DATA A ZNALOSTI 2017

Sborník konference

Editoři

Josef Steinberger
Martin Zíma
Dalibor Fiala
Martin Dostal
Michal Nykl

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

+NTIS

Plzeň, hotel Angelo
Česká republika
5. – 6. října 2017
www.dataznalosti.cz

<KIV> KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Data a znalosti 2017 organizují

Katedra informatiky
a výpočetní techniky
Západočeská univerzita v Plzni
Technická 8
306 14 Plzeň

Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Technická 8
306 14 Plzeň

Nové technologie
pro informační společnost
Západočeská univerzita v Plzni
Technická 8
306 14 Plzeň

Partneři vydání

IntraWorlds s.r.o.
Bělohorská 476/7
301 00 Plzeň

KadeL Data service, spol. s r.o.
Pod Vinicemi 931/2
301 00 Plzeň

RTsoft s.r.o
Pod Všemi svatými 40
301 00 Plzeň

DATA A ZNALOSTI 2017

Sborník konference

Editoři

Josef Steinberger
Martin Zíma
Dalibor Fiala
Martin Dostal
Michal Nykl

Plzeň, hotel Angelo
Česká republika
5. – 6. října 2017
www.dataznalosti.cz

Vydáno: Západočeskou univerzitou v Plzni

Data a znalosti 2017

Editoři

Josef Steinberger, Dalibor Fiala, Michal Nykl
Nové technologie pro informační společnost
Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Technická 8
306 14 Plzeň

Martin Zíma, Martin Dostal
Katedra informatiky a výpočetní techniky
Fakulta aplikovaných věd
Západočeská univerzita v Plzni
Technická 8
306 14 Plzeň

Partneři vydání
IntraWorlds s.r.o.
KadeL Data service, spol. s r.o.
RTsoft s.r.o.

Autoři příspěvků jsou uvedeni v obsahu. Každý příspěvek byl recenzován, recenzenti jsou členy programových výborů konferencí.

Tato publikace neprošla redakční ani jazykovou úpravou.

Vydala
Západočeská univerzita v Plzni
P.O.Box 314, Univerzitní 8, 306 14 Plzeň

Elektronická verze sborníku konference

1. vydání, 248 stran

ISBN 978-80-261-0720-0

Předmluva

Data a informační technologie ovlivňují naši společnost na každém kroku stále více. Přirozeně tak datové, informační a znalostní inženýrství tvoří významnou složku náplně činnosti mnoha výzkumných skupin působících v Česku a na Slovensku. Ve dnech 5. a 6. října 2017 se v Plzni uskutečnilo již 37. setkání odborníků zabývajících se datovými a znalostními technologiemi, které bylo prostorem pro výměnu poznatků a zkušeností mezi výzkumníky z českých a slovenských univerzit, výzkumných ústavů a také vývojových týmů předních nadnárodních firem. Pohodlné zázemí nalezlo více než 60 vědců v příjemném prostředí hotelu Angelo přímo naproti proslulému plzeňskému pivovaru.

Konference *Data a znalosti* navazuje na dlouholetou tradici dvou prestižních konferencí *DATAKON* a *Znalosti*. *DATAKON* existoval od roku 2001, kdy navázal na konferenci *DATASEM* s tehdy dvacetiletou tradicí. Konference *Znalosti* se pořádala rovněž od roku 2001. Tyto dvě konference se v roce 2015 sloučily v jednu a daly tak vzniknout současné podobě největšího setkání výzkumníků v oboru datových a znalostních technologií z oblasti celého bývalého Československa.

Hlavními tematickými okruhy konference v roce 2017 byly:

- získávání, ukládání a zpracování Big Data,
- dolování dat,
- vizualizace velkých dat,
- analýza nestrukturovaných dat,
- strojové učení, klasifikační a prediktivní systémy,
- tvorba, publikování a využívání otevřených a propojených dat,
- indexování a vyhledávání textových a multimediálních dat,
- modelování uživatelů, adaptivní a personalizované systémy,
- pokročilá uživatelská rozhraní softwarových a informačních systémů,
- systémy pro správu znalostí v organizacích,
- expertní, inteligentní a agentní systémy, počítačová inteligence,
- počítačová lingvistika,
- ontologické a konceptuální modely a
- automatické odvozování a plánování.

Autoři posílali příspěvky ve formě rozšířených abstraktů v českém, slovenském nebo anglickém jazyce do následujících kategorií:

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

Celkově bylo obdrženo 50 příspěvků, s největším zastoupením příspěvků o probíhajícím výzkumu. Každý příspěvek posoudili minimálně dva členové programového výboru. Výsledkem recenzování bylo rozhodnutí o přijetí 42 příspěvků. Autoři prezentovali své dílo buď formou kratší prezentace (6 příspěvků), nebo živou diskuzí při posterech (25 příspěvků), které předcházela minutová upoutávka na každou práci. Jeden příspěvek byl vybrán, kvůli jeho zaměření a velmi pozitivnímu hodnocení, pro delší prezentaci a diskuzi na páteční ráno. 10 příspěvků zaslaných do doktorandské sekce bylo představeno formou prezentace. Mladší kolegové tak měli šanci sbírat zkušenosti a také cenné rady od těch zkušenějších. Na programu bylo také pět vyzvaných přednášek na zajímavá a aktuální témata od expertů z firem zvučných jmen.

Zájem o konferenci *Data a znalosti* byl tento rok velký a téměř se vyrovnal loňskému rekordnímu a velmi podařenému setkání na Smolenickém zámku na Slovensku. Jsme velmi potěšeni, že do Plzně přijeli zástupci všech skupin působících v předmětné oblasti v Česku a na Slovensku.

Děkujeme členům programového výboru, kteří se ochotně podíleli na posuzování příspěvků. Zároveň děkujeme Martinu Zímovi za kvalitní práci při přípravě tohoto sborníku a také všem členům organizačního výboru, kteří vynaložili nemalé úsilí o to, aby se Plzeň, „Evropské hlavní město kultury“ roku 2015, stala poprvé dějištěm vědeckých diskuzí konference *Data a znalosti*.

Plzeň, říjen 2017 Josef Steinberger a Dalibor Fiala

Organizace konference

Řídící výbor

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

Programový výbor

Předseda: Josef Steinberger, FAV ZČU, Plzeň
Členové: František Babič, FEI TU, Košice
Michal Barla, FIIT STU, Bratislava
Roman Barták, MFF UK, Praha
Vladimír Bartík, FIT VUT, Brno
Václav Belák, MSD IT Innovation Center s.r.o.
Miroslav Benešovský, BenSoft s.r.o.
Petr Berka, FIS VŠE, Praha
Mária Bieliková, FIIT STU, Bratislava
Přemysl Brada, FAV ZČU, Plzeň
Radek Burget, FIT VUT, Brno
Peter Butka, FEI TU, Košice
Peter Dolog, Aalborg University
Martin Dostal, FAV ZČU, Plzeň
Marie Duží, FEI VŠB-TU, Ostrava
Igor Farkaš, FMFI UK, Bratislava
Dalibor Fiala, FAV ZČU, Plzeň
Ján Genči, FEI TU, Košice
Peter Gurský, PrF UPJŠ, Košice
Petr Hanáček, FIT VUT, Brno
Zdeněk Havlice, FEI TU, Košice
Ladislav Hluchý, ÚI SAV, Bratislava

Martin Holeňa, ÚI AV ČR, Praha
Irena Holubová, MFF UK, Praha
Martin Homola, FMFI UK, Bratislava
Tomáš Horváth, PrF UPJŠ, Košice
Tomáš Hruška, FIT VUT, Brno
Petr Hujňák, Per Partes Consulting, Praha
Jozef Hvorecký, UK, Bratislava
Dušan Chlapek, FIS VŠE, Praha
Daniela Chudá, FIIT STU, Bratislava
Karel Ježek, FAV ZČU, Plzeň
Jana Klečková, FAV ZČU, Plzeň
Jiří Kléma, FEL ČVUT, Praha
Tomáš Kliegr, FIS VŠE, Praha
Jakub Klímek, FIT ČVUT, Praha
Tomáš Knap, MFF UK, Praha
Michal Kompan, FIIT STU, Bratislava
Pavel Kordík, FIT ČVUT, Praha
Stanislav Krajči, PrF UPJŠ, Košice
Jaroslav Král, UK Praha
Pavel Král, FAV ZČU Plzeň
Michal Krátký, VŠB-TU Ostrava
Petr Křemen, FEL ČVUT Praha
Miroslav Kubát, University of Miami
Petr Kučera, Komix
Martin Labský, IBM TJW Praha
Peter Lacko, FIIT STU Bratislava
Michal Laclavík, ÚI SAV Bratislava
Vitaly Levashenko, FRI ŽU Žilina
Lenka Lhotská, FEL ČVUT Praha
Aleš Limpouch, TopoL Software s.r.o.
Marian Mach, TU Košice
Jan Martinovič, FEI VŠB Ostrava
Karol Matiaško, ŽU Žilina
Peter Mikulecký, Univerzita Hradec Králové
Martin Molhanec, FEL ČVUT Praha
Roman Mouček, FAV ZČU Plzeň
Iveta Mrázová, MFF UK Praha
Pavol Návrat, FIIT STU Bratislava
Martin Nečaský, MFF UK Praha
Giang Nguyen, ÚI SAV Bratislava
Vít Nováček, Insight @ NUI Galway
Marek Obitko, Rockwell Automation, Praha
Ján Paralič, TU Košice
Robert Pergl, FIT ČVUT Praha
Tomáš Pitner, FI MU Brno

Jaroslav Pokorný, MFF UK Praha
Lubomír Popelínský, FI MU Brno
Jaroslav Porubän, FEI TU Košice
Jan Rauch, VŠE Praha
Václav Řepa, FIS VŠE Praha
Karel Richta, FEL ČVUT Praha
Viera Rozinajová, FIIT STU Bratislava
Hana Rudová, FI MU Brno
Hana Řezanková, VŠE Praha
Pavel Smrž, FIT VUT Brno
Václav Snášel, FEI VŠB Ostrava
Ivan Srba, FIIT STU Bratislava
Vojtěch Svátek, VŠE Praha
Petr Šaloun, VŠB-TU Ostrava
Olga Štěpánková, FEL ČVUT Praha
Július Štuller, ÚI AV ČR Praha
Jozef Tvarožek, FIIT STU Bratislava
Henrieta Telepovská, TU Košice
Michal Valenta, FIT ČVUT Praha
Tomáš Vlk, ČVUT Praha
Peter Vojtáš, MFF UK Praha
Ondřej Zamazal, VŠE Praha
Jaroslav Zendulka, FIT VUT Brno
Filip Železný, FEL ČVUT Praha
Jan Žižka, MU Brno

Organizační výbor

Předseda:	Dalibor Fiala, FAV ZČU Plzeň
Členové:	Martin Dostal, FAV ZČU Plzeň
	Karel Ježek, FAV ZČU Plzeň
	Peter Krejzl, FAV ZČU Plzeň
	Michal Nykl, FAV ZČU Plzeň
	Martin Zíma, FAV ZČU Plzeň

Obsah

Příspěvky o probíhajícím výzkumu

Doktorandské symposium

Zvané přednášky (abstrakty)

Sémantická analýza ve forenzním vyšetřování

Kateřina Veselovská

Deloitte ČR

Abstrakt. Včasné vyhodnocení závažnosti případu a správná identifikace klíčových slov jsou v rámci forenzního vyšetřování dlouhodobě zásadními kompetencemi. Objem analyzovaných textových dat ale stále roste a častěji se také setkáváme s případy zahrnujícími data z mnoha různých jazyků. Tradiční přístupy k prohledávání dat zastarávají, ukazují se jako nákladné a nepraktické. Při hledání vzorců a trendů v datech již není možné spoléhat se výhradně na manuální analýzy. V tomto příspěvku ukážeme, jak využíváme nejnovější přístupy z oblasti počítačového zpracování přirozeného jazyka pro účely automatické analýzy nestrukturovaného obsahu a identifikaci vyhledávaných slov ve forenzní analytice, a uvedeme konkrétní příklady ilustrující úsporu času a nákladů při uplatnění sémantické analýzy ve forenzním vyšetřování.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 2-2.

Data Science a umělá inteligence v O2

Jan Romportl

O2 ČR

Abstrakt. Data Science tým O2 pracuje nad rozsáhlými heterogenními zdroji dat, které integruje ve své Big Data platformě a která pokrývají poměrně široké spektrum od signalizací z mobilní sítě přes webový provoz, geolokační data, IPTV či zákaznické chování, až po různé druhy textových dat. Přednáška tedy nejprve stručně představí roli Data Science týmu uvnitř struktury O2 Czech Republic, dále podrobněji popíše konkrétní zdroje dat a platformy, s nimiž pracujeme, a pak se zaměří hlavně na metody, které nejčastěji používáme, jakou roli v nich hraje aplikovaná umělá inteligence, kde využíváme deep learning či kde je naopak nutno použít lineární, snadno interpretovatelné modely. Vzhledem k tomu, že přednáška proběhne mezi akademiky, tak se nebude moci vyhnout ani představení toho, jak v O2 využíváme metody strojového zpracování přirozeného jazyka. Na závěr přednáška zmíní, jak je to s bezpečností dat a ochranou citlivých informací.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 3-3.

Bot framework a Cognitive services aneb sestavte si vlastního inteligentního bota

Lukáš Kohut

Microsoft CEE

Abstrakt. Microsoft investuje v poslední době nemalé finanční prostředky do rozvoje kognitivních služeb umožňujících rozvíjet umělou inteligenci v netradičních řešeních. Jedno z aktuálních témat dneška je využití umělé inteligence v podobě bota, který umožňuje vést konverzaci formou textu nebo hlasu. Díky spojení umělé inteligence a jakékoliv konverzační platformy se otevírají úplně nové možnosti, jak lze pomocí přirozeného jazyka lépe vyhledávat informace nebo řešit specifické úlohy. Cílem přednášky je představit kognitivní služby pracující s textovým obsahem a blíže si představit využití botů ve společnosti.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 4-4.

Jak se pečou data v Socialbakers

Milan Lepík

Socialbakers

Abstrakt. Socialbakers pomáhají firmám měřit úspěšnost na sociálních sítích jako Facebook, Twitter, Instagram či Youtube. Mezi zákazníky firmy patří polovina největších světových firem z žebříčku Fortune 500. Základem analýz jsou data. Každou vteřinu se odešle 1.000 požadavků na veřejně dostupná API sociálních sítí, takto získaná data jsou v systému dostupná okamžitě po uložení do databáze, včetně výsledku výpočtu agregací klíčových metrik. Fulltext hledání nad 1 miliardou příspěvků v clusteru se pohybuje pod 3s. Přednáška je zaměřená na popis silných a slabých stránek vyzkoušených databázových technologií z pohledu potřeb Socialbakers.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 5-5.

Moderní přístupy ke strojovému generování textů

Jiří Materna

Machine learning guru

Abstrakt. V poslední dekádě bylo možné zaznamenat obrovský nárůst popularity strojového učení a zejména umělých neuronových sítí. Je to dáno především exponenciálním nárůstem dostupného výpočetního výkonu, který umožňuje trénování i velice komplexních sítí s hlubokou strukturou. Díky tomu je možné strojovým učením řešit úlohy, u kterých to dříve bylo nepředstavitelné. Jednou z nich je i plně automatické generování textů, které nevyžaduje žádnou explicitně definovanou slovní zásobu ani gramatiku.

V přednášce budou představeny principy algoritmů založených na hlubokých rekurentních neuronových sítích a Long-short Term Memory sítích, které stojí za vytvořením sbírky básní *Poezie umělého světa*[1]. Jedná se o první knižně vydanou sbírku české poezie, která byla kompletně vygenerována počítačem.

[1] www.kosmas.cz/knihy/216522/poezie-umeleho-sveta/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 6-6.

Výzkumné příspěvky

Relační a NoSQL databáze: dvě strany téže mince?

Jaroslav Pokorný

Katedra softwarového inženýrství, MFF UK Praha
Malostranské nám. 25, 118 00 Praha

pokorny@ksi.mff.cuni.cz

Abstrakt. Analýza vlastností relačních a NoSQL databází vede k závěru, že tyto systémy pro zpracování dat jsou do jisté míry komplementární. V současných aplikacích pro Big Data, speciálně tam, kde jsou nutné rozsáhlé analýzy, se pak ukazuje, že je netriviální navrhnout infrastrukturu zahrnující software obojího typu. Z hlediska výkonu může být dokonce přínosné transformovat schéma SQL databáze do NoSQL anebo provádět oboustrannou migraci dat mezi relační a NoSQL databází. Cílem článku je diskutovat tyto možnosti a zejména některé nové metody návrhu takových databázových architektur stojící na dědictví tříúrovňové ANSI/SPARC architektury.

Klíčová slova: Relační databáze, NoSQL databáze, Big Data, Big Analytics

1 Úvod

V poslední době se zdá, že se většina velkých podniků minimálně stará o údržbu podnikových aplikací stávajících systémů. To způsobuje, že se používají "špatná" databázová schémata a obecně dochází k "úpadku databází" [10]. Autoři tvrzení vychází z diskusí s téměř dvaceti správci databází (DBA) u tří velkých podniků. Databáze se mění v závislosti na podmínkách byznysu, běžně jednou za čtvrtletí i více. Heterogenní a dynamické datové prostředí vede k tomu, že často mizí role centrálního správce a objevuje se spíše decentralizovaný přístup s více skupinami DBA zabývajícími se databázemi v podniku.

Databáze jsou většinou relační. Od zavedení relačního modelu dat bylo sice zavedeno několik databázových modelů, jako je objektově-orientovaný (OO), objektově-relační (OR), XML či RDF. OO a OR SŘBD reagovaly na objektové přístupy v softwarovém inženýrství z 90. let. Tyto prostředky, však nikdy na trhu nebyly skutečně konkurenceschopné. Důvody by mohly být v nedostatku jejich teoretických základů a omezené výkonnosti.

Dnes situaci v databázovém světě ovlivňují tzv. Big Data. Jejich základní *V*-charakteristiky jsou objem (Volume), rychlost (Velocity) a různorodost (Variety). Autor práce [9] uvádí dokonce 14 takových *V*. Ty zásadně ovlivňují infrastrukturu ukládání a zpracování Big Dat. Efektivní využívání systémů zahrnujících zpracování velkých objemů dat vyžaduje v mnoha aplikačních scénářích odpovídající nástroje

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 8-14.

pro jejich ukládání na nízké úrovni a analytické nástroje ve vyšších úrovních. Zdá se, že z pohledu uživatele je nejdůležitějším aspektem zpracování velkých objemů dat na počítači právě jejich analýza, jak se dnes říká - Big Analytics. Bohužel, velké kolekce dat obsahují data v různých formátech, např. relační tabulky, XML data, textová data, multimediální data nebo RDF trojice, což může působit potíže při jejich zpracování algoritmy pro dolování dat (DM). Rovněž zvyšující se objem dat v úložišti a počet jeho uživatelů vyžaduje spolehlivé řešení škálování v těchto dynamických prostředích a pokročilejší prostředky pro zajištění vysokého výkonu, než nabízejí tradiční databázové architektury.

Je zřejmé, že Big Analytics se provádí i nad velkým množstvím transakčních dat rozšířením metod používaných v datových skladech (DW). Technologie DW ale vždy byla zaměřena na strukturovaná data ve srovnání s bohatší variabilitou typů dat, tak jak je dnes aktuální pro Big Data. Analytické zpracování velkých objemů dat proto vyžaduje nejen nové databázové architektury, ale také nové metody pro analýzu dat.

Pro ukládání a zpracování Big Dat lze dnes volit tradiční SŘBD, paralelní DBS, distribuované souborové systémy (např. HDFS), datová úložiště typu klíč-hodnota (NoSQL databáze) a nové databázové architektury (NewSQL databáze). Pro volbu technologie jsou rozhodující aplikace, které mohou být jak transakční tak analytické. Požadují obvykle různé architektury software i hardware, často v jedné infrastruktuře.

Cílem článku je diskutovat vztah SQL a NoSQL databází v tomto polyglotním světě, a to hlavně směrem k Big Analytics. Důležitá je skutečnost, že tyto databáze mají komplementární vlastnosti [4], což i motivuje používat je v jedné infrastruktuře. V sekci 2 stručně popíšeme koncept Big Analytics. Sekci 3 podává stručný přehled technologií NoSQL databází. V sekci 4 ukážeme dualitu mezi SQL databázemi a NoSQL databázemi. Sekce 5 obsahuje závěry a výzvy pro databázovou komunitu.

2 Analytické zpracování Big Dat

Big Analytics slouží k proměňování informací ve znalosti pomocí kombinace stávajících a nových přístupů aplikovaných na Big Data. K souvisejícím technologiím patří:

- správa dat (uvažující nejistotu, zpracování dotazu v téměř reálném čase, extrakci informací, explicitní správu časové dimenze),
- nové programovací modely,
- strojové učení (ML) a statistické metody,
- komponentové architektury systémů ukládání a zpracování dat,
- vizualizace informací.

Obvyklé se rozlišují dva typy zpracování: realtimové zpracování „dat v pohybu“ (*data-in-motion*) a dávkové zpracování dat získaných z různých zdrojů do např. jedné databáze (*data-at-rest*). Dávková analýza pak může být: *malá* (Small Analytics), tj. OLAP a DW, a *velká* (Big Analytics), tj. DM, ML, e-science.

Problémy, které se v této souvislost vyskytují, vycházejí z faktu, že požadavky na Big Data jsou často dynamičtější než klasické zpracování dat v DW. Dalším problémem je, jak analyzovat Big Data pocházející z relačních DB. Velký objem je nejen

problémem pro ukládání dat, ale ovlivňuje také Big Analytics. S nárůstem složitosti dat je rovněž složitější i jejich analýza. Chceme-li využívat Big Data, musíme škálovat jak infrastrukturu, tak i standardní techniky jejich zpracování. Rychlost může být také problémem, protože hodnota analýzy (a často i dat) se snižuje s časem. Pokud je potřeba více průchodů proudu dat, musí být údaje vloženy do DW, kde lze provést další analýzy. Data mohou být uložena a zpracována pomocí např. NoSQL databáze.

Big Data jsou často zmiňována pouze v souvislosti s BI, nicméně nejen vývojáři BI, ale také vědci v e-science analyzují velké kolekce dat. Výzvou pro počítačové odborníky nebo datové vědce je poskytnout těmto lidem nástroje, které mohou efektivně provádět složitou analytiku s přihlédnutím ke zvláštní povaze zpracování velkých objemů dat. Big Analytics také nezahrnuje pouze fáze analýzy a modelování. Roli hraje zkreslený kontext, heterogenita dat a interpretace výsledků. Tyto aspekty ovlivňují škálovatelné strategie a algoritmy, proto je zapotřebí účinné předzpracování dat (filtrování a integrace) a pokročilé paralelní výpočetní prostředí. Variabilita dat se dnes stává součástí celkového návrhu systému, nicméně výkon je stále požadavkem první kategorie.

Kromě těchto spíše klasických témat DM velkých objemů dat se v posledních letech objevily další zajímavé požadavky, jako rozpoznávání pojmenovaných entit, analýza názorů a mínění (např. pozitivní, negativní, neutrální) a jejich dolování (sentiment analysis). Jejich řešení využívají zejména metody vyhledávání informací a analýzy webových dat. Porovnávání vzorů grafů se běžně používá při analýze sociálních sítí, kde grafy např. zahrnují miliardu uživatelů a stovky miliard odkazů. Technické problémy současných technik DM používaných pro Big Data pak v každém případě pocházejí z jejich nedostatečné škálovatelnosti a paralelizace.

3 NoSQL databáze

Pro ukládání a zpracování velkých kolekcí dat jsou často používány NoSQL databáze. NoSQL znamená "ne pouze SQL" nebo "žádné SQL vůbec", což dělá tuto kategorii databází velmi různorodou a ne příliš jasně specifikovatelnou. NoSQL databáze, jejichž vývoj začíná od konce 90. let, poskytují v porovnání s relačními databázemi jednodušší škálovatelnost a vyšší výkon. Popíšeme stručně jejich vlastnosti a klasifikaci včetně použitelnosti pro zpracování Big Dat. Detailnější diskusi těchto témat jsou věnovány v české literatuře např. články [6], [8], či kniha [3], škálovatelnost je diskutována v [7].

To, co je hlavní v klasických přístupech k databázím - (logický) datový model - je v NoSQL databázích popsáno spíše intuitivně, bez jakýchkoliv formálních základů. Terminologie NoSQL je také velmi rozmanitá a rozdíl mezi konceptuálním a databázovým pohledem na data je většinou rozmazaný. Nejznámější NoSQL databáze mohou být podle použitého datového modelu klasifikovány jako:

- úložiště typu klíč-hodnota, např. Redis[1],
- sloupcově-orientované, např. CASSANDRA[2],

[1] https://redis.io/

- dokumentově-orientované, např. MongoDB[3].

Zdá se, že všechny uvedené datové modely jsou v podstatě typu klíč-hodnota. Odlišují se především v možnostech agregace dvojic (klíč, hodnota) a zpřístupňování těchto hodnot. Obecněji se mezi NoSQL databáze řadí i grafové databáze, XML databáze, RDF databáze a další. Pro naše úvahy vystačíme s třemi výše uvedenými typy.

Tím, že jsou NoSQL určeny hlavně pro ukládání Big Dat, musí být tyto databáze škálovatelné. Významnou roli pak hraje jejich indexace. V NoSQL databázích se používají speciální datové struktury, např. LSM-strom (Log Structured Merge Tree) [5], požívaný např. v Cassandra a MongoDB. LSM-strom je tvořen kaskádou B-stromů. LSM-strom je vhodný speciálně pro data, kde převládá operace `INSERT`.

NoSQL bývají částí cloudových, datově intenzivních aplikacích (hlavně webových). Patří sem zábavné aplikace, obsluha stránek webových míst s vysokým provozem, doručování médií proudovým způsobem, či data vyskytující se v sociálních sítích. Google využívá sloupcově-orientovanou BigTable ve více než 60 aplikacích.

Zkušenosti s NoSQL databázemi ukazují, že je lze použít i na „malá“ data a zejména na aplikace nepožadující transakční sémantiku, např. pro adresáře, blogy nebo systémy zpracování obsahu, rovněž pro Big Analytics i dat v reálném čase (např. proudy kliknutí na webové místo). V prostředí mobilního zpracování dat jsou navíc transakce ve větším rozsahu technicky nemožné. NoSQL systémy jsou tedy vhodné spíše pro prostředí s interaktivními datovými službami. Vynucování schématu a uzamykání na úrovni řádků jako v relačních databázích může překomplikovat tyto aplikace. Absence některých vlastností ACID pak dovoluje význačné zrychlení a decentralizaci NoSQL databází.

Existuje mnoho diskusí o roli NoSQL databází při poskytování informačních služeb. ProNoSQL tábor tvrdí, že tato technologie je budoucností databází. Na druhé straně prorelační databázový tábor tvrdí, že databáze NoSQL mají velkou nevýhodu v tom, že neposkytují korektní zacházení s integritou dat. To souvisí s nedostatkem sémantiky způsobeným jejich základní vlastností – nemají schéma. Nedostatek metadat zabraňuje aplikačnímu systému vědět, která data jsou uložena a jak jsou vzájemně propojena.

V databázovém světě však NoSQL zaujímají významné místo. V hodnotícím DB-Engines Ranking se v květnu 2017 sledovalo 328 systémů. V prvních 10 místech se objevují MongoDB (5. místo), Cassandra (8. místo) a Redis (9. místo).

4 Dualita mezi SQL a NoSQL

V práci [4] autoři argumentují, že NoSQL databáze jsou spíše komplementem tradičních transakčních databází. Neměly by se spíše jmenovat „co-relational“[4]? Možná přirozenější je říkat coSQL místo NoSQL. V Tab. 1 je uvedeno devět takových rozdílů.

---

[2] http://cassandra.apache.org/

[3] https://www.mongodb.com/

[4] Pozor, v češtině pojem „korelační“ znamená něco jiného.

Důležité jsou rozdíly 1 a následně 8. Díky normalizaci mohou být data o jednom objektu v relační databázi rozložena do více relací. Např. data o zákazníkovi jsou v jedné tabulce, data o bankách, kde má zákazník účet, jsou v druhé tabulce. Propojení je realizováno přes cizí klíče. V NoSQL databázi je toto možné realizovat tak, že každý „řádek“ od banky může obsahovat pro každého zákazníka jeho data i čísla účtu. NoSQL jsou denormalizované, tj. ukládají na místě objektu nikoliv objekt, ale kopii objektu. Ten může být dokonce kompozicí řádků (hnízděná data), což v klasickém SQL není možné, není totiž kompozitní. To vede k horším možnostem aktualizace dat.

Fundamentálním rozdílem je nedostatek schémat dat (sémantiky) u NoSQL databáze. Ten brání analytikům porozumění struktuře dat a tím i vytváření seriózních analýz. Tendencí je proto vytvářet víceúrovňové modelovací přístupy zahrnující relační i NoSQL architektury včetně jejich integrace v jedné infrastruktuře. Vyvíjejí se tak metody společného návrhu pro relační a NoSQL databáze založené na modifikaci 3-úrovňového ANSI/SPARC přístupu (konceptuální, logický, fyzický návrh) [2]. Z hlediska uložení a přístupu k datům se pak hovoří o *polyglotní perzistenci* či o *polyglotních databázích*. Počítá se i s vývojem konceptuálního/databázového schématu v celkové infrastruktuře. Konceptuální návrh ovšem předpokládá korektnost současné znalosti aplikační domény. Silnou motivací proto je i fakt, že při návrhu databáze je třeba uvažovat vzory pro DM/ML, shlukování některých atributů pro zajištění výkonu systému apod.

V praxi se objevují i další možnosti, jako model konverze schématu, ve kterém se schéma SQL database konvertuje do schématu NoSQL databáze [11].

Tab. 1: Duální vlastnosti SQL a NoSQL databází

	SQL	NoSQL
1	data závislých relací ukazují k rodičům (přes cizí klíče)	od dat rodičů se ukazuje k dětem
2	uzavřený svět	otevřený svět
3	entity mají identitu (primární klíč)	identitu určuje prostředí
4	data jsou silně typovaná	potenciálně dynamicky typovaná
5	synchronní (ACID) aktualizace přes více řádků	asynchronní (BASE) aktualizace v jednotlivých hodnotách
6	změny (transakce) koordinuje prostředí	entity odpovědné reagovat na změny (případná konzistence)
7	referenční integrita založená na hodnotách	slabá referenční integrita založená na výpočtu
8	není kompozitní	je kompozitní
9	optimalizátor dotazů	vývojář/vzor

5 Závěry a výzvy

Klíčové problémy pro budování infrastruktury zpracování dat jsou v lidských rozhodnutích týkajících se NoSQL databází. Zahrnují zejména volbu správných produktů a návrh vhodné databázové architektury pro danou třídu aplikací.

Role člověka je významná i v Big Analytics. Dnes je proces DM řízen analytikem či datovým vědcem. V závislosti na aplikačním scénáři ten určuje část dat, odkud mohou být např. užitečné vzory extrahovány. Lepší řešení by bylo mít k dispozici automatický proces DM s cílem získat přibližné syntetické informace jak o struktuře, tak o obsahu velkého množství dat.

Současné výzvy pro databázový výzkum zahrnují:

- modelování polyglotních databází (relační i NoSQL v jedné infrastruktuře) [1].
- zlepšení kvality a škálovatelnosti metod DM. Formulace dotazu - zejména při absenci schématu - a prezentace a interpretace odpovědí může být netriviální.
- transformaci obsahu do strukturovaného formátu pro pozdější analýzu, protože mnoho dat není nativně strukturovaných. Filtrací lze i zmenšit objem dat.
- vývoj smysluplného a použitelného formalismus pro modelování NoSQL databází a následně silný a použitelný uživatelský dotazovací jazyk.

Vztah SQL a coSQL databází lze charakterizovat v pojmech jin a jang [4]. V čínské filosofii jde o dva související a protikladné pojmy, pomocí nichž se dá popsat vzájemný poměr rovnováhy jak v těle, tak i v dějích okolo nás (např. noc a den). Rovněž coSQL a SQL nejsou v konfliktu. Jsou to dva protiklady, které koexistují v harmonii, vzájemně se doplňují, podporují a mohou se vzájemně měnit. Kéž by tomu tak bylo i v praxi.

Literatura

1. Abelló, A.: Big Data Design. In: Proc. of DOLAP (2015) 35-38.
2. Herrero, V., Abelló, A., Romero, O.: NOSQL Design for Analytical Workloads: Variability Matters. Proc. of ER Conf. (2016) 50-64.
3. Holubová, I., Kosek, J., Minařík, K., Novák, D.: Big Data a NoSQL databáze. Grada, 2015.
4. Meijer, E., Bierman, G.M.: A co-relational model of data for large shared data banks. Commun. ACM 54(4) (2011) 49-58.
5. O'Neil, P. E., Cheng, E., Gawlick, D., O'Neil, E.: The Log-Structured Merge-Tree (LSM-Tree). Acta Inf., 33(4) (1996) 351-385.
6. Pokorný, J.: NoSQL databáze. In: Proc. of the Annual Database Conf. DATAKON'2011, J. Zendulka, M. Rychlý (eds.), Mikulov, VUT Brno (2011) 71-82.
7. Pokorný J.: NoSQL Databases: a step to databases scalability in Web environment. International Journal of Web Information Systems, 9 (1) (2013) 69-82.
8. Pokorný, J.: Big Data: jejich ukládání, zpracování a použití. Proc. of the 34th Ann. Database Conference DATAKON'2014, P. Šaloun, D. Chlapek (Eds.), VŠB Ostrava (2014) 3-16.
9. Pokorný, J.: Big Data Storage and Management: Challenges and Opportunities. In: Proc. of 12th IFIP WG 5.11 Int. Symp. on Environmental Software Systems, IFIP AICT 507, Springer (2017)
10. Stonebraker, M., Deng, D., Brodie, M.L.: Database decay and how to avoid it. In: Proc. of 2016 IEEE Int. Conference on Big Data, IEEE Explore (2016) 7-16.
11. Zhao, G., Lin, Q., Li, L., Li, Z.: Schema Conversion Model of SQL Database to NoSQL. In: Proc. of the 3PGCIC, IEEE (2014) 355-362.

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

Annotation:

Relational and NoSQL databases: two sides of the same coin?

The analysis of relational and NoSQL databases leads to the conclusion that these data processing systems are to some extent complementary. In current Big Data applications, especially where extensive analyses are needed, it turns out that it is non-trivial to design an infrastructure involving software of both types. In terms of performance, it may even be beneficial to transform the SQL database schema into NoSQL or to perform double-sided data migration between relational and NoSQL databases. The aim of the article is to discuss these possibilities and some new methods of designing such database architectures standing on the legacy of the three-level ANSI/SPARC architecture.

Interaktívna vizualizácia výsledkov vyhľadávania informácií pomocou konceptových zväzov

Veronika Novotná, Peter Butka, Miroslav Smatana

Katedra kybernetiky a umelej inteligencie,
Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

veronika.novotna@student.tuke.sk,
{peter.butka,miroslav.smatana}@tuke.sk

Abstrakt. Klasický pohľad na výsledky vyhľadávania získaných z vyhľadávača je často vo forme usporiadaného zoznamu, poznáme však aj systémy poskytujúce štruktúrovanú formu výsledkov. Cieľom tohto príspevku je popísať implementovaný systém pre interaktívnu vizualizáciu výsledkov vyhľadávania vo forme konceptového zväzu. Tento poskytuje štruktúrovanú formu hierarchicky usporiadaných podmnožín vrátených dokumentov na základe metódy známej ako FCA (Formal Concept Analysis). Vytvorený nástroj umožňuje zvoliť dopyt a získať hierarchickú štruktúru zhlukov vrátených výsledkov na základe analýzy obsahu ich krátkych popisov – tzv. snippet-ov. Takto vytvorený konceptový zväz sa zobrazuje ako interaktívny graf, na ktorý je následne možné aplikovať redukcie zvyšujúce prehľadnosť výstupnej vizualizácie. Okrem grafickej vizualizácie je možné prehliadať aj jednotlivé výsledky a odkazy na nich, analyzovať ich vzťahy a odlišnosti vzhľadom na ich pozíciu v hierarchickej štruktúre zhlukov, ako aj využiť interaktívne zobrazenie pre rozšírenú navigáciu v sub-doméne odpovedajúcej množine vrátených výsledkov.

Kľúčové slová: formálna konceptová analýza, vyhľadávanie informácií, vizualizácia, konceptový zväz

1 Úvod

V rámci oblasti vyhľadávania informácií je jedným z riešených problémov poskytnutie výsledkov vyhľadávania v štruktúrovanej forme, v podobe ktorá lepšie zohľadňuje vzťahy medzi výsledkami a umožňuje lepšie pochopenie prehľadávanej domény dokumentov. Jednou z možností ako tento problém riešiť je použiť metódy z oblasti formálnej konceptovej analýzy (Formal Concept Analysis – FCA) [1]. V tomto prípade používateľ zadá dopyt a systém mu poskytne k nájdeným dokumentom hierarchickú štruktúru zhlukov dokumentov reprezentujúcich podmnožiny dokumentov zdieľajúcich atribúty definované v danej doméne. V klasickom rámci FCA sa pracuje s binárnou vstupnou tabuľkou (objekt má alebo nemá atribút, napríklad dokument obsahuje alebo neobsahuje daný term). V praktickej analýze samozrejme často existu-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 15-19.

je potreba spracovať rôzne typy atribútov. Aj preto bolo navrhnutých viacero fuzzy prístupov. Jedným z nich je aj model tzv. zovšeobecneného jednostranne fuzzy konceptového zväzu (Generalized One-Sided Concept Lattice - GOSCL) [2]. Tento model bol použitý v rámci našej práce, jeho výhodou je možnosť spracovávať vstupné tabuľky s rôznymi typmi atribútov. Pre detaily k modelu, implementácii a použitiu GOSCL odporúčame preštudovať aj ďalšiu literatúru (okrem uvedenej), napríklad [3].

Cieľom tejto práce bolo vytvorenie nástroja pre interaktívnu vizualizáciu výsledkov vyhľadávania a poskytnúť ich používateľovi práve vo forme modelu GOSCL. Vstupom do analýzy boli krátke popisy výsledkov (tzv. snippet), tieto boli následne predspracované do formy vstupného kontextu pre tvorbu modelu GOSCL, kde objekty sú výsledky vyhľadávania a atribúty sú početnosti výskytu termov v popise výsledku.

V ďalšej kapitole popíšeme motiváciu nášho návrhu a jeho implementácie, následne sa budeme venovať navrhnutému systém a zhrnieme výsledky jeho testovania.

2 Motivácia

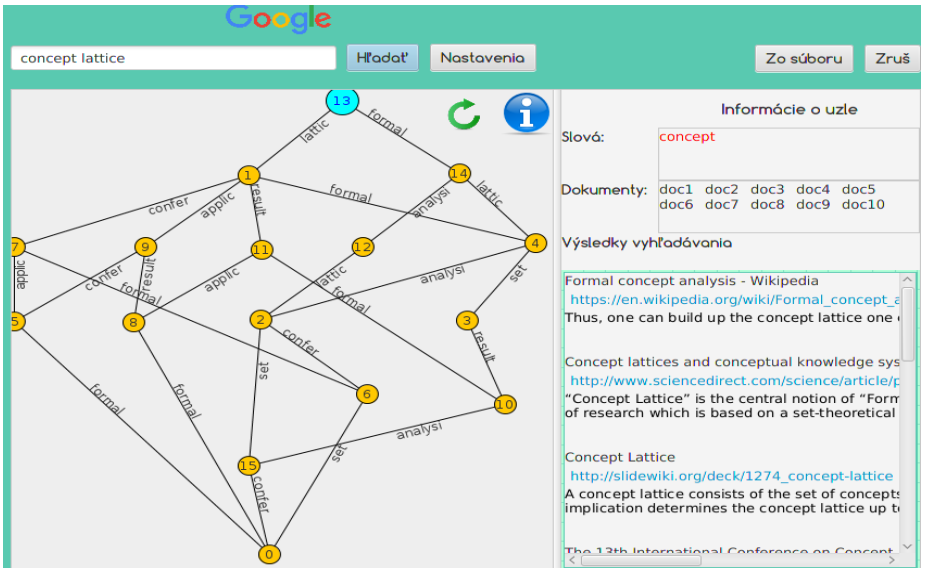
FCA už bolo samozrejme aplikované aj v doméne vyhľadávania informácií, a to v rámci riešenia rôznych problémov dopytovanie a navigácie v množine dokumentov. Väčšinou je výsledok z vyhľadávacieho stroja na nejaký zvolený dopyt vrátený vo forme usporiadaného (lineárneho) zoznamu výsledkov, pričom tieto obsahujú názov, krátky popis (snippet), či linku URL, atď. V oblasti vyhľadávania informácií boli vytvorené napríklad systémy využívajúce princípy FCA ako CREDO, respektíve rozšírená verzia CRE-CHAIN-DO [4]. Základný prvkom je analyzovať vrátené výsledky vyhľadávania na daný dopyt a vybudovať vstupný kontext pre FCA analýzu, kde objekty sú dokumenty a výskyt termov v názve alebo snippet-e predstavuje atribúty. Systém CREDO resp. CRE-CHAIN-DO bol vytvorený s cieľom analyzovať iba binárne vstupné kontexty popisujúce iba výskyt termu. Z tohto kontextu bola potom vytvorená dvojúrovňová hierarchia zobraziteľná ako klasická navigačná hierarchia v množinách dokumentov (vo forme stromovej štruktúry). Tento prístup je špecifický aj tým, že výsledná stromová štruktúra už nepredstavuje pôvodný konceptový zväz, takisto aj rozdelenie dokumentov bolo už špecificky upravené. My sme sa rozhodli zostať pri pôvodnej koncepcii FCA a ponúknuť radšej graf konceptového zväzu, pričom náš systém sme sa rozhodli zamerať na dva aspekty:

1. Poskytnúť interaktívnu vizualizáciu výstupného modelu v podobe konceptového zväzu (alebo jeho redukcie).
2. Využiť model GOSCL pre použitie rôznych typov atribútov.

3 Návrh a implementácia nástroja

Navrhovaný nástroj predstavuje aplikáciu pre používateľa, ktorý zadá dopyt a dostane výsledky Google vyhľadávania v podobe konceptového zväzu, alebo jeho redukcie. Proces vytvorenia výstupu pozostáva zo 6 hlavných krokov:

1. Získanie výsledkov z vyhľadávača – po zadaní dopytu, systém použije GoogleSearch API pre získanie množiny výsledkov, je možné zvoliť počet výsledkov do vizualizácie, výstup je uložený v JSON formáte.
2. Spracovanie výsledkov vyhľadávania – na základe nastavených parametrov sa predspracujú vrátené výsledky, t.j., title a snippet sú predspracované (tokenizácia, transformácia na malé písmená, odstránenie stop slov).
3. Vytvorenie vstupného kontextu – na základe výskytu termov je vytvorený formálny kontext, pričom tento môže byť binárny (0/1 ak sa term vyskytol /nevyskytol), alebo vektorový (fuzzy) (0 ak sa term nevyskytol, inak frekvencia termu).
4. Vytvorenie konceptového zväzu – algoritmus pre tvorbu GOSCL vytvorí výstupný model pre ľubovoľný kontext.
5. Redukcia výstupu – pre zlepšenie prehľadnosti a lepšiu interpretáciu sa často môže používateľ zvoliť redukciu, v našom prípade dve redukcie: orezanie konceptu s počtom objektov 0, alebo odstrániť koncepty s malou podporou (ponecháme len koncepty obsahujúce aspoň prahový počet objektov).
6. Vizualizácia konceptového zväzu – výstupný graf zväzu, vytvorený s pomocou rámca JUNG (http://jung.sourceforge.net/), je poskytnutý používateľovi.



Obr. 1. Interaktívna vizualizácia konceptového zväzu k zvolenému dopytu.

Získaný výsledok sa môže zobraziť, tak ako to je napríklad na Obr.1. Zobrazenie má 3 hlavné časti:

- Vstupná časť (horná časť) – zadanie dopytu, otvorenie nastavení parametrov, načítanie výsledkov zo súboru.

- Výstup grafu (časť vľavo) – interaktívny graf konceptového zväzu alebo jeho redukcie pomocou rámca JUNG, informácie o rozdieloch medzi uzlami (termy ktoré odlišujú uzly) sú zobrazené na hrane.
- Informačná časť (vpravo) – poskytuje detaily o vybranom prvku grafu, t.j., obsah konceptu, prehľad príslušných výsledkov k danej podmnožine objektov, dôležité termy daného, odlišujúce termy od rodičovského uzla, atď.

4 Experimenty

V rámci testovania sme realizovali experimenty so systémom pre rôzne dopyty, konkrétne boli experimenty spustené pre 10 rôznych dopytov do Google, pričom počet analyzovaných výsledkov pre jeden dopyt sa v závislosti od experimentu mohol meniť medzi 10 až 100 (parameter N). Okrem toho bolo jedným z nastavení aj to, či bol použitý iba binárny kontext, alebo fuzzy kontext (v tomto prípade početnosti termov v rámci snippet-u). Pre každý experiment sme sledovali minimálnu, strednú a maximálnu hodnotu výsledkov meraní vzhľadom k množine dopytov. Treba však zdôrazniť, že išlo o relatívne jednoduché prvotné experimenty so systémom, ktorých cieľom bolo nájsť základné obmedzenia fungovania aplikácie z pohľadu počtu konceptov, či náročnosti výpočtu modelu.

V prvej sérii experimentov sme sa zamerali na sledovanie počtu konceptov v závislosti od použitého počtu výsledkov vyhľadávania vstupujúcich do tvorby konceptového zväzu. Nastavenie experimentu bolo nasledovné: minimálny počet výskytu slova v rôznych dokumentoch bol 1 a nebola použitá žiadna redukcia. Snažili sme sa odhadnúť veľkosť vzniknutého konceptového zväzu pre zobrazenie pri takýchto nastaveniach. Vzhľadom k tomu, že vstup je v tejto doméne riedka matica, bol nárast počtu konceptov s pribúdajúcim počtom analyzovaných vrátených snippet-ov (dokumentov) približne lineárny. V praxi to napríklad pre prípad fuzzy konceptového zväzu znamenalo zhruba 20 konceptov pre N=10, približne 200 konceptov pre N=60, až po maximum okolo 400 konceptov pre N=100.

V ďalšom experimente sme analyzovali vplyv redukcie na báze podpory (support) na výslednú početnosť konceptov. V tomto prípade sme zobrali do úvahy 100 vstupných dokumentov a zisťovali sme vplyv redukcie. Ukázalo sa, že redukcia je výrazná, nakoľko už pri hodnote 0.1 je počet konceptov na úrovni 5-6% (z absolútneho počtu cca 400 konceptov). Pri podpore 0.3 zostáva po redukcii cca 1% konceptov.

Následne sme sa snažili experimentálne odhadnúť aká množina konceptov by ešte mohla byť pre používateľa zrozumiteľná a aké nastavenia počtu výsledkov a redukcie umožňujú tento výsledok dosiahnuť. Ukázalo sa, že rozumne prehľadný výstup bez aplikácie redukcie je možné dosiahnuť len ak náš vstup nepresiahne 10-15 dokumentov. Pri použití redukcie (podpora) je rozumný výstup dosiahnutý už pri podpore 0.2 aj pre 100 vstupných dokumentov.

Z hľadiska časovej náročnosti sa ukázalo, že najviac problematické je samotné získanie výsledkov cez API, nasleduje tvorba zväzu. Pre menšie množiny výsledkov (medzi 20 až 50, v závislosti od redukcie), je čas spracovania dostatočne krátky pre používanie aplikácie (pod pol sekundy).

5 Záver

V tejto práci sme sa venovali jednej z možností ako poskytnúť štruktúrovaný pohľad na výsledky vyhľadávania získaných z klasického vyhľadávača, a to poskytnutím interaktívnej vizualizácie hierarchickej štruktúry zhlukov získaných výsledkov v podobe konceptového zväzu. Okrem grafickej vizualizácie je možné prehliadať aj jednotlivé výsledky, odkazy na nich, analyzovať ich vzťahy a odlišnosti, ako aj využiť graf pre navigáciu v doméne.

Literatúra

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag, Berlin (1999)
2. Butka, P., Pocs, J.: Generalization of One-Sided Concept Lattices. Comput. Informat. 32(2), 355–370 (2013)
3. Butka, P., Pócsová, J., Pócs, J.: Design and implementation of incremental algorithm for creation of generalized one-sides concept lattices, Proc. of CINTI 2011, 373-378 (2011)
4. Nauer, E., Toussaint, Y.: Dynamical modification of context for an iterative and interactive information retrieval process on the web, CLA 2007, CEUR-WS proceedings (2008)

Poďakovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.025TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Interactive visualization of information retrieval results using concept lattices

This paper describes motivation and details of the tool designed and implemented for interactive visualization of information retrieval results using concept lattices, which are created using one-sided fuzzy extension of Formal Concept Analysis. The concept lattice is then shown as an interactive graph, which provides a structured view on the domain of query to the user.

Včasná identifikácia trendov v správaní používateľov elektronického zľavového portálu

Ondrej Kaššák, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva,
Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

ondrej.kassak@stuba.sk, maria.bielikova@stuba.sk

Abstrakt. Správanie používateľov služieb na webe sa v čase mení. Napríklad v zľavovom portáli používatelia reagujú na jednotlivé ponuky rozlične. Našim cieľom je dokázať včas identifikovať, ktoré ponuky sa stanú trendami (vysoko predávanými) a ktorým naopak treba pomôcť napríklad dodatočnou propagáciou. Hlavnou výzvou tejto úlohy je veľké množstvo dát o nákupoch, ktoré prichádzajú formou kontinuálneho prúdu. Riešenie, ktoré v práci navrhujeme je založené na časovo a výpočtovo efektívnom jednoprechodovom spracovaní dát umožňujúcom prácu v online čase. Týmto spôsobom sme schopní pomerne skoro identifikovať, ktoré zľavy sa stanú trendami. Opisované riešenie sme overili na reálnej množine dát zľavového portálu, kde sme ukázali, že časť trendov je možné identifikovať už na základe prvých dní, kedy sú dané zľavy v ponuke.

