokáže dobre identifikovať malý počet tých najdôležitejších slov kvôli spôsobu ako funguje.



Obr. 1. Výsledky porovnanie našej metódy a metód LIME a LRP. Na určenie interakcií používame kombináciu kosínovej podobnosti s hranicou 70% a 2 prechádzajúce a nasledujúce slová.

5 Záver

V našej práci sme navrhli novú atribučnú metódu založenú na narušení vstupu, ktorá berie do úvahy interakcie v dátach, čím dokáže zlepšiť úspešnosť. Túto metódu je možné použiť na vysvetľovanie individuálnych rozhodnutí neurónových sietí.

Metódu sme overovali pomocou porovnania s inými atribučnými metódami s využitím spätnej väzby od ľudí na vytvorenie zlatého štandardu. Výsledky indikujú, že naša metóda dokáže lepšie identifikovať dôležité slová skryté kvôli interakciám v dátach, avšak nedokáže dobre identifikovať malý počet tých najdôležitejších slov. Tento problém sa dá vyriešiť sofistikovanejším určovaním interakcií, alebo kombináciou našej metódy s metódou v ktorej nie sú zahrnuté interakcie, čo bude vykonané v ďalšej práci.

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV-17-0267, APVV-15-0508 a VG 1/0725/19.

Literatúra

1. Doshi-Velez, F., Been K.: Towards a rigorous science of interpretable machine learning. 2017.
2. Andrews R., Diederich J., Tickle A.B.: Survey and critique of techniques for extracting rules from trained artificical neural networks. In: Knowledge Based Systems, vol. 8, pp. 378–389, 1995.
3. Tickle A.B., et al.: The truth will come to light: Directions and challenges in extracting rules from trained artificial neural networks. In: IEEE Transactions on Neural Networks, vol. 9, pp. 1057–1068, 1998.
4. Zilke J. R., Mencía E. L., Janssen F.: Deepred – rule extraction from deep neural networks. In: International Conference on Discovery Science, pages 457–473. Springer, 2016.
5. Ribeiro, M. T., Sameer S., Carlos G.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp. 1135–1144, 2016.
6. Bach, S., et al.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, vol. 10, no. 7, p.e0130140, 2015.
7. Zintgraf, L.M., et al.: Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. In: 5th International Conference on Learning Representations (ICLR 2017), 2017.

Odporúčanie založené na lokálnych temporálnych aspektoch

Elena Štefancová, Ivan Srba

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava
{meno}.{priezvisko}@stuba.sk

Abstrakt. Táto práca sa zaoberá odporúčaním založeným na časových aspektoch v doméne sociálnych sietí založených na bodoch záujmu. Takýmito sieťami sú napríklad Yelp alebo Foursquare. Navrhujeme metódu, ktorá odporúča body záujmu na základe sezónnosti. Avšak, na rozdiel od existujúcich metód, modelujeme tieto temporálne aspekty špeciálne pre jednotlivé geografické oblasti, nie globálne. Výsledky ukazujú, že zohľadnenie lokálnej sezónnosti je účinnejšie ako jej globálna alternatíva.

Kľúčové slová: odporúčanie, odporúčacie systémy, čas, kontext, sezónnosť

1 Úvod

Odporúčacie systémy sú dôležitou súčasťou webových služieb a ich popularita je na vzostupe. Zohľadnenie kontextu sa ukázalo ako prospešné pre zvýšenie úspešnosti vo viacerých scenároch. V prípade temporálneho kontextu, odporúčanie môže zohľadniť cyklické vzory správania ľudí. Ako príklad cyklických vzorov si môžeme predstaviť sezónnosť v prípade nákupu oblečenia – športovo orientovaný človek si nakúpi v obchode športové oblečenie počas celého roka, ale iné prvky uprednostní cez leto (napr. krátke turistické nohavice) a iné počas zimy (napr. lyžiarsku bundu). Pritom tieto vzory sú často podobné naprieč spoločnosťou – väčšina populácie ráno uprednostní návštevu kaviarne a naopak bar vo večerných hodinách.

My sa zameriavame práve na doménu na lokalite založených sociálnych sietí (angl. Location Based Social Networks - LBSN), ktoré umožňujú používateľom vyhľadať a ohodnotiť body záujmu (angl. Points of Interests - POI). Na úspešnosť týchto sietí ma odporúčanie veľký vplyv. Odporúčajú používateľom, ktorý obchod, reštauráciu atď. navštíviť. Je to zároveň doména, kde geografický a temporálny kontext zohráva signifikantnú úlohu. Sústredíme sa primárne na temporálny kontext v odporúčaní. Viaceré práce v minulosti sa temporálnym kontextom v tejto doméne zaoberali, ale zvyčajne len na globálnej úrovni. Takéto práce teda nezohľadňujú lokálnu špecifickosť (teda špecifickú sezónnosť pre konkrétne oblasti). Práve tento otvorený problém sa snažíme riešiť, primárne pritom skúmajúc rozdielne časové vplyvy v oblastiach s

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 54-59.

veľkými výkyvmi počasia na základe ročných období a oblastí, kde sú tieto výkyvy miernejšie. Využívame pritom pred-filtrovanie a modelovanie s maticovou faktorizáciou, ktorá zohľadňuje temporálne aspekty v odporúčaní POI. Pre spresnenie výsledkov sa vykonáva aj geografické po-filtrovanie.

Náš hlavný prínos pozostáva z výskumu vplyvov temporálnych aspektov lokality, pričom sme zistili, že pri využití sezónnosti je skutočne vhodnejšie zohľadňovať lokálne trendy, najmä keď sa jedná o oblasti, kde je premenlivosť počasia počas roka výrazná.

2 Kontextuálne odporúčanie

V rámci klasických odporúčacích techník nie je zvykom brať kontext v úvahu. Avšak, v mnohých prípadoch je zahrnutie relevantné – napr. používatelia pozerajú iné filmy počas Vianoc než počas zvyšku roka [1]. Najčastejšie zohľadnené aspekty kontextu bývajú čas, geografické prvky a prítomnosť iných osôb. V takomto prípade je odporúčanie obohatené o ďalší vstup (Rovnica 1, U používatelia, I položky, C kontext).

$$\mathrm{R} : \mathrm{U} \times \mathrm{I} \times \mathrm{C} \rightarrow \text{Hodnotenie položiek} \tag{1}$$

Zahrnutie kontextu môže byť vykonané ako [2, 3]:

- *Pred-filtrovanie* – dáta sú rozdelené na základe kontextu na segmenty (napr. položka na viacero inštancií), nad ktorými sa následne odporúča takmer samostatne.
- *Po-filtrovanie* – vykoná sa odporúčanie bez kontextu, výsledky sa preusporiadajú na základe kontextu.
- *Modelovanie* – kontext vstupuje do samostatného procesu odporúčania.

Temporálne aspekty používateľov a položiek môžu byť založené na aktuálnosti (nedávne transakcie majú väčšiu váhu ako staršie) alebo cyklickosti (napr. sezónnosť, cykly počas dňa) [3, 5].

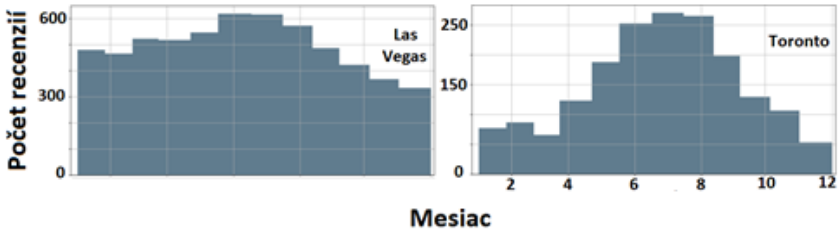
2.1 Odporúčanie POI na základe časových aspektov

Častou doménou, v ktorej sa kontextuálne odporúčanie uplatňuje, sú na lokalite založené sociálne siete (LBSN) [4], akými sú Foursquare či Yelp. Základným prvkom týchto sociálnych sietí je bod záujmu (POI). Transakciami v týchto systémoch sú tzv. prihlásenia [7] alebo recenzie používateľov [6] na POI. Rozdiel je v charaktere recenzií ako zvyčajne jednorazových aktivít, zatiaľ čo používateľ má spravidla viacero prihlásení aj na tom istom POI.

Kontext sa v LBSN využíva na odporúčanie POI v dosahu používateľa (geografický aspekt) a v momente, kedy sa daný POI zhoduje so záujmy používateľa (časový aspekt) [7]. Z prístupov je najčastejšie využívané kolaboratívne filtrovanie, keďže záujmy používateľov sa zvyknú opakovať v spoločnosti – napríklad väčšina zvykne navštíviť kino vo večerných hodinách. Keďže zapojenie kontextu v dátach sa často prejaví na následnej riedkosti dát [7], metódy odhaľujúce latentné črty sa stali populárnymi. Jednou z najpopulárnejších je maticová faktorizácia [8].

3 Návrh metódy

Navrhli sme metódu, ktorá sa zaoberá sezónnosťou kategórií POI na základe ich geografickej lokality (Obrázok 1). Na rozdiel od príbuzných metód, odporúčame prvú návštevu zatiaľ nenavštíveného POI (nepoužívame prihlásenia, ale recenzie). Ako základný cyklus sezónnosti sme si vybrali rok, pričom samotnými „sezónami" sú jednotlivé mesiace. Metóda kombinuje temporálne pred-filtrovanie, modelovanie časových aspektov a geografické po-filtrovanie.



Obr. 1. Počet recenzií v meste so slabými (Las Vegas) / silnými (Toronto) zimami rozdelenými podľa mesiacov

Základná odporúčacia metóda je založená na kombinácii kolaboratívneho filtrovania (matica používatelia x položky) a maticovej faktorizácii. Zadané hodnotenia (vrámci recenzií) sú normalizované pre každého používateľa ako neutrálne (0 – nenavštívené položky používateľom), negatívne (-1 - najnižší kvantil hodnotení daného používateľa) a pozitívne (zvyšok).

Prvým krokom je pre-filtrovanie, kedy sa transakcie každého POI rozdelia podľa mesiacov a vytvorí sa tak nových 12 inštancií POI (jedna pre každý mesiac, každá inštancia s transakciami z daného mesiaca).

Následne je vykonané modelovanie (vďaka odporúčaniu založenom na obsahu a kolaboratívnom filtrovaní) s maticovou faktorizáciou. Vektory čŕt pre položky slúžia na zachytenie kontextuálnych informácií. Ako črty slúžia informácie o aktuálnom mesiaci (cyklická hodnota alebo číslo), oblasti (zakódovaný štát využitím techniky *one-hot encoded*), kategória (zakódovaná technikou *one-hot encoded* alebo ako skóre kategórie v danom mesiaci).

Pritom skóre kategórie sa ráta ako pomer jej priemerného počtu transakcií v danom mesiaci ku priemernému počtu jej transakcií počas roka. Ak POI podlieha viacerým kategóriám, skóre vzniká ako vážený priemer všetkých kategórií. Čím viac podnikov podlieha tejto kategórii, tým je vplyv jej skóre menší. Tento prístup bol zvolený, aby sa oslabil vplyv generických kategórií (napr. jedlo) a dal dôraz na špecifickejšie kategórie POI (napr. indická kuchyňa).

Na záver sa uprednostnia odporúčania, ktorú sú v blízkosti už navštívených položiek používateľom (geografický po-filter).

4 Vyhodnotenie experimentu

Na extrahovanie použitých dátových sád sme využili dátovú sadu Yelp. Na trénovanie sme použili dve dátové sady – jednu so silným výskytom sezónnosti, druhú s jej minimálnymi prejavmi. Overenie prebiehalo na obdobne vybraných dvoch dátových sadách a ich kombinácii. Každá z dátových sád obsahovala približne 3 200 POI, 39 000 recenzií a 3 000 používateľov.

Ako najužitočnejšie črty sa prejavili mesiac ako cyklická hodnota, zakódovaný štát (len v prípade lokálnej sezónnosti) a skóre kategórie. Skóre sme vytvárali pre lokálny a globálny prístup zvlášť. Lokálne skóre na základe hodnotení kategórií POI v danej oblasti v danej časti roka) a globálne ako spoločne pre všetky oblasti.

Najlepšie výsledky dosahovalo modelovanie v kombinácii s pred-filtrovaním. Výsledky v Tabuľke 1 ukazujú, že sezónnosť špecifická pre danú lokalitu dosiahla výrazne lepšie výsledky ako globálna. Najmä pre mestá, kde sa počasie v priebehu roka sa výrazne líši. Pre kombináciu všetkých miest, aj tých bez výrazne cyklickosti počas roka, má lokálna sezónnosť tiež lepšie výsledky, ale zaostáva za základným odporúčaním bez sezónnych vplyvov. Toto je spôsobené mestami, ktoré nepodliehajú cyklickosti počasia počas roka, teda majú stabilnejšiu klímu. Sezónnosť nad týmito dátami dávala extrémne nízke hodnoty.

Tab. 1. Porovnanie odporúčania bez temporálnych aspektov, so zohľadnením globálnej sezónnosti a so zohľadnením lokálnej sezónnosti (zvýraznené hodnoty kurzívou sú lepšie ako odporúčanie bez temporálnych vplyvov, hodnoty zvýraznené hrúbkou sú najlepšie výsledky zo všetkých troch možností)

	Základná MF		**Lokálne temporálne črty**		**Globálne temporálne črty**	
k	**Presnosť**	**Pokrytie**	**Presnosť**	**Pokrytie**	**Presnosť**	**Pokrytie**
	Dátová sada miest s rôznou intenzitou výkyvov počasia v priebehu roka					
1	1,01	0,52	**1,11**	**0,83**	1,32	1,1
3	1,5	1,41	1,22	1,22	1,15	1,15
5	2,11	2,07	1,22	1,22	1,15	1,15
10	2,36	2,36	1,31	1,31	1,15	1,15
	Dátová sada miest s výraznými výkyvmi počasie v priebehu roka					
1	1,78	0,41	**3,29**	**2,41**	*2,42*	*1,77*
3	1,12	0,84	**2,82**	**2,82**	*2,41*	*2,41*
5	1,3	1,23	**2,82**	**2,82**	*2,5*	*2,5*
10	1,79	1,78	**2,82**	**2,82**	*2,68*	*2,68*

Zistili sme, že sezónnosť funguje lepšie pre používateľov s vyšším počtom predchádzajúcich transakcií, zatiaľ čo výkon pre začínajúcich používateľov je podobný odporúčaniu založenému na popularite.

Z týchto výsledkov môžeme vyvodiť záver, že sezónnosť špecifická pre danú lokalitu môže skutočne dosahovať lepšiu výkonnosť ako globálna sezónnosť. Toto zistenie sa však nemusí pozitívne odraziť vo všetkých prípadoch. Ako sa dá očakávať,

došlo k významnému zlepšeniu pre lokality a kategórie s výraznými sezónnymi výkyvmi počasia. Preto odporúčame sezónnosť špecifickú pre danú lokalitu používať výlučne pre prípad lokalít a kategórií so značným sezónnym charakterom a naopak ignorovať ju pri zvyšku.

Naše výsledky sme porovnali s prácou Zhang et al. [9], ktorých metóda nazvaná GeoSoCa využíva geografický kontext, spoločenské korelácie medzi používateľmi a vplyv kategórií. Ich dataset obsahoval údaje z datasetu Yelp, z mesta Pheonix, Arizona, USA. Dosiahli presnosť@k 1% a pokrytie@k 1,5% (k = 10), ktoré sú výrazne nižšie ako naše výsledky 3,29 %, resp. 2,41.

5 Záver

Zatiaľ čo časový kontext je široko analyzovaný v oblasti odporúčania POI, neidentifikovali sme žiadnu prácu, ktorá by sa pokúsila explicitne modelovať časové aspekty špecifické pre danú lokalitu. Preto sme navrhli a experimentálne overili systém odporúčania, ktorý explicitne modeluje sezónnosť pre danú lokalitu a porovnávať ich s ich globálnymi verziami. Môžeme potvrdiť pozitívny vplyv tohto prístupu, najmä pri oblastiach s výkyvmi počasia počas roka tento prístup umožňuje využiť plný potenciál sezónnosti špecifickej pre danú lokalitu.

Zistenia poskytujú príležitosti pre budúci výskum. Bolo by zaujímavé experimentovať s rôznymi jednotkami lokality (regiónmi, krajinami) a sezónnosti (mesiac, týždeň, deň). Nápomocnou metódou by mohlo byť zhlukovanie lokalít podľa sezónnosti.

Navyše, keďže navrhovaná metóda nie je v závislosti od konkrétnej domény, môže byť použitá aj v inom prostredí než LBSN.

Poďakovanie: Tento príspevok vznikol vďaka čiastočnej podpore v rámci projektov č. APVV-15-0508, VG 1/0725/19 a KEGA 028STU-4/2017.

Literatúra

1. Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach. ACM Trans. Inf. Syst. 23, 1 (Jan. 2005), 103–145. https://doi.org/10.1145/1055709.1055714
2. Gediminas Adomavicius and Alexander Tuzhilin. 2008. Context-aware Recommender Systems. Proceedings of the 2008 ACM Conference on Recommender Systems (2008), 335–336, dostupné na: https://doi.org/10.1145/1454008.1454068arXiv:arXiv:1011.1669v3
3. Charu C. Aggarwal. 2016. Recommender Systems - The Textbook. Vol. 40. Springer International Publishing, Cham. 56–58 pages. https://doi.org/10.1007/978-3-319-29659-3 arXiv:arXiv:102.1112v1
4. Jie Bao, Yu Zheng, and Mohamed F. Mokbel. 2012. Location-based and Preference-aware Recommendation Using Sparse Geo-social Networking Data. In Proceedings of the 20th International Conference on Advances in Geographic Information Systems (SIGSPATIAL '12). ACM, New York, NY, USA, 199–208. https://doi.org/10.1145/2424321.2424348

5. Justin Basilico and Yves Raimond. 2017. DéJà Vu: The Importance of Time and Causality in Recommender Systems. Proceedings of the Eleventh ACM Conference on Recommender Systems (2017), 342. https://doi.org/10.1145/3109859.3109922
6. Jungkyu Han and Hayato Yamana. 2017. Geographical Diversification in POI Recommendation: Toward Improved Coverage on Interested Areas. In Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17). ACM, New York, NY, USA, 224–228. https://doi.org/10.1145/3109859.3109884
7. Xin Li, Mingming Jiang, Huiting Hong, and Lejian Liao. 2017. A Time-Aware Personalized Point-of-Interest Recommendation via High-Order Tensor Factorization. ACM Trans. Inf. Syst. 35, 4, Article 31 (June 2017), 23 pages. https://doi.org/10.1145/3057283
8. Francisco J. Peña. 2017. Unsupervised Context-Driven Recommendations Based On User Reviews. Proceedings of the Eleventh ACM Conference on Recommender Systems - RecSys '17 (2017), 426–430. https://doi.org/10.1145/3109859.3109865
9. Zhang, J.D., Chow, C.Y.: Geosoca: Exploiting geographical, social and categoricalcorrelations for point-of-interest recommendations. In: Proceedings of the 38thInternational ACM SIGIR Conference on Research and Development in Infor-mation Retrieval. pp. 443–452. SIGIR '15, ACM, New York, NY, USA (2015).https://doi.org/10.1145/2766462.2767711

Data Analytics in Sports Statistics

Martin Gajdoščík, František Babič

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia
martin.gajdoscik@student.tuke.sk, frantisek.babic@tuke.sk

Abstract. Nowadays, data analytics plays a crucial role in the sports industry today. It can be used for different purposes like a draft selection, game-day decision making, or player evaluation. In our work, we aim to understand the data about the National Hockey League and to generate the best possible model for salaries prediction. The whole analytical process we managed by the most widely used methodology, called CRISP-DM. In the data preparation phase, we dealt with a feature selection to find the most influencing attributes from the initial dataset. Next, we used the stepwise regression to find the best combination of these attributes. Finally, we generated the expected models within Polynomial regression. In the case of goalies, the best model's precision exceeds 95%.
On the other hand, the models for forwards and defenseman's together have the precision just over 50%. In this case, we derived new data samples based on the contract type (RFA, UFA, Rookie). This operation helped us to achieve higher precision, such as 92.46% for defenseman Rookie or 82.92% for forward UFA. The final set of models can be used as a simple decision support system for players, managers, agents, team owners, or general managers.

Keywords: hockey statistics, regression, feature selection

1 Introduction

Data analytics is revolutionizing the world of sports. The teams use suitable analytical methods and approaches for different purposes from an analysis of sensor measurements placed on players during training, to predicting their opponents' strategy, or even their own best strategy. Data analytics has been applied to other major sports with some successes, but minimal research activities were performed within ice hockey statistics. Some authors argue that it is complicated to divide the hockey match into a sequence of separate plays.

The website of the National Hockey League (NHL) keeps detailed player statistics dating back to 1997, and appends these basic statistical categories with penalty minutes, power-play goals, shorthanded goals, game-winning goals, game-tying goals, overtime goals, shots, shooting percentage, time on ice, shifts per game, and face-off win percentage.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 60-65.

One group of people involved in the management of the ice hockey team that can benefit from the stats analytics are the team executives. For example, they can propose a trade for a player which is undervalued in the other team. Also, they can use the models for salaries prediction to plan the team's budget per year and in the short-term future.

The paper is organized as follows: a short introduction with brief state-of-the-art and used methodology. Next, the whole analytical process is described step by step with the conclusion.

1.1 Related work

We briefly describe selected works of application of selected analytical methods in various sports. Wang and Wang [1] used the Apriori algorithm to determine association rules between different types of technical movements in soccer. Chai [2] applied data mining techniques to the serial data to decide what chains of possession between players often happened, incorporating both teams with specific players into each pattern. Nunes and Sousa [3] applied different data mining and data visualization techniques to available data relating to the European championship soccer matches over the history of the tournaments. Hipp and Mazlack [4] investigated the situation of mining ice hockey data based on existing work and appropriate methods. McDonald [5] generated models to predict goal using variables like faceoffs, hits, and other statistics as predictor variables in addition to goals, shots, missed shots, and blocked shots. Buttrey et al. [6] proposed a model to estimate the rates at which NHL teams score and yield goals. In low-scoring sports such as ice hockey, the first goal is critically essential to the game outcome. Jones [7] investigated the home advantage for the scoring of a first goal and its effect for both teams. Chuckers [8] used historical data available when players were eligible to be selected in the National Hockey League (NHL) Player Entry Draft to build a statistical prediction model for their performance in the NHL. Weissbock et al. [9] combined traditional statistics, such as goals for and against, and performance metrics such as possession and luck, to build a classification model.

1.2 CRISP-DM

The Cross-Industry Standard Process for Data Mining methodology includes descriptions of the typical phases of a project, the tasks involved with each step, and an explanation of the relationships between these tasks [10], [11]. The life cycle model consists of six phases, with arrows indicating the most important and frequent dependencies between steps.

The first phase deals with a specification of business goal and its transformation to the data mining context. The second phase focus on detailed data understanding through various graphical and statistical methods. Data preparation is usually the most complex and most time-consuming phase, including data aggregation, cleaning, reduction, or transformation. In modeling, different machine learning algorithms are applied to the preprocessed datasets. Traditionally, this dataset is divided into training and testing sample, or analysts use 10-cross validation. The obtained results are evaluated by traditional metrics like accuracy, ROC, precision, or recall. Also, the analysts verify the accomplish of the specified business goals. The last phase is devoted to the deployment

of the best results into practice and identification of the best or worst practices for the next analytical processes.

2 Analytical Process

The whole analytical process was performed through the two most popular programming languages in the domain of data analytics: R and Python.

2.1 Business Understanding

From a business perspective, it is essential to understand the data about each player as best as possible and to predict their future salary as accurately as possible. The analytical point of view covers an application of suitable feature selection and regression methods to generate the prediction model.

2.2 Data Understanding and Preparation

The input data sample contains 6 908 records of which 6 266 were player records, and the remaining 642 were goalies. We analyzed seven seasons from 2011. This decision meant that we had a different number of variables per season. Most of the attributes were numeric.

We divided the goalie's stats into ten and player's stats into twelve categories by the same characteristics. In the beginning, the number of attributes was high (79 for goalies, 216 for the players) and therefore, depending on the correlation between them, we started their reduction. For example, the pairs of attributes like iFOW – Fow.Up, G.Tip – G.Wrap, iPent – iPENT or Misc – G.Misc represents a situation that one of them is calculated from the value of the second. We also used the Bayesian information criterion (BIC) to remove less-important categories and keep only relevant data [12].

Other preprocessing operations represented synchronization the attributes names and measuring units; deletion of all players and goalies that have played nine or fewer matches, creation of new twelve attributes based on existing data, such as goals per game played, assists per game played, shorthanded goals, etc.

Next, we calculated the importance of numerical attributes relative to the target attribute (CapHit) in both data samples (players, goalies) by the univariate attribute selection [13]:

- Top 5 goalies attributes: Wt (weight), HT (height), PPSA (power play shots against), GAA (goals-against average), SHSA (shorthanded shot against).
- Top 5 players attributes: WR (weight), HT (height), TOI / G (time on ice per game), Shift / G (shifts per game), TOI (time on ice).

Also, we applied the Sequential Forward Selection [14] to identify the most important attributes. Sequential Feature Selector algorithms belong to a family of greedy search algorithms adding or removing one attribute at the time based on the classifier performance until a variables subset of the desired size is reached:

- Top 16 goalies attributes: GS, OT, GR, MIN, SO, StMIN, StSV%, StGAA, QS, ReMIN, SHSV%, PPSV%, SOW, SOL, SSA and, SGA (the prediction accuracy only 54%, but the best from the all related experiments).
- Top 8 players attributes: HT, Wt, TOI/G, A/G, PIM/G, Shifts/G, Diff/G and Bl/G (the prediction accuracy 67%).

The results of the second algorithm better covered the target task.

Finally, we divided the data into nine tables by type of contract and by the player's position. Goalies data was split into three small subfiles by contract type to RFA (restricted free agents), UFA (unrestricted free agents) and rookie categories with salary less than 1 million. Player attributes were initially divided by position, namely to the defenders and the forwards. Then again by contract type to UFA, RFA and rookie with salary less than 1 million. Overall, we have 9 data files. The goalie's samples contained 47 columns and 62-261 rows, the players' data 59 attributes and 308-1423 rows (players).

2.3 Modeling and Evaluation

We used the preprocessed data as an input to generate the models within the polynomial regression methods (Table 1).

Table 1. Precision and RMSE (Root Mean Square Error) for all tables at polynomial regression (precision is 100*score, R-RFA, U-UFA, N-Rookie)

Table	Precision	RMSE	Table	Precision	RMSE
GoaliesN	84.80	0.04689	ForwardN	41.50	0.13733
GoaliesU	79.73	0.89543	ForwardU	82.92	0.84808
GoaliesR	84.78	0.49851	ForwardR	71.46	0.82824
DefensemanN	54.30	0.12428	Goalies	95.47	0.45397
DefensemanU	53.99	1.11923	Forward	54.34	1.36579
DefensemanR	92.46	0.43649			

As we can see from the table, all three goalie subcategories were less accurate than the whole set. The situation was the opposite for the players; the subgroups had higher precision than the entire sample. Figure 1 visualises the result for defenseman (unrestricted free agents)

2.4 Deployment

The final stage of the whole analytical process represents a deployment of the best models or extracted knowledge into practice. In our case, we were able only to simulate this procedure, i.e., to design and implement a decision support systems offering selected experimental steps in a simple, understandable form.

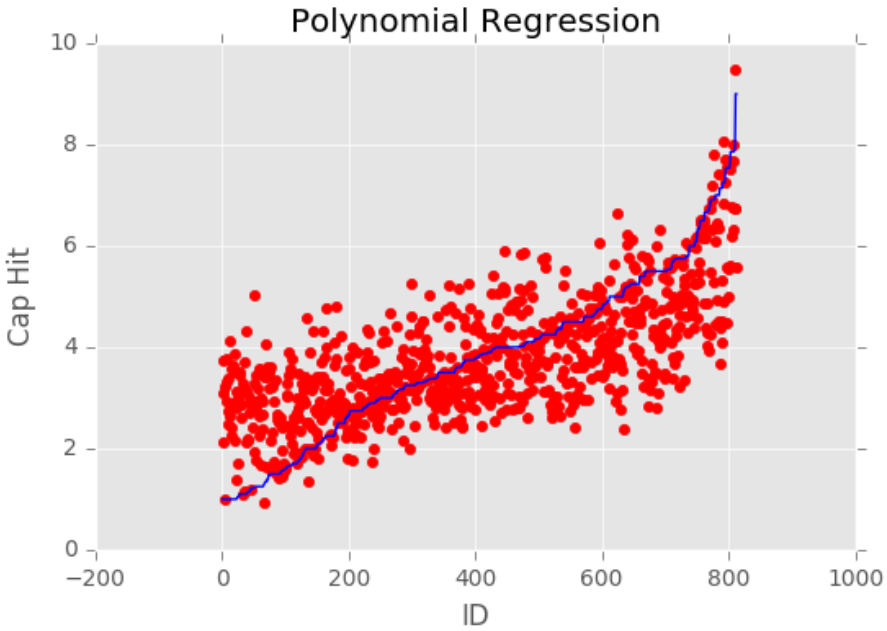


Fig. 1. Visualization of DefensemanU (UFA) with precision of 53,995%

3 Conclusion

The use of appropriate data analytic methods in sports statistics has enormous potential. It is essential to respect the unique characteristics of each sport before we decide which task we want to solve. In our case, we started with a more in-depth investigation of the state-of-the-art. Our motivation was to prevent potentially unsuccessful procedures and got inspiration on how to understand the data about NHL players.
The final model is not perfect. It is essential to continue in this direction, to improve the data sample quality, to try different methods or their combinations.

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

References

1. Wang, B., Wang, L.: Research of Association Rules in Analyzing Technique of Football Match. Second International Conference on Power Electronics and Intelligent Transportation Systems, 178-180 (2009).
2. Chai, B.: Time Series Data Mining Implemented on Football Match. Applied Mechanics and Materials, vol. 26-28, 98-103 (2010).

3. Nunes, S., Sousa, M.: Applying Data Mining Techniques to Football Data from European Championships. Conferência de Metodologias de Investigação Científica (CoMIC'06), 4-16 (2006).
4. Hipp, A., Mazlack, L.: Mining ice hockey: Continuous data flow analysis. In: IMMM 2011, The First International Conference on Advances in Information Mining and Management, 31-36 (2011).
5. MacDonald, B.: An expected goals model for evaluating NHL teams and players. In: Proceedings of the 2012 MIT Sloan Sports Analytics Conference (2012).
6. Buttrey, S.E., Washburn, A.R., Price, W.L.: Estimating NHL scoring rates. Journal Quantitative Analysis in Sports 7 (2011).
7. Jones, M.B.: Responses to scoring or conceding the first goal in the NHL. Journal of Quantitative Analysis in Sports 7(3), 15 (2011).
8. Schuckers, M., Statistical Sports Consulting, L. L. C.: Draft by Numbers: Using Data and Analytics to Improve National Hockey League (NHL) Player Selection. MIT Sloan Sports Analytics Conference (2016).
9. Weissbock, J., Viktor, H., Inkpen, D.: Use of Performance Metrics to Forecast Success in the National Hockey League. Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics co-located with 2013 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2013), 39-48 (2013).
10. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R.: CRISP-DM 1.0 Step-by-Step Data Mining Guide (2000).
11. Shearer, C.: The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Ware-housing 5(4), 13-22 (2000).
12. Schwarz, G.: Estimating the Dimension of a Model. The Annals of Statistics 6(2), 461-464 (1978)
13. Jovic, A., Brkič, K., Bogunovič, N.: A review of feature selection methods with applications. Proceedings of 38th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia (2015)
14. Phuong, T. M., Lin, Z., Altman, R. B.: Choosing SNPs using feature selection. Proceedings IEEE Computational Systems Bioinformatics Conference, 301-309 (2005)

Detekce anomálií v otevřených datech o znečištění ovzduší polétavým prachem

Ondřej Podsztavek[1], Jaroslav Kuchař[1]

[1]Fakulta informačních technologií, České vysoké učení technické v Praze
Thákurova 9, 16000 Praha 6
{podszond,jaroslav.kuchar}@fit.cvut.cz

Abstrakt. Senzorická síť veřejného osvětlení na pražském Karlínském náměstí poskytuje měření znečištění ovzduší polétavým prachem PM_{10} jako otevřená data. V této práci v nich detekujeme anomálie pomocí algoritmů strojového učení pro predikci časových řad a prahování. Chceme, aby se algoritmus strojového učení naučil pravidelnosti v datech a pokud se stane něco neočekávaného, tak to prahováním odhalíme. Experimentovali jsme s lineární regresí a LSTM rekurentní neuronovou sítí, které jsme mezi sebou porovnávali střední kvadratickou chybou. Ukázalo se, že lineární regrese, která predikuje z posledních dvou měření, dosahuje lepších výsledků. Anomálie jsme detekovali z rozdílů predikovaných a skutečných hodnot. Práh pro detekování anomálií jsme vypočítali z histogramu rozdílů predikcí a skutečně naměřených hodnot. Testování ukázalo, že takto navržená metoda dokáže odhalit některé anomálie v měřeních polétavého prachu PM_{10}, ale mnoho anomálií (například postupně nabíhajících) nedetekuje.

Klíčová slova: anomálie, otevřená data, znečištění ovzduší, strojové učení

1 Otevřená data o znečištění ovzduší

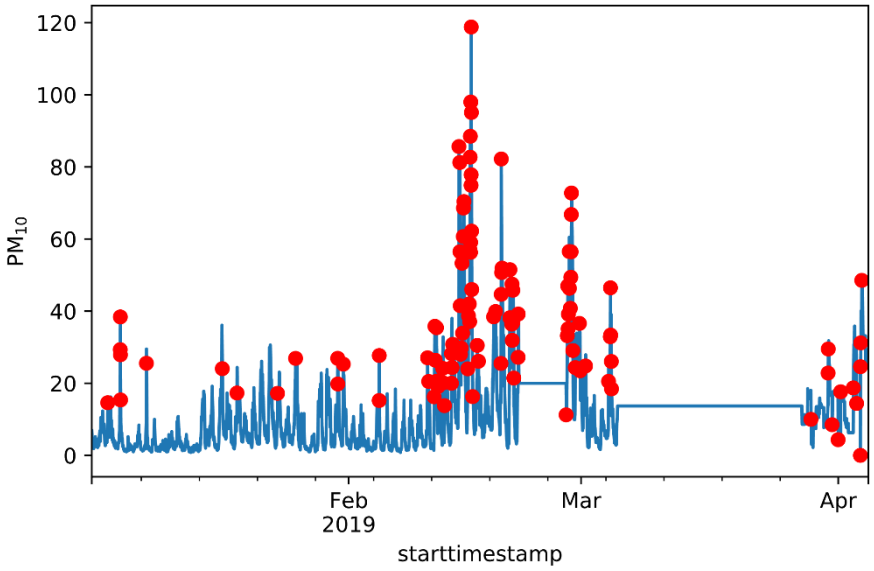
Otevřená data přináší velký prostor pro statistickou datovou analýzu, která může sloužit ke zlepšení kvality životního prostředí. V této práci na otevřená data aplikujeme metody strojového učení. Zabýváme se detekcí anomálií [1, 2] v měřeních polétavého prachu z chytrých lamp, které jsou poskytovány hlavním městem Praha jako otevřená data.

Pražská datová platforma Golemio poskytuje data z pilotního provozu Senzorické sítě veřejného osvětlení, v rámci kterého bylo nainstalováno 92 chytrých pouličních LED lamp v blízkosti Karlínského náměstí na Praze 8. Některé z těchto lamp mají senzory pro měření a sběr dat o hluku, prašnosti ($PM_{2,5}$ a PM_{10}) a množství dalších polutantů (O_3, NO_2 a SO_2). Pevné částice (nebo polétavý prach, angl. *particulate matter*, PM) jsou tuhé nebo kapalné částice v ovzduší o velikosti v rozsahu 1 nm až 100 μm. PM_{10} jsou částice menší než 10 μm a $PM_{2,5}$ částice menší než 2,5 μm. $PM_{2,5}$ je tedy podmnožina měření PM_{10}, a proto hledáme anomálie pouze v PM_{10}.

Naším cílem je prozkoumat detekci anomálií predikcí časové řady následovanou prahovnáním. Chceme, aby se model strojového učení naučil pravidelnosti v datech a

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 66-71.

pokud se stane něco neočekávaného, abychom to pomocí prahování dokázali odhalit. Jako algoritmy strojového učení jsme vyzkoušeli lineární regresi a LSTM rekurentní neuronovou síť [3]. Vyhodnocení na validační množině ukázalo, že lineární regrese dosahuje v našem případě lepších výsledků.[1]



Obr. 1. Detekované anomálie (označené červenými body) v testovacích datech.

2 Předzpracování dat

Náš datový set obsahuje pět různých příznaků (měření $PM_{2,5}$ a PM_{10}, O_3, NO_2 a SO_2), a protože v této práci pracujeme s časovou řadou, zachováme datum a čas měření. Pro potřebu verifikace natrénovaných modelů jsme rozdělili data na trénovací a testovací množinu podle roku měření. Data z druhého pololetí roku 2018 s 32872 záznamy (72,1 %) jsme použili jako trénovací množinu a 12709 záznamů z první poloviny roku 2019 (27,9 %) jako množinu testovací.

Původní data ale obsahují problémy, které by mohly zamezit správnému natrénování algoritmů strojového učení. Data nejsou odečítána v rovnoměrných intervalech, a proto je v rámci předzpracování přesamplujeme každých 15 minut. Dále musíme data přetransformovat do podoby vhodné pro předpovídání časových řad, abychom potom mohli detekovat anomálie, takže z příznaků měření v čase t se budeme snažit předpovědět hodnotu PM_{10} v čase $t + 1$. Nakonec použijeme standardní škálování, takže průměr příznaků je nula a standardní odchylka jednotková. Takto předzpracovaná data umožní správnou selekci příznaků lineární regresí s L1 regularizací.

[1] Kód analýzy s podrobnějšími výstupy a dalšími vizualizacemi je dostupný na webové službě GitHub: https://github.com/podondra/lampy.

2.1 Selekce příznaků

Pro snížení komunikační zátěže při například online monitorování může být přínosná selekce podmnožiny příznaků, kdy není potřeba skrz API dotazovat všechny příznaky. Vhodné příznaky pro lineární modely můžeme vybrat natrénováním lineární regrese s L1 regularizací (tzv. *lasso*), která koeficienty u nevhodných příznaků drží blízko nule, pro různé hodnoty regularizačního parametru. Pro měření úspěšnosti predikce je vhodná odmocnina ze střední kvadratické chyby (anglicky *root mean squared error* neboli RMSE), kterou zvolený model minimalizuje.

Z výše zmíněných důvodů jsme zkusili natrénovat *lasso* pro 23 různých hodnot regularizačního parametru *alpha* na logaritmické škále mezi 10^{-10} a 10^{1}. Ukázalo se, že důležitý je pouze příznak PM_{10}. Hodnota koeficientu PM_{10} byla zhruba 0,9, pokud byla RMSE nízká (0.6387), zatímco koeficienty ostatních příznaků byly stále blízké nule, tj. v intervalu −0,02 až 0,02.

3 Predikce a detekce anomálií

V této práci detekujeme anomálie pomocí predikce časové řady, a to stejně jako v článku [5]. Při detekování anomálií v časových řadách se používá metoda, kdy je daná řada modelem predikována dopředu a následně je porovnána se skutečným měření. Z tohoto porovnání jsou určeny anomálie. Jako vhodné modely jsme zvolili následující:

- základní model (popsaný níže, anglicky *baseline*),
- lineární regresní model
- a rekurentní neuronovou síť (konkrétně dnes nejpopulárnější LSTM).

Tab. 1. Porovnání RMSE pro použité metody na validační množině.

Metoda	RMSE
Základní model	0,6492
Lineární regrese	0,6402
Lineární regrese se dvěma měřeními z historie	0,6380
LSTM	1,0588

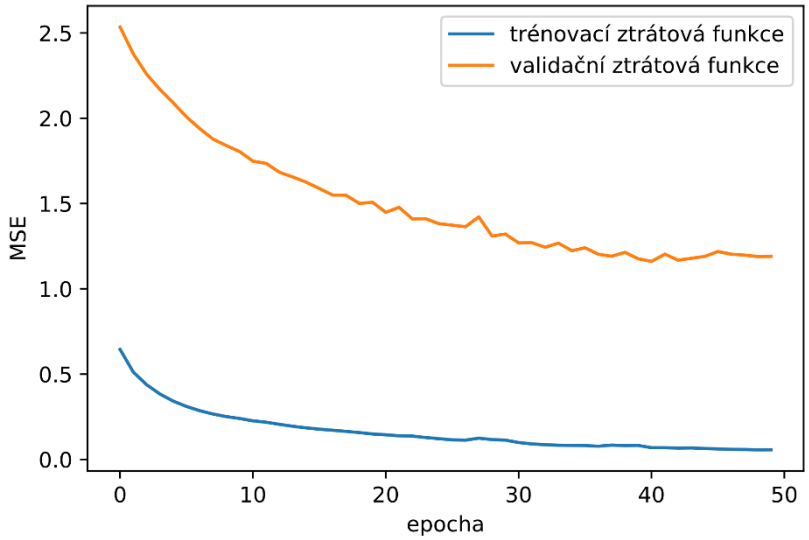
3.1 Základní model

Základní model predikuje vždy předchozí naměřenou hodnotu, a tak poskytuje odrazový můstek, ke kterému budeme moci vztahovat výsledky ostatních modelů. Navíc je jednoduchý na implementaci a dosahuje relativně dobrého RMSE: 0,6492.

3.2 Lineární regrese

Druhým modelem je lineární regrese bez regularizace, protože selekce příznaků ukázala, že stačí příznak PM_{10}. Lineární regrese je jednoduchý model, který bohužel nedokáže zachytit nelineární vztahy a nemůže používat informace z libovolně dlouhé historie (v porovnání s LSTM). Přesto může být vhodným kandidátem pro predikci časové řady znečištění ovzduší, protože a priori neznáme jejich povahu.

Naše pokusy ukázaly, že pokud lineární regrese předpovídá pouze z bezprostředně předcházející hodnoty, její úspěšnost v RMSE je 0,6402. Lepší je nechat lineární regresi předpovídat ze dvou předcházejících měření, kdy je RMSE rovna 0,6380. Větší velikosti historie už výhodné nejsou, protože jejich RMSE je vždy vyšší než 0,7 (experimentovali jsme s velikostí historie 2 až 100).

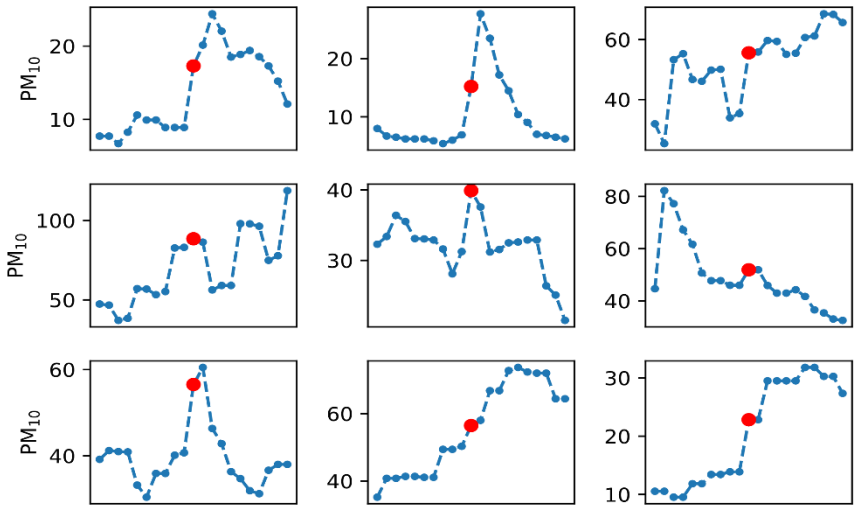


Obr. 2. LSTM byla trénována 50 epoch, kdy již došlo ke konvergenci ztrátové funkce. Rozdíl mezi trénovací a validační hodnotou ztrátové funkce nemusí v případě tohoto regresního problému znamenat přeučení, ale může ukazovat, že ve validačních datech jsou větší anomálie.

3.3 Long Short-Term Memory

Nejpokročilejším prediktivním modelem, který jsme na data aplikovali, je LSTM [3]. Rekurentní sítě se ukázaly být velice efektivní i v aproximování nelineárních vztahů dat, proto nyní zachováme všech pět příznaků jako vstup rekurentní sítě. LSTM by měla být vhodná pro predikci časové řady, protože je navržena tak, aby byla schopna si zapamatovat závislosti z libovolně vzdálené historie. Jako konkrétní architekturu jsme zvolili jednovrstvou LSTM s dimenzí skrytého vektoru 64. Výstupem LSTM musí být pouze jedno číslo, které predikuje měření PM_{10}, a proto je na vlastní buňce LSTM ještě lineární vrstva, která produkuje predikci PM_{10} v následujícím čase.