Kľúčové slová: frekventované prvky, posuvné okno, prúd dát, trendy v správaní používateľov webovej služby

1 Úvod

Predmetom nášho záujmu je pojem trend. Pod týmto pojmom v doméne online zľavového portálu rozumieme položku nakupovanú používateľmi v určitom časovom období (prípadne celkovo) viac ako iné položky. V oblasti dolovania dát trend typicky predstavuje frekventovaný vzor, teda sekvenciu (asociačné pravidlo), prvok alebo štruktúru (strom alebo graf) [1]. Pre potreby našej úlohy, ktorou je včasná identifikácia nakupovaných položiek, sme vybrali reprezentáciu prostredníctvom frekventovaných prvkov. Tie poskytujú jednoduchú reprezentáciu a zároveň umožňujú identifikáciu trendov v doménach kde používatelia typicky nakupujú jedinú položku (napr. zľavový portál).

Trendom môže byť *top-n* najkupovanejších položiek prípadne položky, ktorých nákup tvorí viac ako *m%* všetkých realizovaných nákupov za sledované obdobie [1]. Keďže *top-n* prvkov nedokáže dynamicky reflektovať počet aktuálne existujúcich trendov, ktorý sa v čase môže meniť, rozhodli sme sa za trendy pokladať položky nakupované vo viac ako *m%* nákupov.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 20-26.

Témou sa zaoberáme z dôvodu, že nakupovanie používateľov výrazne podlieha sezónnosti, rozličné položky sú v určitých obdobiach kupované viac, inokedy menej. Pre obchodníka je dôležité dokázať obslúžiť všetkých a preto by mal mať možnosť včas doplniť skladové zásoby, prispôsobiť reklamu náladám zákazníkov a podobne. Trendy v nakupovaní je možné predikovať len do určitej miery, maximálne na úrovni typov tovaru (napr. prezutie zimných pneumatík), nie však konkrétnych položiek (prezutie v pneuservise XY). Je však možné identifikovať ich na základe vývoja nákupov a včas tak odhadnúť, ktoré položky sa budú predávať veľa, resp. viac ako v súčasnosti.

V doméne online zľavových portálov sa položky dynamicky menia. Rýchlo vznikajú, zanikajú či predlžujú platnosť. Na tieto zmeny je potrebné dokázať rýchlo zareagovať. Položiek a nákupov je navyše pomerne veľa, čo spolu s predchádzajúcou vlastnosťou vylučuje dávkové spracovanie. Rýchlosť je v tomto prípade kľúčová aby mal obchodník dosť času zareagovať. Preto je potrebné prúdové spracovanie dát.

Prúdové spracovanie je založené na jednoprechodovom prístupe k dátam [7], kedy nie je možné opätovne prechádzať všetky dáta ale nanajvýš určitú najaktuálnejšiu podmnožinu, prípadne agregované štatistiky. Týmto spôsobom dosiahneme výpočtovo efektívne spracovanie, ktoré však vyžaduje upustenie od presnosti výsledku.

V sekcii 2 opisujeme existujúce prístupy identifikácie trendov v prúde dát. V sekcii 3 na základe zistených poznatkov identifikujeme 2 metódy využívajúce posuvné respektíve skáčuce okno. Tieto metódy overujeme a porovnávame v sekcii 4. Zistené poznatky diskutujeme v závere práce, v sekcii 5.

2 Prístupy identifikácie frekventovaných prvkov

Pre identifikáciu frekventovaných prvkov prúde dát, existuje viacero prístupov. Medzi základné radíme počítadlové, skicové a oknové prístupy, ktoré v tejto sekcii stručne opisujeme a diskutujeme z pohľadu vhodnosti pre náš problém.

2.1 Počítadlové prístupy a skicové prístupy

Prvé dve skupiny prístupov sú založené na postupnom prechádzaní dát a pamätaní si počtu nákupov jednotlivých položiek [3]. Oba prístupy a tiež jednotlivé ich algoritmy sa líšia spôsobom pamätania, udržiavania pamäte prípadne zabúdania nedôležitých prvkov, avšak princíp ostáva rovnaký. Trendy sú identifikované za celú históriu, neaktuálne prvky sú evidované s rovnakou dôležitosťou ako aktuálne. Nové trendy sa len ťažko a postupne presadia voči dlhodobým, ktoré sú už známe a teda ich v skutočnosti identifikovať vôbec netreba a podobne. Z tohto dôvodu pokladáme počítadlové a skicové prístupy za nie vhodné pre riešenie nášho problému včasnej identifikácie trendov v správaní sa používateľov online zľavového portálu.

Princípom počítadlových prístupov je vytvorenie určitého počtu počítadiel – dvojíc položka, početnosť (počet počítadiel je zadaný pevne, prípadne je odvodený z počtu existujúcich položiek). Po nákupe položky sa táto pridá medzi počítadlá, prípadne ak sa tam už nachádza, tak počítadlo inkrementuje. V závislosti od použitého

algoritmu sa líši akcia po nájdení novej položky. Pokiaľ sú všetky počítadlá obsadené inými položkami, napr. algoritmus Frequent dekrementuje všetky počítadlá, algoritmus Space-Saving nahradí najmenej frekventovanú položku, pričom zachová jej početnosť [8].

Skicové prístupy zefektívňujú princíp počítadiel tým, že jednotlivé položky ukladajú formou hešovacích tabuliek, čo zvyšuje pamäťovú efektívnosť a znižuje čas prístupu k položkám, čo má zmysel najmä pri spracovávaní prúdov veľkých dát. Medzi najznámejšie skicové algoritmy patria CountSketch, či CountMinSketch [3].

2.2 Oknové prístupy

Pokiaľ pri identifikácii trendov pracujeme len s aktuálnymi dátami, je vhodné použiť tretí prístup, algoritmy založené na oknách. V tomto prípade pracujeme len s obmedzenou podmnožinou najaktuálnejších dát, vďaka čomu je proces spracovania taktiež menej pamäťovo náročný. Existuje viacero algoritmov využívajúcich okná, ktoré sa líšia spôsobom tvorby okien, ich počtom a spôsobom udržiavania.

Najjednoduchším prístupom je skáčuce okno (Landmark Window), ktorý pracuje s dátami v okne od pevne daného časového medzníka. Okno sa postupne napĺňa dátami, v ktorých sa hľadajú trendy. Po dosiahnutí ďalšieho časového medzníka sa okno premaže a dáta sa začínajú zbierať odznova [6]. Výhodou tohto prístupu je, že okno netreba udržiavať a priebežne kontrolovať aktuálnosť dát, stačí len sledovať nastatie časového medzníka. Nevýhodou je, že po premazaní okna sú trendy identifikované na základe malého a nereprezentatívneho množstva dát.

Druhým prístupom je posuvné okno (Sliding Window), kedy sú trendy v ľubovoľnom čase identifikované z dát za určité stanovené obdobie (napr. posledný deň, hodina) alebo počet nakúpených položiek. Výhodou voči predchádzajúcemu prístupu je, že trendy sú vždy identifikované na základe dostatočného časového obdobia alebo počtu dát. Nevýhodou je, že o jednotlivých nákupoch je potrebné si pamätať čas ich vzniku, prípadne poradie, a po posune okna mimo ne ich odstrániť, čo vyžaduje určitú réžiu [5].

Medzi pokročilé oknové prístupy patrí napríklad algoritmus tlmené okno (Damped Window), ktorý je založený na sérii okien zachytávajúcich rozlične vzdialenú minulosť, pričom okná sú uvažované s rozličnou dôležitosťou klesajúcou smerom k starším dátam. Týmto spôsobom je možné simulovať proces zabúdania informácií v čase prípadne poklesu ich dôležitosti [4]. Tento algoritmus zachytáva viac informácií v porovnaní s predchádzajúcimi, avšak za cenu výrazne väčšej pamäťovej a výpočtovej réžie.

3 Návrh metód včasnej identifikácie trendov

Našim cieľom bola včasná identifikácia trendov v prúde dát zachytávajúcich nákupy používateľov online zľavového portálu. V rámci riešenia sme sa zaoberali využitím prístupu založeného na posuvnom okne, resp. skáčucom okne. V oboch prípadoch sme sa zamerali na také riešenie, ktoré bude výpočtovo efektívne (lineárna zložitosť).

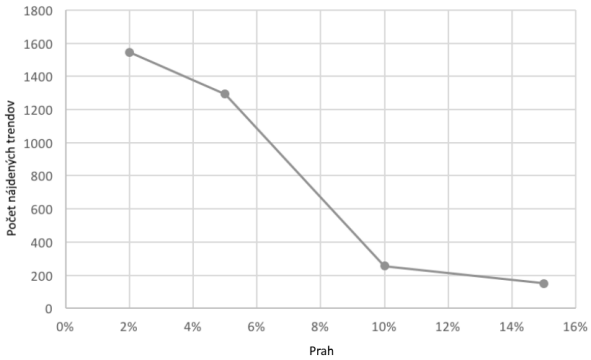
Použité metódy boli v prvej fáze inicializované nastavením vstupných parametrov metódy:

1. Nastavením prahu, pri ktorom nakupovanú položku pokladáme za trend. Presnejšie, sledujeme či podiel nákupov danej položky a celkového počtu nákupov prekonáva stanovený prah.
2. Stanovením veľkosti okna (pre posuvné okno) respektíve vzdialenosti míľnikov (pre skáčuce okno), v rámci ktorých identifikujeme v nákupoch trendy.

Po inicializácii metód sú tieto schopné postupne prijímať prúd dát a spracúvať ho nasledujúcim spôsobom. Pre každý spracovávaný záznam o nákupe:

1. Pokiaľ je časová pečiatka najstaršieho nákupu staršia ako hranica okna (vypočítaná zo spracovávaného záznamu) dekrementuj počítadlo pre túto položku a tú vymaž. Tento krok opakuj pokým nenájdeš najstaršiu položku, ktorú netreba vymazávať.
   a. V prípade skáčuceho okna je postup rovnaký, avšak najstaršia položka sa porovnáva voči poslednému míľniku pred aktuálnou položkou.
2. Inkrementuj počítadlo pre nakúpenú položku.
3. Identifikuj aktuálne trendové položky
   a. Tento krok nie je potrebné vykonávať vždy, ale len na požiadanie, čím sa zníži výpočtová náročnosť algoritmu.

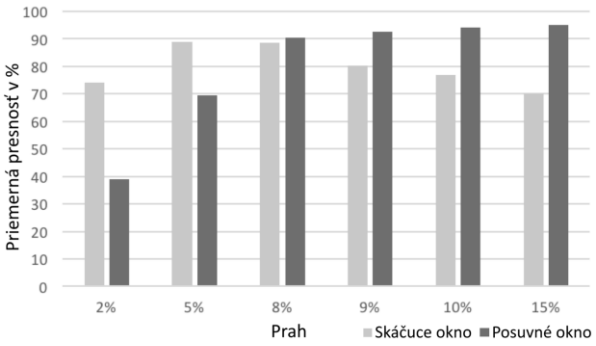
Pri inicializovaní algoritmov sme prah, kedy položku považujeme za trend, nastavili na základe pozorovania, kedy sme zistili, že existuje sigmoidálny pomer medzi hodnotou vzťahu a počtom nájdených trendov (Obr. 1). Na základe tohto zistenia sme experimentovali pri overovaní metódy s viacerými hodnotami zvoleného prahu. Pri vyššom nastavení totiž odfiltrujeme dostatok nezaujímavých položiek, pri nižšom naopak identifikujeme dostatočne širokú základnú položiek, ktoré môžeme neskôr používať pre ďalšie úlohy. Pre účely tohto príspevku sme však zafixovali veľkosť uvažovaného okna na 30 dní, čo zodpovedá typickej dobe ponuky zľavovej položky.



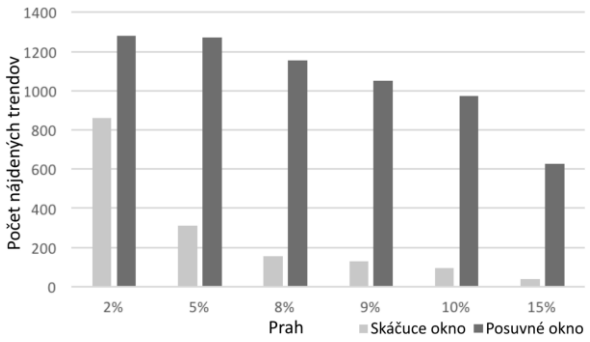
Obr. 1.Závislosť prahu nakupovanosti položky voči ostatným v pomere k počtu položiek (trendov) nakupovaných nad daný prah.

4 Overenie použitých metód

V rámci experimentálneho overenia sme identifikovali trendy v rámci datasetu zo zľavového portálu získaného ako súčasť projektu HIBER [2]. Pracovali sme s 1 mil. záznamov pochádzajúcim z obdobia 6 mesiacov. Tu sme sa zamerali na sledovanie presnosti identifikácie trendov (Obr. 2) a taktiež počet reálne nájdených trendov (Obr. 3). Ako môžeme vidieť, metóda využívajúca posuvné okno dokáže identifikovať výrazne viac trendov ako metóda využívajúca skáčuce okno. Tieto sú však identifikované s menšou presnosťou (okrem prípadov, kedy je prah nastavený na 5% a menej). Metóda skáčuceho okna naopak nedokáže identifikovať toľko trendov, jej presnosť je však vyššia, najme s rastúcim prahom akceptovateľnosti. Dôvodom je to, že táto metóda má k dispozícii menšie množstvo dát a teda na začiatku intervalu vymedzeného medzníkom dokáže definovať len skutočne výrazné trendy.



Obr. 2 Priemerná presnosť identifikácie trendov metódami využívajúcimi okná pri rozlične nastavenom prahu akceptovania trendov.



Obr. 3 Počet nájdených trendov pri rozlične nastavenom prahu akceptovania trendov.

5 Záver

V tomto príspevku sme sa zaoberali problémom včasnej identifikácie trendov v prúde dát. Dostatočne skoro identifikovaná informácia o tom, ktoré prvky budú v krátkej budúcnosti dobre predávané má vysokú hodnotu pre poskytovateľov webových portálov (prípadne akýchkoľvek predajní). Na základe takejto informácie totiž môžu nastaviť svoju predajnú politiky, naskladniť tovar alebo nastaviť výšku zliav.

Na riešenie problému efektívneho spracovania veľkého množstva dát, ktoré ostávajú aktuálne len po obmedzenú dobu, sme použili dva oknové algoritmy. Metóda posuvného okna bola schopná aj na obmedzenej množine dostupných dát identifikovať vysoké percento skutočných trendov. Na druhej strane však za trend označila aj množstvo dát, ktoré trendami neboli. Pri striktnejšom nastavení prahu, pri ktorom boli prvky označované za trend, sme však aj pri tejto metóde dokázali dosiahnuť presnosť 95%. Druhou použitou metódou bolo metóda skáčuceho okna, ktorá aj napriek menšiemu objemu dát, ktoré mala v priemere k dispozícii v porovnaní s predchádzajúcou metódou, dokázala odhaľovať trendy s vysokou presnosťou (70-89%). Počet nájdených trendov bol však pri tejto metóde malý a identifikované boli len veľmi výrazné položky. Jej výhodou je najmä jednoduché udržiavanie dát v okne, keďže stačí sledovať dosiahnutie časového míľnika a následne naraz vymazať všetky dáta. Z hľadiska užitočnosti pre úlohu včasnej identifikácie trendov, však pokladáme za vhodnejšie použiť metódu posuvného okna so striktne zadaným prahom, nakoľko týmto spôsobom dokážeme identifikovať dostatočné množstvo trendov s vysokou presnosťou ich identifikácie.

Literatúra

1. Aggarwal, C. C., Wang, J.: Data Streams: Models and Algorithms. Data Streams, 31, s. 9–38, 2007.
2. Bieliková M. a kol.: Projekt HIBER: hlbšie poznávanie správania sa človeka v digitálnom priestore, WIKT & DaZ 2016, s. 141 – 144, 2016.
3. Cormode, G., Hadjieleftheriou, M.: Finding the frequent items in streams of data, Communications of the ACM, 52(10), s. 97-105, 2009.
4. Gaber, M. M., Zaslavsky, a., Krish-naswamy, S.: Data Stream Mining overview. Data Mining and Knowledge Discovery Handbook, s. 759–787, 2009.
5. Giannella, C., Han, J., Yan, X., Yu, P. S.: Mining Frequent Patterns in Data Streams at Multiple Time Granularities. Next generation data mining, s. 191–212, 2003.
6. Golab, L., DeHaan, D., Demaine, E. D., Lopez-Ortiz, A., Munro, J. I.: Identifying frequent items in sliding windows over online packet streams. Proc. of the 2003 ACM SIGCOMM conference on Internet measurement - IMC '03, s. 173, 2003.
7. Krempl, G. a kol.: Open challenges for data stream mining research. SIGKDD Explor. Newsl. 16(1), s. 1-10, 2014
8. Metwally, A., El Abbadi, A.: Efficient Computation of Frequentand Top-k Elements in Data Streams s. 398–412, 2005.

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektov APVV-15-0508 a VG 1/0646/15. Autori článku chcú poďakovať Natálii Čulákovej, ktorej výskum a výsledky bakalárskej práce poslúžili ako základ pre tento článok.

Annotation:

An Early Identification of Trends within Behaviour of Online Discount Portal Users

The paper focuses on an early identification of trends within online discount portal. As the users' behaviour changes in time and they react differently to specific discount, our aim is to say in advance which items will be bought the most. The main challenge is the volume of the data. For this reason we process them as a stream and identify trends by window methods considering only selected subset of the most recent data. We show that in this way it is possible to identify the trends based on first days of their selling.

Towards User-friendly and High-performance Analytics with Big Data Historian

Martin Possolt[1], Václav Jirkovský[1], Marek Obitko[2]

[1]Czech Institute of Informatics, Robotics and Cybernetics
Czech Technical University in Prague
Zikova 4, Prague, Czech Republic
[2]Rockwell Automation Research and Development Center
Argentinská 1610/4, 170 00 Prague, Czech Republic

{martin.possolt,vaclav.jirkovsky}@cvut.cz
mobitko@ra.rockwell.com

Abstract. We are witnessing the trend of increasing data production in various domains including industrial automation. This trend requires means for data capturing, storing, and analyzing. Furthermore, a versatile data model is needed to enable easy knowledge representation as well as change management. In this paper, we utilize Semantic Big Data Historian, which can cope with previously mentioned requirements, for a demonstration of promising analytic approach combining Big Data methods and a user-friendly modular platform. The approach is demonstrated on data from a hydroelectric power station. The station has been dealing with the interesting problem of prediction when to momentarily stop their turbine to increase generated power after the restart. In this contribution, we discuss several approaches how to process and analyze data from power station sensors for achieving the best results.

Key words: Multilayer perceptron, Big Data, Ontology, Hydroelectric power station.

1 Introduction

Nowadays, we are witnessing the trend of increasing data production in various domains including industrial automation problem of data processing in industrial automation domain. This trend requires means for data capturing, storing, and analyzing. Many companies are facing the problem of processing this huge amount of data and a company capability to process data in the shortest time, to provide faultless processing, and to derive new previously unknown knowledge represents a competitive advantage. These data are produced by sensors and machines from a shop floor as well as by high-level systems, e.g. MES[1] and ERP[2].

[1] Manufacturing Execution Systems
[2] Enterprise Resource Planning

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 27-30.

The essential requirement is the proper understanding of given data models (sensors, machines, etc.) together with an understanding and a utilization of knowledge coming from various surrounding systems across a factory or external data sources. This requirement may be expressed as semantic integration problem [1], and the suitable solution is the employment of Semantic Web technologies and model description in ontologies [2].

We proposed and developed Semantic Big Data Historian (SBDH) [3] to cope with previously mentioned requirements. SBDH was designed to overcome common deficiencies of a legacy historian software which is usually optimized to allow fast and compressed storage of data. However, a legacy historian software does not pay much attention to analytics nor to heterogeneous data models integration. Thus, SBDH stores data according to a global data model in the form of ontology knowledge base and particular data are represented as RDF triples. This approach facilitates proper understanding of given data, and subsequently, it enables better data integration as well as data querying. On the other hand, a read/write time could be a bottleneck in the case of semantic data description and therefore SBDH stores RDF triples corresponding to time series in the form of Hybrid SBDH Model [4]. The historian architecture is divided into four layers – data acquisition (collects data from sensors and other related systems), transformation layer (transforms data to the unified semantic form according to the ontology), data storage layer (reads and writes data from storage – implemented by means of Apache Spark[3] and Apache Cassandra[4]), and analytic layer (represented by Apache Spark and KNIME[5]).

KNIME Analytics Platform is an open source data analysis toolbox. It offers various components for machine learning and data mining from data preprocessing, through modeling and data analysis to visualization. One of its biggest advantages is a graphical user interface which allows building a workflow very easily by connecting nodes, each with single operation task. There is no need to have any programming background. An example of a workflow will be illustrated later.

2 How to Analyze Big Data from Industrial Automation

The fast, high-quality, and versatile approach to coping with analytical tasks over production data may mean a significant advantage for many manufacturing companies. Nowadays, the Big Data paradigm and Cloud technologies become widely accepted by many companies and we can utilize such technologies for facilitating our task. Moreover, a processing of data stored in the form of ontologies (i.e. RDF triples) brings bigger demands for data handling. We have tried to overcome this issue by storing RDF triples in specialized structures (e.g. Hybrid SBDH Model [4]) but still RDF data handling is a very demanding task. Thus, we separated the processing of an analytical task into two layers – processing of all data with the help of Apache Spark;

[3] http://spark.apache.org
[4] http://cassandra.apache.org
[5] https://www.knime.org

and a utilization of KNIME for enabling the user-friendly analytic interface, where a user may conduct very complex task without any complex programming skills.

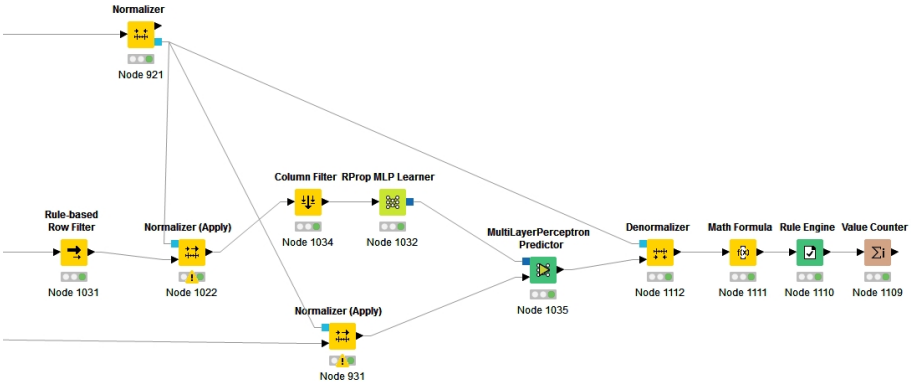
This two layers processing can be applied in different ways. First, in the case of ad-hoc task where the decomposition of the task should be near to the optimum for the best processing times, the prevalent part of processing is dedicated to the Spark layer (if it is reasonable from the nature of the task, i.e., degree of parallelism) and KNIME serves only as a presentation layer.

In the second case, SBDH contains set of pre-defined analytical methods for Apache Spark (e.g., filtering, application of various parameterized methods as computation of fast furrier transformation from this time interval and with a given time window). Then, KNIME is used for finishing preceding processing from Apache Spark and for providing visualization methods as well.

More detailed overview of the first approach to this two layer analytics is provided in [5].

3 KNIME for User Friendly Analytics

As it was mentioned, data analysis was tested on the data from the hydroelectric power station. There seems to be a problem with waste deposition on the blades of the turbine which causes a decrease of generated power. It is known from experience that if the turbine is momentarily stopped and restarted, the shock wave of water cleans the blades and increases the power to the normal level depending on other conditions. The power is dependent on water density and flow, turbine efficiency and water level difference before and after the station. Because the restart of the turbine is not very good for it, we analyze the data from station sensors in order to predict when it is desirable to stop it, i.e., when the effect of restart is greater than its abrasion.



Picture 1. Example of a workflow from KNIME

In the Picture 1 there is a screenshot of the workflow from KNIME. After some preprocessing the data from the whole year of operation are normalized and sent to MultiLayer Perceptron (MLP) Learner node where they are used as training data. The

neural network was then tested on the data from the following year (MLP Predictor node). The next nodes serve for the evaluation of the result. With this approach, we achieve over 80% accuracy of predicting of the right time to stop the turbine.

One of the reasons that the accuracy is not higher can be a small amount of training data. We had data from only a few years available (there were not many stops) and the difference between these years in the conditions can be substantial.

4 Conclusions

The requirements to improve the performance and versatility of analytics are pervading many domains including industrial automation domain. As in many cases, the trade-off between versatility (it is related also to user friendliness) and good performance has to be established.

In this paper, we shortly introduced the solution how to analyze big amount data with the help of SBDH and KNIME. KNIME does not allow to process big amounts of data, but it offers many pre-built analytic blocks for easy composing of a complex task and as we illustrated, can be used in combination with Big Data storage

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

References

1. A. Doan a A. Y. Halevy, „Semantic integration research in the database community: A brief survey," *AI magazine*, sv. 26, č. 1, p. 83, 2005.
2. M. Obitko a V. Mařík, „Integrating transportation ontologies using se-mantic web languages," v *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, Springer Berlin Heidelberg, 2005.
3. V. Jirkovský, M. Obitko a V. Mařík, „Understanding Data Heterogeneity in the Context of Cyber-Physical Systems Integration," *IEEE Transactions on Industrial Informatics*, pp. 660-667, 2017.
4. V. Jirkovský, M. Possolt a M. Obitko, „RDF Storage for Semantic Big Data Historian," v *Proceedings of WIKT & DaZ 2016*, Bratislava, Slovakia, 2016.
5. M. J., *Transformace uživatelských dotazů pro analýzu dat v průmyslové automatizaci*, BS thesis. České vysoké učení technické v Praze. Vypočetní a informační centrum., 2016.

Ontology Learning for Facilitating Ontology Matching in Automotive

Ondrej Šebek[1], Václav Jirkovský[1], Nestor Rychtyckyj[2], Petr Kadera[1]

[1]Czech Institute of Robotics, Informatics, and Cybernetics
Czech Technical University in Prague,
Zikova 4, 166 36 Prague, Czech Republic
[2]Ford Motor Company
Dearborn, MI, USA

{ondrej.sebek,vaclav.jirkovsky,petr.kadera}@ciirc.cvut.cz
nrychtyc@ford.com

Abstract. All manufacturing companies need to monitor a large number of devices and from which critical data must be captured and analyzed. The increasing complexity of these ecosystems emphasizes the requirement for a flexible and versatile data model architecture. Ontologies may facilitate a proper understanding of the problem domain as well as the interoperability with surrounding systems using ontology matching approach. However, data models of surrounding systems are not always ontologies. Thus, concepts and relations among them have to be extracted from the models to enable their integration with the ontology. The definition of concepts, their hierarchy, relations between concepts, and properties from a general architecture is a complex task and has to be tailored to an application's needs. In this paper, we propose an involvement of the ontology learning approach to the process of ontology matching in the automotive.

Keywords: Ontology, Heterogeneity, Ontology Learning, Ontology Matching,

1 Introduction

All manufacturing companies need to monitor a large number of devices from which critical data must be captured and analyzed. The increasing complexity of these ecosystems emphasizes the requirement for a flexible and versatile data model architecture. Furthermore, common data models are becoming insufficient for conducting analytical tasks due to the systems complexity and corresponding exacting integration of new devices due to a complicated understanding of system data model.

Thus, the essential requirement is the proper understanding of given data models (sensors, machines, etc.) for ensuring a faultless processing of a huge amount of data. Moreover, the solution should also allow for easy maintainability as there will be frequent additions and modifications to the data model. The mentioned semantic integration problem [1], as well as the problem of easy maintainability, may be solved by

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 31-34.

the employment of Semantic Web technologies and model description in ontologies [2]. This approach also supports knowledge expressiveness and reasoning as well as the ability to keep track of the source of each data item.

The Semantic Web technologies were originally intended for the representation of various highly heterogeneous data models (e.g. web page metadata) within the Internet. Thus, these technologies were developed mainly for facilitating an integration of various data models. This research branch is named ontology matching and mapping [3] and many promising methods have already been developed. On the other hand, many of these methods were proposed as fully-automatic approaches – mainly for the integration of large ontologies. This approach is not suitable for all domains because it is required to achieve the highest possible precision in domains such as automotive, medicine, etc. In our previous work, MAPSOM [4] framework was developed for semi-automatic ontology matching based on Kohonen's self-organizing map[1] and active learning[2].

In this paper, we introduce a utilization of MAPSOM framework for an integration of heterogeneous data models in automotive. Within the framework of this work, we utilized our previous experiences with the development of manufacturing ontologies and will be building upon those ontologies in this work [5][6]. We provide a short overview of the problem of matching MS Excel sheet containing data about Ford spare parts to Ford Supply Chain ontology. The main obstacle in such matching may be found mainly in curtness and hard meaning understanding of data descriptions in catalogs.

If we would like to integrate data sources which are unstructured (e.g. text files, etc.), then we may exploit ontology learning [7] algorithms to derive new relations, concepts, and relations among entities. On the other hand, this approach has only limited applications in the case of structured data sources such as catalogs – there are no explicitly defined relations among entities, and there are typically no suitable additional text data for a widespread utilization of ontology learning algorithms. Nevertheless, we briefly describe conducted experiments which demonstrate that even limited utilization of ontology learning approaches may improve previously mentioned ontology matching task.

2 Ontology Learning for Facilitating Ontology Matching

The goal of ontology matching is to find correspondent entities expressed in different ontologies. The subsequent goal is an enrichment of captured knowledge in the first ontology by knowledge from the second one and vice versa. In this paper, we use a hybrid matching system prototype [8] which is responsible for matching elements from an MS Excel file (representing Ford spare parts) to the Ford Supply Chain ontology.

The Ford supply chain ontology captures the risk managements in the Ford global supply chain – every car model depends on many different suppliers, and important

[1] https://en.wikipedia.org/wiki/Self-organizing_map

[2] https://en.wikipedia.org/wiki/Active_learning_(machine_learning)

capability is to be able to determine which vehicles at which plant would be impacted by a potential shortage. The XLS file contains Ford spare part records and has about 62 various columns identifying particular parts. A predominant number of columns contain specific numerical codes or strings composed of abbreviated labels. A manual integration of such a data would be very exacting because of the big volume of records and their attributes.

The interesting tasks such as data preprocessing, searching acronyms in external vocabularies, and a process of elements matching are not mentioned in this paper due to the scope limitations, but it is presented in [8].

The outcome of common matching process results in the situation when some of the spare part records are mapped to individuals of the concept “BPNO” – base part number object (e.g. BPNO6C358) or to the concept “Part”. Unfortunately, many correspondent individuals have no additional properties (i.e., data and object properties) in the case of “BPNO”. In this situation, we have only limited understanding of a meaning of records gained after matching. Similarly, records mapped to the concept “Part” are mapped only to this concept and any other specialized sub-concepts are missing for some records. For example, there are no specialized sub-concepts like “break” or “crankshaft”, etc.

Thus, we need to derive new specialized sub-concepts or more detailed relationships between concepts for example with the help of ontology learning methods. These methods are usually applied to, for example, text documents (manuals) containing required information. In this task, the question is what information should be used for ontology learning task – part number, part description, and what else?

As the first approach, the base part number categories may be used for deriving new concepts and their relations. In our experiments, we used WordNet dictionary where hyponyms/hypernyms and holonyms/meronyms information may be found.

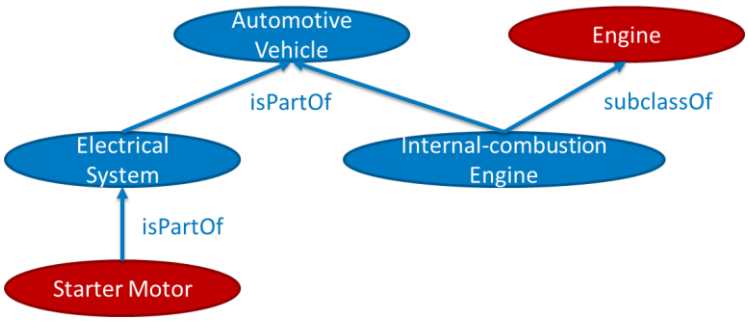


Figure 1. Deriving relationship between the concept "Starter Motor" and the concept "Engine"

Fig. 1 illustrates how the relationship may be found between “Starter Motor” and “Engine”. There is not only the new relationship but also new derived concepts such as “Electrical System”, “Automotive Vehicle”, and “Internal-combustion Engine”.

The main disadvantage is that there should be user verification (there are many various candidate relationships) as well as the search space is very extensive in some cases.

3 Conclusions

In this paper, we have shortly introduced how new concepts, as well as relationships between them, may be discovered even with very limited sources.

The future work will reside in deriving a more detailed/complex concept structure and relations with the help of external sources (text files – manuals, etc.) and part description in the form of abbreviated terms from the spare parts Excel sheet.

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

References

1. A. Doan and A. Y. Halevy, "Semantic integration research in the database community: A brief survey," *AI magazine*, vol. 26, no. 1, p. 83, 2005.
2. M. Obitko and V. Mařík, "Integrating transportation ontologies using semantic web languages," in *International Conference on Industrial Applications of Holonic and Multi-Agent Systems*, Springer Berlin Heidelberg, 2005.
3. J. Euzenat and P. Shvaiko, Ontology matching, Heidelberg: Springer, 2007.
4. V. Jirkovský and R. Ichise, "Mapsom: User involvement in ontology matching," in *Joint International Semantic Technology Conference*, Seoul, 2013.
5. D. Ostrowski, N. Rychtyckyj, P. MacNeille and M. Kim, "Integration of big data using semantic web technologies," in *IEEE Tenth International Conference on Semantic Computing (ICSC)*, 2016, 2016.
6. N. Rychtyckyj, V. Raman, B. Sankaranarayanan, P. S. Kumar and D. Khemani, "Ontology Reengineering: A Case Study from the Automotive Industry," *AI Magazine*, vol. 38, no. 1, 2017.
7. A. Maedche and S. Staab, "Ontology learning," in *Handbook on ontologies*, Springer Berlin Heidelberg, 2004, pp. 173-190.
8. V. Jirkovský, P. Kadera and N. Rychtyckyj, "Semi-Automatic Ontology Matching Approach," in *HoloMAS 2017*, Lyon, (forthcoming 2017).

Aplikační příspěvky

Minimal Transportation Disruptions Model and Ontologies for Modelling of Disruptive Events

Josef Petrák

Some move

me@jspetrak.name

Abstract. Carriers and operators seek effective ways to distribute structured information about transportation disruptions to passengers, drivers and other users of the infrastructure. Transit Alerts, a conceptual application from *Some move*, utilizes all available means of distribution to deliver timely updates about disruptive events. Apart from REST resources and JavaScript Web API, an interlinked repository is proposed to support intelligent assistive technologies. Several ontologies were compared to model the events and share the information from the application repository to third parties in a vendor-independent schema.

Keywords: Transportation Disruptions, Linked Data, Advanced Traveller Information System, ATIS, Information System Integration

1 Introduction

Public transportation is a key economic factor that enables workforce mobility and development of smart cities. Buses, trains and other transportation modes, however, face both scheduled and unexpected disruptive events that negatively impact availability, attractiveness and quality of the service. Carries need adequate means to integrate reports from various sources, inform travellers and staff, and accurately deliver updates with minimal delays and inaccuracies. Furthermore, they strive to analyse, predict and avoid such events in the future. Interoperability is the main challenge for integrating existing reporting systems, as well as devices and facilities used to present updates and responsive measures to stakeholders.

The carriers' immediate concern is to address the distribution of such events. Dispatchers receive unstructured voice or textual information reporting a new disruption or an update on an existing one. Upon acknowledgement, a comprehensive description is compiled and distributed over voice or text to on-board personnel and passengers. Quality of the information process depends on multiple factors:

1. Availability of input information,
2. Speed of reaction,
3. Implemented mitigation measures and quality of information provided to recipients (passengers, passenger-facing employees)
4. Distribution to information channels followed by the intended audience

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 36-40.

Information systems addressing this process must integrate data with different textual formats – extract from plain or RDFa-annotated HTML, XML, JSON[1], RESTful services [3], SOAP, GraphQL[2] or proprietary structured formats – and integrate them with voice information, geospatial data detailing specific points or sections within the transportation network, as well as the current position of affected vehicles. The same integration requirements are applied to distribution channels where third-party distribution systems or devices expose different interfaces to present information.

Transit Alerts is an information system designed in a way that addresses seamless integration of reporting channels and distribution systems in a single dispatcher interface. An internal data model is used across the system. Inbound reports are mapped into the internal data model and outbound messages are transformed to a format that third-party distribution systems understand, assuming they cannot consume the internal data model directly.

Addressing interoperability is the challenge that Semantic Web and Linked Data overcome by nature. When a data model is shared not only by internal subsystems but also exposed to third parties, it enables them to query, reason and analyse structured information from the system.

2 Information System Model

While designing the system, existing information sources[3] were reviewed to define sets of information currently shared with stakeholders, to be designated as required or optional. Reporting structures from major Czech and Slovak rail and bus operators were cross-referenced to determine common denominators and identify a minimum set of attributes required to conceivably report a traffic disruption. *Table 1* provides an overview of the carriers' reporting templates.

Tab. 1. Required (R) and optional (O) attributes of disruptions published by studied carriers

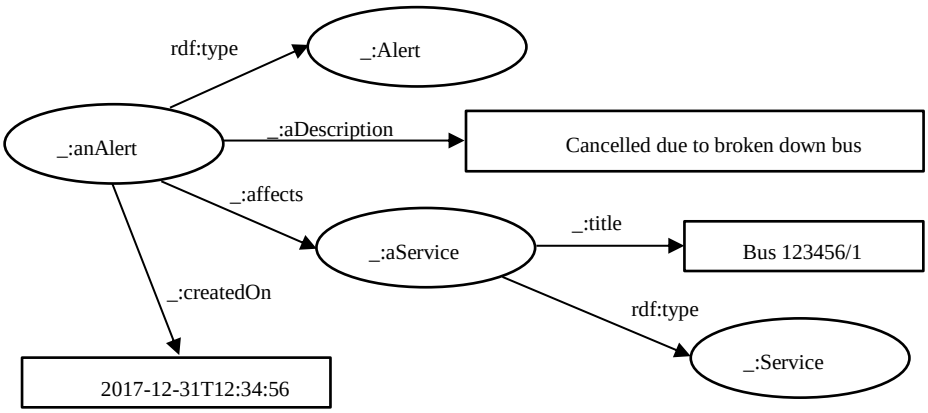
Information	ČD	RJ	LE	ARRIVA	ÖBB	ZSSK
Route	O	O	O	O	O	O
Service	O	O	R	R	O	R
Timestamp From	R	R	R	R	R	O
Timestamp To	R	O	O	O	O	O
Description	R	R	R	R	R	R
Event Type	R	O	R	O	R	R
Measure	O	O	O	O	O	O

[1] ECMA-404 The JSON Data Interchange Format

[2] http://graphql.org

[3] www.cd.cz/mimo, www.le.cz (Actual traffic restrictions), www.regiojet.com/en/news-from-railway-network/, www.arriva-vlaky.cz/jizdni-rad/mimoradnosti-a-zmeny/, www.arrivaexpress.cz/zmeny-v-provozu/ to name a few.

As a result of this analysis, the designed internal data schema consists of at least three required attributes, which are key to providing actionable information: services affected by the event, unstructured textual description and a timestamp of occurrence. Optional information includes estimated time of resolution, root cause, geospatial location of the event within the transportation network and implemented mitigation measures (non-planned rotation of vehicles, substitution services etc.).



Picture. 1. RDF graph for a sample alert (disruptive event)

None of the sources examined provide information in any semantic format, as plain HTML is generally used instead. Respecting the Linked Data best practices[4], we looked for an existing RDF schema or ontology capable of expressing both minimal and structured information about disruptive events in different modes of transportation. Compatibility with ontologies that describe timetables, geospatial information or network information was considered as an advantage.

3 Ontologies Examined

GTFS[5] has been adopted by various systems to process timetable information in a format interoperable with Google Maps and Open Trip Planner [5]. Linked GTFS[6] specifies the feed vocabulary for RDF. GTFS Realtime[7] enables the systems to push trip updates, service alerts or vehicle positions, though this extension is not covered by an RDF schema.

[4] http://www.w3.org/TR/ld-bp
[5] http://gtfs.org
[6] http://vocab.gtfs.org
[7] https://developers.google.com/transit/gtfs-realtime/

The DATEX[8] standard was developed for information exchange between traffic management centres, traffic information centres and service providers in Europe. Its schema[9] specifically covers road transportation.

Service Interface for Real Time Information (SIRI)[10] is a CEN protocol to distribute real time information about public transportation services. This standard was designed for real-time data distribution and utilizes SOAP messaging over web services.

The Transport Disruption Ontology [2] formalizes a framework for modelling of disruptive events regardless of transportation mode. Other ontologies like FOAF, DC or LinkedGeoData are compatible with TD ontology. A general `td:DisruptiveEvent` class represents any event though TD ontology and provides several possible event types.

```
_:aService
  a sj:Service;
  dc:title "Bus 123456/1".

_:aPlaceOfFailure
  a geo:SpatialThing;
  geo:lat "50.076005";
  geo:lon "14.4304176".

_:aPlan
  a td:Plan;
  transit:service _:aService.

_:anAlert
  a td:DisruptiveEvent;
  dt:causeOf td: BrokenDownBus;
  dc:description "Cancelled due to broken down bus";
  event:time [ tl:beginsAtDateTime "2017-12-
31T12:34:56"^^xsd:dateTime ];
  event:place _:aPlaceOfFailure;
  td:impactsOn _:aPlan.
```

Figure 1: Sample of a discruptive event information about a broken bus

Whereas *Picture 1* describes a general RDF graph of a single disruptive event, *Figure 1* demonstrates an example with an applied TD ontology combined with the Social Journeys[11] ontology to express the impacted bus service. This listing defines a public bus service and an alert that the service is impacted by a vehicle failure at known GPS coordinates and a known point in time.

[8] http://www.datex2.eu
[9] http://vocab.datex.org/terms
[10] http://www.siri.org.uk
[11] https://github.com/SocialJourneys/SocialJourneysOntologies

The above information enables profound analysis, such as identifying the most failure-prone services, vehicles or critical points in the transportation network. Having further information about the measures implemented by dispatchers, an analysis of measure effectiveness, as well as automated recommendations is possible.

4 Conclusion

Transport Disruption Ontology[12] has been selected to express internal information in the Linked Data Repository. Apart from the integration of RDF-consuming third-party systems, an expert system may be implemented.

Network managers can identify faulty infrastructure points; carriers and transportation planners may use such data to improve timetables and service reliability, and avoid points of frequent congestion. Smart travel planning systems can query the data source to provide estimates on on-time arrival or plan for alternative routes once real-time data is available.

References

1. Colpaert, P., Llaves, A., Verborgh, R., Corcho, O., Mannens, E., Van de Walle, R.: Intermodal public transit routing using Linked Connections. In: Proc. of International Semantic Web Conference – ISWC 2015.
2. Corsar D., Markovic M., Edwards P., Nelson J. D.: The Transport Disruption Ontology. In: Proc. of International Semantic Web Conference – ISWC 2015.
3. Fielding, R.T.: Architectural Styles and the Design of Network-based Software Architectures. University of California, Urvine, 2000.
4. Grant-Miller, S. M., Gal-Tzur A., Minkov E., Nocera S., Kuflik, T., Shoor I.: Enhancing transportation data collection through social media sources: methods, challenges and opportunities for textual data. The IET Inteligent Transport Systems Journal of Institution of Engineering Technology (2014).

[12] https://transportdisruption.github.io/

Porovnanie algoritmov na analýzu sekvencií pohľadu

Róbert Móro, Michal Melúch, Martin Mokrý, Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva
Fakulta informatiky a informačných technológií
Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

{robert.moro, xmeluch, xmokry, maria.bielikova}@stuba.sk

Abstrakt. Sledovanie pohľadu si v súčasnosti nachádza uplatnenie v rôznych oblastiach v praxi (napr. pri testovaní použiteľnosti aplikácií), ako aj vo výskume, kde je využiteľné ako jeden zo zdrojov implicitnej spätnej väzby pri modelovaní interakcie a vlastností používateľa. V tomto príspevku sa venujeme porovnaniu existujúcich algoritmov na analýzu sekvencií pohľadu, ktorá slúži na hľadanie opakujúcich sa vzorov, umožňuje identifikovať rôzne stratégie používateľov alebo ich skupiny pri interakcii s Webom. Vytvorili sme systém, ktorý spracúva dáta zo sledovania pohľadu a poskytuje prostredie pre porovnanie týchto algoritmov na rôznych dátových množinách. V práci opisujeme doteraz nazbieranú dátovú množinu, identifikované problémy existujúcich algoritmov a možnosti ich ďalšieho vylepšenia.

Kľúčové slová: sekvencia pohľadu, sledovanie pohľadu, oblasti záujmu, algoritmy, testovanie použiteľnosti, UX.

1 Existujúce spôsoby spracovania pohľadu

Sledovanie pohľadu sa využíva pri testovaní použiteľnosti a používateľského zážitku (angl. user experience, UX) aplikácií. Existujúce práce sa zameriavajú na automatické odhaľovanie chýb použiteľnosti v rozhraniach [2], sľubným smerom je tiež výskum individuálnych rozdielov medzi používateľmi pri interakcii s počítačom, ktoré spočívajú napr. v ich kognitívnych vlastnostiach (rýchlosť vnímania, veľkosť pracovnej pamäte a i.) [9], kognitívnej záťaži [8] alebo vizuálnom hľadaní [1] a využitie týchto rozdielov pri modelovaní používateľov a prispôsobovaní rozhraní ich potrebám.

Spracovanie pohľadu prebieha vo viacerých krokoch: súradnice pohľadu na obrazovke sú spracúvané na tzv. fixácie (miesta spracovania vizuálneho vnemu, kedy je oko statické; trvá typicky okolo 200 až 300 ms) a sakády (rýchle pohyby oka medzi jednotlivými fixáciami) [7]. Takto predspracované dáta je možné kvantifikovať v podobe metrík, ako je napr. počet fixácií, priemerná dĺžka fixácie či uhol medzi po sebe idúcimi sakádami [2]. Často sú metriky počítané nie pre celý vizuálny podnet (napr. webovú stránku), ale pre konkrétne oblasti záujmu (napr. hlavička, päta stránky, menu). Rozdielne hodnoty metrík indikujú rôzne problémy pri interakcii (napr.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 41-45.

väčšia dĺžka fixácií indikuje ťažkosti pri spracovaní a pochopení informácie) [2]. Metriky založené na fixáciách a sakádach však nepostihujú komplexnejšie vzory správania používateľov; na tento účel sa využívajú *sekvencie pohľadu*, ktoré zachytávajú aj temporálne hľadisko (postupnosť fixácií v čase).

Na analýzu sekvencií pohľadu sme sa zamerali aj v tejto práci. Za účelom porovnania existujúcich algoritmov sme vytvorili systém, ktorý umožňuje výskumníkom spracovať dáta zo sledovania pohľadu, aplikovať na ne zvolené algoritmy analýzy sekvencií pohľadu, vizualizovať a exportovať výsledky.

2 Analýza sekvencií pohľadu

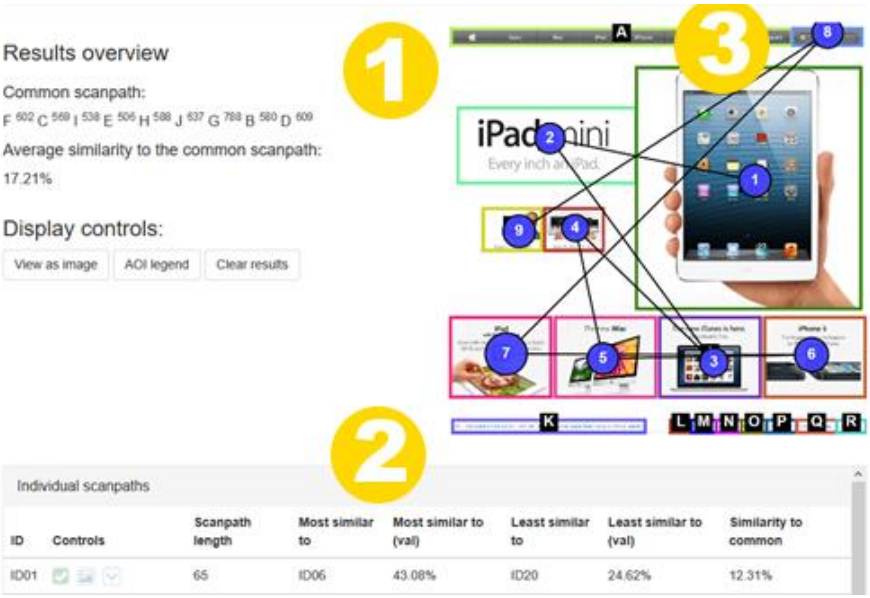
Pri analýze sekvencií pohľadu rozlišujeme viacero prístupov, my sme sa v práci zamerali na algoritmy na *identifikáciu spoločnej sekvencie*, ktoré agregujú sekvencie pohľadu jednotlivých používateľov do jednej, predstavujúcej „typický" prechod pohľadu po obrazovke [4]. Cieľom je identifikovať takú spoločnú sekvenciu, ktorá sa čo najviac podobá všetkým individuálnym sekvenciám pohľadu.

Tieto prístupy sa väčšinou skladajú z dvoch krokov: v prvom kroku sa identifikuje dvojica najviac podobných sekvencií, v druhom kroku sa táto dvojica nahradí ich spoločnou reprezentáciou; toto sa iteratívne opakuje, dokým nezostane len jedna spoločná sekvencia. Na tomto princípe pracujú napr. algoritmy eMINE [3] a Dotplot [6]; líšia sa v konkrétnom použitom algoritme na výpočet podobnosti (Levenstheinova vzdialenosť v prvom prípade a algoritmus založený na tzv. bodkovom diagrame v druhom) a v spoločnej reprezentácii tejto dvojice (najväčšia spoločná podpostupnosť v prípade algoritmu eMINE a najdlhšia diagonála v bodkovom grafe v prípade Dotplots algoritmu). Iný pokročilejší prístup predstavuje algoritmus STA [5], ktorý je založený na identifikácii najčastejšie fixovaných oblastí a mal by byť robustnejší voči prípadnému šumu v dátach.

Všetky tieto tri algoritmy sme poskytli v rámci implementácie nami navrhnutého systému na analýzu sekvencií pohľadu. Implementácia STA bola priamo od autorov algoritmu, eMINE sme implementovali podľa opisu v [3] a Dotplots podľa [4]. Používateľ môže do systému nahrať dáta zo sledovania pohľadu; momentálne podporujeme formát dát zo systému Tobii Studio[1], ktorý je súčasťou softvérového vybavenia *Výskumného centra používateľského zážitku a interakcie*[2] na našej fakulte. Okrem fixácií používateľ potrebuje nahrať definíciu oblastí záujmu a samotné vizuálne stimuly pre potreby vizualizácie. Nahraté dáta sú následne spracované do jednotlivých sekvencií pohľadu, ktoré môže používateľ jednotlivo skúmať a vizualizovať pomocou štandardného grafu sekvencií pohľadu. Pre zvolené jednotlivé sekvencie môže následne vypočítať spoločnú sekvenciu; systém prezentuje priemernú mieru podobnosti spoločnej sekvencie voči všetkým individuálnym (na základe Levenshteinovej vzdialenosti), ako aj podobnosť pre jednotlivé sekvencie zvlášť (pozri Obr. 1). Okrem toho je možné zobraziť prehľad porovnávajúci všetky systémom podporované algoritmy

[1] https://www.tobiipro.com/product-listing/tobii-pro-studio
[2] http://uxi.sk

naraz. Vypočítané sekvencie a ich podobnosti je možné stiahnuť v CSV formáte pre ďalšiu analýzu vo zvolenom štatistickom softvéri.



Obr. 1. Snímok obrazovky systému na analýzu sekvencií pohľadu. Systém počíta spoločnú sekvenciu na základe zvoleného algoritmu (1), pričom používateľ môže prehliadať aj individuálne sekvencie a zvoliť, ktoré sa majú pri výpočte spoločnej sekvencie použiť (2). Tieto si vie vizualizovať pomocou grafu sekvencií pohľadu (3).

3 Overenie a ďalšie smerovanie

Overenie realizovaného systému sme vykonali v dvoch krokoch. Najprv sme overili správnosť implementácie algoritmov na dátovej množine prezentovanej v [5], pričom sme dosiahli vo väčšine prípadov totožné výsledky. Následne sme tieto algoritmy aplikovali na nami zozbieranej dátovej sade – išlo o záznam interakcie 53 používateľov stránky FIIT STU[3], ktorých úlohou bolo nájsť na stránke odkaz na program Dňa otvorených dverí, pričom na nahrávanie pohľadu sme použili zariadenie Tobii X2-60. Vzhľadom na to, že účastníci experimentu neboli časovo obmedzení pri vykonávaní úlohy, sú medzi dĺžkami sekvencií značné rozdiely; maximálna podobnosť medzi dvojicou sekvencií v dátovej množine bola 66,67%, minimálna 0. Vzhľadom na túto variabilitu neboli algoritmy eMINE ani Dotplot schopné nájsť žiadnu spoločnú sekvenciu. Jediným úspešným algoritmom tak bol STA, ktorý mal priemernú podobnosť 19,7% (so štd. odchýlkou 6,53%). Tieto čísla nie sú uspokojivé; presnosť by sme mohli zvýšiť odstránením používateľov s príliš nepodobnými sekvenciami, prípadne tých, ktorým sa danú úlohu nepodarilo splniť. Iné riešenie spočíva v aplikácii algorit-

[3] http://www.fiit.stuba.sk

mu zhlukovania na identifikáciu skupín podobných sekvencií, ktoré môžu reprezentovať rôzne stratégie riešenia úloh (aj v prípade testovanej úlohy bolo možné nájsť cieľový odkaz na dvoch miestach, a to v menu aj v príspevku uverejnenom na stránke); existujúce prístupy síce využívajú metódy zhlukovania, ale väčšinou s cieľom skonštruovať jednu spoločnú sekvenciu (pozri napr. [5] a [6]).

Do budúcnosti plánujeme ďalšie experimenty s týmito algoritmami – jednak ich modifikáciu, aby boli robustnejšie na odchýlky v sekvenciách pohľadu (odstránením odchýlok v kroku predspracovania, resp. modifikáciou funkcie na výpočet podobnosti a spoločnej reprezentácie sekvencií), ako aj ich aplikáciu na ďalšie dátové sady nazbierané v rámci výskumného centra v spolupráci s Katedrou psychológie na FiF UK.

Pracujeme tiež na využití analýzy sekvencií pohľadu pri určovaní miery oboznámenosti používateľa s webovou stránkou. Zrealizovali sme prvotný experiment so 14 účastníkmi, ktorí plnili rôzne úlohy na stránke vybraného elektronického obchodu, pričom sme sledovali mieru ich webovej gramotnosti, ako aj oboznámenosti s danou stránkou; vyhodnotenie experimentu a otestovanie väčšieho počtu účastníkov v rámci neho predstavuje našu budúcu prácu.

Literatúra

1. Dragunova, M., Moro, R., Bielikova, M.: Measuring Visual Search Ability on the Web. In: Proc. of the 22nd Int. Conf. on Intelligent User Interfaces Companion - IUI ’17 Companion, ACM Press, NY, USA, (2017), 97–100.
2. Ehmke, C., Wilson, S.: Indentifying Web Usability Problems from Eye-Tracking Data. In: Proc. of the 21st British HCI Group Annual Conf. on People and Computers - BCS-HCI ’07, BCS Learning & Development Ltd., Swindon, UK, (2007), 119 128.
3. Eraslan, S., Yesilada, Y., Harper, S.: Identifying patterns in eyetracking scanpaths in terms of visual elements of web pages. In: Proc. of the 14th Int. Conf. on Web Engineering - ICWE 2014, Springer International Publishing, (2014), 163 180.
4. Eraslan, S., Yesilada, Y., Harper, S.: Eye Tracking Scanpath Analysis Techniques on Web Pages: A Survey, Evaluation and Comparison. J. of Eye Movement Res. 9(1) (2016) 1 19.
5. Eraslan, S., Yesilada, Y., Harper, S.: Scanpath Trend Analysis on Web Pages: Clustering Eye Tracking Scanpaths. ACM Transactions on the Web 10(4) (2016) 1 35.
6. Goldberg, J.H., Helfman, J.I.: Scanpath clustering and aggregation. In: Proc. of the 2010 Symposium on Eye-Tracking Research & Applications - ETRA ’10, (2010), 227 234.
7. Holmqvist, K. et al.: Eye tracking: A comprehensive guide to methods and measures. Oxford University Press, 2011.
8. Juhaniak, T., Hlavac, P., Moro, R., Simko, J., Bielikova, M.: Pupillary Response: Removing Screen Luminosity Effects for Clearer Implicit Feedback. In: UMAP 2016: Posters, Demos, Late-breaking Results and Workshop Proc. of the 24rd Conf. on User Modeling, Adaptation, and Personalization. CEUR-WS, Aachen, (2016), 2.
9. Steichen, B., Carenini, G., Conati, C.: User-adaptive information visualization - Using eye gaze data to infer visualization tasks and user cognitive abilities. In: Proc. of the 18th Int. Conf. on Intelligent User Interfaces - IUI ’13, ACM Press, NY, USA, (2013), 317 328.

Poďakovanie: Tento článok vznikol vďaka čiastočnej podpore projektu APVV-15-0508 a vďaka podpore v rámci projektu „Rozvoj výskumnej infraštruktúry STU”,

projekt č. 003STU-2-3/2016 zo zdrojov štátneho rozpočtu prostredníctvom dotácie z Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky.

Annotation:

The Comparison of the Algorithms for Scanpath Analysis

Eye tracking is currently applied in various areas of industry (e.g., for usability testing of applications) as well as of research, where it can be utilized as one of the sources of implicit feedback for modeling of interaction and user characteristics. In our work, we focus on a comparison of the existing algorithms of scanpath analysis that is used for identification of recurring interaction patterns, different user strategies or groups of similar users based on their interaction on the Web. We developed a system that processes eye tracking data and provides an environment for algorithms comparison on various datasets. In the paper, we describe a dataset that we collected, problems with the existing algorithms that we identified and possibilities for their improvement.

OLAP Recommender: Supporting Navigation in OLAP Cubes Using Association Rule Mining

Bohuslav Koukal[1], David Chudán[2], Vojtěch Svátek[3]

Department of Information and Knowledge Engineering, Faculty of Informatics and Statistics
University of Economics, Prague
nám. W. Churchilla 1938/4, 130 67 Praha 3

[1]bohuslav.koukal@yahoo.co.uk
[2]xchud01@vse.cz
[3]svatek@vse.cz

Abstract: The OLAP Recommender tool automates the multidimensional data exploration process and recommends potentially interesting views on the data to the user. It integrates two data analytics methods – OLAP visualisation and data mining, the latter being represented by GUHA association rule mining. Algorithms implemented in the tool include automated data discretization, setup of dimensions' commensurability, automatic design of the data mining task based on the data structure, and mapping between the mined association rules and the corresponding OLAP visualisation. The system was tested with real retail data and with EU structural funds data. The experiments indicate that complementary usage of association rule mining and OLAP analysis identifies relationships in the data with higher success rate than the isolated use of both techniques.

Keywords: OLAP navigation, GUHA association rules, OLAP Recommender, guided analytics

1 Introduction

Self-service Business Intelligence and visualisation tools, operating over multidimensional data cubes specifying the target numerical values, *measures*, using combinations of *dimensions*, are today widely used for reporting, creating dashboards and answering the users' questions about their data. However, manual browsing and visualising the data brings some obvious problems. The user usually has certain prior understanding of the problem area, which guides him/her into the portion of the data s/he considers interesting. If there are other interesting portions in the data that s/he does not know about, it is very hard or even impossible to discover them. Another problem is a lack of completeness – it is impossible for a human to manually identify all interesting relationships in large multidimensional data just by browsing the data visualisations.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 46-50.

OLAP Recommender copes with these problems by recommending interesting views on multidimensional data, thus revealing strong relationships (indicating both trends and abnormalities) identified in the data. Compared to other proposals to combine data mining and OLAP [1, 2, 3, 4], it is designed to find these relationships in more specific subsets of the data and offers the possibility of extending to full potential of GUHA association rules. In contrast to general concept of recommending systems, OLAP Recommender is not based on the users' behaviour but solely on the relationships found in the data.

2 Searching for strong relationships in the data

2.1 GUHA association rules

For identifying strong and potentially interesting relationships in multidimensional data, OLAP Recommender uses association rule mining, specifically, the GUHA 4-ft method, described by Rauch and Šimůnek [5]. GUHA association rules can be considered as an extension and generalization of the association rules, originally introduced by Agrawal et al. [6]. As an addition to the traditional association rules, GUHA offers three important extensions useful for mining multidimensional data. Firstly, GUHA offers seventeen different interest measures in comparison to two (confidence and support) defined for the original association rules. Then, GUHA offers conditional association rules, making it possible to mine the rules in a subset of the data, while still maintaining the interest measures in relation to the whole dataset. Last, but not least, GUHA offers different types of coefficients (attribute value restrictions in the hypotheses) – e.g. multiple elements subsets, sequences or cuts, while traditional rules can use only a single-element subset as a coefficient.

2.2 Mining multidimensional aggregate data

Association rules were originally introduced for transactional data (market basket analysis) [6]. Chudán [7] introduced a different approach for using the association rules for mining aggregate data and based on his work, Koukal defined additional steps needed for automatically setting up the association rule mining task for the OLAP data [8]:

1. Measures discretization: Based on our experiments we proposed to discretize the measures to equal frequency intervals. We can discretize quite finely (to 20 and more intervals, based on the data volume), taking advantage of sequence coefficients offered by GUHA method.
2. Interest measures: We use the *base* and *above average dependency* (as GUHA counterparts to the more common support and lift) interest measures because of their easy interpretability and understandability.

3. Dimensions commensurability: Which data subsets can be compared together and which should not be compared in association rule mining?[1] Our research on automatic setting of the commensurability levels is currently ongoing; thus, this setting is done by the user in the current version of OLAP Recommender.

2.3 Visualising the results

OLAP Recommender does not visualise the association rules themselves; rather it visualises the part of the original data defined by the association rules. We use 2D column charts for the visualisation, thus constraining the maximal association rule antecedent length to 2 (as dimensions on the x-axis and in the legend) and the consequent length to 1 (as the measure on the y-axis) with a condition of unlimited length (as the data filter – slice or dice).

3 OLAP Recommender tool

OLAP Recommender was implemented as a web application in the ASP.NET framework and is running at http://connect-dev.lmcloud.vse.cz/Recommender. The workflow consists of three steps:

1. The user uploads the data as a CSV table or as RDF data semantically described by the Data Cube Vocabulary,[2] and defines their structure (dimensions and measures).
2. The user runs an association rule mining task. In this step the user can set the interest measures values and commensurability levels to get the most appropriate results.
3. OLAP Recommender displays the mined association rules, serving as a link to the data visualisation.

4 Experiments

OLAP Recommender was tested with two different datasets, whose attributes are displayed in Table 1: a retail dataset, described in more detail in Chudán's thesis [7], and a dataset[3] about European structural and investment funds (ESIF), prepared in the OpenBudgets.eu project.[4] We ran 5 mining tasks for the retail dataset and 8 tasks for the ESIF dataset with different settings. The tasks returned 46 association rules for the former and 114 rules for latter. Almost 70 % of the results in the retail dataset and

[1] Considering sales in a hypermarket as an example, not setting the commensurability levels would lead to many useless rules, pointing out the obvious: e.g. high sales of pastry and low sales of electronics. After proper commensurability level setting, we can receive more valuable rules, e.g., which pastry type sales prevail among the pastry products.

[2] https://www.w3.org/TR/vocab-data-cube/

[3] Available at https://github.com/openbudgets/datasets/tree/master/ESIF/2014/dataset

[4] http://openbudgets.eu

40 % in the ESIF dataset identified the highest or the lowest columns of the chart visualisations (i.e., anomalies and outlying values in the data).

5 Conclusions

By experiments with two different real datasets we proved the usefulness of the association rule mining and OLAP analysis combination, as the OLAP Recommender for the most test cases found more relationships, more interesting relationships and relationships in more parts of the cube, compared to using only self-guided OLAP analysis or using both methods separately.

Tab. 1: Retail and ESIF datasets differences summary

Characteristic	Retail dataset	ESIF dataset
Row count	34360	7039
Row interpretation	Sales of one product in one day	Single funded project
Dimension count	8	3
Measure count	1	3
Hierarchy	Two dimension hierarchies (product and time) with depth 4 and 2	Flat structure
Time dimension	Yes	No
Domain	Retail	Public fiscal data
Data form	Single table in .csv file	RDF data

The research has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833, OpenBudgets.eu.

References

1. HAN, J. OLAP mining: *An integration of OLAP with data mining.* In the Proceedings of the 7th IFIP Conference on Data Semantics, Leysin, Switzerland, 1997, pp 1-9.
2. IMIELIŇSKI, T., KHACHIYAN, L. and A. ABDULGHANI. *Cubegrades: Generalizing Association Rules.* Data Mining and Knowledge Discovery, v. 6 n.3, 2002, pp 219-257.
3. KAMBER, M., HAN, J. and J. CHIANG. Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes. In: *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining (KDD 1997).* Newport Beach, CA, USA, 1997, pp 207-210.
4. ZHU, H. *On-Line Analytical Mining of Association Rules*. Master thesis. Simon Fraser University, Burnaby, British Columbia, Canada 1998. Available at: http://www.cin.ufpe.br/~jtalr/Mestrado/Thesis/zhu98line.pdf
5. RAUCH, J. and M. ŠIMŮNEK. *Dobývání znalostí z databází, LISp-Miner a GUHA*. 1st ed. Praha: Oeconomica, nakladatelství VŠE, 2014. ISBN 9788024520339.
6. AGRAWAL, R., IMIELIŃSKI , T. and A. SWAMI. Mining association rules between sets of items in large databases. *Proc. SIGMOD '93,* 1993, pp 204-216.

7. CHUDÁN, D. *Association rule mining as a support for OLAP*. Dissertation thesis. University of Economics, Prague, Faculty of Informatics and Statistics, 2015.
8. KOUKAL, B. *OLAP Recommender: Supporting Navigation in OLAP Cubes Using Association Rule Mining*. Master thesis. University of Economics, Prague, 2017.

Recommending News Articles using Rule-based Classifier

Christián Golian and Jaroslav Kuchař

Web Intelligence Research Group, Faculty of Information Technology,
Czech Technical University in Prague
Thákurova 9, 160 00 Prague 6, Czech Republic

{goliachr, jaroslav.kuchar}@fit.cvut.cz

Abstract. In this paper we summarize our experiments with a rule-based classifier as a recommender within CLEF NewsREEL 2017 challenge. Systems that recommend news articles are suitable to solve information overflow in digital editions of newspapers, when users have problems choosing what they want to read. They face challenges unknown to the systems recommending books or movies such as a frequency of producing the new content. This paper deals with an approach based on association rules acting as a classifier. In our approach we experimented with settings that allow reducing the amount of rules used for the classification and increasing the performance that is crucial for real recommendations.

Keywords: news recommender, association rules, CLEF NewsREEL.

1 Introduction

The enormous number of available news causes information overflow resulting in users having problems choosing what they want to read. News recommendation systems should solve this problem and offer them an article or a collection of articles, which they could find worth a read.

In this paper we summarize our experiments with a rule-based classifier as a recommender within CLEF NewsREEL 2017 challenge[1] [1]. The challenge enables to compare and evaluate news recommendation systems both offline and online. This challenge allows to participate in two tasks: 1) NewsREEL Live [4] which uses real-time information about interactions between users and items. It is realized by redirecting a part of internet traffic to a recommender system of a participant in the challenge and 2) NewsREEL Replay [3] which uses historical data. The provided benchmarking system simulates the real stream of events by replaying of recorded historical data. Those tasks allow benchmarking of algorithms in terms of quality of recommendations (measured by the Click-Through Rate) and technical aspects (measured by reliability and response time)[5].

[1] www.clef-newsreel.org

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 51-55.

Over the years, many different approaches to news recommendation have been developed. Given that the rule-based approach yielded promising results [6] we decided to do further work in this area. The advantage of the rule-based approach is the possibility to easily explain the recommendations, since rules are considered as on of the most understandable representation of models. In our algorithm we focused on the rule-based classifier *CBA* [6] that is also focused on reducing amount of rules using a pruning of available rules. The proposed solution applies several existing algorithms and approaches that together build a competitive recommender system.

2 Approach

In our approach we decided to mainly use the contextual information about the reader or the article being read. During offline evaluation, we made several experiments to explore, which available features would perform best. Based on these experiments, we selected settings and a subset of twelve attributes for the online recommendation task.

The reasoning behind the use of rules was following: If a certain number of user interactions with an item often include values of attributes repeating themselves, they may be interesting either for users sharing these attributes or for users reading articles sharing these attributes. In order to provide an example: if an article was read frequently during evening hours, it may be interesting to someone else reading news in the evening or late at night.

To get list of relevant rules we use standard existing algorithms (e.g. *Apriori*, *FPGrowth*). Each rule is composed from a left-hand side, right-hand side and it is described by its quality measures: support and confidence. To prefer certain rules to other, we sort rules in the same way as in *CBA* according to the confidence, support and length of the rule. Since the amount of rules returned by the standard implementation can be huge, we use the rule pruning: it removes rules that can be never used for subsequent classification, usually due to their redundancy, lower significance etc. Our algorithm works as follows: an article (article id on the right hand side of the rule) is recommended only when values of attributes contained on the left hand side (contextual information) of a rule are equal to values of attributes in the recommendation request. If there are more matching rules, we use all unique article identifiers as a list of recommended articles. If no recommendation was made using a matching rule, implementation of baseline algorithm provided by organizers of CLEF News-REEL was used. Examples of rules from the classifier (values of features are anonymized):

```
{browser:40052, geo_user_zip:61958} => {itemId:341743113}
    supp = 0.02 , conf = 1.0
{isp:6, category:420949} => {itemId:367259468}
    supp = 0.01 , conf = 1.0
{browser:16064801, device:504182} => {itemId:315791779}
    supp = 0.1 , conf = 0.7
{} => {itemId:322334534}
    supp = 0.01 , conf = 0.01
```

In every domain there was a rule with an empty left hand side called the default rule. This means that every recommendation request from this domain matched it, and so at least one recommendation was always made using association rules.

The technical solution is built on top of the provided *Java SDK*[2]. Our implementation uses only the latest *n* interactions for rule mining. It enables to reflect outdating of news items and dealing with performance issues as well. Main implementations of rule mining algorithms and corresponding operations are available for the programming language R; we thus use a binary server *Rserve*[3] to provide communication between Java and R. To create association rules from interactions between users and news articles *arules* [2] library for R was used. The *rCBA* [7], a classification based on association classifier for R was used to prune and remove redundant association rules.

3 Evaluation

For the offline evaluation we selected subsets from the very large dataset provided by organizers. We compared results of our approach with the provided baseline implementation that recommends the most recent articles. The comparison of both approaches for selected news portals is on the Table 1. The rule-based approach can provide slightly better results. It takes into account the contextual features and temporal aspects at the same time.

Tab. 1. Offline evaluations – Click-Through Rate

News Portal (id)	Baseline algorithm	Rule-based algorithm
418	0.07%	0.14%
1677	0.40%	0.45%
35774	1.30%	1.62%
All	1.04%	1.29%

Table 2 presents the experiments with setting of parameters that influence the building of the rule-based classifier. It allows getting brief insights to the influence of specific settings. Those experiments helped us to find appropriate settings for further experiments. Please note, that only subset of experiments is presented.

[2] https://github.com/plista/orp-sdk-java
[3] https://rforge.net/Rserve/

Tab. 2. Offline evaluations – parameters setting

Parameter/Portal	418	1677	35774	All
conf: 1%, sup: 0.5%	0.14%	0.45%	1.62%	1.29%
conf: 1%, sup: 1%	0.16%	0.43%	1.59%	1.27%
conf: 5%, sup: 2%	0.13%	0.46%	1.67%	1.34%
max. clicks 5000	0.14%	0.41%	1.65%	1.31%
max. clicks 100000	0.14%	0.40%	1.65%	1.31%
max. rule length 2	0.16%	0.42%	1.66%	1.32%
max. rule length 10	0.13%	0.42%	1.60%	1.30%
max. rule length 12	0.13%	0.43%	1.62%	1.29%
pruning disabled	0.18%	0.53%	1.75%	1.40%

Several conclusions were drawn from these experiments. The prediction accuracy increased together with increasing of maximum length of association rule only up to a certain length. Increasing the maximum number of clicks (size of data that is used to mine rules) did not significantly increase the prediction accuracy. Disabling pruning of rules can bring better results, but at a cost of higher number of rules leading to higher number of erroneous responses and longer computing time. It can lead to the decreased reliability and increased response times.

Tab. 3. Online evaluations – Selected teams from the official results [5]

Team	Clicks	Impressions	CTR
BL2Beat	726	193014	0.376%
B	879	244334	0.360%
WIRG	600	154419	0.389%
I	764	236332	0.323%
N	1268	255663	0.496%
O	896	130221	0.688%
Q	12	1380	0.870%

In the online evaluation, our algorithm took 13th place of 21 contestants based on the CTR that relates to impressions - how often the recommendations of a system have been shown to readers (Algorithm called WIRG in the Table 3). The table summarizes results from the final evaluation period. The setting of our algorithm we selected according to the best results from the offline evaluation and two online testing periods allowing tuning of participated algorithms. We were able to beat the baseline (BL2Beat) with our approach. Since we do not know any details about other participating algorithms, it is not possible to state any conclusions yet. The important fact is that our algorithm was able to handle incoming messages, process data and provide recommendations (up to 100 messages per second with responses within 100ms). The recommendations are at the same time easily explainable.

4 Conclusion

In this paper our news recommender system based on association rules was examined. The main idea of the algorithm is to build a rule based classifier using contextual features and study influence of several settings. The advantage is that the recommendations are explainable and the recommender is technically designed as a scalable solution allowing real-time recommendations. Our algorithm took part in both tasks of the CLEF NewsREEL 2017 challenge. The algorithm managed to beat baseline. In our future work we would like to address the detected issues and limitations of our solution. One aspect that we would like to try to overcome is the need for repetitive computation of models on the background using stream-based version of the rule-mining algorithm.

References

1. Golian Ch., Kuchar J.: News Recommender System based on Association Rules @ CLEF NewsREEL 2017. In: Working Notes of CLEF 2017 – Conference and Labs of the Evaluation forum, Dublin, Ireland, 11-14, 2017.
2. Hahsler M., Gruen B.,Hornik K.: arules – A computational environment for mining association rules and frequent item sets. Journal of Statistical Software, 14(15):1–25, 2005.
3. Hopfgartner F., Brodt T., Seiler S., Kille B., Lommatzsch A., Larson M., Turrin R., Sereny A.: Benchmarking news recommendations: The CLEF newsreel use case. SIGIR Forum, 49(2):129–136, 2015.
4. Hopfgartner F., Kille B., Lommatzsch A., Plumbaum T., Brodt T., Heintz T.: Benchmarking news recommendations in a living lab. In: CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative, LNCS, pages 250–267.Springer Verlag, 09 2014.
5. Kille, B., Lommatzsch, A., Hopfgartner, F., Larson, M. and Brodt, T.: CLEF 2017 NewsREEL Overview: Offline and Online Evaluation of Stream-based News Recommender Systems. In: Working Notes of CLEF 2017: Conference and Labs of the Evaluation Forum, Dublin, Ireland, 11-14, 2017.
6. Kliegr T., Kuchar J.: Benchmark of rule-based classifiers in the news recommendation task. In: Proceedings of the Sixth International Conference of the CLEF Association, CLEF'15, pages 130–141, 2015.
7. Vojir S., Zeman V., Kuchar J., Kliegr T.: Easyminer/r preview: Towards a web interface for association rule learning and classification in r. In: Proceedings of the RuleML 2015 Challenge, Berlin, Germany, August 2-5, 2015., 2015.

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

Využití EasyMiner API v projektu OpenBudgets.eu

Stanislav Vojíř[1], Václav Zeman[1], Jaroslav Kuchař[1, 2], Tomáš Kliegr[1]

[1]Katedra informačního a znalostního inženýrství, FIS, Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 13067 Praha 3 – Žižkov
[2]Web Intelligence Research Group, FIT, České vysoké učení technické v Praze
Thákurova 2700/9, 160 00 Praha 6 - Dejvice

{stanislav.vojir|vaclav.zeman|tomas.kliegr}@vse.cz
jaroslav.kuchar@fit.cvut.cz

Abstrakt. V souvislosti s rostoucí popularitou využívání data miningových dat lze registrovat také rostoucí poptávku po možnosti integrace data miningových algoritmů a systémů do komplexnějších, uživatelsky přívětivějších aplikací. Tento příspěvek prezentuje novou verzi systému EasyMiner, integrovanou do softwarového řešení vyvíjeného v rámci evropského projektu OpenBudgets.eu, který je zaměřen na zpřístupňování a analýzy finančních dat samospráv. EasyMiner je webový data miningový systém podporující dolování asociačních pravidel, tvorbu klasifikačních modelů a v současné verzi nově také detekci outlierů. Příslušná funkcionalita je k dispozici nejen prostřednictvím grafického uživatelského rozhraní, ale také prostřednictvím komplexního REST API.

Klíčová slova: asociační pravidla, klasifikace, detekce outlierů, data mining, REST API, EasyMiner, OpenBudgets.eu.

1 Úvod

EasyMiner (*http://easyminer.eu*) je webový data miningový systém dlouhodobě vyvíjený na Katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze. Starší verze tohoto systému [1] byly zaměřeny zejména na zpřístupnění dolování asociačních pravidel a posléze tvorbu klasifikačních modelů v přehledném, grafickém uživatelském rozhraní, fungujícím ve všech moderních webových prohlížečích. Tato možnost je stále dostupná, avšak pro možnost automatizovaných analýz většího množství dat a propojitelnost s dalšími systémy bylo nutné celý systém EasyMiner dále rozvíjet.

Tento příspěvek popisuje novou verzi systému EasyMiner, použitou pro analýzu finančních dat v rámci evropského projektu OpenBudgets.eu. Aktuální verze podporuje nejen dolování asociačních pravidel, ale také *detekci outlierů*[1] založenou na dolo-

[1] *Outlier* – z hlediska překladu do češtiny se jedná o „odlehlé hodnoty" – instance dat charakterizované atributy s konkrétními hodnotami, které se odlišují od zbytku datové matice; míru odlišnosti charakterizuje *outlier score.*

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 56-60.

vání častých vzorů (*frequent patterns*). Zároveň jsou všechny funkce EasyMineru dostupné prostřednictvím nového REST API.

Zahraniční komunitě byla tato nová verze systému EasyMiner prezentována v rámci konference RuleML 2017 [2].

2 EasyMiner API

Předchozí verze systému EasyMiner podporovala jednoduché dolování asociačních pravidel v grafickém uživatelském rozhraní. Ačkoliv byla tato varianta vhodná pro koncové uživatele, nebylo možné využívat funkcionality systému EasyMiner v rámci rozsáhlejších projektů. Za tímto účelem bylo vytvořeno nové REST API podporující jednak veškerou původní funkcionalitu systému EasyMiner a zároveň také nově implementované algoritmy pro předzpracování dat a detekci outlierů.

Prostřednictvím daného API je integrována funkcionalita systému EasyMiner do projektu OpenBudgets.eu. Obdobně je možné jej využít také v dalších projektech – pro tvorbu *mashup aplikací*, začlenění do vlastních skriptů atp., což je podpořeno jeho plnou dokumentovaností a open source licencí celého systému EasyMiner.

2.1 Data mining prostřednictvím API

Komplexní REST API pro koncové uživatele je dostupné na adrese *<easyminer-server>/easyminercenter/api*. Tato adresa je rozcestníkem celého API a zároveň je na ní k dispozici také kompletní dokumentace v syntaxi *Swagger*.[2]

Pro využití API musí mít uživatel nejprve vytvořený vlastní unikátní API klíč, který získá prostřednictvím registrace uživatelského účtu. Tento API klíč musí být zasílán ve všech jednotlivých požadavcích na API. Následný postup přípravy dat a dolování pomocí API je obdobný jako při použití grafického rozhraní – viz následující odstavce.

Postup pro dolování asociačních pravidel: 1. nahrání dat ve formátu CSV, **2.** vytvoření *instance mineru*,[3] **3.** předzpracování dat (vytvoření atributů z nahraných dat za využití jednoduchých definic předzpracování), **4.** zadání data miningové úlohy (vzor asociačních pravidel, požadované míry zajímavosti), **5.** spuštění dolování, vyčkání na výsledky a **6.** zpracování výsledků (export ve formátu PMML[4] či JSON).

Postup pro detekci outlierů: Kroky 1.-3. jsou totožné jako u dolování asociačních pravidel. Následující kroky jsou: **4.** definice úlohy detekce outlierů (minimální podpora [support]), **5.** spuštění úlohy detekce outlierů a vyčkání na výsledky, **6.** procházení datových řádků uspořádaných podle *outlier score*.[1]

Předzpracování dat: Pro použití algoritmů založených na dolování frekventovaných vzorů je nutné nejprve předzpracovat (diskretizovat) hodnoty číselných dato-

[2] *Swagger* – framework pro tvorbu dokumentace pro REST API, *https://swagger.io/*

[3] EasyMiner podporuje dvě verze backendů – pro dolování pomocí systému R či pomocí Hadoop serveru. Při vytvoření *instance mineru* dochází k přípravě databáze a inicializaci příslušného backendu.

[4] PMML – XML formát pro záznam data miningových modelů

vých sloupců. Pro řadu případů použití je uživateli vyžadováno také předzpracování hodnot sloupců výčtových. EasyMiner API podporuje všechny základní metody předzpracování dat – převzetí původních hodnot, uživatelsky definované množiny hodnot či intervaly, automaticky vygenerované intervaly (*equidistant, equifrequent, equisized*).

V rámci kroku předzpracování dat musí uživatel předzpracovat všechny datové sloupce, které chce použít pro následnou data miningovou úlohu (tj. vytvořit z nich *atributy*, které se následně nacházejí ve výsledcích). Při dolování asociačních pravidel využívá uživatel zvolené atributy pro definování vzoru hledaných pravidel – má tedy možnost využívat atributy opakovaně v rámci většího množství úloh. V případě úlohy hledání outlierů jsou využity všechny připravené atributy.

2.2 Použité algoritmy

Základní úlohou podporovanou systémem EasyMiner je dolování asociačních pravidel, volitelně s možností využití jejich prořezání za účelem tvorby klasifikačního modelu. Za tímto účelem je v současné době využíván algoritmus *apriori* [3], implementovaný v balíčku *arules* pro systém R. Z hlediska architektury systému EasyMiner je tento algoritmus spouštěn prostřednictvím *R serveru*, jehož funkcionalita je zprostředkována dolovací službou. Pro velké datasety je využíván také *Hadoop server*, konkrétně algoritmus *FP-Growth* [4]. Protvorbu klasifikačních modelů je využívána vlastní implementace algoritmu *CBA*. [5]

Výsledkem dolování jsou asociační pravidla ve tvaru *antecedent* → *konsekvent*, ve kterých mohou být *antecedent* i *konsekvent* tvořeny konkjunkcí atributů s konkrétními hodnotami. V případě tvorby klasifikačních modelů platí je konsekvent omezen na jeden cílový atribut. Použitelnými měrami zajímavosti jsou *konfidence*, *podpora* a *lift*.

V současné době podporuje EasyMiner také detekci outlierů. Za tímto účelem byl Jaroslavem Kuchařem implementován balíček *fpmoutliers* pro systém R [6]. Podporovanými algoritmy jsou *FPCOF*, *FPOF*, *MFPOF*, *WCFPOF*, *WFPOF*, *LFPOF* a nový inovativní přístup *FPI*. [7] Výsledkem detekce outlierů jsou datové řádky z předzpracovaného datového souboru seřazené podle *outlier score*.[1]

3 Využití EasyMineru v systému OpenBudgets.eu

OpenBudgets.eu (H2020-645833, *http://openbudgets.eu*) je evropský projekt zaměřený na zpřístupňování a analýzy finančních a rozpočtových dat samospráv. Součástí tohoto projektu, v rámci pracovního balíčku *WP2 Infrastructure for Data Collection and Mining* byly zkoumány možnosti analýzy finančních dat data miningovými algoritmy a nástroji. Jedním z vybraných nástrojů je také systém EasyMiner. Na základě analýzy požadavků byla do systému EasyMiner implementována podpora dolování outlierů a komplexní REST API.

Všechna data analyzovaná v projektu OpenBudgets.eu jsou dostupná ve formátu RDF.[5] Prostřednictvím nástrojů pro přípravu dat jsou data získaná od samospráv konvertována do formátu RDF, rozšířena o data z veřejně dostupných zdrojů (například z registrů ekonomických subjektů) a posléze zpřístupněna pro analýzy. Pro použití EasyMineru jsou data následně konvertována do CSV nástrojem *LinkedPipes ETL* [8].

Z hlediska integrace do softwarové architektury vyvíjené v projektu OpenBudgets.eu je funkcionalita EasyMineru využívána prostřednictvím API, jehož funkce jsou volány z integračního nástroje *DAM* (*http://github.com/openbudgets/DAM*).

4 Závěr

Data miningový systém EasyMiner je experimentálním akademickým projektem, dostupným pod open source licencí *Apache License, Version 2.0*. Jeho nová verze disponuje nejen grafickým uživatelským rozhraním pro dolování asociačních pravidel, ale podporuje také integraci do dalších projektů prostřednictvím API. Komplexní příklad využití EasyMiner API je dostupný na *http://www.easyminer.eu/api-tutorial*.

Z hlediska reálného nasazení je systém EasyMiner v současné době využíván v rámci data miningové architektury vytvářené v projektu OpenBudgets.eu, zaměřeném na analýzy finančních a rozpočtových dat.

V rámci budoucího vývoje by měly být systém dále rozšířen o podporu dolování nad RDF daty a mělo by dojít k posunu současných limitů týkajících se velikosti použitelných dat.

Literatura

1. Vojíř, S., Zeman, V., Kuchař, J., Kliegr, T.: EasyMiner/R Preview: Towards a Web Interface for Association Rule Learning and Classification in R. In: Proceedings of the RuleML 2015 Challenge, the Special Track on Rule-based Recommender Systems for the Web of Data, the Special Industry Track and the RuleML 2015 Doctoral Consortium hosted by the 9th International Web Rule Symposium (RuleML 2015), Berlin, Germany (2015). http://ceur-ws.org/Vol-1417/paper10.pdf
2. Vojíř, S., Zeman, V., Kuchař, J., Kliegr, T.: Using EasyMiner API for Financial Data Analysis in the OpenBudgets.eu Project. In: Proceedings of the Doctoral Consortium, Challenge, Industry Track, Tutorials and Posters @ RuleML+RR 2017 hosted by International Joint Conference on Rules and Reasoning 2017 (RuleML+RR 2017), London, UK (2017). http://ceur-ws.org/Vol-1875/paper21.pdf
3. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: SIGMOD (1993) 207-216.
4. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. In: Data Mining and Knowledge Discovery 8 (2004) 53–87.

[5] *RDF* – datový formát pro sémantický web, *https://www.w3.org/RDF/*

5. Kliegr, T., Kuchař, J., Sottara, D., Vojíř, S.: Learning Business Rules with Association Rule Classifiers. In: International Workshop on Rules and Rule Markup Languages for the Semantic Web, Springer (2014) 236–250.
6. Kuchař, J.: jaroslav-kuchar/fpmoutliers. https://github.com/jaroslav-kuchar/fpmoutliers [15. 6. 2017].
7. Kuchař, J,; Svátek, V.: Spotlighting Anomalies using Frequent Patterns. In: Proceedings of the KDD 2017 Workshop on Anomaly Detection in Finance, PMLR, Halifax, Nova Scotia, Canada (2017).
8. Klímek, J., Škoda, P., Nečaský, M.: LinkedPipes ETL: Evolved linked data preparation. In: International Semantic Web Conference, Springer (2016) 95-100.

Poděkování: Tento článek vznikl díky podpoře z projektů OpenBudgets.eu (H2020-645833) a IGA 29/2016 Vysoké školy ekonomické v Praze.

Annotation:

Using EasyMiner API in the OpenBudgets.eu Project

Related to the increasing popularity of data mining there is a growing effort to integrate data mining algorithms and systems into user-friendly applications and information systems. This paper introduces a new version of web-based data mining system EasyMiner and its integration into a software solution developed within the European project OpenBudgets.eu. This project is aimed at publication and analysis of financial data of municipalities. The current version of EasyMiner supports mining of association rules, building of classification models and newly also outlier detection. Its functionality is available not only via a graphical user interface, bus also via REST API. The API can be easily used also from third party applications.

Hodnocení (ne)zajímavosti asociačních pravidel za využití báze znalostí

Přemysl Václav Duben, Stanislav Vojíř

Katedra informačního a znalostního inženýrství, FIS, Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 13067 Praha 3 - Žižkov

{xdubp00|stanislav.vojir}@vse.cz

Abstrakt. Dolování asociačních pravidel je jednou z populárních data miningových metod, prostřednictvím které lze objevovat zajímavé vztahy v datech. Asociační pravidla jsou využitelná nejen pro analýzu chování zákazníků, ale také pro tvorbu klasifikačních modelů či pro detekci výjimek. Algoritmy pro hledání asociačních pravidel mají však také jednu nevýhodu – velké množství nalézaných výsledků. Při ručním řešení analytické otázky se musí data miningový expert probrat velkým množstvím pravidel a vybrat z nich jen ta opravdu zajímavá. V rámci tohoto příspěvku je představena metoda výběru (ne)zajímavých asociačních pravidel dle jejich podobnosti s pravidly dříve vybranými do báze znalostí. Tato metoda postprocessingu je implementována ve webovém data miningovém systému EasyMiner.

Klíčová slova: asociační pravidla, postprocessing, podobnost asociačních pravidel, báze znalostí, EasyMiner.

1 Motivace

Dolování asociačních pravidel je jedním ze základních typů data miningových úloh, vhodných nejen pro analýzu nákupních košíků, ale také pro objevování složitějších vztahů v datech. Nevýhodou algoritmů pro dolování asociačních pravidel jsou však velké nároky na následné zpracování výsledků. I v rámci jednoduchých úloh lze nalézt opravdu velké množství pravidel, která je následně nutné vyhodnotit z hlediska jejich zajímavosti pro řešenou analytickou otázku. Ruční vyhodnocení všech nalezených pravidel data miningovým či doménovým expertem je časově velmi náročné, neřkuli nemožné. Dlouhodobě jsou tedy hledány možnosti filtrování nalézaných pravidel ať již v průběhu zadání data miningové úlohy, v průběhu procesu dolování či v rámci zpracování jejích výsledků.

V rámci výzkumu realizovaného na Katedře informačního a znalostního inženýrství Vysoké školy ekonomické v Praze v projektech SEWEBAR a posléze EasyMiner (*http://easyminer.eu*) byly analyzovány možnosti zjednodušení procesu řešení úloh dolování asociačních pravidel a výběru adekvátních výsledků. V minulosti šlo o hledání obdobných asociačních pravidel za využití jejich zápisu ve formátu XML a jazyka *XQuery* [1]. Posléze byla v rámci projektu EasyMiner navržena a implementována

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 61-66.

báze znalostí umožňující shromažďování zajímavých a nezajímavých asociačních pravidel do ručně vytvářených *rule setů*, využitelných jak pro ruční vytváření klasifikačních modelů, tak pro sdílení doménových znalostí o potenciální zajímavosti pravidel. [2][3]

V rámci tohoto příspěvku je představena nově implementovaná metoda pro vizuální zjednodušení procházení asociačních pravidel nalezených v rámci data miningové úlohy na základě hodnocení jejich podobnosti s pravidly dříve přidanými do zvoleného rule setu. Daná metoda je implementována v grafickém uživatelském rozhraní webového data miningového systému EasyMiner (*http://easyminer.eu*).

Z hlediska zasazení do kontextu aktuálního výzkumu jsou v rámci dolování asociačních pravidel v systému EasyMiner využívány metody omezení nalézaných pravidel při zadávání data miningové úlohy a práce s doménovou znalostí (uloženou v bázi znalostí). Uživatel má zároveň možnost volitelně využít prořezávání pravidel za využití algoritmu CBA [4]. Z relevantních alternativních přístupů lze jmenovat například metody dolování asociačních pravidel s omezeními [5], již zmíněné metody prořezávání nalézaných pravidel, nalézání zajímavých pravidel za využití ontologií [6] či metody vizualizace a shlukování obdobných pravidel [7].

2 Báze znalostí

V rámci projektu EasyMiner je implementována komplexní znalostní báze shromažďující informace o atributech využívaných pro řešení data miningových úloh a jejich metodách předzpracování a také o pravidlech shromážděných v rámci rule setů uložených taktéž v bázi znalostí. Báze znalostí byla poprvé představena v [2]. Posléze byl návrh dokončen a implementován v [3].

Báze znalostí se vnitřně skládá z trojice základních typů uložených informací: 1. metainformace o zpracovávaných datech – meta-atributy a jejich formáty, 2. informace o vhodných metodách předzpracování konkrétních typů dat (formátů meta-atributů) na atributy použitelné pro data mining, 3. konkrétní uložená pravidla (a jejich seskupení do rule setů).

Jednotlivé *meta-atributy* označují základní typy dat, ve své podstatě jde o abstraktní skupiny atributů – např. „věk“. Dané typy dat se posléze vyskytují v konkrétních *formátech* – např. „věk v letech“.

Pro jednotlivé *formáty* jsou definovány *metody předzpracování* uživatelem nahraných dat do podoby datových *atributů* použitelných pro dolování asociačních pravidel. Báze znalostí obsahuje informace o již definovaných metodách seskupení hodnot pomocí množin či intervalů.

V rámci řešení data miningové úlohy nejprve uživatel nahraje svá vlastní data do aplikace EasyMiner. Následně je vyzván k jejich předzpracování – vytvoření atributů z datových sloupců obsažených v nahraných datech. K předzpracování jsou využívány metody ručně definovaných množin či ručně či automaticky definovaných intervalů.

V rámci předzpracování jsou data *namapována na meta-atributy* uložené v bázi znalostí (a jejich *konkrétní formáty*). Ve výchozím stavu jsou formáty automaticky

vytvářeny na základě nahraných dat, volitelně má však uživatel možnost znovu-použít již existující metodu předzpracování – čímž dojde k namapování datového sloupce na již existující formát/meta-atribut. Na základě mapování na formáty meta-atributů poté mohou být vyhodnoceny jako podobné (či totožné) také atributy, které se sice liší ve svých názvech, ale vznikly ze stejných datových sloupců.

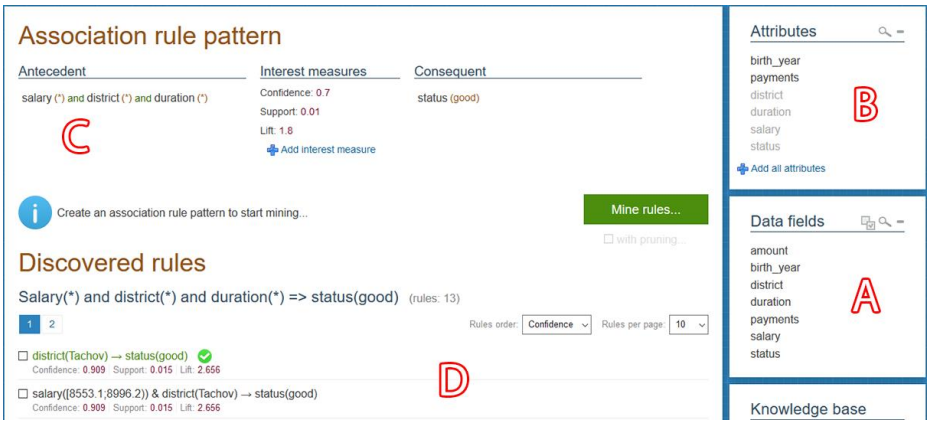
V průběhu dolování má následně uživatel možnost přidávat vybraná asociační pravidla do báze znalostí – za účelem vytváření klasifikačních modelů či za účelem hodnocení (ne)zajímavosti u dalších, podobných pravidel.

3 Hodnocení zajímavosti asociačních pravidel

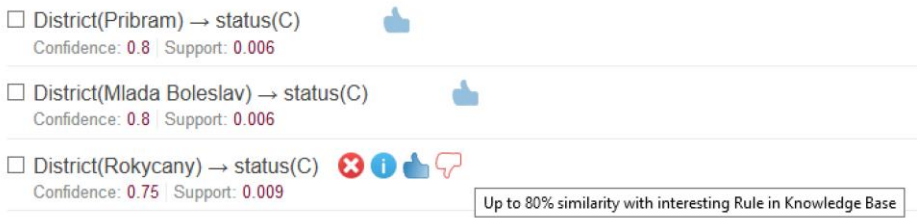
V rámci procesu dolování asociačních pravidel v grafickém rozhraní systému Easy-Miner (**Obr. 1**) uživatel nejprve předzpracuje datové sloupce na atributy (**Obr. 1** A, B). Následně definuje data miningovou úlohu vytvořením vzoru hledaných asociačních pravidel (**Obr. 1** C) – tj. umístí zvolené atributy do antecedentu či konsekventu vzoru pravidel a definuje minimální hodnoty požadovaných měr zajímavosti.[1] Po spuštění dolování jsou nalezené výsledky zobrazeny přímo v rámci uživatelského rozhraní v sekci *Discovered rules* (**Obr. 1** D). V rámci této sekce má uživatel možnost pravidla řadit dle použitých měr zajímavosti a také označovat pravidla, která považuje za zajímavá či naopak za nezajímavá.

Při procházení výsledků má uživatel možnost uložit vybraná pravidla do báze znalostí – pomocí jejich označení za zajímavá či nezajímavá. Následně jsou všechna ostatní pravidla hodnocena z hlediska podobnosti s takto uloženými pravidly a dle jejich podobnosti je k nim doplněna informace o tom, zda je nejpodobnější pravidlo uloženo jako (ne)zajímavé. Dané porovnání je realizováno nejen ve výsledcích jedné dílčí data miningové úlohy, ale také u všech úloh následujících. Tj. pokud uživatel např. označí zvolené pravidlo za nezajímavé, bude jako nezajímavé označeno i v případě, že se vyskytne ve výsledcích některé z následujících úloh. Označování zajímavosti pravidel je znázorněno na **Obr. 2**. K aktivaci vyhodnocení podobnosti daného jednotlivých pravidel je využíván AJAX.

[1] Uživatel má k dispozici míry zajímavosti *konfidence*, *podpora* a *lift*. Volitelně lze zapnout také automatické prořezání zobrazených výsledků algoritmem *CBA* [4].



Obr. 1. Uživatelské rozhraní systému EasyMiner



Obr. 2. Ukázka vyhodnocení podobnosti asociačních pravidel s pravidly ve znalostní bázi

Pro porovnání s obsahem báze znalostí je využívána množina atributů nacházejících se v antecedentu či konsekventu konkrétního asociačního pravidla a posléze jejich konkrétních hodnot. Celé porovnání je za účelem rychlosti realizováno jako dvoufázové. V první fázi dochází u pravidel nalezených nad jedním konkrétním datovým souborem. V tomto případě dochází nejprve k porovnání textové podobnosti antecedentu a konsekventu daných pravidel.

V rámci podrobného porovnání je pravidlo: 1. rozděleno na jednotlivé atributy obsažené v antecedentu/konsekventu, 2. k jednotlivým atributům jsou v rámci báze znalostí dohledány odpovídající formáty meta-atributů (na základě informace o předzpracování dat), 3. dochází k dekomponování hodnot. Tímto způsobem jsou porovnávána i pravidla získaná na základě analýz rozdílných datových souborů, pokud dané atributy využívají stejné formáty meta-atributů.