Takto navrženou LSTM jsme trénovali v současnosti nejpoužívanějším optimalizátorem Adam [4] a MSE ztrátovou funkcí, která je běžná pro trénování rekurentních sítí na regresních problémech, na vstupních sekvencích délky 100 po celkem 50 epoch, kdy již došlo ke konvergenci ztrátové funkce (viz Obr. 2). Výsledné RMSE na validační množině (polovina trénovací množiny) je 1,0588.



Obr. 3. Náhodný výběr 9 z celkem 113 detekovaných anomálií v testovacích datech. (Ostatní anomálie nejsou v grafech zobrazeny.)

3.4 Detekce anomálií

Nejlepších výsledků dosahuje lineární regresní model, který má na vstupu dvě měření z historie, a proto tento model použijeme pro detekci anomálií. Nyní po výběru vhodného prediktivního modelu detekujeme anomálie postupem stejným jako v článku [5]:

1. lineární regresní model s oknem 2 predikuje další měření PM_{10},
2. predikovanou hodnotu $\hat{y}_i$ porovnáme se skutečně naměřenou hodnotou y_i (vypočítáme rozdíl v absolutní hodnotě: $|\hat{y}_i - y_i|$)
3. a pomocí prahování určíme, jestli se jedná o anomálii.

V měření úspěšnosti detekce anomálií je problém absence označení anomálií v datech. Detekce tedy musí být kontrolovány nejlépe doménovým expertem, který dokáže určit také správný práh. Protože takového experta nemáme k dispozici, zvolili jsme jako hodnotu pro prahování hodnotu okraje prvního binu z histogramu s deseti biny: 3,9617. Takto zvolený práh odhalil ve validačních datech celkem 9 anomálií a v trénovačích datech celkem 113 anomálií.

Obr. 1 ukazuje 113 detekovaných anomálií v testovacích datech a Obr. 3 náhodně zobrazuje 9 z nich. Veškeré abnormálně vysoké špičky byly detekovány správně. Také

je vidět, že některé špičky detekovány nebyly, to může být tím, že nárůst je tak pozvolný a za dlouhý časový okamžik, proto by zřejmě bylo vhodnější detekovat skupinové anomálie.

4 Diskuze

Ačkoliv námi popsaný přístup dokáže detekovat některé anomálie, není pro využití v praxi dostatečně robustní. Náš přístup není schopen detekovat anomálie, které nastávají pozvolně, a při použití historie pouze z poslední půl hodiny jsme se mohli stát lehce objetí šumu (naše metoda nerozliší anomálii od chyby v datech). Pro vylepšení naší metody by bylo vhodné přidat další atributy (například den v týdnu, čas, meteorologické informace, předpověď počasí apod.) a získat větší množství dat, protože nemáme měření pro všechny roční období, a to může mít výrazný vliv na přesnost predikce. Uzavíráme, že smysluplnější se určitě jeví hledání skupinových anomálií.

5 Závěr

V této práci jsme se zabývali detekcí anomálií v otevřených datech z chytrých lamp o polétavém prachu PM_{10}. Jako nejlepší se ukázalo předpovídat časovou řadu pomocí lineární regrese (LSTM bylo horší), která na vstup dostane dvě poslední měření. Následně jsme předpověděnou hodnotu porovnali se skutečnou a prahováním určili, jestli hodnota je z hlediska modelu normální nebo ne. Na testovacích datech se ukázalo, že tento postup je schopen detekovat některé anomálie, ale zdaleka ne všechny. Například není schopen detekovat anomálie, které nastávají postupně malými přírůstky nebo často nedokáže rozlišit šum v datech od anomálie. Proto by bylo vhodné do budoucna sesbírat více dat, rozšířit množinu příznaků (např. informace o času nebo počasí) a detekovat anomálie skupinové.

Poděkování: Tento příspěvek vznikl s podporou Fakulty informačních technologií, Českého vysokého učení technického v Praze.

Literatura

1. Aggarwal, C.C.: Outlier Analysis. Springer International Publishing, 2017.
2. Chandola, V., Banerjee, A. and Kumar, V.: Anomaly Detection: A Survey. ACM Computing Surveys, Vol 41(3) (2009), 15:1–15:58.
3. Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory. Neural Computation, Vol 9(8) (1997), 1735–1780.
4. Kingma, D.P. and Ba, J.: Adam: A Method for Stochastic Optimization. 2015. In: Proc. of 3rd International Conference on Learning Representations, San Diego (2015).
5. Shipmon, D.T., Gurevitch, J.M., Piselli, P.M., Edwards S.T.: Time Series Anomaly Detection. arXiv e-prints (2017), arXiv:1708.03665.

Automatické titulkovanie spravodajských relácií v slovenskom jazyku

Ján Staš[1], Martin Lojka[1], Peter Viszlay[2], Daniel Hládek[1], Jozef Juhár[1]

[1]Katedra elektroniky a multimediálnych telekomunikácií, FEI TU v Košiciach
Boženy Němcovej 32, 042 00 Košice
{jan.stas, martin.lojka, daniel.hladek,
jozef.juhar}@tuke.sk
[2]Oddelenie dizajnu a implementácie dátových skladov, Tatrabanka a.s.
Černyševského 50, 851 01 Bratislava
peter.viszlay@gmail.sk

Abstrakt. V tomto článku sú v krátkosti opísané výstupy projektu aplikovaného výskumu APVV-15-0517 zameraného na návrh a vývoj pilotného systému na automatické titulkovanie audiovizuálneho obsahu z oblasti televízneho (ale aj rozhlasového) spravodajstva. Jedným z hlavných cieľov projektu bola eliminácia bariér spôsobených nedostatkom informácií u osôb so sluchovým postihnutím a navýšenie celkového podielu živých spravodajských relácií sprevádzaných skrytými titulkami pomocou moderných rečových technológií. V závislosti od hlasových charakteristík hovoriaceho, štýlu rečového prehovoru a prostredia, v ktorom sa hovoriaci nachádza, navrhnuté riešenie na automatické titulkovanie dosahuje mieru chybovosti rozpoznaných slov v rozmedzí 8,33 až 13,13%.

Kľúčové slová: audiovizuálny obsah, automatické titulkovanie, časové zarovnávanie textu, rozpoznávanie plynulej reči

1 Úvod

V súčasnosti sme obklopení mnohými rečovými aplikáciami, ktoré nám uľahčujú požívanie technológií, ako sú napr. rôzne diktovacie systémy, dialógové systémy na vyhľadávanie informácií, aplikácie ovládané hlasom, a i. Pomocou jednoduchého hlasového príkazu vieme ovládať napr. chytré (smart) telefóny, televízory a pod.

Rečové technológie si našli cestu aj k sofistikovanejším aplikáciám. Žijeme v čase, kedy základom úspechu sú rôzne zdroje informácií a rýchly prístup k nim. Medzi takéto zdroje informácií patria aj audiovizuálne (AV) nahrávky, obsahujúce napr. záznamy akademických prednášok, rozhlasových a televíznych rozhovorov, diskusných relácií, parlamentných debát, pracovných stretnutí, rokovaní zastupiteľstiev a pod. Efektívne a rýchle vyhľadávanie a prezentovanie informácií v AV obsahu nie je triviálna úloha. Základom riešenia úlohy je schopnosť rýchlo a čo najpresnejšie získať informáciu vo forme textového prepisu. So zvyšujúcou sa presnosťou súčasných rečových technológií

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 72-77.

môžeme nahradiť alebo zefektívniť manuálny prepis a vytvoriť plne automatizovaný písomný záznam, či indexovať AV záznamy v rozsiahlych archívoch.

Jednou zo zaujímavých aplikácií rečových technológií je automatizovaná tvorba titulkov pre osoby so sluchovým postihnutím, čím sa táto technológia stáva atraktívnou pre TV vysielanie. Napr. výsledky projektu FP7 SAVAS poskytujú takéto riešenie pre majoritné jazyky EÚ, ako je angličtina, francúzština, nemčina, taliančina, španielčina, či portugalčina [1]. Keďže jadrom týchto systémov sú akustické a jazykové modely, ktoré možno získať len učením na rozsiahlych anotovaných databázach, výsledky tohto projektu nemožno aplikovať na minoritné jazyky, akým je slovenčina. Z toho dôvodu sme pristúpili k návrhu vlastného systému na automatické titulkovanie AV obsahu.

Predchádzajúca verzia systému na automatický prepis mítingových audiozáznamov [7] umožňovala spracovanie viackanálového audiovizuálneho záznamu s automatickou segmentáciou reči, rozpoznávaním pohlavia hovoriacich a paralelným rozpoznávaním reči s viacerými akustickými modelmi, ktoré sa adaptujú na vstupné charakteristiky hovoriaceho. Na zlepšenie presnosti rozpoznávania reči sa využívala kombinácia výstupných hypotéz z viacerých dekodérov, pracujúcich súčasne na vysokovýkonnom serveri. Nedostatkom tejto verzie systému bolo, že pridávaním nových komponentov sa systém stával stále komplexnejším, súčasne ale vykazoval nízku efektivitu prepisu. Navrhli sme preto novú verziu systému s distribuovanou architektúrou, moderným jadrom na automatické rozpoznávanie reči a inovovaným používateľským rozhraním.

V nasledujúcej časti opíšeme základnú architektúru pilotnej verzie systému na automatické titulkovanie spravodajských relácií navrhnutej pre slovenský jazyk, jej tri hlavné moduly, jednotlivé komponenty a predstavíme aj nové používateľské rozhranie.

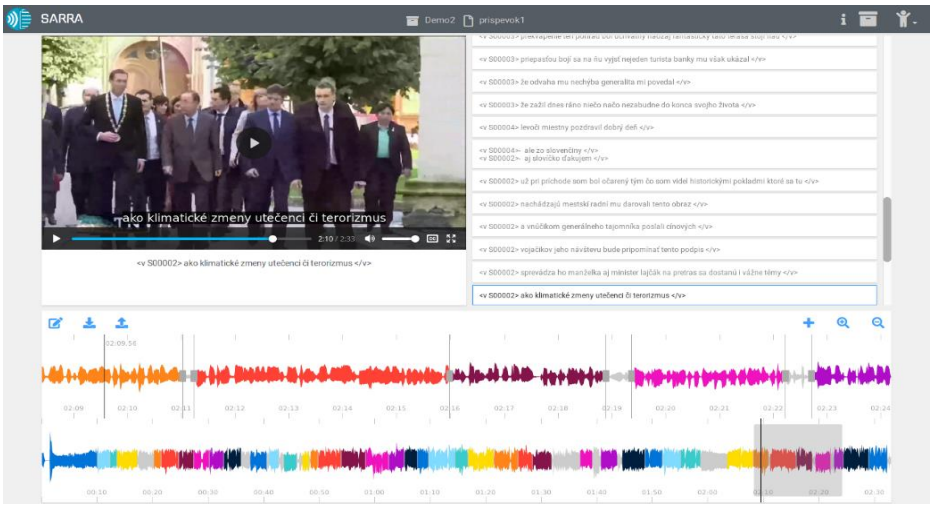
2 Automatické titulkovanie spravodajských relácií

Zaužívané delenie TV vysielania v spravodajstve je na živé vysielanie a vysielanie zo záznamu. V rámci riešenia projektu sme predpokladali aj s ďalším typom AV obsahu, ktorý je síce vysielaný naživo, no existuje k nemu úplný textový prepis. Ide napr. o scenáre k seriálom, divadelným hrám, ale aj texty čítané z čítačky – telepromptera. Výsledkom riešenia projektu bolo preto navrhnúť a vytvoriť systém na automatický prepis spravodajských relácií vysielaných naživo a zo záznamu [2] a systém na časové zarovnávanie AV obsahu s existujúcim textovým prepisom.

2.1 Automatické titulkovanie spravodajských relácií vysielaných zo záznamu

Výhodou offline titulkovania je možnosť vykonať dôkladnú analýzu vstupného AV obsahu a na jej základe zostaviť čo najpresnejší systém na automatický prepis reči do textu a tvorbu titulkov. Z dôvodu vysokej náročnosti procesov vedúcich ku konečnému výsledku bolo najvhodnejším riešením navrhnúť a vytvoriť serverovo-založený systém a poskytovať jeho funkcie v podobe služieb. Používateľ svoj AV obsah sprístupňuje systému cez grafické rozhranie, ten je automaticky spracovaný a výsledok sa vracia používateľovi vo forme titulkov. Bol vytvorený celistvý systém zložený z troch častí:

1. **Úlohový server na styk s používateľom**: Hlavnou funkciou servera je poskytovať interakciu používateľa so systémom na spracovanie úloh pomocou súkromnej databázy úložiska súborov. Systém umožňuje používateľovi vytvoriť novú úlohu, nahrať AV obsah a dopytovať sa na stav a výsledky spracovania úlohy. Server obsahuje tzv. „priehradkový systém“ na zaradenie vybraných úloh do skupín, pričom umožňuje zdieľať tieto priehradky medzi oprávnenými používateľmi. Administrátor systému má k dispozícii prístup k správe používateľov. Úlohou systému je potom zaraďovať jednotlivé úlohy do fronty na ich spracovanie. S používateľom server komunikuje pomocou rozhrania REST, ktoré umožňuje prípadný vývoj ďalších aplikácií, resp. môže poskytovať svoje služby aj iným aplikáciám.
2. **Vykonávacie servery**: Serverové aplikácie bežia nezávisle na vysokovýkonnom serveri a ich úlohou je vyberať z fronty jednotlivé úlohy a vykonávať spracovanie AV obsahu. Ide o jednoduché a univerzálne aplikácie, ktoré na svoj výkon používajú externú aplikáciu naprogramovanú v ľubovoľnom jazyku. Program na spracovanie je tvorený pomocou skriptu, čím je zabezpečená jednoduchá úprava a rozšíriteľnosť vybraných funkcií. Tento koncept zabezpečuje jednoduchú škálovateľnosť systému, kde počet aplikácií bežiacich paralelne ovplyvňuje výkon webovej služby a možno ho prispôsobiť momentálnemu vyťaženiu používateľmi.
3. **Používateľské rozhranie**: Používateľské rozhranie je tvorené webovou aplikáciou, ktorá sa napája na rozhranie REST s úlohovým serverom. Pre administrátora poskytuje prístup k správe používateľov a pre používateľov prístup k jednotlivým úlohám spracovania AV obsahu. Grafické rozhranie obsahuje aj jednoduchý editor titulkov (pozri Obr. 1), ktoré boli systémom vopred vytvorené.



Obr. 1. Vstavaný editor titulkov

Automatické spracovanie AV obsahu potom možno rozdeliť do nasledujúcich úloh, ktoré sú spracované samostatnými vykonávacími serverami:

- automatická segmentácia reči na báze analýzy hlavných komponentov (PCA) [8];

- diarizácia hovoriacich na báze modelov Gaussových zmesí (GMM) [5] a i-vektorov;
- automatické rozpoznávanie plynulej reči na báze hlbokých neurónových sietí (DNN) [3,4] s pokročilou adaptáciou akustických a jazykových modelov [3,6];
- postspracovanie textu po rozpoznaní;
- tvorba titulkov.

Miera chybovosti takto navrhnutého systému na offline titulkovanie spravodajských relácií sa pohybovala v rozmedzí 8,33 až 13,13% WER (word error rate). Hodnotenie bolo vykonávané na dvoch rozdielnych množinách rečových nahrávok z večerných televíznych novín. Prvá množina obsahovala hodinové záznamy televíznych novín s označením hovoriacich v jednotlivých rečových segmentoch o celkovej dĺžke 12 hod. a druhá 10 náhodne vybraných redakčných príspevkov bez označenia hovoriacich v jednotlivých segmentoch a bez možnosti adaptácie systému na hlas hovoriaceho.

2.2 Automatické titulkovanie spravodajských relácií vysielaných naživo

Cieľom titulkovania živého prenosu je poskytnúť titulky s čo najmenším oneskorením. Použitie offline titulkovania by bolo z dôvodu vysokého oneskorenia nemysliteľné. Rozpoznávanie plynulej reči je preto prispôsobené behu naživo. Vstupom je živý zvuk zaznamenaný mikrofónom. Dekódovanie prebieha súčasne s prichádzajúcim obsahom. V každom časovom okamihu je dostupná predbežná hypotéza toho, čo bolo reálne vyslovené (resp. rozpoznané). Počas online titulkovania sú vykonávané tieto kroky:

- automatická segmentácia reči a rozpoznávanie plynulej reči so sledovaním pauzy v definovanom časovom úseku;
- adaptácia akustických modelov na báze i-vektorov (bez adaptácie modelu jazyka);
- postspracovanie textu po rozpoznaní;
- tvorba titulkov.

Miera chybovosti systému online titulkovania dosahuje úroveň približne 20% WER. Problémovými sú najmä segmenty obsahujúce množstvo mimoslovníkových slov.

2.3 Časové zarovnávanie (synchronizácia) titulkov s audiovizuálnym obsahom

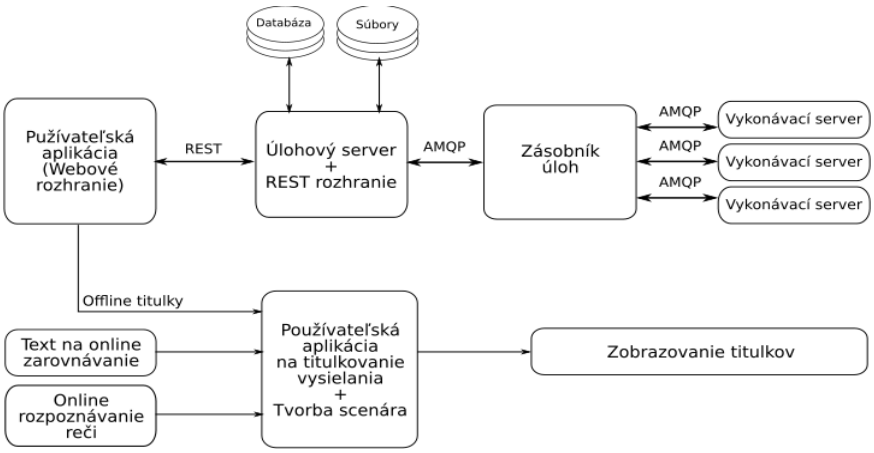
Tak ako v prípade online prepisu, aj v tomto prípade ide o rozpoznávanie reči naživo. Rozpoznávacia sieť je ale limitovaná obsahom textu v titulkoch, ktorý je už vopred známy. Výsledkom je zníženie výpočtových nárokov, zlepšenie presnosti prepisu a rýchlejšia odozva systému (titulky sa zobrazujú so začiatkom rečového prehovoru). Na riadenie ďalších krokov sa využíva sledovanie aktuálnej hypotézy z rozpoznávania:

- tvorba rozpoznávacej siete z predspracovaných titulkov;
- automatická segmentácia reči sledovaním výstupných hypotéz;
- adaptácia akustických modelov na báze i-vektorov.

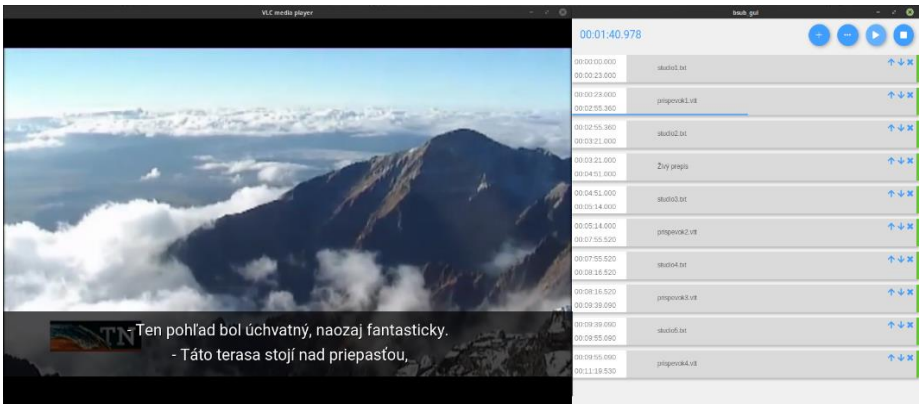
Presnosť časového zarovnávania v závislosti od typu AV obsahu, kvality rečového prehovoru v danom rečovom segmente a dĺžky segmentov dosahuje úroveň až 98%.

2.4 Používateľská aplikácia

Na prepojenie uvedených čiastkových riešení (online a offline titulkovania a časového zarovnávania textu na základe architektúry systému zobrazenej na Obr. 2) sme vytvorili aplikáciu (na Obr. 3), ktorá môže byť prínosná hlavne pre poskytovateľov AV obsahu, či už ide o živé vysielanie alebo zo záznamu. Aplikácia pozostáva z rozhrania, ktoré je možné použiť na vytvorenie scenára televízneho vysielania spravodajstva. Dochádza tu k automatickému prepínaniu navrhnutej trojice systémov. Výsledkom je potom neprerušené vysielanie s titulkami zobrazenými v polopriehľadnom okne. Pripravenosť scenárov a konfigurácií je indikovaná farebne. Jednotlivé scenáre je možné presúvať potiahnutím, pričom sa automaticky mení aj časovanie scenára. Aplikácia demonštruje, čo je možné s vytvorenými rečovými technológiami v súčasnosti dosiahnuť a možno ju prispôsobiť požiadavkám používateľa, príp. zabudovať do už existujúcich systémov.



Obr. 2. Architektúra systému na automatické titulkovanie spravodajských relácií



Obr. 3. Používateľská demonštračná aplikácia na prepojenie offline a online titulkovania a časového zarovnávania textu s audiovizuálnym obsahom na tvorbu scenára TV spravodajstva

3 Záver

Vytvorili sme tri pilotné vezie systému na automatický prepis spravodajských relácií, ktoré ďalším vývojom môžu dospieť k jednoduchej aplikácii na každodenné používanie redaktormi v TV vysielaní, publikujúcimi na webových portáloch, či rôznymi inými používateľmi, ktorí vytvárajú napr. vlastné video blogy. Prvá verzia je zameraná na prepis spravodajských príspevkov, ktoré sú vytvorené s dostatočným predstihom pred vysielaním redaktormi pracujúcimi v teréne. Tieto sú prepísané offline systémom na rozpoznávanie reči v čo najlepšej kvalite, pričom redaktor ich má možnosť upraviť a prispôsobiť vysielaniu pomocou vytvorenej webovej služby, ktorá obsahuje vstavaný editor titulkov. Druhá verzia uvažuje so živými vstupmi redaktorov v TV vysielaní. Vytvorili sme rýchly online rozpoznávací systém s logikou prerozdeľovania a zobrazovania titulkov s čo najmenším oneskorením. Tretia verzia systému umožňuje živé zarovnávanie textu s AV obsahom a zobrazovaním titulkov v časovom predstihu. Používateľská aplikácia prepája čiastkové riešenia vo forme scenára TV vysielania.

Poďakovanie: Tento výskum bol realizovaný vďaka podpore Vedeckej grantovej agentúry MŠVVaŠ SR a SAV realizáciou projektu VEGA 1/0511/17 a vďaka podpore Agentúry na podporu výskumu a vývoja realizáciou projektu aplikovaného výskumu APVV-15-0517 a projektu bilaterálnej medzinárodnej spolupráce SK-TW-2017-0005.

Literatúra

1. Álvarez, A., Mendes, C., Raffaelli, M., Luís, T., Paulo, S., Piccinini, N., Arzelus, H., Neto, J., Aliprandi, C., Del Pozo, A.: Automating live and batch subtitling of multimedia contents for several European languages. Multimedia Tools and Appl. 75 (2016), 10823-10853.
2. Lojka, M., Viszlay, P., Staš, J., Hládek, D., Juhár, J.: Slovak broadcast news speech recognition and transcription system. In: Proc. of NBiS, Bratislava, Slovakia (2019).
3. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Proceedings of INTERSPEECH, Dresden, Germany (2015), 3214-3218.
4. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: Proc. of ASRU, Waikoloa, Hawaii, USA (2011).
5. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meignier, S.: An open-source state-of-the-art toolbox for broadcast news diarization. In: Proceedings of INTERSPEECH, Lyon, France (2013), 1477-1481.
6. Staš, J., Hládek, D., Lojka, M., Juhár, J.: Dual-space re-ranking model for efficient document retrieval, user modeling and adaptation. In: Proc. of Elmar, Zadar, Croatia (2018), 203-206.
7. Staš, J., Viszlay, P., Lojka, M., Koctúr, T., Hládek, D., Juhár, J.: Automatic transcription and subtitling of Slovak multi-genre audiovisual recordings, In: Vetulani, Z., Mariani, J., Kubis, M. (eds): Human Language Technologies: Challenge for Computer Science and Linguistics, LNCS 10930, Springer, Cham (2018), 42-56.
8. Vavrek, J., Viszlay, P., Kiktová, E., Lojka, M., Juhár, J., Čižmár, A.: Query-by-example retrieval via fast sequential dynamic time warping algorithm. In: Proc. of TSP, Berlin, Germany (2014), 469-473.

Využitie nositeľných zariadení na rehabilitáciu pacientov trpiacich symptómom zmrazenia chôdze

Pavol Šatala[1], Vladimír Haň[2], Petra Levická[3], Peter Butka[1]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
{pavol.satala, peter.butka}@tuke.sk
[2]Neurologická klinika, Lekárska fakulta, Univerzita Pavla Jozefa Šafárika v Košiciach
vladimir.han@upjs.sk
[3]Lekárska fakulta, Univerzita Pavla Jozefa Šafárika v Košiciach
petra.levicka@student.upjs.sk

Abstrakt. V tejto práci sa zaoberáme využitím a účinnosťou nositeľných zariadení pri rehabilitácií pacientov trpiacich Parkinsonovou chorobou. Pomocou nositeľných zariadení vytvárame pomôcky, ktoré uľahčujú pacientom prekonať epizódy zmrazenia chôdze. V práci prinášame prvé výsledky dosiahnuté počas testovaní týchto pomôcok. Výsledky preukazujú ich účinnosť.

Kľúčové slová: nositeľné zariadenia, Parkinsonova choroba, zmrazenie chôdze

1 Úvod

Zmrazenie chôdze patrí medzi typické motorické symptómy Parkinsonovej choroby. Ide o náhlu neschopnosť pokračovať v chôdzi bez zjavnej vonkajšej príčiny [1]. Prejavuje sa predovšetkým pri prechode zúženými miestami, otočkách, začiatkoch chôdze prípadne pri zmene povrchu, po ktorom pacient kráča [2]. Pacient počas zmrazenia nestojí nehybne. Jeho nohy sa pokúšajú urobiť ďalší krok, avšak vyzerajú ako prilepené k zemi. Na prekonanie tejto udalosti lekárska veda navrhla niekoľko kognitívnych pomôcok: vizuálne (kolmé čiary na zemi, pohybujúci sa svetelný bod, alebo loptička, ktorú ma pacient nasledovať), akustické (rytmické zvuky, hudba), vibračné **(**rytmické vibračné impulzy).

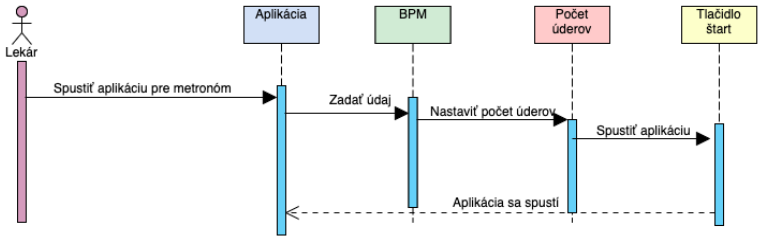
V tejto práci sme sa rozhodli spomínané pomôcky implementovať do nositeľných zariadení a následne otestovať ich účinnosť.

2 Implementácia

Pomôcky boli implementované ako Android aplikácia pre inteligentné telefóny s OS Android. Komunikácia prebiehala pomocou Bluetooth komunikácia, kde z pohľadu ISO/OSI modelu išlo o výmenu balíčku na vrstve TCP/IP. Aplikáciu ovládal zaškolený medicínsky pracovník. Na pacienta boli umiestnené inteligentné hodinky, ktoré

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 78-81.

poskytovali vibračné impulzy a slúchadlá, ktoré zabezpečovali akustické pomôcky. Schéma fungovania aplikácie je viditeľná na Obr. 1.



Obr. 1. Schéma využitia aplikácie

3 Experiment

Účinnosť nášho riešenia bola testovaná počas meraní na Neurologickej klinike univerzitnej nemocnice v Košiciach. Testovacia trasa bola tvorená dvoma miestnosťami. Úlohou pacienta bolo postaviť sa zo stoličky v prvej miestnosti, prejsť 4m, prejsť cez zúžený priestor, ktorý bol tvorený dverami a pokračovať do druhej miestnosti. V druhej miestnosti pacient znova po 4m urobil otočku o 180 stupňov a vrátil sa naspäť po rovnakej trase. Táto trasa bola navrhnutá tak, aby sa na čo najkratšej vzdialenosti vyskytovalo čo najviac kritických miest. Konkrétne išlo o začiatok chôdze, dvere, otočka a dvere na ceste naspäť. Cely priebeh sledoval zdravotnícky pracovník, ktorý značil časy, výskyty zmrazenia ako aj celkovú kvalitu chôdze.

4 Vyhodnotenie

Získané údaje sme vyhodnocovali v dvoch ukazovateľoch. Prvým ukazovateľom bolo skrátenie času chôdze. Pri zachovaní rovnakej trasy s konštantnou dĺžkou znamená zníženie času chôdze zvýšenie jej rýchlosti. Druhým ukazovateľom bolo zníženie početnosti výskytu FoG. Keďže namerané údaje boli pre každého pacienta výrazne odlišné a výsledné dáta mali veľmi veľký rozptyl pre štatistické spracovanie sme vypočítali percentuálne zlepšenie parametrov pre jednotlivých pacientov pomocou nasledujúceho vzorca:

$$x = \frac{(t_k - t_p)}{t_k}$$

kde *x* nám udáva, koľko percent z času chôdze počas kontrolnej chôdze tvorí čas o ktorý sa nám podarilo skrátiť chôdzu pacienta. Premenná t_k nám udáva dĺžku kontrolnej chôdze a t_p dĺžku chôdze s pomôckou. Rovnaký spôsob je použitý aj na výpočet percentuálneho zníženia výskytu prípadov zmrazenia chôdze.

Testovanie prebehlo na vzorke 11 pacientov. Sedem pacientov bolo mužského pohlavia a štyri ženského. Priemerný vek pacientov bol 66,44 ± 4.59 rokov. Priemerné trvanie chôdze počas kontrolnej chôdze bolo 42.43 ± 19.43s. Podrobný prehľad výsledkov uvádza tabuľka č. 1 a 2. Pri interpretácií výsledkov treba poukázať na malý počet meraní pri vibračnej pomôcke s frekvenciou 2p.

Tabuľka č. 1: Absolútne hodnoty parametrov nameraných počas experimentu

Typ merania	Počet meraní	Dosiahnutý čas		Výskyt FoG	
		Hodnota	SD	Hodnota	SD
kontrolná	11	42,43	19,43	1,95	1,27
akustická 60BPM	5	28,03	8,61	0,31	0,64
akustická 90BPM	7	31,89	13,16	0,20	0,35
akustická 120BPM	6	25,27	7,26	0,33	0,40
vibračná p/2	5	35,21	15,40	0,45	0,93
vibračná p	5	34,12	15,34	0,18	0,40
vibračná 2p	2	26,23	9,93	0,00	0,00

Tabuľka č. 2: Priemerné relatívne vyjadrenia zlepšení pre jednotlivé typy pomôcok oproti kontrolným meraniam bez pomôcok

Typ merania	Počet meraní	Zlepšenie čas v %		Zlepšenie FoG %	
		hodnota	SD	hodnota	SD
akustická 60BPM	5	9,29	15,81	64,28	41,64
akustická 90BPM	7	22,16	9,89	71,87	36,44
akustická 120BPM	6	23,02	9,92	58,33	37,63
vibračná p/2	5	2,45	15,50	30,00	57,00
vibračná p	5	5,88	7,03	60,00	0,41
vibračná 2p	2	11,10	13,80	100,0	0,00

Následne sme pre ďalšie štatistické testovanie formulovali hypotézy:

- H0: Akustické pomôcky nemajú vplyv na rýchlosť chôdze pacienta trpiaceho FoG,

- H1: Akustické pomôcky majú vplyv na rýchlosť chôdze pacienta trpiaceho FoG.

Na testovanie sme vybrali množinu údajov nameraných s akustickou pomôckou s o frekvencii 90BPM z dôvodu najväčšieho rozsahu vzoriek. Pred samotným testovaním hypotézy sme pomocou Shapiro-Wilkonovho testu overili, že výberový súbor pochádza z normálneho rozdelenia. Následne sme vykonali t-test strednej hodnoty na potvrdenie, alebo vyvrátenie stanovených hypotéz. Vykonane T testy potvrdili, že akustické pomôcky majú štatisticky významný vplyv na rýchlosť chôdze ($p = 0.001$). Priemerne zlepšenie pri nami testovanej akustickej pomôcke o frekvencii 90BPM bolo 22%. Rovnako pozitívny výsledok sa preukázal aj pri testovaní zníženia počtu výskytov FoG ($p=0.002$) s priemerným znížením o 62,5%.

Z vibračných pomôcok sme pre testovanie vybrali pomôcku s polovičnou frekvenciou bežnej frekvencie krokov. Pri tej sa nám na stanovenej hladine významnosti a dosiahnutom rozsahu meraní nepodarilo jednoznačne štatisticky potvrdiť zvýšenie rýchlosti ($p = 0.1348$). Priemerne zrýchlenie dosahuje 6%. Zníženie výskytu FoG sa však na hladine významnosti 0.05 štatisticky potvrdilo ($p = 0.016$). Priemerne zníženie výskytu FoG dosahuje až 80%.

5 Záver

Ako ukázal náš experiment pomôcky vytvorené pomocou inteligentných zariadení sú funkčné a porovnateľné s bežne lekármi odporúčanými pomôckami. Rovnako však experiment ukázal, že účinnosť pomôcky je výrazne individuálna a pre každého pacienta môže byť najvýhodnejší iný typ pomôcky. To otvára dvere pre ďalšiu prácu a vývoj algoritmu, ktorý bude schopný prispôsobiť použitú pomôcku individuálne pre každého pacienta.

Poďakovanie: Táto práca bola podporená Agentúrou na podporu výskumu a vývoja v rámci grantu č. APVV-16-0213 a APVV-17-0550. Taktiež tento príspevok vznikol s podporou VEGA 1/0493/16.

Literatúra

1. Velik, R.: Effect of On-Demand Cueing on Freezing of Gait in Parkinson's Patients. International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering no. 6, (2012), 280 - 285.
2. Schaafsma, J. D., Balash, Y., Gurevich, T., Bartels, A. L., Hausdorff, J. M., Giladi, N.: Characterization of freezing of gait subtypes and the response of each to levodopa in Parkinson's disease. European Journal of Neurology no. 4, (2003) 391-398
3. Rocha, P.A., Porfírio G.M., Ferraz, H.B., Trevisani, V.F.M.: Effects of external cues on gait parameters of Parkinson's disease patients: A systematic review in Clinical Neurology and Neurosurgery 124 (2014) 127–134

Empirical Evaluation of Explainability of Topic modelling and Clustering Visualizations

Oliver Genský, Jiří Žárský, Tomáš Kliegr

Fakulta informatiky a statistiky, VŠE v Praze
Nám. W Churchilla 4, 13067 Praha
Genoli978@gmail.com, {xzarj08,tomas.kliegr}@vse.cz

Abstract. In this paper, we report on two user studies aimed at evaluating comprehensibility of several frequently used visualizations of models generated by non-hierarchical clustering algorithms. The first user study was performed on numerical data from the telco domain and the second study was performed on textual data from the news domain (disinformation tweets). Survey methodology was adopted from study of comprehensibility of decision trees by Piltaver et al, 2016, and included classify, explain and validate tasks and three measures (correct answers, subjective comprehensibility, and time required). The visualizations investigated in the first case study were cluster centroids (shown as bar charts), Z-scores (shown as bar charts and heatmaps) and decision trees learnt to predict cluster membership. Based on responses from 55 participants, decision trees were found to be most effective for numerical data. The visualizations investigated in the second case study were word clouds computed from TF-IDF frequencies, word clouds from TF-IDF Z-scores, and Z-scores shown as bar charts. For this study, we pre-registered our hypotheses, procedure, and planned analyses prior to data collection at osf.io and participants were recruited via crowdsourcing. Based on responses from 188 participants, word clouds generated from z-scores were found to produce most effective visualizations of clusters generated from textual data.

Keywords: topic modeling, clustering, explainability, machine learning crowdsourcing, disinformation, fake news

1 Introduction

There are numerous ways for presenting results of clustering and topic modeling algorithms, including word (tag) clouds, decision trees, and z-scores visualized as heatmaps. So far, there has been little empirical research on understanding the strengths and weaknesses of the individual visualizations. In this paper, we report on two user studies investigating which visualizations are most interpretable depending on type of data (textual or numerical) and the task at hand.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 82-86.

This paper is organized as follows. The methodology used for measuring comprehensibility of clustering visualizations is covered in Section 2. Section 3 briefly describes the visualizations, which were subject to the analysis. Section 4 reports on the user studies. The conclusions summarize the results.

2 Measuring comprehensibility of clustering

There is a paucity of prior work on measuring comprehensibility of clustering through human-subject experiments. The closest applicable study that we identified was performed by Piltaver et al., 2016 on comprehensibility of decision trees. Their methodology is composed of the four types of tasks, which are participants required to complete, along with corresponding metrics, such as answer correctness or time required to work out the answers. We adapted three of these tasks for the evaluation of comprehensibility of clustering:

- **classify** – assign a new instance to a cluster or a topic,
- **explain** – identify cluster, which is the best match for a given question,
- **validate** – answer a closed question based on the presented clustering.

The measures used to gauge participants' performance was also adapted from Piltaver et al., 2016:

- **percentage of correct answers,**
- **time taken to complete the questionnaire,**
- **subjective comprehensibility stated by the participants.**

Note that Piltaver et al., 2016 also included a fourth task, discover, which asked participants to determine which property is unusual for a given task. We did not transpose the discover task into our methodology, as we found it difficult to translate to the unsupervised learning context.

3 Visualizations of clustering algorithms

The clustering visualizations, which were subject of our research, were initially selected based on Linoff, 2004 and Tsiptsis and Chorianopolous, 2010. In the following, we provide their brief description.

3.1 Numerical datasets

We consider the following visualizations: (1) *Bar charts with centroids.* There is one bar chart for each attribute displaying its average values for all clusters. The bar chart also shows whether the value for the given cluster is above or below the average in the complete dataset. (2) *Z-score bar chart* visualization with one bar chart per cluster. The bar for an individual attribute is the higher the more the average of attribute values computed from instances in the cluster differs from the average computed from the

complete data set. (3) *Heat map of Z-scores* is presented as a matrix with clusters on the horizontal axis and attributes on the vertical axis. The higher the Z-score for an attribute-cluster combination, the darker the value of the corresponding cell. (4) *Decision tree* trained to predict cluster labels with leaf nodes depicting the per cluster probability distribution.

3.2 Textual data

From the visualizations described above, we did not consider decision trees to be suitable for textual data. The remaining visualizations were adapted for textual data as follows: (1) *Word clouds of average values. A*verage TF-IDF values for individual words for instances within cluster. There was one cloud per cluster with size of words increasing with the average. *(2) Z-score bar charts as outlined above.* Only attributes with five highest and five lowest Z-score values within each cluster were shown. *(3) Word clouds generated from Z-scores.*

4 Research questions

The primary goal of our research was to identify visualizations that are most suitable for individual tasks (classify, explain, validate) in combination with the type of data analysed (numerical, textual).

Additionally, we wanted to empirically validate the following two hypotheses:

1. Correctness of answers rises with increasing levels of perceived subjective comprehensibility.
2. Time needed to complete the survey decreases with increasing levels of perceived subjective comprehensibility.

5 Case studies

To validate the research questions, we analysed data from two user studies, which primarily differed in the type of attributes in the input data (only numerical or only textual). A case with mixed attribute value types was left for future work.

5.1 Telco case study - only numerical attributes

The input dataset was retrieved from the online platform BigML (BigML.com, 2013). It contains information about customer behavior of a telecommunication company, such as the number and duration of phone calls by time of day (evening, day, night, …). This data set was used to train a clustering model. Experimentation with several algorithms and their parameters led to the choice of the k-means algorithm with k=5. We then performed a user study assessing the four visualizations listed in section 3.1 in which 55 students from the University of Economics participated.

Results and discussion. As we expected, the percentage of correct answers for the classify task was the highest for the decision tree visualization, which had also overall most correct answers across all three tasks. Participants working with decision trees also took on average the shortest time to complete all tasks, this was most pronounced for the classify task, where they were the fastest overall.

For the two additional hypotheses, we found a statistically significant correlation between correctness of answers and subjective comprehensibility for all visualizations independently ($p < 0,05$). The second hypothesis was not confirmed, as time to complete *rose* with increasing subjective comprehensibility ($r = 0,318$, $p = 0,018$). More detailed discussion of results can be found in Genský, 2018.

5.2 Twitter case study - textual dataset

The second case study used Twitter data for document clustering. The data set used contained approximately three million tweets known as "Russian troll tweets", which were posted during the U.S. Presidential Election campaign in 2016 (U.S. House of Representatives, 2017). As in the previous case study, after experimentation with several algorithms and settings, the dataset was clustered with k-means (k=6). We sought to make our research transparent by pre-registering our hypotheses, procedure, and planned analyses prior to data collection with the Open Science Foundation at osf.io.

We then performed a user study assessing the three visualizations listed in section 3.3 in which 188 respondents recruited via the Prolific Academic crowdsourcing platform participated. Pre-screening of participants assured that only respondents with higher education (BA, MSc or PhD) and living in English-speaking countries (Ireland, US and UK) were eligible. Additional filtering was performed by three simple validation questions about U.S. Presidential Election campaign in 2016.

Results and discussion. We confirmed our assumption that *Word cloud generated from Z-scores* outperforms other two visualizations in terms of the ratio of correct answers (83 % vs. 70 % and 64 %). The largest difference was observed for the validate and classify tasks, the explain task had similar ratio for all visualizations. The type of visualization affects the correctness of answers ($p < 0,01$) and the type of visualization plays a statistically significant role in the time taken to complete the questionnaire too ($p < 0,05$). Participants working with *Word cloud generated from Z-scores* took on average the shortest time to complete the questionnaire (12,91 minutes) but participants with basic *Word cloud* took a similar time (13,04 minutes). On the contrary, participants presented with *Z-score charts* took on average the longest time (15,3 minutes).

For the two supplementary hypotheses, we found a statistically significant correlation between correctness of answers and subjective comprehensibility for all visualizations independently ($r = 0,227$, $p < 0,01$). The second hypothesis was also confirmed. Time needed to complete the survey was found to decrease with increasing levels of perceived subjective comprehensibility ($r = 0,25$, $p < 0,01$).

6 Conclusion

We conducted two human-subject experiments to find the most suitable visualization techniques for presenting results of non-hierarchical clustering algorithms. Our findings indicate that for numerical datasets, decision tree provides the best option. For textual data, our results lead us to recommend word clouds computed from z-scores. In future work, we plan to extend this preliminary report with a more detailed account of the user studies and their results.

Acknowledgment: This paper was supported by grant IGA 33/2017.

Literature

1. Alnajran, N., Crockett, K., McLean, D., Latham, A., 2017. Cluster Analysis of Twitter Data: A Review of Algorithms:, in: Proceedings of the 9th International Conference on Agents and Artificial Intelligence. Porto, Portugal, pp. 239–249. https://doi.org/10.5220/0006202802390249
2. BigML.com, 2013. Churn in Telecom's dataset [WWW Document]. BigML.com. URL http://bml.io/16N3u9z (accessed 7.2.19).
3. Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.
4. Genský, O., 2019. Evaluation of cluster comprehensibility. University of Economics in Prague.
5. Gordon S. Linoff. Data mining techniques: for marketing, sales, and customer relationship management. John Wiley & Sons, 2004.
6. Piltaver, R., Luštrek, M., Gams, M., Martinčić-Ipšić, S., 2016. What makes classification trees comprehensible? Expert Systems with Applications 62, 333–346. https://doi.org/10.1016/j.eswa.2016.06.009
7. Tsiptsis, Konstantinos K., and Antonios Chorianopoulos. Data mining techniques in CRM: inside customer segmentation. John Wiley & Sons, 2011.
8. U.S. House of Representatives, 2017. Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements [WWW Document]. URL https://intelligence.house.gov/social-media-content/ (accessed 6.30.19).

Normalization of Business Processes

Václav Řepa[0000-0001-9656-5564]

Fakulta informatiky a statistiky, Vysoká škola ekonomická v Praze
Nám. W. Churchilla 1938/4, 13067 Praha
repa@vse.cz

Abstract. In spite of the unquestioned importance of business process modeling in both the information systems development and the enterprise development, the methodology standards in this field are still insufficient. One of the most important aspects that the process models have to cover is the essential unity of object oriented and process oriented views of a business system. In this paper, we introduce so-called Process Normalization technique, a part of the MMABP methodology, as a particular way of methodical covering of this essential unity. This technique is freely inspired with the famous 'ancient' Normalization of Data Structures technique, which we regard as relevant right because of the essential unity of objects and processes in the business system.

Keywords: business process model, normalization, structured programming, object orientation, process orientation.

1 Introduction

The ideas expressed in this paper come from the Methodology for Modeling and Analysis of Business Processes – MMABP [1]. Its Principle of Modeling expresses the presumption that every organization as an implementation of some business system (business idea) must be based on the model of the relevant part of the Real World. MMABP also distinguishes between the Real World: structure (object view) and behavior in the Real World (process view). In both dimensions, there are two basic types of model: global (system) view on the system as a whole and detailed (particular) view on just one element of the system. Sufficient business system modeling methodology thus has to support both types of models (global and detailed) in both dimensions (structural and behavioral) and allow working with both dimensions in their mutual interconnections. At the same time, the methodology has to integrate also both essential components of a business system: its managerial and technology contents. Proposed Process Normalization technique tries to cover all those aspects offering a certain way of integration of the global process map with the process details, intentional processes with conceptual objects, and the process structure with their managerial meaning.

To understand the 'business essence' of the collaboration of processes in terms of ideas of process-driven management [4], one primarily has to differentiate between two basic functional types of processes. **Key processes** are those processes in the

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 87-92.

organization that are *linked directly to the customer*, covering the whole business cycle from expression of the customer need to its satisfaction with the product / service. **Support processes** are *linked to the customer indirectly* - by means of key processes, which they are supporting with particular products / services. This way, every process is ultimately connected to the value for the customer either directly (key process) or through its services for other processes. Key processes thus represent a *specific enterprise's way* of satisfying the customer needs while support processes represent rather *standard functionality* often connected with some technology. Consequently, key processes are *very dynamic*, often changing, permanently developing, every instance is an original. Support processes are *mostly static*, stable, offering standardized and multiply usable services. Therefore, the main effort in the process of creating the conception of the system of processes must be establishing the *equilibrium of needed dynamics of key processes on one hand and the necessary stability of the system ensured with its maximally standard support processes*. The proposed technique for process normalization contributes to the keeping of the above-mentioned equilibrium. It forces the modeler to respect the essential relationships among particular types of process structures and corresponding structures of business objects, what is explained in more detail in the following sections.

2 Process normalization technique

Normalization of processes is a technique freely inspired with the Normalization of Data Structures technique firstly introduced by E.F.Codd in [2], then elaborated in further detail with R.F.Boyce in [3] and comprehensively explained in [9]. Although the original Codd's intention was mainly technical in terms of a database system design, this technique started uncovering the essential Principle of Modeling in the field of information systems development that has been later defined by P.Chen in [5]. As this principle is essential also in terms of its validity in all dimensions of the Real World models, it has to be valid even for process descriptions. Regarding this fact together with the aforementioned essential unity of objects and processes in the business system as it is defined in MMABP [1], the goals of the Normalization of Processes are: (a) To reduce redundancy of process activities. (b) To ensure that all activities and non-initial events are dependent on the initial event. (c) To eliminate unnecessary hidden dependency relationships within a process.

Redundancy of process activities means unnecessary repeating of activities with essentially the same content and meaning in different processes. Unrecognized redundant occurrence of activities with the same contents in different processes is one of typical consequences of hierarchical organization mentioned by Hammer in [4] as a 'symptom of broken processes'. The goal (b): unconditional clear dependency of all activities on the initial event is a certain way to ensuring the relationship to the customer-oriented value of the performance of the company. This goal, together with the goal (c): no hidden dependencies support mainly the principle of key processes as centerlines of the final meaning of all activities in the organizations. Using the Process Normalization technique the key processes are relieved of all activities, which signalize the existence of possible process goals, other than the primary process goal, expressed

with the dependency on the initial event. These supportive process sub-goals usually signalize the existence of other, more general, sub-processes, which are focused on some specific goal and are hidden in the body of the normalized process. These processes represent the set of supporting actions with more general meaning that are necessary as a step on the way to the final process target. Their general meaning is then the reason for their removal from the body of the normalized process and establishing them as standalone support processes. This way the normalized process is freed of all non-essential activities that are removed to more general support processes and the necessary relationships between the original process and the new ones are uncovered in terms of the meaning of support activities in the context of the goal of the normalized process.

3 The procedure of process normalization

The procedure of process normalization is a simple sequence of steps by particular normal forms. The input assumptions for the normalizing procedure are:
(i) the logical process represents a part of the Real World consisting of natural process chains and their relationships.
(ii) Each activity in the process represents an activity from some natural process chain or relationship among process chains.
(iii) Each event in the process represents an activity of some external actor or related (collaborating) process chain.
(iv) Each natural process chain hidden in the logical process can be uniquely identified by some event or by a logical structure of events. Such event (structure of events) we call 'initial event'.
The initial condition for each step is that the process is in the previous normal form (i.e. fulfills its required characteristics).

1st Normal Form (iterative generalizable structures free). The process is in the 1st Normal Form if the bodies of all its repeating non-elementary structural parts (iterations) have been removed to standalone processes and replaced with process states.

2nd Normal Form (alternative generalizable structures free). The process is in the 2nd Normal Form if it is in the 1st Normal Form and the bodies of all its mutually alternative non-elementary structural parts (selections) have been removed to standalone processes and replaced with process states.

3rd Normal Form (parallel generalizable structures free). The process is in the 3rd Normal Form if it is in the 2nd Normal Form and the bodies of all its mutually parallel non-elementary structural parts (simultaneities) have been removed to standalone processes and replaced with process states.

4th Normal Form (hidden generalizable sub-structures free). The process is in the 4th Normal Form if it is in the 3rd Normal Form and the bodies of all its non-elementary structural parts (sequences) which are not fully specific to the starting event of the process have been removed to standalone processes and replaced with process states.

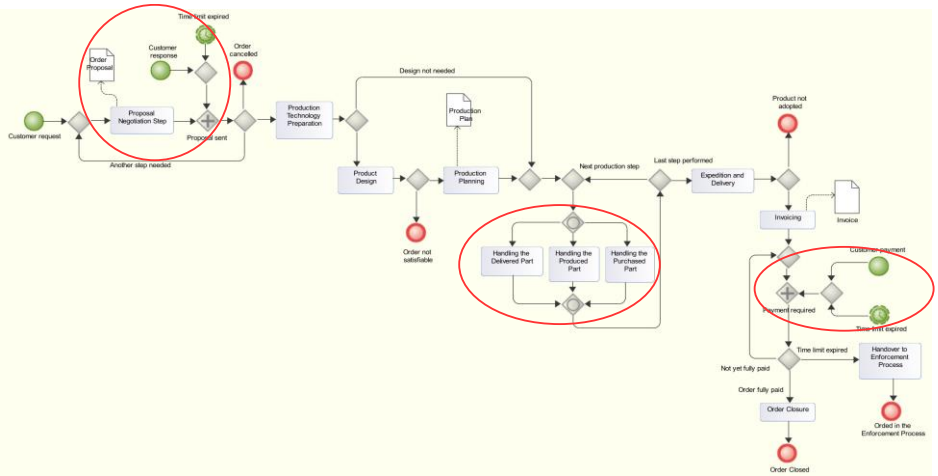


Figure 1. Example of the unnormalized process.

After removal from the original process, the removed part of the process always represents some business system object. Its starting event is the request from the original process, and its product represents the requested service together with the corresponding return event. This way the original single process transforms to the system of mutually collaborating processes where the rest of the original (source) process plays the role of the key process and other processes, removed parts of the source process, are its supporting processes. All processes in the system are mutually tied with the services, what is the consequence of the fact that originally they all came from the same single process. The example of source, not yet normalized process at Figure 1 shows the identified repeating parts of the process in red circles.
Figure 2 shows the process transformed to the 1st normal form. Repeating parts have been removed and replaced with the process states. Each state represents the waiting for the service from the newly created supporting process. Example also shows that each removed part of the process corresponds with some business object (Order, Production, Payment).

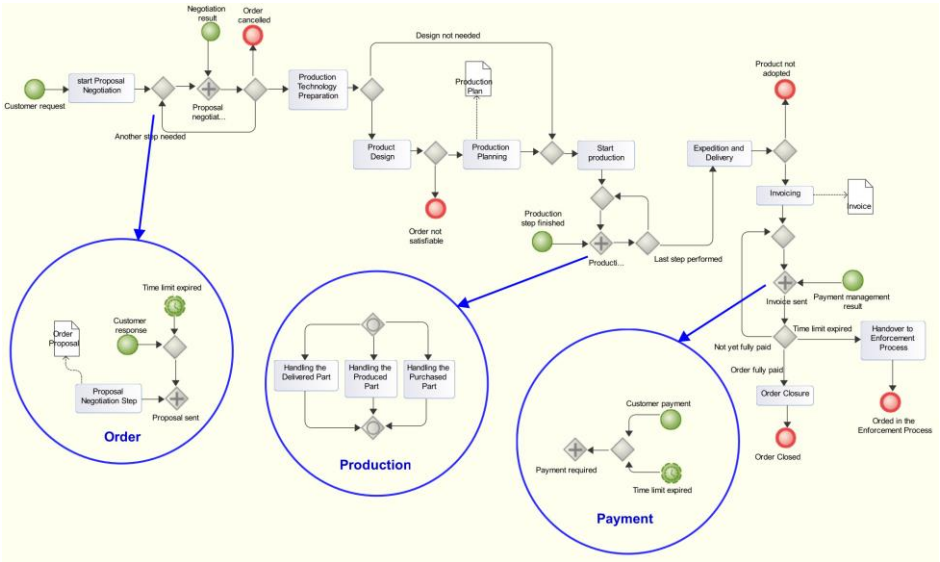


Figure 2 Example of the process in the 1st normal form.

Figure 3 shows the resulting process system after the full normalization of the source process. Firstly removed parts are represented by support processes Order Step Management, Production Step Management and Payment Management. Other support processes have been created during the transformation to higher normal forms either from the source process or from newly created support processes.

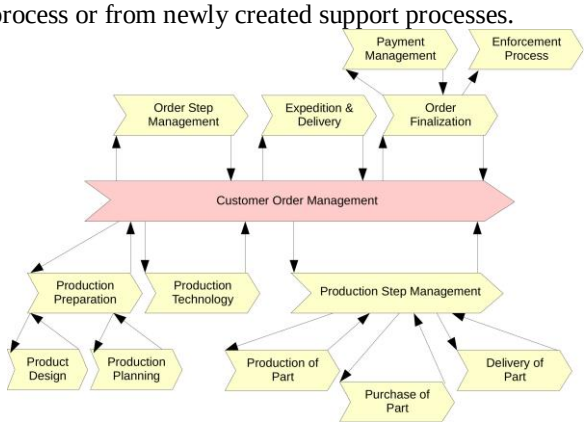


Figure 3 Resulting process system

4 Conclusions

From the point of view of the process system, the Process Normalization is as a way of uncovering the natural supporting processes, hidden in the body of the given process.

The ordering of the steps is important; to be able to correctly interpret alternative sub-structures (branches) in the process one has to remove all repeated structures of activities first (1st to 2nd Normal Form) as the decision about the repetition (end of the loop) might be incorrectly interpreted as a fork signalizing mutually alternative sub-structures. Seeking for parallel structures does not make sense between different alternative structures so the structure has to be in the 2nd normal form before it can be transformed to the 3rd normal form. Similarly, identified sub-structure cannot overstep the border between alternative nor parallel branches of the process. These rules generally follow from the theory of structured thinking most comprehensively described in [6] and [7]. More information about the process normalization technique together with the example and deeper discussion of its essential consequences can be found in [8].

References

1. Business System Modeling Specification: http://opensoul.panrepa.org/metamodel.html (2000-2016).

2. Codd, E. F.: A Relational Model of Data for Large Shared Data Banks. In Communications of the ACM, vol. 13, No. 6, (June 1970).

3. Codd, E. F. "Recent Investigations into Relational Data Base Systems". IBM Research Report RJ1385 (April 23, 1974).

4. Hammer, M., Champy, J.: Reengineering the Corporation: A Manifesto for Business Revolution. London: Nicholas Brealey Publishing (1993).

5. Chen, P P-S.: The Entity-Relationship Model-Toward a Unified View of Data. In ACM Transactions on Database Systems, Vol. 1, No. 1, (March 1976).

6. Jackson, M.A.: Principles of Program Design, Academic Press, London, UK, (1975).

7. Jackson, M.A.: System Development, Prentice Hall International Series in Computer Science, Prentice Hall, London, UK, (1983).

8. Řepa, V. Building a Process Driven Organization with the Process Normalization Technique. in Complex Systems Informatics and Modeling Quarterly, no. 14 (April 30, 2018): 22–37. https://doi.org/10.7250/csimq.2018-14.02.

9. Kent, W.: A Simple Guide to Five Normal Forms in Relational Database Theory, Sept. (1982).

KnowING IPR:
projekt podpory inovací znalostními prostředky

Karel Ježek, Dalibor Fiala, Martin Dostal, Štěpán Baratta, Pavel Herout,
Ladislav Pešička, Markéta Včalová, Pavel Král, Michal Nykl, Ladislav Lenc

Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni
Univerzitní 8, 306 14 Plzeň
{jezek_ka, dalfia, madostal, stepanb, herout, pesicka,
mkucova, pkral, nyklm, llenc}@kiv.zcu.cz

Abstrakt. V příspěvku jsou popsány cíle a současný stav řešení české části mezinárodního projektu, který sdružuje země podunajského regionu. Za podpory Evropské unie a jejího programu pro regionální rozvoj Interreg má projekt za úkol zdokonalit podmínky pro inovace. Na základě zmapování současného stavu využívání práv duševního vlastnictví (IPR), zejména z průmyslové oblasti, klade si projekt za cíl vytvořit nadstavbu k současným bázím dat, kterými disponují patentové úřady, univerzitní knihovny, technické časopisy a obdobné datové zdroje zúčastněných zemí. Tato nadstavba propojí jednotlivé databáze a umožní získávat informace o technických řešeních uživatelsky příjemnějším způsobem a v komplexnější formě.

Klíčová slova: duševní vlastnictví, přenos technologií, patenty, průmyslové vzory

1 Úvod

Cílem projektu je zlepšení rámcových podmínek inovačních procesů v zemích patřících do povodí Dunaje. Projekt je součástí mezinárodního programu INTERREG DANUBE, jehož programové území pokrývá čtrnáct zemí střední a jihovýchodní Evropy od jižního Německa po Bulharsko. Konkrétně se jedná o část programu „Innovative and socially responsible Danube region“, která má rozpočet téměř 76 000 000 €. Název projektu KnowING IPR je akronymem pro „ Fostering Innovation in the Danube Region through Knowledge Engineering and Intelectual Property Rights (IPR) Management“. Projekt je důsledkem přiznání významu IPR pro ekonomický rozvoj, prezentovaný např. v [2], [4] nebo [5].

KnowING IPR poskytne platformu volně přístupných prostředků pro analýzu IPR a návody pro zlepšení a harmonizaci IPR podmínek v regionu Dunaj. Zajistí bohatší informační zdroje o stávajících inovacích, výsledcích výzkumu, patentech a IPR znalostech a zlepší podmínky pro komercionalizaci výsledků výzkumu a transfer technologií (TT). Jedná se o pionýrský počin: využití pokročilých technologií znalostního inženýrství v oblasti IPR. Umožní se tím sdílené využívání existujících

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 93-96.

inovací a zlepšení příležitostí pro spolupráci založenou na IPR (předávání a přebírání znalostí a licencí). Projekt otevře dosud přiškrcený trh IPR a podnítí investice do inovací. Věříme, že také vytvoří konkurenční výhodu pro menší podniky, vysoké školy a výzkumné instituce dunajského regionu. Konečným produktem projektu bude volně přístupný „znalostní hub", sdružující jednotlivé databáze, umožňující jejich simultánní dotazování a poskytující pokročilé analytické funkce. Tento znalostní systém bude pracovat v on-line režimu a bude umět zodpovědět dotazy typu „Kolik stojí patentová přihláška v Rumunsku?", „Kdy vyprší platnost patentu X v Maďarsku?" apod.

2 Způsob řešení

Do projektu se zapojilo celkem šestnáct partnerů z univerzit, výzkumných institucí a vládních úřadů ze třinácti různých zemí. Řešení bylo zahájeno 1. 7. 2018 a bude probíhat do 30. 6. 2021 s rozpočtem 2 149 800 €, z něhož 270 000 € připadne na ZČU v Plzni.

Celý projekt je rozfázován do několika etap, z nichž ta současná končí říjnem 2019. Tato část má za úkol popsat v jednotlivých zemích aktuální stav IPR, TT, zejména pak datové kolekce a jejich současnou i potenciální použitelnost v projektu. Obdobným způsobem je prováděno ohodnocení mezinárodních databází a významných databází institucí z třetích zemí, jako je např. PATSTAT (European Patent Office), PatentScope (World Intellectual Property Organization), či USPTO (United States Patent and Trademark Office). Za relevantní jsou považovány i vědecké a firemní databáze, jako jsou ACM Digital Library, CiteSeer, DBLP, Google Scholar, Microsoft Academic Graph, PubMed, Scopus, SemanticScholar, Web of Science, apod.

Po etapě vyhodnocující vhodnost a dosažitelnost dat, následuje jejich akvizice. Touto činností je pověřena část nazvaná *Data Acquisition Module* (DAM). Data jsou ke stažení z webových stránek v různých formátech, nejčastěji ve formátu XML a JSON, ale také HTML, CSV, XSL, TXT, ZIP nebo PDF. Je proto nutné před ukládáním provést jejich konverzi a parsing do formátu JSON importovatelného do nerelační DB (viz níže). Významné patentové databáze přitom mají rozsah dat i přes 100 GB.

Další funkcí související se stažením a uložením dat je jejich aktualizace. Administrační modul musí v pravidelných intervalech kontrolovat URL stahovaných dat, pomocí časových razítek rozpoznat, zda neobsahují nová data a případně rovnou provést jejich stažení, či upozornit na potřebu manuálního stažení administrátora systému.

Vývoj a struktura akvizičního modulu jsou významně ovlivněny zvoleným databázovým systémem (DBMS), který obhospodařuje hlavní, tzv. zdrojovou databázi s údaji o patentech a článcích. Po zvážení předností i nedostatků tradičních relačních versus NoSQL databází byl jako primární DBMS vybrán MongoDB, který se již dříve osvědčil při zpracování nestrukturovaných dat ve znalostních aplikacích [1]. Zdrojová databáze, modelovaná v nerelační architektuře MongoDB bude obsahovat dva typy kolekcí: pro publikace a pro patenty.

Každá z kolekcí bude ukládat dokumenty z odlišných datových zdrojů, jejichž struktura se sice bude lišit, ale části jako osoby, firmy apod. budou součástí obou. Jelikož MongoDB je dokumentografická databáze, jejími základními prvky jsou dokumenty, které mohou mít různou strukturu. Stažená a extrahovaná data, případně

převedená do JSON formátu, jsou částečně restrukturována před uložením do cílové databáze systému MongoDB. DAM provádí také kontrolu a odstranění duplicit v záznamech a určení klíčových slov pro každý záznam.

Dva základní typy kolekcí znamenají také dva způsoby řešení cílové databáze:

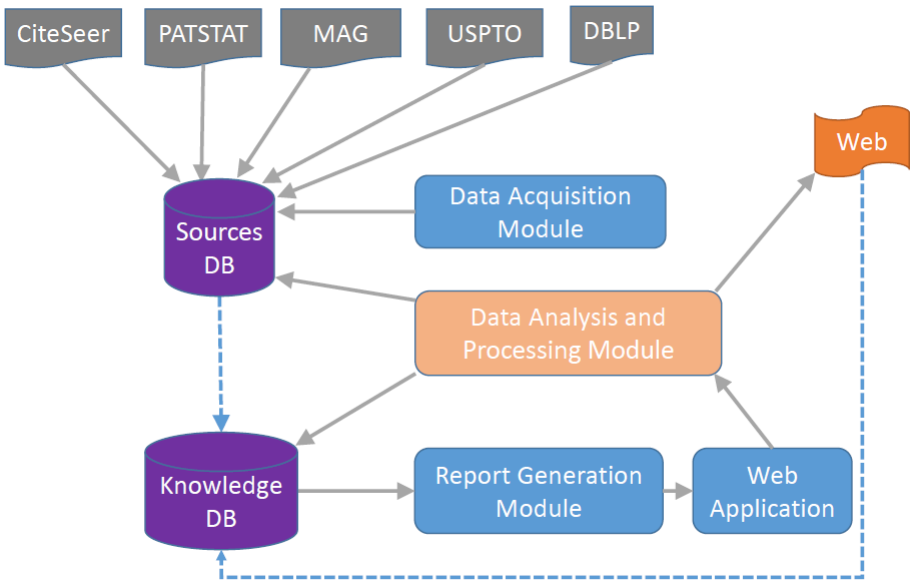
- Separátní MongoDB kolekce pro publikace a pro patenty pro každý datový zdroj.
- Jedinou společnou cílovou kolekci pro každý typ dat.

Zvolen byl druhý způsob, pro jednodušší vkládání i dotazování, pracující pouze se dvěma globálními kolekcemi: *patents* a *publications*.

S daty uloženými ve zdrojové databázi nadále pracuje *Data Analysis and Processing Module* (DAPM). Oproti databázově orientovanému DAM, je DAPM pověřen znalostními funkcemi. Vyhodnocuje a zpracovává IPR dotazy uživatelů a ukládá odezvy na dotazy uživatelů prostřednictvím *Report Generation Module* (RGM) a *Core Communication Controller* (CCC) do znalostní databáze (*Knowledge Database* - KD). KD je SQL databáze realizovaná na bázi systému MariaDB. DAMP, RGM a CCC tvoří tzv. *Knowledge Generation Core* (KGC).

Po zadání dotazu uživatelem přes webovou aplikaci probíhají následující aktivity:

4. Předání dotazu do KGC, jeho lemmatizace a extrakce klíčových slov.
5. DAPM ověří, zda dotaz byl již zadán v minulosti.
6. Pokud byl, vybere se odpověď z KD a je předána uživateli.
7. Pokud nebyl, je dotázána zdrojová (NoSQL) databáze a případně web. Ze seznamu navrácených výsledků je vytvořena odpověď v JSON formě a je zaslána tazateli.
8. Nově získaná odpověď se spolu s dotazem zaznamená do KD (viz Obr. 1).



Obr. 1: Blokový diagram se schématem projektových modulů a databází

3 Současný stav projektu

Řešitelé projektu byli poměrně přesně instruováni, jak postupovat při analýze situace IPR v jednotlivých zemích. Instrukce zahrnovaly anketní otázky, na které odpovídali experti z akademické, průmyslové i politické sféry. Pro vyhodnocení anketní části byl uspořádán workshop s účastí expertů, který formuloval závěry hodnocení stavu IPR v ČR a jeho vliv na výzkum a vývoj. Provedená SWOT analýza věnovala zvláštní pozornost těm slabým stránkám využití IPR, které jsou ovlivnitelné realizací projektu KnowING IPR. Jedná se zejména o:

- nedokonalé rešeršní patentové služby,
- nákladné právní poradenství,
- rozmanitost informačních zdrojů a jejich obtížnou přístupnost.

Z těchto důvodů bylo rozhodnuto v akviziční fázi stáhnout data z národních patentových databází alespoň z části dostupných v anglickém jazyce. Na této úloze se podílí každý z národních týmů. My jsme se zaměřili na získání a propojení dat z české a evropské databáze patentů.

4 Závěr

Propojení databází je pouze prvým krokem k vytvoření znalostního systému, jenž dataminingové komunitě zpřístupní velké množství heterogenních dat z oblasti patentů. Báze dokumentů i znalostí se však musí průběžně doplňovat a měnit. Vytvoření dynamického znalostního systému, který bude obdobně jako v [3] pracovat v režimu „Never Ending Learning“ je proto naším dalším cílem.

Poděkování: Tento příspěvek vznikl s podporou programu Interreg Danube, projektu „KnowING IPR: Fostering Innovation in the Danube Region Through Knowledge Engineering and IPR Management“ (číslo projektu DTP2-076-1.1).

Literatura

1. Jabbari S., Stoffel K.: Ontology Extraction from MongoDB Using Formal Concept Analysis, In: Proc. 2nd International Conference on Knowledge Engineering and Applications (ICKEA 2017), London, UK, 21 - 23 October 2017, 178-182.
2. Kogan, L., Papanikolaou, D., Stoffman, A.S.N.: Technological Innovation, Resource Allocation, and Growth. The Quarterly Journal of Economics 132 (2017), 665–712.
3. Mitchell, T., Cohen, W., Hruschka, E. et al.: Never-Ending Learning. Communications of the ACM 61 (2018), 103-115.
4. Tolstaya, A.M., Suslina, I.V., Tolstaya, P.M.: The role of patent and non-patent databases in patent research in universities. AIP Conference Proceedings 1797 (2017), no. 020017.
5. Van Raan, A.F.J.: Patent Citations Analysis and Its Value in Research Evaluation: A Review and a New Approach to Map Technology-relevant Research. Journal of Data and Information Science 2 (2017), 13-50.

MISDEED – Odhaľovanie medicínskych dezinformácií s využitím tvrdení a expertov

Róbert Móro, Ivan Srba, Matúš Tomlein, Mária Bieliková, Daniela Chudá, Peter Lacko, Marián Šimko, Jakub Šimko, Jakub Ševcech, Andrea Hrčková

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava
{meno}.{priezvisko}@stuba.sk

Abstrakt. Šírenie medicínskych dezinformácií predstavuje jeden z typov antisociálneho správania, ktorý má vážny presah do fyzického sveta (napr. ľudia môžu pod vplyvom falošne interpretovaných informácií odmietnuť konkrétny typ liečby). Hoci v doméne správ a politických dezinformácií boli dosiahnuté prvé výsledky v ich automatickom odhaľovaní, výskum falošných informácií v doméne medicíny je stále len na začiatku. V príspevku predstavujeme náš prístup k tomuto problému, na ktorý sa zameriavame v rámci prebiehajúceho projektu MISDEED. Využívame pritom overené medicínske tvrdenia, ktoré sa snažíme mapovať na články zo spoľahlivých, ale predovšetkým nespoľahlivých zdrojov. V práci sa zameriavame aj na zapojenie odborníkov (lekárov) do procesu učenia dátovo-založených metód, ako aj na stratégie argumentácie cielené na laickú verejnosť.

Kľúčové slová: projekt MISDEED, medicínske dezinformácie, antisociálne správanie, medicínske tvrdenia, detekcia, argumentačné stratégie

1 Falošné informácie v medicíne

Antisociálne správanie a konkrétne šírenie falošných informácií je aktuálny spoločenský, ako aj výskumný problém, o čom svedčí aj počet prehľadových článkov vydaných v posledných dvoch až troch rokoch [2, 4, 6, 7]. Väčšina pozornosti sa pritom zameriava na doménu falošných (politických) správ a menej na doménu medicíny, ktorá má však priamy dopad na zdravie a život ľudí.

Avšak aj v tejto doméne existujú prvé práce zamerané predovšetkým na charakterizáciu falošných medicínskych informácií a čiastočne aj na ich automatickú detekciu. Dhoju, Rony a Hassan analyzovali rozdiely medzi medicínskymi článkami zo spoľahlivých a nespoľahlivých zdrojov [1], čo predstavuje dobrý základ pre identifikáciu vhodných čŕt pre automatickú detekciu falošných (nespoľahlivých) informácií v medicíne. Ghenai a Mejova sa zamerali na identifikáciu nespoľahlivých (nepotvrdených) medicínskych správ na sociálnej sieti, ako aj na identifikáciu používateľov, ktorí ich šíria [3]. Zaujímavým výstupom ich práce je predovšetkým

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 97-101.

slovník nepotvrdených spôsobov liečby rakoviny, ktorý použili pri identifikácii správ na sociálnej sieti[1]. Vplyv falošných a nespoľahlivých medicínskych informácií pri vyhľadávaní na tvorbu názorov používateľov o efektívnosti rôznych spôsobov liečby bol skúmaný v [5], pričom sa potvrdil ich vplyv, keďže ľudia majú tendenciu dôverovať tomu, čo nájdu na webe (a je to vysoko vo výsledkoch vyhľadávania).

Na falošné a nespoľahlivé informácie v doméne medicíny sa zameriavame aj v bilaterálnom projekte MISDEED (*Detekcia falošných informácií a dezinformácií v doméne zdravotníctva*)[2], ktorý riešime v spolupráci s Univerzitou Bar-Ilan v Tel Avive v Izraeli. Na základe analýzy súčasného stavu poznania sme si v projekte zadefinovali nasledovné ciele:

1. *Vyvinúť metódy detekcie medicínskych falošných a nespoľahlivých informácií a spôsobov ich vyvracania* – dali sme si za cieľ zlepšiť automatickú detekciu medicínskych falošných informácií s využitím informácií v podobe dynamiky ich šírenia, zdrojov, z ktorých sa šíria a pod. Zameriavame sa tiež na využitie overených medicínskych tvrdení, ktoré môžu pomôcť pri detekcii. V rámci tohto cieľa plánujeme vytvoriť rozsiahlu dátovú množinu medicínskych článkov a k nim prislúchajúcich overených medicínskych tvrdení.
2. *Preskúmať možnosti spolupráce medzi ľudskými expertmi (lekármi) a dátovo-orientovanými metódami* – zameriame sa predovšetkým na výskum aktívneho učenia. Aby sme podnietili interakciu a zdieľanie znalostí medzi dátovo-orientovanými metódami a expertnými, ako aj laickými používateľmi, plánujeme vyvinúť komunitnú platformu otázok a odpovedí (CQA - Community Question Answering).
3. *Zmierniť dopad falošných medicínskych informácií* – plánujeme skúmať spôsoby prezentovania overených medicínskych tvrdení a ďalších argumentov laickej verejnosti s využitím optimálnych (a potenciálne personalizovaných) stratégií argumentácie.

Vo zvyšku článku prezentujeme náš aktuálny pokrok na projekte po zhruba polroku jeho riešenia, ako aj najbližšie plánované ďalšie kroky.

2 Detekcia s využitím overených medicínskych tvrdení

Keďže doména medicínskych informácií má viaceré špecifiká (vyplývajúce napr. z častého preberania tých istých informácií viacerými zdrojmi a ich opakovaným výskytom v čase, ťažkosťami pri automatickom overovaní pravdivosti medicínskych tvrdení, ale aj z prítomnosti autoritatívnych zdrojov), vytvorili sme v prvom kroku *konceptuálny model medicínskych (dez)informácií*, ktorý zahŕňa nasledovné koncepty:

– *Medicínske tvrdenie*. Ide o opakujúci sa lekármi a výskumom podporovaný alebo nepodporovaný medicínsky výrok (napr. „Vakcinácia spôsobuje autizmus.“).

[1] https://cs.uwaterloo.ca/~aghenai/data.html
[2] https://misdeed.fiit.stuba.sk/

Tvrdenie môže byť definované striktne v podobe výroku, ale aj voľnejšie pomocou kľúčových slov (resp. dopytu).

- *Relevantný dokument*. Je to dokument, ktorý obsahuje daný výrok, pričom k nemu môže zaujímať rôzne stanovisko (pozitívne, t. j. podporné, neutrálne alebo negatívne, t. j. odmietavé). Môže ísť pritom buď o dokumenty overujúce fakty[3] alebo o výskumné články z domény medicíny, ktoré sú dostupné napr. v databáze PubMed[4], alebo o články z rôznych zdrojov, pričom cieľom je overiť ich pravdivosť.
- *Zdroj*. Ide o portál, na ktorom bol článok uverejnený. Môžeme modelovať dôveryhodnosť zdroja, pričom nám môžu poslúžiť zoznamy nedôveryhodných zdrojov[5].

V nadväznosti na navrhnutý konceptuálny model sme identifikovali čiastkové úlohy, ktoré sú potrebné pri verifikácii pravdivosti článku, resp. detekcii (klasifikácii) falošných informácií. Z nich sú pre nás relevantné najmä *klasifikácia (ohodnotenie) relevantnosti dokumentov* – určenie, či je daný dokument relevantný pre zvolené medicínske tvrdenie, t. j. či dokument obsahuje zvolené tvrdenie (resp. či zodpovedá dopytu, ktorým je reprezentované zvolené tvrdenie) a *klasifikácia (identifikácia) stanoviska*, kedy je cieľom rozlíšiť, aké stanovisko (pozitívne, t. j. podporné, neutrálne alebo negatívne, t. j. odmietavé) dokument relevantný pre zvolené tvrdenie zaujíma k tomuto tvrdeniu. Ak dokument podporuje nepravdivé (resp. neoverené, nepodporované) tvrdenie, môžeme o článku prehlásiť, že je nepravdivý (resp. nedôveryhodný).

Aby sme naplnili navrhnutý model, identifikovali sme zoznam 45 nespoľahlivých zdrojov medicínskych informácií prevažne v anglickom jazyku (napr. sieť stránok *Natural news* alebo *Health impact news*), ako aj zoznam asi 24 spoľahlivých zdrojov (napr. *NY Times, World Health Organization* alebo *Health Advocate*). Články z týchto zdrojov sťahujeme pomocou platformy *Monant* [8] vyvinutej na FIIT STU v Bratislave. Aktuálne (september 2019) máme stiahnutých vyše 135 000 článkov zo spoľahlivých a nespoľahlivých zdrojov, pričom tento počet v čase neustále narastá. Okrem toho sme identifikovali vyše 10 fakty overujúcich portálov (angl. *fact-checking sites*, napr. *Snopes* alebo *Metafact*), z ktorých sme stiahli viac ako 1200 overených tvrdení z oblasti medicíny.

Ďalším krokom je mapovanie týchto tvrdení na jednotlivé články, kde v prvom kroku vychádzame z práce [9]. Vychádzajúc z tejto práce sme navrhli metódu identifikácie prítomnosti tvrdenia v článku. Každé medicínske tvrdenie získané z fakty overujúcich portálov je spracované do podoby *dopytu*, pričom využívame transformáciu tvrdenia na *N*-gramy a dopyt rozširujeme o synonymá (resp. súvisiace slová) získané pomocou vektorovej reprezentácie slov s využitím knižnice *fastText*[6]. Všetky dokumenty (články), ktoré zodpovedajú (nad zvolenú prahovú hodnotu skóre) takto získanému dopytu (dopytom), označujeme ako obsahujúce vyhľadávané tvrdenie.

Krok identifikácie prítomnosti tvrdení v článkoch (t. j. mapovanie tvrdení na články) nám umožňuje identifikovať, aká je frekvencia výskytu jednotlivých tvrdení,

[3] Napr. https://www.snopes.com/
[4] https://www.ncbi.nlm.nih.gov/pubmed/
[5] https://www.konspiratori.sk/
[6] https://fasttext.cc/

ale môžeme ďalej skúmať aj dynamiku šírenia tvrdení (v čase, medzi stránkami a pod.). Získavame tak tiež významnú črtu pre automatickú detekciu falošných medicínskych informácií a tiež informáciu použiteľnú pre argumentáciu, prečo je daná informácia označená ako falošná alebo minimálne nespoľahlivá.

3 Ďalšia práca

Ďalšia práca vyplýva z cieľov projektu. Plánujeme overiť navrhnutú metódu mapovania získaných tvrdení na množinu stiahnutých článkov, pričom aktuálne pracujeme na získaní expertne označkovanej dátovej vzorky. Ďalším krokom bude určenie stanoviska článkov k danému tvrdeniu. Takto označkované články (na prítomnosť tvrdenia a postoj článku k nemu) nám budú slúžiť ako trénovacia vzorka pre učenie metód automatickej detekcie falošných medicínskych informácií.

Ďalej sa zameriame na metódy aktívneho učenia, aby sme využili znalosti expertov (lekárov). Na projekte spolupracujeme s lekárom z Lovcov šarlatánov[7], ktorého by sme chceli zapojiť do značkovania pravdivosti článkov, či už vo fáze prípravy trénovacích dát, resp. verifikácie namapovaných tvrdení na jednotlivé články alebo vo fáze trénovania (t. j. v procese aktívneho učenia). Predpokladáme však aj zapojenie ďalších expertov, prípadne aj laických používateľov v rámci využitia múdrosti davu.

V rámci zmierňovania dopadu falošných medicínskych správ plánujeme preskúmať rôzne stratégie argumentácie, prečo je daný článok označený za falošný alebo nespoľahlivý. Dôležitým aspektom je pritom vysvetliteľnosť použitých metód strojového učenia, čo je tiež jeden z výskumných problémov, na ktoré sa zameriavame v rámci projektu.

Poďakovanie: Tento príspevok vznikol vďaka čiastočnej podpore Agentúry na podporu výskumu a vývoja v rámci projektu č. APVV SK-IL-RD-18-0004.

Literatúra

1. Dhoju, S., Rony, M., Hassan, N.: Differences between Health Related News Articles from Reliable and Unreliable Media, (2018). http://arxiv.org/abs/1811.01852.
2. Fernandez, M. Alani, H.: Online Misinformation: Challenges and Future Directions. In: Companion of the The Web Conference 2018 - WWW '18. New York, USA, ACM Press (2018), pp. 595–602.
3. Ghenai, A., Mejova, Y.: Fake Cures: User-Centric Modeling of Health Misinformation in Social Media. In: Proc. ACM Hum-Comput. Interact 2, (2018), p. 20.
4. Kumar, S., Shah N.: False Information on Web and Social Media: A Survey. In: Social Media Analytics: Advances and Applications. CRC press (2018).
5. Pogacar, F.A., Ghenai, A., Smucker, M.D., Clarke, C.: The Positive and Negative Influence of Search Results on People's Decisions about the Efficacy of Medical Treatments. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval - ICTIR '17, (2017), pp. 209–16. New York, USA, ACM Press.

[7] http://www.lovcisarlatanov.sk/

6. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating Fake News: A Survey on Identification and Mitigation Techniques. ACM Transactions on Intelligent Systems and Technology (TIST) 10(3), (2019), p. 21.
7. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter 19(1), (2017), pp. 22–36.
8. Srba, I., Móro, R., Šimko, J., Ševcech, J., Chudá, D., Návrat, P., Bieliková, M.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. In: ROME'19: Workshop on Reducing Online Misinformation Exposure, Paris, France. 7 pages.
9. Wang, X., Yu, C., Baumgartner, S., Korn, F.: Relevant Document Discovery for Fact-Checking Articles. In: Companion of the The Web Conference 2018 - WWW '18, pp. 525–33, (2018), New York, USA, ACM Press.

REBELION – Charakterizácia, detekcia a mitigácia antisociálneho správania

Ivan Srba[1], Róbert Móro[1], Pavol Návrat[1], Daniela Chudá[1], Mária Bieliková[1], Marián Šimko[1], Jakub Šimko[1], Michal Kompan[1], Jakub Ševcech[1], Irina Malkin Ondik[1], Peter Lacko[1], Alena Martonová[1], Gabriela Grmanová[1], Kristína Machová[2], Ján Paralič[2], Peter Butka[2], Peter Bednár[2], Martin Sarnovský[2], Barbora Mesárošová[3], Radoslav Blaho[3], Lucia Sabová[3]

[1]Ústav informatiky a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
{meno}.{priezvisko}@stuba.sk
[2]Katedra kybernetiky a umelej inteligencie, FEI, Technická univerzita v Košiciach
Letná 9, 04200 Košice
{meno}.{priezvisko}@tuke.sk
[3]Katedra psychológie, FiF, Univerzita Komenského v Bratislave
Šafárikovo námestie 6, 81499 Bratislava
{meno}.{priezvisko}@uniba.sk

Abstrakt. Výskum v oblasti detekcie, charakterizácie a mitigácie antisociálneho správania sprevádza množstvo otvorených problémov a výskumných výziev. Najväčšie z nich súvisia s nedostatkom dát, absenciou platformy pre nasadenie navrhnutých metód, ako aj nedostatočné porozumenie, ako používatelia interagujú a konzumujú antisociálny obsah. V príspevku predstavujeme tri výskumné smerovania, v ktorých adresujeme tieto otvorené problémy v rámci prebiehajúceho projektu REBELION. Pre získanie väčších a atribútovo bohatších dátových sád predstavujeme experimentálnu platformu Monant. Dáta zozbierané touto platformou následne umožňujú vykonávať výskum nových typov metód pre charakterizáciu a detekciu antisociálneho správania. V neposlednom rade sme vykonali štúdie správania používateľov pri interakcii a konzumácií antisociálneho obsahu.

Kľúčové slová: projekt REBELION, antisociálne správanie, platforma Monant, charakterizácia a detekcia antisociálneho správania, strojové učenie, mitigácia následkov prejavov a následkov antisociálneho správania

1 Úvod

Antisociálne správanie v online prostredí (šírenie dezinformácii, trolovanie, hejtovanie, kyberšikana, atď.) sa vzhľadom na kritický dopad pre spoločnosť stalo predmetom multidisciplinárneho výskumu, ktorý prirodzene zahŕňa aj oblasť informačných technológií a psychológie. Veľké množstvo výskumu v tejto oblasti dokumentuje

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 102-107.

viacero nedávnych prehľadových článkov [1,2,5,6]. V rámci existujúcich výskumných publikácií je možné identifikovať tri hlavné výskumné úlohy:

- *Charakterizácia*. Úlohou prvej skupiny prístupov je charakterizovať antisociálne správanie analyzovaním jeho prejavov a opisovaním jeho základných charakteristík. Primárnym predmetom skúmania je obsah a používatelia, ktorí ho vytvárajú. V menšej miere sa skúma aj kontext (napr. zdroje alebo šírenie falošných správ).
- *Detekcia*. Druhá najväčšia skupina prístupov sa zameriava na automatickú alebo semi-automatickú detekciu antisociálneho správania. Využívajú sa tu vo veľkej miere techniky umelej inteligencie (predovšetkým strojového učenia a spracovania prirodzeného jazyka).
- *Mitigácia*. Úlohou mitigačných prístupov je regulovať alebo eliminovať výskyt antisociálneho správania a jeho negatívnych dôsledkov. Doterajší výskum mitigačných prístupov je len v počiatkoch a poskytuje široký potenciál pre ďalšie výskumné aktivity.

Charakterizácia, detekcia a mitigácia antisociálneho správania je predmetom skúmania aj v našom projekte REBELION (*Automatizované rozpoznávanie antisociálneho správania v online komunitách*)[1]. Projekt sa zameriava na výskum nových modelov a metód pre automatizované rozpoznávanie antisociálneho správania v online prostredí. Ciele a zameranie projektu REBELION sme detailnejšie predstavili v predchádzajúcom príspevku [4].

2 Výskumné smerovanie a prvotné výsledky

Na základe existujúcich publikácií, otvorených problémov a výskumných výziev sme identifikovali ako kľúčové pre ďalší rozvoj výskumu charakterizácie, detekcie a mitigácie antisociálneho správania tri výskumné smery:

- *Adresovanie veľkého množstva neoznačkovaných a dynamických dát*. Väčšina existujúcich metód vyžaduje označkované dáta a využíva najmä strojové učenie s učiteľom. Existujúce označkované dátové sady však nedosahujú potrebnú veľkosť, príp. sú anotované automaticky s využitím veľmi zjednodušenej heuristiky (napr. správy sú označkované ako falošné len na základe domény, na ktorej boli publikované). Tieto dátové sady sú navyše statické a nereflektujú vysokú dynamickosť antisociálneho správania. Rovnako v súčasnosti absentuje platforma, v rámci ktorej by bolo možné vytvorené metódy detekcie aplikovať.
- *Širšie využitie dát o obsahu, používateľoch a kontexte*. V súčasnosti existujúce metódy charakterizácie a detekcie antisociálneho správania využívajú len malú časť dostupných dát o obsahu, používateľoch a kontexte. Zapojenie takýchto dát však prináša možnosť výskumu nových typov metód, ktoré napr. využívajú dáta z viacerých zdrojov, vo viacerých jazykoch, z viacerých modalít (text, obrázky,

[1] https://rebelion.fiit.stuba.sk/

video), zo širšieho kontextu (napr. šírenie antisociálneho správania, reakcie používateľov).