Z hlediska dekomponování je například pravidlo

salary_intervals([10000;20000]) & district(Tachov) → status(good)

nejprve rozděleno na jednotlivé atributy. Následně jsou k těmto atributům dohledány jejich konkrétní formáty meta-atributů – antecedent: *salary/czk([10000,20000]), district/towns(Tachov)*, konsekvent: *status/sets(good)*. V posledním kroku jsou hodnoty

atributů nahrazeny identifikátory intervalů či množin hodnot definovaných v bázi znalostí. Z hlediska daného příkladu je hodnota atributu *good* definována jako množina původních hodnot datového sloupce {„A“, „B“}.

Pro stanovení podobnosti je nejprve vyhodnocena shoda atributů v antecedentu/konsekventu pravidla a posléze shoda jejich hodnot. Každá část pravidla má váhu 50 % celkového hodnocení. Pro uživatelskou srozumitelnost je následně podobnost převedena na hodnoty ze škály 0 – 100 %. V současné variantě nejsou porovnávány hodnoty použitých měr zajímavosti.[2]

4 Závěr

V rámci tohoto příspěvku byla představena metoda hodnocení zajímavosti nalézaných asociačních pravidel za využití pravidel dříve ručně přidaných do báze znalostí. Daná metoda byla implementována v systému EasyMiner.

V rámci dalšího vývoje by mělo dojít k podrobnějšímu uživatelskému a výkonnostnímu testování implementovaných algoritmů, testování vhodnosti (ne)zahrnutí hodnot měr zajímavosti pro stanovení podobnosti pravidel a k usnadnění ručního mapování nahraných dat na existující formáty meta-atributů báze znalostí.

Literatura

1. Kliegr, T., Hazucha, A., Marek, T.: Instant feedback on discovered association rules with PMML-based query-by-example. In: International Conference on Web Reasoning and Rule Systems, Springer Berlin Heidelberg (2011), pp. 257-262.
2. Vojíř, S.: Concept of Semantic Knowledge Base for Data Mining of Business Rules. In: Znalosti 2014 Exhibice, Edukace a nacházení Expertů - Exhibition, Education and Expert finding. Praha: KIZI FIS (2014) pp. 132-136.
3. Vojíř, S.: Učení business rules z výsledků dolování GUHA asociačních pravidel. Vysoká škola ekonomická v Praze (2016).
4. Kliegr, T., Kuchař, J., Sottara, D., Vojíř, S.: Learning Business Rules with Association Rule Classifiers. In: International Workshop on Rules and Rule Markup Languages for the Se-mantic Web, Springer (2014), pp. 236–250.
5. Dan Nguyen, Bay Vo: Mining Class-Association Rules with Constraints. In: {Knowledge and Systems Engineering, Springer (2014) pp. 307-318.
6. Saravanam, D., Vijayalakshmi, S., Joseph, D.: Finding of Interesting Rules from Association Mining by Ontology and Page Ranking. International Journal for Modern Trends in Science and Technology (2017) 03 01, pp. 46-50
7. Hahsler, M., Karpienko, R.: Visualizing association rules in hierarchical groups. Journal of Business Economics (2017) 87 3, pp. 317-335.
8. Duben, P. V.: Filtrování zajímavých pravidel v systému EasyMiner. Vysoká škola ekonomická v Praze (2017).

[2] Dle základního uživatelského testování není nezahrnutí měr zajímavosti problémem. V rámci dalšího výzkumu by měly být ověřeny možnosti stanovení priorit pro podobnost antecedentu/konsekventu.

Poděkování: Tento článek vznikl díky podpoře z projektu IGA 29/2016 Vysoké školy ekonomické v Praze.

Annotation:

Rating of (Non)Interestingness of Association Rules using Knowledge Base

Data mining of association rules is one of popular data mining method used to discovering of interesting relationships hidden in data. Association rules are applicable not only for customer behavior analysis, but also for building of classification models and for detection of exceptions. However, algorithms for discovery of association rules have one disadvantage – many results, founded even in simple data mining tasks. During the solving of analytical question, the data mining expert must deal with a big count of rules and choose the interesting of them. This paper introduces a method of selection of (non) interesting association rules according to their similarity to rules previously stored in knowledge base. This post-processing method is implemented in the web data mining system EasyMiner.

Efektivní analýza velkých dat pomocí Apache Spark a samoučících neuronových sítí na jediném počítači

David Andrešič, Petr Šaloun

Katedra informatiky, FEI VŠB-TUO v Ostravě
17. listopadu 15/2172, 708 33 Ostrava - Poruba

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

Abstrakt. Apache Spark je běžně používaná platforma pro analýzu velkých dat na velkých počítačových clusterech, kde pro svou práci využívá především hlavní paměť počítače. Pokusili jsme se přidat softwarovou knihovnu samoučící se neuronové sítě do jednoho takového analytického celku pro big data. Výsledek je efektivní a rychlý dokonce na jediném běžném počítači.

Tento přístup je přínosem pro výzkumníky s omezenými zdroji, kterým přináší možnost analýzy velkých dat. Náš nápad byl experimentálně ověřen a je popsán zde. Jako případovou studii pro naši metodu jsme použili dostupná data ze sociální sítě Twitter, konkrétně tweety pro hashtag #Brexit a jejich analýzu sentimentu, přičemž jsme hledali korelace s burzovními daty.

Klíčová slova: Apache Spark, samoučící neuronové sítě, big data, Twitter, brexit, burza

1 Úvod

Korelace tweetů a burzy statistickými metodami je častým předmětem výzkumu i publikací. Náš přístup pro sociální síť Twitter a vývoj burzy zahrnoval agregování veřejných dat a shlukovou analýzu metodami strojového učení (samo-organizujících mapy – SOM) pomocí Apache Spark[1] určeného pro zpracování velkých dat na počítačových clusterech a to na jediném, standardním počítači pro pre-processing velkých dat. Jelikož se Spark snaží zpracovávat data primárně v paměti RAM, čelili jsme omezeným prostředkům. S podobným problémem jsme se potýkali v analýze shluků pomocí SOM, kde byla efektivita implementací velmi rozdílná (viz část 3.2).

2 Metodologie a datové sady

Celý postup vypadal takto:

– *Analýza sentimentu* za účelem zjištění jak pozitivní, či negativní daný tweet byl.

[1] Apache Spark homepage: http://spark.apache.org/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 67-71.

- *Spojení a agregování dat,* zahrnující výpočet korelačních koeficientů a transformaci dat do podoby vhodné pro SOM, vše s použitím Apache Spark.
- *Shluková analýza* pomocí metody SOM ve snaze najít shluky podobných společností, které korelovaly nejvíce.

2.1 Analýza sentimentu dat z Twitteru

Tweety pro hashtag *#brexit* byly k dispozici pro několik dní z období 29.4.2016 až 2.7.2016. Analýza sentimentu ukázala, že před referendem (23.6.2016) byly tweety spíše kritikou EU, zatímco po něm k Brexitu samotnému. Celkový počet analyzovaných tweetů byl 21137 (10799 před referendem a 10338 po referendu).

2.2 Burzovní data a data společností

Zdrojem věrohodných burzovních dat bylo Yahoo Finance[2]. Poskytuje historická data z mnoha burz, včetně námi zvolené London Stock Exchange (1292 společností). Yahoo Finance umožňuje stažení těchto burzovních dat: datum, hodnota akcie při otevření a zavření burzy v daný den, maximum a minimum hodnoty daný den a objem zobchodovaných akcií. K těmto jsme přidali další atributy z Google Finance (adresa, oblasti působnosti, město, země).

3 Transformace a agregace dat

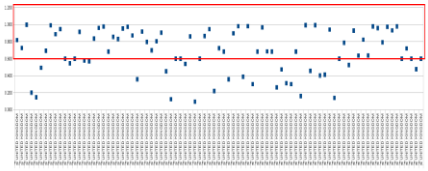
Nejprve jsme použili data podrobněji popsaná v části 2.1 s interpolací chybějících hodnot. SOM vyžadovala, aby pro každou společnost byly informace o střední hodnotě ceny akcie při uzavření burzy a počtu zobchodovaných akcií, jejich korelační koeficient se sentimentem, město, zemi a oboru působnosti. Výpočty byly rozděleny do tří period: před a po referendu a celé období. Spojením obou burzovních datových sad jsme získali informace o 962 společnostech. Následně byl spočítán korelační koeficient pro hodnoty burzovních ukazatelů a sentiment v daný den pro každou periodu pomocí Sparku v módu clusteru o jediném uzlu (počítač Intel Core i3 se dvěma jádry/čtyřmi vlákny, 8GB RAM). Jde o současný komoditní hardware, pro který je Spark navržen (škálování do šířky). Pokusili jsme se všechny operace spojení datových sad a selekce provést najednou v RAM. Bohužel, každý výběrový dotaz z důvodu velikosti datového rámce (stovky řádků pro každou společnost, které byly dále spojeny s jejími dodatečnými atributy) zabral několik minut a výpočty ani po několika dnech nedoběhly. Klasický přístup relačních databází s indexem na atribut burzovního symbolu (tickeru) ve světě Spark SQL nelze použít (pouze transformací do RDD). Vyladili jsme proto části kódu tak, aby se výpočty prováděly iterativně na menších blocích. Pomocí této optimalizace jsme se vyhli drahým výběrovým operacím v rozsáhlém datovém rámci a dokončili výpočty za necelé dvě hodiny. Jelikož se při výpočtu pro každou společnost používají pouze její data, je možné tento postup

[2] Yahoo Finance: https://finance.yahoo.com/

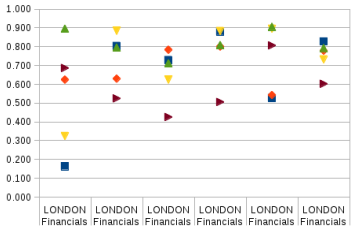
paralelizací dále zefektivnit. Výsledná tabulka tak navíc obsahovala střední hodnoty akcií a jejich zobchodovaného objemu na konci dne pro všechny periody normalizované v rozsahu 0-1. Kategorické (nominální) atributy byly „binarizovány“ (např. „London“ v atributu „town“ vytvořil nový atribut „is_town_London“ nabývající hodnot 0, nebo 1).

3.1 Analýza shluků

Zde jsme zvolili SOM [3], které jsou na rozdíl od dále uvažovaných k-means méně náchylné k lokálním optimům [2]. Jednou z oblíbených implementací je Weka[3]. V té nicméně ani po několika hodinách pro 100 vzorků z naší datové sady výpočty nedoběhly. Vyzkoušeli jsme tedy implementaci *java-ml*[4] a skončili s obdélníkovou mřížkou 5x5, 1000 iteracemi, koef. učení 0.5 a počátečním radiusem 8, která vygenerovala 24 nejprůkaznějších shluků během několika minut. Pro vyhodnocení jsme využili bodové diagramy (BD; pomocí *java-ml* nelze sestavit U-matici). Shluky ukazovaly korelaci finančního sektoru (FS) v Londýně se sentimentem (viz Obr. 1, 2 a 3). Toto není překvapující, Londýn je brán za fin. centrum Velké Británie (VB) s mladší a vzdělanější populací aktivní na sociálních sítích [4]. Lze také vidět korelaci společností z FS VB po referendu (Obr. 1). To naznačuje, že FS reflektoval sentiment na Twitteru. Naopak technologický sektor nekoreloval vůbec (pouze společnosti přímo z VB vykazovaly mírně větší korelační koeficienty - KK), viz Obr. 4.

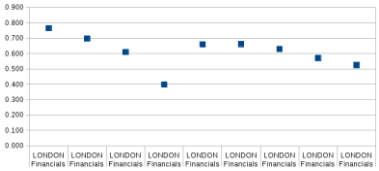


Obr. 1. BD shluku 3. Většina bodů (společnosti ve VB a KK pro hodnoty akcií na konci dne po referendu) má KK větší, než 0.6.

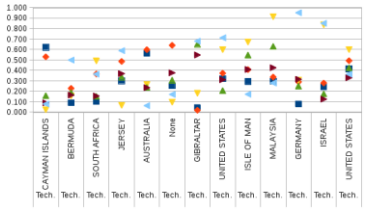


Obr. 2. BD shluku 8. Většina bodů (pro téměř všechny KK finančních společností ve VB) má KK větší, než 0.6.

[3] Weka homepage: http://www.cs.waikato.ac.nz/ml/weka/
[4] java-ml homepage: http://java-ml.sourceforge.net/



Obr. 3. BD shluku 9. Většina bodů (společnosti ve VB a KK pro hodnoty akcií na konci dne před referendem) mají KK větší, než 0.6.



Obr. 4. BD shluku 11. Většina bodů (tech. společnosti mimo VB a jejich KK) mají KK menší, než 0.5.

4 Závěr

Popsaná případová studie Londýnské burzy ukázala největší korelace ve finančním sektoru VB, zjevné u společností sídlících v Londýně. Naopak, technologický sektor nevykazoval korelaci žádnou. Jak je popsáno v [4] a [5], data z Twitteru (a sociálních sítí obecně) často postrádají demografické, věkové, či vzdělanostní údaje. Pro další výzkum bychom měli uvažovat doplnění těchto informací. Je také nutné uvažovat, že uživatelé sociální sítě představují specifickou skupinu obyvatelstva [5].

Je-li Apache Spark v clusteru s jedním uzlem, můžou omezené zdroje (především RAM) způsobit pomalý výběr z datových rámců. Přepsáním kódu spojení datových rámců jsme se vyhnuli paměťově náročným operacím, což umožnilo zredukovat výpočetní čas z několika dní na dvě hodiny. Tento postup byl dále paralelizovatelný.

Použití SOM umožnilo analýzu fenoménu Brexitu se stabilními a prokazatelnými výsledky napříč iteracemi. Lišila se efektivita knihoven, (implementace *java-ml* byla mnohem rychlejší, než Weka, což může být rovněž přínosné pro další výzkumníky).

Výzkum byl v plné verzi publikován anglicky na mezinárodní konferenci.

Literatura

1. A. J. Awan, M. Brorsson, V. Vlassov and E. Ayguade, "Performance Characterization of In-Memory Data Analytics on a Modern Cloud Server," *Big Data and Cloud Computing (BDCloud), 2015 IEEE Fifth International Conference on*, pp. 1-8, 2015.
2. F. Bação, V. Lobo and M. Painho, "Self-organizing Maps as Substitutes for K-Means Clustering," *Computational Science -- ICCS 2005: 5th International Conference, Atlanta, GA, USA, May 22-25, 2005, Proceedings, Part III*, pp. 476-483, 2005

3. T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52-65, January 2013. [Online]. Available: http://aisii.azc.uam.mx/mcbc/Cursos/IntCompt/Lectura 8. SOM.pdf. [Accessed March 10, 2017].
4. L. Vasiliu, R. McDermott, M. Zarrouk, M. Hürlimann, B. Davis, T. Daudert, M. B. Khaled, D. Byrne, S. Fernández, A. Freitas, F. Caroli, S. Handschuh and Angelo Cavallini, "In or Out? Real-Time Monitoring of BREXIT sentiment on Twitter," *CEUR Workshop Proceedings*, 2016.
5. M. Hürlimann, B. Davis, K. Cortis, A. Freitas, S. Handschuh and S. Fernández, "A Twitter Sentiment Gold Standard for the Brexit Referendum," *Proceedings of the 12th International Conference on Semantic Systems*, pp. 193-196, 2016.

Poděkování: Tento článek vznikl díky finanční podpoře grantů: Grantová agentura ČR- GACR P103/15/06700S, Grant SGS č. SGS 2017/134, VŠB-Technická univerzita Ostrava, MŠMT Národní program udržitelnosti (NPU II) projekt ”IT4Innovations excellence in science - LQ1602”.

Annotation:

Apache Spark is a common big data analysis platform on large computer clusters. It uses primarily the main memory. We added a SOM library to such big data analytical stack. The result is effective and fast enough even on a single computer. This approach brings the possibility of big data analysis even for researchers with limited resources. Our idea was experimentaly tested and is described here. We used the Twitter data (tweets for #brexit hashtag) and their sentiment analysis for finding correlations with stock exchange data as a case study for our approach.

Získávání dat z bibliografických databází

Dalibor Fiala

Katedra informatiky a výpočetní techniky
Západočeská univerzita v Plzni
Univerzitní 8, 306 14 Plzeň

dalfia@kiv.zcu.cz

Abstrakt. Známé bibliografické databáze splňují několik funkcí a mohou v zásadě sloužit jako vyhledávače odborných publikací, citační indexy nebo kalkulačky bibliometrických indikátorů. Mezi nejznámější z nich patří Web of Science, Scopus, ACM Digital Library, DBLP, CiteSeer^X a Google Scholar. Větší množství dat z těchto databází je možno s úspěchem použít mj. pro bibliometrická měření, analýzu citačních sítí a sítí spolupráce a pro vizualizaci produktivity a kvality vědeckého výzkumu. Tato data lze v některých případech získat pouze ručně, ale v jiných i automatizovaně. V tomto příspěvku si uvedeme přehled bibliografických databází a možnosti získávání dat z nich a podrobněji se zaměříme na databáze Web of Science a Scopus.

Klíčová slova: WoS, Scopus, DBLP, CiteSeer, ACM DL, Google Scholar.

1 Úvod

V tomto příspěvku budeme kvůli zjednodušení nazývat bibliografickou databází každý systém umožňující v různé míře vyhledávání odborných publikací, poskytování bibliografických informací o nich, procházení referencí a citací a přístup k jejich abstraktům nebo dokonce plným textům. V zásadě je tedy podle funkcionality můžeme rozdělit do tří vzájemně se nevylučujících skupin: vyhledávače odborných publikací, citační indexy a „kalkulátory" bibliometrických indikátorů. Databáze v první skupině umožňují typicky vyhledávání mj. v názvech článků, jménech autorů i jejich adresách, klíčových slovech, abstraktech a někdy i plných textech. Citační indexy poskytují především možnost nalézat citované a citující publikace a sledovat tak vývoj nějaké úzce vymezené problematiky určitého vědního oboru. Poslední kategorie bibliografických databází je zaměřena na měření nejrůznějších aspektů publikační činnosti, např. počty publikací a citací, h-index, impaktní faktor aj.

Mezi nejznámější bibliografické databáze patří Web of Science (dříve ISI, Thomson Reuters, nyní Clarivate Analytics [1]), Scopus [2], ACM Digital Library (rozšířená verze známá jako ACM Portal nebo Guide [3]), DBLP [4], CiteSeer^X (dříve CiteSeer [5]) a Google Scholar [6]. Některé jejich vlastnosti přehledně shrnuje tabulka 1 a jimi se budeme dále zabývat. Mimo ně samozřejmě existují i mnohé další, které

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 72-76.

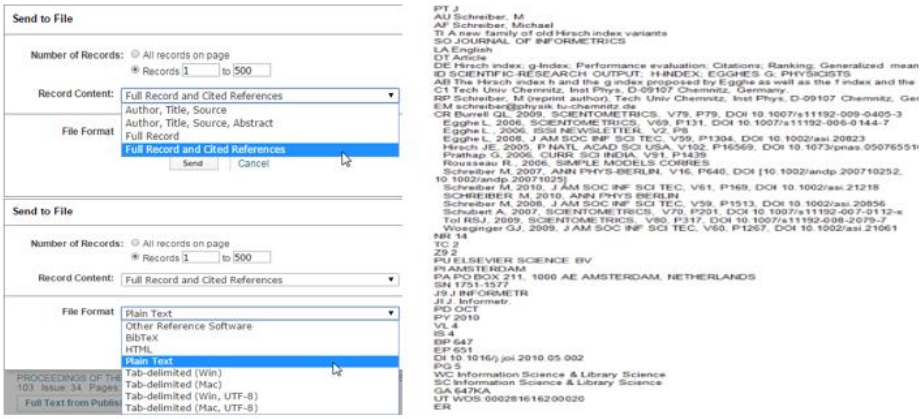
nejsou tématem tohoto příspěvku, jako např. IEEE Xplore [7], PubMed [8], arXiv [9], Microsoft Academic [10], SemanticScholar [11] atd.

Tab. 1. Tabulka vlastností vybraných bibliografických databází (červen 2017)

	ACM DL (Guide)	CiteSeer^{X}	DBLP	Google Scholar	Scopus	Web of Science
Zdarma	částečně	ano	ano	ano	ne	ne
Automatizovaný	ne	ano	ne	ano	ne	ne
Počet záznamů	2,68 mil.	10 mil.	3,81 mil.	100+ mil.	75 mil.	62,25 mil.
Vše ke stažení	ne	ano	ano	ne	ne	ne
Propojení referencí	ano	ano	ne	ne	ano	ano
Propojení citací	ano	ano	ne	ano	ano	ano
Počet citací článku	ano	ano	ne	ano	ano	ano
Počet citací autora	ano	nepřímo	ne	nepřímo	ano	ano
Vědní obor	informatika	informatika	informatika	všechny	všechny	všechny

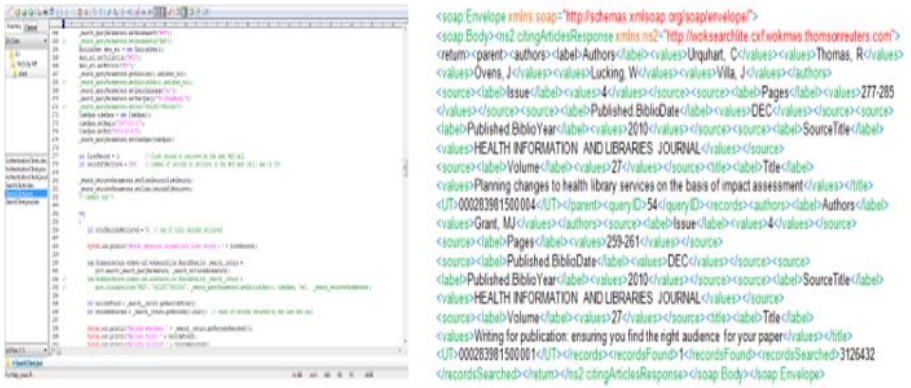
2 Export dat

Některé uvedené databáze jsou zdarma a poskytují dokonce všechna svá data ke stažení ve formě jednoho obřího souboru XML (DBLP) nebo na vyžádání jako celý *dump* databáze MySQL (CiteSeer^{X}). Google Scholar, jenž je rovněž zdarma, žádnou takovou možnost nenabízí a dokonce (zřejmě záměrně) neposkytuje ani žádné programátorské rozhraní (API) pro přístup ke svým datům. To samé platí i pro částečně zpoplatněnou ACM DL. V obou případech by tedy získání většího množství dat bylo možné pouze programově automatizovaným webovým pavoukem (robotem) se všemi problémy a omezeními s tím spojenými. U dvou zbývajících placených databází Web of Science (WoS) a Scopus je situace pochopitelně odlišná. Na obr. 1 je vidět ruční export bibliografických záznamů (maximálně 500 najednou) do čistého textu ve Web of Science:



Obr. 1. Ruční export záznamů z databáze Web of Science

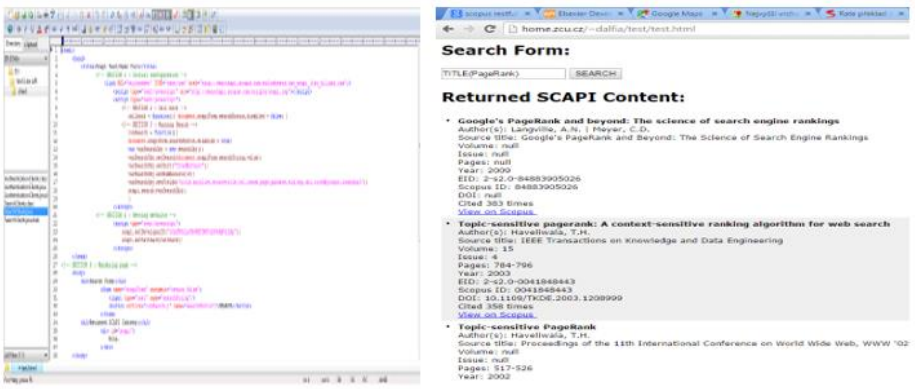
Omezení počtu najednou exportovaných záznamů je i ve Scopusu: 2000. Pokud se spokojíme s jen minimalistickými informacemi o publikacích (v zásadě jen název, autor, místo a rok vydání), zvyšuje se tento limit až na 20000 ve Scopusu a nově též na 5000 ve WoSu. Pro export řádově vyšších počtů bibliografických záznamů bude tedy vhodné využít API, které obě databáze nabízejí, a záznamy stahovat automaticky pro tento účel vytvořeným programem. Na obr. 2 je ukázka programového kódu a výsledného kusu stažených dat ve formátu XML, jak ho poskytovala starší „odlehčená“ verze WoS API, tzv. *Web Services Lite*.



Obr. 2. Automatický import záznamů z databáze Web of Science (*Web Services Lite*)

Kromě této odlehčené verze je rovněž k dispozici API *Web Services Expanded* poskytující i mnoho dalších údajů o článcích jako jsou např. počty citací, adresy autorů nebo abstrakty [12]. Všech rozhraní lze užívat jen po registraci a získání přístupového klíče. Podobně na obr. 3 je ukázka programového kódu, který přes Scopus API [13],

kdysi označované jako *SCAPI*, zobrazuje importované záznamy ve webové aplikaci, jež je na rozdíl od WoS API povinným cílovým místem využití stahovaných dat:



Obr. 3. Automatický import záznamů z databáze Scopus (*SCAPI*)

3 Závěr

Dosud nezmiňovanou možností získání dat je jejich nákup přímo od provozovatele databáze. Tuto možnost sám autor tohoto příspěvku vyzkoušel u databáze Web of Science. Výhodou je dodání dat „na míru“ podle přesně zadaných kritérií ve formátu XML. Nevýhodou je vysoká cena, a to i pro pracovníka instituce, která je standardním předplatitelem této databáze a má k ní přístup přes webové rozhraní. V každém případě mají získaná data cenné využití, ať už v bibliometrických studiích, grafových analýzách nebo obecně jakýchkoliv jiných experimentech nad rozsáhlými daty.

Literatura

1. Web of Science: http://clarivate.com/scientific-and-academic-research/research-discovery/web-of-science/. Získáno 28. 6. 2017.
2. Scopus: https://www.scopus.com/. Získáno 28. 6. 2017.
3. ACM Digital Library: http://dl.acm.org/. Získáno 28. 6. 2017.
4. DBLP: http://dblp.org/. Získáno 28. 6. 2017.
5. CiteSeer^{X}: http://citeseerx.ist.psu.edu/. Získáno 28. 6. 2017.
6. Google Scholar: https://scholar.google.com/. Získáno 28. 6. 2017.
7. IEEE Xplore: http://ieeexplore.ieee.org/. Získáno 28. 6. 2017.
8. PubMed: https://www.ncbi.nlm.nih.gov/pubmed/. Získáno 28. 6. 2017.
9. arXiv: https://arxiv.org/. Získáno 28. 6. 2017.
10. Microsoft Academic: https://academic.microsoft.com/. Získáno 28. 6. 2017.
11. Semantic Scholar: https://www.semanticscholar.org/. Získáno 28. 6. 2017.
12. Web of Science Data Integration: http://ip-science.interest.thomsonreuters.com/data-integration/. Získáno 28. 6. 2017.
13. Scopus APIs: https://www.elsevier.com/solutions/scopus/features/api/. Získáno 28.6. 2017.

Poděkování: Tato publikace byla podpořena projektem LO1506 Ministerstva školství, mládeže a tělovýchovy ČR.

Annotation:

Data Acquisition from Bibliographic Databases

The established bibliographic databases have a number of functionalities and can, in principle, serve as search engines of academic papers, citation indices, or calculators of bibliometric indicators. Web of Science, Scopus, ACM Digital Library, DBLP, CiteSeer^{X}, and Google Scholar belong to the best known ones. Larger amounts of data from these databases can be successfully used for bibliometric measurements, citation and collaboration networks analysis, and for the visualization of the production and quality of scientific research. These data can be acquired only manually in some cases but also automatically in some others. In this short paper we will give an overview of bibliographic databases and the possibilities of data acquisition from them and we will focus in more detail on Web of Science and Scopus.

SISel: Aviation Safety Powered by Semantic Technologies

Martin Ledvinka, Petr Křemen, Bogdan Kostov and Miroslav Blaško

Department of Cybernetics, FEE CTU in Prague
Karlovo náměstí 13, 121 35 Praha 2

martin.ledvinka@fel.cvut.cz

Abstract. Aviation safety is a complex domain with a potential of big losses in human lives and property. It encompasses not only media-followed accidents, but also minor incidents, regulation violations and, most importantly, their prevention. It is necessary to constantly monitor and improve the safety. In this paper, we show how we built SISel - a safety intelligence system based on a conceptual description of the domain. Semantic technologies played a key role in the development of this system – the conceptual model is based on the Unified Foundational Ontology, the data is stored in an RDF triple store and accessed using the ontological persistence framework JOPA. The system is already deployed at the Czech Civil Aviation Authority, where it is used to gather and analyze safety data.

Keywords: Aviation safety, Conceptual model, Ontology, Information system.

1 Introduction

Aviation safety supervision, carried out by the Civil Aviation Authority (CAA) in the Czech Republic, is a complex domain. CAA performs safety audits in organizations like operators or aerodromes, it receives reports about safety occurrences including accidents, rule violations and incidents. It also coordinates with the European Aviation Safety Agency (EASA). All these agendas were done separately, with almost no coordination between the corresponding departments in the CAA. Each department maintained its own documentation, often in different formats (mostly text documents or Excel files).

It is clear that the aforementioned agendas influence each other and their integration would provide a more general, unified view of the safety management and could lead to improvements in the area. Our research group at FEE CTU in Prague, together with the Department of Air Transportation at FT CTU in Prague and dolphin consulting s.r.o. embarked on a journey that would provide foundations for this unification.

We developed SISel – a safety intelligence system, which unifies some of the agendas of CAA and provides a more complex overview of the domain. In this paper, we briefly introduce SISel and show how semantic technologies helped in creating the system.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 77-82.

2 SISel

SISel is a safety intelligence system built upon conceptualization of the aviation safety domain developed with the help of domain experts and expressed using an ontology. The overall architecture of the system can be seen in Figure 1.

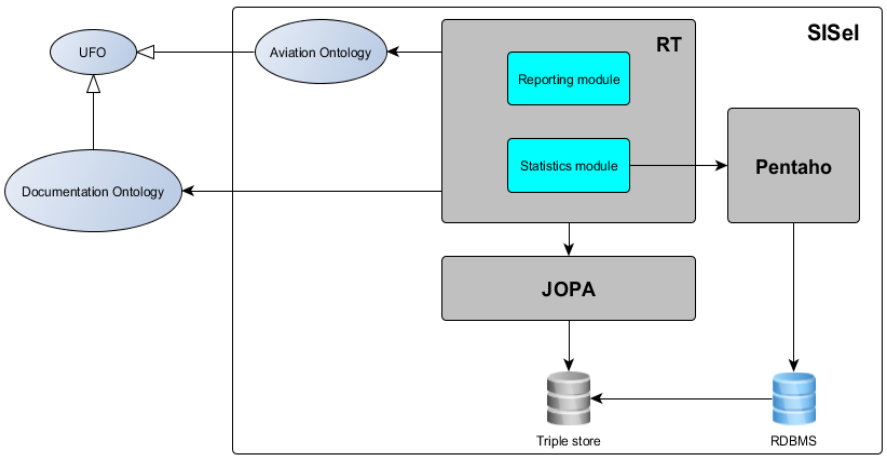


Figure 1. Overall architecture of SISel.

2.1 Conceptual Model of the Domain

The conceptual model of the aviation safety domain (the Aviation ontology in Figure 1) is built upon the Unified Foundational Ontology (UFO) [1]. It represents safety occurrences as UFO *events*, which are documented by *reports*. Reports themselves are a kind of *documents*. Events have *participating agents*, which can be for example persons or organizations, and a number of other data and object properties, for instance temporal and spatial information and sub-events.

This conceptualization allows a better understanding of the whole domain. Building the conceptual model based on the upper level UFO ontology makes it more interoperable. In fact, the conceptual model of SISel stems from the *documentation ontology* [2], which is not specific to the aviation domain, so reports from SISel can be integrated with other reports for example from the railroad transportation domain, provided their conceptualization is also built upon the documentation ontology. A detailed discussion of the aviation safety ontology can be found in [2].

2.2 SISel Reporting Tool

The main (or most visible) part of SISel is the Reporting Tool (RT). It is a safety reporting information system built on top of the conceptual model of the domain. The RT supports several kinds of reports:

- Occurrence reports – reports of safety occurrence like bird strikes, separation minima infringements etc. (mandatory in the Czech Republic),
- Audit reports – reports of audits performed by the CAA auditors containing findings discovered by the audits and their corrective measures,
- Safety issues – reports of problematic data patterns. Safety issues are usually created based on occurrence reports or audit findings and represent patterns likely to be problematic for the aviation safety.

An important possibility of the RT is to model chains of factors. This is most relevant to the occurrences, where the factors represent events, which were a part of or did influence the occurrence. The RT contains a simple chain designer component, in which these events can be added, arranged and connected using links representing relationships such as causality. Detailed information about the events can then be filled in. SISel RT uses the context of the type of the occurrence (or event) to require only relevant information (this relevance is determined by domain experts). To support this kind of dynamic selection, the RT uses forms generated from a declarative ontological description. This means that various kinds of events provide the user with different forms. Neither the chain designer, nor the dynamic forms are supported by reporting tools provided by EASA.

The RT allows standard report management and editing functions. In addition, it allows importing the reports from the email box, which receives occurrence reports from participants in the Czech aviation and occurrence and audit reports from the European systems ECCAIRS [3] and SAFA[1]. These reports can also be manually imported into the system. Another possibility is to synchronize the reports directly with the ECCAIRS system.

SISel also contains a statistical module, which is built using the open-source Pentaho business intelligence platform. The data is exported into Pentaho's database every night. Analytical dashboards created by dolphin consulting s.r.o. are then populated by the exported data. These dashboards include statistics on the frequency of occurrence types, the average severity of occurrences, the most frequent factors etc.

Semantic Technologies in SISel

As we have already mentioned, the conceptual model of the aviation domain is in fact an ontology. We use this ontology to generate portions of the object model of the reporting tool.

The domain also contains a number of taxonomies and predefined value lists, for instance there are predefined sets of occurrence categories and event types. All these value lists have been transformed into an RDF form and stored in a triple store. There is an internal service, which uses SPARQL queries to retrieve relevant lists of possible values from these value lists. Output of this service is then used to populate selection and autocomplete components in the RT user interface.

[1] http://www.eraa.org/sites/default/files/SAFA%20Forum%202012%20EASA%20presentation.pdf, accessed 2017-06-13.

The RT is a standard web application written in Java using the Spring framework. It has a JavaScript-based user interface (UI), which communicates with the backend via REST services. Such setup is common in today's Java web applications. The unusual thing about the RT is that it is based on the ontological conceptual model and stores its data in an RDF triple store.

To access the data, the RT uses a framework called JOPA [4], which maps ontological data to a Java object model and vice versa. Thanks to JOPA, the overall architecture of the RT very closely resembles any other Java application running on top of a relational database. This makes the development and maintenance of the RT easier.

In addition, the form descriptions are also stored in an ontology and the form structure is sent to the client in the JSON-LD format.

3 Deployment

SISel is currently deployed at the Czech CAA. The CAA employees use it mainly to model factor chains of occurrences reported to them. The statistical reports from SISel are then examined by the Safety Action Group – a board of safety experts established at the CAA. The system imports several reports every day; usually they are received in the CAA's email box, which is connected to the SISel RT. The total number of reports in the system as of June 2017 is over 1700 and it has been in daily use for the last eight months.

4 Example

We shall illustrate SISel usage on a real world example which demonstrates the usual way CAA employees work with SISel. An accident report sent by the Air Navigation Services (ANS) of the Czech Republic was received by CAA via the dedicated email box. The email contained a file in the E5X format. E5X is a file format used by the European Co-ordination Centre for Accident and Incident Reporting Systems (ECCAIRS), which is a safety occurrence reporting system whose usage is mandated by EASA throughout Europe. Technically, E5X is a compressed archive which contains one or more XML documents.

SISel RT imported the report from the received E5X attachment. A responsible person at CAA then reviewed the report, conducted an independent classification of the occurrence and its severity and designed a chain of factors that contributed to the accident. An excerpt of data from the resulting report can be seen in Figure 2. Data from this report are then manifested in the output of the analytical module of SISel.

The accident was later investigated (completely independent of CAA) by the Air Accident Investigation Institute (AAII) and a report was published[2].

As was stated, EASA mandates use of ECCAIRS for safety occurrence reporting throughout Europe. Therefore, a report of the incident was also entered into ECCAIRS by the employees at AAII, who manage the ECCAIRS instance in the

[2] http://www.uzpln.cz/incident/506 (Czech only), accessed 2017-06-03.

Czech Republic. Since SISel supports integration with ECCAIRS, CAA inspectors are able to review the latest version of the report of the same occurrence in ECCAIRS and compare it to the version they created in SISel.

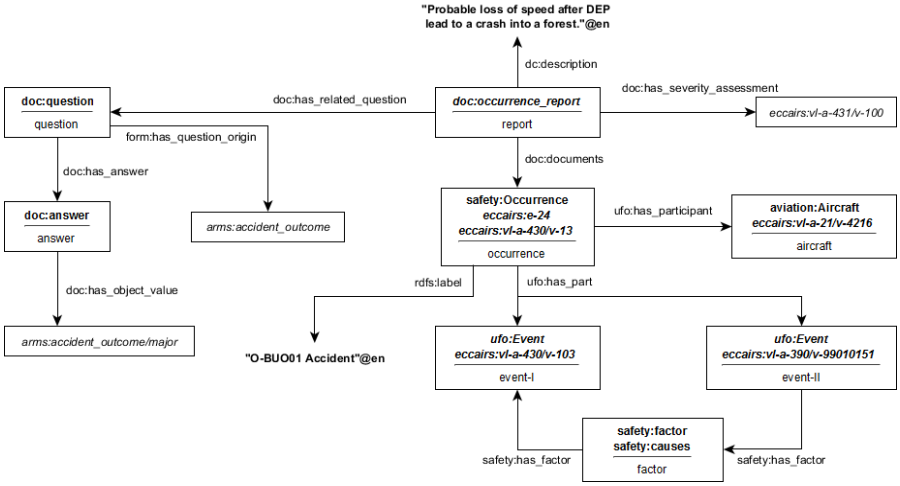


Figure 2. Report example visualization. IRIs are shortened, prefixes correspond to the respective ontologies/vocabularies. Bold labels above the horizontal line in nodes represent types of the instances, labels in italics refer to external vocabularies. Nodes without borders represent literal values.

5 Conclusions

We presented a case of semantic technologies being employed in an information system used on a daily basis by a non-academic organization. The system shall serve as a foundation of a larger ecosystem of applications and processes, which is planned for future development and will concern aviation safety in the Czech Republic. This ecosystem has also a potential of expansion into other domains and other countries.

Acknowledgement: This work was supported by grant No.SGS16/229/OHK3/3T/13 Supporting ontological data quality in information systems of the Czech Technical University in Prague and by grant TB0400MD010 of the Technology Agency of the Czech Republic.

References

1. Guizzardi, G: Ontological Foundations for Structural Conceptual Models. Ph.D. thesis, University of Twente (2005).
2. Kostov, B., Ahmad, J. and Křemen, P.: Towards Ontology-Based Safety Information Management in the Aviation Industry. In: On the Move to Meaningful Internet Systems:

OTM 2016 Workshops, I. Ciuciu, Ch. Debruyne, H. Panetto, G. Weichhart, P. Bollen, A. Fensel and M.-E. Vidal (Eds.), Springer International Publishing, 2017.

3. Křemen, P., Kostov, B., Blaško, M., Ahmad, J., Plos, V., Lališ, A., Stojić, S. and Vittek, P.: Ontological Foundations of European Coordination Centre for Accident and Incident Reporting Systems, Journal of Aerospace Information Systems, Vol. 14, No. 5 (2017), pp. 279-292.
4. Křemen, P., Kouba, Z.: Ontology-Driven Information System Design. IEEE Transactions on Systems, Man, and Cybernetics: Part C 42(3), 334–344 (May 2012).

Projektové příspěvky

Analýza zpravodajských textů a jejich komentářů napříč jazyky

Josef Steinberger

Katedra informatiky a výpočetní techniky, Nové technologie pro informační společnost, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 8, 306 14 Plzeň

jstein@kiv.zcu.cz

Abstrakt. Tento příspěvek představuje projekt MediaGist, jehož cílem je vytvoření online systému, který analyzuje a propojuje zpravodajské články a jejich komentáře v pěti jazycích. Umožňuje novinářům detekovat a prozkoumávat zpravodajská témata, která jsou kontroverzně reportována nebo diskutována napříč různými jazyky/zeměmi. Sumarizace a analýza polarity textu jsou dvě hlavní technologie použité v textové analytice. Polarita slouží k výpočtu kontroverze a souhrny pomáhají zkoumat rozdíly.

Klíčová slova: analýza textu, sumarizace, analýza polarity textu.

1 Úvod

Zpravodajské portály publikují tisíce článků každý den v různých jazycích. Vyznat se v takto ohromném množství informací je bez automatických nástrojů nemožné. Existuje řada agregátorů a analyzátorů zpravodajství a každý má své silné stránky. Google News shlukuje články a zobrazuje příběhy podle zájmů čtenáře. IBM Watson News Explorer přináší analytický způsob čtení zpráv skrze vizualizaci pomocí *linked data*. Europe Media Monitor (EMM) shlukuje zpravodajství téměř v reálném čase ve více než padesáti jazycích [1]. Ovšem existuje ještě další hodnotný zdroj informací na zpravodajských portálech: komentáře k článkům, ze kterých lze dolovat názory veřejnosti na témata zpravodajství. Zahrnutí komentářů do analýzy přináší mnoho nových možností pro novináře, agentury, které studují veřejné mínění a částečné také pro čtenáře. Kontroverzní témata, jako např. migrační krize nebo skandál s emisemi VW, a jejich vnímání v různých zemích může být zdrojem pro další zpravodajství. Zaměření na tyto témata přináší více čtenářů a bohaté diskuze na zpravodajské portály. Mezinárodní agentury a politické instituce mohou využít srovnání vnímání témat v různých zemích k analýze veřejného mínění. Krosjazyčně organizované články a komentáře mohou také využít sami čtenáři, kteří žijí v multikulturním prostředí. Mohou rychle najít a pochopit různé pohledy na kontroverzní témata.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 84-88.

MediaGist[1] [8] je online systém, který je postaven na funkcionalitě agregátorů zpráv, navíc ale přidává dimenzi komentářů. To, že propojuje shluky článků v různých jazycích a analyzuje také komentáře, umožňuje objevovat a prozkoumávat témata, která jsou kontroverzně reportována nebo diskutována v různých zemích.

Následující sekce je věnována technologii, na které je MediaGist postaven, třetí sekce ukazuje stručně jeho funkcionalitu a závěr rozebírá další směry vývoje.

2 Technologie

Zpracování dat v MediaGistu začíná crawler. Ten sbírá články a komentáře pod nimi z předdefinovaných zpravodajských portálů[2]. Pro každý článek je vytvořen RSS soubor, který putuje dále přes další moduly zpracování přirozeného jazyka.

Nejprve jsou rozpoznány pojmenované entity (NER), jak v článku, tak v jeho komentářích, a je jim přiřazeno krosjazyčné ID. NER modul je založen na JRC-Names[3], což je vícejazyčný seznam jmen osob a organizací. Různé varianty téhož jména jsou propojeny stejným ID [9]. Množina výskytů entit je rozšířena detektorem referencí, který rozpoznává dva druhy referencí: části jmen (např. „Zeman“ nebo definitní popisy, např. „český prezident“) [5].

Dalším krokem je přiřazení polarity ke každému článku, komentáři a také každému výskytu entity. Skóre polarity je v intervalu <-100,+100>. Analyzátor polarity používá vícejazyčné a porovnatelné slovníky vzniklé automatickou triangulizací (viz [7]). V případě přiřazení polarity článku nebo komentáři počítá subjektivní slova. V případě přiřazení výskytu entity se omezuje pouze jeho okolí. Navíc obsahuje pravidla pro měření intenzity subjektivního výrazu nebo negace [6]. Přestože strojového učení by lépe predikovalo polaritu, v současné době nemáme k dispozici vhodná trénovací data ve všech cílených jazycích.

U každého článku můžou být i tisíce komentářů. Dalším krokem je tedy jejich automatická sumarizace. Sumarizace v MediaGistu je založena na extraktivním přístupu založeném na latentní sémantické analýze, který používá jak slovní tak entitní příznaky [3]. Tento krok výrazně redukuje velikost dat, která se posílají dalším modulům. Obohacené RSS soubory pak vstupují do fáze shlukování.

Každé čtyři hodiny, pro každý jazyk samostatně, načte shlukovací modul soubory článků publikované během aktuálního týdne a vytvoří jednojazyčné shluky. Je použito aglomerativní hierarchické shlukování se strategií *group average* [2]. Články jsou reprezentované *log-likelihood vektory* a podobnostní funkcí je *kosinus*. Od tohoto kroku obsahuje RSS informace o všech článcích shluku. Krosjazyčný linker pak propojí nejpodobnější shluky mezi jazyky. Linker používá dva typy příznaků: přítomnost

[1] MediaGist je dostupný na adrese: *http://mediagist.eu*. Video, které jej představí lze shlédnout zde: *https://www.youtube.com/watch?v=ONtKw_l6_X4*.

[2] Momentálně MediaGist sbírá data z 8 zdrojů v pěti jazycích: angličtina (theguardian.com), čeština (idnes.cz, ihned.cz, novinky.cz), italština (corriere.it, repubblica.it), francouzština (lemonde.fr) a němčina (spiegel.de).

[3] https://ec.europa.eu/jrc/en/language-technologies/jrc-names.

stejných entit a výskyt stejných deskriptorů z thesauru *EuroVoc*[4] [10]. Posledním krokem je vytvoření souhrnu článků a souhrnu komentářů (které již byly sumarizovány na úrovni článku) pro každý shluk.

RSS soubor tak obsahuje všechny informace pro prezentační vrstvu založenou na technologii java servletů a JSP: http://mediagist.eu.

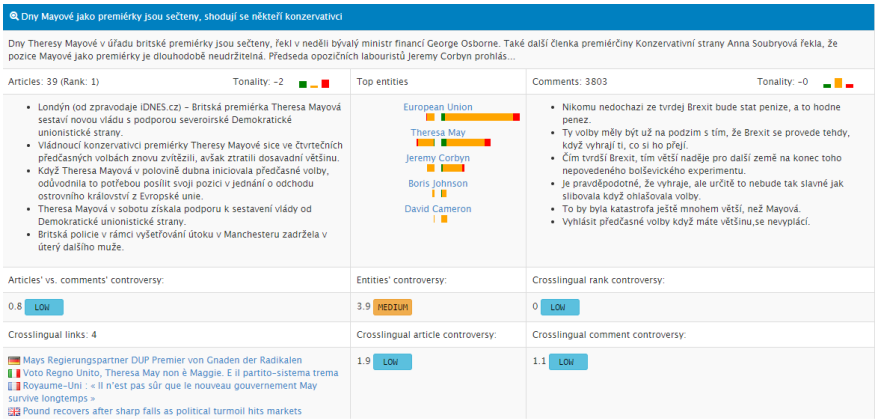
3 Funkcionalita

Systém obsahuje dva pohledy k prozkoumávání dat z médií: pohled na shluky článků (~témata) a pohled na entity. Po výběru jazyka, týdne a řazení dat se zobrazí detaily pohledu a také nejvýznamnější témata/entity analyzovaného výběru v levém panelu.

U každého tématu je zobrazen název a popis, který odpovídá centrálnímu článku. (viz obr. 1). V levé části jsou informace o článcích a v pravé o komentářích. Na obou stranách jsou zobrazeny vygenerované souhrny a agregované údaje o polaritě. Centrální část pohledu ukazuje entity a distribuci jejich polarity v článcích a v komentářích. Dole se nachází odkazy na asociované shluky v dalších jazycích.

MediaGist počítá také několik skóre kontroverze:

- Články vs. komentáře: směrodatná odchylka polarit na obou stranách.
- Entitní kontroverze: porovnává polaritu entit v článcích a komentářích.
- Kontroverze pořadí: porovnává významnost tématu v různých jazycích (dle počtu článků).
- Kontroverze mezi články v různých jazycích: vysoké číslo odpovídá velkému rozdílu polarit článků mezi jazyky.
- Kontroverze mezi komentáři v různých jazycích: vysoká hodnota indikuje témata, která jsou diskutovaná s různou polaritou napříč jazyky (~zeměmi).



Obr. 1. Ukázka hlavního shluku v češtině v týdnu 11.-17. června 2017.

[4] http://eurovoc.europa.eu.

U entit systém ukazuje nalezené varianty jména a jejich frekvence, agregovanou polaritu v článcích a v komentářích a také nejčastější zabarvené pojmy, které pomáhají vysvětlit detekovanou polaritu. Protože máme entity také propojené mezi jazyky, lze spočítat jejich kontroverzi mezi reportováním o dané entitě (články) a veřejném mínění (komentáře) v různých jazycích (~zemích).

4 Závěr

V projektu MediaGist se pracuje na jazykových technologiích, které pomohou detekovat kontroverzi v mezinárodním zpravodajství. Detekce polarity textu identifikuje kontroverzní témata a entity mezi jazyky a skrze sumarizaci je možné data detailněji prozkoumávat. Mimo vylepšení jednotlivých technologií (detekce polarity textu, sumarizace, NER, detekce referencí, shlukování, propojení mezi jazyky, výpočet kontroverze) je cílem dalšího vývoje rozšířit množství dat jak vertikálně (množství zdrojů/jazyků), tak horizontálně (historická data). Systém momentálně konzumuje nefiltrované komentáře. Odstranění osobních útoků, nerelevantních k tématu, a textů od trolů [4] povede ke zvýšení přesnosti detekce názorů uživatelů online světa.

Literatura

1. Atkinson, M. and E. van der Goot: Near real time information mining in multilingual news. In: Proceedings of the 18th International WWW Conference (2009), 1153-1154.
2. Hastie, T., R. Tibshirani, and J. Friedman: The Elements of Statistical Learning. Springer-Verlag, 2009.
3. Kabadjov, M., J. Steinberger, and R. Steinberger: Multilingual statistical news summarization. In: Multilingual Information Extraction and Summarization, volume 2013 of Theory and Applications of Natural Language Processing, Springer (2013), 229-252.
4. Mihaylov, T., G. Georgiev, and P. Nakov: Finding opinion manipulation trolls in news community forums. In: Proceedings of the 19th CoNLL, ACL (2015), 310-314. ACL.
5. Steinberger, J., J. Belyaeva, J. Crawley, L. Della-Rocca, M. Ebrahim, M. Ehrmann, M. Kabadjov, R. Steinberger, and E. Van der Goot: Highly multilingual coreference resolution exploiting a mature entity repository. In: Proceedings of the 8th RANLP Conference, Incoma Ltd. (2011), 254-260.
6. Steinberger, J., P. Lenkova, M. Kabadjov, R. Steinberger and E. van der Goot: Multilingual entity-centered sentiment analysis evaluated by parallel corpora. In: Proceedings of the 8th RANLP Conference, Incoma Ltd. (2011), 770-775.
7. Steinberger, J., M. Ebrahim, M. Ehrmann, A. Hurriyetoglu, M. Kabadjov, P. Lenkova, R. Steinberger, H. Tanev, S. Vzquez, and V. Zavarella: Creating sentiment dictionaries via triangulation. In: Decision Support Systems (2012), 53(4), 689-694.
8. Steinberger, J.: MediaGist: A cross-lingual analyser of aggregated news and commentaries. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, ACL (2016), 145-150.
9. Steinberger, R., B. Pouliquen, M. Kabadjov, J. Belyaeva, and E. van der Goot: JRCNames: A freely available, highly multilingual named entity resource. In: Proceedings of the International RANLP Conference. Incoma Ltd. (2011).

10. Steinberger, R.: Multilingual and cross-lingual news analysis in the europe media monitor (EMM). In: Multidisciplinary Information Retrieval, LNCS 8201, Springer (2013), 1-4.

Poděkování: Tento článek byl podpořen projektem MediaGist, EUs FP7 People Programme (Marie Curie Actions), č. 630786.

Annotation:

A crosslingual analyser of news and their commentaries

The paper introduces project MediaGist, which builds an online system for analysing and linking news and commentaries in five languages. It is designed to assist journalists to detect and explore news topics, which are controversially reported or discussed in different countries. Sentiment analysis and summarization are key technologies used for text analytics. Sentiment analysis provides a basis to compute controversy scores and summaries help to explore the differences.

First Insight into the Processing of the Historical Documents from the Period of Totalitarian Regimes

Lucie Skorkovská[1], Petr Neduchal[2], Zbyněk Zajíc[1], Pavel Ircing[2], Luděk Müller[2], Lukáš Bureš[2]

[1]University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society
Univerzitní 8, 306 14 Plzeň
[2]University of West Bohemia, Faculty of Applied Sciences, NTIS - New Technologies for the Information Society and Dept. of Cybernetics
Univerzitní 8, 306 14 Plzeň

lskorkov@ntis.zcu.cz, zzajic@ntis.zcu.cz
{neduchal, ircing, muller, lbures}@ntis.zcu.cz

Abstract. In this paper, we describe the goals and the initial stages of the project "System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes". The main goal of this project is to create an integrated archive of the recordings, documents, and photographs that would be accessible online and would provide multifaceted search capabilities (spoken content, biographical information, relevant time period, etc.). The recordings contain retrospect interviews with witnesses of the totalitarian regimes in Czechoslovakia; the other documents are copies of relevant text material and photographs mainly from home archives. This paper focuses on the processing of the historical documents with the optical character recognition and describes the initial experiments.

Keywords: historical sources processing, optical character recognition, document processing

1 Introduction

The main objective of the project "System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes" is the research and development of software tools for archiving and providing access to the historical resources gathered within the documentary mission of the Institute for the Study of Totalitarian Regimes (USTR)[1].

The Institute for the Study of Totalitarian Regimes studies and impartially evaluates the two times of non-freedom periods of the history of the Czech Republic: the

[1] https://www.ustrcr.cz/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 89-92.

time of the Nazi occupation (1939-1945) and the time of Communist totalitarian power (1948-1989), examines the anti-democratic and criminal activity of state bodies, especially its security services, as well as other organizations based on its ideology. For that purpose, USTR secures and makes accessible to the public the documents related to the time of non-freedom and the time of Communist totalitarian power and converts acquired documents into the electronic form.

Within the documentation activities of the USTR in the years 2008-2015 at least 1000 hours of the audio recordings of the interviews with the witnesses of the totalitarian regimes in Czechoslovakia and 50000 scanned textual documents were created. Nowadays, these documents and recordings are stored on the internal storage of the USTR and are made accessible for the researchers on DVD or through some digital storage services. For only about 160 recordings the text transcription is available, for the rest the researchers must manually go through the whole video recording.

Despite these imperfections, the historic resources gathered in this collection are being used by history experts and researchers from not only the Czech Republic but also from other European countries and USA.

2 Goal of the Project

The main goal of this project is to create an integrated archive of the recordings, searchable documents, and photographs that would be accessible online and would provide multifaceted search capabilities (including the actual spoken content, name and other biographical information, relevant time period, etc.). The archive created in such way would make the work of the researchers more efficient and also would allow a wider scope of interested persons to access these historic resources.

In order to achieve this goal, the methods of automatic speech recognition, automatic indexing, search in recognized recordings, the optical character recognition (OCR) and related techniques of natural language processing will be employed. The rest of the paper describes the first stages of the processing of the scanned documents, which is a challenging task in these circumstances since the source documents are old and of low quality.

3 Optical Character Recognition of the Scanned Typewritten Documents

One of the goals of this project is a transformation of old scans of typewritten documents into searchable text documents. It consists of several tasks that have to be solved. The first one is a development or a selection of an existing Optical Character Recognition (OCR) engine. In the first phase of the project, we have chosen the Google Tesseract OCR engine [3]. The second one is a development of the preprocessing methods that would clear noise and other artifacts in the documents and improve the results of the OCR engine.

There are two additional goals that should also be addressed. The first one is the decomposition of a scanned folder - i.e. group of scans that belong to the same topic or person - into clusters of related documents and the creation of PDF files based on the clusters. The last goal is searching for important meta-information about documents such as its type, title or mentioned persons.

In the first year of the project, we have performed several experiments. In the first experiment, the influence of preprocessing methods on the OCR engine was examined. The second one was focused on the clustering of related documents.

3.1 Preprocessing

During the preprocessing experiment, several methods and their combination were used [4]. One of the most important preprocessing methods is a deskewing (estimation of a skew angle) algorithm. It searches for the rotation that has to be applied in order to get a document with no skew of contained text. A method based on the Fourier Transform was used. The results of this method are promising but it has approximately the same results as the intern deskew algorithm in the Google Tesseract OCR.

We also experimented with the color spaces of the input documents. In the first test, three color variants were compared - particularly red-green-blue (RGB), LAB and gray. The results were approximately 80 % in all cases. In the second phase, the influence of binarization was tested. Results of the binarized documents were slightly better. In the last phase, we propose a different binarization algorithm. Each component of a RGB image is binarized independently. The final image is composed of the binarized components as follows:

$$B_{rgb}(i,j) = \begin{cases} 0 & if\ B_r(i,j) = B_g(i,j) = B_b(i,j) = 0, \\ 1 & otherwise, \end{cases} \quad (1)$$

where $B_{rgb}(i,j)$ is a value of binarized image at coordinates *i,j*. Values B_r, B_g and B_b respectively contained binary value of the image component at coordinates *i,j*. The noise and image artifacts are reduced by this approach. The score of OCR on small dataset was 84% using this approach.

Several other methods were tested during this experiment, particularly histogram equalization [2], [6] and image smoothing algorithms [1]. None of them were significantly successful. All experiments were performed on the small dataset of 25 annotated scans. The score of the used methods was computed using the Levenshtein Distance metric [5].

3.2 Clustering of Consecutive Documents

In the second set of experiments, the decomposition of a document folder into PDF files containing related scans was examined. It is worth mentioning that related scans are usually consecutive. In the first phase, the color similarity was addressed. We have created a feature vector that was composed of the differences of RGB and HSV components computed of the consecutive documents. The K-Nearest Neighbors algo-

rithm (KNN) was trained on the part of a document folder. The test was then performed on two document folders. The first one is the rest of the folder from which KNN was trained. The results are excellent in this case. The second test data was chosen from different document folder. The results are worse in this case - there are clearly non-consecutive scans in one PDF and vice versa. The results seem to be a good starting point for a couple of next experiments. Experiments focused on decomposition based on text extracted by OCR and structural features will be performed in the next phase of the project. But it is not straightforward to pick the right features because consecutive documents can be different in structure and color and vice versa.

4 Conclusion

This paper described the goals of the project ”System for permanent preservation of documentation and presentation of historical sources from the period of totalitarian regimes” and the first stage of the processing of the scanned documents. Based on the results of these first experiments, the subsequent research on the contained text classification and processing can be initiated.

For the OCR experiments, the results were improved by the proposed binarization algorithm. For the scans clustering, it is not straightforward to choose the good features because the consecutive documents can be different in structure and color and vice versa. On the other hand, the results based on the color similarity seem to be a good starting point for further experiments.

References

1. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. International Journal of Computer Vision 81(1), 24 52
2. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., Romeny, B.T.H., Zimmerman, J.B.: Adaptive histogram equalization and its variations. Comput. Vision Graph. Image Process. 39(3), 355-368
3. Smith, R.: An overview of the tesseract ocr engine. In: Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. vol. 2, IEEE (2007), 62 633
4. Sonka, M., Hlavac, V., Boyle, R.: Image processing, analysis, and machine vision. Cengage Learning
5. Yujian, L., Bo, L.: A normalized levenshtein distance metric. IEEE transactions on pattern analysis and machine intelligence 29(6), 109 1095
6. Zuiderveld, K.: Contrast limited adaptive histogram equalization. In: Graphics Gems IV, Heckbert,P.S. (eds.), Academic Press Professional, Inc., San Diego, CA, USA, 47 485

Acknowledgments: This research was supported by the LINDAT/CLARIN, project of the Ministry of Education of the Czech Republic No. CZ.02.1.01/0.0/0.0/16_013/0001781, by the Ministry of Education, Youth and Sports of the Czech Republic project No. LO1506 and by the Ministry of Culture Czech Republic, project No.DG16P02B048.

Projekt MONSOON – návrh platformy pre analýzu veľkých dát v priemysle

Martin Sarnovský[1,2], Peter Bednár[1,2]

[1]Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická Univerzita v Košiciach, Letná 9, 042 00 Košice
[2]Ekonomická fakulta, Technická Univerzita v Košiciach, Letná 9, 042 00 Košice

matin.sarnovsky@tuke.sk, peter.bednar@tuke.sk

Abstrakt. Predkladaný článok predstavuje základnú myšlienku a ciele projektu MONSOON, ktorý je zameraný na oblasť využitia analýzy veľkých dát v priemysle. Hlavným cieľom je predstaviť projekt vrátane jeho hlavných oblastí nasadenia a popísať konceptuálnu architektúru softvérovej platformy používanej v projekte. Charakterizujeme jednotlivé prípady použitia, ktorými sú prediktívne úlohy v dvoch hlavných doménach, produkcii hliníka a výroby plastových výrobkov. Jadrom projektu bude softvérová platforma postavená na technológiách pre spracovanie veľkých dát, ktorá umožňuje ukladanie a spracovanie dát z prevádzok z rôznych odvetví priemyslu. Platforma bude integrovaná so systémami prevádzok a umožní dátovým analytikom vyvíjať, testovať a nasadzovať prediktívne metódy pre optimalizáciu výrobných procesov prevádzok.

Klíčová slova: procesný priemysel, big data technológie, analýza dát.

1 Úvod

MONSOON (MOdel based coNtrol framework for Site-wide OptimizatiON of data-intensive processes)[1] je 3-ročný výskumný SPIRE projekt ktorý je zameraný na návrh a implementáciu infraštruktúry pre využitie analýzy veľkých dát v oblasti procesného priemyslu. Hlavnou úlohou projektu je aplikovanie dátovo-orientovanej metodológie pre optimalizáciu produkčných procesov, ktorá je založená na budovaní prediktívnych modelov. Projekt je založený na koncepte zdieľanej medzi-sektorovej dátovo analytickej platformy, ktorá zhromažďuje a spracováva veľké dáta z viacerých prevádzok. Zdieľanie analytickej platformy medzi jednotlivými odvetviami priemyslu potom umožní, okrem optimalizácie procesov a šetrenia výrobných nákladov, aj transfer najlepších praktík a znalostí medzi jednotlivými doménami. V rámci projektu sú zahrnuté dve prostredia pre validáciu a demonštrovanie možností navrhovanej platformy: spoločnosť pre produkciu hliníka vo Francúzsku a spoločnosť pre výrobu plastových dielov v Portugalsku. V nasledujúcich kapitolách bližšie predstavíme

[1] https://www.spire2030.eu/monsoon

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 93-97.

hlavné prípady použitia v oboch doménach a popíšeme konceptuálnu architektúru navrhovanej analytickej platformy.

2 Prípady použitia

2.1 Produkcia hliníka

V súčasnosti je proces výroby hliníka na veľmi vysokej technologickej úrovni, ktorej optimalizácia má vyše 130 ročnú históriu. Aj napriek tomu však existujú oblasti, v ktorých je možné využiť moderné metódy dátovej analýzy na optimalizovanie výrobného procesu, ako napr. prevencia procesných anomálií v elektrolytických peciach, prediktívna detekcia defektov na anódach a zníženie emisií v ovzduší. Práve produkcia anód odstraňovanie defektov v anódach predstavuje v súčasnosti najväčšiu variabilnú zložku produkčných nákladov pri výrobe hliníka. Ciele pri optimalizovaní môžu byť zamerané na zníženie nákladov pri produkcii kvalitných anód resp. pri optimalizácii prevádzky elektrolytických pecí s cieľom čo najlepšieho vyťaženia kvality anód pri zredukovaní energetických nákladov. V rámci projektu boli stanované dva ciele:

- Prediktívna údržba zariadení – hlavným cieľom je odhalenie procesných odchýlok a zlyhaní zariadení pri výrobe anód, ktoré majú vplyv na ich výslednú kvalitu (napr. na hustotu a homogenitu)
- Predikcia kvality anód – hlavným cieľom je detegovanie chybných anód a ich vyradenie z výrobného procesu.

Prediktívna údržba zariadení je jedným z najbežnejších prípadov využitia analýzy veľkých dát pri optimalizácii výrobných procesov. Hlavným cieľom je skrátenie časov pri výpadkoch. Ako vstupné dáta sa používajú senzorické dáta zozbierané z jednotlivých zariadení, ktoré tvoria produkčnú linku pri výrobe anód. Vstupné dáta sú korelované s historickými záznamami o poruchách zariadení. Cieľom je nielen detegovať známe poruchy, ale aj detegovať neočakávané správanie mimo bežnej prevádzky. V prípade prediktívnej kvality anód je cieľom zahrnúť parametre celého výrobného procesu anód spolu a odvodiť z nich indikátor kvality anódy.

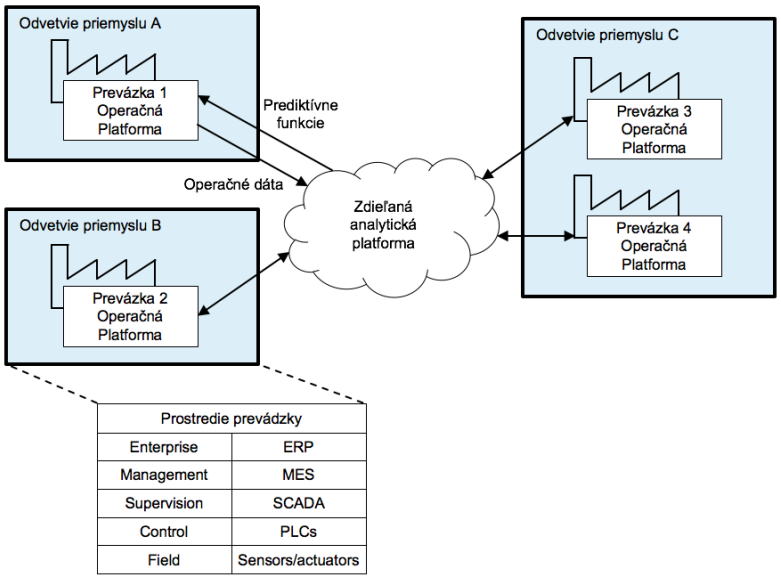
2.2 Výroba plastových dielov

V prípade výroby plastových dielov je projekt zameraný na výrobu dielov technológiou injekčného vstrekovania. Pomocou tejto technológie je možné vyrábať veľké množstvo rôznych typov dielov určených napr. pre medicínsku oblasť, automobilový priemysel, športové potreby, hračky a pod. Produkcia plastových dielov pokrýva masovú výrobu dielov pri ktorých sa nepožadujú vysoké kvalitatívne a estetické normy výrobkov až po kusovú výrobu vysokokvalitných dielov. Charakteristika dielov závisí jednak na použitom materiály a jednak na parametroch výrobného transformačného procesu pri vstrekovaní. V rámci projektu je výskum zameraný na dva prípady. V prvom ide o výrobu kapsúl pre kávomaty, ktoré sú vyrábané vo veľkých počtoch

s nízkou variabilitou a relatívne nízkymi požiadavkami na kvalitu. Do výrobného procesu je zaradený systém kontroly kvality, ktorý pokrýva 100% výrobkov. Druhý prípad v doméne výroby plastových výrobkov reprezentuje výrobu technických dielov pre automobilový priemysel, ktoré zahŕňajú inštaláciu kovových dielov vo výrobku. Okrem procesu vstrekovania, výrobný proces zahŕňa aj proces osádzania kovových dielov. V oboch prípadoch je hlavným cieľom predikovanie porúch zariadení a notifikácia riadiacich pracovníkov, čo umožňuje efektívnejšie operatívne odstavenie výroby s cieľom zamedziť produkcii chybných výrobkov. Ďalším cieľom je optimalizácia procesných parametrov počas vstrekovania s cieľom skrátiť výrobný cyklus jedného výrobku pri zachovaní jeho kvality.

3 Architektúra riešenia

MONSOON platforma pozostáva z dvoch hlavných častí, ktoré sú zobrazené na Obr. 1. *Operačná platforma* je množina komponentov, ktoré sú nasadené priamo v prostredí konkrétnej prevádzky. Hlavnou úlohou operačnej platformy je spracúvať operačné dáta z prevádzky v reálnom čase. Tento komponent je integrovaný a prepojený s rôznymi, existujúcimi a už nasadenými systémami v produkčnom prostredí ako napr. ERP, manažérske systémy, dátami zo zariadení a senzorov, atď.. Relevantné dáta sú následne z operačnej platformy prenášané do zdieľanej analytickej platformy.



Obr. 1. Architektúra MONSOON platformy

Zdieľaná, medzi-sektorová dátová analytická platforma je jadrom MONSOON riešenia. Predstavuje škálovateľné distribuované prostredie, ktoré sa používa pre zber a ukladanie dát z rôznych prevádzok z rôznych odvetví priemyslu. Táto platforma slúži

predovšetkým pre dátových analytikov alebo procesných manažérov. Je postavená na technológiách pre veľké dáta, vrátane distribuovaného úložiska a technológií pre spracovanie veľkých dát. Okrem toho poskytuje viaceré nástroje pre podporu pokročilých analytických procesov. Jej súčasťou je množina vývojárskych nástrojov, ktoré umožňujú navrhovať a implementovať prediktívne metódy pre potreby optimalizácie procesov jednotlivých prevádzok. Vývojárske nástroje sú podporované množinou simulačných nástrojov, ktorých primárnou úlohou je vyhodnotenie implementovaných metód v testovacích prostrediach a ich nasadenie do reálneho prevádzkového prostredia. Pri implementácii riešenia sme použili viacero existujúcich technológií pre spracovanie veľkých dát. Ako jadro sme použili rámec Hadoop v distribúcii Hortonworks HDP, ktorý sme konfigurovali Apache Ambari (používaný aj na manažment). Používame Apache YARN [1] pre manažment zdrojov, (CPU, pamäť, disk, atď.), Hadoop Distributed File System (HDFS) ako distribuované a škálovateľné úložisko [2]. Pre spracovanie dát v reálnom čase používame Apache Spark [3], vývoj prediktívnych modelov bude podporovaný knižnicami strojového učenia ako napr. MLlib [4] alebo Mahout. Pre streamovanie dát používame Apache Kafka a Apache ZooKeeper slúži pre konfiguračný manažment.

4 Záver

Článok mal za úlohu predstaviť projekt MONSOON, jeho hlavné oblasti nasadenia, popísať konkrétne jednotlivé prípady nasadenia a architektúru riešenia. Architektúra riešenia obsahuje z funkčného pohľadu dva hlavné komponenty celej platformy - operačnú platformu zbierajúcu dáta z prevádzky reálnom čase, ktoré sú potom zozbierané, ukladané a analyzované v zdieľanej analytickej platforme. Tá navyše umožňuje jej využitie naprieč rôznymi sektormi, čo môže viesť k transferu znalostí alebo šíreniu skúseností medzi jednotlivými odvetviami. Prvotná verzia analytickej platformy bola implementovaná a nasadená použitím technológií pre spracovanie veľkých dát v doménach produkcie hliníku a výroby plastových komponentov.

Literatura

1. Kumar Vavilapalli, V., Murthy, A.C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., Saha, B., Curino, C., O 'malley, O., Radia, S., Reed, B., Baldeschwieler, E.: Apache Hadoop YARN: Yet Another Resource Negotiator. SOCC '13 Proc. 4th Annu. Symp. Cloud Comput. . 13, 1–3 (2013).
2. White, T.: Hadoop: The definitive guide. (2012).
3. Han, Z., Zhang, Y.: Spark: A Big Data Processing Platform Based on Memory Computing. In: Proceedings - International Symposium on Parallel Architectures, Algorithms and Programming, PAAP. pp. 172–176 (2016).
4. Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., Freeman, J., Tsai, D., Amde, M., Owen, S., Xin, D., Xin, R., Franklin, M.J., Zadeh, R., Zaharia, M., Talwalkar, A.: [seminal] MLlib: Machine Learning in Apache Spark. J. Mach. Learn. Res. 17, 1–7 (2016).

Poděkování: Tento článok bol podporený projektom H2020-SPIRE-2016 Project MONSOON "MOdel based control framework for Site-wide OptimizatioN of data-intensive processes"

Annotation:

Project MONSOON – design of the big data analysis platform for process industries

The main objective of the paper is to describe the main objectives of the MONSOON project. MONSOON (MOdel-based coNtrol framework for Site-wide OptimizatiON of data-intensive processes) is a 3-year integrated SPIRE project, that aims to establish data-driven methodology to support identification and exploitation of optimization potentials by applying model-based predictive controls so as to perform plant and site-wide optimization of production process. MONSOON is based on the concept of a cross-sectorial big data platform, a scalable analytical platform that will support collection, storage and processing of data from multiple industry domains. The analytical platform will contain development environment to build these functions and simulation environment to evaluate the models. The platform will be shared among multiple sites from different industry sectors. Cross-sectorial sharing will enable transfer of knowledge across different domains. The project consider two main processs industry domains, the aluminium production factory from France and the plastic molding industry from Portugal.

Visionářské příspěvky

Využití formálních gramatik v automatickém plánování – na cestě k sjednocujícímu modelu

Roman Barták

Univerzita Karlova, Matematicko-fyzikální fakulta
Malostranské náměstí 25, 118 00 Praha

bartak@ktiml.mff.cuni.cz

Abstrakt. V práci navrhujeme použít atributové gramatiky jako sjednocující model pro existující modely plánovacích domén. Takový sjednocující model přinese do plánování řadu nových možností využitím existujících technik vytvořených pro formální gramatiky. Můžeme například ověřit, zda plán odpovídá danému doménovému modelu. Můžeme dokonce verifikovat, zda je daný model vnitřně konzistentní, což je problém, který plánovací komunita dosud ani neřešila. Můžeme využít techniky učení se gramatiky z příkladů slov jazyka pro automatické učení doménových modelů, tedy jednu z důležitých schopností autonomních systémů, která je v případě plánování zatím na velmi nízké úrovni. Samozřejmě je možné pomocí gramatik plány přímo generovat, případně naopak lze z pozorované části plánu odvodit, jaké akce budou následovat nebo které byly přehlédnuty. Formální gramatiky tam mohou posloužit jako vhodné médium pro transport technik a znalostí mezi tak vzdálenými komunitami jako je zpracování jazyka a plánování.

Klíčová slova: plánování, modelování, formální gramatiky.

1 Úvod

Plánování neboli uvažování o budoucích akcích je jednou z klíčových součástí umělé inteligence. Nejintenzivnější výzkum probíhá především v oblasti klasického (STRIPS) plánování, které je založené na ploché struktuře akcí provázaných kauzálními vazbami (předchozí akce poskytují předpoklady pro běh dalších akcí). Přes pokrok v efektivitě tzv. doménově nezávislých plánovačů se tyto plánovací systémy jen zřídka používají v aplikacích např. v robotice nebo v počítačových hrách. Zde jsou mnohem populárnější přístupy umožňující kódovat doménově specifickou informaci, která dramaticky zvyšuje efektivitu plánování. Často používaný je model hierarchických sítí úloh (HTN), kde je plánování založené na dekompozici úloh. V článku navrhujeme pro popis plánovacích domén použít abstraktní model atributových gramatik, který vychází z HTN, ale zároveň umožňuje popsat další modely plánovacích domén jako jsou STRIPS či procedurální modely. Popíšeme výhody a možná použití tohoto

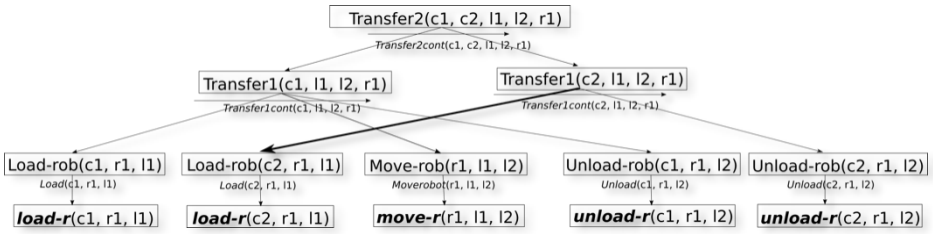
J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 99-104.

přístupu a nastíníme výzkumný program vedoucí k tomu, aby navržený přístup mohl sloužit jako sjednocující model pro reprezentaci plánovacích domén.

2 Použité pojmy a existující práce

Plánování se zabývá hledáním posloupnosti akcí vedoucí ke splnění zadaného cíle [8]. Existuje řada plánovacích modelů, od klasického plochého STRIPS modelu, přes hierarchické plánování (HTN) až například po popis plánů formou nedeterministického programu. Klasické STRIPS plánování popisuje akce jako operátory, které mění vlastnosti světa. Akce má své předpoklady, typicky množinu atomických tvrzení, které musí ve světě platit, aby šlo akci aplikovat, a efekty popsané atomy, které budou nebo naopak nebudou platné po provedení akce. Cílem je nalézt posloupnost akcí, která převede daný stav světa na stav, kde platí požadovaná množina atomických tvrzení. Hierarchické plánovaní [6] používá podobné primitivní akce s předpoklady a efekty, ale skupiny akcí jsou sdružovány do úloh tvořících hierarchickou strukturu. Plánování potom spočívá v rozkladu zadané úlohy na podúlohy a dále až na primitivní akce, které lze seřadit tak, že tvoří klasický plán. Hierarchická struktura tak poskytuje návod, jak najít akce pro dosažení zadaného cíle.

Výzkumníci si již dávno všimli podobnosti bezkontextových gramatik a hierarchického plánovacího modelu založeného na popisu plánu jako dekompozice úloh. Formální gramatiky se zatím využily při důkazech složitosti HTN plánovacích problémů [10] a při rozpoznávání plánů [7]. Překvapivě ale zatím nikdo neukázal převod plného HTN modelu na formální gramatiku. Základním problémem je zde prolínání akcí/terminálů vzniklých z různých úloh/neterminálů, které bezkontextové gramatiky nemohou podchytit, viz obrázek 1. K takovému prolnutí dochází, když akce vzniklá z jedné úlohy slouží pro poskytnutí předpokladů akce z jiné úlohy. V uvedeném příkladu byla akce move-r(r1,l1,l2) vytvořena pro úlohu převozu kontejneru c1, ale zároveň lze stejnou akci použít pro přepravu kontejneru c2, který ovšem musí být naložen před přejezdem a vyložen až po něm. Z toho vznikne prolnutí akcí pocházejících z různých úloh.



Obr. 1. Příklad dekompozice úlohy pro převoz dvou kontejnerů na dvě podúlohy převozu jednotlivých kontejnerů a dále na primitivní akce. Tučná čára ukazuje prolnutí dílčích plánů.

Formální gramatiky slouží pro popis jazyků formou přepisovacích pravidel [9]. Jazyk je množina slov skládajících se z písmen, kterým se v gramatikách říká terminální

symboly (terminály). Pro jejich generování se používají pomocné symboly – neterminály. V této práci vycházíme z bezkontextových gramatik, ve kterých mají všechna přepisovací pravidla tvar X → w, kde X je neterminál a w je slovo složené z terminálních i neterminálních symbolů. Generování slova začínající z daného neterminálu lze popsat formou derivačního stromu, který se velmi podobá stromu rozkladu úlohy na podúlohy až na ono prolínání akcí. Pro jeho správné zachycení budeme používat atributové gramatiky [11] přidávající k symbolům atributy, které mohou být provázány podmínkami. Tyto podmínky potom umožní uspořádání akcí, tak aby odpovídalo kauzálním vazbám.

3 Atributové gramatiky jako plánovací model

V práci [3] jsme ukázali, že HTN, klasický STRIPS model a procedurální model [1] lze převést na atributovou gramatiku se specifickou podmínkou *timeline*. Hierarchická struktura se zde používá podobně jako rozklad úlohy na podúlohy případně pro generování sekvence úloh, zatímco *timeline* podmínka zajišťuje platnost kauzálních vazeb mezi primitivními operátory. Stručně řečeno, při rozkladu úlohy shromažďujeme události vážící se na změny platnosti daného atomického tvrzení a *timeline* podmínka potom zajišťuje, že tyto události lze konzistentně uspořádat. Konkrétně, pokud akce požaduje platnost nějakého atomu, tak jiná akce v pořadí před ní musí atom nastavit na platný. Zde je ukázka takového dekompozičního pravidla (množina TL sdružuje události pro *timeline* podmínku a množina I pořadová čísla akcí):

$$\begin{aligned} &\mathit{Transfer1}_{c,l,l1,l2,r}(I,TL) \quad \rightarrow \quad \mathit{Load\text{-}rob}_{c,r,l1}(I_1,TL_1). \\ &\qquad\qquad\qquad\qquad\qquad\quad \mathit{Move\text{-}rob}_{r,l1,l2}(I_2,TL_2). \\ &\qquad\qquad\qquad\qquad\qquad\quad \mathit{Unload\text{-}rob}_{c,r,l2}(I_3,TL_3) \\ &[TL = TL_1 \cup TL_2 \cup TL_3,\ I = I_1 \cup I_2 \cup I_3,\ max(I_1) < min(I_2),\ max(I_2) < min(I_3)]. \end{aligned}$$

Pokud bychom věděli, že v plánu se již vyskytuje akce pro přesun robota mezi místy l1 a l2, která byla například vygenerována z jiné úlohy, nemuseli bychom úlohu $\mathit{Move\text{-}rob}_{r,l1,l2}$ do rozkladu vůbec zahrnovat. V takovém případě je ale potřeba přidat podmínku, že před vyložením kontejneru bude robot na příslušném místě:

$$\begin{aligned} &\mathit{Transfer1}_{c,l,l1,l2,r}(I,TL) \quad \rightarrow \quad \mathit{Load\text{-}rob}_{c,r,l1}(I_1,TL_1). \\ &\qquad\qquad\qquad\qquad\qquad\quad \mathit{Unload\text{-}rob}_{c,r,l2}(I_3,TL_3) \\ &[TL = TL_1 \cup TL_3 \cup \{at(r,l2)@\,min(I_3)\},\ I = I_1 \cup I_3,\ max(I_1) < min(I_3)]. \end{aligned}$$

Pro jednu úlohu tak může existovat více dekompozičních pravidel a při plánování se vyberou takové dekompozice, kterými lze získat korektní posloupnost akcí. O správné uspořádání akcí se stará podmínka *timeline*, která zajišťuje navázání efektů a předpokladů akcí. Předpoklady jsou popisovány pomocí tzv. *before* událostí, zatímco efekty pomocí *after* událostí. Podmínka zajišťuje, že pro každý atomický výrok je *before* událost předcházena příslušnou *after* událostí resp. jinou *before* událostí stejného

typu. Obrázek 2 ukazuje posloupnosti událostí pro jednotlivé atomické výroky z plánu zobrazeném obrázkem 1. Plus index indikuje požadavek na platnost daného výroku (v předpokladu) resp. jeho nastavení (v efektu), mínus index požadavek na neplatnost výroku resp. jeho smazání.

index akce	0	1 load-r(c1, r1, l1)	2 load-r(c2, r1, l1)	3 move-r(r1, l1, l2)	4 unload-r(c1, r1, l2)	5 unload-r(c2, r1, l2)
loaded(r1, c1)	a^-	b^-a^+			b^+a^-	
loaded(r1, c2)	a^-		b^-a^+			b^+a^-
at(r1, l1)	a^+	b^+	b^+	b^+a^-		
at(r1, l2)	a^-			a^+	b^+	b^+
in(c1, l1)	a^+	b^+a^-				
in(c1, l2)	a^-				a^+	
in(c2, l1)	a^+		b^+a^-			
in(c2, l2)	a^-					a^+

Obr. 2. Ukázka vývoje platnosti atomických výroků z plánu na obrázku 1 pomocí before a after událostí. Nultá vrstva popisuje počáteční stav.

4 Výzkumný program

Pokud můžeme převést různé plánovací modely na model jediný reprezentovaný atributovou gramatikou, lze navrhovat algoritmy pracující jen s atributovou gramatikou a řešit tak problémy pro původní modely, což je základní vize použití jednotného modelu. Navíc pro řešení těchto problémů lze používat techniky vyvinuté pro (atributové) gramatiky a tím přirozeně využít existující výsledky v jiné oblasti.

4.1 (Automatické) modelování

Pro modelování plánovacích problémů existuje minimum softwarových nástrojů a uživatelé jsou často odkázáni na textový editor a vytvoření modelu ručně v jazyce PDDL [12]. Při modelování problémů pro HTN model je situace ještě náročnější v nutnosti navrhnout vhodnou hierarchickou strukturu dekompozice úloh. Ideální by tedy bylo, pokud by systém dokázal navrhnout model sám ze sady ukázkových plánů.

V práci [4] jsme ukázali, jak lze rozpoznat akce z dat získaných ze sensorů drona. Jedná se tak o přínos k problému mostu mezi analogovým světem, typickým pro robotiku, a světem symbolickým, typickým pro umělou inteligenci. Otevřenou otázkou je, jak se naučit parametry takových akcí, například že přímý let může být na různou vzdálenost. To má souvislost s učením se předpokladů a efektů akce, tj. jaké vlastnosti musí mít svět, aby šlo akci realizovat, a jak se svět po provedení akce změní. Přirozeně je také potřeba se učit, jak danou akci následně provést. Pokud získáme primitivní akce, je dalším krokem jejich organizace do hierarchické struktury. Zde je možné provádět různé úlohy, pozorovat příslušné posloupnosti akcí, v nich hledat opaku-

jící se sekvence, které lze označit jako úlohy. Zde se vlastě jedná o učení se gramatiky na základě zadaných slov a šlo by tedy použít postupy z formálních gramatik.

4.2 Práce s modely

Máme-li k dispozici model v podobě atributové gramatiky, lze s ním řešit zajímavé problémy, které byly dosud v plánování složité. Prvním z nich je ověření, zda daný plán odpovídá modelu. Pro HTN model zatím existuje jediný výpočtově poměrně náročný přístup převodem na problém Booleovské splnitelnosti [5]. V případě realizace modelu v atributové gramatice je možné použít techniky rozpoznání, zda slovo patří do daného jazyka, například známý CYK algoritmus [13]. Ten je potřeba upravit pro atributové gramatiky, a hlavně vzít v úvahu prolnutí akcí v plánu (obrázek 1). Stejný algoritmus, který je v podstatě založený na sdružování akcí do úloh dle pravidel metodou zdola-nahoru (použití gramatiky analytickým způsobem), lze potom použít na rozpoznání, jakou úlohu agent řeší, analýzou pozorované (i neúplné) posloupnosti akcí. Ještě zajímavější a dosud nikým neřešený problém je otázka vnitřní konzistence doménového modelu. Tato otázka je důležitá zvláště pro automaticky získané modely. V práci [2] jsme navrhli základní přístup pro verifikaci atributové gramatiky s rekurzí, kde atributy mohou nabývat hodnot z konečných domén. Zde je použita technika odvozená od redukce bezkontextové gramatiky. Otevřenou otázkou je, zda lze takto verifikovat také atributové gramatiky s *timeline* podmínkou, které se používají pro plánovací modely.

5 Závěr

Článek představuje vizi jednotného modelu plánovacích problémů použitím atributových gramatik a jejich použití pro řešení existujících i nových problémů kolem plánovacích modelů.

Literatura

1. Baier, J.A., Fritz, Ch. and McIlraith S.A.: Exploiting Procedural Domain Control Knowledge in State-of-the-Art Planners. In: Proc. of the Seventeenth International Conference on Automated Planning and Scheduling (ICAPS 2007). M.S. Boddy, M. Fox, and S. Thiébaux (Eds.). AAAI (2007), 26-33.
2. Barták, R. Dvořák, T.: On verification of workflow and planning domain models using attribute grammars. In: Proc. of the 15th Mexican International Conference on Artificial Intelligence. Springer (2017).
3. Barták, R., Maillard, A.: Attribute Grammars with Set Attributes and Global Constraints as a Unifying Framework for Planning Domain Models. In: Proc. of the ICAPS Workshop on Knowledge Engineering for Planning and Scheduling (KEPS 2017), 45-53.
4. Barták, R., Vomlelová, M.: Using Machine Learning to Identify Activities of a Flying Drone from Sensor Readings, In: Proc. of Thirtieth International Florida Artificial Intelligence Research Society Conference (FLAIRS-30), AAAI (2017), 436-441.

5. Behnke, G., Höller, D., Biundo, S.: This Is a Solution! (... but Is It though?) Verifying Solutions of Hierarchical Planning Problems In: Proc. of the Twenty-Seventh International Conference on Automated Planning and Scheduling (ICAPS 2017). L. Barbulescu, J.D. Frank, Mausam, S.F. Smith (Eds.). AAAI (2017), 20-28.
6. Erol, K., Hendler, J., Nau, D.S.: Semantics for Hierarchical Task-network Planning. Technical Report. University of Maryland at College Park, USA (1994).
7. Geib, Ch.W., Steedman, M.: On Natural Language Processing and Plan Recognition. In Proc. of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007), M. Veloso (Ed.), IJCAI (2007). 1612-1617.
8. Ghallab, M., Nau, D.S., Traverso, P.: Automated planning - theory and practice. Elsevier, 2004.
9. Hopcroft, J.E., Ullman, J.D.: Introduction to Automata Theory, Languages and Computation, Addison-Wesley, 1979.
10. Höller, D., Behnke, G., Bercher, P., Biundo, S.: Language Classification of Hierarchical Planning Problems. In: Proc. of 21st European Conference on Artificial Intelligence (ECAI 2014). T. Schaub, G. Friedrich, and B. O'Sullivan (Eds.), Vol. 263. IOS Press (2014), 447-452.
11. Knuth, D.E.: Semantics of Context-Free Languages. Mathematical Systems Theory 2 (1968): 127-145.
12. McDermott, D.: The planning domain definition language manual. CVC Report 98-003, Yale Computer Science Report 1165, 1998.
13. Younger, D. H.: Recognition and parsing of context-free languages in time n^3. Inform. Control. 10 (2) (1967): 189-208.

Poděkování: Tento článek vznikl díky podpoře projektu GAČR P103-15-19877S.

Annotation:

Exploiting formal grammars in automated planning – towards the unifying model

The paper suggests using attribute grammars as a unifying framework for planning domain models. It is already possible to translate existing STRIPS, HTN, and PDCK models to attribute grammars with the timeline constraint. The paper discusses how the techniques developed for formal grammars can be used to solve problems with domain models, for example to validate if a plan conforms the model and if the model is internally consistent, and to guess which task is being performed by observing (even incomplete) sequence of actions. An interesting challenge is automated learning of the grammar from observed plans.

Kripke style Dynamic model for Web Annotation with Similarity and Reliability

M. Kopecký[1], M. Vomlelová[2], P. Vojtáš[1]

Faculty of Mathematics and Physics Charles University
Malostranske namesti 25, Prague, Czech Republic

[1]{kopecky|vojtas}@ksi.mff.cuni.cz
[2]marta@ktiml.mff.cuni.cz

Abstract. In this "visionary contribution" (term from conference organizers) we deal with the web semantization as a (semi) automated process of enriching web data in a way understandable for algorithms. It turns out that similarity and dynamic aspects of web data play a role here. We propose a web data extension by a Kripke style dynamic model to describe this process for future extractions.

Keywords: information extraction, semantic annotation, Kripke dynamic logic

1 Motivation

No human can read, understand, and synthesize the whole web information on an everyday basis. So we need automated web data processing. Our dream is to have a multicriterial web search supporting a customer (user) looking for a product (service, resource). Our interest is on web content which is not semantically annotated by owner. To distinguish – we see a difference between schema.org annotation (enabling search engines to improve their job) and our annotation which tries to make (and index upon request) part of the web relevant to user search more like an integrated database.

Closest to our approach is the work of G. Gottlob: Lixto [1](see e.g. [2]) – that is wrapper generation, web data extraction … just to mention a few. On the other side, to our knowledge, Dublin Core Metadata Initiative[2] methods, serve a different purpose – namely for mapping of ontologies to integrate sources annotated by an owner already.

Our motivation example (Fig.1, [4]) shows a web page of a travel agency consisting of several data records (hotels) inside a data region. Our first task is to recognize these areas (in the DOM tree) and make them available for further extraction of e.g. price, quality. The main idea is to use the fact that there is a certain repetition as the

[1] http://www.lixto.com/
[2] http://dublincore.org/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 105-109.

page was constructed using a template (using Levenshtein similarity and threshold depending on the domain, see [4] with experiments in notebooks, cars and hotels).

So we have the task, algorithm (with preconditions and post conditions), training sets and parameters ensuring best result depending on application domain and possibly a similarity measure. Now we can imagine two scenarios:

- to reuse the algorithm (on a similar page, e.g. created by same template or similar in more general sense) it would be good to have records on previous experiences,
- maybe extracted data are no more available (e.g. many LOD points do not function properly) and we have to extract data again or on demand.

In both cases it would be helpful to have semantic data extended with this information.

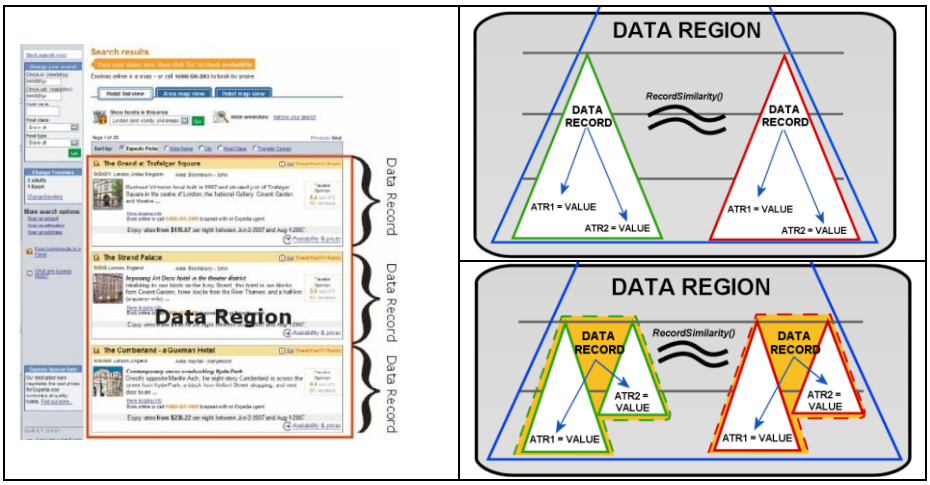


Fig. 1. Web resource, SW extracting data region and data records (in one cell, several cells), [4]

In [5] we presented several methods for mining web information and assisted annotations as we believe this should be the first steps towards the semantic web. Then several methods for processing the gathered data are described. The proposed methods mainly aim on modeling user and his/her preferences and then helping them with reaching their goals. We considered also the connection with a social network (see e.g. SoSIReCR[3]) and friends annotating an initial golden standard.

Our idea her is: we would like to have a formal model to remember origin of web extracted data and the means of this extraction (for possible future reuse and/or re-extraction and comparison of quality of alternative extraction tools).

[3] http://www.sosirecr.cz/index_en.php

2 A model for description of extraction/annotation circumstances

For effective using of changing and/or increasing information we have to evolve tools (e.g. inductive methods) used for creation of specific web service (e.g. recommendation). Our goal is to extend the semantic web foundations to enable describing creation, dynamics and similarities on data. To describe the reliability of extraction algorithms we propose a "half-a-way" extension of dynamic logic [3]. Programs (typically extractors) remain propositional whereas formulas will be more predicate-like (describing properties/attributes of web resources).

Kripke states correspond to different representation of content on the web and results of our extraction and annotation. So, we have two forms of states – input states and output states (with a possible overlap, to be able to describe chaining of extraction algorithms). Programs will usually act on various forms of web content representation, e.g. XML, (X)HTML, tables, DOM, texts, …). Today we can be challenged also by big data downloaded from the web and stored in a cloud. Output of our programs are data in various machine readable forms carrying semantic information, e.g. RDFa enrichment of (X)HTML, relations, FOL, RDF, texts (with PoS, morphology, dependency).

We aim to define a language working with data, hence we need to give our formulas a meaning. To specify formulas of our dynamic logic on each of states the respective semantics is defined using appropriate query language, e.g. XQuery, XPath, FOL, SPARQL, SQL, keyword search, … E.g. an empty SELECT or ASK can give me information on validity of certain statement hidden in FROM, WHERE conditions. New development ([1]) in area of multimodal databases gives extensions of SQL able to handle different forms of data.

Our "half-a-way" extension of dynamic logic has expressions of two sorts (and each sort is/can be typed): Statements about web data: atomic e.g. Φ_0^{RDF}, Φ_0^{FOL}, Φ_0^{RDB}, Φ_0^{XML}, Φ_0^{DOM}, Φ_0^{BoW}, Φ_0^{PoS}, $\Phi_0^{DepTree}$, … and Φ more complex φ^{RDF}, ψ^{FOL}, … with corresponding data model, query language based semantics – all can be subject of uncertainty, probability extensions.

Programs remain propositional: atomic, e.g. Π_0^{σ} for subject extraction, Π_0^{π} for property extraction, Π_0^{ω} for object value extraction in case of html, xhtml, xml data; Π_0^{ner} for named entity extraction in case of text data, and Π more complex $\alpha^{\sigma\pi\omega}$, $\beta^{\sigma\pi\omega}$, $\gamma^{\sigma\pi\omega}$, …

Statements are typically accompanied by information about program creation (data mining tool used for extraction, training data, metric (e.g. precision, recall) …) and there is a lot of reification describing the training and testing data and the metrics of learning. In place of ontologies we assume usage of user created dictionaries (usually very simple). Our model is based on dynamic logic, calculates similarity of Kripke states and describes uncertain/stochastic character of our knowledge.

Hence we are able to express our experience using extraction algorithms in statements like $\{\varphi\}\ \alpha\ \{\psi\}$ or $\varphi \rightarrow [\alpha]_x\ \psi$, where φ is a statement about data D_1 before extraction (preconditions), ψ is a statement about data/knowledge D_2 , K_2 after extrac-

tion (post conditions), α is the program used for extraction. Modality $[\alpha]_x$ can be weighted, describing stochastic aspects of learning. Lot of reification about learning can be helpful.

The main idea of our vision is that if there are some data D_1' similar to D_1 and φ is true in some degree (e.g., because both resources were created using the same template) then after using α we can conclude with high certainty/probability that the statement ψ will be true in some degree on data D_2' (knowledge K_2').
So one has first to train extractors but afterwards check how these extractors are resistant to data changes (with information on specific similarity measure).
We have already provided some experiments with extraction and similarity resistance, but this is out of scope of this "visionary" paper.

In our motivating example from Figure 1, ($\{\varphi\}\ \alpha\ \{\psi\}$ or $\varphi \rightarrow [\alpha]_x\ \psi$) can look like:

α_{Mar} is the software from [4], maybe available free on a URL with training data D_1

φ says: α_{Mar} was trained on data D_1^{ntb}, D_1^{car}, D_1^{hot} in notebooks, cars and hotels domain with parameters of learning (parameters of α_{Mar}, cross validation, metric, …) and precision, recall, … of extraction.

ψ can describe output, e.g. if D_2^{ntb}, D_2^{car}, D_2^{hot} and the similarity resistance i.e. if $D'_1{}^{ntb}$, $D'_1{}^{car}$, $D'_1{}^{hot}$ are similar to D_1^{ntb}, D_1^{car}, D_1^{hot} (in same/close domains) in similarity $\approx^{ntb}$, $\approx^{car}$, $\approx^{hot}$ in degree x^{ntb}, x^{car}, x^{hot} then after running α_{Mar} one can expect P/R …

To conclude, we have presented our vision on how to enable remembering origin of web extracted data and the means of their extraction for future reuse and/or re-extraction. We propose a formal dynamic model for automated web annotation with similarity and reliability (a Kripke style dynamic logic model).

We expect to face several challenges. First is that our doing tries to improve both extractor and resistance against similarity. These are two parameters and it is not clear what the optimum is. The Pareto optimal point – a very good extractor which is highly resistant to similarity changes is probably hard to construct for nontrivial data. Classical data mining optimizes this by cross validation and trying to cover the whole data space. This is difficult to achieve on the web, so we can expect dynamic strategies as web content evolves. To optimize similarity resistance is a new challenge.