- *Hľadanie nových prístupov k mitigácii*. Popri charakterizácií a detekcii, hľadanie nových prístupov k mitigácií antisociálneho správania predstavuje ďalšie dôležité výskumné smerovanie. Priestor pre výskum predstavuje napr. systém pre skoré varovanie pred výskytom antisociálneho správania alebo edukačné nástroje pre trénovanie používateľov v jeho rozpoznávaní.

V rámci prvého roku riešenia projektu REBELION sme v každom tomto výskumnom smere priniesli prvé výsledky, ktoré predstavujeme v nasledujúcich podkapitolách.

2.1 Monant: Platforma pre výskum antisociálneho správania

Problém nedostatku dostatočne veľkých a atribútovo bohatých dátových sád sme sa rozhodli vyriešiť návrhom a implementáciou experimentálnej platformy nazvanej Monant. Platforma Monant dokáže priebežne zbierať dáta z médií na webe (novinové portály, diskusie, sociálne siete, atď.) a následne ich ukladať v jednotnej podobe. Následne umožňuje priamo zaintegrovať široké spektrum metód pre charakterizáciu a detekciu antisociálneho správania v zozbieraných dátach. V neposlednom rade poskytuje možnosť rozšírenia o viacero typov používateľských aplikácií.

Architektúra platformy Monant sa skladá z piatich modulov (jednotlivé moduly sú detailnejšie opísané v príspevku [8]): 1) *centrálne úložisko 2) monitoring webu*, 3) *metódy umelej inteligencie*, 4) *manažment platformy*, 5) *koncové používateľské aplikácie*.

Vyvinutý prototyp platformy Monant obsahuje prvú verziu implementácie prvých štyroch vyššie uvedených modulov. Pri zbere dát sme sa zamerali špecificky na dezinformácie, ktoré v rámci antisociálneho správania predstavujú pravdepodobne najkritickejší problém.

Platforma Monant je už aktuálne nasadená v produkčnej prevádzke, v rámci ktorej sa používa na monitorovanie dezinformácií v oblasti medicíny. Modul monitoringu webu bol nakonfigurovaný pre získavanie článkov a diskusií zo 69 zdrojov - novinových a blogovacích portálov poskytujúcich obsah v anglickom jazyku (medzi zdrojmi sú zastúpené portály šíriace ako pravdivý, tak aj nepravdivý obsah). Aktuálne (september 2019) obsahuje platforma 135 226 článkov a 455 836 k ním naviazaných diskusných príspevkov. Platforma kontinuálne jednotlivé zdroje monitoruje, a tak sa dátová sada neustále rozširuje a poskytuje neustále aktuálne dáta, čím sa nám podarilo vyriešiť problém dynamicky meniacich sa a pribúdajúcich dát.

V rámci platformy Monant adresujeme aj problém neoznačkovaných dát. Analogicky ako je možné monitorovať novinové články a blogy, dokážeme monitorovať aj články overujúce fakty (angl. fact-checking articles) z portálov ako je napr. Snopes[2]. V rámci týchto portálov skupina expertov overuje konkrétne tvrdenia a poskytuje k nim hodnotenie ich pravdivosti aj s príslušným zdôvodnením. V aktuálnej verzii platformy Monant sme zozbierali viac ako 1 200 takýchto článkov z medicínskej domény a v nich verifikovaných tvrdení. Aktuálne vyvíjame sadu metód,

[2] https://www.snopes.com/

ktoré automaticky vyhodnotia prítomnosť tvrdenia v článku (t. j. či sa tvrdenie v článku nachádza, nachádza len okrajovo alebo nenachádza vôbec) a postoj článku k tomuto tvrdeniu (t. j. či článok tvrdenie podporuje, oponuje mu alebo ho len neutrálne diskutuje). Pre natrénovanie týchto metód aktuálne prebieha ručná anotácia časti všetkých možných mapovaní. Následne plánujeme všetky zozbierané články oanotovať na základe vytvorených metód a automaticky získaných značiek pre prítomnosť a postoj k tvrdeniu (t. j. ak článok podporuje nepravdivé tvrdenia alebo oponuje pravdivým tvrdeniam, bude považovaný za nepravdivý a naopak). Takto vytvorené značky budú môcť byť následné použité v rámci metód charakterizácie a detekcie šírenia dezinformácií.

2.2 Metódy charakterizácie a detekcie antisociálneho správania

Počas prvého roku riešenia projektu REBELION sme navyše navrhli aj prvé verzie metód pre charakterizáciu a detekciu falošných správ a trolovania. V rámci charakterizácie sme napr. poukázali na rozdiel v sentimente, témach alebo čitateľnosti textu obsiahnutom v pravdivých a falošných správach. V rámci detekcie falošných správ sme skúmali napr. možnosti využitia širšieho kontextu (konkrétne vytvorenie nových čŕt odvodených zo správ s rovnakým obsahom šírených na ďalších dôveryhodných/nedôveryhodných stránkach).

Venovali sme sa tiež detekcii trolov a autorít v diskusných príspevkoch. Vyvinuli sme viacero modelov a metód založených na regresnej analýze a genetickom programovaní [3].

2.3 Štúdie konzumácie antisociálneho správania

Pre umožnenie výskumu inovatívnych spôsobov mitigácie antisociálneho správania sme zrealizovali štúdiu, ktorej sa zúčastnilo 44 stredoškolských študentov. V simulovanom prostredí pripomínajúcom portál sociálnej siete mali participanti možnosť prezerať kolekciu ako falošných, tak aj pravdivých správ. Počas experimentu sme zbierali implicitnú a explicitnú spätnú väzbu a využili sme aj zariadenia na sledovanie pohľadu. Výsledkom štúdie je niekoľko zistení, ako používatelia konzumujú falošné/pravdivé správy a ako táto konzumácia súvisí s mierou ich záujmu o témy obsiahnuté v poskytnutých správach (samotný experiment ako aj jednotlivé zistenia sú detailnejšie opísané v príspevku [7]).

Oslovili sme tiež pedagogických a odborných zamestnancov škôl a zrealizovali sme neformálne diskusné stretnutia s cieľom získania skúseností s prejavmi antisociálneho správania u žiakov: jeho príčin, podôb výskytu a vplyvu na dynamiku vzťahov. Na základe týchto zistení môžeme stanoviť konkrétnejšie ciele pri sledovaní implicitných motívov a dôsledkov takéhoto správania (predovšetkým prejavov trolovania, flamingu, kyberšikany) u populácie detí a dospievajúcich, ktoré patria v tomto kontexte medzi najviac ohrozené skupiny.

3 Ďalšia práca

Ďalšia práca na projekte REBELION bude zameraná na rozširovanie výsledkov v troch uvedených výskumných smeroch. V rámci platformy Monant plánujeme rozšírenie platformy o modul koncových aplikácií. Pôjde predovšetkým o aplikácie slúžiace širokej verejnosti ako sprievodca a pomocník pri rozlišovaní dezinformácií na webe. V rámci týchto aplikácií plánujeme preskúmať možnosť využitia tzv. chatbotov ako v procese detekcie pre komunikáciu medzi metódami strojového učenia a expertmi (tzv. aktívne učenie, angl. active learning) ako aj pre komunikáciu s laickou verejnosťou.

Už zozbierané dáta nám poskytujú možnosť pokračovať vo výskume metód charakterizácie a detekcie antisociálneho správania. Špecificky sa plánujeme zamerať na skúmanie multilingválnej detekcie, detekcie názorov, analýzy sentimentu a ďalších pokročilých techník spracovania prirodzeného jazyka. V rámci mitigácie nadviažeme na vykonané štúdie, ktoré plánujeme rozšíriť o experimenty s expertnými participantmi, ktorí sa profesionálne venujú detekcii antisociálneho správania (predovšetkým dezinformácií). Dosiahnuté zistenia následne zohľadníme pri návrhu rôznych argumentačných stratégií, ako informovať používateľov webu o možnom výskyte dezinformácií v obsahu, s ktorým práve interagujú. V neposlednom rade, ďalšiu už rozbehnutú aktivitu tvorí príprava prehľadového článku zaoberajúceho sa prehľadom výskumu antisociálneho správania z pohľadu informatiky.

Poďakovanie: Tento príspevok vznikol vďaka čiastočnej podpore Agentúry na podporu výskumu a vývoja v rámci projektu č. APVV-17-0267.

Literatúra

1. Fernandez, M. Alani, H.: Online Misinformation: Challenges and Future Directions. In: Companion of the The Web Conference 2018 - WWW '18. New York, USA, ACM Press (2018), pp. 595–602.
2. Kumar, S., Shah N.: False Information on Web and Social Media: A Survey. In: Social Media Analytics: Advances and Applications. CRC press (2018).
3. Machová, K., Mikula, M., Szabóová, M., Mach, M.: Sentiment and Authority Analysis in Conversational Content. In *Computing and Informatics*, Vol. 37, No. 3(2018), pp. 737-758.
4. Návrat, P. et al.: REBELION – odhaľovanie antisociálneho správania na Webe. In: *Data a znalosti & WIKT 2018*, Brno 11.-12.10.2018, pp. 65-69.
5. Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., Liu, Y.: Combating Fake News: A Survey on Identification and Mitigation Techniques. ACM Transactions on Intelligent Systems and Technology (TIST) 10(3), (2019), p. 21.
6. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter 19(1), (2017), pp. 22–36.
7. Simko, J., Hanakova, M., Racsko, P., Tomlein, M., Moro, R., Bielikova, M.: Fake News Reading on Social Media: an Eyetracking Study. In *HT '19: ACM Conference on Hypertext and Social Media*, September 17–20, 2019, Hof, Germany. ACM, New York, NY, USA, 10 pages. To appear.

8. Srba, I., Móro, R., Šimko, J., Ševcech, J., Chudá, D., Návrat, P., Bieliková, M.: Monant: Universal and Extensible Platform for Monitoring, Detection and Mitigation of Antisocial Behaviour. In: ROME'19: Workshop on Reducing Online Misinformation Exposure, Paris, France. 7 pages.

Zpracování přirozeného jazyka v rámci projektu InteCom

Pavel Král[1,2], Ladislav Lenc[2], Josef Steinberger[1,2], Tomáš Brychcín[2], Pavel Přibáň[2], Jakub Sido[2]

[1] Katedra informatiky a výpočetní techniky, FAV ZČU v Plzni,
Univerzitní 8, 306 14 Plzeň
{meno}.{priezvisko}@stuba.sk
[2] Nové technologie pro informační společnost – NTIS, FAV ZČU v Plzni
Technická 14, 306 14 Plzeň
{pkral,llenc,jstein,brychcin,pribanp,sidoj}@kiv.zcu.cz

Abstrakt. Cílem tohoto příspěvku je seznámení čtenáře s úlohami, kterými se zabývá naše výzkumná skupina v rámci projektu „Výzkum a vývoj inteligentních komponent pokročilých technologií pro plzeňskou metropolitní oblast (InteCom)“ a jejich širším kontextem. Jedná se především o sémantickou analýzu textu a její využití v návazných úlohách v oblasti zpracování přirozeného jazyka (NLP). V rámci článku budou rovněž představeny relevantní problémy v dané oblasti a možnosti aplikací výsledků výzkumu. Dále bude čtenář seznámen s cíli a výstupy, které jsou naplánovány v rámci tohoto projektu. Na závěr popíšeme výsledky, které byly dosaženy během prvního roku a půl řešení projektu.

Klíčová slova: InteCom, sémantická analýza, zpracovaní přirozeného jazyka

1 Představení projektu

Cílem výzkumného záměru Inteligentní komponenty pokročilých technologií je vyvinout IT technologie škálovatelné cenou a výkonovými parametry, které převedou stávající teorie, poznatky, principy, metody a algoritmy z oblasti automatizace, robotiky, umělé inteligence, monitorování, diagnostiky a zpracování signálů do formy, která významně urychlí a zviditelní možnost jejich uplatnění v praxi. Záměr se skládá ze tří výzkumných témat:

- robotické a řídící technologie,
- diagnostické a rozhodovací technologie,
- monitorovací technologie.

Výsledky projektu jsou orientovány k aplikacím ve výrobě, dopravě, energetice, zdravotnictví, veřejné správě, telekomunikacích a dalších oblastech lidské činnosti. Jejich úspěšné uplatnění v produktech a procesech konkrétních podniků a institucí významně posílí a rozšíří spolupráci výzkumného centra NTIS s aplikační sférou.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 108-112.

Projekt je financovaný Ministerstvem školství, mládeže a tělovýchovy z operačního programu EU Výzkum, vývoj a vzdělávání v rámci výzvy Předaplikační výzkum pro ITI plzeňské metropolitní oblasti.

Zpracování přirozeného jazyka se řeší v rámci druhého výzkumného tématu, diagnostické a rozhodovací technologie.

2 Zpracování přirozeného jazyka

Klíčovou oblastí, kterou se budeme zabývat v rámci řešení tohoto projektu, je sémantická analýza textu [10,7] a její využití v návazných úlohách zpracování přirozeného jazyka (NLP). Sémantická analýza se zabývá způsoby reprezentace významu přirozeného jazyka a lepší reprezentace významně přispěje ke zlepšení výsledků řady dalších NLP úloh např. klasifikace dokumentů, rozpoznávání pojmenovaných entit, analýza polarity textu, automatická sumarizace, strojový překlad a další.

Sémantická analýza (porozumění) jednotlivých slov v textu a zároveň strojové učení s učitelem dnes dosahují velmi dobrých výsledků. Tyto metody bohužel mají své limity: sémantická analýza jednotlivých izolovaných slov neuvažuje kontext, který je ale pro význam a pochopení textu klíčový. Strojové učení s učitelem vyžaduje ruční tvorbu trénovacích dat a dalších jazykových prostředků, což je časově i finančně velmi nákladné. V obou případech nastává problém při změně cílového jazyka nebo i při adaptaci na jinou doménu. Výše uvedené limity je možno překonat pomocí metod bez učitele [11] (nebo s jeho částečnou pomocí) a s využitím modelů schopných analyzovat více jazyků [6]), které jsou zároveň jazykově / doménově téměř nezávislé. Výborných výsledků lze také dosáhnout s využitím (hlubokých) neuronových sítí [3].

3 Relevantní problémy

Většina přístupů v oblasti analýzy textu (vč. automatické klasifikace dokumentů, přiřazení klíčových slov, detekce pojmenovaných entit, apod.) využívá sémantickou informaci pouze ve velmi omezené míře. Tyto metody zpravidla neberou v úvahu kontext
a jeho strukturu (vztahy mezi slovy a jejich slovosled) a sémantiku textu často vnímají jen jako bag-of-words.

Tyto přístupy jsou zpravidla založené na pravidlech nebo na strojovém učení s učitelem, protože obojí má v současné době dostatečný aplikační potenciál. Nedostatkem těchto metod analýzy přirozeného jazyka je proto ale jejich jazyková / aplikační závislost.

Dalším důležitým problémem je analýza specifických textů. Jedná se zejména o komentáře v sociálních mediích, které jsou zpravidla velmi krátké, obsahují velké množství překlepů a nespisovných slovních spojení. V případě češtiny často chybí diakritika a správná kapitalizace. Dále sem patří analýza textu z obrazového (zpravidla PDF) formátu, kde je třeba před samotnou analýzou rozpoznat text. Převod do textu

může být problematický v případě nízké kvality naskenovaných stránek nebo v případě velkého množství obrázků a tabulek.

4 Aplikační potenciál

S rostoucím množstvím textových dat (na Internetu, ale i v interních systémech) je stále více důležité jejich inteligentní zpracování. Enormní množství textu generované uživateli skrývá velmi hodnotné informace, které analýza textů může odhalit. Mnoho aplikací dnes používá fulltextové vyhledávání, detekci spamů, filtrování obsahu webových stránek, strojový překlad do jiných jazyků, detekci polarity textu recenzí atd. Analýza velkých textů je stále více populární kvůli velkému potenciálu jak v privátním tak ve veřejném sektoru. Firmy se potřebují vyznat v obrovském množství interních dat (často v podobně PDF), potřebují také chápat potřeby svých zákazníků, kteří je píší online, aby lépe cílily své marketingové kampaně a vytvářely lepší produkty. Vládní organizace musí monitorovat globální problémy společnosti (např. uprchlické krize). Řešení si navíc musí poradit s různými jazyky. Analýza textu je také „core business" hlavních internetových společností, např. Facebook, Google, Twitter, Baidu, Yahoo. Sémantická analýza textu výrazně přispěje ke zlepšení výsledků ve všech výše uvedených aplikačních oblastech.

5 Cíle a výstupy

V rámci projektu se zaměříme na překonání potřeby vytvářet a trénovat jazykově závislé modely, zaměříme se na sémantickou cross-linguální analýzu delších textů (slovních spojení, vět, odstavců, apod.) pomocí metod učení bez učitele (příp. s jeho částečnou pomocí) nebo s využitím neuronových sítí. Výstupem budou metody, které v experimentech prokáží schopnost zpracovávat více jazyků a~zároveň se sníží potřeba vytvářet označená trénovací data. Zvýší se tak potenciál pro nasazení našich aktuálních metod do praxe. Použitelnost vyvinutých metod v praxi bude ověřena na datech z reálného prostředí, viz např. následující usecase.

Řada IT subjektů v současné době intenzivně řeší digitalizaci dokumentů (tištěných, ale i ručně psaných) s následným ukládáním do databáze. Další skupina organizací již dokumenty v databázi uložené má. V obou případech ale zpravidla chybí jejich jednoduché zpřístupnění uživatelům (rychlé nalezení dokumentu dle různých kritérií, tzv. inteligentní vyhledávání). Naším cílem je vyvinout metody pro rozpoznání textu z naskenovaných dokumentů, jejich analýzu, zaindexování do fulltextové databáze a umožnění rozšířeného inteligentního vyhledávání nad jejich obsahem vč. metadat. Zaměříme se nejen na tištěné dokumenty, ale i na dokumenty ručně psané, u kterých předpokládáme využití zejména hlubokých neuronových sítí. V rámci analýzy obsahu dokumentů bude automaticky provedena kategorizace dokumentů, přiřazena klíčová slova, určeny pojmenované entity (jako např. osoby, instituce, data, apod.), vytvořen automatický souhrn obsahu dokumentu a v případech, kde to je relevantní, bude určen sentiment dokumentu. Při zpracování textu (tj. napříč všemi použitými metodami analýzy dokumentů) bude provedeno sémantické zpracování obsahu dokumentů, kde budou využity nejnovější poznatky z metod strojového učení a výpočetní lingvistiky.

6 Dosažené výsledky

Nejprve jsme navrhli novou metodu pro reprezentaci textu, která rozšiřuje stávající metody reprezentace slov o globální informaci z Wikipedie [9]. Překonali jsme tak state-of-the-art v oblasti mapování slov na vektory reálných čísel. V oblasti vícejazyčné (cross-linguální) sémantické analýzy jsme vymysleli metodu transformace, která doplňuje nejlepší transformační metody o vážení [1]. Náš přístup překonává ostatní metody v úloze sémantické podobnosti vět na několika souborech dat v různých jazycích. Nové metody byly úspěšně využity ve dvou NLP úlohách a to pro automatické rozpoznávání dialogových aktů [5] a pro vícejazyčné slovní analogie [2].

Dále jsme navrhli a implementovali základní metody pro segmentaci a rozpoznávání tištěných textů (OCR) založené na hlubokých neuronových sítích typu CNN a LSTM [4]. V současné době probíhá ověřování funkčnosti těchto metod na reálných historických datech. Zároveň jsme vytvořili systém pro predikci emocí v krátkých textech, který rovněž využívá neuronové sítě typu LSTM. Funkčnost systému byla ověřena na reálných datech, tj. krátkých zprávách (tweetech) ze sociální sítě Twitter [8].

7 Závěr

V rámci tohoto článku jsme seznámili čtenáře s úlohami, kterými se zabývá naše výzkumná skupina v rámci projektu InteCom a jejich širším kontextem. Zároveň byly představeny relevantní problémy v dané oblasti a možnosti aplikací výsledků výzkumu. Dále byl čtenář seznámen s cíli a výstupy, které byly naplánovány v rámci tohoto projektu. Na závěr jsme popsali výsledky, které byly dosaženy během prvního roku a půl řešení projektu.

Poděkování: Tento článek vznikl za podpory projektu "VaV inteligentních komponent pokročilých technologií pro metropolitní oblast Plzeňského kraje (InteCom)" reg. č.: CZ.02.1.01/0.0/0.0/17_048/0007267 finacovaného z EFRR.

Literatura

1. Brychcín, T.: Linear transformations for cross-lingual semantic textual similarity. Knowledge-Based Systems (2019). https://doi.org/10.1016/j.knosys.2019.06.027
2. Brychcín, T., Taylor, S., Svoboda, L.: Cross-lingual word analogies using linear transformations between semantic spaces. Expert Systems with Applications **135**, 287–295 (2019). https://doi.org/10.1016/j.eswa.2019.06.021
3. Kim, Y.: Convolutional neural networks for sentence classification. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (October 25-29 2014)
4. Lenc, L., Martínek, J., Král, P.: Tools for semi-automatic preparation of training data for ocr. In: MacIntyre, J., Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds.) Articial Intelligence Applications and Innovations. pp. 351–361. Springer International Publishing, Cham (24-26 May 2019). https://doi.org/10.1007/978-3-030-19823-7_29

5. Martínek, J., Král, P., Lenc, L., Cerisara, C.: Multi-lingual dialogue act recognition with deep learning methods. In: Interspeech. Graz, Austria (15-19 September 2019)
6. McDonald, R., Petrov, S., Hall, K.: Multi-source transfer of delexicalized dependency parsers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11). pp. 62–72. Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
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. Přibáň, P., Martínek, J.: UWB at IEST 2018, emotion prediction in tweets with bidirectional long short-term memory neural network. p. 224–230. Association for Computational Linguistics, Brusel, Belgie (2018), https://www.aclweb.org/anthology/W18-6232
9. Svoboda, L., Brychcín, T.: Improving word meaning representations using wikipedia categories. Neural Network World 28(6), 523–534 (2018). https://doi.org/10.14311/NNW.2018.28.029
10. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of articial intelligence research 37, 141–188 (2010)
11. Zanzotto, F.M., Korkontzelos, I., Fallucchi, F., Manandhar, S.: Estimating linear models for compositional distributional semantics. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1263–1271. Association for Computational Linguistics (2010)

Vyhlížení ostrovů pravidelnosti ve znalostních grafech RDF dalekohledem metamodelu OWL

Vojtěch Svátek, Daniel Vodňanský, Jiří Ivánek

Katedra informačního a znalostního inženýrství VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
{svatek,daniel.vodnansky,ivanek}@vse.cz

Abstrakt. Pro efektivní indexování dat RDF, i pro jejich zpřístupňování uživateli ve srozumitelné podobě, je často třeba je převést do podoby tabulek nebo stromů. Na podkladě zobecňujícího modelu se pokoušíme předběžně analyzovat některé z konstruktů OWL, schématového jazyka pro RDF, z hlediska toho, zda podporují vznik „ostrovů pravidelnosti" dobře převoditelných do tabulek nebo stromů nebo ne. Na tuto úvodní fázi výzkumu bude navazovat empirická analýza konkrétních datasetů, i projektů, kde byl takový převod použit.

Klíčová slova: reprezentace dat, RDF, OWL, ontologie, pravidelnost

1 Úvod

Datový model RDF je (velmi zjednodušeně řečeno) založen na reprezentaci faktů pomocí elementárních trojic „*subjekt – predikát - objekt*", kde je objekt vždy hodnotou predikátu přiřazenou pro daný subjekt. Příkladem je trojice vyjadřující, že pro subjekt „Česká republika" nabývá predikát (v terminologii RDF nazývaný též *vlastnost*) „má hlavní město" objektu (tj. hodnoty) „Praha". Prvky trojice jsou zpravidla (resp. v případě predikátu vždy) globálně unikátními identifikátory, tzv. IRI, přes které je možné jednotlivé trojice (shodující se v IRI některého svého členu) propojovat do libovolně rozsáhlých grafů.

V posledních letech je formát RDF stále více využíván k vystavování propojených dat na webu, resp. obecněji k tvorbě znalostních grafů. Jeho předností oproti tradičním technologiím jsou vysoká míra flexibility vyjádření, minimální náklady na změnu schématu, a snadnost, s jakou lze nezávisle vzniklé datasety propojovat. Za tuto flexibilitu a otevřenost však platíme vysoké náklady ve chvíli, kdy je třeba rozsáhlá data efektivně indexovat a vyhledávat v nich, ale i při vývoji aplikací, která mají data zpřístupňovat uživateli v přehledné vizuální podobě.

V těchto situacích je vhodné opřít se o „ostrovy pravidelnosti"[1] v grafových datech, které je možné indexovat a/nebo vizualizovat tradičními prostředky určenými pro data strukturovaná tabulkově nebo stromově. Pokud ovšem takové prostředky použijeme na méně pravidelné části dat RDF, projeví se to nárůstem velikosti datové reprezentace, z důvodu opakování shodných dat (redundance). Proto je důležité umět „ostrovy"

[1] Termín zavádíme jen jako metaforický, bez přesného formálního vymezení.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 113-118.

detekovat, což je z praktického hlediska nutné provádět na úrovni konkrétních dat samotných. Přesto se nabízí otázka, do jaké míry je pravidelnost dat ovlivněna tím, jak vypadá schéma dat – standardně vyjádřené pomocí jazyka OWL [3] nebo jeho podjazyků – včetně toho, které z konstruktů OWL jsou vůbec ve schématu (ontologii) použity.

V tomto příspěvku se na reálná data RDF prvotně díváme „z dálky", tedy optikou metamodelu jazyka OWL používaného k tvorbě schémat pro RDF grafy (tzv. ontologií). Analyzujeme na obecné rovině, které z konstruktů jazyka podporují vznik „ostrovů pravidelnosti" a které naopak umožňují pravidelnost narušovat, což v případě převodu do „čistých" stromů a tabulek vede k redundanci, popřípadě ke ztrátě informace.

2 Související výzkum

Úloha převodu struktur RDF do jednodušší, zejména tabulkové podoby byla řešena v mnoha kontextech. Pro účely rozpoznání tabulkových struktur (s „emergujícím schématem") uvnitř dat RDF, které mohou být efektivněji indexovány než obecné grafy, vyvinuli pro *RDF databáze* vzorkovací techniku Pham a Boncz [4]. Attard a kol. [1] navrhli sadu nástrojů převádějících data RDF do tabulek a stromů primárně pro lepší orientaci *lidského uživatele* při vizualizaci; transformace je sice ztrátová, ale informace lze zpětně dohledat v původních grafových datech. Jeden z hlavních znalostních grafů, *DBpedia*[2] (vzniklá automatickou extrakcí z článků Wikipedie) také umožňuje vedle nativního formátu RDF stažení dat v podobě tabulek (CSV) a souborů JSON, tzv. *DBpedia as Tables*,[3] aby s nimi mohli pracovat i vývojáři bez znalosti RDF. Žádný z uvedených projektů se ovšem nezaměřoval na aparát schématového jazyka OWL.

Pokud jde o transformaci grafu RDF na stromovou strukturu, triviálně se toto děje např. při tzv. *serializaci* dat do formátu RDF/XML [2]. Jde však o velmi jednoduché stromy tvořící nosnou vrstvu pro data stále inherentně síťová. Z obsahového hlediska lze z objektů RDF vytvořit strom jejich „navěšením" na *hierarchickou strukturu tříd* (za předpokladu, že je tato sama o sobě čistě stromová a neobsahuje multihierarchie) a zanedbáním všech vztahů těchto objektů mezi sebou. Tato možnost se při vizualizaci dat RDF běžně využívá. Naopak si nejsme vědomi žádného projektu, který by se systematicky věnoval analýze vztahů hierarchického charakteru vyjádřených pomocí *vlastností* propojujících objekty charakteru instancí, přestože se v řadě ontologií používají nativně hierarchické vlastnosti (binární relace) typu „partOf" npod.

3 Abstraktní model tabulkových a stromových struktur

Abychom mohli postihnout, které prvky jazyka OWL používané v ontologiích (a následně v jim odpovídajících datasetech) podporují transformaci do tabulkových a stromových struktur, a které nikoliv, potřebujeme vyjádřit pojem tabulky a stromu v obecné reprezentaci, kterou lze namapovat na grafovou reprezentaci RDF. Pro tento

[2] https://dbpedia.org/

[3] https://wiki.dbpedia.org/services-resources/downloads/dbpedia-tables

účel jsme předběžně navrhli jednoduchý *datový model* nazvaný OARV, umožňující popsat vzory datových struktur složených ze čtyř druhů prvků: *objektů*, *atributů*, *vztahů* (mezi objekty) a *hodnot* (kterých nabývají atributy).

Na *tabulku* se v OARV můžeme dívat jako na množinu trojic spojujících objekty buď pomocí vztahů s jinými objekty (trojice (*o*, *r*, *o*')), nebo pomocí atributů s hodnotami (trojice (*o*, *a*, *v*)). Charakteristikou vysoce pravidelné tabulky pak je existence právě jednoho objektu *o*' pro každou dvojici (*o*, *r*) a právě jedné hodnoty *v* pro každou dvojici (*o*, *a*). Pokud objekt/hodnota chybí, nebo jich je víc, pravidelnost je narušena.

Strom je v OARV popsán nepatrně složitěji. Je složen z trojic stejného typu jako tabulka, ovšem vztah *r* je vyčleněného („hierarchického") typu: objekt na první pozici trojice je chápán jako podřazený objektu na třetí pozici trojice. Z hlediska pravidelnosti dále u stromu požadujeme existenci jen jednoho nadřazeného objektu („rodiče"), jediného kořenu, a absenci cyklů.[4]

4 Vstupní analýza metamodelu OWL

Metamodel jazyka OWL definuje mnoho desítek logických konstruktů. V současné počáteční fázi výzkumu jsme se omezili na obecné rozvažování nad těmito konstrukty z hlediska vztahu k tabulkové a stromové reprezentaci, jakou je možné získat buď z ontologie pomocí konstruktů vytvořené, nebo z dat (na úrovni instancí), která jsou pomocí takové ontologie namodelována. Některé z úvah jsou uvedeny v Tab. 1. Rozbor vlivu je nutno v tuto chvíli chápat jako velmi předběžný, a shromážděné postřehy jsou značně různorodé. Přesnější určení a roztřídění vlivů konstruktů na pravidelnost bude závislé na konkrétních transformacích, které budou rovněž vyjádřeny pomocí modelu OARV, a na kontextu, v rámci kterého bude vliv zkoumán (zejména efektivita strojového zpracování na jedné straně a přehlednost zobrazení pro uživatele na druhé straně).

Provedenou analýzu budeme pro většinu konstruktů z tabulky ilustrovat na souhrnném příkladě – samozřejmě velmi zjednodušeném oproti reálným, optimálně modelovaným ontologiím. Mějme následující ontologické schéma zapsané ve formě trojic; prefix *v1* identifikuje určitý slovník („vocabulary") navržený pro data RDF určitého typu – akademické události a publikace s nimi spojené:

v1:Event rdfs:subClassOf owl:Thing .
v1:Document rdfs:subClassOf owl:Thing .
v1:Seminar rdfs:subClassOf v1:Event .
v1:Publication rdfs:subClassOf v1:Document .
v1:PublishedArticle rdfs:subClassOf v1:Publication .
v1:PublishedTextbook rdfs:subClassOf v1:Publication .
v1:partOf rdf:type owl:ObjectProperty .
v1:sessionOf rdf:type owl:ObjectProperty .
v1:yearOfEvent rdf:type owl:DatatypeProperty .
v1:yearOfEvent rdf:type owl:FunctionalProperty .

[4] Formální vyjádření těchto podmínek je v modelu přímočaré, pro stručnost je zde neuvádíme.

Tab. 2. Vybrané konstrukty OWL a jejich možný vliv na pravidelnost struktur RDF

Skupina konstruktů	Vliv na tabulární a stromovou pravidelnost
Podtřída (r*dfs*:*subClassOf*), a instanciace (*rdf*:*type*)	Jednotlivá použití konstruktů definují strukturu *podřazenosti*. Navázáním objektů (instancí) pomocí *rdf:type* na hierarchicky uspořádané třídy můžeme získat *stromovou* strukturu. Hierarchické uspořádání tříd ale může být narušené multihierarchií, pak nemusí jít o strom.
Podvlastnost (r*dfs*:*subPropertyOf*)	V případě použití na vlastnosti hierarchického typu může vztah podvlastnosti komplikovat její stromovou strukturu: stejné objekty budou spojeny jak podvlastností, tak i nadvlastností.
Ekvivalence tříd (*owl:equivalentClass*)	Pokud jsou dvě třídy ekvivalentní, instance jedné je také instancí druhé. Ekvivalencí je jednak formálně narušena stromová struktura *podřazenosti* na úrovni schématu, jednak tabulková regularita na úrovni instancí, pokud budeme chtít třídu reprezentovat sloupcem tabulky.
Identita instancí (*owl:sameAs*)	Narušuje strukturu *podřazenosti*; pro *tabulární* data může vyvolat *vícehodnotovost* (spojí-li se data z různých zdrojů, hodnoty pro daný atribut se často mohou odlišovat)
Funkčnost vlastností (*owl:FunctionalProperty*)	Vylučuje *vícehodnotovost* pro *tabulární* data.
Disjunktnost tříd/vlastností (*owl:AllDisjointClasses*, *owl*:*AllDisjointProperties*)	Omezuje přiřazování instancí více třídám, tím podporuje *tabulkovou* pravidelnost.

Prvních šest trojic definuje hierarchickou strukturu tříd (s univerzálním kořenem *owl:Thing*). Zbylé čtyři trojice zavádějí tři vlastnosti („properties“, v terminologii RDF), přičemž *v1:partOf* a *v1:sessionOf* jsou objektové vlastnosti (neboli vztahy v terminologii OARV) a *v1:yearOfEvent* je datová vlastnost (neboli atribut v terminologii OARV), která je navíc funkční, tj. platí o ní, že danému zdroji v subjektu přiřazuje nejvýše jeden objekt.

Konkrétní báze faktů o specifické konferenci pak obsahuje mj. trojice (se zdroji – instancemi ze jmenného prostoru identifikovaného prefixem *d1*):

d1:paper12345 rdf:type v1:Publication .
d1:paper12345 rdf:type v1:PublishedArticle .
…
d1:DZW19ProgramBrochure rdf:type v1:Publication .
d1:DZW19 rdf:type v1:Seminar .
d1:InvitedTalk23 v1:sessionOf d1:DZW19 .
d1:WIKT19 rdf:type v1:Event .
d1: WIKT19 v1:partOf d1:DZW19 .
d1:DZW19 v1:yearOfEvent 2019 .

Navázáním (pomocí *rdf:type*) instancí odpovídajících jednotlivým příspěvkům (*d1:paper12345* atd.) a programové brožuře (*d1:DZW19ProgramBrochure*) na základní hierarchii tříd o publikacích tvořenou spojením třídy *v1:PublishedArticle* a její nadtřídy *v1:Publication*[5] pomocí *rdfs:subClassOf*. Tím vznikne společná stromová struktura obsahující třídy i instance.

Uvažujme následně obohacení schématu datasetu novým slovníkem, pro jehož entity použijeme odlišný prefix *v2*:

v2:ScientificText rdfs:subClassOf v1:Document .
v1:PublishedArticle rdfs:subClassOf v2:ScientificText .
v2:Workshop owl:equivalentClass v1:Seminar .
v1:sessionOf rdfs:subPropertyOf v1:partOf .

Po přidání první a druhé trojice (opět využívající *rdfs:subClassOf*) již schéma tříd není stromové (*v1:PublishedArticle* má dvě přímé nadtřídy), a tudíž přestává být stromem i struktura rozšířená o instance, což může vést např. ke zhoršené orientaci v datech při jejich grafickém zobrazování. Třetí trojice (v případě provedení logické inference využívající *owl:equivalentClass*) způsobí, že bude instance přímo podřazena dvěma třídám (vedle původní *v1:Seminar* i *v1:Workshop*), v tomto případě však bude možné zachování *stromové* struktury zobrazení dosáhnout sloučením uzlů odpovídajících ekvivalentním třídám, a obdobně, v případě *tabulkového* zobrazení sloučením sloupců. Poslední trojice pak (opět na základě inference, tentokrát nad *rdfs:subPropertyOf*) způsobí, že bude vedle původního vztahu *d1:InvitedTalk23 v1:sessionOf d1:DZW19* . platit i *d1:InvitedTalk23 v1:partOf d1:DZW19* . I zde lze zachování stromové struktury sekundárně zajistit, a to sloučením hran.

Předpokládejme dále, že dojde k rozšíření datasetu pomocí propojení na jiná data o stejné konferenci. Tato data (instancemi s prefixem *d1*) ne zcela korektně vycházejí ze zaindexování sborníku až v r. 2020, a vztahují tento rok na celou konferenci:

d2:DataZnalostiWikt2019 rdf:type v1:Seminar .
d2:DataZnalostiWikt2019 v1:yearOfEvent 2020 .
d2:DataZnalostiWikt2019 owl:sameAs d1:DZW19 .

Vlastnost *v1:yearOfEvent* teď bude (v inferenčním uzávěru datasetu) pro ztotožněné instance nabývat dvou odlišných hodnot, což představuje mj. problém pro tabulkové zobrazení. Ovšem díky tomu, že je vlastnost deklarována jako funkční, lze tento problém automaticky (opět pomocí formálně-logického odvození) detekovat jako chybu a následně opravit.

[5] Poznamenejme, že pro různorodé publikace typu brožur nebo pozvánek není nutné zavádět novou podtřídu, ale lze je řadit přímo pod nadřazenou třídu: z hlediska modelování v jazyce OWL není důvod přiřazovat instance jen třídám na nejnižší úrovni hierarchie.

5 Závěr

Ačkoliv je problematika transformace mezi grafově a tabulkově nebo stromově strukturovanými daty předmětem řady specifických projektů, systematická analýza metamodelu jazyka OWL z hlediska podmínek pro takovou transformaci podle našich informací dosud provedena nebyla.

V rámci další fáze výzkumu chceme na teoretické rovině formálně ukotvit model OARV v některém etablovaném formálním kalkulu, a navrhnout v jeho rámci konkrétní vzory transformací. Na empirické rovině plánujeme ověřit, nakolik se obecné rysy konstruktů OWL uplatnily na úrovni konkrétních slovníků a datasetů RDF, a také, které vzory transformací se uplatnily ve kterých praktických projektech.

Poděkování: Tento příspěvek vznikl s podporou projektu GAČR č. 18-23964S, "Focused categorization power of web ontologies".

Literatura

1. Attard, J., Orlandi, F., Auer, S.: ExConQuer: Lowering barriers to RDF and Linked Data reuse. Semantic Web, 9 (2), (2018), 241–255.
2. Gandon, F., Schreiber, G.: RDF 1.1 XML Syntax. W3C Recommendation, 25 February 2014. Online https://www.w3.org/TR/rdf-syntax-grammar/.
3. Motik, B., Parsia, B., Patel-Schneider, P. F.: OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax. W3C Recommendation. W3C (2009).
4. Pham, M.-D., Boncz, P.A.: Exploiting Emergent Schemas to Make RDF Systems More Efficient. In: Groth, P.T. et al. (eds.) International Semantic Web Conference (2016), 463–479.

Podpora vícekriteriálního recenzního řízení akademických prací složenou obrázkovou metaforou

Vojtěch Svátek, Petr Strossa

Katedra informačního a znalostního inženýrství VŠE
Nám. W. Churchilla 4, 130 67 Praha 3
{svatek,kizips}@vse.cz

Abstrakt. Akademické práce, zejména konferenční články, bývají evaluovány podle dílčích kritérií zahrnujících i numerické hodnoty. Na vzorku konferencí ze stejné tematické oblasti jsme nalezli vysokou míru shody kritérií, což dává prostor pro návrh obecných recenzních metrik. Hodnocení článku recenzenty pak můžeme vyjádřit pomocí obrázkové metafory představující např. auta s odlišenou velikostí/charakterem dílčích komponent. Takový obrázek může hypoteticky být výrazně přehlednější než číselná tabulka s desítkami hodnot. Implementaci podpory předpokládáme formou webové služby využívající ontologie.

Klíčová slova: recenzní proces, vizualizace, metafora, ontologie

1 Úvod

Vědecké práce, zejména konferenční články, jsou zpravidla hodnoceny několika recenzenty podle dílčích kritérií, která mohou zahrnovat i číselná skóre. V rámci souhrnného hodnocení by nadřazený hodnotitel (zejména předseda programového výboru nebo meta-recenzent) měl ideálně vzít všechny dílčí strukturované údaje do úvahy. Překážkou je ovšem jak jejich prostý počet (často několik desítek), tak i skutečnost, že se označení dílčích kritérií napříč konferencemi liší, a jsou identifikována jen textově.

Výzkum, jehož raná fáze je popsána v tomto příspěvku, si klade za cíl ověřit dvě výzkumné otázky:

1. Lze napříč větší množinou konferencí ze stejné nebo obdobné tematické oblasti identifikovat kritéria se stejným nebo velmi blízkým významem?
2. Lze promítnout takovou sadu kritérií do komponent obrázkové metafory, která by souhrnnému hodnotiteli umožnila rychleji/lépe posoudit danou vědeckou práci (ve smyslu toho, jak je vnímána množinou primárních recenzentů a numericky jimi ohodnocena podle dílčích kritérií), než by tomu bylo v případě tabulkového zobrazení číselných hodnot?

Částečnou odpověď na první otázku poskytuje provedená mikrostudie konferencí z oblasti sémantických technologií. Zodpovězení druhé otázky vyžaduje rozsáhlejší

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 119-123.

výzkum na testovacích subjektech; příspěvek nicméně formuluje variantu obrázkové metafory, která by podle názoru autorů mohla mít potenciál zmiňované hodnocení usnadnit, a demonstruje její aplikaci na konkrétním, reálném případě.

2 Recenzní kritéria konferencí o sémantických technologiích

Pro úvodní studii bylo vybráno 9 renomovaných konferencí, které se zabývají problematikou sémantických technologií, a to, v detailnějším členění, sémantického webu a propojených dat (ISWC, ESWC, SEMANTiCS), aplikované ontologie (FOIS), reprezentace a zpracování znalostí (KR, K-CAP, EKAW), ale i souhrnně umělé inteligence, pod kterou velká část sémantických technologií spadá (IJCAI, ECAI). Přestože jsou mezi komunitami těchto konferencí průniky, nelze předpokládat, že by všechny jejich recenzní formuláře vycházely z jednoho zdroje. Pokud se tedy podaří nalézt jednotný systém kritérií, bude tím silně podpořena kladná odpověď na první výzkumnou otázku. Autoři proto detailní analyzovali jak samotné recenzní formuláře (které měli k dispozici díky účasti na recenzním řízení konferencí ať už v roli recenzenta nebo autora), tak i případné doprovodné materiály (tzv. „Reviewer Guidelines“ apod.). Kritéria, jejichž význam vnímali jako (téměř nebo zcela) shodný, vždy zastřešili metrikou[1] s jednotným pracovním označením. Výsledkem je devět metrik.

Rozbor ukázal, že napříč všemi devíti konferencemi je jednota pouze v přítomnosti kritérií odpovídajících metrikám *Overall score* a *Reviewer's confidence*. Metriky *Relevance*, *Novelty*, *Technical quality* a *State of the art* byly přítomny sedmkrát, *Presentation* šestkrát, *Significance* čtyřikrát a *Evaluation* dvakrát; lze předpokládat, že na těch konferencích, kde pro rozsah a kvalitu evaluace není samostatné kritérium (např. IJCAI a ECAI), je implicitně zahrnuta do "technické kvality" článku. Pouze v jediném případě (ESWC) byla naopak dvě kritéria přiřazena stejné metrice (*Technical quality*): šlo na jedné straně o úplnost a správnost navrženého řešení, a na druhé straně o to, jak jsou vlastnosti tohoto řešení názorně demonstrovány a diskutovány. Druhé z kritérií lze také chápat jako "technickou kvalitu", i když spíše textu než výzkumu samotného; na druhou stranu jde stále ještě o obsahovou a nikoli prezentační stránku textu, proto druhé kritérium nebylo vztaženo k metrice *Presentation*.

Pokud se na výsledky podíváme ze strany konferencí: šest z devíti vyžadovalo po recenzentovi vyplnit do formuláře minimálně sedm numerických kritérií odpovídajících identifikovaným metrikám (se započtením *Overall score* a *Reviewer's confidence*). Sedmá konference (FOIS) má kritérií šest; „odlehlé hodnoty“ jsou pouze u EKAW (tři, z dílčích jen *Relevance*) a K-CAP (žádné dílčí kritérium) – tyto konference vycházejí z tradice „workshopů“ (mj. bez paralelních sekcí) a jejich recenzní formulář je zřejmě i proto méně strukturovaný.

Souhrnně provedená mikrostudie indikuje kladnou odpověď na první výzkumnou otázku. Ve směru druhé otázky naznačuje, že pro znázornění hodnocení podle dílčích kritérií pomocí obrázkové metafory může často být k dispozici přiměřený objem dat, a že lze takovou metaforu aplikovat unifikovaným způsobem. Poznamenejme, že pokud

[1] Termín používáme v obecném („inženýrském“), nikoli matematickém smyslu.

některá metrika chybí, stále lze pro všechny recenze použít její „střední“, popřípadě „průměrnou“ hodnotu.

V příkladu vizuální metafory uvedeném v dalším textu jsou použita reálná data z recenzování článku na jedné z uvedených konferencí (SEMANTiCS), z r. 2018.

3 Recenzovaný článek jako auto složené z komponent

Ze spektra objektů reálného světa, jejímiž obrázky lze recenzovaný článek reprezentovat, jsou pro nás zajímavé ty, které:

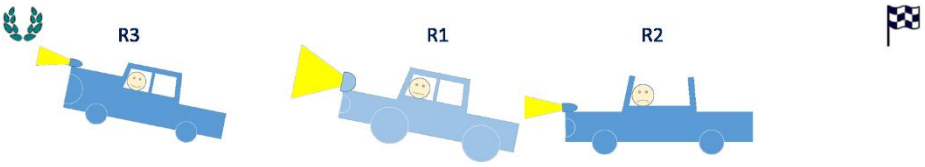
- Zahrnují dostatečný počet komponent, které lze na obrázku snadno rozlišit.
- Komponenty jsou všeobecně známé, tudíž na většinu uživatelů (hodnotitelů) nekladou speciální nároky na pochopení.
- Existuje alespoň částečná (byť metaforická) spojitost mezi významem dané komponenty pro zobrazovaný objekt a významem určité metriky pro článek.

Tyto požadavky ve značné míře splňuje koncept auta. Tab. 1 je příkladem, jak mohou být identifikované metriky mapovány na komponenty auta, s tím, že číselné hodnoty metrik budou odpovídat „vizuálním proměnným“.

Tab.1. Mapování recenzních metrik na vizuální proměnné obrázku auta

Metrika	Vizuální proměnná	Škála
Relevance	Rotace auta	Lineární, od 90° do 0°
Novelty	Velikost motoru	Lineární, od „žádný“ do „velký“
Technical quality	Velikost kol	Lineární, od „žádná“ do „velká“
State of the art	Čelní světlo	Lineární, od „žádné“ do „velké“
Evaluation	Čelní sklo a střecha	Nic / Jen sklo / „Targa“ / Celá střecha
Significance	Zavazadlový prostor	Lineární, od „žádný“ do „velký“
Presentation	Tvář řidiče	Uslzená / Zamračená / Neutrální / Úsměv
Confidence	Sytost obrázku	Lineární, od „bledý“ k „sytý“
Overall score	Vodorovná pozice	Lineární, zprava (praporek) doleva (vavřín)

Obr. 1 znázorňuje tři recenze stejného článku. Původní numerické hodnoty (na škále 1-5, jen u *Overall score* jde o Likertovu škálu od -3 do 3) jsou přímočaře, proporcionálně, převedeny na hodnoty vizuálních proměnných z Tab. 1. Celkově se jedná o 27 numerických hodnot. Zobrazení by mělo hodnotiteli metaforicky sdělit např. to, že Recenzent 3 (R3) hodnotí článek nejlépe, a to s ohledem na inovativnost (velký motor – „velký tah kupředu“), kvalitní evaluaci (kompletní střecha – „důvěryhodnost, bezpečí“) a prezentaci (úsměv – „pohodlí a spokojenost čtenáře“). R1 si zase cení povědomí autorů o existujícím výzkumu (silné světlo – „dobrý přehled“) a technického provedení výzkumu (velká kola – „robustnost“); je si však svým názorem méně jistý.



Obr. 1. Vizuální metafora trojice recenzí stejného článku

Použité metafory do určité míry reflektují výzkum v oblasti kognitivní psychologie [1, 2]. Ten zmiňuje jednoduchou metaforu „MORE IS BIGGER"[1], i sofistikovanější "LINEAR SCALES ARE PATHS" (pozice aut na dráze podle celkového hodnocení), „DIFFICULTIES ARE IMPEDIMENTS TO MOTION" (malá kola při nízké technické kvalitě) nebo „THOUGHT IS MOTION" (originalita myšlenky jako velikost motoru) [2]. Na druhou stranu lze nalézt příklady, kdy zvolený přístup s uznávanou metaforou plně nekoresponduje, např. podle „DIFFICULTIES ARE BURDENS" by mohl velký zavazadlový prostor evokovat pomalu jedoucí auto. V každém případě bude efektivní využití obrázkových metafor vyžadovat určitou fázi zácviku.

4 Nástin architektury webové služby a ontologie

Předpokládáme, že pro implementaci příslušného nástroje je nejvhodnější forma webové služby. Ta v jednoduchém případě obdrží přes REST rozhraní od recenzního systému identifikátory (IRI) metrik popsaných ve *sdílené ontologii*, a jejich hodnoty; následně vrátí obrázek s odpovídající vizuální metaforou. Alternativně by provozovatelé recenzního systému mohli navíc zaslat soubor s *přímým mapováním* svých lokálních recenzních kritérií na vizuální proměnné, a upravit si tak vizuální metaforu „na míru".

Některé části sdílené ontologie bude možné převzít z ontologií existujících, např. *FAIR*[2] zachycuje aktéry recenzního procesu a *BIDO*[3] škály používané v hodnocení. Pro samotné recenzní metriky však bude zřejmě nutné navrhnout modul nový.

5 Závěr

Předložený výzkum má vysoce praktický cíl: usnadnit a zkvalitnit posuzování konferenčních článků v situaci s velkým počtem dílčích skóre. Intuitivně se obrázková metafora pro tento cíl zdá být vhodná. Pro reálné ověření a realizaci přínosu je ale nutné jednak prakticky implementovat referenční webovou službu a ontologii, jednak provést kognitivní experimenty na testovacích subjektech.

Poděkování: Tento příspěvek vznikl s podporou projektu GAČR č. 18-23964S, "Focused categorization power of web ontologies".

[2] https://sparontologies.github.io/fr/current/fr.html
[3] https://sparontologies.github.io/bido-review-measures/current/bido-review-measures.html

Literatura

1. Hiniker A., Hong S., Kim Y.-S., Chen N.-C., West J. D., Aragon C. R.: Toward the operationalization of visual metaphor. JASIST 68(10), (2017), 2338-2349.
2. Lakoff, G.: The contemporary theory of metaphor. In A. Ortony (Ed.), Metaphor and thought. Cambridge, MA: Cambridge University Press (1993), 202–251.

Filtering outliers to improve classification. First results

Dušan Hetlerovič, Ľuboš Popelínský

K KD Lab, FI MU
Brno, Czech republic
{xhetler,popel}@fi.muni.cz

Abstract. In this work we describe preliminary experiments for exploring influence of outlier elimination to classifier accuracy. A workflow is a couple consisting of some pre-processing methods - outlier detection and filtering in our case - and a classifier. We describe a method for testing various workflows and bring first results.

Keywords: outlier detection, data pre-processing, classification

1 Introduction

When working with data, we often come across instances known as outliers. From the definition in [5], an outlier can be seen as an observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism. While outlier detection (OD) itself [2] is commonly used for various purposes, such as fraud detection, the impact of eliminating outliers from data before classification has not been studied so thoroughly.

In this paper we examine outlier detection / elimination methods and ex- amine whether it is useful to add them to the preprocessing phase. Although previous work presented rather negative results or showed that accuracy of ensembles overcame accuracy after outlier filtering [11] there are some aspects that have not been investigated, namely OD parameter and hyperparameter (e.g. % of outliers to remove) settings and their influence to a change of accuracy.

We selected outlier detection methods the most frequently used, and also representative classifiers including Bayesian ones, Decision trees and rules, Logistic regression, Support Vector Machine, Random trees and Forest and Multilayer Perceptron. We have conducted a set of experiments with the aim of identifying useful combinations of a particular outlier detection / elimination method with a particular classifier.

Apart from classifier selection, e.g. SVM, users may achieve an improvement in performance by applying different (hyper)parameter settings or data processing

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 124-128.

methods. The combination of these three factors pre-processing, hyperparameter settings and a classifier is called a workflow.

In the following section we bring an overview of related work. Section 3 describes a method that has been used for experiments. Summary of results can be found in Section 4.

2 Related work

One pioneering work in the area of outlier detection was the work of John on the so-called robust decision trees [8], who studied the effects of label noise. After learning a tree, all misclassified instances were removed from the learning set and a new tree was learned. It was repeated until the learning set was consistent. Although it did not result in accuracy increase, the resulting tree was much smaller. Similar approach for kNN was presented in [7, 13]. In [12] instance hard- ness is introduced (how difficult it is to classify an instance) both theoretically and practically - an instance is hard if classifiers disagree. This value is then compared with 9 hardness measures. Instances of 64 datasets, mostly from UCI, and 9 classifiers from Weka [4] were analyzed. It was shown that class overlap is a main contribution to instance hardness. In [11] filtering misclassified instances is explored. Misclassification was studied both in conjunction with a single classifier or an ensemble of classifiers. The same 64 datasets and 9 supervised learning algorithms were used. In both cases, misclassified instances were removed. When the same learning algorithm was used to filter misclassified instances and to learn a model, only three algorithms displayed accuracy increase – LWL lazy learner, Neural net and Ripper. In all cases, using an ensemble of learning algorithms for filtering resulted in a greater increase in classification accuracy than when using only a single learning algorithm. However, if compared with majority voting ensemble of those 9 classifiers, the majority voting ensemble reached, on average, the highest accuracy. Further references can be found in [11,12].

3 Experiments

To landmark and explore the area of filtering outliers we performed experiments with following datasets, classifiers and outlier detection & elimination methods.

Datasets. We used all of the datasets that appeared either in [11] or in [1], together 58 datasets.
Classifiers. 20 classification algorithms (CL) was employed with default parameter settings, *BayesNet, IBk, LWL, J48, JRip, Logistic regression and Simple logistic regression, Multilayer Perceptron, Naive Bayes, OneR, PART, Random tree and Random Forest, and Support Vector Machine (SMO)*
Outlier detection. We also extended the selection of outlier detection algo rithms from [12] with CODB [6], and with several OD algorithms from scikit-learn [10] kNN, Isolation Forest [9] and LOF [3]. All 17 OD algorithms were again employed with

default parameter settings, similarily as in [12]. The list of OD method contained *Local Outlier Factor (LOF), NearestNeighbors, Isolation Forest, ClassLikelihood (CL), ClassLikelihoodDifference (CLD), F2 (Max individual feature efficiency), F3 (Collective Feature Efficiency), T1 (Fraction of maxi- mum covering spheres), T2 (Ave number of points per dimension), MV (Minority value), CB (Class balance), KDN (K-Disagreeing Neighbors), DS (Disjunct size), DCP (Disjunct class percentage), TD, TDWithPrunning (Tree Depth with and without prunning), and CODB.*

Each outlier method has a hyperparameter indicating the percentage of outliers to be eliminated. For this study we set it to top 5 and 10% outliers. The total number of outlier methods and its variants is then 34. Regarding the classifiers, we have used the same 20 classifiers from the Weka toolkit [4] as mentioned in [11]. The total number of workflows was 680. As each workflow was run on 58 datasets, the total number of experiments was 39440.

For each combination of workflow (OD method + classifier) and dataset, the dataset was split into train and test sets with a ratio of 90/10. Afterwards, corresponding OD method was run on the train set and based on hyperparameter setting % of the top outlying instances were removed. Finally, the classifier was trained on the newly cropped train set and then evaluated on the test set. This whole process was repeated 10 times, in manner of 10-fold cross-validation. These results were then compared against the results of workflows with no outlier detection or elimination (also obtained via 10-fold cross-validation).

4 Results

First experiments. We took all OD methods and all classifiers. Hyperparameter was either 5 or 10% of top outliers to be eliminated. Overall, we usually see a slight decrease in performance, which could possibly be attributed to a rather significant reduction in the number of training instances. Furthermore, the removal of noisy instances might lead to overfitting of the trained models. We also observed that there is no OD method that significantly outperforms the others, neither a workflow of CL+OD. For the results of a particular OD method see Table 1 where Gain stands for difference in accuracy when no OD method was used. It confirms a hypothesis that there is no OD/filtering method that is useful for any dataset.

Second experiment. In the next step we focused on four OD methods that looked the most promising, with the highest number of datasets where filtering increased accuracy. It was *kNN, Isolation Forest, LOF and CODB*.

Now the hyperparameter was set to 1, 2, 3, 4, and 5%. We also narrowed the list of datasets: we removed those with less than 100 instances. The best workflows can be found in Table 2. Out. % stands for the percentage of top outliers removed. Base, Extend and Gain is accuracy of CL, accuracy of CL+OD, and their difference. The most important is the last column that says for how many datasets CL+OD outperformed CL in accuracy.

OD	Accuracy	Gain
CB	76.021	-1.426
CODB	76.797	-3.383
CL	75.805	-3.369
CD	76.419	-4.880
DCP	75.802	-9.903
DS	76.767	-4.151
F2	78.749	-1.178E-15
F3	75.862	4.218E-16
IsolationForest	76.804	-3.808
KDN	76.704	-7.982
LOF	76.893	-3.033
MV	76.088	-0.028
NearestNeighbors	76.565	-6.184
T1	76.635	1.874E-03
T2	76.275	-8.761E-04
TD	76.644	-0.349
TDWithPrunning	76.043	-0.097
Total	76,458	-3.283

Table 1. Average accuracy for OD

Classifier	OM	Out. %	Base Acc.%	Extend. Acc.%	Gain Acc.%	#Wins in 32
IBk	kNN	3	76.04	74.64	-1.41	12
IBk	kNN	4	76.04	74.54	-1.50	11
IBk	kNN	5	76.04	74.36	-1.69	12
RandTree	I-Forest	4	74.68	75.22	0.53	14
PART	CODB	5	77.83	72.63	-5.10	18
LWL	CODB	5	67.82	66.18	-1.64	17
RandTree	CODB	5	74.68	69.57	-5.11	18
RandTree	CODB	4	74.68	69.39	-5.30	19
JRip	LOF	1	77.20	77.78	0.58	13
RandTree	CODB	1	74.68	69.04	-5.64	16

Table 2. Some of the most promising workflows

Regarding Table 2, we can point out a few interesting things. The first is that CODB, a class-based outlier detection method, is rather dominant. It reaches an improvement in over half the cases top 5 workflows employed CODB even though mean gains in accuracy are still negative, Another thing to note is that there are two cases where gain was positive - *Random Tree/Isolation Forest and JRip/LOF*.

Especially, the first fact is worth exploring. To recognize a dataset that is promising for filtering is the main goal for future.

5 Conclusion

Despite the first results being rather negative, it is apparent that adding outlier detection and elimination to the preprocessing phase could lead to increased accuracy in certain cases. Future work could be focused on finding patterns that bring improvements so that we could get a clearer idea of when and how to eliminate outliers ahead of classification. Another potential direction is finding and using these promising workflows in automated machine learning.

Acknowledgment: We thank to Pavel Brazdil LIAAD - INESC TEC Porto for permanent and fruitful discussions and support. We also thank to Robert Kolcun and Ondrej Kurak for implementation of the test framework, and to students of a Machine learning and Data mining course. This work has been partially supported by Faculty of Informatics, Masaryk University Brno.

References

1. Salisu Abdulrahman and Pavel Brazdil. Measures for combining accuracy and time for meta-learning. *CEUR Workshop Proceedings*, 1201:49–50, 01 2014.
2. Charu C. Aggarwal. *Outlier Analysis (2nd edition)*. Springer, 2017.
3. Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
4. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009.
5. Douglas M. Hawkins. *Identification of outliers / D.M. Hawkins*. Chapman and Hall London ; New York, 1980.
6. Nabil Hewahi and Motaz Saad. Class outliers mining: Distance-based approach. *International Journal of Intelligent Technology*, 2(1):5568, 2007.
7. Tomek I. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(6):448–452, June 1976.
8. George H. John. Robust decision trees: Removing outliers from databases. In *In Knowledge Discovery and Data Mining*, pages 174–179. AAAI Press, 1995.
9. Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Trans. Knowl. Discov. Data*, 6(1):3:1–3:39, March 2012.
10. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
11. Michael R. Smith and Tony R. Martinez. The robustness of majority voting compared to filtering misclassified instances in supervised classification tasks. *Artif. Intell. Rev.*, 49(1):105–130, 2018.
12. Michael R. Smith, Tony R. Martinez, and Christophe G. Giraud-Carrier. An instance level analysis of data complexity. *Machine Learning*, 95(2):225–256, 2014.
13. D. Randall Wilson and Tony R. Martinez. Reduction techniques for instance-basedlearning algorithms. *Mach. Learn.*, 38(3):257–286, March 2000.

Sémantické modely pre popis dátovo-analytických procesov

Juliana Ivančáková[1], Peter Butka[1], Peter Bednár[1,] Lukáš Kandrik[1]

[1]Katedra kybernetiky a umelej inteligencie, FEI TU v Košiciach
Letná 9, 042 00 Košice
{juliana.ivancakova,peter.butka,peter.bednar}@tuke.sk
lukas.kandrik@student.tuke.sk

Abstrakt. Cieľom práce, ktorá je odprezentovaná v tomto článku je vytvorenie systému, ktorí umožní automatické generovanie skriptov pre úlohy procesov dolovania v dátach. Tento článok sa zameriava na sémantické modely popisujúce procesy v dátovej analytike. V úvode článku je priblížený sémantický web, jeho technológie (RDF, RDFS, OWL) a ontológie, ako sú DSO, EXPO, LABORS a OntoDM. Keďže cieľom je automatické generovanie skriptov pre úlohy procesov dolovania v dátach, tak je stručne popísaná aj metodológie CRISP-DM. Tvorba automatického generovania skriptu je prekladaná zo sémantického grafu na kód v prostredí Python a pri vyhodnotení je tento kód porovnávaný s reálnym kódom algoritmov, ktoré sú typické pre dátovú analýzu.

Kľúčové slová: dolovanie v dátach, DSO, OntoDM, ontológia, sémantický model

1 Úvod

V uplynulých desaťročiach sa uskutočnil veľký počet výskumov nielen v oblastiach klinického výskumu, medicíny, ale aj prírodných vied a mnohých ďalších odvetví. Dnešná moderná doba plná informácií a údajov zaplavuje svet „Big Dátami“ aj vďaka značnému technologickému pokroku. To vedie k exponencionálnemu nárastu tvorby dát, a preto si procesy, zamerané na prácu s dátami, vyžadujú dostatok prostriedkov práve na analýzu a získavania dát. Podniky s technickými znalosťami riadenia veľkých dát nahrádzajú svoje obvyklé odhady a prácne postupy založené na modelovaní dát týkajúcich sa „Big Data“ [1]. Jednou z technológie sémantického webu [2] sú aj ontológie, ktoré pomáhajú interpretovať heterogenitu veľkých dát pomocou spájania konceptov údajov s triedami ontologických údajov. Ontológie organizujú doménové koncepty v hierarchii prostredníctvom logických vzťahov medzi nimi. Preto sémantické mapovanie pomocou slovníkov, lexikónov a mapovania tém prepojuje dátové koncepty s ontologickými triedami. Tým nielen pomáha počítačom interpretovať heterogénne dáta, aby pochopili príslušný kontext, ale môže tiež pomôcť odhaliť anomálie pri veľkých dátach a doplniť chýbajúce informácie.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 129-132.

2 Ontológie

V dnešnej dobe sa pojem ontológia stáva čo raz viac populárnym a to hlavne v oblastiach informačných vied. V súvislosťou s touto terminológiou vidí informatika ontológiu ako technológiu, ktorá sa zameriava na kategorickú analýzu a odpovedá na otázky „Aké sú kategórie subjektov?“ alebo „ Aké sú entity?“ a mnohé ďalšie. Hlavný zmysel použitia ontológie v informatike je vytváranie modelov inžinierskej reality, ktoré budú použité softvérom, priamo interpretované a odôvodnené inferenčnými motormi. V súčasnosti existuje niekoľko špecifických ontológií, ktoré popisujú medicínske, ale aj iné experimenty. Ontológia v oblasti medicíny, OBI, opisuje pojmy medicínskej terminológie a vzťahy medzi nimi, čím umožňuje zdieľanie lekárskych poznatkov a poskytuje štandard pre reprezentáciu biologických a biomedicínskych vyšetrení. Experiment je postup, ktorý sa vykonáva s cieľom overiť, falšovať alebo potvrdiť platnosť hypotézy. EXPO ontológie predstavujú návrh na formalizáciu vedeckých experimentov a rozšírením tohto návrhu je ontológia LABORS.

2.1 Ontológie pre popis dátovo-analytických procesov

Základným prvkom dátovej vedy je formálny opis experimentov pre efektívnu analýzu, anotáciu a zdieľanie výsledkov. Ontológie popisujúce dátovo-analytické procesy, zahrňujú základné informačné subjekty pre reprezentáciu DM a KKD. OntoDM [3] definuje koncepty pre popis scenárov a pracovných tokov pri získavaní dát. Jeho charakteristickým znakom je, že používa BFO ako ontológiu na vyššej úrovni a šablónu, množinu formálne definovaných vzťahov z relačnej ontológie a opätovne používa triedy a vzťahy OBI. Zabezpečuje kompatibilitu a prepojenie s inými ontológiami. Pre popis DM a procesov objavovania znalostí sme využili OntoDM-KDD, ktorý je založený na modeli CRISP-DM.

2.2 Data Science Ontology(DSO)

DSO [4] je znalostná podstata o dátovej vede so zameraním na počítačové programovanie. Okrem katalogizácie a organizovania konceptov dátovej vedy poskytuje ontológiu sémantickej anotácie bežne využívaných softvérových knižníc ako pandas a scikit-learning. Anotácie mapujú typy a funkcie knižníc na univerzálne koncepty ontológie.

Účelom DSO je umožňovať: sémantické dotazy na analýzu dát, porovnanie sémantickej podobnosti medzi dátovými analýzami, automatizovaná štatistická meta-analýzy, meta-učenie pre strojové učenie.

DSO definuje základné koncepty z ontológie ako sú typy úloh, algoritmy, ich nastavenia, operácie pri predspracovaní. DSO disponuje rôznymi typmi úloh z DM, ako napríklad: klasifikácia, regresia, zhlukovanie, detekcia anomálie, asociačná analýza a podobne.

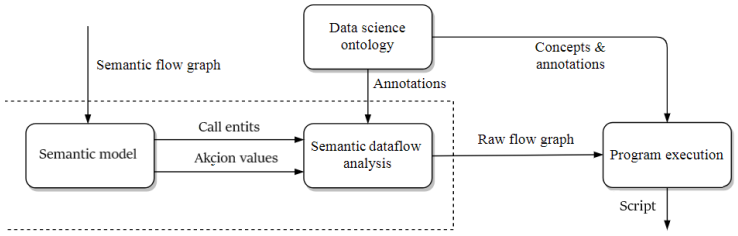
3 Sémantický model pre popis dátovo analytických procesov

Našou hlavnou motiváciou je vytvoriť sémantický model, ktorý by mal umožňovať formálny popis dátovo-analytických procesov, zabezpečiť replikovateľnosť dátovo-analytických procesov a automatické generovanie skriptov pre zadanú úlohu objavovania znalostí v databázach.

Pri definovaní modelu sme vychádzali z ontológie DSO, ktorú sme postupne rozširovali podľa postupu ktorý zahŕňal nasledujúce kroky:

1. **Vytvorenie bázy prípadových štúdií** - popis prípadových štúdií obsahu popis úlohy a dát a skripty na vyriešenie danej úlohy, ktoré boli vytvorené manuálne dátovým analytikom v programovacom jazyku Python.
2. **Mapovanie konceptov na fragmenty** kódu - jednotlivé príkazy kódu skriptov boli anotované konceptami z ontológie DSO.
3. **Zovšeobecnenie mapovania** - fragmenty kódu ktoré boli anotované tým istým konceptom DSO a ktoré vykonávali nad dátami alebo modelmi tie isté operácie boli zovšeobecnené do parametrizovateľnej šablóny, ktorá slúži na automatické generovanie fragmentu kódu pre operáciu definovanú daným konceptom DSO.
4. **Vytvorenie zovšeobecnených modelov pre dátovo-analytické procesy** - pre rôzne typy úloh ako napr. klasifikácia, regresia, zhlukovanie a pod. sme vytvorili zovšeobecnené procesné sémantické grafy zložené z konceptov DSO, ktoré prepájajú operácie potrebné na riešenie daných úloh.

Výsledný model umožňuje pre zadanú úlohu automaticky odvodiť procesný model ktorý je aplikovateľný na riešenie danej úlohy a automaticky vygenerovať skript pre implementáciu tohto procesu v jazyku Python. Celý proces zostavenia a použitia nášho znalostného modelu je uvedený na nasledujúcom obrázku.



Obr. 1 Návrh modelu

4 Testovanie

Testovanie prebehlo na vybraných prípadových štúdiách aplikácie analýzy dát. Pre zvolené úloh bol odvodený sémantický graf, ktorý bol automaticky prevedený na skript v programovacom jazyku Python a porovnaný s kódom vytvoreným manuálne dátovým analytikom. Testovanie sme vyhodnotili rôznymi metrikami pre porovnanie kódu ako napr. celkový počet riadkov kódu, počet zhodujúcich sa funkcií, počet modifikovaných

funkcií, počet pridaných funkcií a celkové pokrytie kódu. Okrem vyhodnotenia pokrytia kódu sme testovali aj presnosť naučených modelov pričom neboli badané výrazne rozdiely v kvalite naučených modelov medzi generovaným a pôvodným kódom.

4.1 Vyhodnotenie analýzy

Model bol aplikovaní na 2 vybrané dátové sety. Jeden dátový súbor obsahoval údaje o poskytovaní pôžičiek na autá. Cieľom bolo zistiť či si noví zákazníci zakúpia poistenie na auto. Úlohou druhého dátového súboru bolo predpovedať počasie. V oboch prípadoch bola teda riešená klasifikačná úloha a súbory mali nominálne a numerické atribúty a obsahovali aj chýbajúce hodnoty.

Tab. 1 Porovnanie

	Počet riadkov kódu	Celkový počet funkcií	Počet rovnakých funkcií	Počet modifikovaných funkcií	Počet pridaných funkcií	Celkové pokrytie kódu (v%)
1.dataset	239/183*	47/35*	9	12	14	74
2.dataset	79/147*	19/24*	2	12	3	92

**porovnávaný kód/náš kód*

5 Záver

Cieľom práce bolo vytvoriť model na automatické generovanie skriptu, ktorý popisuje procesy sémantických modelov pre úlohy dátovej analytiky. Pri prvej analýze dosiahol náš model pokrytie kódu 74% pričom bolo za potreby pridať niekoľko funkcií, ktoré náš model nemohol vypísať. Pri druhej analýze dosiahol náš model značne lepšie pokrytie kódu až 92%. Opakované spúšťanie nášho modelu dosahovalo priaznivé výsledky s odchýlkou +/- 1,43%.

Poďakovanie: Tento príspevok vznikol s podporou APVV v rámci projektov APVV-16-0213 a SK-AT-2017-0021, ako aj v rámci podpory VEGA projektu č. 1/0493/16.

Literatúra

1. Inmon, W. H. & Linstedt, D.: Data Architecture: a Primer for the Data Scientist 49–55 (Morgan Kaufmann, 2015).
2. Hitzler, P., Krotzsch, M. & Rudolph, S.: Foundations of Semantic Web Technologies. (CRC Press, 2009).
3. Panov, P., Džeroski, S., and Soldatova, L.N. (2010). Representing Entities in the OntoDM Data Mining Ontology. In Inductive Databases and Constraint-Based Data Mining, S. Džeroski, B. Goethals, and P. Panov, eds. (New York, NY: Springer New York), pp. 27–58.
4. PATTERSON E., BALDINI I., MOJSILOVIC A., VARSHNEY K. What is the Data Science Ontology? [online] [cit. 28.06.2019] Dostupné z: https://www.datascienceontology.org/help.

Vylepšení klasifikace textových dokumentů algoritmem n-gramů pomocí crowdsourcingu

Petr Šaloun[2], David Andrešič[1], Barbora Cigánková[1], Milan Klement[2]

[1]Vysoká škola báňská – Technická univerzita Ostrava
17. listopadu 2172/15, 708 00 Ostrava – Poruba
[2]Univerzita Palackého v Olomouci
Křížkovského 511/8, 771 47 Olomouc
{david.andresic,barbora.cigankova.st}@vsb.cz
{petr.saloun,milan.klement}@upol.cz

Abstrakt. Běžnou úlohou zpracování přirozeného jazyka je klasifikace. Tato úloha je nejlépe vykonávána člověkem, ačkoliv v některých aplikacích si můžeme dovolit mírnou ztrátu přesnosti výměnou za rychlost. Zde najde uplatnění zpracování přirozeného jazyka, které zpracuje text do podoby srozumitelné některému z klasifikátorů, jako např. k-nearest neighbor, rozhodovací strom, umělá neuronová síť, nebo SVM. Ke zlepšení přesnosti těchto automatických výsledků však můžeme použít lidský element prostřednictvím crowdsourcingu. Cílem této práce je vytvořit a uvést do praxe klasifikátor textových dokumentů (algoritmus n-gramů) a připravit rozhraní pro vyhodnocování a vylepšování klasifikace pomocí crowdsourcingu. Jeho úkolem je totiž kromě sběru dat také vyhodnocení přesnosti klasifikace, což dále rozšiřuje tréninkovou sadu klasifikátoru. Náš postup jsme otestovali na dvou datových sadách, kde dosahoval slibných výsledků napříč různými jazyky. To vedlo k jeho uvedení do běžného provozu na začátku roku 2019 v kooperaci univerzit VŠB-TUO a OU.

Klíčová slova: klasifikace, textové dokumenty, zpracování přirozeného jazyka, n-gramy, Crowdsourcing, WordPress

1 Úvod

Navzdory pokrokům v klasifikací textů pomocí Natural language processing (NLP) jsou lidé stále přesnější, což otevírá prostor pro lidskou asistenci, např. pomocí crowdsourcingu. V této práci popisujeme naše zkušenosti a předběžné výsledky klasifikace textových dokumentů psaných v češtině a obecný přehled dnešních algoritmů se zaměřením na češtinu. Dále také přinášíme stručné shrnutí výhod crowdsourcingu. Náš postup jsme testovali na dvou datových sadách a porovnali s použitím crowdsourcingu.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 133-136.

2 Zpracování textů v přirozeném jazyce a přehled klasifikátorů

Při klasifikaci textových dokumentů je potřeba vždy provést následující kroky [1]:

- Odstranění stop slov, tagů a další pre-processing (viz část 2).
- Extrakce různých vlastností textu [1].

Dnešní klasifikátory používají buď statistické postupy, nebo strojové učení. *Naivní Bayesův klasifikátor* [2] má např. výhodu v dobré přesnosti při malé učící sadě. *Term Frequency-Inverse Document Frequency (TF-IDF)* je statistická metrika, která měří důležitost slov v daném dokumentu [1, 3]. *Latentní sémantická analýza* je založena na analýze vztahů mezi sadou dokumentů a jejich slovy [4, 5]. *Support Vector Machines* je metodou strojového učení založenou na principu rozhodovacích ploch [6, 7]. *N-gram* je uspořádaná N-tice prvků patřících do nějaké sekvence např. slov, nebo písmen používaná mj. pro klasifikaci dokumentů, kde začátek a konec slova je označen speciálním znakem [8]. Velkou výhodou je nezávislost na jazyku dokumentu, protože neprobíhá žádný pre-processing a jistá tolerance k překlepům a gramatickým nepřesnostem.

Celý proces NLP má několik fází [10], především však [9] morfologickou, syntaktickou a sémantickou analýzu. Zpracování textu vyžaduje určitý pre-processing: převod na malá písmena, odstranění speciálních znaků, stop slov a tokenizace. Dále často stemmatizaci a lemmatizaci [11, 12**].** Čeština patří ke složitějším jazykům pro zpracování. Jedním z mála použitelných frameworků je Apache Lucene[1] obsahující česká stop slova, stemmer a tokenizér. Lemmatizér lze kompenzovat analyzátorem Majka [13].

2.1 Crowdsourcing

Může být definován jako způsob práce, ve kterém je aktivita outsourcována davu [14] jako všeobecná výzva [15]. Základní myšlenkou je tedy kolektivní inteligence [16]. Klady a zápory crowdsourcingu nabízí [17]. Způsoby, jak dav motivovat pak [18].

3 Prototyp a předběžné výsledky

Z dostupných algoritmů jsme zvolili n-gramy implementované dle [8], avšak s odstraněním stop slov umožňujícím začít v profilu od začátku seznamu. Náš klasifikátor také kvůli psychologickým textům s tenkou hranicí mezi třídami pracuje s delšími profily. Kromě klasifikace dále určí klíčová slova každé třídy. Pět s největší váhou pak doporučí vybraným uživatelům, což by mělo potvrdit hypotézu o přesnější klasifikaci pro příspěvky s předdefinovanými klíčovými slovy. Kalkulace klíčových slov je realizována pomocí TF-IDF modifikovaným dle [19] pro účely třídy. Chyby klasifikace opraví přímo uživatel. Takovýto příspěvek je přidán do tréninkové sady.

Jazyková datová sada obsahovala texty v češtině, slovenštině a angličtině. Každá kategorie obsahovala 40 textů s 20 až 60 slovy. Tréninková i testovací sada obsahovala

[1] Apache Lucene: https://lucene.apache.org

20 textů pro každou třídu. Zdrojem byly technické texty z konference DATAKON (2010, 202, 2013 a 2014). Výsledky můžeme vidět v tab. 1, kde naše očekávání předčily výsledky pro češtinu a slovenštinu. Jazyková nezávislost algoritmu byla potvrzena.

Psychologická sada obsahující méně vyvážené a strukturované texty obtížně klasifikovatelné i pro člověka. Celkem obsahovala 87 textů v kategoriích osobních problémů a nemoci (63), práce, financí a školy (8) a vztahů, rodiny a zaměstnání (14). Poměr tréninkových a testovacích dat byl 80:20. Výsledky vidíme v tab. 1. V každé třídě byl jeden špatně klasifikovaný dokument, což přisuzujeme struktuře textu (úvod popisoval něco trochu jiného), nebo marginálním případům obtížným i pro člověka.

Tab.1. Přesnost klasifikace pro jazykovou a psychologickou datovou sadu.

Jazyková sada		Psychologická sada	
Třída	Úspěšnost	Třída	Úspěšnost
Angličtina	20/20	Osobní problémy, nemoci	12/13
Slovenština	20/20	Práce, finance, škola	1/2
Čeština	20/20	Vztahy, Rodina, zaměstnání	2/3

Pro crowdsourcing bylo navrhované téma příspěvků „*život neformálních pečujících a jeho ovlivnění jako následek této péče*“ s následujícími třídami: *motivace, benefity a následky*, *podpora pečovatelů* a *potřeby pečovatelů*. Tréninková sada obsahovala 180 textů s max. 2 větami a 4 třídami. Příspěvky tvořili a ověřovali studenti LF OU. Špatně klasifikované rozšířily tréninkovou sadu. Bylo přidáno 8 příspěvků (2 na třídu) jedním autorem (tzn. podobný slovník). Nejdříve do každé třídy 1 příspěvek, což vedlo k nulové přesnosti způsobené rozdílnou povahou oproti tréninkovým datům. Po synchronizaci dat byla provedena stejná druhá fáze s přesností 50%. I když je vidět zlepšení přesnosti a učení klasifikátoru, není vzorek dostatečně rozsáhlý pro závěry.

4 Závěr

Hlavním cílem této práce bylo vytvořit prototyp klasifikátoru textových dokumentů založený na algoritmu n-gramů s crowdsourcingem pro zlepšení přesnosti. První datová sada potvrdila jazykovou nezávislost algoritmu n-gramů. S druhou (psychologickou) si algoritmus také poradil velmi dobře (špatně klasifikoval pouze případy složité i pro člověka). Nakonec s klasifikací pomáhal člověk, čímž se do tréninkové sady dostaly dříve chybně klasifikované texty. S ohledem na malé množství těchto textů nelze dělat žádné závěry, i když zlepšení zde vidět je. Další práce bude zahrnovat potvrzení hypotézy o zvýšení přesnosti pomocí doporučených klíčových slov a také rozšiřování tréninkové sady o chybně klasifikované texty. Tato práce úzce souvisí s naší další činností popisovanou v [20]. Zde jsme se zaměřili především na algoritmus n-gramů a zakomponování crowdsourcingu. Výsledky vedly k realizaci celého projektu v kooperaci VŠB-TUO a OU na počátku roku 2019 a uvedení do reálného provozu.

Poděkování: Výzkum byl částečně podpořen projektem TAČR č. TL02000050.

Literatura

1. KARMAN, S. Senthamarai; RAMARAJ, N. Similarity-Based Techniques for Text Document Classification. Int. J. SoftComput, 2008, 3.1: 58-62.
2. OPITKA, P.; ŠMAJSTRLA, V. „PRAVDĚPODOBNOST A STATISTIKA," [In Czech] (Probability and statistics) 2013. [Online]. Available: https://homen.vsb.cz/~oti73/cdpast1/KAP02/PRAV2.HTM. [Accessed on 4. 3. 2018].
3. „Tf-idf : A Single-Page Tutorial - Information Retrieval and Text Mining," [Online]. Available: http://www.tfidf.com/. [Accessed on 25. 12. 2017].
4. LANDAUER, Thomas K.; FOLTZ, Peter W.; LAHAM, Darrell. An introduction to latent semantic analysis. Discourse processes, 1998, 25.2-3: 259-284.
5. HÁJEK, Petr, et al. Možnosti využití přístupu indexování latentní sémantiky při předpovídání finančních krizí. POLITICKÁ EKONOMIE, 2009, 6: 755.
6. „Support Vector Machines (SVM)," TIBCO Software Inc, [Online]. Available: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Accessed on 28. 12. 2017].
7. ŽIŽKA, J. „Studijní materiály předmětu FI:PA034," (Study materials to FI:PA034) [Online]. Available: https://is.muni.cz/el/1433/podzim2006/PA034/09_SVM.pdf. [Accessed on 29. 12. 2017].
8. CAVNAR, William B., et al. N-gram-based text categorization. Ann arbor mi, 1994, 48113.2: 161-175.
9. HABROVSKÁ, P. „Vybrané kapitoly z počítačového zpracování přirozeného jazyka," 2010. [In Czech] (Seleted chapters from natural language processing) [Online]. Available: http://www.inflow.cz/kratce-o-zpracovani-prirozeneho-jazyka.
10. SCAGLIARINI, L.; VARONE, M. „Natural language processing and text mining," 11 April 2016. [Online]. Available: http://www.expertsystem.com/natural-language-processing-and-text-mining/. [Accessed on 15. 12. 2017].
11. KODIMALA, Savitha. Study of stemming algorithms. 2010.
12. RISUENO, T. „The difference between lemmatization and stemming," 28. 1. 2018. [Online]. Available: https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/. [Accessed on 4. 3. 2018].
13. ŠMERK, P.; RYCHLÝ, P. „Majka – rychlý morfologický analyzátor," (Majka - quick morphological analyzer) 2009. [Online]. Available: https://www.muni.cz/vyzkum/publikace/935762. [Accessed on 15. 12. 2017].
14. ESTELLÉS-AROLAS, E.; GONZÁLEZ-LADRÓN-DE-GUEVARA, F. Towards an integrated crowdsourcing definition. Journal of Information science, 2012, 38.2: 189-200.
15. SCHENK, Eric; GUITTARD, Claude. Crowdsourcing: What can be Outsourced to the Crowd, and Why. In: Workshop on Open Source Innovation, Strasbourg, France. 2009.
16. AITAMURTO, T.; LEIPONEN, A.; TEE, R. The promise of idea crowdsourcing–benefits, contexts, limitations. Nokia Ideasproject W.P., 2011, 1: 1-30.
17. KALSI, M. „Crowdsourcing through Knowledge Marketplace," 3. 3. 2009. [Online]. Available: http://blog.spinact.com/knowledge_as_a_service/2009/03/crowdsourcing-through-knowledge-marketplace-.html. [Accessed on 2018 3. 4.].
18. KAUFMANN, N.; SCHULZE, T.; VEIT, D. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. In: AMCIS. 2011. p. 1-11.
19. VRL, NICTA. An unsupervised approach to domain-specific term extraction. In: Australasian Language Technology Association Workshop 2009. 2009. p. 94.
20. Andrešič, D., Ondrejka, A, Šaloun, P., Cepláková, R.: Webový portál pro identifikaci poruchy osobnosti z psaného textu. In: Proc. of 12th Workshop on Intelligent and Knowledge oriented Technologies 2017 (2017).

Využitie strojového učenia v stávkovom systéme: predikcia výsledku futbalového zápasu

Marek Ružička[1], Juraj Gazda[1]

[1]Katedra počítačov a informatiky, FEI TU v Košiciach
Letná 9, 042 00 Košice
marek.ruzicka@tuke.sk,
juraj.gazda@tuke.sk

Abstrakt. Využitie strojového učenia pre predikciu najpravdepodobnejšieho výsledku zápasu je v tejto dobe stále považované za náročnú úlohu. Stávkové spoločnosti využívajú tieto dáta pre určovanie stávkových kurzov a generovanie zisku. Výber vhodných dát použitých pre trénovanie modelov je veľmi dôležitý. V tomto príspevku sa pokúšame určiť najvhodnejší dátový model ako aj algoritmus spomedzi vybraných. Popri tom analyzujeme aj to, ako vplýva počet sezón použitých pre predikciu na celkovú úspešnosť. V neposlednom rade analyzujeme aj to, či je lepší prístup predikcie troch možných výsledkov v jedinom kroku, alebo metódou oddelenej predikcie. Presnosť natrénovaných modelov bola hodnotená testovaním na dátovej množine jednej celej sezóny, kde sme zaznamenali výsledky v niektorých prípadoch cez 60%, čo presahuje výsledky súčasných stávkových kancelárií. Pre validáciu sme využívali dáta získané z anglickej futbalovej súťaže Premier League.

Kľúčové slová: stávkovanie, strojové učenie, predikcia výsledku zápasu

1 Úvod

Využitie strojového učenia v športových systémoch je stále dôležitou oblasťou výskumu. V minulosti už vyšlo v tejto oblasti viacero prác, no stále je obrovský priestor pre zlepšenia.

Jednou z prvých prác v tejto oblasti bola práca Michaela Puruckera, Neural network quarterbacking [1]. V tejto práci sa pokúšal pomocou strojového učenia predikovať výsledky zápasov americkej Národnej futbalovej (rugby) ligy (NFL). Dátová sada však obsahovala iba zápasy z poslednej sezóny a testovacia sada obsahovala iba 14 zápasov. Najlepšiu úspešnosť, ktorú dosiahol, bola 61%.

V tomto výskume pokračoval v roku 2003 Kahn, ktorý využitím viacerých atribútov dvihol úspešnosť na 75% [2]. Pravdepodobnosť remízy je však pri rugby veľmi nízka, a preto sa dá v tomto prípade dosiahnuť relatívne vysoká úspešnosť.

McCabe a Trevathan pokračovali v podobnom výskume v práci Artificial Intelligence in Sport Prediction [3]. Skúmali štyri rôzne ligy, mimo iné aj NFL a Premier league. Pri Premier league dosiahli výsledky v priemere 54,6%, pričom

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 137-141.

najlepšie skóre bolo 58,9%. Pri NFL sa pohybovala úspešnosť medzi 63 a 67%. V tejto práci sa vyúžili výsledky z viacerých sezón.

Tax a Joustra využili vo svojej práci výsledky z 13 rokov nemeckého futbalu [4]. Porovnávaním viacerých algoritmov dosiahli najlepší výsledok 56,1%.

V diplomovej práci Machine learning for Soccer analytics autor použil cez 300 atribútov týkajúcich sa tímu a hráčov [5]. V tomto prípade autor dosiahol najlepší výsledok 53,4% a odôvodnil to zložitosťou predikcie remízy.

Vo webovom portáli kickoff.ai, ktorý využíva umelú inteligenciu na predpovedanie výsledku zápasov rôznych futbalových líg sme analýzou posledných 100 zápasov zistili, že správne predikovaných bolo 56% a z toho ani jeden predikovaný výsledok nebola remíza.