Another challenge is as with all semantic web ideas: how to convince people to use them. One way is to concentrate on well-organized communities (medicine, pharmacology) or governmental data sources and publishing regulated by a law. Our vision should act also out of these well-organized communities, so we can expect either positive influence of social network friends or customer creation methods (e.g. Research Gate). We have acquaintance with two sides of the coin working with our social network of IT specialists in the regions of the Czech Republic (https://www.sitit.cz/).

As a future work we will concentrate on experiments in specific domains on bigger and variable data.

Acknowledgement: Research was supported by Czech project Progres Q48.

References

1. M. Aslett. Neither fish nor fowl: the rise of multi-model databases, February 8th, 2013, The 451 Group,
2. R. Baumgartner, G. Gottlob, M. Herzog: Scalable Web Data Extraction for Online Market Intelligence. PVLDB 2(2): 1512-1523 (2009)
3. D. Harel, D. Kozen, J. Tiuryn. Dynamic Logic (Foundations of Computing) The MIT Press, 2000
4. R. Novotny, P. Vojtas, D. Maruscak. Information extraction from web pages. In Proc. 2009 IEEE/WIC/ACM WI&IAT-workshops, 121-124, dl.acm.org id=1632266
5. L. Peska, I. Lasek, A. Eckhardt, J. Dedek, P. Vojtas, D. Fiser: Towards web semantization and user understanding. In EJC 2012, Y. Kiyoki et al Eds. Frontiers in Artificial Intelligence and Applications 251, IOS Press 2013, pp 63-81

Příspěvky o probíhajícím výzkumu

UWB at SemEval 2014 and 2016

Tomáš Brychcín[1], Tomáš Hercig[2], Lukáš Svoboda[2], and Michal Konkol[1]

[1]NTIS – New Technologies for the Information Society,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň
[2]Department of Computer Science and Engineering,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň

{brychcin, tigi, svobikl, konkol}@kiv.zcu.cz

Abstract. International Workshop on Semantic Evaluation (SemEval) is an ongoing series of evaluations of NLP (Natural Language Processing) algorithms, organized by Association for Computational Linguistics (ACL), the international scientific society which hold the major NLP conferences. The evaluations are intended to explore different aspects of meaning in a natural language. The results of NLP algorithms are compared with human judgments. The submitted systems from research teams across the world are compared in terms of performance. Our research team actively participates in the SemEval exercises. This paper summarizes our results in the area of semantic textual similarity and aspect-based sentiment analysis. In 2014 and 2016 our systems were among the best performing in both mentioned tasks.

Key words: SemEval, Semantic Textual Similarity, Aspect-based Sentiment Analysis, Distributional Semantics

1 Introduction

Natural language processing (NLP) is a progressive research field of computer science, artificial intelligence, and computational linguistics. Challenges in NLP involve human-computer interaction and the natural language understanding.

During the last years, NLP research has focused mainly on the semantic analysis, which investigates the ways, how to represent and how to automatically infer the meaning of a text. It has become the core NLP task and can be seen at prestigious conferences as the main topic. A better semantic models result in better performance of the particular NLP tasks (named entity recognition [4], language modeling [5], sentiment analysis [2, 3], document classification, summarization, stance detection, machine translation, and many others).

International Workshop on Semantic Evaluation (SemEval) is a shared task for evaluation of semantic models. It is organized by Association for Computational Linguistics (ACL), the society which holds the major NLP conferences. The results of semantic models are compared with human judgments. Our NLP research team actively participates in the SemEval tasks. This paper describes our results achieved at

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 111-115.

SemEval 2014 and 2016, concretely, in the tasks semantic textual similarity (Section 3) and aspect-based sentiment analysis (Section 4).

2 Distributional Semantics

The basic idea behind many modern semantic models is known as *Distributional Hypothesis*. It states that the meaning of a word is defined by the contexts where the word appears. This allows us to compare the meaning of words based on their contexts; words with similar contexts have similar meaning.

Distributional semantics models process huge amount of data in order to recognize contextual patterns. The meaning of words, phrases, or sentences is usually represented by vectors. Each word is associated with a vector, which captures all the information hidden in the contexts, including syntactic, semantic, pragmatic, or morphological information. The vectors form a k-dimensional vector space referred to as *semantic space*. The similarity of words can be measured based on their similarity (or distance) in the vector space. The most common method for measuring similarity is *cosine similarity*, which measures the cosine of the angle between the vectors of two words.

In recent years, many distributional semantics models were proposed, e.g. neural network based models (Continuous Bag-Of-Words and Skip-Gram) [6] and log-bilinear model called GloVe [7].

3 Semantic Textual Similarity

Semantic textual similarity (STS) is one of the core tasks at SemEval. Given the two textual fragments (word phrases, sentences, paragraphs, or full documents), the goal is to estimate the degree of their semantic similarity. STS systems are compared with the manually annotated data, consisting of sentence pairs and the corresponding score between 0 and 5 (higher score means higher semantic similarity). STS at SemEval 2016 were divided into English-English monolingual subtask (Section 3.1) and English-Spanish cross-lingual subtasks (Section 3.2). More information can be found in [1].

3.1 Monolingual Semantic Textual Similarity

We participated with two monolingual STS systems (the results are shown in Table 1):

- **UWB-sup**: Supervised system based on SVM regression with RBF kernel. We use state-of-the-art algorithms for the meaning representation as features. These methods benefit from various sources of information, such as lexical, syntactic, and semantic. Together, we have 301 STS features. The system is trained on all SemEval datasets from prior years (i.e. the data from 2012 up to 2015).

- **UWB-unsup**: Unsupervised system based on weighted word alignment. The method finds and aligns the words that have similar meaning and similar function in the pair of sentences.

3.2 Cross-lingual Semantic Textual Similarity

Our cross-lingual STS system for Spanish-English bilingual sentence pairs is based on two steps. Firstly, we translate Spanish sentences into English via *Google translator*. The English sentences are left untouched. Secondly, we use the same STS systems as for monolingual task. The results are shown in Table 2.

Table 1. Pearson correlations on SemEval 2016 monolingual STS evaluation data.

Team Run	Ans.-Ans.	HDL	Plagia-rism	Post editing	Ques.-Ques.	Mean	Run Rank	Team Rank
Samsung Poland	69.2	82.7	84.1	83.5	68.7	77.8	1	1
UWB-sup	62.1	81.8	82.3	82.0	70.1	75.7	2	2
UWB-unsup	64.4	79.3	82.7	81.2	53.3	72.6	21	2

Table 2. Pearson correlations on SemEval 2016 cross-lingual STS evaluation data.

Team Run	News	Multi Source	Mean	Run Rank	Team Rank
UWB-sup	90.6	81.8	86.3	1	1
UWB-unsup	91.2	80.8	86.0	2	1

4 Aspect-based Sentiment Analysis

The objective of aspect-based sentiment analysis (ABSA) is to identify the aspects of a given target entity and to estimate the sentiment polarity for each mentioned aspect. The definition of the ABSA task from SemEval 2014 distinguishes between two aspects of sentiment: aspect terms and aspect categories. The whole task is divided into four subtasks. The later SemEval ABSA tasks further distinguish between more detailed aspect categories and associate aspect terms (targets) with aspect categories.

4.1 SemEval 2014

For each subtask we propose both constrained (no external knowledge) and unconstrained approach. The constrained versions of our system are based purely on machine learning techniques. The unconstrained versions extend the constrained feature set by LDA, semantic spaces and sentiment dictionaries. The proposed approaches achieved very good results. More information can be found in [3]. Some of our results are shown in Table 3.

4.2 SemEval 2016

Our constrained submission was based on lexical and syntactic features and machine learning. The unconstrained submission additionally contained semantics features and dictionaries. We achieve state-of-the-art results in 9 experiments among the constrained systems and in 2 experiments among the unconstrained systems. We participated in four languages on both text and sentence levels. More information can be found in [2]. Some of our results are shown in Table 3.

Tab. 3. Achieved ranks and results on SemEval 2014 and 2016 ABSA task. F1 denote F_1 score in percentages.

				Constrained				**Unconstrained**			
Domain	**Year**	**Lang.**	**Level**	**Category**		**Sentiment**		**Category**		**Sentiment**	
				Rank	**F1**	**Rank**	**F1**	**Rank**	**F1**	**Rank**	**F1**
Restaurants	2016	EN	Sentence	3.	68	2.	82	8.	68	9.	82
Laptops	2016	EN	Sentence	1.	48	3.	74	7.	47	10.	74
Restaurants	2016	EN	Text	1.	81	1.	81	3.	80	1.	82
Laptops	2016	EN	Text	1.	61	1.	75	2.	60	1-2.	75
Restaurants	2014	EN	Sentence	12.	76	12.	72	7.	79	4.	78
Domain	**Year**	**Lang.**	**Level**	**Target**		**Sentiment**		**Target**		**Sentiment**	
Restaurants	2016	EN	Sentence	1.	67	4.	41	3.	67	6.	41
Restaurants	2014	EN	Sentence	1.	67	4.	41	3.	67	6.	41
Laptops	2014	EN	Sentence	5.	81	9.	73	-	-	4.	67

5 Conclusion

In this paper we presented our UWB systems for STS task at SemEval competition. We use distributional semantics models as a core part of our methods. We were ranked #2 out of 113 systems in monolingual STS and #1 out of 26 systems in cross-lingual STS. Our system for ABSA was ranked as one of the bests on both SemEval 2014 and 2016.

References

1. Brychcín, T. and Svoboda, L.: UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016.
2. Hercig, T., Brychcín, T., Svoboda, L., and Konkol, M.: UWB at SemEval-2016 Task 5: Aspect Based Sentiment Analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pages 342-349.
3. Brychcín, T., Konkol, M., and Steinberger J.: UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, 2014, pages 817-822.

4. Konkol, M., Brychcín, T., and Konopík, M.: Latent semantics in named entity recognition. *Expert Systems with Applications*, 2015, pages 3470-3479.
5. Brychcín, T., and Konopík, M.: Semantic spaces for improving language modeling. *Computer Speech & Language*, 2014, pages 192-209.
6. Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient estimation of word representations in vector space, 2013.
7. Pennington, J., Socher, R., and Manning, C.: Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pages 1532-1543.

Univerzální řešení domén v relační databázi

Martin Zíma, Michal Nykl, Martin Dostal

Katedra informatiky a výpočetní techniky,
Západočeská univerzita v Plzni,
Univerzitní 8, 306 14 Plzeň, Česká republika

{zima, nyklm, madostal}@kiv.zcu.cz

Abstrakt. Tento příspěvek je zaměřen na relační databáze a představuje náš inovativní přístup k problematice využívání atributů s výčtovým datovým typem (dále označovány jako doménové atributy). Velkým nedostatkem existujících řešení daného problému je např. nezbytný zásah do navrženého schématu databáze při každém požadavku na přidání nového doménového atributu. Námi navržený způsob správy doménových atributů využívá některé dobré vlastnosti dřívějších řešení, není závislý na zvoleném relačním systému řízení báze dat a nevyžaduje zásah do schématu databáze kromě přidání nového doménového atributu. Navržený způsob správy doménových atributů v relační databázi bude využit v námi řešeném projektu *Zpřístupnění dotazů jazykové poradny v lingvisticky strukturované databázi*.

Klíčová slova: databázové systémy, integritní omezení, doménové atributy.

1 Úvod

Nezbytnou součástí komplexních schémat relačních databází bývají také atributy tabulek, které mohou nabývat pouze hodnoty z omezeného výčtu hodnot. Přesto, že každý atribut je definován svým jménem a doménou zastupující datový typ [2], budeme tyto atributy nazývat *doménové atributy*. Pokud je výčet hodnot doménového atributu neměnný, tak je jeho definice jednoduchá. Prostřednictvím klausule `CHECK` jazyka SQL vytvoříme pro daný doménový atribut integritní doménové omezení, přičemž hodnoty dané domény definujeme operátorem `IN` [3]. Některé databázové systémy nabízí pro řešení tohoto problému speciální datové typy, např. MySQL zavádí datový typ `ENUM` [4]. Problém ale nastává, pokud chceme doménový atribut přidat do tabulky nebo pokud chceme změnit výčet jeho hodnot. Zde žádné existující řešení není dostatečně univerzální, a proto v tomto článku představujeme náš inovativní přístup ke správě doménových atributů, který umožňuje přidávat do tabulek doménové atributy, aniž by vyžadoval změnu schématu databáze.

V sekci 2 shrneme existující způsoby správy doménových atributů a v sekci 3 představíme vlastní způsob řešení tohoto problému. Na závěr shrneme kladné vlastnosti navrženého způsobu správy doménových atributů a představíme projekt, ve kterém bude tento způsob použit.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 116-119.

2 Existující způsoby správy doménových atributů

Jedním z nejznámějších způsobů správy doménových atributů, který nabízí možnost měnit, přidávat či odebírat položky domény, je tvorba tzv. *číselníků*, či též doménových tabulek. Každý číselník obsahuje minimálně dva atributy: celočíselný primární klíč a (obvykle) textovou hodnotu dané domény. Číselník je na tabulku využívající doménový atribut vázán relací 1:N, přičemž doménový atribut je v tabulce nahrazen cizím klíčem odkazujícím do číselníku. Uvedený přístup má několik výhod: změna hodnoty v číselníku se okamžitě promítá do stávajících dat a stejnou doménu lze použít v jiném doménovém atributu libovolné tabulky.

Nevýhodou uvedeného přístupu je nutnost založení nového číselníku vždy, když přijde požadavek na přidání nové domény, tj. je zde nutný zásah do schématu relační databáze. Nicméně snahou databázových architektů je navrhnout schéma tak, aby výše uvedené změny nevyžadovaly žádný zásah do schématu a vše bylo řešeno jen uloženými daty v databázi.

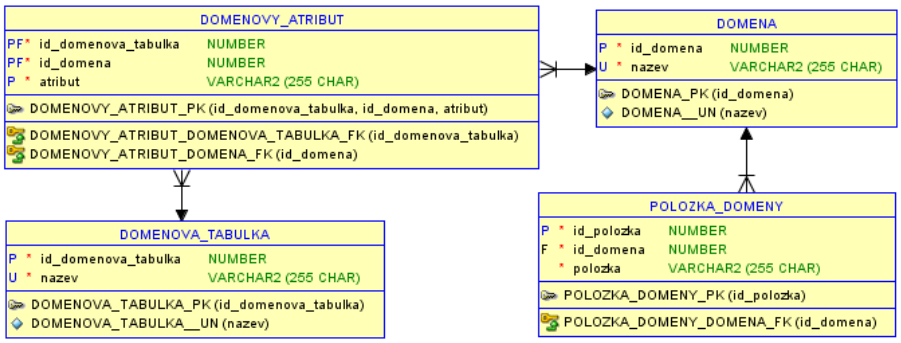
Vlastní způsob správy doménových atributů nabízí také společnost Oracle, která do produktu *Oracle Designer* zabudovala speciální tabulku `CG_REF_CODES` [1]. Ta je navržena tak, aby zahrnula všechny zakládané domény (výčty hodnot) v databázi, ale není propojena žádnou relací s ostatními tabulkami. Doménový atribut musí být textového datového typu. Při definici doménového atributu tabulky je nezbytné vygenerovat tzv. *table API* pro danou tabulku. *Table API* se skládá z uloženého balíku procedur a funkcí a sady triggerů, které při nastalé události volají metody z tohoto balíku. Jedna z metod hlídá, že daný doménový atribut může nabývat pouze zvolených hodnot z tabulky `CG_REF_CODES`.

3 Navržený univerzální způsob správy doménových atributů

Při řešení větších projektů je stěžejní správný a univerzální návrh schématu relační databáze. Nicméně, i po dokončení návrhu schématu mohou vznikat požadavky na nové doménové atributy, což vede k rozšiřování databáze o nové číselníky. Naším cílem je zabránit rozšiřování schématu relační databáze a přitom uspokojit nové požadavky na doménové atributy. Za tímto účelem jsme navrhli nový způsob práce s doménami, který využívá dobré vlastnosti výše popisovaných přístupů a není závislý na zvoleném databázovém systému.

3.1 Schéma navrženého způsobu správy doménových atributů

Námi navržený univerzální způsob správy doménových atributů využívá vlastní schéma databáze obsahující čtyři tabulky, jak ukazuje obrázek 1. Schéma je vytvořeno nástrojem *Oracle SQL Developer Data Modeler*. Atributy mohou být označeny písmeny `P`, `F` či `U`, pokud je daný atribut prvkem množiny definující integritní omezení. Těmi mohou být primární klíč (`P`, *primary key*), cizí klíč (`F`, *foreign key*) a unikátní hodnota (`U`, *unique*). Hvězdička u atributu označuje integritní omezení `NOT NULL`.



Obr. 1. Navržené schéma databáze pro podporu doménových atributů

Jak to celé funguje, si ukážeme na příkladu. Při návrhu schématu relační databáze si představme např. tabulku `NAHRAVKA`, do které chceme vložit doménový atribut `FORMAT`, který může nabývat pouze hodnot *mp3, wma* a *wav*. Dále pro jednoduchost výkladu předpokládejme, že tabulky z obrázku 1 jsou prázdné. Nyní budeme postupovat podle těchto bodů:

1. Do tabulky `DOMENOVA_TABULKA` vložíme nový záznam o tabulce, tj. `NAHRAVKA`.
2. Do tabulky `DOMENA` vložíme nový záznam o doméně, tj. `FORMAT_NAHRAVKY`.
3. Do tabulky `POLOZKA_DOMENY` vložíme odpovídající tři hodnoty dané domény, tj. *mp3, wma* a *wav*.
4. Do tabulky `DOMENOVY_ATRIBUT` vložíme nový záznam o tom, ve které tabulce byla použita která doména a jak se daný doménový atribut v tabulce jmenuje, tj. tabulka `NAHRAVKA`, doména `FORMAT_NAHRAVKY` a atribut `FORMAT`.

Nyní víme, že atribut `FORMAT` tabulky `NAHRAVKA` je doménovým atributem, ale aktuálně není nikterak zajištěno, že může nabývat pouze hodnot dané domény.

3.2 Validace hodnot doménových atributů

Aby nově přidaný doménový atribut mohl v tabulce nabývat pouze hodnot ze své domény, je nezbytné vytvořit uloženou funkci pro kontrolu vstupu a řádkový trigger pro danou tabulku. Funkce bude mít 4 parametry: název tabulky a jejího doménového atributu, název domény tohoto atributu a do tabulky vkládaná či aktualizovaná hodnota atributu. Trigger při vkládání či aktualizaci hodnot v tabulce dovolí zapsat do doménového atributu pouze hodnotu, kterou obsahuje příslušná doména. K tomu účelu bude volat připravenou funkci, která svou návratovou hodnotou stanoví, zda vkládaná či aktualizovaná hodnota doménového atributu je validní, tj. jedná se o hodnotu správné domény.

4 Závěr

Navržený univerzální způsob správy doménových atributů přebírá dobré vlastnosti existujících řešení a eliminuje vlastnosti nepříznivé. Použití navrženého způsobu je jednoduché, protože je potřeba pouze rozšířit stávající schéma relační databáze o čtyři tabulky a v databázi zkompilovat uloženou funkci. Pro každý doménový atribut je nutné tyto tabulky odpovídajícím způsobem naplnit a současně každému atributu vytvořit v databázi řádkový trigger pro jeho kontrolu. Výhodou námi navržené správy doménových atributů je také skutečnost, že do doménového atributu je ukládána příslušná hodnota místo hodnoty cizího klíče odkazujícího na záznam v číselníku.

Navržený způsob správy domén v relační databázi bude využit v projektu *Zpřístupnění dotazů jazykové poradny v lingvisticky strukturované databázi*, jehož cílem je vytvořit podpůrný nástroj pro pracovníky Jazykové poradny Ústavu pro jazyk český AV ČR. Nástroj by měl usnadnit pracovníkům poradny jejich poradenskou činnost a současně by měl osobám z široké veřejnosti umožnit nalezení správných informací o aktuálních jazykových jevech, které stávající jazykové příručky dosud neobsahují. Tento projekt bude dokončen v roce 2019.

Literatura

1. COXALL, Malcolm a CASWELL, Guy. *Oracle Quick Guides Part 2 - Oracle Database Design.* Cornelio Books, Spain, 2014. ISBN 978-84-940853-5-2
2. CONNOLLY, Thomas a BEGG, Carolyn. *Database Systems: A Practical Approach to Design, Implementation, and Management.* Pearson Addison Wesley, 2004. ISBN 978-0-321-21025-8
3. STEPHENS, Ryan a PLEW, Ron: *Sams teach yourself SQL in 21 days.* 4th ed. Indianapolis, SAMS, 2003. ISBN 978-0672324512
4. The ENUM Type. *MySQL 5.7 Reference Manual* [online]. [cit. 2017-06-07]. Dostupné z: https://dev.mysql.com/doc/refman/5.7/en/enum.html

Poděkování: Tento příspěvek byl podpořen grantem Ministerstva Kultury České republiky číslo DG16P02B009 *Zpřístupnění dotazů jazykové poradny v lingvisticky strukturované databázi*.

Annotation:

Universal solution of domains in relational database

This article describes the problem of domain attributes management in relational databases and it describes our innovative approach to this issue. The core problem of other solutions is in a need to extend relational database schema when a new domain attribute has to be added. Our solution is RDMS independent and it not requires modifications in relational database schema. The proposed solution will be used in our project: *Access to a Linguistically Structured Database of Enquiries from the Language Consulting Centre*.

Data integration for customer preference learning

M. Kopecký[1], M. Vomlelová[2], P. Vojtáš[1]

Faculty of Mathematics and Physics
Charles University
Malostranske namesti 25,
Prague, Czech Republic

[1]{kopecky|vojtas}@ksi.mff.cuni.cz
[2]marta@ktiml.mff.cuni.cz

Abstract. We describe the process and challenges of integration of movie data from Movie Lens, Netflix and RecSys Challenge 2014 with IMDB and DBPedia. Thanks of this integration we can enhance information by semantic data and improve prediction of customer preferences and recommendation. These data were collected in different situation by different methodologies. We want to use these data to be able to extend and further enhance our machine learning approaches developed for individual datasets to other datasets.

Keywords: Applications using data extracted from web, computer annotation, data, experiments and metrics

1 Introduction, motivation, recent work.

No human can comprehend any large collection of multi-dimensional data in his/her mind and choose the optimal item according to complex and often difficult to formulate criterion. For this purpose can be helpful recommender systems, that can learn user's preferences from his/her both explicit and implicit actions. The goal of the recommender system is then suggest suitable and often surprising proposals. Different collections of the otherwise similar data can often require different approaches simply due to different semantic data available about items and users in datasets. Because these approaches cannot be directly executed on all datasets, they can be compared only with complications. In this ongoing research report we thus concentrate on synergy effect of annotation and integration of data for user preference learning, and consequently for recommendation. The optimal are such domains where individual items can be identified and where additional data are publicly available. As a basic domain we choose the domain of movies.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 120-124.

2 Extracting and integrating data from movie domain

In this chapter we first describe data creation, interchanging annotation and data integration. We use *Flix* data i.e. enriched *Netflix* competition data, *RecSys* 2014 challenge data [3] and *RuleML* Challenge data [1].

We started with three available independent datasets: *MovieLens 20M* dataset, *Twitter dataset* and *Flix dataset*

Sizes of all datasets are summarized in Table 1.

Table 1– original datasets

Dataset	Ratings	Rated /all movies	Rating users
MovieLens 20M	20 000 263	26 744 27 278	137 493
Twitter dataset	168 880	13 616 14 542	22 073
Flix dataset	90 217 939	12 031 17 770	479 870

The datasets are quite different. Still they have few things in common. Movies have their title and usually also the year of their production. Ratings are equipped by timestamp that allows us to order ratings from individual users chronologically.

To be able to map movies from different datasets, we wanted to enhance every movie record by the corresponding IMDb[1] identifier **TT** with format 'ttNNNNNNN'.

We observed that the *Twitter* dataset uses as their internal **MOVIEID** the numeric part of the IMDb identifier. So the movie "Midnight Cowboy" with **MOVIEID**=64665 corresponds to the IMDb record with ID equal to 'tt0064665'.

To be able to assign IMDb identifiers to movies from other datasets, we had to use the search capabilities of the IMDb database. For both of them we used an HTTP interface for searching movies according to their name. The HTTP response then – among others – contains a table in form:

```
<table><tr>
    <td><a href="/title/ttNNNNNNN/?ref_=fn_ft_tt_1" ><img
src="..."></a></td>
    <td><a href="/title/ttNNNNNNN
/?ref_=fn_ft_tt_1">Title of the movie</a> (YEAR) ...</td>
</tr></table>
```

To be able to maintain both *MovieLens* and *Flix* dataset equally – regardless different formats of movie titles in them – and potentially in other future datasets, we needed to transform each movie title to the proper form expected by the IMDb interface. The basic algorithm can be described in steps:

[1] http://www.imdb.com/

- Convert all letters in movie title to lower case.
- If the movie title contains year of production at its end in brackets remove it.
- If the movie title still contains text in brackets at its end, remove it. This text usually contained original name of movie in original language.
- Move word ”the”, respectively ”a”/”an” from the end of the title to the beginning.
- Translate characters ”_”, ”.”, ”?” and ”,” to spaces
- Translate ”&” and ”&” in titles to word ”and”

For example, the transformation changes title ”Official Story, The (La Historia Oficial) (1985)” from the *MovieLens* dataset to its canonical form ”the official story” which can be identified as movie with the ID=’tt0089276’. Similarly the title ”Seventh Seal, The (Sjunde inseglet, Det) (1957)” from the same dataset is transformed to the form ”the seventh seal” with ID=’tt0050976’.

The successfulness of this approach to map movies from both *MovieLens* and *Flix* datasets is in first line of Table 2.

In optimal case, the table returning from the IMDb search contains exactly one row with the requested record. For this situation the algorithm behaves well and is able to retrieve the correct IMDb identifier. In many other cases the result contained more rows and the correct one or the best possible one had to be identified. For this purpose we enhanced the algorithm by additional steps:

- The correct record should be from the requested year, so the returned table should be searched only for records from this year and other records should be ignored
- The IMDb search provides more levels of tolerance in title matching. Try to use them from the most exact one to the most general. If the matching record from requested year cannot be found using stricter search, the other search level is used.

Currently, we have 13 081 out of all 17 770 *Flix* movies mapped onto the IMDb database. Even all 27 278 out of 27 278 movies from the *MovieLens* set are mapped to the equivalent IMDb records. So the current results provided by the combination of most advanced versions of algorithms are promising.

The diagram in the Figure 1 shows the amount of movies associated to the IMDb record in different intersections after the integration. For each movie registered in the IMDb database we then retrieved XML data from the URL address `http://www.omdbapi.com/?i=ttNNNNNNN&plot=full&r=xml`
and then from the XML data we retrieved following movie attributes. Among others *title*, *rating*, *avards*, *year*, *country*, *language*, *genres*, *director* and *actors*.

Another source of semantic data we use is the *DbPedia*. For this purpose we implemented the mapping technique described in [K] and assigned DbPedia[2] identifiers and associated semantic data to *IMDb* movies.

[2] http://wiki.dbpedia.org/

The *DbPedia* identifier of movie is a string, for example ”The_Official_Story” or ”The_Seventh_Seal”. This identifier can then be used to access directly the *DbPedia* graph database or retrieve data in an XML format through the URL address in form `http://dbpedia.org/page/DbPediaIdentifier`.

Table 2 –IMDb search by title name – the successfulness of IMDb title search for original – seven steps – algorithm and the final – enhanced – version.

	MovieLens	**Flix**	**Twitter**
IMDb search by title name	45,4%	70,9%	Not needed
Final enhanced version	100.0%	73.6%	Not needed

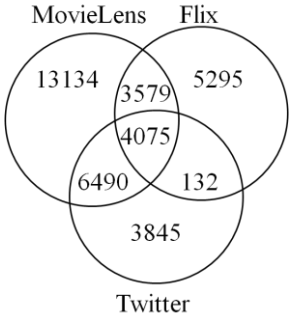


Figure 1 – Integration of movies in datasets based on the IMDb mapping

3 Conclusions, future work

We illustrated our approach to integration of five datasets – three movie datasets and two movie databases containing semantics data.

The future challenge is twofold:

- provide deeper analysis of data mining and use interconnection of datasets and their semantic enhancements for identifying and using possible dataset similarities.
- In future research we would like to continue in approaches in [2].
- extend this approach to other domains

Acknowledgement: Research was supported by Czech project Progres Q48.

References

1. Kuchar J.: Augmenting a Feature Set of Movies Using Linked Open Data, Proceedings of the RuleML 2015 Challenge, Berlin, Germany. Published by CEUR Workshop Proceedings, 2015
2. L. Peska, I. Lasek, A. Eckhardt, J. Dedek, P. Vojtas, D. Fiser: Towards web semantization and user understanding. In EJC 2012, Y. Kiyoki et al Eds. Frontiers in Artificial Intelligence and Applications 251, IOS Press 2013, pp 63-81
3. Twitter data from RecSys 2014 challenge http://2014.recsyschallenge.com/

Interaktívna vizualizácia hierarchických štruktúr

Miroslav Smatana, Peter Butka

Katedra kybernetiky a umelej inteligencie,
Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

miroslav.smatana@tuke.sk, peter.butka@tuke.sk

Abstrakt. V súčasnosti existuje množstvo dát v digitálnej podobe. Tieto dáta obsahujú informácie, ktoré môžu byť potencionálne užitočné napr. pre spoločnosti pri podpore rozhodovania. Na extrakciu informácií z dát je možné použiť širokú škálu metód analýzy dát. Problémom pri analýze dát je interpretácia jej výsledkov koncovým používateľom, tak aby získané informácie dokázali pochopiť a následne vhodným spôsobom využiť. Jednou z možností ako čo najrýchlejšie porozumieť výsledkom je ich transformácia do grafickej podoby. Preto sa v tomto článku zameriame na prezentáciu rozličných vizualizačných techník určených na vizualizáciu výsledkov získaných z hierarchických metód analýzy dát akou je napr. formálna konceptová analýza.

Kľúčové slová: dynamická vizualizácia, hierarchické štruktúry, formálna konceptová analýza

1 Úvod

V súčasnosti s príchodom sociálnych sieti, e-shopov, on-line novín, časopisov a pod., keď je v digitálnej podobe dostupné enormné množstvo dát, nastáva potreba ich spracovania a analýzy. Tieto dáta môžu obsahovať dôležité informácie, ktoré môžu využiť napríklad firmy na upevnenie svojho postavenia na trhu a získania konkurenčnej výhody a pod.

Doposiaľ bolo vyvinuté nemalé množstvo metód analýzy takýchto dát jednou z nich je aj formálna konceptová analýza (FCA) [1]. Cieľom analýzy dát pomocou FCA je vytváranie konceptového zväzu, ktorý predstavuje hierarchicky organizovanú štruktúru objektov (konceptov) na základe nimi zdieľaných atribútov. Metódy FCA [2-4] našli uplatnenie v oblastiach ako vyhľadávanie informácií, manažment znalostí a pod.

Aj keď FCA pomáha používateľovi lepšie pochopiť vstupné dáta a vzťahy medzi nimi, tak jedným z problémom s ktorým sa táto metóda stretáva je veľké množstvo generovaných konceptov v konceptovom zväze. Preto nastáva otázka ako tieto výsledky vhodne interpretovať používateľovi, aby ich dokázal čo najrýchlejšie pochopiť a využiť. To sa aj stalo jednou z hlavných tém v spojení s FCA. Jednou z možností je použiť redukčné metódy (výber len najdôležitejších konceptov) [5-6], vhodná vizuali-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 125-130.

zácia alebo dynamická vizualizácia (kde sa sústredíme len na podčasť celého konceptováho zväzu). V tejto práci sa zameriame práve na dynamické metódy vizualizácie konceptových zväzov, avšak predstavené metódy je možné využiť aj pre vizualizáciu iných hierarchických štruktúr.

2 Vizualizácia hierarchických dát

Vizualizácia dát predstavuje grafickú reprezentáciu špecifických informácií, ktorej cieľom je ponúknuť ich efektívnu reprezentáciu pre ich čo najrýchlejšie porozumenie a prezentáciu. Predstavuje kľúčový element v rôznych oblastiach.

V súčasnosti existuje množstvo vizualizačných techník, z toho dôvodu Lenger a Eppler [7] predstavili rozsiahlu tabuľku vizualizačných techník. Táto tabuľka obsahuje 100 vizualizačných techník rozdelených podľa typu ich použitia.

Taktiež bolo predstavených niekoľko metód v spojení s vizualizáciou hierarchických dát. Wills [8] vo svojej práci prezentoval niekoľko "node-edge" a "space-filling" vizualizačných techník. V práci [9] bol prezentovaný prístup založený na použití "tree-ring metaphor". Ďalšie z techník sú popísané v prácach [10-12].

Všetky predchádzajúce spomínané prístupy sú určené pre vizualizáciu hierarchií, kde každý uzol v grafe (prvok v hierarchii) má len jedného rodiča. Avšak v kontexte konceptových zväzov môže mať každý prvok hierarchie viacerých rodičov. Medzi najčastejšie používané metódy pre takýto typ vizualizácie sa používa Hasseho diagram [13] alebo vizualizácia popísaná v práci [14].

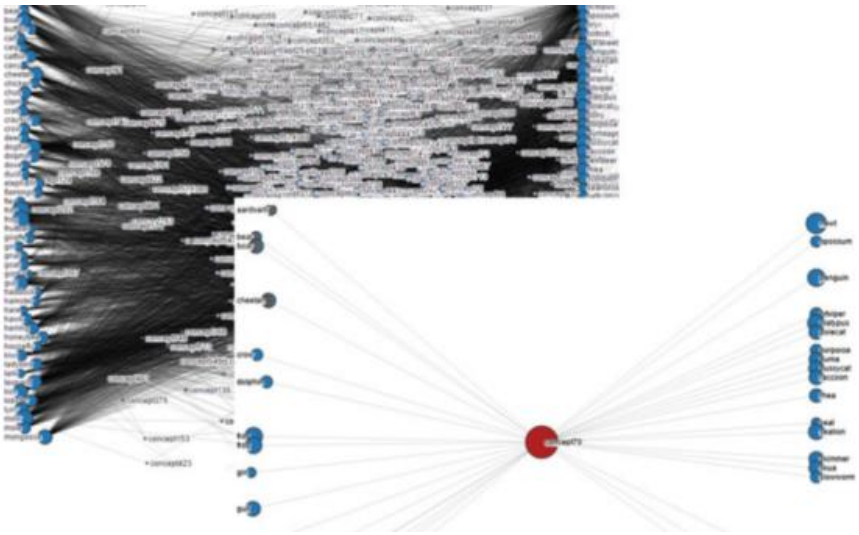
3 Vizualizácia konceptových zväzov

Vizualizácia je dôležitá časť analýzy dát pomocou konceptových zväzov. Avšak štandardné statické vizualizačné techniky sú zvyčajne vhodné len na vizualizáciu konceptových zväzov obsahujúcich len malé množstvo konceptov a prepojení medzi nimi. Pri rozsiahlych konceptových zväzoch sú tieto techniky neefektívne a neprehľadné. Preto v tomto článku prezentujeme niekoľko vizualizačných techník, ktoré by mali používateľovi pomôcť a sprehľadniť vizualizáciu veľkých konceptových zväzov.

3.1 Dynamický Hasseho diagram

Ako už bolo spomenuté Hasseho diagram predstavuje prirodzenú techniku pre zobrazovanie hierarchie, kde môže mať uzol viacero rodičov. Avšak problém je pri zobrazovaní veľkých štruktúr, preto sme sa rozhodli implementovať dynamickú verziu tejto vizualizačnej techniky, kde si používateľ môže priblížiť špecifický uzol a zobraziť jeho detail ako aj jeho prepojenia na iné uzly (viď. **Obr. 1**).

Obr. 1. Dynamický Hasseho diagram pre zobrazenie konceptového zväzu



Obr. 2. Dynamický prehľad konceptov.

3.2 Dynamický prehľad konceptov

Predstavuje odlišnú techniku ako Hasseho diagram. Dynamický prehľad konceptov je vhodný na vizualizáciu a prehľadávanie veľkých konceptových zväzov (vzhľadom

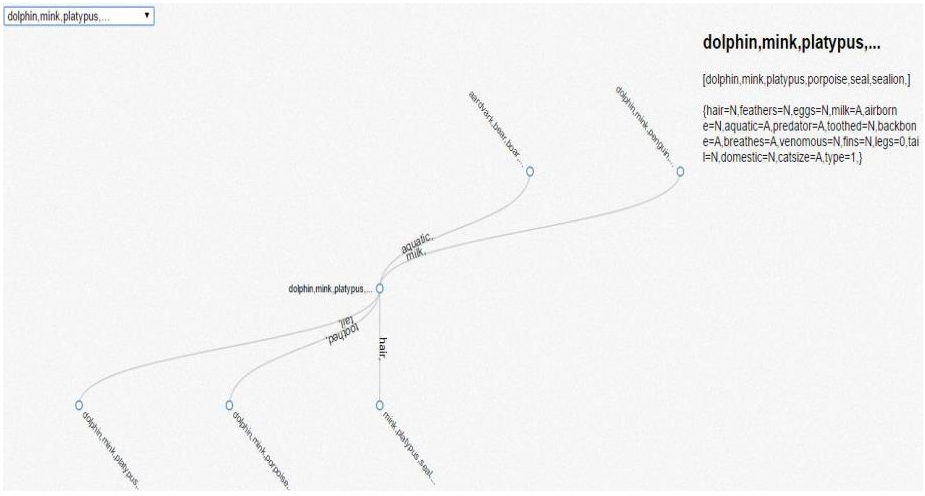
k objektom a ich výskytu v konceptoch). Príklad takéhoto zobrazenia je znázornený na **Obr. 2**, kde uzly vpravo a vľavo predstavujú objekty a uzly v strede predstavujú koncepty a hrany medzi nimi predstavujú výskyt objektu v koncepte. Používateľ si môže vybrať na zobrazenie len niektoré koncepty a im prislúchajúce objekty, čo robí túto techniku veľmi transparentnú.

3.3 Dynamický prehľad konceptov

Predstavuje odlišnú techniku ako Hasseho diagram. Dynamický prehľad konceptov je vhodný na vizualizáciu a prehľadávanie veľkých konceptových zväzov (vzhľadom k objektom a ich výskytu v konceptoch). Príklad takéhoto zobrazenia je znázornený na **Obr. 2**, kde uzly vpravo a vľavo predstavujú objekty a uzly v strede predstavujú koncepty a hrany medzi nimi predstavujú výskyt objektu v koncepte. Používateľ si môže vybrať na zobrazenie len niektoré koncepty a im prislúchajúce objekty, čo robí túto techniku veľmi transparentnú.

3.4 Double tree

Double tree technika predstavuje podobnú vizualizačnú techniku ako Hasse diagram, avšak v našom prípade sme použili jej modifikáciu. Tá spočívala v zobrazovaní len podčasti konceptového zväzu, kde bol zobrazovaný len jeden hlavný koncept a jemu prislúchajúci rodičia a potomkovia, pričom používateľ bol následne schopný prehľadávať daný konceptový zväz kliknutím na jeden zo zobrazovaných uzlov (konceptov), ktorý sa následne nastavil ako hlavný uzol. Príklad Double tree vizualizácie je zobrazený na **Obr. 3**.



Obr. 3. Double tree vizualizácia

4 Záver

V práci sme predstavili niekoľko vizualizačných techník a ich využitie pre zobrazovanie konceptových zväzov. V rámci našich skúsenosti sa ukázalo, že dynamická vizualizácia len lokálnej časti konceptového zväzu je veľmi vhodná pre prehľadávanie, navigáciu a pochopenie konceptového zväzu.

Literatúra

1. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer Verlag, Berlin (1999)
2. Krajci, S.: A generalized concept lattice. Logic Journal of IGPL 13(5), 543–550 (2005)
3. Medina, J., Ojeda-Aciego, M., Ruiz-Calvino, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. 160, 130–144 (2009)
4. Butka, P., Pocs, J.: Generalization of one-sided concept lattices. Comput. Inf. 32(2), 355–370 (2013)
5. Butka, P., Pocs, J., Pocsová, J.: Reduction of concepts from generalized one-sided concept lattice based on subsets quality measure. Adv. Intell. Syst. Comput. 314, 101–111 (2015)
6. Antoni, L., Krajci, S., Kridlo, O.: Randomized fuzzy formal contexts and relevance of one-sided concepts. LNAI (Subseries of LNCS) 9113, 183–199 (2014)
7. Lengler, R., Eppler, M.: Towards a periodic table of visualization methods for management. In: Proceedings of the International Conference on Graphic and Visualization in Engineering (GVE 2007), Clearwater, Florida, pp. 83–88 (2007)
8. Wills, G.: Visualizing hierarchical data. In: Encyclopedia of Database Systems, pp. 3425–3432 (2009)
9. Theron, R.: Hierarchical-temporal data visualization using a tree-ring metaphor. In: Smart Graphics. Springer, Berlin, pp. 70–81 (2006)
10. Itoh, T., Yamaguchi, Y., Ikehata, Y., Kajinaga, Y.: Hierarchical data visualization using a fast rectangle-packing algorithm. IEEE Trans. Visual Comput. Graphics 10(3), 302–313 (2004)
11. Neumann, P., Schlechtweg, S., Carpendale, S.: ArcTrees: Visualizing relations in hierarchical data. In: Proceedings of EuroVis 2005, pp. 53–60 (2005)
12. Jadeja, M., Shah, K.: Tree-map: A visualization tool for large data. In: Proceedings of 1st International Workshop on Graph Search and Beyond (GSB 2015), pp. 9–13 (2015)
13. Crampes, M., Oliveira-Kumar, J., Ranwez, S., Villerd, J.: Visualizing social photos on a hasse diagram for eliciting relations and indexing new photos. IEEE Trans. Visual Comput. Graphics 15(6), 985–992 (2009)
14. Holten, D.: Hierarchical edge bundles: visualization of adjacency relations in hierarchical data. IEEE Trans. Visual Comput. Graphics 12(5), 741–748 (2006)

Poďakovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.014TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Interactive visualization of hierarchical structure

Currently, there exist a large amount of data in digital form. They contain information which can be useful for companies or users in many possible ways (in making decisions, competitive intelligence, etc.). For that reason, a lot of methods has been proposed for their analysis. However, problem is to present result of the analysis in an appropriate form. Visualization is one of the possible ways to solve that problem. In this paper, we focus on the presentation of different visualization techniques of hierarchical structures, which can be obtained for example by Formal Concept Analysis.

Vyhľadávanie významných konceptov v rámci konceptuálnej analýzy dát

Miroslav Smatana, Peter Butka, Zuzana Čabalová

Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

{miroslav.smatana, peter.butka}@tuke.sk,
zuzana.cabalova@student.tuke.sk

Abstrakt. Existuje množstvo prístupov a nástrojov, ktoré slúžia na extrakciu konceptuálnych štruktúr zo vstupného datasetu. Ich hlavným cieľom je pomôcť používateľovi lepšie porozumieť vstupným dátam a vzťahom medzi nimi. Jednou z takýchto metód je formálna konceptová analýza (FCA), ktorá je schopná spracovať a analyzovať vstupné dáta v tvare objekt-atribútov tabuľky na základe ich vzťahov. FCA obsahuje niekoľko modelov, v tejto práci budeme pracovať s modelom zovšeobecneného jednostranne fuzzy konceptového zväzu (GOSCL), ktorý je schopný pracovať s rozličnými typmi atribútov. Avšak jedným z problémov GOSCL je množstvo konceptov, ktoré generuje. Existuje niekoľko prístupov, ktoré riešia problém generovania veľkého množstva konceptov. V tejto práci sa zameriame na získavanie len tých konceptov, ktoré môžu byť pre používateľa potencionálne užitočné na základe ním zadaného dopytu.

Kľúčové slová: formálna konceptová analýza, vyhľadávanie informácií, Levenshteinova vzdialenosť

1 Úvod

Jeden z prístupov používaných v oblasti analýzy dát je tzv. teórie konceptových zväzov. Tento prístup je známy ako formálna konceptová analýza (Formal Concept Analysis, FCA) [1,2] a je určený pre analýzu dát vo forme objekt-atribút modelu (formálny kontext). Výsledkom analýzy dát pomocou FCA je reprezentovaný pomocou konceptového zväzu, ktorý predstavuje hierarchicky organizovanú štruktúru skupín objektov (konceptov) so spoločne zdieľanými atribútmi. Analýza s využitím FCA našla svoje uplatnenie v oblastiach ako vyhľadávanie informácií, dolovaní znalostí, spracovaní prirodzeného jazyka, manažment znalostí a pod.

Existuje niekoľko metód FCA [3-5], avšak v našej práci sme sa zamerali na prácu s modelom zovšeobecneného jednostranne fuzzy konceptového zväzu (GOSCL) [6,7]. Narozdiel od ostatných metód FCA je GOSCL schopný spracovávať formálny kontext, ktorý môže obsahovať atribúty rozličných typov.
GOSCL zjednodušuje interpretáciu analýzy dát, ale ak je použitý na veľké alebo stredne veľké datasety, tak výsledný konceptový zväz obsahuje enormné množstvo

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 131-135.

konceptov a tým sa stáva neprehľadným. Bolo predstavených niekoľko redukčných metód [8-9] a rozličných vizualizácií, ktoré sa snažia riešiť tento problém. V tejto práci sme sa rozhodli použiť odlišný prístup, ktorý je založený na princípoch vyhľadávania informácií, kde používateľ zadá dopyt (čo ho v konceptovom zväze najviac zaujíma) a výsledkom sú koncepty, ktoré najviac vyhovujú jeho požiadavke.

2 Zovšeobecnený jednostranne fuzzy konceptový zväz

V tejto časti poskytneme len veľmi stručný popis základných charakteristík zovšeobecneného jednostranne fuzzy konceptového zväzu - GOSCL (viac informácií sa dozviete v práci [7]).

GOSCL predstavuje zaujímavý model v kontexte FCA. Používa jednostrannú fuzzifikáciu, čo znamená že jedna strana je ostrá (či sa prvok nachádza alebo nenachádza v množine) a druhá strana je fuzzy (atribúty nadobúda hodnoty v podobe fuzzy nožiny). Tento model ma taktiež aj iné výhody ako:

- Dokáže generovať konceptový zväz z objektov, ktoré sú reprezentované rozličnými typmi atribútov ako nominálne, ordinálne, numerické atď.
- Pracuje inkrementálne.

3 Navrhovaný prístup

V tejto sekcii je popísaný nami navrhovaný prístup. Hlavnou myšlienkou je aplikovanie procesu vyhľadávania informácií pre vyriešenie problému generovania veľkého množstva konceptov metódou GOSCL, kde sa snažíme získať len najzaujímavejšie koncepty vzhľadom k používateľom definovanému dopytu.

Celý proces pozostáva zo 4 základných krokov, kde najprv je generovaný konceptový zväz zo vstupného datasetu pomocou metódy GOSCL. Následne používateľ zadá dopyt, čo je pre neho vo vstupnom datasete zaujímavé. Ďalším krokom je nájdenie najzaujímavejších konceptov pre daný dopyt. To sa vykonáva na základe porovnania podobnosti konceptu k dopytu (keďže GOSCL je schopný spracovať rozličné typy atribútov, nie je možné použiť štandardné podobnostné metriky, preto sme v rámci práce navrhli modifikovanú Levenshteinovu vzdialenosť popísanú v kapitole 3.1). Posledným krokom je vizualizácia získaných výsledkov.

3.1 Modifikovaná Levenhsteinova vzdialenosť

Pre konceptové zväzy s ostrými hodnotami je jednoduché nájsť najzaujímavejšie koncepty na základe dopytu. Je to možné vykonať pomocou metrík podobnosti ako Hamming alebo Jaccard. Taktiež pre konceptové zväzy využívajúce fuzzy model, ktoré obsahujú len intervalové atribúty, je to možné dosiahnuť pomocou metrík ako Euklidovská a kosínusovú. Problém nastáva pri konceptových zväzoch, ktoré obsahujú viacero typov atribútov, kde môžeme mať aj hodnoty, ktoré sú neporovnateľné.

Preto sme sa rozhodli modifikovať Levenshteinovu vzdialenosť (http://www.levenshtein.net/) pre nájdenie najzaujímavejších konceptov k dopytu. Pseudokód výpočtu vzdialenosti:

1. inicializácia vzdialenosť = 0, n = počet atribútov, Q = dopyt používateľa, C = vektor atribútov pre koncept
2. pre i = 1:n
   a. porovnaj Q[i] s C[i]
   b. ak je Q[i] menšie ako C[i] potom vzdialenosť += 1
   c. ak je Q[i] neporovnateľné s C[i] potom vzdialenosť +=2
3. vráť vzdialenosť

4 Experimenty

V tejto časti sú popísané experimenty, kde sme navrhovaný prístup testovali v dvoch fázach. V prvej fáze sme mali vstupný dataset pozostávajúci z 50 objektov a 5 atribútov (intervalového typu) a kvalitu navrhovaného prístupu sme porovnávali pomocou metrík presnosť a návratnosť [10], kde sme mali pre každý dopyt definované, ktoré objekty sú relevantné a ktoré nie. Následne sme vypočítali skoré (zhodu s dopytom) pre každý koncept v konceptovom zväze a vybrali sme N objektov z konceptov s najvyšším skóre. Výsledky je možné vidieť v **Tab. 1**, kde nami navrhovaná metóda dosahuje porovnateľné výsledky ako ostatné štandardne používané metriky na našej testovacej vzorke.

Tab. 1. Výsledky prvej fázy experimentov - presnosť a návratnosť pre rozličné dopyty Q a rôzne typy vzdialenosti (MLD predstavuje našu modifikovanú Levenshteinovu vzdialenosť).