Podobné výsledky vzhľadom na remízy sme pozorovali aj v dátach využívaných v našej práci. Zo 7830 zápasov v 21 sezónach bolo správne predikovaných približne 54%, z čoho iba 9 zápasov bolo predikovaných ako remíza. Z týchto 9 zápasov iba dva naozaj skončili remízou.

Z analýzy prác a riešení ktoré sme mali k dispozícii vyplýva, že v predikcii výsledkov zápasov je veľmi kritickým bodom predikcia remízy.

2 Metódy

Hlavnými cieľmi práce bolo určiť, ktorý algoritmus je najvhodnejší pre daný problém, ako vplývajú na predikciu dlhodobé dáta, aký model vybrať a pri koľkých sezónach v trénovacích dátach dosiahneme najlepšie výsledky. Na dosiahnutie týchto cieľov sme potrebovali najprv vytvoriť správnu dátovú sadu. Začali sme s dátovou sadou obsahujúcou iba základné informácie ako názvy tímov, počet strelených gólov, výsledok zápasu, herný týždeň, dátum konania a v minulosti určené kurzy na jednotlivé výsledky. Distribúcia výsledkov domáceho tímu v 7830 zápasoch bola nasledovná: výhry 46,45%, prehry 27,98% a remízy 25,57%. Ďalej sme z dostupných zdrojov získali informácie o finálnych umiestneniach tímov v jednotlivých rokoch. Z týchto dát sme boli schopní vypočítať rôzne atribúty prislúchajúce aktuálnej forme ako napríklad výsledky posledných piatich zápasov, série výhier či prehier, body vo forme, strelené a inkasované góly a podobne. Tieto atribúty reprezentovali formu oboch tímov v danom zápase. Forma sa vzťahuje na výkon tímov od začiatku aktuálnej sezóny.

Neskôr sme pridali do dátovej sady aj informácie o dlhodobých štatistikách. Z oficiálnej stránky Premier league sme získali informácie o celkovom pôsobení tímov v lige. Agregáciou týchto hodnôt sme boli následne schopní získať množstvo ďalších atribútov platných k danému zápasu. Príkladom takýchto atribútov sú počty odohratých sezón, prehier, výhier, inkasovaných a strelených gólov a pod., za celé pôsobenie v súťaži až po daný zápas. Z týchto hodnôt sme taktiež vypočítali priemery a rozdiely, ktoré definovali, ktorý tím je z dlhodobého hľadiska lepší. Príkladom môže byť pomer medzi výhrami a prehrami tímu po daný zápas, inkasovaných a strelených gólov tímu, priemerné počty inkasovaných a strelených gólov na zápas, priemerné postavenie v sezónach a podobne. Tieto priemerné hodnoty sme určili agregáciou hodnôt od začiatku pôsobenia tímu v súťaži po daný zápas.

Na určenie vyhovujúceho počtu sezón pre trénovanie modelov sme sa rozhodli pre jednoduché cyklické testovanie modelov tak, aby sme pri každej iterácii odobrali z dátovej sady najstarších 380 zápasov, čiže jednu sezónu. Vzhľadom na množstvo testovacích dát, ktoré sme využívali, a to 380 zápasov, sme skončili pri 10 sezónach. V opačnom prípade by bola testovacia sada príliš veľká a výsledky by mohli byť skreslené.

Využitím predchádzajúceho cyklu sme taktiež určovali aj úspešnosť vybraných algoritmov – lineárna regresia, logistická regresia, metóda podporných vektorov (SVM), náhodný les (RF) a gradient boosting (XGB). V každej iterácii sme skúsili všetkých 5 algoritmov s práve používanou veľkosťou dátovej sady. Pred tým, ako sme testovali jednotlivé algoritmy, sme sa pokúsili určiť najvhodnejšie parametre pre jednotlivé klasifikátory.

Posledným testom sme sa snažili určiť vhodnú metódu predikcie. Do tohto bodu sme predikovali jednoducho všetky tri možné výsledky naraz. V tomto teste sme navrhli metódy, kde by sme zjednodušili predikciu na binárny problém a snažili sa určiť pravdepodobnosť iba výhry, alebo iného výsledku. V druhom kroku sme sa snažili určiť pravdepodobnosť prehry alebo iného výsledku. Spojením týchto pravdepodobností sme nakoniec získali výslednú úspešnosť.

3 Výsledky simulácií

Testovanie sme započali s dátovou sadou obsahujúcou iba informácie o zápasoch a súčasnej forme tímov. Výsledkom týchto simulácií však boli hodnoty nižšie ako výsledky v skúmaných prácach. Preto sme dátovú sadu rozšírili o dlhodobé štatistiky tak, ako sme spomínali v predchádzajúcej kapitole.

Po rozšírení o dlhodobé štatistiky sme okamžite boli schopní badať prvé výsledky. V tabuľke 1 môžeme vidieť výsledky simulácií pri znižujúcom sa množstve trénovacích dát. Algoritmus lineárnej regresie nedosiahol v žiadnom prípade výsledky lepšie ako 35%, preto sme ho v tabuľke neuviedli. Algoritmus SVM spomedzi vybraných dosahoval výsledky pod úrovňou publikovaných prác, preto tvrdíme, že nie je pre daný problém vhodný. Zvyšné algoritmy vykazovali výsledky nad priemerom, menovite 59,17% v priemere u logistickej regresie, 59,67% pri algoritme RF, a najlepšie výsledky dosiahol algoritmus XGB – priemerne 60,2%, najlepšie 61,05%.

Správnosť hypotézy, že odobratím časti dát sa môžu zlepšiť celkové výsledky sa potvrdila len čiastočne, pričom každý algoritmus sa správal inak. Preto nie je úplne jednoznačné, aké množstvo dát vyhovuje trénovaniu najviac. Najlepšie skóre algoritmov sme v tabuľke vyznačili silnejšie.

V tejto chvíli je nutné spomenúť aj spôsob, akým sa modely testovali. Na začiatku sme modely testovali štandardnou metódou náhodného premiešania dátovej sady. Takto sa však do testovacieho súboru nemuseli dostať rovnomerne dáta z každého hracieho týždňa. Preto sme prístup zmenili a testovali sme modely s dátami z jedinej (poslednej) hracej sezóny. Výsledky tejto práce teda vypovedajú, ako by bol systém schopný predpovedať výsledky na súčasných dátach.

Okrem tohto sme v práci testovali aj dátumové atribúty dňa a mesiaca zápasu. Využitím týchto atribútov sme však nezaznamenali signifikantnú zmenu úspešnosti, takže sme ich ďalej netestovali.

Lepšie výsledky sme zaznamenali aj pri predikcii remízy. V skúmanej testovacej vzorke 380 zápasov algoritmus logistickej regresie predikoval remízu 108-krát a z toho 71-krát správne. Obdobné výsledky mali aj ostatné algoritmy. Tieto výsledky sú značne lepšie ako výsledky v skúmaných prácach.

Sezóny	LogR	SVM	XGB	RF
21	0.5868	0.5263	**0.6105**	0.5974
20	0.5947	0.5316	0.5974	0.6026
19	0.5974	0.5368	0.6026	**0.6053**
18	**0.6**	0.5368	0.6026	0.6
17	0.5947	0.5395	0.5974	0.6026
16	0.5974	**0.5684**	0.6	0.5947
15	0.5895	0.5474	0.5974	0.5895
14	0.5737	0.5526	0.6079	0.5816
Priemer	0.591775	0.542425	**0.601975**	0.5967125

Tabuľka 1.: Úspešnosť predikcie jednotlivých algoritmov pri meniacom sa počte sezón

Pri modeli kde sme predpovedali výhru a prehru osobitne sme zistili, že vzhľadom na slabú distribúciu prehier v dátovej sade vzniká veľká chyba. Z tohto dôvodu sme sa nedostali na hodnotu vyššiu ako 52%, a preto nemá zmysel túto metódu využívať.

4 Záver a diskusia

V tejto práci sme dokázali, že aj využitím jednoduchých a dostupných dát ohľadom športových výsledkov sme schopní predpovedať výsledky s vyššou presnosťou ako skúmané práce. Pri troch algoritmoch sme sa dokonca dostali cez 59%. Suverénne najvhodnejším algoritmom je XGB, kde sme dosiahli najvyššiu úspešnosť 61,05%. Zlepšenie sme zaznamenali aj pri predikcii remízy samotnej, kde sme mali správne predpovedaných až 65% predikovaných výsledkov.

Pri metódach sa ukázala ako vhodnejšia metóda priamej predikcie, čo bolo spôsobené nízkou distribúciou prehier a remíz oproti výhram domáceho tímu.

Vplyv dátumových atribútov a veľkosti dátovej sady ostáva otázny a bol by potrebný hlbší výskum v tejto oblasti.

V budúcnosti by bolo možné výsledky tejto práce vylepšiť rôznymi ďalšími štatistikami, či už informáciami o hráčoch, počasí, alebo aj pozorovaním množstva pozitívnych a negatívnych zmienok v správach či na sociálnych sieťach. Nami dosiahnuté výsledky presiahli výsledky skúmaných prác, ale stále je predpoklad možného zlepšenia.

Poďakovanie: Na tomto mieste by som chcel poďakovať Ing. Michalovi Kováčikovi za pomoc so svetom stávkových kancelárií, Ing. Marekovi Kunštárovi za ozrejmenie

niektorých aspektov strojového učenia a Mgr. Petrovi Dubóczimu za rady ohľadom futbalového sveta.

Literatúra

1. Purucker,Michael C.: Neural network quarterbacking, IEEE Potentials 15.3, 1996, 9–15. DOI: 10.1109/45.535226
2. Kahn, Joshua: Neural Network Prediction of NFL Football Games, In World Wide Web electronic publication 2003, 9–15. http://homepages.cae.wisc.edu/~ece539/project/f03/kahn.pdf
3. McCabe, A., Trevathan, J.: Artificial Intelligence in Sports Prediction. In Fifth International Conferenceon Information Technology: New Generations, 2008, 1194–1197. DOI 10.1109/itng.2008.203
4. N. Tax, Y. Joustra.: Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach. Transactions on Knowledge and Data Engineering 10.10, 2015, 1-13. DOI 10.13140/RG.2.1.1383.4729
5. Kumar, G.: Machine learning for soccer analytics.Master's thesis, Katholieke Universiteit, Leuven, Belgium, 2013. DOI: 10.13140/RG.2.1.4628.3761

Identifikácia zmätenia používateľa vo webovej aplikácii

Michal Hucko, Mária Bieliková

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

Abstrakt. Zmätenie požívateľa vo webovej aplikácií je problém, ktorý výrazne zhoršuje používateľský zážitok. Zmätený používateľ nevie ako má v aplikácii postupovať, čo môže viesť až k opusteniu webovej aplikácie. V tomto príspevku sa venujeme identifikácii zmätenia používateľa vo webovej aplikácii v reálnom čase s využitím interakčných dát zo správania. Navrhujeme metódu predikcie zmätenia v reálnom čase založenú na klasifikácii správania počas časového okna. Ukazujeme vplyv veľkosti okna na úspešnosť klasifikácie. Na overenie výsledkov sme zostavili používateľskú štúdiu na dovolenkovom portáli *FiroTour* a v aplikácii poisťovne *Aegon* ako súčasťou aplikácie *YesElf*, určenej na vytváranie sprievodcov vo webových aplikáciách. Náš modul predikcie zmätenia v reálnom čase predstavuje základ pre personalizáciu týchto sprievodcov.

Kľúčové slová: zmätenie, klasifikácia, predikcia v reálnom čase, YesElf

1 Úvod a existujúce prístupy predikcie zmätenia

Webové aplikácie sú dnes prepojené s naším každodenným životom. Pracujeme v nich, nakupujeme či dokonca vypĺňame voľný čas. Problémy začnú, keď sa používateľ dostane do stavu zmätenia. Slovník *Webster* charakterizuje zmätenie ako stav, v ktorom si človek nie je istý čo má urobiť, alebo ako to má urobiť. Zmätený človek vynaloží úsilie na to, aby prekonal tento stav. Toto úsilie je zväčša spojené s hľadaním očakávanej informácie alebo realizovaním služby. U každého však existuje hranica, kedy túto snahu vzdáme a aplikáciu jednoducho opustíme.

Jedným z bežných scenárov je situácia, počas ktorej používateľ navštívi aplikáciu prvýkrát. Neznalosť rozhrania, množstvo obsahu a mnohé iné faktory spôsobujú, že nastane zmätenie. Toto je situácia, ktorú sa snaží riešiť aplikácia *YesElf* (www.yeself.com), ktorá umožňuje pridanie sprievodcov do akejkoľvek webovej aplikácie. Zameriava sa na firemné systémy a prípad, keď prichádzajú noví používatelia. V súčasnosti sa vynakladajú veľké prostriedky na zaškoľovanie nových zamestnancov (angl. *onboarding*). Sprievodcovia majú za úlohu zefektívniť tento proces.

Zmiasť sa však môže aj skúsený používateľ. Vezmime si situáciu, kedy vlastník aplikácia nahrá novú zmenu do aplikácie. Bežná funkcionalita sa zmení a zmiasť sa

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 142-146.

môže aj skúsený používateľ. Ďalšou situáciou môže byť návšteva aplikácie po dlhej pauze.

Identifikáciou stavu zmätenia sa zaoberalo niekoľko prác. Thomas a kol. v práci [6] predstavili metódu na idetifikáciu obtiažnych sedení vo webovej aplikácii. Autori používajú na miesto pojmu zmätenie, obtiažne sedenie (angl. *strugling session*), ale v princípe ide o ten istý stav. Ich metóda je založená na klasifikácii sedenia po jeho uplynutí. Využívajú pri tom črty z myši (počet skrolov, počet klikov a iné) a črty odvodené z HTTP serveru. Jedná sa o črty, ktoré vieme získať spracovaním HTTP záznamov na serveri ako napríklad dĺžka sedenia, počet stránok a iné. Ich klasifikátor dosahuje presnosť až 0,84 avšak neposkytujú metódu, ktorá by sa dala aplikovať v reálnom čase.

Lallé a kol. v práci [5] predstavili metódu predikcie zmätenia používateľa v reálnom čase. Ich prístup je založený na interakčných dátach z okulografu. Na overenie metódy navrhli používateľskú štúdiu so 136 účastníkmi v aplikácii na prenajímanie domov. Na označenie momentu zmätenia autori implementovali softvérové tlačidlo v pravom hornom rohu obrazovky. Na predikciu zmätenia autori rozdelili sedenia na časové okná rovnakej dĺžky, ktoré využili pri klasifikácii. Ich metóda dosiahla presnosť 0,6 pri predikcii zmätených pozorovaní s využitím algoritmu náhodného lesa. Ako najdôležitejšia črta predikcie sa ukázala vzdialenosť hlavy od monitora.

V tomto príspevku prezentujeme metódu na predikciu zmätenia v reálnom čase na základe interakčných dát správania používateľa webovej aplikácie. Naša metóda využíva primárne dáta o správaní myši a snaží sa nájsť presný moment kedy zmätenie nastalo. Riešenie je priamo súčasťou aplikácie *YesElf* a umožňuje personalizované odporúčanie sprievodcov zmäteným používateľom.

2 Metóda predikcie zmätenia používateľa v reálnom čase

Naša metóda predikcie zmätenia v reálnom čase je založená na využití interakčných dát o správaní. Pod pojmom reálny čas rozumieme odhadnutie presného momentu, kedy zmätenie používateľa nastane. V súčasnosti disponujeme výsledkami zo spracovania dát z myši. V dohľadnej dobe však pracujeme na pridaní ďalších zdrojov, ktorými je napríklad klávesnica.

Metóda pozostáva z týchto krokov:

1. *Predspracovanie surových dát.* V tomto kroku čistíme dáta (mazanie duplikátov, mazanie neúplných dát). Pre realizáciu predikcie v reálnom čase nazbierané dáta delíme do časových okien rovnakej veľkosti. Časové okno je súbor záznamov myši s určenou dĺžkou trvania. Tento prístup bol použitý na predikciu zmätenia v reálnom čase v práci [5]. V našej metóde využívame veľkosti okien 3 a 5 sekúnd.
2. *Extrakcia čŕt.* V tomto kroku pristúpime k výpočtu čŕt myši. Ide o črty: horizontálna rýchlosť, vertikálna rýchlosť, rýchlosť, zrýchlenie, akcelerácia, hybný moment, vzdialenosť, počet pohybov, celkové trvanie pohybov. Nami vybrané črty boli použité v prácach [2,3] pri autentifikácii používateľa na základe biometrie. Pod pojmom pohyb rozumieme súbor záznamov myši, kde

časový rozdiel medzi záznamami je menší ako 120ms. Tento pojem bol definovaný v práci [2].

3. *Trénovanie modelu.* V tomto kroku pristúpime k samotnému trénovaniu. V prípade nevyváženého datasetu používame techniku nadvzorkovania minoritnej triedy pomocou algoritmu *SMOTE* založenom na technike k najbližších susedov predstavenom v práci [1]. Minoritnú triedu nadvzorkujeme o 200 a 500 percent. Pre trénovanie sme vybrali algoritmy logistickej regresie a náhodného lesu.
4. *Predikcia.* Po natrénovaní modelov sa vyberie najúspešnejší. Kritérium vyhodnotenia je maximalizovanie metriky presnosti pri predikcii zmätenia. Dôvod výberu je prípadné produkčné nasadenie modelu, kde chceme kde má cenu chceme predikovať zmätenie čo najpresnejšie.

3 Vyhodnotenie

Na overenie metódy sme navrhli používateľskú štúdiu, počas ktorej účastníci riešili úlohy na vybranej webovej aplikácii. Výber úloh reflektuje bežné používanie aplikácie a zameriava sa vždy na problémové miesta v aplikácii.

Počas celého experimentu sme zaznamenávali správanie používateľa pomocou logeru, ktorý zaznamenáva polohu kurzora a správanie klávesnice. Frekvencia zaznamenávania udalostí je 60 vzoriek za sekundu. Na získanie explicitného bodu zmätenia používame softvérové tlačidlo v pravom hornom rohu obrazovky. Využitie tejto techniky na rovnaký problém opisuje Lalle [5]. Podobne aj my poskytujeme účastníkom opis funkcie tlačidla na zaznamenávanie zmätenia pred začatím samotnej štúdie. Tento opis bol dostupný účastníkom v tlačenej forme počas celého experimentu.

Každá z úloh pozostávala vždy z troch častí:

- *Obrazovka s inštrukciami.* Na tejto obrazovke si účastníci môžu prečítať inštrukcie potrebné k splneniu danej úlohy. Inštrukcie sú okrem toho vytlačené na priloženom papieri pre prípadné dohľadanie počas samotného riešenia.
- *Riešenie úlohy.* Účastníci riešia úlohu pričom zaznamenávajú bod zmätenia pomocou softvérového tlačidla.
- *Obrazovka s dotazníkom.* V prvej otázke dotazníku účastník zodpovie informáciu, ktorú bolo potrebné na danej úlohe dohľadať. V nasledujúcej otázke odpovie na dôvod stlačenia tlačidla zmätenia.

Na vyhodnotenie experimentu sme zorganizovali používateľskú štúdiu v dvoch aplikáciách. Prvou z nich bola aplikácia dovolenkového portálu *FiroTour,* kde 59 účastníkov riešilo 6 úloh. Počas riešenia nastalo kliknutie na tlačidlo zmätenia v 95 prípadoch. Druhou aplikáciou bola interná aplikácia poisťovne *Aegon*. 20 zamestnancov firmy riešilo 6 úloh a kliknutie nastalo v 30 prípadoch. Do datasetu sme zaradili len kliky na tlačidlo zmätenia, ktoré zo slovného opisu naozaj súviseli so zmätením.

Dôvodom výberu dvoch odlišných aplikácií je porovnanie úspešnosti metódy v rámci domén. Chceme určiť aj ako veľmi všeobecná môže naša metóda byť. Na implementáciu modelov sme použili jazyk *python* a knižnice rámca *sci-kit learn*. Na

výber optimálnych parametrov modelov sme použili prehľadávanie do mriežky. Na vyhodnotenie modelov sme použili vrstviacu trojskupinovú krížovú validáciu.

Tab. 1 a 2 poskytujú predbežné výsledky natrénovaných modelov. Najúspešnejší model pre aplikáciu je vždy zvýraznený. Môžeme si všimnúť, že v oboch prípadoch sa prejavila vysoká nevyváženosť datasetu. Snažili sme sa ju zmierniť nadvzorkovaním. Najlepší model vznikol algoritmom náhodného lesa v aplikácii poisťovne *Aegon*. Kritérium hodnotenia bolo maximalizovanie metriky presnosti predikcie zmätenosti.

Tab.1. Prehľad výsledkov modelov v aplikácii *Firotour*. Označenie *1* reprezentuje predikciu zmäteného pozorovania. *LR* znamená lineárna regresia a *RF* náhodný les. Písmeno *S* označuje prípady kde bolo aplikované nadvzorkovanie minoritnej triedy a je vždy nasledované percentami.

Model	Presnosť (0)	Presnosť (1)	Úplnosť (0)	Úplnosť (1)
LR	0.98	0.30	0.99	0.11
LR S200	0.98	0.36	0.09	0.09
LR S500	0.98	0.24	0.99	0.11
RF	0.98	0	1	0
RF S200	0.98	0.25	1	0.02
RF S500	0.98	0.27	1	0.07

Tab.2. Prehľad výsledkov modelov v aplikácii *Aegon*. LR znamená lineárna regresia a RF náhodný les. Písmeno S označuje prípady kde bolo aplikované nadvzorkovanie minoritnej triedy a je vždy nasledované percentami.

Model	Presnosť (0)	Presnosť (1)	Úplnosť (0)	Úplnosť (1)
LR	1	0.50	1	0.07
LR S200	0.98	0	1	0
LR S500	1	0.50	1	0.07
RF	1	0	1	0
RF S200	1	0	1	0
RF S500	1	0.50	1	0.12

4 Záver

V tejto práci sme predstavili metódu na predikciu zmätenia založenú na analýze interakčného správania z myši vo webovej aplikácii. Náš výskum nadväzuje na našu predchádzajúcu prácu [4], v ktorej sme predstavili výsledky metódy len nad aplikáciou *Firotour*. Chceme sa však posunúť ďalej, a preto sme zopakovali používateľskú štúdiu nad aplikáciou poisťovne *Aegon*. Ako sme v práci predstavili počiatočné experimenty ukazujú, že metóda dosahuje porovnateľné, ba dokonca lepšie výsledky pri predikcii zmätenia. Otvára sa tu však priestor na aplikáciu nových čŕt odvodených z klávesnice (keďže naše datasety disponujú týmito záznamami). Taktiež sa tu otvára priestor na natrénovanie čo najvšeobecnejšieho modelu. Tu je však otázka, ktoré z vybraných čŕt sú doménovo nezávislé.

Je dôležité dodať, že v dnešnej dobe disponujeme kompletnou infraštruktúrou v rámci projektu *YesElf*, určenou na aplikáciu modelu predikcie v produkčnom prostredí. Stále pracujeme na jej doladení a plánujeme nasadiť naše riešenie. Otvára sa tu priestor na to ako navrhnúť správne overenie v produkčnom prostredí.

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV SK-IL-RD-18-0004, KEGA 028STU-4/2017 a VG 1/0667/18.

Literatúra

1. Chawla, N.W., Bowyer, K.W., Hall, O.W., Kegelmeyer W.P.: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16 (2002), 321–357.
2. Gamboa, H., Fred, A.: A User Authentication Technique Using a Web Interaction Monitoring System. In Iberian Conference on Pattern Recognition and Image Analysis. Springer (2003), 246–254.
3. Gamboa, H., Fred, A.: A behavioral biometric system based on human-computer interaction. In Biometric Technology for Human Identification, Vol. 5404. Int. Society for Optics and Photonics (2004), 381–393.
4. Hucko, M., Gazo, L., Simun, P., Valky, M., Moro, R., Simko, J., Bielikova, M.: YesElf: Personalized Onboarding for Web Applications. In Adjunct Porc. of the 27th Conf. on User Modeling, Adaptation and Personalization (UMAP'19). ACM, New York, NY, USA, 39-44.
5. Lallé, S., Toker, D., Conati, C., Carenini, G.: Prediction of Users' Learning Curves for Adaptation while Using an Information Visualization. In Proc. of the 20th Int. Conf. on Intelligent User Interfaces - IUI '15. ACM Press, New York, New York, USA (2015), 357–368.
6. Thomas, P.: Using interaction data to explain difficulty navigating online. ACM Trans. on the Web 8, 4 (2014), 24.

Metodika pro mapování biomedicínských ontologií

Jana Vataščinová[0000-0001-9656-5564]

Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze
Nám. W. Churchilla 1938/4, 13067 Praha
xvatj00@vse.cz

Abstrakt. Mapování ontologií hraje důležitou roli při integraci různých systémů nebo při propojování dat, která používají různé ontologie. Jednou z předních domén, ve které se ontologie používají a ve které je konstantně generováno velké množství dat, je biomedicína. Cílem dizertační práce je navrhnout metodiku pro mapování biomedicínských ontologií. Stávající projekty a publikace se tímto tématem jako celkem nezabývají. Související projekty poskytují obecnou metodiku pro mapování ontologií nebo se zabývají samotným mapováním biomedicínských ontologií (získáním sady mapování). Biomedicínské ontologie mohou být velmi specifické, díky čemu se naskytuje potřeba pro metodiku pro jejich mapování. K dosažení cíle budou použité nejprve analytické kroky – přehled literatury v oblasti mapování biomedicínských ontologií, použití a evaluace efektivnosti a omezení stávajících metodik, výzkum charakteristik biomedicínských ontologií, evaluace různých mapovacích nástrojů a evaluace kombinovaných výsledků z více mapovacích nástrojů. Po těchto krocích bude následovat samotná formulace metodiky.

Klíčová slova: ontologie, mapování ontologií, biomedicínské ontologie.

1 Úvod

Biomedicína je jednou z oblastí, která v dnešní společnosti hraje významnou roli. Existuje velké množství biomedicínských dat a neustále jsou vytvářena nová data z různých studií, experimentů atd. Příkladem, kde se biomedicínská data používají a vytvářejí, mohou být farmaceutické firmy. Bez kvalitního zpracování dat (kterému rozumí i počítače) může být množství informací ztraceno.

Biomedicína je zároveň jednou z významných oblastí, ve které se používají ontologie. Obsah mnoha biomedicínských ontologií se překrývá [1], co vede k potřebě jejich mapování. Mapování ontologií je hledáním odpovídajících si konceptů v různých ontologiích. Jeho cílem je vyhodnotit podobnost konceptů, vlastností a instancí na základě jejich pojmenování, struktury nebo logické interpretace. [2]

Nalezená mapování mohou být velmi významná. Například, aplikace, která pracuje s ontologií A, bude moct pracovat i s daty popsanými ontologií B za pomoci vztahů (mapování) mezi ontologiemi A a B. Mapování biomedicínských ontologií však může být náročným úkolem. Biomedicínské koncepty mohou mít mnoho synonym, jejich význam může být uměle ohraničen, koncepty se stejnými jmény mohou mít různé významy nebo se význam konceptů částečně překrývá atd.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 147-151.

Praktický příklad využití může být následující. Farmaceutická firma, která provádí různé studie vyvíjených látek, používá pro popis svých dat vlastní interní ontologii. Dále existuje veřejný projekt pro biomedicínská data a experimenty včetně jejich výsledků, který používá jiné ontologie. V zájmu farmaceutické firmy je, aby mohla porovnat výsledky svých studií s danou látkou s výsledky látky ve veřejně zaznamenaných experimentech. Toho může být dosaženo přepisováním dotazů za pomoci mapování ontologií.

Otázkou tedy je, *jak provést proces mapování ontologií? Který algoritmus pro mapování použít? Musí se použít pouze jeden? Jak algoritmus použít?* Když začnou farmaceutické firmy používat ontologie, tyto ontologie mohou být automaticky vygenerované například z Excelových tabulek. Je možné, že budou obsahovat neshody a nebudou dodržovat osvědčené postupy. Pravděpodobně tedy budou v budoucnu nahrazeny jinými ontologiemi a celý proces mapování ontologií tak musí být opakovatelný. Obecně tedy, *jakou strategii zvolit pro mapování ontologií v biomedicíně?*

Cílem disertační práce je poukázat na to, v čem jsou současné metodiky nedostačující, a navrhnout metodiku pro mapování biomedicínských ontologií.

2 Současný stav zkoumané oblasti

Existuje velké množství způsobů, jak může být mapovací úloha provedena (manuálně, automaticky), bylo vyvinuto množství mapovacích nástrojů pro danou oblast (deset nástrojů se zúčastnilo kampaně OAEI 2018 – Large Biomedical Ontologies track, viz dále) a obecná metodika pro mapování ontologií byla navržena. Nejsme si ovšem vědomi existence metodiky pro mapování biomedicínských ontologií. V této části budou různé zdroje pro práci stručně představeny.

Prvním zdrojem je *metodika pro mapování ontologií* od autorů Jérôme Euzenat a Pavel Shvaiko [3]. Tato metodika zahrnuje celý proces mapování ontologií od vymezení potřeb a účelu mapování, přes získání sady mapování a její evaluace, až po samotnou implementaci výsledných mapování. Kroky této metodiky jsou ovšem velmi obecné a neposkytují návod, jak jednotlivé kroky provést pro daný problém.

Dalším zdrojem jsou kampaně *OAEI*[1] (Ontology Alignment Evaluation Initiative) a zprávy z jejich událostí, např. [4]. OAEI je mezinárodní iniciativa pro organizování evaluace nástrojů pro mapování ontologií. OAEI zahrnuje i sady úkolů (tracks) zaměřené na biomedicínské ontologie (Large Biomedical Ontologies track, Anatomy track, Disease and Phenotype track), které představují důležitý zdroj zejména pro hledání vhodného nástroje pro mapování biomedicínských ontologií.

Mezi další zdroje patří článek *Tackling the challenges of matching biomedical ontologies* od autorů Daniel Faria et al [5], který popisuje strategie mapovacích nástrojů pro zdolání výzev mapování biomedicínských ontologií. Mezi výzvy patří velikost biomedicínských ontologií, biomedicína a její komplexní a bohatý slovník nebo rozdílný pohled na doménu, který vede k logicky neslučitelným mapováním kvůli odporujícím si restrikcím.

[1] http://oaei.ontologymatching.org/

Existuje dále řada metod, jak provést mapování biomedicínských ontologií ve smyslu získání sady mapování pro dané vstupní ontologie (část z celého procesu), např. [6, 7]. Další zdroje popisují konkrétní mapovací nástroje [8, 9], technologie nebo projekty [10] (projekt pro sbírání biomedicínských dat).

3 Cíl práce

Hlavním cílem práce je navrhnout metodiku pro mapování biomedicínských ontologií. Metodika by měla zahrnovat celý proces mapování ontologií, počínaje identifikací ontologií pro mapování a charakteristikou potřeb, konče se samotnou implementací výsledných mapování. Metodika bude založena na existující metodice pro mapování ontologií (viz [3]), která bude dále specializovaná pro mapování biomedicínských ontologií (s ohledem na jejich potřeby a charakteristiky).

Hlavními otázkami, které by měly být zodpovězené, jsou:

- *Proč jsou obecné metodiky pro mapování ontologií nedostatečné pro mapování biomedicínských ontologií?*
- *Jaké jsou charakteristiky biomedicínských ontologií a jak ovlivňují proces mapování ontologií?*
- *Které nástroje a techniky jsou vhodné pro mapování biomedicínských ontologií – jak vybrat vhodný nástroj? Jak nástroje kombinovat?*

Metodika by měla představovat návod pro ty, kdo používají biomedicínské ontologie a mají potřebu jejich mapování.

4 Metody a postup práce

Pro dosažení cílu práce, následující metody budou použity:

- Přehled literatury v oblasti mapování biomedicínských ontologií,
- Použití současných metodik a evaluace jejich efektivnosti a omezení,
- Průzkum charakteristik biomedicínských ontologií,
- Evaluace mapovacích nástrojů, evaluace kombinovaných výsledků z více mapovacích nástrojů (OAEI),
- Vývoj metodiky,
- Evaluace metodiky.

5 Předběžné výsledky

V současnosti je možné uvést pouze několik základních charakteristik biomedicínských ontologií, částečné výsledky probíhající analýzy výsledků z OAEI – Large Biomedical Ontologies track a vztah mezi účelem mapování s přístupem k získání sady mapování.

Co se týče charakteristik biomedicínských ontologií, následující tři charakteristiky byly zatím zaregistrované:

- *Značná velikost biomedicínských ontologií*. Tyto ontologie mohou obsahovat desítky a stovky tisíc tříd, nehledě na vlastnosti a axiomy.

Taková velikost hraje rozhodující roli při rozhodování mezi manuálním a automatickým mapováním, kdy manuální mapování není u takového rozsahu možné.

- *Názvy konceptů v kódech*. Pro biomedicínské ontologie je běžné, že jako názvy konceptů jsou použity různé kódy specifické pro jednotlivé ontologie. Mapovací nástroj tedy musí umět pracovat s labely konceptů.
- *Taxonomický charakter biomedicínských ontologií*. U biomedicínských ontologií je časté, že jsou reprezentovány ve formě taxonomií. Příkladem může být i Large Biomedical Ontologies track z OAEI, kde všechny tři použité ontologie mají taxonomický charakter (např. NCIT[2] definuje 145 810 tříd a pouze 97 vlastností). U biomedicínských pojmů je ovšem obtížné jejich zařazení do hierarchie. Tyto hierarchie jsou tedy často uměle vytvořené a v různých ontologiích se liší. Mapovací nástroje jsou tedy odkázané hlavně na lexikální mapování.

Tyto charakteristiky potvrzuje i dosavadní analýza výsledků OAEI Large Biomedical Ontologies tracku, kdy mapování z různých nástrojů jsou založena na výskytu stejných nebo synonymních slov v názvech konceptů. Zároveň jsou použité ontologie velmi rozsáhlé a všechny taxonomického charakteru.

Cíl mapování je výchozím kritériem, které je třeba brát v úvahu. Ukázkou může být úvodní příklad, kdy cílem je nalézt všechny výskyty dané látky pro přepsání dotazu, kdy není potřeba, aby namapované koncepty byly logicky koherentní. Není tedy nutné, aby byly použity mapovací nástroje, které kontrolují logickou koherentnost.

6 Závěr

Návrh metodiky pro mapování biomedicínských ontologií není jednoduchým úkolem. Měla by být ovšem přínosem pro pracoviště, kde se vyskytuje potřeba mapování biomedicínských ontologií a kde mapování mohou umožnit sdílení dat, hledání v datech atd.

Poděkování: Tento příspěvek částečně vznikl s podporou projektu IGA VSE 33/2019.

Literatura

1. Gross, A., Hartung, M., Kirsten, T., Rahm, E.: Mapping Composition for MatchingLarge Life Science Ontologies. In: Proceedings of the Second International Conference on Biomedical Ontology, pp. 109–116. Buffalo, NY (2011)
2. Staab, S., Studer, R.: Handbook on Ontologies. 2nd edn. Springer, Berlin (2009)
3. Euzenat, J., Shvaiko, P.: Ontology Matching. 2nd edn. Springer, Berlin (2013)
4. Euzenat, J., Shvaiko, P.: Ontology Matching. 2nd edn. Springer, Berlin (2013) 6. Algergawy, A., Cheatham, M., Faria, D., et al.: Results of the Ontology Alignment

[2] https://cbiit.cancer.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/terminology/

Evaluation Initiative 2018. In: Ontology Matching OM2018, pp. 76–116. CEUR-WS, Ca´chy (2018)
5. Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F.M.: Tackling the Challenges of Matching Biomedical Ontologies. Journal of Biomedical Semantics 9(4) (2018)
6. Zaveri, A., Dumontier, M.: Ontology Mapping for Life Science Linked Data. In: Proceedings of the First International Workshop on Biomedical Data Integration and Discovery. CEUR-WS, Japan (2016)
7. Sarasua, C., Simperl, E., Noy, N.: CROWDMAP: Crowdsourcing Ontology Alignment with Microtasks. Lecture Notes in Computer Science 7649, 525541 (2012)
8. Jiménez-Ruiz, E., Grau, B.C., Zhou, Y., Horrocks, I.: Large-scale interactive ontology matching: Algorithms and implementation. Proceedings of the 20th European Conference on Artificial Intelligence, 444–449 (2012)
9. Faria, D., Pesquita, C., Santos, E. et al.: The AgreementMakerLight Ontology Matching System. OTM Confederated International Conferences On the Move to Meaningful Internet Systems, 527–541 (2013)
10. PubChem, http://pubchemdocs.ncbi.nlm.nih.gov/rdf. Last accessed 10 February 2019

Knowledge-based anomaly detection

Matej Kloska, Viera Rozinajová

Fakulta informatiky a informačných technológií STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
{xkloskam,viera.rozinajova}@stuba.sk

Abstract. The paper presents our research ideas in the initial stage of the project focusing on the problem of anomaly detection. The main goal of our work is to propose an approach for effective anomaly detection in data streams, specifically for group anomaly detection. The proposed approach encompasses two aspects of the solution. The first is the method and representation based on detection of transitions in data streams between normal and anomalous state. The second one is the support of the bidirectional knowledge transfer between the defined model and the expert who evaluates the model. Initial verification of the proposed change detection approach allows us to assume promising results and provides the basis for further research in this area.

Keywords: anomaly detection, group anomaly, human-in-the-loop

1 Introduction

The quality of decision-making is closely related to the quality of data we have at the input of the decision-making process [1]. Data may contain noise that may be encoded into data naturally or targeted. Extreme variant of noise are anomalies that do not belong to normal data due to their nature. In case the proposed processing method fails to correctly deal with extremes and noise in data, the processing of new data may be, but not always, negatively affected by already processed data extremes. The extremes in data are interesting for us because they represent incorrect, mysterious behaviors in the observed environment, which must be treated with proper consequences. Alternatively, they are gripping for us as they affect accuracy or efficiency of the computational models that work with the data.

Detection of anomalies is the task, which could be found in many real-life scenarios. Extremes can be found in various forms with respect to the applied area. Among the most common areas where we can find the detection, we can certainly involve systems and communication networks security, sustainability of industrial systems, finance and insurance, natural language processing and intelligent audiovisual signal processing.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 152-157.

2 Existing methods

Research on anomaly detection goes back to the very beginning of the data mining research, as these study fields support each other. There are many data mining methods which could be applied for anomaly detection. These approaches include classification, clustering, graphs analysis, statistical methods and many others.

2.1 Classification methods

Subba et al. [2] proposed simple Artificial Neural Network (ANN) model that may be easily used to classify any attempts to break into the monitored system. The proposed neural network uses one hidden layer, making it computationally inexpensive and thus able to be deployed in real-time evaluation which is crucial in intrusion detection systems (IDS).

Another approach for intrusion detection was proposed by Xiao et al. [3]. They tried to utilize Bayesian network as classifier with powerful reasoning capabilities. To overcome the issues of training suboptimal models based on heuristics and incompleteness of datasets they propose Bayesian Network Model Averaging based on k-best Bayesian Network classifiers.

Perdisci et. al [3] exploit one-class SVM classifier for hardening payload-based IDS. The differentiation of the other IDS classifiers is in the learning approach as the proposed method uses unsupervised learning. The method applies intelligent feature clustering algorithm originally proposed for text classification problems for dimensional reduction of the feature space. Ensemble of multiple one-class SVM classifiers for different features from obtained payload according to evaluation improves model accuracy and the hardness attacker's evasion.

2.2 Clustering methods

Leung and Leckie [9] proposed method for unsupervised IDS detection based on clustering. The key benefit of this methods is in unsupervised learning. Proposed method is capable of detecting previously unseen attacks. According to evaluation using KDD Cup 1999 dataset, the methods is near the accuracy of other IDS methods, but with the mentioned advantage of unsupervised learning and smaller computational complexity.

Another research [18] in clustering-based anomaly detection focuses on network traffic anomalies. Approach is novel by means of flow-based detection scheme based on k-means clustering algorithm. Identified cluster centroids from k-means are used in efficient distance-based detection of anomalies in new incoming data.

2.3 Graph analysis-based methods

Calderara et al. [10] applied graph analysis for detecting anomalies in people's trajectories. The approach was based on combination of two state-of-the-art methodologies. The first was represented by an ability to track and label single person trajectories in a crowded area acquired by multiple video cameras. The second was an

application of novelty unsupervised detection algorithms based on spectral analysis of graphs. The work as a proof-of-concept calls for extensions and further work.

Jiang et. al [20] tried to solve suspicious network activities through DNS failure graph analysis. Known methods for network activities analysis work on network traffic data. The analysis is due to a big amount of data used too expensive and reduces effectiveness of these methods. The method is based on analysis of failed DNS queries which form DNS failure domains (graphs). They apply tri-nonnegative matrix factorization technique to iteratively extract coherent co-clusters from DNS failure graphs. Those co-clusters represent a variety of anomalous activities e.g., spamming, bots etc. The method was evaluated on 3-month DNS traces. The results show that 4 unknown groups of hosts were identified in addition to 4 correctly identified already known suspicious groups of hosts.

2.4 Summary

Research in anomaly detection is still active as there are many studies [11,12] that address current anomaly detection issues and have been published in recent years. Neural networks-based methods are used more often than statistical methods. The promotion of neural networks is in the constantly increasing computational capacity, which has limited the full use of its potential recently. We personally hold the view that other methods, not based on neural networks, have their own justification and motivation of use nowadays. One of them is, that the decisions based neural network are hard to be explained if possible. Interpretability is crucial in many areas of everyday file where anomaly detection is necessary. What we see as natural ways of anomaly detection by means of data mining methods are classification and clustering methods. Those methods generally fulfill transparent interpretability constraint and therefore are suitable for our use case. The examples are decision trees, kNN and SVM.

3 Research Goals

One of numerous problems of anomaly detection is frequent false anomalous state notification of the observed system. This issue was discussed in several works [16,17,18]. The higher the volume of false notifications the higher the unnecessary intervention of the operator / expert monitoring the system. Perfect anomaly detection method should notify the user only in case when real anomaly happens and on the other hand should do not skip any single anomaly occurrence. The indirect relationship between high detection rate and low volume of false notifications is one of the key indicators of the quality of the proposed method. We identified challenges considering correct detection of transitions between normal and anomalous state of the observed system. Those challenges make up goals of our further research:

- capturing the patterns of transition in the behavior of the observed system,
- bidirectional knowledge transfer.

Methods primarily try to correctly detect the anomalous state of a particular data instance, possibly by means of rules to define certain groupings of instances creating anomalies. What we see as an opportunity for improvement is the training of models to capture the patterns of observed system behavior. This means more accurately

describing the transitions between the anomalous and the normal state over the data stream. The task would be to construct a method that will automate deduction of transition motifs, categorize them and then use them for a new detection. The result of the categorization with appropriate post-processing method would be a set of transition patterns. In our case, detection would consist of calculating the anomalous score of a particular instance in the observed sequence and then estimating the global state (degree of anomaly) of the system. There are works which focus on particular issues within proposed method such as data stream processing annotation [14] or detection of motifs in data stream [15].

The first subproblem in this task is to effectively represent transition states of the observed system. Nowadays, we work with large volumes of data that do not usually fit into computer's operating memory. One option is to store sampled data, or the other is to store data on slower mass media that are not suitable for real-time processing. Effective representation of data that is not sampled, accessible in the computer's operating memory and directly usable (without any compression) by the detection method can greatly expand the possibilities of using unsupported, rapidly available data in the context of anomaly detection.

Solving the problem of capturing and representing transition states, we come to the problem of how to exploit potentially very long sequences of system state transitions in order to estimate the current state. The challenge is to effectively identify the motifs of normal / anomalous variable length behavior and evaluation how the sequences contribute to the current state of development.

The main goal is to develop methods that can automatically exploit all available data and improve over time without any human intervention. However, there are cases, when human intervention could remove automatic approach related struggles. In this case so called "human-in-the-loop" interactive machine learning process appears to be more suitable. [13]

The anomaly detection process requires in certain stages the intervention of a domain expert who introduces his knowledge into the decision-making process. Most available methods suffer from a lack of deducibility why the method labeled a specific instance or a sequence of the instances as normal or anomalous. This issue is also known as "black-box machine learning problem" [13]. The transfer of knowledge between the method and the expert should become bi-directional in order to achieve the most accurate detection results, thus shifting from "black-box problem" to wished "glass-box problem" [13]. There are two types of knowledge which would be transferred between the model and expert. The model would provide knowledge in form of rules describing the motifs such as transition rules (eg. increase during 2 epochs for 4 points, steady for 5 epochs, decrease during 5 epochs for 6 points, optionally whole in context of another motif A and B). The expert could provide own knowledge in the same manner as the model along with setting hard boundaries such as minimum, maximum or average value over time for any attribute of any single observed instance. Given the first identified challenge, we identified an opportunity for automated deduction of rules for behavioral motifs. Knowledge transfer from the method to the expert would greatly facilitate further training of the model and modeling the system for detection as such while the transfer model to export would help in reasoning about the detection process when model is deployed in testing or production environment.

4 Conclusions

This paper presents our research ideas in the initial stage of our project focusing on the problem of anomaly detection. As stated in the chapter related to existing methods, we have identified many approaches in anomaly detection for various areas, especially but not exclusively in IDS detection. What we identified as key issue of any anomaly detection methods is the interpretability of the model. The ability to explain reasoning behind the model is crucial in many industries like finances, healthcare, medicine and security. Solving the interpretability is closely related to understanding the behavioral changes inside the monitored system. Those challenges make up goals of our further research:

- *capturing the patterns of transition in the behavior of the observed system*: detection of transition motifs and their grouping into patterns,
- *bidirectional knowledge transfer*: transfer of knowledge between trained model (extraction of rules describing motifs) and expert (definition of rules for motifs; definition of hard constraints for any instance attribute).

We hold the view that solving the mentioned challenges could lead to more precise anomaly detection. By capturing the patterns of transition in the observed system, we pursue the capability of gathering the data for interpretability of reasoning behind the detection. We assume, that transitions in data represent the key issue for effective observation of the system and gathering the knowledge. Knowledge without the possibility of deriving an action has no added value. Therefore, we suppose that the logical second step is the support of knowledge transfer between the model and the expert. The ability of knowledge transfer from model to expert in form of motif rules brings more transparency in reasoning about the correctness of detection for the expert and possibly trust in automation solution. On the other hand, easy way for definition of rules for detection would boost the detection as the expert could bring the initial set of domain-dependent knowledge.

References

1. B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer, “Data quality for data science, predictive analytics, and big data in supply chain management: An intro- duction to the problem and suggestions for research and applications,” International Journal of Production Economics, vol. 154, pp. 72–80, 2014.
2. B. Subba, S. Biswas, and S. Karmakar, “A neural network based system for intrusion detection and attack classification,” in Communication (NCC), 2016 Twenty Second National Conference on, pp. 1–6, IEEE, 2016.
3. L. Xiao, Y. Chen, and C. K. Chang, “Bayesian model averaging of bayesian network classifiers for intrusion detection,” in Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International, pp. 128–133, IEEE, 2014.
4. R. Perdisci, G. Gu, and W. Lee, “Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems,” in Data Mining, 2006. ICDM’06. Sixth International Conference on, pp. 488–498, IEEE, 2006.

5. G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with ga-based feature selection," in Proceedings of the 43rd annual Southeast regional conference-Volume 2, pp. 136–141, ACM, 2005.
6. Y. K. Kim, S. Y. Lee, S. Seo, and K. M. Lee, "Fuzzy logic-based outlier detection for bio-medical data," in Fuzzy Theory and Its Applications (iFUZZY), 2014 International Conference on, pp. 117–121, IEEE, 2014.
7. W. Li, "Using genetic algorithm for network intrusion detection," Proceedings of the United States Department of Energy Cyber Security Group, vol. 1, pp. 1–8, 2004.
8. R. Chitrakar and C. Huang, "Anomaly based intrusion detection using hybrid learn- ing approach of combining k-medoids clustering and naive bayes classification," in Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on, pp. 1–5, IEEE, 2012.
9. LEUNG, Kingsly; LECKIE, Christopher. Unsupervised anomaly detection in network intrusion detection using clusters. In: *Proceedings of the Twenty-eighth Australasian conference on Computer Science-V*olume 38. Australian Computer Society, Inc., 2005. p. 333-342.
10. CALDERARA, Simone, et al. Detecting anomalies in people's trajectories using spectral graph analysis. Computer Vision and Image Understanding, 2011, 115.8: 1099-1111.
11. C. C. Aggarwal, "Outlier analysis," in Data mining, pp. 237–263, Springer, 2015.
12. K. Anand, J. Kumar, and K. Anand, "Anomaly detection in online social network: A survey," in Inventive Communication and Computational Technologies (ICICCT), 2017 International Conference on, pp. 456–459, IEEE, 2017.
13. A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?," Brain Informatics, vol. 3, no. 2, pp. 119–131, 2016.
14. S. Kolozali, M. Bermudez-Edo, D. Puschmann, F. Ganz, and P. Barnaghi, "A knowledge-based approach for real-time iot data stream annotation and processing," in Internet of Things (iThings), 2014 IEEE International Conference on, and Green Computing and Communications (GreenCom), IEEE and Cyber, Physical and Social Computing (CPSCom), IEEE, pp. 215–222, IEEE, 2014.
15. M. Shokoohi-Yekta, Y. Chen, B. Campana, B. Hu, J. Zakaria, and E. Keogh, "Discov- ery of meaningful rules in time series," in Proceedings of the 21th ACM SIGKDD in- ternational conference on knowledge discovery and data mining, pp. 1085–1094, ACM, 2015.
16. V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM computing surveys (CSUR), vol. 41, no. 3, p. 15, 2009.
17. S. Sadik and L. Gruenwald, "Research issues in outlier detection for data streams," ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 33–40, 2014.
18. R. Sekar, A. Gupta, J. Frullo, T. Shanbhag, A. Tiwari, H. Yang, and S. Zhou, "Specification-based anomaly detection: a new approach for detecting network intrusions," in Proceedings of the 9th ACM conference on Computer and communications security, pp. 265–274, ACM, 2002.
19. MÜNZ, Gerhard; LI, Sa; CARLE, Georg. Traffic anomaly detection using k-means clustering. In: GI/ITG Workshop MMBnet. 2007. p. 13-14.
20. JIANG, Nan, et al. Identifying suspicious activities through dns failure graph analysis. In: The 18th IEEE International Conference on Network Protocols. IEEE, 2010. p. 144-153.

Aplikácia zložených metód strojového učenia na nevyvážené dátové sady: predikcia bankrotu spoločností

PeterGnip[1], Peter Drotár[1]

[1]Katedra počítačov a informatiky, FEI TU v Košiciach
Letná 9, 042 00 Košice
{peter.gnip,peter.drotar}@tuke.sk

Abstrakt. Aplikácia metód strojového učenia na vysoko nevyvážené dátové sady je v súčasnosti považovaná za náročnú úlohu v oblasti spracovania dát. Spracovanie nevyvážených dátových sád je často spájané s nevyváženým učením, ktoré nachádza svoje uplatnenie v mnohých oblastiach reálneho života, kde zaradzujeme aj problematiku predikcie bankrotu spoločností. Tento príspevok je zameraný na porovnanie výkonnosti niekoľkých zložených (ensemble) metód strojového učenia aplikovaných na dátach spoločností s ručením obmedzeným pôsobiacich na Slovensku. Presnosť natrénovaných modelov bola hodnotená metrikou geometrického priemeru, ktorá v niektorých prípadoch dosiahla pomerne vysokú úspešnosť až nad 97%. Dosiahnuté výsledky boli validované na dvoch rozdielnych odvetviach hospodárstva – poľnohospodárstvo a maloobchody.

Kľúčové slová: detekcia anomálií, zložené metódy, predikcia bankrotu, nevyvážené učenie

1 Úvod

Ľubovoľná spoločnosť sa môže vplyvom nepriaznivej finančnej situácie, nevhodným obchodovaním alebo nesprávnym vedením podniku dostať do značných ťažkostí, ktoré môžu v konečnom dôsledku vyústiť až do samotného bankrotu spoločnosti. Včasná informácia o nastávajúcej kritickej situácií môže zohrávať kľúčovú rolu pri procese rozhodovania sa o ďalšom smerovaní spoločnosti. Dôležitosť tejto problematiky potvrdzuje aj nespočetné množstvo príspevkov, ktoré sú zhrnuté v nasledujúcich prehľadových článkoch [1], [3].

Predikcia bankrotu spoločností je považovaná za problém výrazne nevyváženej dátovej sady, nakoľko počet solventných spoločností v praxi vo výraznej miere prevyšuje počet bankrotujúcich podnikov. Na prekonanie slabej výkonnosti klasifikátorov boli vytvorené sofistikovanejšie techniky. Jedným z najpopulárnejších prístupov je využitie metód nevyváženého učenia (*imbalanced learning*). Prehľadový článok [3] sumarizuje 527 príspevkov zameraných na túto problematiku. Najlepšie výsledky boli zaznamenané pri použití techniky vzorkovania, boostovania, detekcie anomálií a taktiež zložených metód. V tomto príspevku sa zameriavame na porovnanie

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 158-161.

výkonnosti zložených metód strojového učenia pri snahe predikovať bankrot spoločností.

2 Dátová sada

Dátová sada bola tvorená tisíckami účtovných závierok spoločností s ručením obmedzeným pôsobiacich na Slovensku počas obdobia rokov 2010-2016 v rozdielnych odvetviach hospodárstva. Tento príspevok a výhradne zameriava na použitie dát z odvetvia poľnohospodárstva a maloobchodov. Každá účtovná závierka bola reprezentovaná 20 pomerovými finančnými ukazovateľmi, ktorých bližšia charakteristika je uvedená v našom predchádzajúcom príspevku [11].

Predložená dátová sada obsahovala záznamy spoločností pre štyri roky evaluácie: 2013, 2014, 2015 a 2016. Pod pojmom rok evaluácie rozumieme obdobie, v ktorom bola spoločnosť vyhodnotená ako bankrotujúca alebo solventná. Pri trénovaní klasifikačných modelov boli použité dáta z obdobia jedného roka pred samotným rokom evaluácie, ktoré v našej predošlej štúdií [11] dosahovali pomerne uspokojivé výsledky. Distribúcia vzoriek bankrotujúcich a nebankrotujúcich spoločností je znázornená v Tabuľke 1. Pomer vzoriek nebankrotujúcich spoločností vo výraznej miere prevyšuje počet bankrotujúcich vzoriek, čo reprezentuje problém výrazne nevyváženej dátovej sady.

Tab. 1. Pomer bankrotujúcich a nebankrotujúcich vzoriek

	2013	2014	2015	2016
poľnohospodárstvo	6/1442	6/1622	8/1882	8/1991
maloobchody	11/5195	11/6107	7/6327	4/6263

3 Metodológia

V tejto štúdií bolo použitých niekoľko zložených metód strojového učenia, ktorých cieľom je kombinovať niekoľko slabších klasifikátorov za účelom zvýšenia presnosti predikcie výsledného modelu. Bolo použitých niekoľko základných zložených metód: *AdaBoost* (*AB*) [8], *Random Forest* (*RF*) [6] a *Gradient Boosting* (*GB*). V práci boli použité aj zložené metódy využívajúce techniku vzorkovania: *Balanced Bagging* (*BB*) [2], *Easy Ensemble* (*EE*), *Balanced Random Forest* (*BRF*) a *RUSBoost* (*RUB*) [10]. Pre porovnanie boli aplikované dve metódy na detekciu anomálií: metóda podporných vektorov s natrénovaním na jednej triede (*OCS*) [9] a izolačný les (*IF*) [7] a taktiež jednoduchá metóda podporných vektorov (*SVC*) [5].

Pred samotným použitím metód strojového učenia bolo potrebné dáta vhodne spracovať. Chýbajúce hodnoty boli nahradené priemernou hodnotou príslušného finančného ukazovateľa. Dáta boli štandardizované na nulovú strednú hodnotu a jednotkový rozptyl. Na trénovanie modelov bolo použitých 80% dát. Testovanie modelov bolo realizované na zvyšných 20% dát, pričom vzorky oboch tried boli rovnomerne zastúpené v trénovacej aj testovacej časti. Výnimku tvorili metódy na

detekciu anomálií, pri ktorých boli na natrénovanie použité len vzorky majoritnej triedy. Experimentálne pokusy sme realizovali 100-krát. Úspešnosť natrénovaných modelov sme hodnotili metrikou geometrického priemeru (*GM*), ktorá je považovaná za vhodnú metriku pre nevyvážené dátové sady [4], pretože berie do úvahy úspešnosť predikcie oboch tried binárneho klasifikačného problému. Výsledky jednotlivých experimentov boli spriemernené a ich dôveryhodnosť vyjadrená hodnotou štandardnej odchýlky.

4 Dosiahnuté výsledky

Hodnoty metriky *GM* najlepšie natrénovaných klasifikačných modelov pre predikciu bankrotov s príslušnou štandardnou odchýlkou sú znázornené v Tabuľke 2. Pri oboch odvetviach hospodárstva boli najlepšie výsledky dosiahnuté na dátach pre evaluačné roky 2015 a 2016, pri ktorých zložené metódy založené na technike vzorkovania (*RUB, EE, BRF* a *BB*) dosahovali hodnoty metriky *GM* od 82,1% až do 99,9%. Porovnateľné výsledky boli zaznamenané pri použití metód na detekciu anomálií (*OCS* a *IF*). Z pomedzi všetkých použitých metód môžeme vyzdvihnúť klasifikátor *RUB*, ktorý je založený na použití podvzorkovania majoritnej triedy v kombinácií s technikou „*boostovania*“, ktorý dosiahol pri všetkých rokoch evaluácie pomerne uspokojivé výsledky.

Tab. 2. Hodnoty geometrického priemeru pre použité metódy

	poľnohospodárstvo				Maloobchody			
	2013	**2014**	**2015**	**2016**	**2013**	**2014**	**2015**	**2016**
AB	40±49	70±46	88.7±26	**100±1**	9,2±24	15,1±30	11±32	25±44
RF	18±39	26±44	92,8±16	**100±1**	2,8±14	22,8±35	6±24	23±43
GB	24,3±38	24,5±40	89,5±19	**100±1**	28±37	23,2±36	28±45	22±42
SVC	**83,2±13**	70,6±21	**99,7±1**	99,9±1	59,4±35	36±36	74±42	30±46
OCS	57,9±9	61±3	95±1	94,8±1	62,3±1	69,9±4	83,1±4	82,1±1
IF	68,2±3	64,6±5	94,5±2	97,6±1	75,6±1	80,1±2	94,9±1	**96,8±3**
EE	77,5±29	93,3±2	99,6±1	99,9±1	81,1±18	79,3±22	96,8±1	95,2±3
BRF	74,4±28	92,9±2	94,2±11	98,3±1	**84±14**	79±19	96,5±1	86,6±29
BB	75,7±28	92,4±14	99,4±1	98,5±1	82±18	78,5±21	94,4±17	82,1±32
RUB	78,7±21	**94,4±2**	99,5±1	99,9±1	81,8±15	**81,9±11**	**97,4±1**	95,9±2

5 Záver

V tomto príspevku sme analyzovali úspešnosť rozdielnych zložených metód strojového učenia pri ich aplikácií na výrazne nevyváženú dátovú sadu pozostávajúcu z účtovných

závierok spoločností s ručením obmedzeným pôsobiacich v rôznych odvetviach hospodárstva na Slovensku. Dosiahnuté výsledky boli porovnané s metódami zameranými na detekciu anomálií a taktiež metódou podporných vektorov používajúcou váhovanie. Najlepšie výsledky boli pozorované pri použití klasifikátora *RUB*, ktorý dosahoval hodnoty metriky geometrického priemeru v rozsahu od 78,7% až do takmer 100%. Porovnateľné výsledky priniesli zložené metódy využívajúce techniku vzorkovania v kombinácií s technikou „*boostovania*" a „*baggingu*". Aplikácia metód na detekciu anomálií dosiahla porovnateľné výsledky, konkrétne metóda *IF* dosahovala hodnoty geometrického priemeru v rozsahu 68,2% - 97,6%.

Poďakovanie: Tento príspevok bol podporený Agentúrou na podporu výskumu a vývoja projektom číslo APVV-16-0211.

Literatúra

1. Alaka, H.E., et al.: Systematic review of bankruptcy prediction models: Towards a framework for tool selection. Expert Systems with Applications (2018), 164-184.
2. Bühlmann, P., Yu, B.: Analyzing Bagging. The Annals of Statistics (2002), 927-961.
3. Haixiang, G., et al.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications (2017), 220-239.
4. Helal, M.A., Haydar, M.S., Mostafa S.A.M.: Algorithms efficiency measurement on imbalanced data using geometric mean and cross validation. International Workshop on Computational Intelligence (IWCI) (2016).
5. Hsu, C.-W., Chang, C.-C., Lin , C.-J.: A Practical Guide to Support Vector Classification. (2010), 1-16.
6. Chen, Ch., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data. University of California, Berkeley (2004), 1-12.
7. Liu, F.T., Ting, K.M., Zhou, Z.-H.: Isolation Forest. Eighth IEEE International Conference on Data Mining (2008).
8. Schapire, R.E.: A Brief Introduction to Boosting. Sixteenth International Joint Conference on Artificial Intelligence (1999), 1401-1406.
9. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the Support of a High-Dimensional Distribution. Neural Computation (2001), 1443-1471.
10. Seiffert, C., Khoshgoftaar, T.M., Hulse, J.V., Napolitano, A.: RUSBoost: a Hybrid Approach to Alleviating Class Imbalance. In: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans (2010), 185-197.
11. Zoričák, M., Gnip, P., Drotár, P., Gazda, J.: Bankruptcy prediction for small – and medium-sized companies using severely imbalanced datasets. Economic Modeling (2019).

Prieskumná analýza prezidentských volieb 2016

Igor Stupavský, Daniela Chudá

Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
igor.stupavsky@stuba.sk
daniela.chuda@stuba.sk

Abstrakt: V dobe elektronickej komunikácie je ťažké zistiť pravdivosť takto získanej informácie. Komunikácia prebieha cez internetové stránky, diskusie, sociálne siete a podobne. Pomocou rôznych tipov antisociálneho správanie, ako napríklad fake news ich šíritelia chcú získať rôzne výhody, alebo presvedčiť prijímateľa o „svojej pravde“. Teoretická časť práce je zameraná na rôzne prístupy k detekcii antisociálneho správania. V experimentálnej časti tejto práci sa budeme zaoberať prieskumnou analýzou datasetu „Electionday tweets“, ktorý obsahuje značené správy z prezidentskej kampane v USA.

Kľúčové slová: antisociálne správanie, fake news, real news

1 Úvod

Antisociálne správanie je široko koncipovaný medzi-odborový problém. S veľkým presahom do oblasti informačných technológii. Veľká časť súčasnej komunikácie prebieha pomocou sociálnych sietí typu Facebook, Twitter, Instagram a pod. Každá technológia okrem jej využitia určitým spôsobom mení aj pohľad jej užívateľov na okolitý svet. V prípade sociálnych sietí menia tieto spôsoby samotný spôsob medziľudskej komunikácie. Pomerne často sa v rodinnom prostredí stáva bežným zvykom, že komunikácia prechádza iba krátkymi správami a dlhšie rozhovory v rodine predstavujú ojedinelý fenomén.
Z vedeckého hľadiska je antisociálne správanie definované odlišne u každého autora. Napríklad [1] ho definoval *„ako šírenie dezinformácií a reakcie antisociálnych používateľov“*. Pre potreby nášho výskumu však antisociálne správanie chápeme v širšom kontexte ako len šírenie dezinformácii a reakcií. Iný autori [2] chápu *„antisociálne správanie (antisocial behaviour) v podobe relatívne širokého spektra veku neprimeraných činov a postojov, ktoré narúšajú očakávania rodiny, sociálne normy, osobné a majetkové práva ostatných, bude pretrvávať aj v neskorších vývinových obdobiach. Zároveň sa jedná o jeden z typických znakov obdobia adolescencie, pre ktoré je charakteristický nárast výskytu prejavov externalizovanej i internalizovanej psychopatológie.“*. Táto definícia je o poznanie širšia a zahŕňa aj oblasť psychológie, s ktorou je táto téma natrvalo spojená.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 162-166.

Pre lepšie pochopenie problému sme sa zamerali na rôzne prístupy k detekcii antisociálneho správania, konkrétne falošných správ a v experimentálnej časti na prieskumný experiment na konkrétnom datasete.

1.1 Prístupy k detekcii falošných správ

V jednotlivých prístupoch k detekcii rôznych foriem antisociálneho správania sa zameriavajú výskumníci prevažne na špecifické vlastnosti jednotlivých typov správania. Vo všeobecnosti však môžeme rozdeliť prístupy do troch základných:

4. Natural language processing (NLP)
5. Fact-checking
6. User Model

Natural language processing (NLP):

Medzi typických zástupcov z tejto oblasti môžeme uviesť napríklad detekciu citovo zafarbených slov, alebo analýzu vetnej štruktúry.

Pri detekcii citovo zafarbených slov je to založené na predpoklade, že falošné správy ako príspevok zaútočia na pocity príjemcu, čím sa táto osoba stane náchylnejšia na to, aby táto osoba verila, že je daná správa pravdivá. Ako príklad môžeme uviesť prácu Mykhailo Granika a Volodymyra Mesyura 2017 [3] na analýze stavu Facebookových statusov, kde dosiahli úspešnosť 75,40% pri odhaľovaní fake news.

Odhalením štruktúry slov a jej jazykových štýlov sa zaoberá druhá veľká časť NLP. Štruktúra o podvodných dokumentoch v rámci internetovej stránky Wikipedia bola riešená v Kumar et al. aj v roku 2016 [4]. Nedávna práca sa zaoberala Wang 2017 [5] z obľasti politických fake news robila výskum na doméne PolitiFact.

Fact-checking:

Témou v tejto oblasti výskumu sa stali udalosti volieb v USA. Zástupcami jednotlivých prácu sú napríklad práce H. Allcott a M. Gentzkow, 2017 [6], na tému „Sociálne médiá a falošné správy vo voľbách v roku 2016“, kde diskutujú o jednotlivých skutočnostiach, ktoré mohli ovplyvniť voličov.

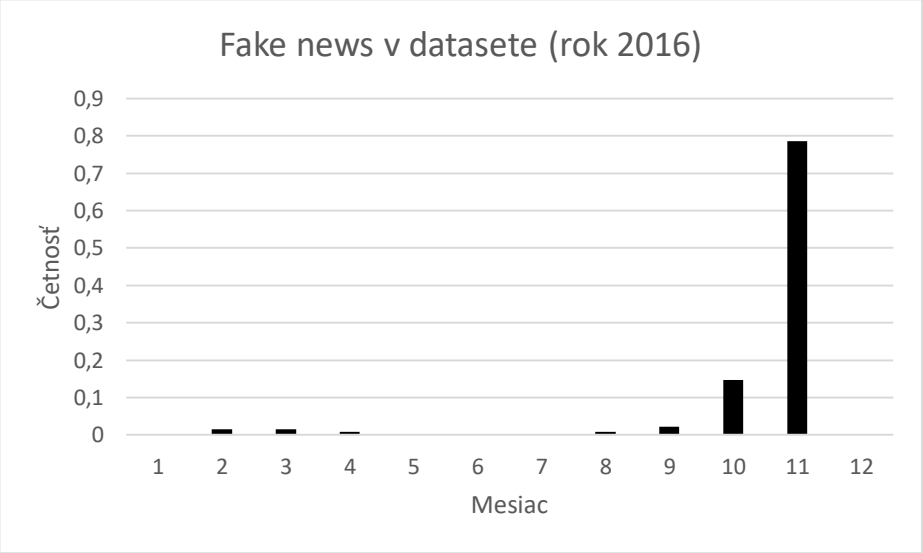
User Model:

Užívateľský model sa používa na detekciu falošných správ porovnaním metadát jednotlivých príspevkov. Tieto meta údaje sú napríklad mená osôb, ktoré uverejnili alebo rozšírili text. Ďalšie meta údaje zahŕňajú dátum účtovania a ďalšie dodatočné údaje. Takáto práca je napríklad práca Zhou et al. 2015 [7], ktorá sa zaoberá vydávaním certifikátov v reálnom čase autorom. Týmto spôsobom si príjemca správy môže byť istý, že to nie je falošná správa.

Touto metódou sme sa v našom výskume vydali aj my v našom prieskumnom experimente, kde sme dataset analyzovali na základe časovej četnosti príspevkov s blížiacim sa termín konania volieb a na základe platformy, ktorá sa na šírenie využívala.

2 Prieskumný experiment na datasete

V rámci nášho výskumu sme analyzovali dataset „Electionday tweets“, ktorý obsahuje 1327 správ zo sociálnej siete Twitter zozbieraných v rokoch 2012 až 2016. Tieto príspevky sa venovali téme prezidentských volieb v USA. Dataset bol značený a obsahoval správy rozdelené podľa doby svojho vzniku.
Analýza odhalila, že príspevky, ktoré boli označené ako falošné vznikli v roku 2016 a to takým spôsobom, že vznikali počas celého roka. Najvýraznejší nárast sa preukázal v období 2 mesiace pred voľbami.



Obr. 1. Normovaná četnosť fake news príspevkov

Analyzovali sme aj zariadenia, ktoré používajú autori fake news na šírenie svojich vymyslených správ. Vychádzame z predpokladu, že sa častejšie používa na šírenie správ osobný počítač s nainštalovanou webovou aplikáciou, alebo webové rozhranie.

Tab. 1. Analýza použitého zariadenia na šírenie

	Fake news (%)	Feal news (%)
App soc. manažmentu	2.21	3.78
Web klient	55.15	35.80
Android	3.68	7.56
Mac	31.62	43.28
Twitter app.	7.35	9.16

Každý príspevok sme podľa použitého spôsobu šírenia zaznamenaného v datasete klasifikovali do jednej z 5 tried: aplikácie sociálneho manažmentu, webový klient, android platforma, mac platforma a Twitter aplikácia.
Ako vidno z analýzy datasetu, ktorý sme mali k dispozícii najčastejšie sa na šírenie fake news v našich údajoch používa webový klient. Tento výsledok vychádza z našej hypotézy. K obdobnému výsledku dospeli aj A. Bovet a H. A. Makse [8].
Podiel požívateľov Mac zariadení sa dá vysvetliť ich obľúbenosťou v rámci USA.

3 Záver

V rámci našej práce sme vychádzali z určitých predpokladov, ktoré sa ukázali v našom datasete ako správne. Analýzou sme zistili, že najväčšia pravdepodobnosť publikovania politických fake news sa predpokladala v období 2 až 3 mesiace pred dátumom konania volieb a najčastejšou formou šírenia je webový klient dostupný na osobných počítačoch, alebo notebookoch.
Je vhodné kombinovať rôzne prístupy k detekcii falošných správ a tak získať čo maximum možných informácií k rozhodovaniu o pravdivosti informácie. Ako doplňujúce informácie môžeme vhodne využiť dostupné meta údaje, ktoré šíriteľ príspevku zanecháva a dajú sa identifikovať. Všetky tieto metódy je vhodné kombinovať napríklad s NLP metódami, ktorým sa chceme venovať v našom ďalšom výskume.

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore Agentúry na podporu výskumu a vývoja v rámci projektu č. APVV-17-0267 - Automatizované rozpoznávanie antisociálneho správania v online komunitách a projektu č. APVV SK-IL-RD-18-0004 - Detekcia dezinformácií v zdravotníctve.

Literatúra

1. Kumar, S., Cheng, J., & Leskovec, J. (2017). Antisocial Behavior on the Web. In *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion* (pp. 947–950). International World Wide Web Conferences Steering
2. Selecká, L., Václavíková, I., Blatný M., Hrdlička, M. (2017). Typológia antisociálneho správania: špecifiká prejavov adolescentných chlapcov a dievčat vo vzťahu k rizikovému sexuálnemu správaniu. *Česká a Slovenská Psychiatrie*, *113*(6), 258–267.
3. Granik, M., Mesyura, V., (2017), "Fake news detection using naïve Bayes classifier", 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON).
4. Kumar, S., West, R., Leskovec, J., (2016), Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In Proceedings of the 25th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, pages 591–602.
5. Wang, W.Y., (2017), "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In Proceedings of the Association for Computational Linguistics Short Papers. Association for Computational Linguistics.

6. Allcott, H., Gentzkow, M., (2017), Social Media and Fake News in the 2016 Election, Journal of Economic Perspectives, num. 2, Volume 31, pag. 211-236.
7. Zhou, X., Cao, J., Jin, Z., Xie, F., Su, Y., Zhang, J., … Cao, X. (2015). Real-Time News Certification System on Sina Weibo, 983–988.
8. Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 1–14.

Optimalizácia využitia elektrickej energie v mikrogride

Miriama Pomffyová[1], Viera Rozinajová[1]

[1]Ústav informatiky, informačných systémov a softvérového inžinierstva, FIIT STU v Bratislave
Ilkovičova 2, 842 16 Bratislava
miriama.pomffyova@stuba.sk, viera.rozinajova@stuba.sk

Abstrakt. Výroba elektrickej energie prostredníctvom obnoviteľných zdrojov je čoraz rozšírenejšia. Rovnako sa stále viac stretávame s uskladňovaním energie pomocou veľkokapacitných batérií. Zapojenie týchto komponentov do siete umožňuje vznik tzv.mikrogridov, teda samostatných komunít odberateľov a producentov elektrickej energie, ktoré poskytujú priestor pre nový spôsob hospodárenia s energiou. Cieľom je efektívne využitie energie vyprodukovanej v rámci mikrogridu a vysoká ekonomická rentabilnosť pre účastníkov tejto mikrosiete. S vybudovaním a efektívnym riadením mikrogridov, ako aj kombináciou s klasickou formou distribúcie energie je spojených viacero problémov. Využitím matematických funkcií a ich ohraničení získame model, ktorý je základom pre optimalizáciu prevádzky mikrogridu. V príspevku analyzujeme a porovnávame rôzne modelové situácie a prístupy k optimalizácii za účelom optimálneho riadenia využívania možností energetických zdrojov s cieľom zvýšenia komfortu používateľov.

Kľúčové slová: viacúčelová optimalizácia, obnoviteľné zdroje energie, evolučné algoritmy

1 Úvod

V oblasti energetiky prevládal centralizovaný spôsob jej distribúcie, ktorý je čoraz častejšie obohacovaný o možnosti výroby elektrickej energie využitím obnoviteľných zdrojov. Pre okamžité vyprodukovanie energie sa využívajú dieselové agregátory, a naopak, pre uskladnenie nadbytočnej energie sa začínajú využívať veľkokapacitné úložiská energie. Práve preto, narastá potreba a význam decentralizácie energetickej sústavy.

Novým komponentom je aktívna spoluúčasť koncových odberateľov na výrobe elektrickej energie (fotovoltaická energia). Aby bola dosiahnutá čo najvyššia možná efektivita distribúcie (výroby a spotreby) elektrickej energie v takejto sieti, je potrebné vyvíjať nové modely riadenia a optimalizácie celej energetickej sústavy, ako aj jej komponentov – mikrogridov. Účastníci mikrogridov, ako producenti elektrickej energie, dokážu z lokálnych obnoviteľných zdrojov energie vyrábať nielen environmentálne prijateľnú, ale zároveň aj ekonomicky udržateľnú elektrickú energiu podľa aktuálnej potreby účastníkov danej lokálnej siete. Sieť mikrogridov by mala byť schopná distribuovať elektrickú energiu bez ohľadu na dispozície hlavnej siete a systém

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 167-172.

decentralizovaného riadenia by mal vedieť na základe optimalizácie ponúknuť kombináciu výkonov oboch sietí.

Na dosiahnutie opísanej nezávislosti je potrebné disponovať modelmi riadenia a optimalizácie oboch sietí s cieľom maximalizácie výroby a efektívnosti distribúcie energie pri súčasnej optimalizácii nákladov, spojených s inštaláciou a prevádzkou zariadení na výrobu elektrickej energie z obnoviteľných zdrojov.

2 Modelovanie procesov v mikrogride

S modelovaním procesov súvisí hľadanie optimálnych riešení problémov [6], súvisiacich s plánovaním – konfigurácie počtu prvkov mikrogridu; prevádzkou – spôsob čerpania energie, optimalizácia nákladov na výrobu elektrickej energie formou znižovania rozdielu medzi množstvom vyrobenej a spotrebovanej elektrickej energie; riadením –zdieľanie energie medzi jednotlivými odberno-odovzdávacími miestami v mikrogride s cieľom zabezpečenia stability siete a s cieľom optimalizácie nákladov používateľa pri dosiahnutí požadovaného komfortu a pri optimálnom plánovaní výroby energie.

Činnosť mikrogridov môžeme modelovať pomocou matematických funkcií a ich ohraničení. Výpočtom maximálnych alebo minimálnych hodnôt stanovenej účelovej funkcie hľadáme optimálne riešenia, ktoré z celej sady možných riešení najviac vyhovujú modelovaným reálnym podmienkam [2].

V projekte sa zameriavame na modelovanie a optimalizáciu vyváženia množstiev spotreby a výroby potrebnej energie. Nami modelované situácie budú riešiť problematiku procesov mikrogridu, ktorý obsahuje fotovoltaické články ako obnoviteľný zdroj elektrickej energie, a úložiská energie.

Hlavným krokom projektu, je definovanie skupiny susedných domov so zdieľaním energie. Následne vyjadríme pre každý dom spotrebu energie z hlavnej siete v čase *i*:

$$g(x_i) = c_i + (x_i - x_{i-1}) - p_i, \tag{1}$$

Kde x_i reprezentuje nabitie batérie v čase i, c_i je spotreba energie v domácnosti v čase i, p_i je vyprodukovaná energia fotovoltaickými článkami v domácnosti v čase i. Kladná hodnota rozdielu nabitia batérie vyjadruje to, že batéria je nabitá. Záporná hodnota rozdielu reprezentuje vybitú batériu.

Pre optimalizáciu problémov sme navrhli nasledovné reprezentácie účelových funkcií:

- Minimalizovanie spotreby energie z hlavnej siete s cieľom maximálneho využitia energie vygenerovanej prostredníctvom fotovoltaiky:

$$minimize\ f(x) = \sum_{i=2}^{N} g(x_i), \tag{2}$$

Kde $f(X)$ je účelová funkcia, vyjadrujúca sumu energie dodanej z hlavnej siete za daný časový interval. Funkciu minimalizujeme za účelom dosiahnutia najnižšej spotreby energie z hlavnej siete v domácnosti. X vyjadruje vektor nabitia batérie v hodinových intervaloch, N je počet hodín, ktorý reprezentuje veľkosť vektora X.

- Minimalizovanie výkyvov spotreby energie z hlavnej siete:

$$minimize\ f(X) = \sum_{i=3}^{N} abs(g(x_i) - g(x_{i-1})), \quad (3)$$

Kde $f(X)$ je účelová funkcia, vyjadrujúca sumu rozdielov energie dodanej z hlavnej siete dvoch po sebe idúcich časových krokoch. Funkciu minimalizujeme za účelom dosiahnutia čo najnižších výkyvov (prudkých poklesov alebo nárastov) hodnôt spotreby energie z hlavnej siete.

- Minimalizovanie ceny za energiu spotrebovanú z hlavnej siete:

$$minimize\ f(X) = \sum_{i=3}^{N} g(x_i) * e_i, \quad (4)$$

Kde e_i je cena energie z hlavnej siete v čase i. Hodnota $f(X)$ je účelová funkcia, vyjadrujúca sumu celkovej ceny za energiu dodanú z hlavnej siete v danom časovom intervale. Funkciu minimalizujeme za účelom dosiahnutia čo najnižšej ceny za energiu dodanú z hlavnej siete.

Na základe tejto optimalizácie formulujeme odporúčania pre:

- rozhodovanie o uskladnení energie pomocou batérií a o jej spotrebovaní na základe aktuálnej ceny energie z hlavnej siete,
- obmedzenie alebo presun spotreby elektrickej energie.

V procese vyhodnocovania výsledkov optimalizácie sa zameriavame hlavne na tieto ukazovatele:

- množstvo energie spotrebovanej z hlavnej siete,
- množstvo energie spotrebovanej z vyprodukovanej energie fotovoltaickými článkami,
- výkyvy a súvislosť priebehu krivky spotreby energie z hlavnej siete,
- celková cena energie spotrebovanej z hlavnej siete.

V nasledujúcej kapitole opisujeme prístupy k riešeniu problémov spojených s optimalizáciou v mikrogride.

3 Optimalizácia prevádzky mikrogridu

Základom pre optimalizáciu simulácie činností mikrogridu je znalosť jeho očakávaného správania sa v budúcnosti, ako napríklad znalosť predpokladanej spotreby odberateľov, očakávanej výrobnej kapacity fotovoltaických panelov, kapacity batérií a predpokladanej doby cyklov nabíjania či vybíjania, a pod.

3.1 Typy prístupov optimalizácie

Najjednoduchším delením prístupov k optimalizácii je delenie na [1, 6]:

- Exaktné prístupy (analytické, matematické) – pri riešení jednoduchých úloh dokážu v krátkom čase nájsť optimálne riešenie. Problémom je potreba hľadania komplexného riešenia, kde čas výpočtu narastá so zložitosťou hľadaného riešenia. Vyskytujú sa aj prípady, kedy týmito analytickými prístupmi nenájdeme optimálne riešenie. Medzi klasické prístupy patrí napríklad lineárne a nelineárne programovanie, dynamické programovanie, či použitie metód, založených na pravidlách (angl. rule-based), a iné.

- Metaheuristické prístupy – dokážu sa prispôsobiť aj problémom s vyššou komplexnosťou. Uprednostňujú nájdenie aspoň nejakého riešenia za kratší čas výpočtu ako exaktné prístupy. Nezaručujú však, že nájdené riešenie bude optimálne. Do tejto skupiny prístupov patria napríklad umelá inteligencia (fuzzy logika, neurónové siete, multi-agentové systémy), prírodou inšpirované algoritmy (genetické algoritmy, algoritmy roja častíc (angl. particle swarm algorithms)), stochastické a robustné programovanie, a iné. Algoritmy sa môžu líšiť napríklad v tom, či pracujú s celou množinou riešení alebo inkrementálne zlepšujú jedno riešenie.

V prípade, že úlohou modelu optimalizácie je riešenie viacerých problémov súčasne, hovoríme o viacúčelovej optimalizácii [4]. Hľadaným riešením chceme dosiahnuť čo najlepšiu stabilitu siete a takú výrobu energie, aby sme minimalizovali objemy prebytočne vyprodukovanej energie. Problémom pri viacúčelovej optimalizácii je to, že to riešenie, ktoré je najúspešnejšie pre jednu účelovú funkciu, nemusí byť najúspešnejšie pre všetky ostatné účelové funkcie. Cieľom je teda nájdenie takého riešenia, ktoré bude viac-menej optimálnym riešením pre všetky účelové funkcie [5].

V našej práci sa chceme venovať najmä metaheuristickým prístupom. Tu možno využiť napríklad prírodou inšpirované algoritmy, ako sú optimalizácia roja častíc, kolóny mravcov či genetické algoritmy. Rýchlosť konvergencie k optimálnym riešeniam je podľa autorov [6] pomocou týchto prístupov vysoká.

3.2 Optimalizácia rojom častíc

Optimalizácia využívajúca roj častíc (angl. Particle Swarm Optimization, abrv. PSO) je založená na tom, že riešenia sú reprezentované ako náhodne vygenerované častice, ktoré majú definovanú svoju polohu, rýchlosť a predchádzajúcu úspešnosť [7]. Tá je reprezentovaná najlepšou dosiahnutou pozíciou (reprezentácia jedinca) a hodnotou fitnes funkcie, ktoré boli v danej polohe dosiahnuté (angl. particle best). Fitnes funkcia vyjadruje úspešnosť riešenia častice (jedinca). V prípade viacúčelovej optimalizácie je súčasťou fitnes funkcie vyhodnotenie úspešností jednotlivých účelových funkcií a úprava ich hodnôt podľa pridelených váh. Vonkajšie ohraničenia modelu optimalizácie sú na sebe nezávislé a zabezpečujú to, aby jedinci nevybehli mimo priestoru riešení. Na druhej strane, vonkajšie ohraničenia neriešia podmienky, ktoré musia spĺňať hodnoty reprezentácie jedinca. Ošetrenie týchto podmienok musí byť súčasťou fitnes funkcie. Ide o podmienky, akou je napríklad to, že nemôžem minúť energiu z batérie, ktorá nebola v predchádzajúcom kroku nabitá.

Okrem čiastkovej úspešnosti, kde si každá častica pamätá svoj najväčší úspech, sa ukladá aj globálne najúspešnejšie riešenie spomedzi všetkých častíc na danej pozícii (angl. global best). V procese vyhodnocovania každej iterácie sa mení rýchlosť častíc smerom k ich vlastnej najúspešnejšej pozícii a smerom k najlepšiemu globálnemu riešeniu. Hodnoty váh sú upravované o náhodné veľkosti.

Hlavným rozdielom optimalizácie využívajúcej roj častíc od genetických algoritmov je to, že časticiam je priradená rýchlosť. Pri genetických algoritmoch sa taktiež navyše využíva kríženie a mutovanie jedincov [7].

Pri návrhu riešenia viacúčelovej optimalizácie je potrebné riešiť viacero otvorených problémov. Jedným z nich je návrh vhodných metód stanovenia váh kritérií. Vymedzením sady účelových funkcií s ich obmedzeniami je možné vytvoriť hierarchiu funkcií s ohľadom na stanovenie priorít ich dôležitosti. Cieľom projektu je obohatenie algoritmu PSO o spôsob prideľovania váh účelovým funkciám, vďaka ktorému chceme dosiahnuť nájdenie optimálnejších riešení pre uskladnenie a využitie elektrickej energie vygenerovanej prostredníctvom fotovoltaiky.

Prideľovanie váh účelovým funkciám vykonávame týmito spôsobmi:

- V počiatočných experimentoch sme váhy výsledkov jednotlivých účelových funkcií inicializovali na konštantné hodnoty, ktoré boli nastavené podľa podielu ich vplyvu na výsledné riešenie optimalizácie. Napríklad, pri experimentoch s štyrmi účelovými funkciami sme každej funkcii priradili váhu 25%.
- Hodnoty váh budú súčasťou reprezentácie jedinca – budú na posledných pozíciách jedinca a pri vyhodnocovaní fitnes funkcie riešenia hodnoty použijeme ako váhy jednotlivých účelových funkcií. Problémom je však nárast počtu atribútov jedinca, čo môže zapríčiniť potrebu zvýšenia počtu iterácií alebo počtu jedincov v populácií pre nájdenie optimálnych riešení. Takto môže vzrásť časová náročnosť prehľadávania priestoru riešení. Cieľom vytvárania fitnes funkcií je to, aby jej časová náročnosť bola čo najnižšia.
- Hodnoty váh nastavujeme samostatne pre každé spustenie optimalizácie a výsledky jednotlivých optimalizácií porovnávame na záver vzhľadom na nastavené hodnoty váh. Riešenie má potenciál mať nižšiu časovú náročnosť. V tomto prípade sa zameriavame na výber vhodnej heuristiky pre nastavovanie hodnôt váh, ktoré vyjadrujú dôležitosti účelových funkcií. V procese nastavovania váh je dôležité brať do úvahy význam funkcií, napríklad oceníme nie len to, že sme ušetrili peniaze, ale aj to, o koľko menej peňazí sme za spotrebovanú energiu minuli.

Vo fáze vyhodnocovania fitnes funkcie jednotlivých jedincov, operujeme nad dátami rôznej granularity:

- na každý deň týždňa zvlášť,
- na spriemerovaný deň,
- na jeden typ dňa (iba pondelok / pracovný deň / víkend).

Hodnoty jedinca vždy optimalizujeme a vyhodnocujeme na základe dát rovnakej granularity. Dáta pozostávajú z meraní v čase hodnôt reálnej spotreby energie, množstva energie dodanej z hlavnej siete a z množstva energie vyprodukovanej fotovoltaikou. Dáta dodané z hlavnej siete sú oddelené podľa typu tarify. Ku každému meraniu je priradená časová jednotka, kedy bolo meranie vykonané (na hodinovej báze), deň v týždni, typ dňa (pracovný / nepracovný). Hodnoty môžeme preškálovať váhami, nastavenými podľa dôležitosti atribútu. Napríklad hodnoty spotreby využívanej na stabilné nočné vykurovanie (tarifa v Austrálii) preškálujeme nižšou hodnotou, pretože tieto zmeny nemajú významný vplyv na dennú spotrebu.

Problém vysokej časovej náročnosti sa objavuje nie len v prípade vysokého počtu atribútov jedinca, ale aj v prípade vysokého počtu účelových funkcií [3]. Práve z tohto dôvodu je problematické použitie viacúčelovej optimalizácie v reálnom čase. Cieľom

našej budúcej práce je navrhnutie úpravy metód tak, aby sme v čase prehľadávania priestoru dokázali paralelizovať vyhodnocovanie fitnes funkcií jednotlivých jedincov populácie, alebo obohatiť model o možnosť dopĺňanie nových dát v procese behu iterácií.