N=10	P/R	Q1	Q2	Q3
Euklidovská	P	0.4	**0.7**	**0.6**
Euklidovská	R	0.2	**0.23**	**0.4**
Kosínusová	P	0.3	0.5	**0.6**
Kosínusová	R	0.15	0.17	**0.4**
Jaccard	P	**0.5**	0.4	0.4
Jaccard	R	**0.25**	0.17	0.27
Hamming	P	**0.5**	0.5	0.5
Hamming	R	**0.25**	0.17	0.33
MLD	P	0.4	0.4	0.5
MLD	R	0.2	0.13	0.33

Modifikovaná Levenshteinova vzdialenosť je vhodnejšia pre komplexnejšie atribúty, preto sme v druhej fáze experimentov aplikovali túto vzdialenosť na dáta s heterogénnym typom atribútov. Tieto dáta pojednávali o poistení aut, kde každé auto pozostávalo z 3 atribútov (typ auta, vodičové skúsenosti, typ poistky). Vyhodnocovanie kvality sa vykonávalo rovnako ako v prvej fáze experimentov. Výsledky boli porovnateľ-

né s výsledkami dosiahnutými na klasických vstupoch, napríklad na dopyte {BMW, Expert, Full} sme dosiahli presnosť 0.6 a návratnosť 0.3.

5 Záver

V práci sme prezentovali prístup pre riešenie problému generovania veľkého počtu konceptov pri formálnej konceptovej analýzy s využitím prístupov vyhľadávania informácií. Náš prístup založený na modifikovanej Levenshteinovej vzdialenosti dosahoval výsledky porovnateľné so štandardnými metrikami, pričom je schopný pracovať s rozličnými typmi atribútov.

Literatúra

1. Wille, R.: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. Springer, Netherlands (1982).
2. Birkhoff, G.: Lattice Theory. American Mathematical Soc. (1940)
3. Belohlavek, R.: Lattices of Fixed Points of Fuzzy Galois Connections. Math. Log. Quart. 47(1), 111–116 (2001).
4. Krajci, S.: A generalized concept lattice. Logic Journal of IGPL 13(5), 543–550 (2005).
5. Medina, J., Ojeda-Aciego, M., Ruiz-Calvino, J.: Formal concept analysis via multi-adjoint concept lattices. Fuzzy Set. Syst. 160, 130–144 (2009).
6. Krajci, S.: Cluster based efficient generation of fuzzy concepts. Neural Netw. World 13(5), 521–530 (2003)
7. Butka, P., Pocs, J.: Generalization of One-Sided Concept Lattices. Comput. Informat. 32(2), 355–370 (2013)
8. Butka, P., Pocs, J., Pocsová, J.: Reduction of concepts from generalized one-sided concept lattice based on subsets quality measure. Adv. Intell. Syst. Comput. 314, 101–111 (2015)
9. Antoni, L., Krajci, S., Kridlo, O.: Randomized fuzzy formal contexts and relevance of one-sided concepts. In: LNAI (Subseries of LNCS) 9113, pp. 183–199 (2014)
10. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proc. of the 23rd international conference on Machine learning. ACM, 233-240 (2006)

Poďakovnie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.025TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Retrieval of Important Concepts in Formal Concept Analysis

Currently, there exist a lot of methods and tools for extraction of conceptual structures from input dataset. The main aim of these methods is to describe input data and relations between them so the user will better understand them. One of such methods is Formal Concept Analysis (FCA), which is used for analysis of object-attribute input data models. FCA contains several methods, but in this work, we focus on Generalized One-Side Concept Lattices (GOSCL), which is able to process attributes of dif-

ferent types. However, the problem of GOSCL method is a number of concepts which it generates. In this paper, we describe our approach suitable for analysis of GOSCL models, where modified Levenshtein distance is used for retrieval of important concepts from output concept lattice based on user query.

Hierarchické prístupy k modelovaniu témy v dokumentoch

Miroslav Smatana, Peter Butka, Matúš Gore

Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach
Letná 9, 042 00 Košice, Slovenská republika

{miroslav.smatana, peter.butka}@tuke.sk,
matus.gore@student.tuke.sk

Abstrakt. Digitálne textové dáta predstavujú v dnešnej dobe dôležitý zdroj informácií. Avšak v súčasnosti je ich počet taký obrovský, že ich manuálne spracovanie a extrakcii informácií by bola časovo veľmi náročná. Existuje niekoľko spôsobov automatickej analýzy textových dát, jednou z nich je modelovanie témy, ktoré ponúka nové možnosti na vyhľadávanie, prehľadávanie a sumarizáciu textových dokumentov. Preto je hlavným cieľom tohto článku predstaviť rôzne metódy hierarchického modelovania téma a taktiež porovnanie vybraných modelov z pohľadu kvality budovania hierarchie tém.

Kľúčové slová: modelovanie témy, hierarchia, Latentná Dirichletova Alokácia

1 Úvod

V súčasnosti s príchodom sociálnych sieti, online časopisov a novín, je veľké množstvo textových dát v digitálnej podobe. Tie predstavujú zaujímavý a dôležitý zdroj informácií. Napríklad na sociálnej sieti Twitter je denne publikovaných okolo 340 miliónov príspevkov, ktoré zvyčajne odrážajú používateľov postoj na niektorú zo svetových udalostí, produkt, či osoby.

Takéto dáta nachádzajú svoje využitie najmä v marketingu, pretože marketing bol stále závislý na dátach a ich správne pochopenie a použitie môže firme priniesť konkurenčnú výhodu a zlepšiť jej postavenie na trhu. Využitie digitálnych textových dát (najmä príspevky zo sociálnych sietí) je možné použiť napríklad pri:

- analýze krízových situácií - príkladom je obdobie vojny, kedy je z týchto dát možné získať reakcie ľudí na danú situáciu;
- zavedenie nového produktu na trh - monitorovanie reakcií používateľov, čo sa im na produkte páči, aké ma chyby;
- v médiách - je možné sledovať o čom ľudia na internete najčastejšie píšu a čo ich zaujíma.

Ako je možné vidieť, analýzu digitálnych textových dát nemožno podceňovať. Avšak ako už bolo spomenuté existuje ich veľké množstvo a preto ich manuálne spracova-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 136-140.

nie, analýza a extrakcia vhodných informácií by bola veľmi časovo náročná. Preto je potrebné tento proces automatizovať.

V súčasnosti existuje niekoľko druhov analýz, ktoré tento problém riešia. Jednou z nich je modelovanie tém, ktoré umožňuje automatickú analýzu textových príspevkov a predstavuje nový spôsob ich hľadania, prehľadávania a sumarizácie. Hlavným cieľom modelovania tém je vytváranie skupín slov (tém), ktoré sa často vyskytujú spoločne vo vstupnej množine textových dokumentov.

Preto sa v článku sa zameriame na prehľad metód modelovania témy a porovnanie metód hierarchického modelovania témy, ktoré ponúka podrobnejšiu analýzu ako klasické metódy.

2 Modelovanie témy

V tejto sekcii predstavíme modelovanie témy. Ako už bol povedané, cieľom modelovania tém je vytváranie skupín (tém), ktoré sa spolu vyskytujú dostatočne často, pomocou hľadania skrytých tematických štruktúr vo vstupnej kolekcii dokumentov. Ako jednou z prvých metód, ktorá sa pokúšala riešiť tento problém môže byť chápaná latentná sémantická analýza [1], ktorá sa snaží odhaľovať skryté sémantické štruktúry z textov. Formálne však nie je predstavovaná ako jedna z metód modelovania témy, ale bola základom pre ďalšie rozšírenia ako pravdepodobnostná latentná sémantická analýza [2], ktorá tvorí základ pre Latentnú Dirichletovu Alokáciu (LDA) [3]. LDA predstavuje jednu z najpoužívanejších metód v oblasti modelovania tém. Z toho dôvodu sa stala základom pre ďalšie metódy modelovania tém [4,5]. Okrem toho vzniklo niekoľko metód, nezávislých od LDA [6,7,8].

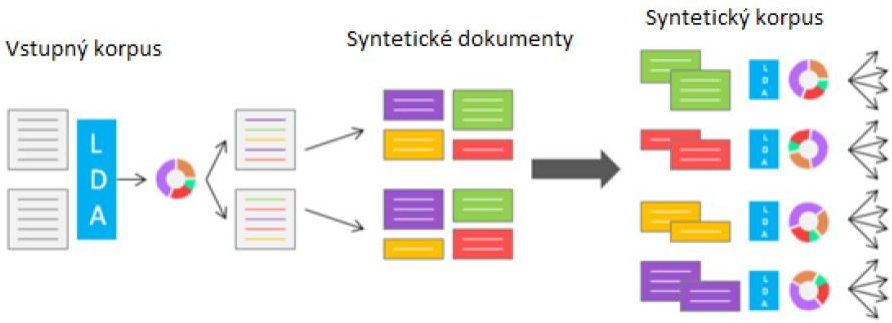
Spomínané práce sa zameriavali na spracovanie dlhých dokumentov. Treba však povedať, že na dátach ako sú príspevky zo sociálnych sieti, kde tieto správy obsahujú len veľmi krátke a stručné texty, nedosahujú dobré výsledky. Preto bolo vyvinutých niekoľko metód [9,10,11], ktoré sú schopne pracovať s krátkymi textami.

Ako je možné vidieť, modelovanie témy je dobre preštudovaná oblasť výskumu, avšak v súčasnosti už tieto štandardné metódy nemusia používateľovi poskytovať dostatočné množstvo informácií. Preto vzniklo niekoľko metód zameraných na budovanie hierarchie tém, ktoré poskytuje podrobnejšie informácie. Jednou z prvých takýchto metód bolo hierarchické LDA (hLDA) [13] a práca [12] v ktorej autori na budovanie tém využívajú proces vnorenej čínskej reštaurácie. Ďalšími z metód hierarchického modelovania témy sú Pachinov alokačný model [14] a metóda HLTA [15].

2.1 hLDA

Metóda hLDA predstavuje jednu z najjednoduchších metód budovania hierarchie tém. Táto metóda využíva prístup "zhoda-nadol" pre vytváranie hierarchie tém pomocou rekurzívneho rozdeľovania a znovu modelovania korpusu pomocou klasické LDA. Výsledkom LDA je model, kde každé slovo v dokumente je priradené téme. HLDA využíva túto informáciu na vytváranie nových syntetických dokumentov pre každú

nájdenú tému zo vstupných dokumentov. Syntetické dokumenty obsahujú iba slová priradené danej téme. Celý tento proces je znázornený na **Obr. 1**.



Obr. 1. Proces vytvárania tém pomocou hLDA

3 Porovnanie metód hierarchického modelovania tém

V tejto kapitole popíšeme porovnanie nami implementovanej metódy hLDA s metódou HLTA. Zvolené metódy sme porovnávali na datasete 20 Newsgroups[1], ktorý obsahuje 20000 dokumentov rozdelených do 20 rôznych tém.

Kvalitu metód sme porovnávali na základe rýchlosti budovania hierarchie a taktiež na základe "coherence score" [16]. V **Tab. 1** je možné vidieť rýchlosť vytvárania hierarchie pre metódy hLDA a HLTA na rôznych vzorkách zo vstupného datasetu (na každej zo vzoriek bolo vykonané jedno meranie). Ako je možné vidieť metóda hLDA je omnoho pomalšia a s narastajúcim počtom vstupných dokumentov sa čas jej spracovania drasticky zvyšuje. Z pohľadu "coherence score" nám, na vzorke 30 dokumentov z 20 Newsgroups datasetu, pre hLDA vyšla hodnota -14.6412 a pre HLTA hodnota 12.6328. Na základe týchto hodnôt je možne povedať, že na zvolenom datasete dosahovali tieto metódy porovnateľné výsledky z hľadiska kvality vytvorených tém.

Tab. 1. Porovnanie metód hierarchického modelovania tém na základe rýchlosti budovania hierarchie

Čas (min)	20NewsGroup (10 (dokumen-tov)	20NewsGroup (20 dokumen-tov)	NewYork Times (40 dokumen-tov)	NewYork Times (50 dokumen-tov)
hLDA	**23,48**	**49,33**	**95,13**	**130,20**
HLTA	**0,036**	**0,078**	**0,25**	**0,32**

[1] http://qwone.com/~jason/20Newsgroups/

4 Záver

V práci sme prezentovali prehľad metód z oblasti modelovania témy ako aj jej podoblasti hierarchického modelovania tém, ktoré predstavujú podrobnejšiu analýzu. Práca taktiež zdôrazňovala potrebu a potenciál využitia týchto metód v reálnom svete, a to najmä v oblasti marketingu. Koniec práce bol venovaný porovnaniu metód hLDA a HLTA, kde sa ukázalo, že z hľadiska času spracovania bola metóda LTA oveľa rýchlejšia, avšak z pohľadu kvality vytvorených tém dosahovali tieto metódy porovnateľné výsledky.

Literatúra

1. Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284
2. Hofmann, T. (1999, August). Probabilistic latent semantic indexing. InProceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 50-57). ACM.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.
4. Petterson, J., Buntine, W., Narayanamurthy, S. M., Caetano, T. S., & Smola, A. J. (2010). Word features for latent dirichlet allocation. In Advances in Neural Information Processing Systems (pp. 1921- 1929).
5. Zhai, K., & Boyd-Graber, J. (2013). Online Latent {D} irichlet Allocation with Infinite Vocabulary. In Proceedings of the 30th International Conference on Machine Learning (pp. 561-569).
6. Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2012). Hierarchical dirichlet processes. Journal of the american statistical association.
7. Li, X. M., Ouyang, J. H., & Lu, Y. (2015). Topic modeling for large-scale text data. Frontiers of Information Technology & Electronic Engineering, 16, 457-465.
8. Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. The Journal of Machine Learning Research, 14(1), 1303-1347.
9. Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th international conference on World Wide Web (pp. 91-100). ACM.
10. Sridhar, V. K. R. (2015, June). Unsupervised topic modeling for short texts using distributed representations of words. In Proceedings of NAACL-HLT (pp. 192-200).
11. Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015, June). Short and sparse text topic modeling via self-aggregation. In Proceedings of the 24th International Conference on Artificial Intelligence (pp. 2270-2276). AAAI Press.
12. Griffiths, D. M. B. T. L., & Tenenbaum, M. I. J. J. B. (2004). Hierarchical topic models and the nested chinese restaurant process. Advances in neural information processing systems, 16, 17.
13. Hofmann, T. (1999, July). The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In IJCAI (Vol. 99, pp. 682-687).
14. LI, Wei a Andrew McCALLUM: Pachinko allocation: Scalable micture models of topic correlations. In: Journal of Machine Learning. s.2-30. 2008.

15. CHEN, Peixian a kol.: Progressive EM for latent tree models and hierarchical topic detection. 13th AAAI Conference on Artificial Intelligence. 2016.
16. MCCALLUM, Andrew: Topic model diagnostics [online]. 2002. [cit. 2017-03-15]. Dostupne z: http://mallet.cs.umass.edu/diagnostics.php

Poďakovanie: Tento príspevok vznikol s podporou VEGA projektu č.1/0493/16, KEGA projektu č.025TUKE-4/2015 a APVV projektu č.APVV-16-0213.

Annotation:

Hierarchical topic modeling

Text documents in digital form represent an important source of documents. However currently is their number so large that their manual processing and extraction of information would be time-consuming. For now, there exist several methods of automatic analysis of textual data, one of them is topic modeling. It offers new ways of searching, browsing and summarizing of textual data. For that reason, the main aim of this work is to present methods of topic modeling and also compare selected models for hierarchical topic modeling.

Automatizace klasifikace evropských projektů pomocí klasifikátoru

Ondřej Zamazal

Fakulta informatiky a statistiky
Vysoká škola ekonomická v Praze
nám. W. Churchilla 1938/4, 130 67 Praha 3, Česká republika

ondrej.zamazal@vse.cz

Abstrakt. Finanční prostředky Evropské unie do vytváření pracovních míst, do evropské ekonomiky a životního prostředí jsou poskytovány prostřednictvím pěti evropských strukturálních a investičních fondů (ESI fondy). Ačkoliv EU má pro kategorizaci jednotlivých projektů jednotný kategorizační systém, tak jednotlivé členské země EU používají různé své vlastní kategorizační systémy. Některé země již přímo používají nový kategorizační systém, ale mnohé země takto stále ještě nepostupují. V dostupných datasetech o evropských projektech tak zůstává značné množství projektů nezařazeno s ohledem na jednotný kategorizační systém EU. Cílem této práce je vyzkoušení možnosti automatické klasifikace evropských projektů pomocí klasifikátoru. Podpora automatickou klasifikací by podpořila fiskální analýzy.

Klíčová slova: evropský projekt, číselník, strojové učení, klasifikace

1 Úvod

Evropská unie podporuje v jednotlivých členských zemích projekty pro vytváření pracovních míst, zdravou evropskou ekonomiku a dobré životní prostředí prostřednictvím pěti evropských strukturálních a investičních fondů (ESI fondy). Pro kategorizaci těchto evropských projektů je k dispozici jednotná kategorizace pro období 2007 – 2013 a zvlášť pro období 2014 – 2020 s explicitním propojením.[1] Členské země EU však při evidování evropských projektů tuto kategorizaci uvádět nemusí a také tak všechny nečiní. Kategorizace projektů se tak musí dělat ručním způsobem až ex post. Jednotný způsob kategorizace evropských projektů umožňuje přímočaré fiskální analýzy. Motivací této práce je podpořit klasifikaci evropských projektů do kategorizačního systému pro období 2014 – 2020 automatickými prostředky strojového učení, což by umožnilo fiskální analýzy. V tomto příspěvku se zabýváme přístupem strojo-

[1] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/data/categorisation_2014_2020_mapping.xls

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 141-145.

vého učení SVM pro automatickou klasifikaci evropských projektů. V části 2 vysvětlujeme použitá data. V části 3 popisujeme celkový postup přístupu a úvodní experimenty jsou popsány v části 4.

2 Otevřená data evropských projektů

Jednotlivé země zveřejňují informace o evropských projektech na úrovni regionů a celých zemí. Přestože data jsou vystavena odpovědnými orgány, tak jejich formát zůstává poněkud „zavřený". Někdy jsou data vystavena ve formátu pdf, byť samotná data jsou zapsána ve formě tabulek, jindy jsou data vystavena rovnou v příhodnějším tabulkovém formátu. Sběrem a sdílením těchto dat se zabývá organizace „Open Knowledge Foundation Deutschland".[2] Datasety jsou ukládané v různých formátech CSV, XSLX nebo JSON. V rámci zpracování dat se také provádí provázání jednotlivých typů informací (atributů) na jednotný fiskální datový model Open Spending.[3] Kromě vystavených datasetů je k dispozici také jednotný integrovaný dataset [2,3], který lze stáhnout (CSV formát, 879 MB, 2,7 miliónů řádků popisů projektů) a analyzovat vlastními silami nebo za použití připraveného analytického nástroje OS Viewever.[4]

Nový kategorizační systém projektů (podobně jako ten starý) je dostupný v příslušných evropských dokumentech a také v tabulkové podobě z webu „Data for research".[5] Podle informací z tabulkových dat jsme z nich vyextrahovali samostatný RDF [4] číselník obsahující kategorizační systém pro staré odbobí 2007-2013, RDF číselník obsahující klasifikační systém pro nové období 2014-2020 a jejich vzájemná mapování.[6]

3 Postup automatické klasifikace evropských projektů

V rámci prvního přístupu ke klasifikaci projektů jsme pracovali s jednotlivými datasety regionů a zemí samostatně. Datasety obsahují popisy projektů s velmi různorodou mírou záběru (např. některé popisy obsahují jen název projektu, dotace EU, region a termín jiné obsahují také slovní popis projektu ad). Kromě numerického popisu projektů tak bývá k dispozici také slovní popis různého druhu. V našem přístupu se zaměřujeme na využití slov uvedených u jednotlivých projektů. V první fázi jsme dávali dohromady dostupná data. Poloautomaticky jsme našli 20 datasetů z celkových 110, ve kterých se objevuje atribut týkající se informace zařazení do kategorie podle intervenčního kódu. Z těchto datasetů jsme v rámci předzpracování pomocí regulárních

[2] https://www.okfn.de/en/

[3] https://github.com/os-data/eu-structural-funds/blob/master/specifications/fiscal.schema.yaml

[4] http://subsidystories.eu

[5] http://ec.europa.eu/regional_policy/sources/docgener/evaluation/data/categorisation_2014_2020_mapping.xls

[6] viz https://github.com/openbudgets/ v části linksets a Code-lists.

výrazů a přesné podoby intervenčních kódů podle RDF číselníku extrahovali intervenční kód a dále samostatná slova jednotně oddělena mezerami. Takto vzniklá data jsme deduplikovali.

Za účelem „sémantického" propojení slov používáme automatický jazykový překlad všech slov do angličtiny, protože jednak automatický překlad do angličtiny bývá nejlepší a jednak angličtina je používána v popiscích jednotlivých intervenčních kódů v RDF číselníku. Pro automatický jazykový překlad jsme vybrali „Microsoft Translator API".[7] Výběr byl proveden na základě dostupnosti API zdarma (2 milióny přeložených znaků za měsíc) a pokrytí všech požadovaných evropských jazyků.

Jednotlivé projekty tak reprezentujeme pomocí vektorů, které mají tolik složek kolik je v datech unikátních slov (po odstranění stop slov) a binárně sledujeme jejich (ne)výskyt. Přirozené úskalí je velká dimenze vektorů, která by mohla být snížena pomocí shlukování slov nebo například latentní sémantickou analýzou, kde by se identifikovaly koncepty složené ze slov a těmi se indexovaly jednotlivé projekty [1]. V rámci našeho přístupu jsme dimenzi vektorů nesnižovali a spíše jsme se zaměřili na algoritmy, které zvládají práci s daty o vysoké dimenzionalitě. Na základě úvodního testování jsme se rozhodli pro vytvoření SVM (Support Vector Machines) klasifikátoru[8], za použití knihovny LibSVM pro Javu, a jeho lineárního kernelu, který oproti jiným kernelům (RBF kernel, polynomiální kernel) dosahoval nejlepší přesnosti na trénovacích datech.

Úvodní analýzy ukázaly, že získaná data jsou rozložena do tříd (123 intervenčních kódů) značně nerovnoměrně. Celkem 10992 instancí patří do 97 tříd, z nichž 23 tříd má pouze jednu instanci, 43 tříd má alespoň 10 instancí, 34 tříd má alespoň 20 instancí a pouze 20 tříd má více než 100 instancí. Nízké zastoupení velkého množství tříd v trénovacích datech a dominance několika tříd by při automatické klasifikaci vedlo k preferování majoritních tříd a znehodnocení výsledků na reálných datech. V rámci našeho přístupu jsme úskalí nevyváženosti tříd řešili pomocí vzorkování.

4 Experiment automatické klasifikace evropských projektů

Pro vyzkoušení našeho přístupu jsme provedli experimenty A, B, C a D se získanými daty s různým přístupem k vytvoření trénovacích a testovacích dat. V rámci přípravy trénovacích dat vždy používáme vzorkování tak, aby všechny třídy byly zastoupeny stovkou instancí. K tomu se používá podle potřeby „oversampling" tak, že se náhodně některé instance duplikují a „undersampling" tak, že se některé náhodně vybrané instance smažou. Experiment A pro trénování uvažuje jen třídy s více než 50 instancemi (celkem 21 tříd). Experiment B pro trénování používá všechny třídy s alespoň jednou instancí (97 tříd). Experimenty C a D uvažují všechny třídy pro trénování (123 tříd), kde pro třídy bez instancí jsou použity popisky z RDF číselníku. Pro sestavení testovacích dat v experimentech A, B a C se vybírá náhodně 20 instancí ze tříd, které

[7] https://www.microsoft.com/en-us/translator/translatorapi.aspx

[8] Dále jsme v nástroji Weka zkoušeli logistickou regresi (neúspěšně kvůli příliš vysoké dimenzionalitě dat) a algoritmus pro učení rozhodovacích pravidel JRip (srovnatelná přesnost jako SVM ale příliš pomalé).

mají více než 50 instancí. Experiment D do testovacích dat zařazuje vždy polovinu instancí od každé třídy s více než jednou instancí. Testovací a trénovací množiny jsou vždy disjunktní. Testování bylo spuštěno 3x a výsledná přesnost klasifikace je jejich průměrem viz tabulka 1 spolu s uvedenými charakteristikami jednotlivých experimentů.

Výsledná přesnost pro všechny experimenty vychází relativně vysoká. Nejlépe vypovídající je experiment D, kde trénujeme a testujeme na nejvíce třídách a nejvíce instancích.

Tabulka 1: experimenty (Tr | Ts znamená trénovací | testovací množina)

Experiment	Tr: #tříd (#instancí)	Ts: #tříd (#instancí)	Přesnost
A	21 (2100)	21 (480)	0.851
B	97 (9700)	21 (480)	0.95
C	123 (12300)	21 (480)	0.951
D	123 (12300)	74 (3653)	0.93

5 Závěr

Provedené experimenty ukazují slibné výsledky za situace, kdy máme získaná data rozdělena na disjunktní množiny trénovacích a testovacích dat. Vypovídající schopnost těchto experimentů je však omezená. Testovací data totiž pocházejí z datasetu zemí, ze kterých byla použita data (sice jiná než ta pro testování) pro trénování klasifikátoru a současně platí, že struktura popisu projektů v rámci jednoho datasetu bývá dosti podobná. Závěrem proto musíme konstatovat, že opravdu vypovídající výsledky testování mohou být dosáhnuty jen na testovacích datech datasetů, ze kterých žádné instance nebyly použity pro trénování. To je také záměrem naší plánované práce s integrovaným datasetem evropských projektů.

Literatura

1. Berka, Petr. Dobývání znalostí z databází. Academia, 2003.
2. SubsidyStories.eu. Dataset Descriptions. https://github.com/os-data/eu-structural-funds/blob/master/documentation/subsidyreport%20-%20dataset%20descriptions.pdf
3. SubsidyStories.eu. Methodology & Variables. 2017. https://github.com/os-data/eu-structural-funds/blob/master/documentation/subsidyreport%20-%20methodology.pdf
4. World Wide Web Consortium. RDF 1.1 concepts and abstract syntax. (2014).

Poděkování: Tato práce byla podpořena z projektu EU H2020 č. 645833 OpenBudgets.

Annotation:

Automatic Classification of European projects

European union funding for job creation and a sustainable and healthy European economy and environment is channelled through the 5 European structural and investment funds (ESIF). Although there is European categorization system for EU projects, EU countries apply their own different categorization systems. Some EU countries already apply European categorization system, but many do not. As a result, many projects are not categorized using European categorization system in available datasets. The goal of this work is to examine an option of an automatic classification of European projects using a machine learning classifier. This would enable straightforward fiscal analyses and interlinking categorization systems of EU countries to European one.

Fokusovaná kategorizační síla webových ontologií

Vojtěch Svátek[1], Ondřej Zamazal[1], Miroslav Vacura[2]

[1]Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze,
nám. W. Churchilla 1938/4, 130 67 Praha 3
[2]Katedra filozofie, Vysoká škola ekonomická v Praze,
nám. W. Churchilla 1938/4, 130 67 Praha 3

{svatek,ondrej.zamazal}@vse.cz
vacuram@vse.cz

Abstrakt: Přepoužívání ontologií je jedním z předpokladů efektivního využívání propojených dat na sémantickém webu. Výběr ontologie pro přepoužití bývá v současnosti založen na textovém vyhledávání a metrikách popularity ontologie. Kritériem přepoužitelnosti ontologie však může být i její schopnost vyjádřit podkategorie klíčových tříd objektů z aktuální datové sady, tzv. fokusových tříd. Tuto schopnost navrhujeme kvantifikovat pomocí fokusované kategorizační síly ontologie: míry odvozené z velikosti množiny binárních možností, které model nabízí pro kategorizaci objektů již přiřazených k obecnější fokusové třídě. Přepoužitelnými podkategoriemi mohou být i složené konceptové výrazy, a to s různou mírou jistoty, která se promítá do váhových koeficientů používaných ve výpočtu kategorizační síly. Pravděpodobnost, s jakou je výraz přepoužitelnou kategorií, můžeme odhadovat mj. na základě hodnocení lidskými respondenty.

Klíčová slova: ontologie, sémantický web, propojená data, kategorizace.

1 Úvod

Webové ontologie lze zhruba charakterizovat jako soubory konceptů (neboli tříd) a typů vztahů (vlastností), označených globálními identifikátory IRI a propojených množinou logických tvrzení v ontologickém jazyce OWL[1] odpovídajícímu relativně expresivní variantě deskripční logiky [1]. Ontologie jsou stále více využívány pro přiřazování strojově čitelné sémantiky strukturovaným datům vystavovaným na webu. Motivací je zejména možnost souběžného automatického zpracování nezávisle vzniklých datových zdrojů. Například, pokud jsou výrobky stejného druhu nabízené různými e-shopy sémanticky popsány pomocí stejných tříd a vlastností (souhrnně, entit), lze relativně snadno implementovat jejich automatické porovnávání a doporučování. To ovšem vyžaduje, aby pojmy z různých ontologií byly *přepoužívány*. Vystavovatel datové sady pak musí při tvorbě jejího schématu, namísto izolovaného návrhu nových

[1] https://www.w3.org/TR/owl2-primer/

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 146-150.

entit, vložit úsilí do nalezení relevantních existujících ontologií a jejich entity do schématu zaintegrovat – ať už přímo, nebo pomocí subsumpčního či ekvivalenčního mapování.

Výběr ontologie pro přepoužití je ovšem netriviální úlohou, pro kterou teprve v posledních letech vznikají exaktní metody. Ty vesměs, vedle *textového vyhledávání*, spoléhají na *metriky popularity* (ontologie nebo jednotlivých entit), např. kolik instancí v kolika různých datových sadách se na ně odkazuje, případně na míry apriorní *důvěryhodnosti* ontologie [2], např. zda jsou ontologie zachyceny v autoritativním katalogu, jako je LOV.[2] Pilotní studie zaměřená na strategie přepoužívání webových ontologií [2] naznačila, že vydavatelé datových sad preferují přepoužít větší počet entit z nižšího počtu ontologií i za cenu nižší průměrné míry jejich popularity. Míry založené na popularitě navíc trpí problémem studeného startu: mnoho kvalitních ontologií je nových a jejich potenciál proto nelze odkazovou popularitou spolehlivě hodnotit.

Navrhovaný přístup ke zlepšení procesu přepoužívání ontologií je postaven na následující intuici:

1. Využití ontologií na webu má často charakter přiřazení objektů k určitým kategoriím, s tím, že již před tímto přiřazením je o objektech známo, že jsou instancemi určité (obecnější) třídy, kterou označíme jako *fokusovou třídu*.
2. Přiřazované kategorie nemusí být nutně v ontologii uvedeny jako pojmenované třídy, ale může se jednat o *složené konceptové výrazy* zkonstruované z pojmenovaných entit pomocí operátorů příslušné deskripční logiky.
3. Počet a „kvalita“ kategorií, které ontologie nabízí pro určitou fokusovou třídu, je *indikátorem přepoužitelnosti* ontologie pro datovou sadu obsahující objekty patřící do této třídy.

Téma bylo in extenso zpracováno v příspěvku na evropské konferenci EKAW 2016 [3] (zařazen v nominaci na *Best Paper Award*). V tomto referativním příspěvku pouze nastíníme navrženou metodu a shrneme hlavní dosažené výsledky.

2 Schéma výpočtu kategorizační síly

Uvažujme množinu n typů konceptových výrazů nepřímo vymezenou formálním jazykem $\mathcal{L}$ (nad konstrukty deskripční logiky) a množinu *vzorů* používaných pro jejich detekci v ontologiích, $P = \{p_1, \ldots, p_n\}$. Odhad fokusované kategorizační síly ontologie O vzhledem k fokusové třídě FC pak můžeme vyjádřit jako

$$\widehat{FOCP}(FC,\mathcal{L},O) = Occ(p_1, FC,O) \,.\, w_1 + \ldots + Occ(p_n, FC,O) \,.\, w_n$$

kde $Occ(p_i, FC,O)$ je funkce vracející počet výskytů vzoru p_i v O, a w_i jsou váhové koeficienty z intervalu (0, 1]. Jedním z typů konceptových výrazů je i „pojmenovaná třída“, reprezentující tranzitivní podtřídy třídy FC. Ta je detekována vzorem vyjádřeným ve formě subsumpčního tvrzení C `rdfs:subClassOf` FC a vyhledávaným nad

[2] http://lov.okfn.org/dataset/lov/

tranzitivním uzávěrem ontologie; *C* je zde proměnná, za kterou se podtřídy dosazují. V případě tohoto vzoru budeme předpokládat hodnotu w_i rovnou 1.

Pro stanovení adekvátních *váhových koeficientů* složených konceptových výrazů se nabízejí dva hlavní zdroje: zpětná vazba od uživatelů, a automatická empirická analýza – jednak samotných ontologií, jednak datových sad, které se na ně odvolávají.

3 Provedené experimenty a jejich výsledky

V první fázi výzkumu byla využívána zejména zpětná vazba od uživatelů k jednotlivým konceptovým výrazům různých typů. Rozlišovány byly tři typy složených výrazů (konkrétní příklady jsou níže), které lze vyjádřit variantami existenční restrikce nad vlastností *R*: $FC \sqcap \exists R.C$ (s hodnotou vlastnosti vymezenou pomocí třídy *C*), $FC \sqcap \exists R.\{i\}$ (tzv. "value restriction" s hodnotou vyjádřenou konkrétní instancí *i*), a $FC \sqcap \exists R.\top$ (s hodnotou vlastnosti "vymezenou" univerzálním "super-konceptem" $\top$, tedy nijak neomezenou). Připomeňme si ovšem, že v ontologiích se nevyskytují samotné konceptové výrazy, ale logická tvrzení. Na ně je nutno aplikovat ontologické vzory z *P*, abychom získali hodnoty jednotlivých $Occ(p_i, FC, O)$. Mírně zjednodušeným příkladem takového "konstrukčního" vzoru pro výraz $FC \sqcap \exists R.C$ je

$\exists D$ (R `rdfs:domain` $FC \wedge R$ `rdfs:range` $D \wedge C$ `rdfs:subClassOf` D)

tj. výraz zkonstruujeme tehdy, jestliže má vlastnost R jako svůj definiční obor (ať už přímo, nebo s využitím dědičnosti) fokusovou třídu *FC*, a zároveň má jako svůj obor hodnot určitou třídu *D* takovou, že třída *C* je její podtřídou.

Uživatelé měli za úkol rozhodnout, zda daný konceptový výraz považují za přepoužitelnou[3] kategorii vzhledem k určité jeho logické nadtřídě *FC*, nebo ne. Příklady konceptových výrazů (tří výše uvedených typů), které byly uživateli, konkrétně, 27 studenty dvou předmětů zaměřených na ontologické inženýrství, resp. propojená data, relativně často vnímány jako přepoužitelné kategorie, jsou:[4]

- $Place \sqcap \exists\ isEquippedBy.AudiovisualEquipment$
- $FridgeFreezer \sqcap \exists\ styleOfUnit.\{SingleDoor\}$
- $ProgramCommitteeMember \sqcap \exists\ writeReview.\top$

V prvním případě je kategorie "místa" upřesněna třídou *AudiovisualEquipment* omezující hodnoty vlastnosti *isEquippedBy*. Ve druhém případě je *SingleDoor* formálně individuem, avšak v realitě jde opět o vyjádření obecné kategorie (stylu chladničky). Ve třetím případě je vymezení kategorie dáno pouze vlastností *writeReview*; omezující kategorie hodnoty vlastnosti ("recenze") je zde přítomna implicitně v jejím názvu, proto kategorie dává smysl i bez explicitního upřesnění hodnoty ve struktuře výrazu.

[3] Za přepoužitelnou kategorii měl uživatel považovat takovou, u které by ho "nepřekvapilo", kdyby byla vyjádřena i jako pojmenovaná třída. Kategorie, které takto vnímá většina uživatelů ("ontologistů") pak označíme jako *ontologistické kategorie*.

[4] Prvním členem konjunkce je vždy fokusová třída, kterou výraz specializuje. Jmenné prostory ontologií používané na webu pro stručnost neuvádíme.

Z četnosti odpovědí uživatelů (na Likertově škále) byla pracovně odvozena empirická pravděpodobnost, že náhodně vybraný konceptový výraz daného typu bude "průměrným uživatelem" chápán jako přepoužitelná kategorie; tato pravděpodobnost může být použita jako váhový koeficient ve výše uvedeném vzorci pro $\widehat{FOCP}$. Nejvyšší hodnoty (cca 0,7) dosáhl vzor $FC \sqcap \exists R.\{i\}$, následován $FC \sqcap \exists R.C$ (cca 0,5); nejnižší, ale stále nezanedbatelná pravděpodobnost (cca 0,3) pak byla zjištěna u vzoru $FC \sqcap \exists R.\top$.

Komplementární analýzou k výše uvedenému ručnímu hodnocení malého vzorku byl dále automatický průzkum výskytu relevantních vzorů nad rozsáhlými kolekcemi (více než 500) ontologií. Tento průzkum prokázal nejvyšší zastoupení vzoru odpovídajícího $FC \sqcap \exists R.\top$, který se ve většině ontologií uplatňuje pro relativně vysoký počet různých fokusových tříd. Další vzory se uplatní jen pro omezený počet fokusových tříd – zřejmě těch, které jsou v dané ontologii skutečně „stěžejní".

4 Závěr

Metoda výpočtu fokusované kategorizační síly představuje zcela nový[5] příspěvek k řešení problému přepoužívání ontologií, přičemž samotný pojem fokusované kategorizace dosud nebyl, přinejmenším v kontextu ontologického inženýrství, explicitně formulován. Navazující výzkum se věnuje mj. návrhu automatické procedury využívající jednoduché techniky analýzy přirozeného jazyka a umožňující na základě jmen prvků složených konceptových výrazů predikovaných jako „ontologistické" automaticky navrhnout jména pro odpovídající pojmenované třídy, např. *PlaceEquippedByAudiovisualEquipment* nebo *SingleDoorFridgeFreezer* (zde by zafungovala heuristika, že pokud vlastnost obsahuje slovo typu „style", „type", „model", npod., není ji třeba do nového názvu zahrnout a stačí název fokusové třídy zleva rozšířit o hodnotu této vlastnosti).

Literatura

1. Baader, F. et al.: The description logic handbook: theory, implementation, and applications. Cambridge University Press New York, NY, USA, 2003.
2. Schaible, J., Gottron, T., Scherp, A.: Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling. In ESWC 2014: 457-472.
3. Stavrakantonakis, I., Fensel, A., Fensel, D.: Linked Open Vocabulary Ranking and Terms Discovery. In: SEMANTICS 2016: 1-8.
4. Svátek, V., Zamazal, O., Vacura, M.: Categorization Power of Ontologies with Respect to Focus Classes. In: EKAW 2016, LNCS 10024, Springer, 2016: 636–650.

[5] Vzhledem k novosti problému i jeho řešení v tomto stručném příspěvku neuvádíme „srovnání s existujícím výzkumem" – dříve řešené projekty s popisovaným souvisí pouze nepřímo.

Annotation:

Focused Categorization Power of Web Ontologies

Ontology reuse is a pre-requisite of effective use of linked data on the semantic web. The selection of an ontology for reuse currently relies on text search and entity popularity metrics. A further criterion may however be the capability of the ontology to express a subcategorization of a given focus class, where the subcategories may include compound concept expressions only implicitly present in the ontology. We propose a method of focused categorization power calculation in which the different compound concept expressions partially contribute via weight coefficients. The weights can be derived, among other, from the assessment by human users – ontologists.

Anotovanie slovníka pre analýzu sentimentu pomocou PSO

Martin Mikula, Kristína Machová

Katedra kybernetiky a umelej inteligencie, TU V Košiciach
Letná 9, 042 00 Košice

{martin.mikula, kristina.machova}@tuke.sk

Abstrakt. Internet v súčasnosti obsahuje veľké množstvo textu, ktorý obsahuje emócie a názory rôznych ľudí. Keďže je veľmi náročné analyzovať ich manuálne, v tomto príspevku sa zameriavame na automatickú analýzu sentimentu použitím slovníkového prístupu. Anotovanie slovníka je často náročný a zdĺhavý proces. Práve preto sa v tejto práci venujeme možnosti nahradenia človeka, ktorý by slovník anotoval, evolučným algoritmom, ktorý by bol použitý na nájdenie optimálnych hodnôt polarity pre slová v slovníku. Vytvorili sme dve verzie slovníka, ktorý bol použitý na analýzu sentimentu. Prvá verzia bola anotovaná človekom a pre anotovanie druhej verzie sme sa rozhodli použiť PSO. Nakoniec sme porovnali výsledky oboch verzií a slovník anotovaný pomocou PSO dosiahol porovnateľné výsledky ako slovník anotovaný človekom.

Kľúčové slová: analýza sentimentu, slovníkový prístup, particle swarm optimization.

1 Úvod

Analýza sentimentu má v dnešnej dobe veľmi významnú úlohu. Pomáha pri rozhodovaní aký produkt si kúpiť, kontrole vlastnej značky, alebo nájdením dobrého filmu. Existuje mnoho prístupov, ktoré je možné použiť na analýzu sentimentu. Niektoré prístupy sú založené na strojovom učení. Tieto prístupy používajú metódy strojového učenia, ako napr. Naivný Bayesov kvalifikátor, metódu Podporných Vektorov, metódu Maximálnej Entropie, či k-Najbližších Susedov [5], aby klasifikovali príspevky do pozitívnej alebo negatívnej triedy. V súčasnosti sú veľmi populárne aj metódy založene na hlbokom učení. V prípade hlbokého učenia, autori používajú Neurónové siete na získanie nových atribútov alebo získanie nových informácií z textov [13, 10]. My sa v tomto príspevku zameriavame na slovníkový prístup, ktorý používa slovník emocionálnych slov na určenie polarity príspevku.

Slovníky je možné rozdeliť do troch skupín, podľa toho, ako boli vytvorené. Manuálne generované slovníky sú kvalitné, keďže boli vytvorené a hodnotené ľuďmi, ale ich generovanie je veľmi časovo náročné. Na druhú stranu, slovníky generované automaticky môžu byť vytvorené veľmi rýchlo, ale ich presnosť a kvalita je často oveľa nižšia ako pri manuálne generovaných slovníkoch. Riešením toho problému sú semi-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 151-156.

automaticky generované slovníky, kedy po automatickom vygenerovaní slovníka človek skontroluje vytvorený slovník a odstráni nepresnosti.

V tomto príspevku sa zameriavame na automatické anotovanie slovníka pomocou Particle swarm optimization (PSO) v slovenskom jazyku. Tento spôsob by mohol zjednodušiť vytváranie slovníkov v nových, resp. menej používaných jazykoch. V takom prípade by stačilo iba preložiť slovník zo svetového jazyka (angličtina) a automaticky anotovať pomocou PSO.

2 Analýza sentimentu pomocou slovníkov a evolučných algoritmov

Slovníky je možné rozdeliť na dve skupiny, podľa toho aké informácie poskytujú pre analýzu sentimentu. Jednoduché slovníky iba delia slová na pozitívne a negatívne [11, 7, 8]. Ak je požadovaná komplexnejšia analýza, je potrebné použiť sofistikovanejšie slovníky. Tie obsahujú aj dodatočné informácie, ako silu polarity [14, 12] alebo skupiny slov a vzťahy medzi nimi [5]. Taktiež môžu obsahovať aj ďalšie slovná potrebné pre analýzu textu ako intenzifikátory a negátory [11, 7, 12]. Pridanie dodatočných informácií o slovách umožňuje komplexnejšiu analýzu spracovaného textu. Pridanie sily polarity umožnilo porovnávať jednotlivé slová v závislosti na zvolenej stupnici (od 0 do 1, alebo od -5 do 5). Taktiež intenzifikácia a negácia umožňujú posúvanie sily polarity pozitívnym alebo negatívnym smerom.

Existuje aj niekoľko prác venujúcich sa použitiu evolučných algoritmov v oblasti klasifikácie textov. Genetické programovanie bolo použité v práci [3] a jeho úlohou bolo nájsť efektívnu schému na váhovanie slov, ktorá by zlepšila presnosť klasifikácie. PSO bolo použité na výber atribútov [10] pre Conditional Random Field, resp. nájdenie najužitočnejších atribútov [1], ktoré najviac prispejú ku zlepšeniu klasifikácie recenzií pomocou metódy Podporných Vektorov.

3 Základná myšlienka PSO

Particle swarm optimization (PSO) je meta-heuristický algoritmus predstavený Kennedym a Eberhartom [2], ktorý napodobňuje chovanie kŕdľa vtákov. V prípade PSO sa potenciálne nazýva jedinec (particle). Skupina jedincov sa v rámci PSO vytvára tzv. populáciu. Každý jedinec sa pohybuje v problémovom priestore. Pritom nasleduje najlepšieho jedinca a uchováva si svoje najlepšie riešenie pre daný problém (*pbest*), ktoré je získané pomocou fitness funkcie. Výberom najlepšieho riešenia spomedzi skupiny jedincov získame lokálne najlepšie riešenie (*lbest*). Globálne najlepšie riešenie (*gbest*) predstavuje najlepšie riešenie z celej populácie. PSO sa skladá z dvoch základných krokov: 1. zmena rýchlosti smerom k *pbest* a g*best* a 2. zmena aktuálnej pozície. Populácia môže byť popísaná ako vektor $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$. Rýchlosť jedinca môžeme popísať ako $V_i = (v_{i1}, v_{i2}, ..., v_{iD})$ a najlepšiu predchádzajúcu pozíciu jedinca ako $P_i = (p_{i1}, p_{i2}, ..., p_{iD})$. Jedinec g reprezentuje najlepšieho jedinca

v populácii a hodnota *w* predstavuje tzv. lenivostný faktor. Zmeny rýchlosti a pozície sa teda vykonávajú na základe rovníc 1, 2:

$$v_{id}^{n+1} = wv_{id}^{n} + c_1 r_1^n (p_{id}^n - x_{id}^n) + c_2 r_2^n (p_{gd}^n - x_{id}^n) \quad (1)$$

$$x_{id}^{n+1} = x_{id}^n + v_{id}^{n+1} \quad (2)$$

kde $d = 1, 2, ..., D$ je rozmer vektora, $i = 1, 2, ..., N$, kde N je počet jedincov, $n = 1, 2, ...$, je počet iterácií. Rovnice tiež obsahujú dve náhodné hodnoty r_1, r_2, ktoré zabezpečujú aby funkcia neupadla do lokálneho optima. Parametre c_1 a c_2 predstavujú faktory ovplyvňujúce pohyb voči *pbest* a g*best*. PSO skončí v momente, keď je dosiahnuté ukončovacie kritérium, alebo sa naplní počet cyklov určených na začiatku.

4 Navrhovaná metóda

V tejto práci sme vytvorili dve verzie slovníka pre analýzu sentimentu v slovenskom jazyku. Slovník bol manuálne preložený z jeho anglickej verzie [7], ktorá obsahovala 6789 slov. Tieto slová sme preložili do slovenčiny a pre každé slovo sme vyhľadali jeho synonymá a antonymá pomocou synonymického slovníka. Finálna verzia slovníka tak obsahuje 598 pozitívnych slov a 772 negatívnych slov. Stupnicu pre určenie sily polarity sme zvolili v rozmedzí od -3 (najviac negatívne slovo) do 3 (najviac pozitívne slovo).

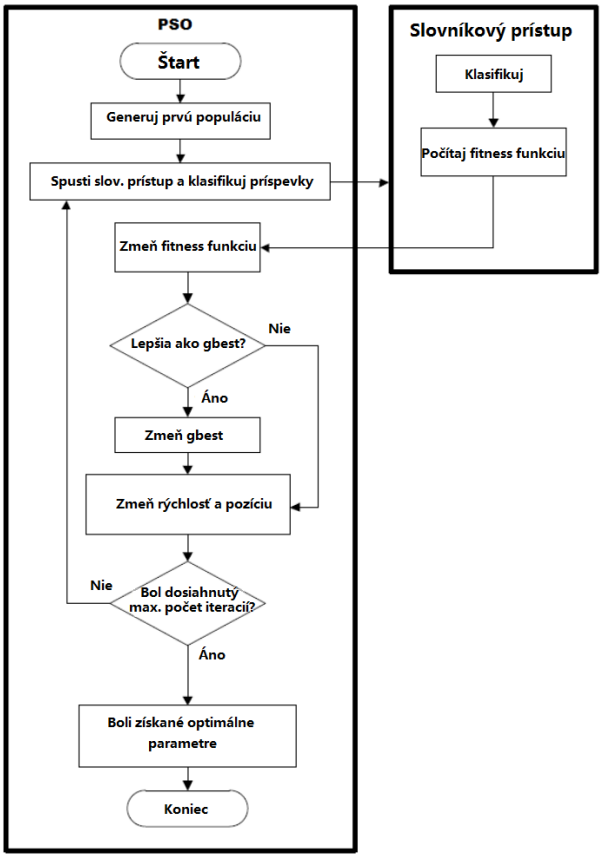
Pre automatickú anotáciu sme sa rozhodli použiť PSO. Ako ukázali viaceré výskumy, jedná sa o efektívny, robustný a jednoduchý optimalizačný algoritmus, ktorý bol aplikovaný na mnoho optimalizačných funkcií. Každý jedinec v našom prípade môže byť zapísaný vo forme vektor $X_i = (x_{i1}, x_{i2}, ..., x_{iD})$ kde $x_{id} \in \{-3,3\}$, $i = 1, 2, ..., N$, kde N je počet jedincov v populácii a $d = 1, 2, ..., D$ označuje počet slov v slovníku. Rozsah od -3 do 3 sme zvolili rovnaký, ako v prípade ľudského anotátora. Celý proces použitia PSO je možné vidieť na obrázku 1.

Prvá generácia bola generovaná náhodne. Ostatné parametre boli získané na základe experimentov a dostupnej literatúry. Tie boli nastavené nasledovne:

- počet jedincov: 15000
- počet opakovaní: 100
- $w = 0.729844$
- $c_1 = 1.49618$
- $c_2 = 1.49618$

Ako fitness funkciu sme použili makro-F1 mieru. Tá sa používa na vyhodnocovanie nevyvážených datasetov. Makro-F1 miera sa vypočíta ako priemer F1 mier pre jednotlivé triedy (pozitívnu a negatívnu). F1 miera je harmonický priemer medzi presnosťou a návratnosťou. Tie boli vyčíslené na základe výsledkov klasifikácie pomocou slovníkového prístupu. Slovníkový prístup skombinoval hodnoty polarity generované pomocou PSO so slovami zo slovníka a vytvoril dočasný slovník, ktorý bol použitý na ohodnotenie príspevkov v testovacom datasete. Na získanie polarity príspevku bol použitý algoritmus, ktorý vyhľadával slová príspevku v slovníku a na základe prira-

denej polarity upravil silu polaritu príspevku. Výsledná polarita bola porovnaná s polaritou uvedenou v datasete. Na základe toho porovnania boli vyčíslené hodnoty presnosti a návratnosti.