4 Záver

Cieľom našej práce je modelovanie a optimalizácia procesov súvisiacich s vhodným využitím elektrickej energie vygenerovanej prostredníctvom fovoltaických článkov v mikrogride.

Porovnaním výhod a nevýhod prístupov pre viacúčelovú optimalizáciu sme dospeli k záveru, že pri vysokom počte účelových funkcií a atribútoch, ktoré reprezentujú jedinca, je vhodné použitie optimalizácie rojom častíc. Cieľom projektu je obohatenie algoritmus PSO o vhodné prideľovanie váh účelovým funkciám. Na základe nájdených optimálnych riešení vytvárame odporúčania pre spôsob uskladnenia a využitia elektrickej vyprodukovanej energie.

Poďakovanie: Tento príspevok vznikol za podpory grantu 002 STU-2-1/2018 "STU ako koordinátor Digitálnej koalície“ Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky a grantu APVV-16-0213 „Znalostné prístupy k inteligentnej analýze veľkých dát“.

Literatúra

1. Aftab Ahmad Khan et al.: A compendium of optimization objectives, con-straints, tools and algorithms for energy management in microgrids. Renewable and Sustainable Energy Reviews 58 (2016), 1664–1683.
2. Cho, Jin-Hee, et al.: A Survey on Modeling and Optimizing Multi-Objective Systems. IEEE Communications Surveys & Tutorials (2017), 1867 – 1901.
3. Bechikh, S., Datta, R. and Gupta, A.: Recent Advances in Evolutionary Multi-objective Optimization. Springer 2017 : Adaptation, Learning, and Optimization 20 (2017), 1-26
4. Deb, K., Singhya, K. and Hakanen, J.: Multi-objective optimization, Decis. Sci. Theory Pract., CRC Press (2016), 145–184.
5. Sharp, Ch.: Evolutionary Computing :Multiobjective Evolutionary Algorithms Multiobjective optimisation problems (MOP) -Pareto optimality, 2019.
6. Fahad Zia, M., Elbouchikhi, E. and Benbouzid, M.: Microgrids energy management systems: A critical review on methods, so-lutions, and prospects, Applied Energy, 222 (2018), 1033–1055.
7. Fan et al.: Multi-objective evolutionary algorithms embedded with machine learning - A survey, (2016), IEEE Congress on Evolutionary Com-putation (CEC) (2016), 1262–1266.

Optimalizace dotazů nad distribuovanou grafovou databází

Lucie Svitáková[1], Michal Valenta[1], Jaroslav Pokorný[2]

[1]Katedra softwarového inženýrství, FIT ČVUT v Praze
Thákurova 9, 160 00 Praha 6, Česká republika
{svitaluc,michal.valenta}@fit.cvut.cz
[2]Katedra softwarového inženýrství, MFF UK
Malostranské nám. 2/25, 118 00 Praha 1, Česká republika
pokorny@ksi.mff.cuni.cz

Abstrakt. Distribuce dat v grafových databázích je dnes zpravidla implementována jako náhodné rozmístění nově příchozích dat na jednotlivé uzly clusteru. Efektivní využití grafové databáze však často vyžaduje přeskupení těchto dat, aby byla komunikace mezi uzly clusteru co nejnižší. Vytvořili jsme modul do frameworku TinkerPop, který získá data o dotazech provedených nad grafovou databází. Tato data slouží jako vstup pro redistribuční algoritmus, který data redistribuuje se snížením potřebné komunikace mezi uzly clusteru (v níže popsaném experimentu o 70–80 %) a s relativně nízkými výpočetními nároky. Do redistribuce chceme dále zahrnout další relevantní informace, stejně tak jako tyto informace využít pro vhodné uložení nově příchozích dat. V příspěvku přiblížíme naše výsledky a představíme oblasti, kterým se chceme v rámci optimalizace dotazů dále věnovat.

Klíčová slova: grafové databáze, NoSQL databáze, distribuované systémy, redistribuce dat, optimalizace dotazů

1 Úvod

Grafové databáze získaly v posledních letech velikou popularitu [2]. Zásluhu na tom má vzrůstající počet dat s charakterem grafu, které jsou součástí například sociálních sítí, map či data lineage. Vzhledem k množství těchto dat je nutné grafové databáze již distribuovat mezi více uzlů clusteru, což přináší nové výzvy v oblasti ukládání těchto dat.

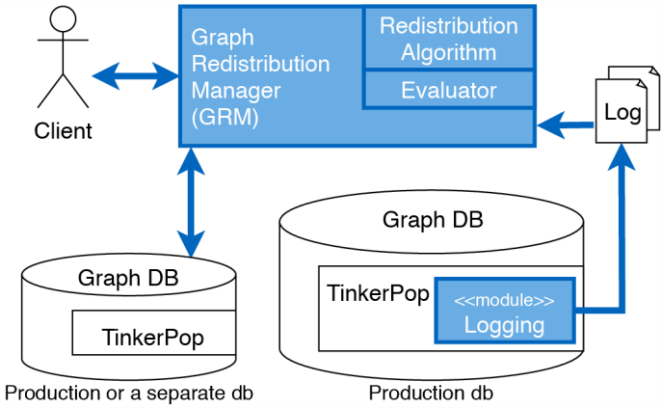
Redistribuci dat v grafových databázích nebylo dosud věnováno příliš pozornosti, náhodné rozmístění dat je zatím považováno za dostatečné. V domácí akademické sféře proto není tomuto tématu věnován žádný prostor. V zahraničních pracích lze nalézt algoritmy poskytující řešení tohoto problému, jedná se například o metodu od Vaquera a spol [6], či algoritmus Ja-be-ja [2]. U existujících prací jsme však nalezli několik nedostatků, proto jsme se jejich diskuzi a dále této problematice věnovali v rámci [4]. V následující kapitole popíšeme naši architekturu pro extrakci potřebných dat o provedených dotazech nad databází a následný redistribuční

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 173-176.

algoritmus. V třetí kapitole uvedeme rozšíření, kterým se chceme v rámci redistribuce věnovat, a ve čtvrté kapitole představíme další oblasti výzkumu, kterými se v blízké budoucnosti budeme zabývat.

2 Architektura řešení redistribuce dat

Naše implementace redistribuce dat v distribuované grafové databázi zahrnuje nejen samotný redistribuční algoritmus, ale také extrakci dat o provedených dotazech nad databází, neboť neexistuje takový nástroj pro framework TinkerPop, který by tyto informace poskytoval. Tato extrahovaná data následně slouží jako vstup pro zmíněný algoritmus. V podsekcích nyní popíšeme jednotlivé části. Celá architektura je znázorněna na obrázku **Obr. 1**.



Obr. 1. Architektura řešení pro redistribuci dat

2.1 Extrakce dat

Extrakce dat je realizována jako samostatný modul frameworku TinkerPop (na obrázku **Obr. 1** „module Logging“), což je výpočetní framework pro grafové databáze podporovaný jejich hlavními zástupci, jakými jsou Neo4j, JanusGraph, OrientDB, Datastax Enterprise Graph a další [5].

Data o proběhnutých dotazech jsou logována ve formě sledu (walk), který musí být proveden, aby byl dotaz zodpovězen (sled je konečná nenulová posloupnost $W = v_0e_{01}v_1e_{12}v_2...e_{(k-1)k}v_k$,, kde se střídají termy uzlů a hran, v jsou uzly a e jsou hrany takové, že $1 \leq i, j \leq k$ a koncové vrcholy hrany e_{ij} jsou v_i a v_j.. [1] Platí, že hrany či vrcholy se mohou v rámci jednoho sledu opakovat.)

2.2 Redistribuční algoritmus

Redistribuční algoritmus je implementován jako součást samostatné aplikace, pracovně nazývané Graph Redistribution Manager (GRM). Ten poskytuje nejen metodu redistribuce, ale také Evaluator, neboli ohodnocení dané distribuce dat v závislosti na

zjištěných dotazech. GRM si pomocná data (váhy hran podle počtu jejich zalogovaných průchodů) ukládá do grafové databáze, která může být shodná s tou, ze které jsme extrahovali data o dotazech. Může se však také jednat o samostatnou databázi, abychom se vyhnuli nadbytečnému zatěžování produkční databáze, způsobeným zpracováním velikých objemů dat distribuovaného grafu.

Jako redistribuční algoritmus jsme použili metodu od Vaquera a spol [6], které jsme dodali tři hlavní rozšíření. Metoda byla upravena pro vážené grafy, byla přidána spodní mez při redistribuci v jednotlivých iteracích algoritmu a byla zakomponována metoda simulovaného ochlazování, abychom předešli brzkému uváznutí v lokálním optimu.

Experimenty probíhaly na dvou rozdílných datasetech, s různými nastaveními vstupních parametrů (faktor ochlazování, maximální povolená dysbalance mezi uzly clusteru, počet uzlů clusteru a další). Výsledkem je snížení komunikace mezi uzly clusteru o 70–80 %, což je obdobný výsledek jako například u známé metody Ja-be-Ja [2], která má však vyšší výpočetní nároky. Více detailů, jakými jsou například podrobný popis úprav původní metody nebo naše jednotlivé výsledky a porovnání, je možné nalézt v původní práci [4].

3 Plánovaná rozšíření řešení pro redistribuci dat

Současné řešení popsané výše zatím nebere v potaz existenci více replik dané informace. Stejně tak počítá s neměnným charakterem dotazů, což však nemusí odpovídat realitě. Typy dotazů se s časem mohou měnit, proto bychom také rádi začlenili temporální složku umožňující amortizaci extrahovaných dat.

Rovněž by bylo vhodné zahrnout velikost vrcholů grafu. V případě vyvažování přístupu k jednotlivým vrcholům by měla být do algoritmu začleněna také extrahovaná informace o vrcholech.

4 Oblasti našeho výzkumu optimalizace dotazů v grafových databázích

V rámci našeho výzkumu optimalizace dotazů nad grafovými databázemi bychom se dále rádi věnovali především následujícím oblastem.

4.1 Predikce

Výše popsaná řešení obsahují nově ukládané informace, které by mohly umožnit vytvoření modelu odhadující využití nově příchozích dat. Tato predikce by byla využita pro nejvhodnější uložení těchto dat v rámci distribuované grafové databáze.

4.2 Specificita

Pro různé domény jsou charakteristické různé topologie grafů. Graf pro sociální síť má například výrazně jinou strukturu než graf pro data lineage. Dotazy a distribuce dat

mohou být optimalizovány pro specifické topologie s lepšími výsledky, než by tomu bylo při obecném přístupu.

4.3 Strojové učení

Aby kvůli předešlému bodu nevznikaly doménově specifické databáze, lze využít strojové učení. Grafová databáze může sama zjistit, jaký typ grafu v sobě uchovává, a podle toho upravit své chování ohledně vyhodnocování dotazů a distribuce dat.

4.4 Indexace

Obdobně jako v jiných typech databází, i v grafových se využívá indexace. V rámci našeho výzkumu bychom se rádi zaměřili na výkon jednotlivých existujících indexovacích nástrojů v grafových databázích. Zároveň chceme zjistit, jaké struktury nejlépe indexovat, případně tedy současné možnosti indexace rozšířit.

5 Závěr

Představili jsme naše řešení pro extrakci dat o dotazech nad distribuovanou grafovou databází a redistribuční algoritmus, který dokáže navrhnout takovou redistribuci uložených dat, aby bylo dotazování nad touto databází co nejefektivnější. Snaží se tedy co nejvíce snížit komunikaci mezi jednotlivými uzly clusteru, zároveň však na těchto uzlech zachovává určitou balanci dat.

Uvedli jsme také vize našeho budoucího výzkumu zahrnující další rozšíření zmíněné implementace, ukládání nově příchozích dat, doménově specifické grafy, strojové učení či indexaci. Součástí naší práce bude i benchmarkování stávajících řešení, ať už grafových databází či indexovacích nástrojů.

Literatura

1. Bondy, J. A.; Murty, U.: Graph Theory With Applications. North-Holland, fifth edition, 1982, ISBN 0-444-19451-7, 12 pp.
2. DB-Engines.com: DBMS popularity broken down by database model, Popularity changes per category, September 2018 [online]. June 2019. [Navštíveno 25. června 2019]. Dostupné z: https://dbengines.com/en/ranking_categories.
3. Rahimian, F.; Payberah, A. H.; et al.: A Distributed Algorithm for Large-Scale Graph Partitioning. ACM Transactions on Autonomous and Adaptive Systems, volume 10, no. 2, June 2015: pp. 1–24, doi:10.1145/2714568.
4. Svitáková, L.: Query Analysis on a Distributed Graph Database. Master's thesis. Czech Technical University in Prague, Faculty of Information Technology, 2019.
5. The Apache Software Foundation: TinkerPop-Enabled Providers [online]. [Navštíveno 25. června 2019]. Dostupné z: http://tinkerpop.apache.org/providers.html.
6. Vaquero, L. M.; Cuadrado, F.; et al.: Adaptive Partitioning for Large-Scale Dynamic Graphs. In 2014 IEEE 34th International Conference on Distributed Computing Systems, Madrid, Spain, 2014, pp. 144–153, doi:10.1109/ICDCS.2014.23.

Vylepšenie odporúčania pomocou podmieňovania modelu neurónových sietí

Martin Mocko, Jakub Ševcech, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva, Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava
{meno.priezvisko}@stuba.sk

Abstrakt. V príspevku prezentujeme výskumné zameranie po prvom roku doktorandského štúdia. V práci identifikujeme obmedzenia deterministických modelov ako veľmi dôležitý motivačný faktor nášho výskumu. Zameriavame sa na odstránenie tohto obmedzenia pomocou podmieňovania modelu. Cieľom je skúmanie nedeterministických prístupov pre odporúčanie položiek pre používateľov. Identifikovali sme metódu, v ktorej vidíme potenciál na vytvorenie prínosu ako v oblasti odporúčacích systémov, tak v oblasti neurónových sietí, sústredením sa na nastavenie parametra latentnej chyby. Našou ambíciou je aj interpretácia latentnej chyby.

Kľúčové slová: odporúčacie systémy, neurónové siete, podmieňovanie modelu

1 Úvod

Medzi ústredné hnacie motory výskumu v poslednom období patria riešenia, ktoré sú založené na deterministickom princípe podmieňovania modelov. Ako príklad môžu poslúžiť rôzne pokroky v hlbokom učení [10,11] (najmä v oblasti spracovávania obrazu, zvuku a prirodzeného jazyka). Pod deterministickým podmieňovaním modelov chápeme trénovanie modelov tak, aby pri rovnakom vstupe **X** generovali vždy rovnaký výstup **y**.

V súčasnosti sa objavujú aj iné spôsoby trénovania modelov a to také, že ich výstup môže byť podmienený nejakou formou nedeterminizmu [3,12, 13]. Takéto riešenie dokáže poskytnúť kontrolu nad faktormi ovplyvňujúcimi výstup, ktoré sa model nedokáže deterministicky naučiť.

Jedným takýmto prístupom ktorý inšpiruje náš výskum je metóda sietí kódujúcich chybu (angl. „Error Encoding Network“, ďalej ako EEN), ktorá poskytuje kontrolu nad výstupom pomocou podmieňovania modelu na základe zakódovania reziduálnej chyby deterministického modelu [3]. Identifikovali sme doménu, v ktorej by bolo vhodné túto metódu aplikovať a tou je doména odporúčacích systémov. Motiváciou tejto aplikácie je potenciál zlepšenia výsledkov (presnosti) odporúčania zapojením nedeterminizmu do modelu. Po aplikácií metódy v tejto doméne máme záujem preskúmať tri výskumné otázky, ktoré neskôr v článku formulujeme.

P. Butka, F. Babič, J. Paralič (eds.)
Data a znalosti & WIKT 2019, Košice, 10-11. október 2019, pp. 177-182.

V našej práci identifikujeme EEN ako jednu z nedeterministických metód, ktorá má potenciál priniesť zaujímavé výsledky v oblasti odporúčacích systémov. Súvisiaci výskum analyzujeme v sekcii 2. Metódu EEN hlbšie diskutujeme v sekcii 3. Návrh metódy pre odporúčanie postavenej na EEN formulujeme v časti 4, kde taktiež uvádzame aj prvé preliminárne výsledky. Vedecký potenciál navrhnutého prístupu diskutujeme v závere sekcie 4.

2 Súvisiace práce

V prvej časti sa venujeme podmieňovaniu modelu v kontexte neurónových sietí ako metóde, ktorá je hlavnou inšpiráciou našej práce. Ďalej stručne opisujeme metódy používané v doméne odporúčacích systémov založené na neurónových sieťach, ktoré tvoria základ pre náš vedecký prínos.

2.1 Podmieňovanie modelu

Podmieňovanie modelu, ktoré nás zaujíma, je také, ktoré dokáže reprezentovať ťažko oddeliteľné faktory týkajúce sa výstupu. Jedná sa o faktory, ktoré sa použitím dát, ktoré máme k dispozícii, nedajú (deterministicky) naučiť; prípadne takéto faktory môžu byť aj zo svojej podstaty náhodné (napr. závislé od náhodného rozhodovania agenta) [3]. Na to, aby bolo podmieňovanie účinné treba, aby model dokázal nejakým spôsobom reprezentovať proces generujúci takéto faktory.

Na učenie sa a reprezentáciu rôznych typov distribúcií dát v doméne neurónových sietí bolo navrhnutých viacero metód [1,2,4,5,6,7,8]. Niektoré metódy sú dokonca tak sofistikované, že nemajú ani obmedzenie na typ distribúcie, ktorú by sa z dát mali naučiť [2,5]. Ďalej opisujeme dve veľmi odlišné metódy – obe z nich však môžu byť použité na zachytenie generujúcej distribúcie a následnú generáciu nových vzoriek. Tieto modely dokážeme taktiež využiť na podmieňovanie iného modelu.

Variačné autoenkódery [5] (angl. „Variational Auto-Encoders“, ďalej VAE) plnia funkciu typických autoenkóderov v tom zmysle, že sa snažia vstup **X** na výstupe znovu zrekonštruovať. VAE sa skladá z časti enkódera a dekódera. Vnútri enkódera sa skrýva mapovanie do málo-dimenzionálneho latentného priestoru, ktorý je reprezentovaný pravdepodobnostným rozdelením (zvyčajne viacrozmerné normálne rozdelenie). Ak je toto rozdelenie dobre natrénované, dokážeme pomocou vzorkovania z natrénovaného rozdelenia generovať nové vzorky, ktoré sa podobajú na vzorky z pôvodných dát. Takýmto spôsobom by sme mohli vytvoriť možno ešte bohatšiu reprezentáciu reziduálnej chyby.

Generatívne súperiace siete [2] (angl. „Generative Adversarial Networks“, ďalej GAN), na druhú stranu nepotrebujú mať určené, akým pravdepodobnostným rozdelením majú nízko-dimenzionálny latentný priestor reprezentovať. Tieto metódy však častokrát majú nevýhodu všeobecného nedostatku plnej podpory naprieč dátami (nezachytenie modusov) – je to problém, na ktorý je vždy potrebné myslieť. Medzi ďalšie problémy patria problémy s optimalizáciou a ťažkosti pri posudzovaní preučenia a zovšeobecnenia [6].

2.2 Neurónové siete v odporúčacích systémoch

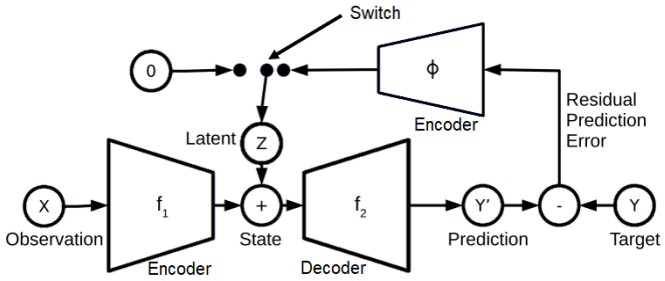
Doména odporúčacích systémov bola v posledných rokoch taktiež ovplyvnená rozvojom, ktorý sa odohráva v metódach neurónových sietí. Tieto metódy sa v doméne odporúčacích systémov zvyčajne snažia riešiť nasledovné problémy: predikcia hodnotenia, predikcia ďalšej položky, predikcia top-N položiek [9]. Vzhľadom na obmedzený rozsah tohto príspevku tu iba menujeme niektoré metódy, ktoré pokladáme za dôležité v tejto doméne: AutoRec, GRU4Rec, Caser, MARank, Mult-VAE, RBM-CF [9].

Otvorený problém z domény odporúčacích systémov, ktorý vnímame ako pre túto prácu relevantný, je problém generovania odporúčaní z riedkych údajov o hodnotení. Domnievame sa, že zapojenie personalizovanej zakódovanej chyby do modelu má potenciál zlepšiť generované odporúčania pre používateľov.

3 Sieť kódujúca chybu (EEN)

Sieť kódujúca chybu [3] (angl. „Error Encoding Network“, ďalej EEN) je metódou a rámcom pre neurónové siete, ktorý bol zavedený v roku 2017 Henaffom a kol. Hlavnou myšlienkou tejto metódy je zakódovanie (multi-dimenzionálnej) reziduálnej chyby do málo-dimenzionálnej latentnej reprezentácie. Predpoklad je, že využitím tohto prístupu sa model naučí rozpoznávať aj faktory, ktoré sa z povahy vstupných dát nedajú deterministicky naučiť, sú príliš náročné na naučenie alebo sú zo svojej podstaty náhodné. Akonáhle dokážeme podmieňovať model aj na základe informácií, ktoré sa nedokázal naučiť, získavame väčšiu kontrolu nad modelom.

Architektonický rámec modelu predpokladá, že sa v architektúre, na ktorú bude aplikovaný, nachádza enkóder a dekóder. Typ dát, pre ktoré dáva využitie modelu najväčší zmysel, sú práve sekvenčné dáta. Model sa skladá z troch hlavných častí: 1) deterministický model, 2) enkóder pre reziduálnu chybu, 3) nedeterministický model, ktorý sme schopní podmieňovať reprezentáciou reziduálnej chyby (viď obr. 1). V tomto prípade na obrázku nie je deterministický a nedeterministický model oddelený, resp. oddelenie týchto modelov reprezentuje „Switch“ – pre deterministický model je to nulový vektor, pri nedeterministickom to môže byť akýkoľvek vektor reálnych čísiel.



Obr. 1. Architektúra siete kódujúcej chybu [3]

Sieť kódujúca chybu sa v podmieňovaní odlišuje od metód VAE a GAN najmä v tom, že ju dokážeme trénovať jednoducho, lacno a nepotrebujeme počítať aproximačné riešenia kvôli neriešiteľnosti problému, ktorý sa snažia riešiť (problém treba relaxovať) [9]. Takýto model bude spravidla používať omnoho menej parametrov ako spomínané alternatívy. Originálna aplikácia metódy EEN bola pre predikciu ďalších obrázkov videa. Metóda bola taktiež úspešne aplikovaná pre syntetizovanie expresívnej reči [14].

4 Návrh nášho riešenia a diskusia

Na základe odprezentovaných súvislostí formulujeme návrh ďalšieho smerovania dizertačnej práce: chceme sa pokúsiť zlepšiť presnosť generovaných odporúčaní pomocou metódy siete kódujúcej chybu. Náš navrhovaný prístup zahŕňa vygenerovanie latentnej chyby, minimalizujúcej chybu v odporúčaní, unikátnej pre každého používateľa odporúčacieho systému. Na základe tejto chyby podmienime model a vygenerujeme odporúčanie. Takúto latentnú chybu zvolíme na základe validačnej vzorky dátovej množiny.

Z našej formulácie vyplývajú nasledujúce výskumné otázky:

- **VO1:** Dokážeme použitím EEN zlepšiť presnosť odporúčania pre používateľov?
- **VO2:** Dokážeme pomocou kódovania chyby zmeniť odporúčania položiek nejakými želanými spôsobmi? (napr. zmena v štýle odporúčaných položiek)
- **VO3:** Ak by sme ponúkli používateľom možnosť zmeniť výstup odporúčaní pomocou zmeny latentnej chyby, viedla by takáto zmena k zlepšeniam v satisfakcii / v relevantných metrikách ?

Pôvodne sme zamýšľali pre účely nášho výskumu použiť dátovú množinu zo zľavového portálu ZľavaDňa. Z dôvodu, že sme si najskôr zvolili implementovať EEN pre typ neurónovej siete, ktorá predikuje hodnotenie používateľov, sme následne podľa tohto problému vybrali aj vhodnú množinu – tou bola množina od MovieLens, konkrétne 1M dataset [15]. Modelom neurónovej siete, ktorý sme sa rozhodli rozšíriť o EEN rámec, je model AutoRec. Je to model založený na autoenkóderoch, ktorý slúži na kolaboratívne filtrovanie. Pre finálnu verziu článku sa nám podarilo EEN rámec úspešne naimplementovať. Zatiaľ máme k dispozícii len predbežné výsledky, na hĺbkové otestovanie sme dostatok času nemali. Každopádne, predbežné výsledky vnímame veľmi priaznivo.

Dátovú sadu sme rozdelili na trénovaciu a testovaciu sadu v pomere 90:10. Oba porovnávané modely, AutoRec a AutoRecEEN, sme trénovali po dobu 1000 epôch, používali sme rovnaký optimalizátor (Adam), rovnakú rýchlosť učenia (0.001) a taktiež rovnaké klesanie rýchlosti učenia. Taktiež sme používali rovnaký počet neurónov na skrytej vrstve (500). V tabuľke 1 môžeme vidieť výsledky porovnania modelu AutoRec a AutoRecEEN na metrike RMSE (angl. „root mean-squared error“). Stĺpce s označením „rs“ označujú východziu hodnotu náhodného stavu, pomocou ktorej bolo robené rozdelenie dátovej množiny na trénovaciu a testovaciu (angl. pomenovanie – „random seed“). V ďalších stĺpcoch sa nachádza priemer nameraných štyroch hodnôt a taktiež ich smerodajná odchýlka. Deterministický model AutoRec nevyžadoval žiadny špeciálny druh nastavovania. Avšak, pri AutoRecEEN sme museli pre každého

používateľa nastaviť aj latentnú reprezentáciu chyby. Pre tento experiment sme si zvolili túto reprezentáciu pre používateľov vypočítať z trénovacích dát. Dôvodom bolo, že trénovacích dát je najviac - a v prípade použitia validačných dát (pre latentnú reprezentáciu chyby) by sme výber tejto reprezentácie robili pre každého používateľa z oveľa menšieho množstva dát. Takáto reprezentácia by mohla byť viac zašumená. Na základe predbežných výsledkov sa domnievame, že použitie EEN pre odporúčanie dokáže zlepšiť presnosť odporúčania pre používateľov.

Tab. 1. Porovnanie metriky RMSE pre klasický AutoRec model a náš model AutoRecEEN

	rs=1000	rs=40	rs=50	rs=99	Priemer	Sm.odch.
AutoRec	0.85262	0.78376	0.78444	0.7823	0.80078	0.03457
AutoRec EEN	**0.84609**	**0.73785**	**0.73248**	**0.74402**	**0.76511**	0.05419

Drvivá väčšina prístupov pre odporúčacie systémy sú implementované deterministickými modelmi, kde nie je možné model ďalej podmieňovať (napríklad na základe latentnej reziduálnej chyby). Z tohto pohľadu vidíme priestor pre aplikáciu spomínanej siete kódujúcej chybu a taktiež vidíme priestor vo vyskúšaní rôznych implementácii tohto rámca – keďže rámec nevyžaduje príliš špecifickú architektúru siete. Zaujímavým problémom v tomto kontexte je určite spôsob, ako správne zvoliť latentnú reprezentáciu chyby pre daného používateľa. Toto je dôležité nielen z hľadiska vygenerovaní čo najlepších odporúčaní, ale taktiež z hľadiska menenia odporúčaní rôznymi spôsobmi, aké si môže želať napríklad majiteľ elektronického obchodu (častejšie odporúčanie určitých typov položiek). Vedecký prínos vidíme v aplikovaní metódy v doméne odporúčacích systémov – pre zlepšenie generovaných odporúčaní a v preskúmaní interpretovateľnosti podmieňujúcej latentnej chyby. Predbežné výsledky aplikácie siete kódujúcej chybu v oblasti odporúčania poukazujú na priaznivý účinok na presnosť odporúčania.

Poďakovanie: Tento príspevok vznikol s podporou grantov VG 1/0667/18, VG 1/0725/19 a APVV-15-0508.

Literatúra

1. Dinh, L., Krueger, D., and Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516. 2014.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets. In Advances in neural information processing systems, 2014 (pp. 2672-2680).
3. Henaff, M., Zhao, J., and LeCun, Y.: Prediction under uncertainty with error-encoding networks. arXiv preprint arXiv:1711.04994. 2017.
4. Hinton, G. E., Osindero, S., and Teh, Y. W.: A fast learning algorithm for deep belief nets. Neural computation, 18(7), 2006, 1527-1554.

5. Kingma, D. P. and Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013.
6. Kingma, D. P. and Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolu-tions. In Advances in Neural Information Processing Systems, 2018 (pp. 10215-10224).
7. Salakhutdinov, R., Mnih, A., and Hinton, G.: Restricted Boltzmann machines for collaborative filtering. In Proceedings of the 24th international conference on Machine learning, 2007 (pp. 791-798).
8. Uria, B., Côté, M. A., Gregor, K., Murray, I., & Larochelle, H.: Neural autoregressive distribution estimation. The Journal of Machine Learning Research, 17(1), 2016, 7184-7220.
9. Zhang, S., Yao, L., Sun, A., and Tay, Y.: Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys (CSUR), 52(1), 5. 2019.
10. Minar, M. R., and Naher J.: Recent advances in deep learning: An overview. arXiv preprint arXiv:1807.08169. 2018.
11. Goodfellow, I., Bengio Y., and Courville A.: Deep learning. MIT press. 2016.
12. Xue, T., Wu, J., Bouman, K., and Freeman, B.: Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In Advances in neural information processing systems (pp. 91-99). 2016.
13. Khan, D. A., Li, L., Sha, N., Liu, Z., Jimenez, A., Raj, B., and Singh, R.: Non-Determinism in Neural Networks for Adversarial Robustness. arXiv preprint arXiv:1905.10906. 2019.
14. Wu, X., Cao, Y., Wang, M., Liu, S., Kang, S., Wu, Z., and Meng, H.: Rapid Style Adaptation Using Residual Error Embedding for Expressive Speech Synthesis. In Interspeech (pp. 3072-3076). 2018.
15. Harper, F.M. and Konstan, J.A. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems 5, 19:1–19:19. 2015.

Krátkodobý kontext odvodzovaný z aktivity používateľa v e-obchode

Miroslav Rác, Michal Kompan, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva, Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

{meno.priezvisko}@stuba.sk

Abstrakt. Preferencie používateľa sa prispôsobujú aktuálnym okolnostiam, ktorým používateľ čelí. Ignorovanie týchto kontextuálnych informácií pri odporúčaní obsahu vedie k suboptimálnej predikcii. V našom príspevku pracujeme s predpokladom, že časť kontextu používateľa sa vynára z aktivity. Preto sú v tomto pohľade kontext a aktivita neoddeliteľné a navzájom sa ovplyvňujú. Veríme, že zmeny kontextu vyplývajúce z aktivity môžeme zachytiť a predikovať na viacerých úrovniach, počínajúc s úrovňou jednotlivých sedení používateľa v e-obchode.

Kľúčové slová: odporúčací systém, e-obchod, zmena preferencií, interakčný kontext, zámer používateľa

1 Úvod

Odporúčacie systémy pomáhajú používateľom prekonať problém informačného zaťaženia najmä využitím informácií o ich preferenciách. Učenie sa takýchto preferencií je náročná úloha. Keďže ich väčšinou používateľ nevkladá do systému explicitným vstupom, musia byť odvodzované z implicitných pozorovaní. Navyše, otázky vzbudzuje vývoj preferencií v čase. Rozlišujeme dlhodobé viac-menej stabilné preferencie a krátkodobé, ktoré sú z ich podstaty veľmi premenlivé a často prechádzajú rýchlymi a zásadnými zmenami.

Odporúčací systém dokáže monitorovať iba zlomok faktorov vyvolávajúcich tieto zmeny. Jeden z očividných faktorov je čas, ktorý sa často využíva ako zástupný kontext pre skutočné, avšak nepozorovateľné dôvody zmien. Množstvo výskumníkov využíva informácie o čase na vytvorenie rôznych mechanizmov zabúdania, ktoré preukázateľne vedú k zlepšeniu odporúčaní [9]. Nevýhoda využívania času ako samostatnej indikácie zmeny však je, že nedokáže povedať akým smerom sa zmeny preferencií uberajú.

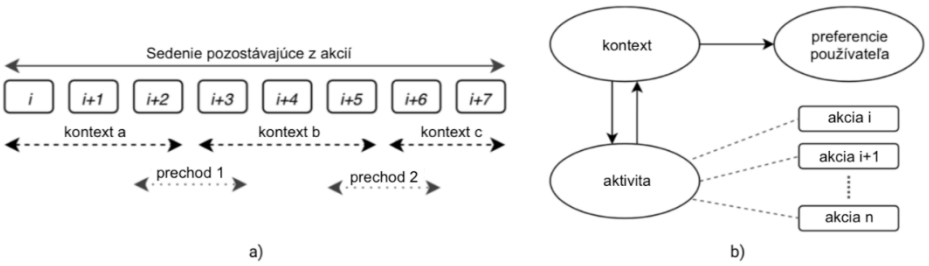
Preferencie môžu byť relevantné z dlhodobej perspektívy alebo môžu závisieť od aktuálneho kontextu, ktorý ich silne ovplyvňuje. Za určitých okolností sa používateľovi nemusia páčiť ani jeho najobľúbenejšie veci. Dobrá interpretácia kontextu môže systému poskytnúť hodnotné indikácie o aktuálnom krátkodobom záujme používateľa. Avšak, veľká časť používateľov vykazuje vysokú premenlivosť záujmov aj v krátkej

dobe [7]. Táto premenlivosť nie je obsiahnutá v rámci jedného kontextu. Predpokladáme, že kontext je nestabilný. Kontext sa konštantne vyvíja a prechody medzi jednotlivými témami obsahujú vzory.

2 Súvisiace práce

Kontextualizované systémy spočiatku využívali reprezentačný kontext vo forme pozorovateľných atribútov. Tie sú obyčajne definované už pred prvou akciou používateľa, ako napr. lokalita alebo počasie [1]. Iný, tzv. interakčný pohľad [2] hovorí o obojsmernom vplyve medzi kontextom a aktivitou. To znamená, že prebiehajúca aktivita používateľa má nepriamy vplyv na jeho preferencie. Využitie sekvencií posledných akcií používateľa na definovanie kontextu signifikantne zlepšuje presnosť odporúčaní [5].

Predpokladáme, že každá interakcia používateľa s položkou sa vzťahuje k nejakému kontextu, pričom každá položka môže byť predmetom interakcie v rámci viacerých odlišných kontextov. Sekvencia akcií teda reflektuje aktuálny kontext používateľa. Avšak, každá akcia je zároveň potenciálnym prechodom k inému kontextu (Obr. 1a). Zmena kontextu môže vyústiť do zmeny aktuálnych preferencií používateľa (Obr. 1b).



Obr. 1. Časť a) zobrazuje ako sa kontext mení v rámci jednej aktivity používateľa. Časť b) zobrazuje vzájomný vplyv medzi aktivitou a kontextom a ako aktivita nepriamo vplýva na aktuálne preferencie používateľa [8].

Podobné predpoklady využíva metóda navrhnutá Hariri a kol. [4]. Ku každej položke pomocou tagov modelujú množinu latentných tém, ktoré považujú za reprezentácie rôznych kontextov. Medzi týmito latentnými reprezentáciami dolujú frekventované vzory, ktoré potom využívajú na predikciu nasledujúcej položky. Latentné informácie umožňujú zachytiť charakteristiky na vyššej úrovni abstrakcie, čo vyžaduje menej dát na trénovanie a zároveň uľahčuje adaptáciu na náhle zmeny v preferenciách používateľa aj v rámci prebiehajúceho sedenia. Gupta a kol. [3] rozdeľujú akcie v reláciách na menšie skupiny po sebe idúcich akcií, medzi ktorými je vysoká podobnosť. Inými slovami, v týchto menších častiach relácie sa preferencie používateľa menia iba veľmi málo. Takéto skupiny položiek autori využívajú na predikciu nasledujúcej položky pomocou jednoduchej nelineárnej neurónovej siete. Tento prístup prekonáva aj zložité riešenia založené na rekurentných neurónových sieťach.

Tieto práce svojimi prvými krokmi ukázali potenciál výskumu kontextu a preferencií na nízkych úrovniach interakcie používateľa s e-obchodom.

3 Kontext riadený aktivitou v e-obchode

V našom výskume v rámci doktorandského štúdia sa zameriavame na odporúčanie v e-obchode. Táto doména má špecifické charakteristiky v porovnaní napr. doménou filmov, hudby alebo článkov. Väčšina používateľov zvyčajne nenavštevuje online obchody často a na periodickej báze. Preto sú pre odporúčanie dôležité krátkodobé záujmy a zámery, ktoré sa zvyčajne odvodzujú z poslednej aktivity používateľa.

RQ1. *Ako môžeme využiť správanie používateľa v modelovaní kontextu a jeho zmien?* V odporúčacích systémoch je štandardným a populárnym spôsobom pristupovať ku kontextu ako k dobre štruktúrovanej informácii. Teda ako k množine atribútov, ktoré sú identifikovateľné, pozorovateľné a ich štruktúra sa príliš nemení. Historicky bol kontext v systémoch opisovaný malým počtom atribútov, ako napríklad lokalita používateľa, aktuálne ročné obdobie a počasie. Neskôr bol rozšírený o používateľské aspekty ako emocionálny stav alebo špecifikovaný zámer. Tento tzv. *reprezentačný pohľad* na kontext predpokladá, že kontext popisuje vlastnosti prostredia používateľa, v ktorom sa odohráva aktivita, a že kontext a aktivita sú ľahko oddeliteľné.

Alternatívny *interakčný pohľad*, ktorý po prvýkrát opísal Dourish [2] má odlišný postoj k vyššie uvedeným predpokladom. Tvrdí že kontext je relačná vlastnosť medzi objektom a aktivitou. Túto vlastnosť nemožno dopredu identifikovať, a potom jednoducho pozorovať, pretože jej rozsah je dynamický. Z pohľadu stability je kontext príležitostný, takže jeho relevancia závisí od konkrétneho nastavenia. Čo je najdôležitejšie, podľa tohto pohľadu nielenže kontext ovplyvňuje správanie používateľa, ale sa konštantne formuje počas prebiehajúcej aktivity. To hovorí o obojsmerných vplyvoch medzi kontextom a aktivitou.

Reprezentačný kontext je založený na určitosti fyzických atribútov kontextu. Ide o objektívnu reprezentáciu interakčných javov. Protichodný *interakčný pohľad* však chápe kontext ako subjektívny a etablovaný aspekt interakcie ľudí. Toto sa ale zatiaľ javí ako nie jednoducho realizovateľné v informačných systémoch.

RQ2. *Môže byť prístup detekcie zmien kontextu aplikovaný na kontextuálne informácie, za účelom modelovania závislostí a prechodov na vyšších úrovniach?* Plnenie jedného cieľa používateľa môže presahovať hranice jedného sedenia zrekonštruovaného pomocou časových heuristík. Môže byť kontext využitý na vytváranie sekvencií akcií vedúcim k riešeniu jedného cieľa používateľa, ktorý je rozdelený do viacero sedení? Presnejšie rozdeľovanie akcií by mohlo viesť k vyššej kvalite odporúčaní.

Predpokladáme, že úplná eliminácia statického kontextu bude viesť k nedostatočným výsledkom. Kontext je náročné odhadnúť z prvých pár akcií. Práve statický kontext by mohol pomôcť ako počiatočný stav kontextu, ktorý by bol ďalej prispôsobovaný vplyvmi prebiehajúcej aktivity. Na druhej strane, aj v kontexte existuje hierarchia [6]. Takže kontext odvodený v predošlých sedeniach alebo na vyšších úrovniach (napr. založený na závislostiach medzi sedeniami) by tiež mohol byť využitý ako počiatočný stav.

4 Diskusia

V našej práci sa venujeme problémom vyplývajúcim zo silne dynamického online prostredia. Existujúce riešenia v odporúčacích systémoch sa zaoberajú predovšetkým modelovaniu preferencií používateľov bez súčasného pozorovania charakteristík správania. Vychádzajúc z podstaty dynamiky správania ľudí, preferencie používateľov podliehajú prechodným ale zásadným zmenám aj na úrovni akcií v rámci jedného sedenia. Bez zohľadňovania správania používateľov tieto zmeny nemožno zachytiť.

Na základe ponímania interakčného kontextu, nedávna aktivita a okolnosti môžu byť generalizované a reprezentované v latentnom priestore. Latentná kontextualizácia začína byť horúcou témou. Ide o generalizáciu okolností, ktorá robí detekciu zmien jednoduchšou a zároveň čiastočne redukuje problém riedkosti dát. Keďže ide o informácie odvodzované zo správania používateľov, táto metóda môže byť efektívnejšia po stránke akvizície, dostupnosti a tiež voči potenciálnym hrozbám ochrane súkromia.

V našom doktorandskom projekte skúmame vlastnosti interakčného kontextu používateľov v e-obchode. V našej predošlej práci sme skúmali dynamiku správania ľudí, kde sme skúmali predpoklad, že záujmy používateľov v rámci jedného sedenia sú variabilné. Okrem toho sme našli vzory správania tesne pred a počas nákupu [7]. Preukázali sme, že veľká časť používateľov si splnenie jedného cieľa (napr. nákupu) rozdeľuje do viacerých návštev e-obchodu, teda do viacerých sedení.

Veríme, že dynamické modelovanie kontextu založené na správaní používateľa je efektívnejšie ako používanie manuálne predefinovaných a statických atribútov. A to nielen kvôli tomu, že ich hodnoty nemusia byť kompletne pozorovateľné a môžu podliehať ľudskej dezinterpretácie konceptov. Okrem toho predpokladáme, že pod vplyvom aktivity sa kontext môže meniť aj na mikro-úrovni, teda v rámci prebiehajúceho sedenia, a to aj niekoľko krát. Ak by sa nám podarilo využiť tieto efekty, veríme že je možné ich zovšeobecniť a aplikovať aj na vyššie úrovne, teda napríklad na modelovanie závislostí medzi sedeniami používateľov.

Poďakovanie: Táto práca bola podporovaná grantami VG 1/0667/18, VG 1/0725/19, APVV-15-0508 a je čiastkovým výsledkom výskumného projektu 002STU-2-1/2018.

Literatúra

1. Peter J Brown, John D Bovey, and Xian Chen. 1997. Context-aware applications: from the laboratory to the marketplace. IEEE Personal Communications 4, 5 (1997), 58–64.
2. Paul Dourish. 2004. What we talk about when we talk about context. Personal and Ubiquitous Computing 8, 1 (2004), 19–30.
3. Kartik Gupta, Noveen Sachdeva, and Vikram Pudi. 2018. Explicit modelling of the implicit short term user preferences for music recommendation. In Proc. of European Conf. on Information Retrieval. Springer, 333–344.
4. Negar Hariri, Bamshad Mobasher, and Robin Burke. 2012. Context-aware music recommendation based on latenttopic sequential patterns. In Proc. of the 6th ACM Conf. on Recommender systems. ACM, 131–138.

5. Dietmar Jannach, Lukas Lerche, and Michael Jugovac. 2015. Adaptation and evaluation of recommendations for short-term shopping goals. In Proc. of the 9th ACM Conf. on Recommender Systems. ACM, 211–218.
6. Cosimo Palmisano, Alexander Tuzhilin, and Michele Gorgoglione. 2008. Using context to improve predictive modeling of customers in personalization applications. IEEE Transactions on Knowledge and Data Engineering 20, 11 (2008), 1535–1549.
7. Miroslav Rac, Michal Kompan, and Maria Bielikova. 2019. Preference Dynamics and Behavioral Traits in Fashion Domain. In Proc. of SMAP 2019, IEEE Press, Accepted.
8. Miroslav Rac. 2019. User's Activity Driven Short-term Context Inference. In 13th ACM Conference on Recommender Systems, September 16–20, 2019, Copenhagen, Denmark. ACM, New York, NY, USA, [Accepted].
9. Hongyi Wen, Longqi Yang, Michael Sobolev, and Deborah Estrin. 2018. Exploring recommendations under user-controlled data filtering. In Proc. of the 12th ACM Conf. on Recommender Systems. ACM, 72–76.

Peter Butka, František Babič, Ján Paralič (editori)

Data a Znalosti & WIKT 2019

Zborník konferencie

1. vydanie
188 strán
Vydala: Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach, 2019

ISBN 978-80-553-3354-0