Obr. 1. Diagram učenia sily polarity slov pomocou PSO

5 Experimenty a výsledky

Predstavený prístup sme testovali na dvoch datasetoch. Prvý (všeobecný dataset) obsahuje 4720 príspevkov z rôznych oblastí (hodnotenie filmov, kníh a elektroniky, politika, atď.), ktoré boli získané z viacerých webových stránok. V datasete sa nachádza 2455 pozitívnych a 2265 negatívnych príspevkov. Druhý dataset (filmový dataset) bol preložený z angličtiny pomocou Google tranlator[1] a jedná sa o dataset vytvorený v práci Pang a Lee [9]. Tento dataset obsahuje 1000 pozitívnych a 1000 negatívnych

[1] https://translate.google.sk/?hl=sk&tab=wT

filmových hodnotení zo stránky rottentomatos.com. Datasety boli v procese učenia nahodne rozdelené v pomere 90:10, teda PSO hľadalo optimálne riešenie na 90% príspevkov a následne bolo toto riešenie použité na 10% neanalyzovaných dát. Výsledky experimentov je možné vidieť v tabuľke 1.

Tab. 1. Porovnanie F1 mier pre obidve verzie slovníka.

	všeobecný dataset	filmový dataset
verzia anotovaná človekom	**0,7668**	0,6294
verzia anotovaná pomocou PSO	0,7647	**0,7292**

Výsledky ukazujú, že PSO dosiahlo výsledky porovnateľné s človekom, resp. ho dokázalo trochu prekonať. Na všeobecnom datasete dosiahlo PSO výsledok veľmi blízky ľudskému anotátorovi. Na filmovom datasete dokázalo PSO prekonať človeka, čo znamená, že dokázalo nájsť vhodnejšie sily polarity pre jednotlivé slová ako človek.

6 Záver

V tomto príspevku sme predstavili prístup na analýzu sentimentu pomocou slovníkovej metódy, ktorá použila slovník anotovaný pomocou PSO. Tento slovník bol vytvorený prekladom z jeho angličtiny a v prvej verzii bol anotovaný manuálne, čo bolo časovo náročné. Druhá verzia slovníka bola anotovaná pomocou evolučného algoritmu, konkrétne PSO. Ako ukázali výsledky, verzia anotovaná pomocou PSO dosiahla v prípade všeobecného datasetu porovnateľné výsledky ako verzia anotovaná človekom a v prípade filmového datasetu lepšie výsledky ako verzia anotovaná človekom.

Do budúcna by sme chceli vylepšiť PSO použitím ďalších metód úpravy pozície a polohy a taktiež otestovať na väčšom datasete.

Literatúra

1. Basari, S. H., a kol.: Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering. 53, 453--462 (2013).
2. Eberhart, R., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, (1995), pp. 39--43.
3. Escalante, H. J., a kol.: Term-weighting Learning via Genetic Programming for Text Classification. Know.-Based Syst. 83, 176--189 (2015).
4. Esuli, A., Sebastiani, F.: SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06). 417--422 (2006).
5. Go, A.: Twitter Sentiment Classification using Distant Supervision. Stanford University, Available on: http://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf, (2013).

6. Gupta, D. K., a kol.: PSO-ASent: Feature Selection Using Particle Swarm Optimization for Aspect Based Sentiment Analysis. NLDB. 220--233 (2015).
7. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04). ACM, New York 167--177 (2004).
8. Mohammad, S. M., Turney, P. D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29, 436--465 (2013).
9. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA, (2004).
10. dos Santos, C. N., Gattit, M.: Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics. 69--78 (2014).
11. Stone, P., a kol.: The General Inquirer: A Computer Approach to Content Analysis. The MIT Press., (1966).
12. Taboada, M., a kol.: Lexicon-based Methods for Sentiment Analysis. Computational Linguistics. 267--307 (2011).
13. Tang, D., a kol.: Coooolll: A Deep Learning System for Twitter Sentiment Classification. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 208--212 (2014).
14. Warriner, A. B., Kuperman, V., Brysbaert, M.: Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods. 45, 1191--1207 (2013).

Poďakovanie:

Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0213 a Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR, projekt č. 025TUKE-4/2015.

Annotation:

Annotation of dictionaries using PSO

The social web contains a huge amount of text with human emotions and opinions. It is very difficult to analyze them manually and in this paper, we use automatic sentiment analysis based on dictionaries. We created two versions of the dictionary. The first version was prepared manually and annotated by human, which was time-consuming. The second version used Particle Swarm optimization (PSO) for the annotation of words in the dictionaries. Both versions were tested on two datasets, a generated dataset and a movie dataset. We compared the results of the two versions and the version annotated by PSO achieved comparable results with the version annotated by human.

Pokroky v analýze heterogenních neuroinformatických dat

Ondřej Klempíř, Václav Čejka, Jan Tesař, Radim Krupička

Katedra biomedicínské informatiky, Fakulta biomedicínského inženýrství ČVUT v Praze
Nám. Sítná 3105, 272 01 Kladno

{ondrej.klempir, vaclav.cejka, jan.tesar, krupicka}@fbmi.cvut.cz

Abstrakt. V současné době roste počet dat získaných při měření pato/fyziologických procesů mozku. Aplikací analytických metod jsou z dat získávány nové poznatky o mechanizmech neurologických onemocnění. Cílem výzkumu je dosažení pokroku v analýze různých typů neuroinformatických záznamů (jednotkové mikroelektrodové neuronové aktivity, transkraniální magnetické stimulace a blízké infračervené spektroskopie) za účelem objektivní identifikace významných biomarkerů metodami strojového učení. Do studie byla zahrnuta data z Neurologické kliniky 1. LF a VFN UK v Praze. V průběhu řešení výzkumu byly vytvořeny skripty pro ukládání, předzpracování a analýzu 3 typů výše zmíněných neuroinformatických datových zdrojů. Pokroky v analýze budou využity pro hodnocení naměřených dat a testování hypotéz s přínosem pro klinickou neurologickou praxi.

Klíčová slova: mikroeletrodové záznamy, transkraniální magnetická stimulace, blízká infračervená spektroskopie, algoritmus třídění pálení neuronů.

1 Úvod

Světová zdravotnická organizace zaznamenává zvyšující se výskyt neurologických onemocnění. Více než 50 milionů Evropanů trpí neurologickou nemocí, což vyžaduje náklady téměř 400 miliard euro [4]. Současně také roste počet dat získaných při studiu (pato)fyziologických procesů mozku. Aplikací analytických výpočetních metod jsou z dat získávány nové poznatky o funkci nervového systému či např. mechanizmech neurologických onemocnění. Ukazuje se, že počítačové učení dokáže u některých nemocí na základě extrahovaných příznaků a biomarkerů rozlišit zdravou normu od patologie přesněji než expert. Dále např. s úspěchem predikovat odpověď na léčbu.

Ve studii se zaměřujeme na dvě skupiny neurologických onemocnění, dystonie a Parkinsonovu nemoc (PN). Dystonie je heterogenní skupina syndromů, které se projevují křečovitým mimovolním pohybem jedné nebo více částí těla [2]. PN je chronické progresivní onemocnění, přičemž mezi základní symptomy patří třes, ztuhlost a

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 157-161.

pohybové zpomalení [5]. Jednou z metod, která se pro léčbu symptomů pokročilých forem těchto nemocí používá, je hluboká mozková stimulace (Deep Brain Stimulation; DBS), která se implantuje během neurochirurgické operace do specifických jader v mozku [2]. Data, která byla naměřena na Neurologické klinice 1. lékařské fakulty Univerzity Karlovy a Všeobecné fakultní nemocnice v Praze (1. LF UK a VFN), jsou ve světovém kontextu unikátní. A to hlavně z důvodu vyšetření pacientů za různých podmínek, které dříve měřeny nebyly (např. při experimentech s nastavením DBS). Datové soubory byly získány třemi neurovědními vyšetřovacími technikami, které snímají informace na různých úrovních mozku - od aktivity jednotlivých nervových buněk, až po aktivitu makroskopickou. Příspěvek představuje podstatu geneze a zpracování tří typů neuroinformatických datových zdrojů.

2 Analýza dat mikroelektrodových záznamů

Mikroelektrodové encefalografické záznamy (microEEG) se používají pro diagnostiku správného umístění DBS elektrody. Technologicky jde o sadu mikroelektrod (průměr hrotu měří okolo 5 μm), které prochází skrz mozek a nahrávají aktivitu (tzv. pálení, spikes) okolních několika neuronů. Dosažení optimální pozice DBS je klíčové z hlediska terapeutického efektu u dystonií i PN. V současné době je klinický efekt posuzován primárně neurologem, a to vizuálně či akusticky, na základě specifického projevu pálení neuronů. Cílem offline analýzy microEEG v probíhajícím výzkumu je nalezení biomarkeru optimální pozice s využitím metod strojového učení.

Stěžejní úlohou následující po základním předzpracování (explorační analýze, filtraci či detekci artefaktů) je automatický spike sorting (třídění spiků). Neboli, automaticky v microEEG detekovat namodulované projevy neuronů (jednotlivá pálení) v okolí mikroelektrody a roztřídit je s ohledem na individuální charakteristiky včetně extrakce dalších užitečných informací o jednotlivých neuronech. Vyvinutá spike sorting procedura sestává z 5 fází:

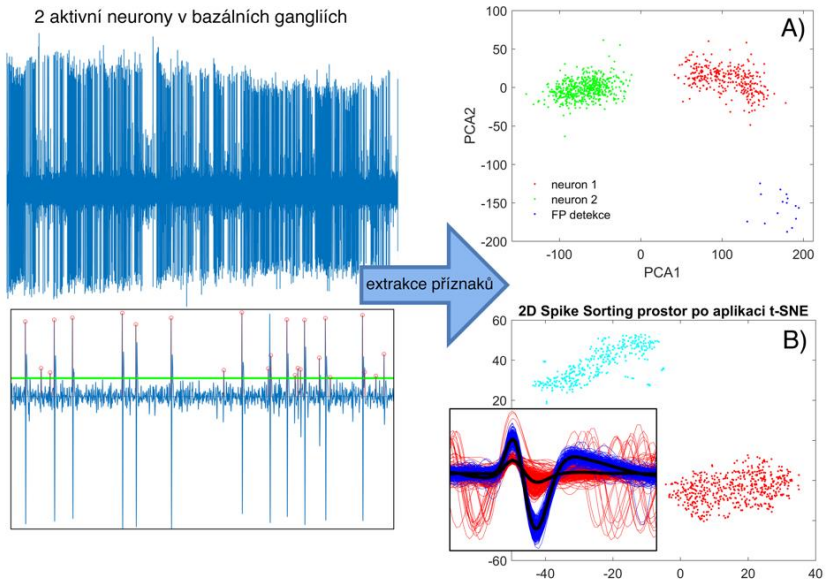
1. stanovení prahu pro odlišení aktivity pálení od šumu pozadí (pravidlo 3σ),
2. výpočet mnohorozměrných tvarově založených příznaků pro detekce z bodu 1,
3. redukce dimensionality pomocí PCA a t-SNE do 2D/3D prostoru,
4. shluková analýza K-means pro rozlišení jednotlivých neuronů,
5. výpočet odvozených charakteristik (např. průměrná frekvence pálení / s).

Metoda t-SNE (t-Distributed Stochastic Neighbor Embedding) zachycuje v porovnání s PCA (Principal Component Analysis) i nelineární strukturu v datech [1]. Na konkrétním experimentu (Obr. 1A) lze vidět, že PCA vede k rozpoznání 3 neuronů, kdy jeden shluk představuje falešnou detekci, která nesouvisí se skutečnou charakteristickou podobou projevu neuronu a musel by být analytikem poloautomaticky vyřazen. T-SNE může fungovat lépe při úplné automatizaci, na Obr. 1B je příklad s chybou 1.6 %. T-SNE přiřadí málo četné hodnoty do bližšího shluku za cenu poměrně malé chyby odhadu výstupních parametrů skutečných neuronů. Ukazuje se také, že t-SNE vede k neproblematickému automatickému určení počtu shluků.

3 Analýza dat z transkraniální magnetické stimulace

Transkraniální magnetické stimulace (TMS) je neinvazivní metoda sloužící k funkčnímu vyšetření centrálního a periferního nervového systému. Je využívána nejenom k rutinním vyšetřením, ale také v oblasti výzkumu funkce jednotlivých mozkových a míšních struktur za účelem identifikace elektrofyziologických biomarkerů poruch řízení hybnosti u dystonií i PD [3].

Pomocí série pulzů aplikovaných do oblasti motorického kortexu byly z registrace odpovědi v horní končetině získány sady motorických evokovaných potenciálů (MEP) (Obr. 2B), které byly průměrovány a filtrovány. Pomocí vlnkové transformace byly v signálech automaticky detekovány MEP a byly vypočteny doby jejich trvání. Jako mateřská vlnka byla použita adaptovaná vlnka zkonstruovaná z průměru motorických potenciálů kontrolních subjektů. Latence potenciálu a jeho trvání byly určeny z pozice maxima (v abs. hodnotě) v matici obsahující koeficienty transformace.

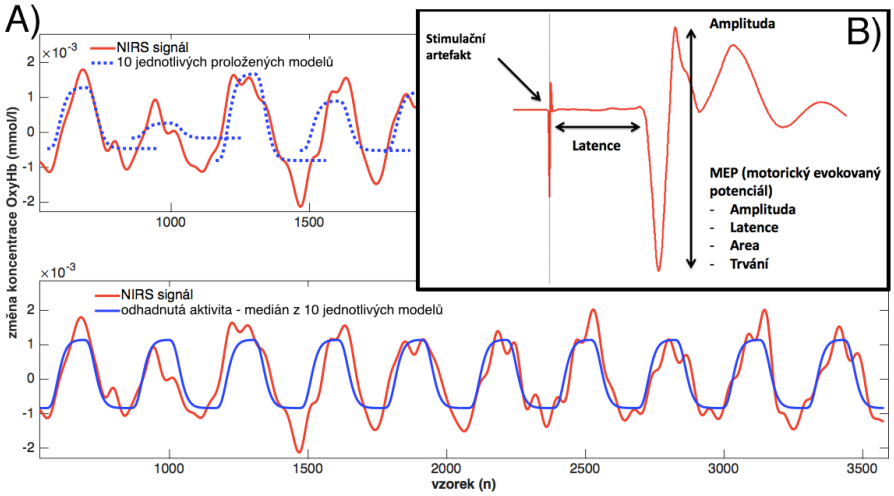


Obr. 1. Metodika detekce jednotlivých neuronů v jednom konkrétním microEEG záznamu bazálních ganglií (modrá časová řada). Vlevo dole - přiblížený úsek jednotlivých pálení po amplitudové detekci (zelený práh). A) 2D příznakový prostor po PCA redukci tvarově definovaných příznaků detekovaných pálení. Aplikace 3-means algoritmu pro nalezení shluků, kdy modrý shluk odpovídá artefaktu. B) 2D příznakový prostor po t-SNE redukci tvarově definovaných příznaků. Artefakty z A) byly přiřazeny do červeného shluku. Na výřezu lze vidět jednotlivé průběhy pálení včetně černé průměrné skupinové hodnoty.

4 Analýza signálů blízké infračervené spektroskopie

Blízká infračervená spektroskopie (Near-InfraRed Spectroscopy; NIRS) je zobrazovací metoda aktivity kůry mozku využívající nízkoenergetické optické záření pro detekci lokálních změn koncentrace oxyhemoglobinu. Charakter geneze je podobný jako u funkční magnetické rezonance. Mozková aktivita byla snímána fotodetektory umístěnými na skalpu při provádění rytmického dotyku palce a ukazováku pravé ruky v průběhu 10 cyklů, během kterých se po 15 sekundách střídala pohybová a klidová fáze úlohy.

Změny koncentrace oxyhemoglobinu byly vypočteny ze surového signálu NIRS pomocí modifikované Lambert-Beerovy rovnice. Pro stanovení hodnot mozkové aktivity vyfiltrovaných NIRS signálů byla využita regresní metoda založená na fyziologickém modelu očekávané odezvy neurální tkáně. Byl uvažován medián z 10 cyklů proložení očekávaným modelem (Obr. 2A). Zjednodušeně, analýza NIRS spočívá v proložení modelu známé odezvy mozku při pohybu ruky na konkrétní naměřený signál. Výsledkem proložení je zisk jednoho charakteristického parametru aktivity signálu, který je vstupem pro navazující statistické testy.



Obr. 2. A) Regresní mediánová metoda (modrý průběh) pro odhad aktivity kůry mozku (červený průběh). Nárůsty v červeném NIRS signálu odpovídají 10 cyklům klidové a pohybové fáze úlohy klepání prsty. B) Typický příklad MEP po automatické detekci vlnkovou transformací.

5 Závěr a budoucí práce

V průběhu řešení výzkumu byly vyvinuty různorodé algoritmy - automatický spike sorting, detekce motorického potenciálu TMS a metoda pro stanovení aktivity NIRS, včetně skriptů pro načítání a předzpracování. Z konkrétních pilotních výstupů navržených metod použitých na reálné signály vyplývá jejich vhodnost: spike sorting procedura s využitím t-SNE vykázala chybu menší než 2 %, automatické detekce MEP dle

experta probíhají správně a mediánový regresní odhad u NIRS vykazuje dobré vlastnosti. Prezentované metody jsou založeny a dále rozšiřují metodologické principy autorů/pracovišť v rámci současného stavu daného oboru (state of the art). Experimentální hodnocení výsledků odhadů jednotlivých metod pro dávkové zpracování mnoha souborů bude předmětem navazující práce. Implementované metody budou využity pro hodnocení pacientů a ověření klinických hypotéz. Cílem článku bylo stručně informovat o probíhajícím výzkumu na FBMI ČVUT a Neurologické klinice 1. LF a VFN UK a přiblížit možnosti zpracování 3 typů neuroinformatických dat.

Literatura

1. Dimitriadis, G.: T-SNE vizualization of large-scale neural recording. BioRxiv (2016) 1 22.
2. Jech, R.: Hluboká mozková stimulace u dystonií. Neurologie pro praxi 14(5) (2013) 232 236.
3. Kobayashi, M.: Transcranial magnetic stimulation in neurology. The Lancet neurology 2(3) (2003) 145 156.
4. Ptáček, R.: Etické problémy medicíny na prahu 21. století. Grada Publishing a.s., Česká republika, 2014.
5. Ulmanová, O.: Parkinsonova nemoc – základy terapie a diferenciální diagnostiky. Psychiatrie pro praxi 2 (2007) 60 62.

Poděkování: Tento článek vznikl díky podpoře projektu SGS17/114/OHK4/1T/17 Zpracování a analýza heterogenních neuroinformatických dat na KBI FBMI ČVUT v Praze.

Annotation:

Advances in analysis of heterogenous neuroinformatics data

The number of the neuroinformatics datasets is currently increasing. Application of the analytical methods is essential to discover new neuroscience knowledge. The aim of this research is to make progress in the analysis of various types of neuroinformatics recordings (single neuronal activity, Transcranial Magnetic Stimulation and Near Infrared Spectroscopy) in order to objectively identify some significant biomarkers by the machine learning methods. This study includes data from the Department of Neurology, 1st Faculty of Medicine and General University Hospital in Prague. The advances in analysis will be used for the evaluation of the measured data and consequently for the hypotheses testing, with expected benefit in clinical neurology.

Predikcia spotových cien elektriny

Róbert Magyar, Viera Rozinajová

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

xmagyarr@is.stuba.sk
viera.rozinajova@stuba.sk

Abstrakt. Každodenné rozhodovanie účastníkov v biznise si vyžaduje informácie o budúcom vývoji určitej premennej dôležitej pre biznis. Dostatočne presná predikcia prináša nové príležitosti modifikácie nákupných a predajných stratégií a lepšiu možnosť kontrolovania rizika. Aplikáciou prístupov objavovania znalostí a umelej inteligencie dokážeme vytvárať zaujímavé prediktívne modely využiteľné v mnohých doménach. Naozajstná výzva sa ale nachádza v zakomponovaní týchto nástrojov do skúmania správania sa predpovedanej spojitej premennej, ktorá je charakteristická krátkodobými skokmi. V tomto článku prezentujeme úvodný návrh riešenia problému krátkodobých skokov v doméne spotových cien elektriny.

Kľúčové slová: predikcia, umelá inteligencia, objavovanie znalostí, spotové ceny elektriny

1 Úvod

Predikovaná spojitá premenná sa často vyznačuje určitými charakteristikami. Tieto charakteristiky môžu zásadne pozitívne alebo negatívne ovplyvniť schopnosť aplikovanej metódy a zvoleného prístupu predikcie dosahovať uspokojivé výsledky.

Cena elektriny sa vyznačuje charakteristikami, ktorých správne pochopenie vytvára príležitosť na tvorbu presnejších predikčných systémov. Medzi charakteristiky vývoja ceny elektriny patrí tendencia návratu ceny k priemeru, sezónnosť, vysoká volatilita a rapídne cenové skoky [3].

Nekontrolované, krátkodobé a rapídne zmeny ceny zapríčiňujú nárast rizika pre nakupujúcich aj predávajúcich. Výskyt krátkodobých cenových skokov je ovplyvnený faktormi ako je teplota, vlhkosť, porucha nízko nákladových elektrárni, obmedzenia prenosu a výroby elektriny [2][4]. Elektrina ako komodita sa neskladuje v signifikantných množstvách, čo je spôsobené nákladmi na skladovanie [3]. Kombinácia týchto faktorov podporuje zvýšenie volatility a krátkodobé cenové skoky, ktoré vytvárajú potrebu čoraz presnejšej predikcie cien a dopytu po elektrickej energii nie len zo strany dodávateľa ale aj ďalších účastníkov trhu, ktorí vstupujú do procesu nákupu a predaja.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 162-165.

Cieľom tohto článku je nielen ukážka prístupu predikcie spojitej premennej, ktorá sa vyznačuje krátkodobými a rapídnymi zmenami vo vývoji ale aj navrhnutie vylepšení v tomto unikátnom prístupe k predikcii. Predpovedanie tohto javu má za cieľ zvýšiť presnosť celkovej predikcie.

2 Predikcia spotových cien a krátkodobých skokov

Od roku 2000, kedy vzrástol záujem o výskum v oblasti predpovede spotových cien elektriny sa začali využívať rôzne modely na tvorbu predikčných nástrojov, medzi ktoré patria multi-agentové systémy, fundamentálne modely, reduced-form modely, štatistické modely a modely založené na umelej inteligencii [1]. V zásade delíme práce na tie, ktoré sa zaoberajú celkovou predikciou spotových cien elektriny [5–7] a tie, ktoré predpovedaním cenových skokov získavajú určitú výhodu pri celkovej predikcii cien [4,8,9].

Zapracovanie predikcie cenových skokov do celkovej predikcie spotových cien elektriny sa ukázalo ako správny krok v dosahovaní cieľov predikčnej úlohy aj keď efekt predikcie krátkodobých skokov na presnosť finálneho riešenia je zatiaľ iba marginálny z dôvodu náročnosti správnej identifikácie cenových skokov [10] [11]. Množina dostupných prístupov identifikácie krátkodobých skokov je dominantná nasledujúcou identifikáciou [4][11]:

$$P_{prah} = \mu + 2\delta \tag{1}$$

Kde P_{prah} reprezentuje prahovú cenu, μ reprezentuje priemer hodnoty ceny (zväčša priemer cien z predchádzajúceho roka), δ je smerodajná odchýlka hodnôt časového radu. Okrem tohto prístupu identifikácie môžeme spozorovať aj identifikáciu krátkodobých skokov pomocou fixného prahu, prahu stanoveného fixnou alebo variabilnou percentuálnou zmenou ceny, identifikáciou skokov pomocou wavelet transformácie a iné [9].

V práci [4] sa krátkodobé skoky identifikovali prístupom (1) a natrénovali tri dopredné neurónové siete (ANN), pričom jedna ANN slúžila ako klasifikátor skokov a zvyšné dva predpovedali hodnoty časového radu. V práci [8] s rovnakou identifikáciou skokov sa autori zamerali aj na tvorbu atribútov ako je existencia skoku 24 hodín pred predpoveďou alebo zachytávanie informácie o počte skokov v určitom intervale.

V našom návrhu sa zameriavame na automatizáciu identifikácie krátkodobých cenových skokov a na využitie sekvenčnej informácie časového radu v hybridnom predikčnom modely. Máme za cieľ transformovať časový rad spôsobom, ktorý by mal napomôcť efektívnejšie extrahovať sekvenčnú informáciu z časového radu, čo by v konečnom dôsledku mohlo pozitívne vplývať na presnosť finálneho predikčného modelu.

3 Návrh

Náš výskum je v počiatočnom štádiu, čo znamená, že úspešnosť navrhovaného prístupu sa bude v blízkej budúcnosti overovať na reálnych dátach.

Hlavnou myšlienkou tohto návrhu je separácia a reprezentácia časového radu spôsobom, ktorý zefektívni extrakciu sekvenčnej informácie z časového radu a využije danú informáciu vo finálnej predikcii s cieľom zvýšiť presnosť finálneho predikčného modelu.

Vhodná separácia úsekov časového radu, vyznačujúcich sa odlišným vývojom, môže vytvárať príležitosť špecializovať predikčné modely na tieto úseky a takto pozitívne ovplyvňovať schopnosť vyhnúť sa nedoučeniu algoritmu v problémových častiach časového radu. V doméne elektriny, vhodnou aplikáciou separácie skokového vývoja ceny od vývoja bez skokov, môžeme pozitívne ovplyvniť presnosť predikčného modelu vo fázach trhu, ktoré sa vyznačujú prudkými krátkodobými skokmi.

Nová reprezentácia časového radu vznikne zgrupovaním podobných hodnôt časového radu do zhlukov. Zaradenie hodnôt časového radu do tried sa prvotne vykoná zhlukovacím algoritmom, kde finálny počet zhlukov určuje počet potrebných predikčných modelov zameraných na predikciu rozdielnych úsekov časového radu. Správnym zaradením cien do zhlukov získame možnosť špecializovať predikčné modely na určité úseky časového radu, čo by malo pozitívne ovplyvniť modelovanie cien v prudkých krátkodobých zmenách predikovanej premennej. Identifikácia skokov vykonaná zgrupovaním cien podľa podobnosti má výhodu oproti identifikačným metódam popísaných v kapitole 2 z dôvodu schopnosti naučenia sa reprezentácie skoku samotným algoritmom, čím sa vytratí potreba manuálne definovať krátkodobé skoky. Výsledkom identifikačnej fázy je transformovaný časový rad popisujúci sekvenciu úsekov pôvodného časového radu.

Transformovaný časový rad sa stáva vstupnou množinou pre natrénovanie rekurentnej neurónovej siete RNN-LSTM (Long Short Term Memory) s konečným cieľom extrahovania a využitia sekvenčnej informácie z novej reprezentácie časového radu. Finálna predikcia sa stanoví kombináciou jednotlivých predikčných modelov, modelov dedikovaných pre predikciu určitých časových úsekov a rekurentnej neurónovej siete, ktorá predpovedá sekvenciu týchto úsekov.

Finálne riešenie má za cieľ dosiahnuť presnejšie výsledky v podobe zmenšenia priemernej absolútnej percentuálnej chyby ako riešenie bez navrhovaných zmien, teda i) bez využitia zhlukovania pri extrakcii sekvenčnej informácie vo finálnej predikcii a ii) bez využitia sekvenčnej informácie vo finálnej predikcii. Prínos návrhu je v automatickej identifikácii krátkodobých skokov, v reprezentácii časového radu pomocou úsekov a v snahe o zefektívnenie extrakcie sekvenčnej informácie v časovom rade.

4 Záver

V tejto práci sme sa zamerali na dôležitosť problému predikcie spotových cien elektriny, poukázali na možný spôsob riešenia krátkodobých skokov a navrhli sme vylepšenia doterajšieho prístupu k predikcii v doméne cien elektriny. Veríme, že navrhova-

né zmeny v predspracovaní a v modelovaní budú viesť k dosiahnutiu presnejších výsledkov. Overenie prínosu týchto návrhov je ďalším krokom v našom výskume.

Literatúra

1. Weron, R. : Electricity price forecasting: A review of the state-of-the-art with a look into the future, Int. J. Forecast. 30 (2014) 1030–1081.
2. European Wind Energy Association : Wind Energy and Electricity Prices Exploring the "merit order effect" (n.d.). http://www.ewea.org/fileadmin/files/library/publications/reports/MeritOrder.pdf (cit 16. máj 2017).
3. Weron, R. : Modeling and forecasting electricity loads and prices: A statistical approach, Chichester, 2006.
4. Sandhu, H.S.,Fang, L., Guan, L. : Forecasting day-ahead price spikes for the Ontario electricity market, Electr. Power Syst. Res. 141 (2016) 450–459.
5. Lahouar, A., Slama, J.B.H. : Comparative study of learning machine predictors for half-hour and day-ahead electricity price forecast in deregulated markets, v: 2016 7th Int. Renew. Energy Congr., IEEE, 2016: s. 1–6.
6. Shiri, A., Afshar, M., Rahimi-Kian, A., Maham, B. : Electricity price forecasting using Support Vector Machines by considering oil and natural gas price impacts, v: 2015 IEEE Int. Conf. Smart Energy Grid Eng., IEEE, 2015: s. 1–5.
7. Dudek, G. : Multilayer perceptron for GEFCom2014 probabilistic electricity price forecasting, Int. J. Forecast. 32 (2016) 1057–1060.
8. Amjady, N., Keynia, F. : A new prediction strategy for price spike forecasting of day-ahead electricity markets, Appl. Soft Comput. J. 11 (2011) 4246–4256.
9. Janczura, J., Truck, S., Weron, R., Wolff, R.C. : Identifying spikes and seasonal components in electricity spot price data: A guide to robust modeling, Energy Econ. 38 (2013) 96–110.
10. Christensen, T.M., Hurn, A.S., Lindsay, K.A. : Forecasting spikes in electricity prices, Int. J. Forecast. 28 (2012) 400–411.
11. Zhao, J.H., Dong, Z.Y., Li, X., Wong, K.P : A framework for electricity price spike analysis with advanced data mining methods, IEEE Trans. Power Syst. 22 (2007) 376–385.
12. Paralič, J. : Objavovanie znalostí v databázach, Košice, 2003.

Annotation:

Prediction of electricity spot prices

In this paper, prediction of electricity spot prices incorporating different way of preprocessing data and prediction of price spikes is briefly described. Proposed enhancements of classic prediction process in electricity domain should improve overall accuracy of prediction. Goal was to separate different phases of development of electricity prices by creating models focusing only on parts of price movement. Phases are defined by probability of price spike occurrences. Proposed improvements of classic prediction method in electricity domain are going to be tested separately in future work.

Exploračná analýza medicínskych záznamov

František Babič, Michal Vadovský, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita v Košiciach, Letná 9/B, 042 00, Košice, Slovenská republika

{frantisek.babic, michal.vadovsky, jan.paralic}@tuke.sk

Abstrakt. Medicínska diagnostika predstavuje komplexný proces pozostávajúci z množstva vstupov a potenciálnych závislostí, ktoré môžu v konečnom dôsledku ovplyvniť správnosť výsledku a následnú liečbu. Dátová analytika a príbuzné domény ako štatistika alebo umelá inteligencia môžu byť v tomto smere nápomocné, ak sú k dispozícii medicínske záznamy v elektronickej podobe. V našom prípade sme sa zamerali na jedno z vážnych civilizačných ochorení s názvom Metabolický syndróm a jeho výskyt u hypertenzívnych žien v menopauze. V rámci tejto úlohy sme už realizovali viacero predikčných experimentov, ale v tomto článku popíšeme práve analýzu potenciálne existujúcej závislosti medzi cieľovou diagnózou a skupinami vstupných faktorov určenými spolupracujúcim medicínskym expertom. Na tento účel sme použili logistickú regresiu a dosiahnuté výsledky sme overili prostredníctvom experta a existujúcich prác.

Kľúčové slová: medicínske zoznamy, logistická regresia, interval spoľahlivosti

1 Úvod

Podpora medicínskej diagnostiky predstavuje jednu z hlavných výziev, ktorá momentálne stojí pred oblasťou dátovej analytiky. K dispozícii máme metódy strojového učenia, umelej inteligencie, štatistiky alebo exploračnej analýzy, ktoré je možné aplikovať na medicínske záznamy dostupné v elektronickej podobe.

Metabolický syndróm (MS) sa definuje ako nenáhodný spoločný výskyt porúch metabolizmu cukrov súvisiacich s inzulínovou rezistenciou, centrálnou obezitou, dyslipidémiou spojenou so zvýšením hladiny triacylglycerolov a znížením lipoproteínov s vyššou denzitou, artériovou hypertenziou a ďalších faktorov, ktoré sa podieľajú na zvýšenom riziku ischemickej choroby srdca a cukrovky 2. typu [2]. O dôležitosti tejto problematiky svedčí aj fakt, že diagnostické kritériá pre MS spĺňa 20 % slovenskej populácie, čo potvrdila nedávna multicentrická skríningová štúdia s názvom „Prevalencia diabetes mellitus a metabolického syndrómu na Slovensku“ [3].

V našich výskumných aktivitách sa venujeme analýze dostupných medicínskych vzoriek, či už s cieľom vytvoriť čo najpresnejšie predikčné modely alebo pochopiť vstupné vzorky prostredníctvom metód exploračnej analýzy alebo štatistiky. V článku na minuloročných Dáta a Znalosti 2016 sme sa venovali identifikácii kľúčových fak-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 166-170.

torov pre diagnostiku ochorenia s názvom Mierne kognitívne zhoršenie prostredníctvom vybraných štatistických testov a identifikácii kľúčových hodnôt pomocou ROC krivky a Youdenovej metódy [1]. V tomto článku popisujeme podobne orientovanú analýzu, ale zamerali sme sa na MS; konkrétne sme na skúmanie potenciálnych vzťahov medzi vstupnými faktormi a cieľovou diagnózou použili logistickú regresiu. Realizované experimenty nadväzujú na už našu ďalšiu publikovanú prácu [6].

2 Pochopenie dát

Hypertenzívne ženy v menopauze predstavujú rizikovú skupinu pre vývoj MS alebo kardiovaskulárnych ochorení [4]. Vzorka dát obsahuje informácie o 200 pacientkach z klinickej praxe vo veku od 47 do 59 rokov, 69 zdravých a 131 s potvrdenou diagnózou MS na základe kritérií medzinárodnej federácie diabetikov [5]:

- Základné kritérium – abdominálna obezita (obvod pásu nad 80cm) a k tomu aspoň dve zo 4 ďalších kritérií:
  - Glykémia nalačno nad 5.6 mmol/l alebo predtým diagnostikovaný diabetes mellitus 2. typu.
  - Triacylglycerol > 1.7 mmol/l
  - HDL cholesterol < 1.3 mmol/l
  - Zvýšené hodnoty krvného tlaku > 130/85.

Každý pacient je zároveň charakterizovaný hodnotami 62 faktorov, ktoré predstavujú potenciálne dôležité vstupy pre diagnostiku MS. Vybrané atribúty sú popísané v Tab.1.

Tab. 1. Vybrané vstupné atribúty

Názov	Popis
Kedy bola diagnostikovaná hypertenzia (v rokoch)	{<5, <5,10>, >10}
Regulácia hypertenzie	{áno, nie}
Menopauza a jej trvanie (v rokoch)	{nie, <1, <1,3>, >3}
Diagnostika cukrovky	{áno, nie}
Komplikácie pri liečení cukrovky	{nie, mac, mic}
Liečba cukrovky (orálne antidiabetiká, orálne antidiabetiká + inzulín, inzulín)	{nie, alebo, a/alebo, a}
Trvanie cukrovky (v rokoch)	{<5, <5,10>, >10}
Užívanie statínov viac ako 3 mesiace	{áno, nie}
Užívanie beta-blokátorov viac ako 3 mesiace	{áno, nie}
Užívanie antikoagulačných liekov viac ako 3 mesiace	{áno, nie}
Užívanie analgetík viac ako 3 mesiace a v minulom roku	{áno, nie}
Užívanie antibiotík viac ako dvakrát za rok	{áno, nie}
Užívanie lieku Metformin na liečbu cukrovky 2.typu	{áno, nie}

Užívanie ACE-inhibítorov viac ako 3 mesiace	{áno, nie}
Metabolický syndróm	<0, 1>

3 Analýza dát

Ako sme už spomenuli vyššie, na analýzu sme použili logistickú regresiu. Vykonali sme niekoľko experimentov s dvoma vybranými množinami dát. V prvom prípade sme brali do úvahy atribúty reprezentujúce diagnostiku a následné liečenie hypertenzie a cukrovky. V druhom prípade sme analyzovali vplyv anamnézy pacientiek a paralelného výskytu viacerých ochorení, čo dokumentuje užívanie rôznych typov liekov.

Dosiahnuté výsledky sú popísané pomocou parametrov ako z-hodnota, pomer šancí (OR) a interval spoľahlivosti (CI 95%). OR vyjadruje pomer šance zaradenia objektu do 1. cieľovej skupiny ak sa hodnota danej premennej zvýši o 1, pričom hodnoty ostatných premenných v modeli zostanú nezmenené, k pôvodnej šanci jej zaradenia do tejto cieľovej skupiny. V prípade binárnych premenných je interpretácia jednoduchšia, t.j. OR predstavuje pomer šance zaradenia objektu do 1. cieľovej skupiny, ak hodnota premennej = 1 ku šanci jeho zaradenia, ak hodnota premennej = 0, pri rovnakých hodnotách ostatných premenných.

diagnóza/ premenná	áno	nie
áno	a	b
Nie	c	d

$$OR = \frac{(a * d)}{(b * c)} \tag{1}$$

Hodnota OR v intervale <0, 1> znamená menšiu šancu zaradenia objektu do 1. cieľovej skupiny, hodnota >1 znamená opačnú situáciu a OR =1 definuje rovnakú šancu zaradenia do prvej alebo druhej cieľovej skupiny, t.j. nezávislosť.

Takisto sme použili McFaddenov koeficient determinácie, ktorý slúži na určenie kvality modelu logistickej regresie. Model s najvyšším R^2 je podľa tohto kritériá tým najlepším [7].

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2}$$

V prvom prípade najlepší model dosiahol hodnotu koeficientu determinácie 0.28, čo znamená že vytvorený model má určitú predikčnú silu smerom k cieľovej diagnóze, ale hodnota je bližšia skôr k 0 ako k 1. Hodnota 0 znamená, že daný model logistickej regresie nemá predikčnú silu. Na úrovni spoľahlivosti 95% môžeme identifikovať ako významný parameter len *trvanie menopauzy <1,3>* (z-hodnota = 2.18, OR = 11.2, CI = <1.2, 69.2>). Ak by sme hranicu posunuli na 90%, tak by pribudli parametre *trvanie menopauzy >3* (1.73, 6.9, <1.1, 44.5>) a *regulácia hypertenzie = normálna* (-1.9, 0.3, <0.1, 0.8>).

V druhom prípade najlepší model dosiahol hodnotu R^2 0.41, t.j. má väčšiu predikčnú silu ako predchádzajúci. Ako významné parametre na úrovni spoľahlivosti sme identifikovali užívanie *statínu* (5.9, 15.8, <7.3, 34.2>), *Metforminu* (3.5, 44.5, <7.6, 259.5>) a *beta-blokátorov* (2.04, 2.7, <1.2, 5.9>).

4 Záver

V našej práci sme sa venovali analýze dostupnej vzorky dát pomocou logistickej regresie, ktorej výsledok predstavuje zoznam významných vstupných premenných pre cieľovú diagnostiku Metabolického syndrómu. Dosiahnuté výsledky v oboch skupinách experimentov sme konzultovali s expertom a overovali prostredníctvom existujúcich prác a štúdií. Tieto potvrdili vyššiu prevalenciu MS u žien po menopauze, ale ju samotnú nepovažovali za rizikový faktor. Skôr sa zamerali na kardiovaskulárne faktory, ktorých včasná diagnostika môže prispieť k správnej liečbe a prevencii.

Literatúra

1. Vadovský M,, Babič, F., Muchová, M.: Systém na podporu rozhodovania pomocou jednoduchého a efektívneho pochopenie medicínskych záznamov. In: WIKT and DaZ 2016, Bratislava: STU (2016) 89 93.
2. Eckel, R.A., Grundy, S.M., Zimmet, P.Z.: The metabolic syndrome. Lancet 365 (2005) 1415 1428.
3. Galajda, P.: Metabolický syndróm, kardiovaskulárne a metabolické riziká. Via practica 4 (2007) 5 9.
4. Carr M. C..: The emergency of the metabolic syndrome with menopause. The Journal of Clinical Endocrinology & Metabolism 88 (2003) 2404 2011.
5. Alberti K. G., Zimmet P., Shaw J.: Metabolic syndrome – a new world-wide definition. A Consensus Statement from the IDF. Diabet Med 23 9(2006) 469 480.
6. Babič F., Majnarić L., Lukáčová A., Paralič J., Holzinger A.: On Patient's Characteristics Extraction for Metabolic Syndrome Diagnosis: Predictive Modelling Based on Machine Learning. In: Information Technology in Bio-and Medical Informatics, Springer, LNCS 8649 (2014) 118 132.
7. McFadden D.: Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (ed.), Frontiers in Econometrics, New York: Academic Press (1974).

Poďakovanie: Táto publikácia vznikla vďaka čiastočnej podpore projektu č.1/0493/16 financovaného Vedeckou grantovou agentúrou MŠVVaŠ SR a SAV (VEGA), projektom č. 025TUKE-4/2015 a č. 005TUKE-4/2017 financovaných Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR (KEGA).

Annotation:

The medical diagnostic is a complex process consisting of inputs and potential dependencies affecting the correctness of the outcome and subsequent treatment. Data analytics and related domains, such as statistics or artificial intelligence, can be ap-

plied to electronic medical records supporting the doctor's decision process. We focused on one of the major civilization diseases called Metabolic Syndrome and its occurrence in a specific target group. We oriented our experiments to investigate potentially existing dependence between the target diagnosis and the input factor groups determined by the cooperating medical expert. We used logistic regression and the results were evaluated by the expert as interesting and usable in clinical practice.

Doktorandské symposium

Procedurální znalosti expertů a model GLIF

Ondřej Říha

Katedra informačního inženýrství, PEF ČZU v Praze
Kamýcká 129, 165 21 Praha 6 - Suchdol

ondrej.riha.praha@gmail.com

Abstrakt. Text příspěvku má za cíl seznámit čtenáře s problematikou expertních procedurálních znalostí. Tyto znalosti mohou být zaznamenány pomocí modelu GLIF, kde při práci s reálnými záznamy vznikají specifické problémy. Řešení problémů jako posouzení času v oborových doporučeních a porovnávání reálných záznamů s doporučeními v podobě nedeterministického modelu GLIF může znamenat zlepšení v procesech využívajících expertní znalosti. Komplikovanost porovnávání s modelem GLIF nevychází pouze z jeho nedeterministické formy. V praxi často nejsou dostupné kompletní záznamy provedených postupů, případně mohou být některé informace ztraceny při převodu na strukturovanou formu záznamu. Text představuje některé z možností řešení uvedených problémů jako je práce s časem v modelu GLIF a porovnávání nekompletních záznamů pomocí nahrazení chybějících hodnot případně kontrolu všech relevantních možností dle dostupných dat.

Klíčová slova: GLIF, oborová doporučení, EHR, nekompletní záznam

1 Úvod

Díky moderním komunikačním prostředkům jsou dnes informace v obrovské míře dostupné široké veřejnosti. Bez ohledu na kvalitu dostupných informací s nimi umíme efektivně pracovat. Mnohem složitější je to v případě záznamu znalosti. Experti při řízení složitých systémů používají tzv. procedurální znalosti, tedy znalosti často velmi komplexní, které reagují na průběžné změny a chování systému. Pro takové znalosti je obtížné definovat matematický model chování. Procedurální znalosti se používají například v humánní či veterinární medicíně, zemědělství, biotechnologii a dalších zejména pro diagnostiku a řešení určitých situací.

2 Metody

K formalizaci znalostí je dostupných mnoho prostředků od predikátového počtu, přes sémantické sítě, produkční systémy až k rámcům a mnoha dalším způsobům. Všechny tyto způsoby mohou být složitě aplikovatelné a nesnadné k porozumění expertům v jiných oblastech než jsou počítačové vědy. Procedurální znalosti lze formalizovat

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 172-175.

prostředkem, který je zápisem, respektive zobrazením bližší široké veřejnosti. Takovým prostředkem může být model GLIF (GuideLine Interchange Format). Tento model je silným formalizačním prostředkem pro jeho přehlednost a snadnou pochopitelnost [1].

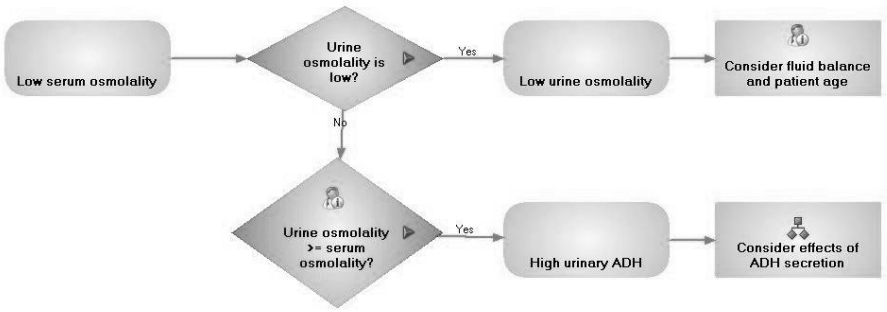
Cílem práce je nalézt vhodnou metodu zavedení časového faktoru do modelu GLIF a navrhnout postup srovnávání nekompletních záznamů o postupu expertů s nedeterministickým modelem GLIF.

2.1 Model GLIF

Model GLIF byl vyvinut ve spolupráci InterMed (Portland), Kolumbijské univerzity, Harvardské univerzity (Brigham and Women's Hospital a Massachusetts General Hospital) a Stanfordské univerzity [2]. GLIF model je procesně a objektově zaměřený a je representován, jako orientovaný graf chronologicky udávající posloupnost úkonů. Samotné doporučení lze vytvářet ve specializovaných editorech nebo psát rovnou ve formě XML souboru (Extensible Markup Language), který je univerzální formát pro přenos informací. Současná verze modelu GLIF 3.5 používá OCL (Object Constraint Language) ve formě jazyka GELLO (‘‘guideline expression language, object-oriented’’). Model GLIF podporuje UMLS (Unified Medical Language System) [4].

GLIF model mapuje informace do tří vrstev abstrakce [3]

- jádro modelu GLIF
- RIM (Reference Information Model)
- lékařské znalosti



Obr. 1. Sublevel postupu léčby hyponatrémie [5]

Tento graf je skládán z pěti různých prvků, některé je možné vidět na obrázku.

Prvek ROZHODNUTÍ umožňuje několik subvariant [4].

- Utility Choice
- Weighted Choice
- Rule In Choice

Přes zdánlivou podobnost modelu s vývojovým diagramem je zde několik podstatných rozdílů a nejvýznamnější je ten, že model GLIF nemusí být deterministický.

3 Výsledky

Autor se ve své práci zabývá posouzením použití faktoru času v modelu, který je při záznamu některých procedurálních znalostí kritický a model GLIF nebyl přímo vybaven pro posouzení času v modelu jiným prostředkem, než prvkem *Rozhodnutí*, který může být v záznamech komplexních postupů značnou nevýhodou pro svoji schopnost zhoršit přehlednost modelu. Řešením může být rozšíření specifikace modelu GLIF o časový parametr či pod jednotnou ontologií definovat časové platnosti jednotlivých stavů systému a ty následně opakovaně používat bez nutnosti je v postupu pokaždé znovu určovat.

Dále je řešena otázka srovnávání existujících záznamů postupů expertů s modelem, který představuje doporučení v konkrétní oblasti užití. Zde je největší výzvou porovnávání záznamů s nedeterministickým modelem GLIF a porovnávání nekompletních záznamů.

Pro srovnání oborových doporučení s nekompletním záznamem autor zvážil užití tří metod.

- Empty metoda
- Imputace hodnot
- Komplexní varianta

Empty varianta zahrnuje triviální doplnění chybějící hodnoty prázdným znakem.

Pro imputaci hodnot bylo zváženo několik metod a zdrojů dat pro odhad chybějící hodnoty. Stejně tak bylo posuzováno, jaké hodnoty je možné doplnit a u jakých je metoda imputace nepřípustná. Výsledkem aplikace metody imputace bylo také zpřesnění dělení prvků v modelu GLIF.

Komplexní metoda představuje úplné prohledání všech dostupných variant průchodu modelem vzhledem k dostupnosti záznamu a posouzení zda mezi několika finálními stavy je alespoň jeden, který odpovídá doporučenému postupu.

4 Diskuze a závěr

Po zvážení několika možností zavedení času do modelu GLIF, bylo na dostupných oborových doporučeních zjištěno, že vhodnost rozšíření specifikace modelu je vhodná až od určité složitosti modelu.

Pro porovnávání záznamů a doporučení je nutná unifikace strukturálních záznamů a oborových doporučení pod jednotnou ontologii. Právě použití jednotné ontologie umožňuje řešit problematiku času v modelu definicí časových platností jednotlivých parametrů nutných pro rozhodovací kroky modelu.

Empty metoda a imputace chybějících hodnot se prokázaly jako metody s velice omezenou možností použití a s výsledkem mající pouze pravděpodobnostní charakter.

Komplexní metoda je schopna zjistit s naprostou jistotou některé chyby v postupu expertů i v případě nekompletních záznamů. Při nekompletním záznamu ovšem nelze určit, zda postup byl zcela správný, ale pouze nalézt záznam, který neodpovídá doporučenému postupu.

Tato metoda může být po dalším zpřesnění využita pro hledání chyb například v postupu léčby pacientů, kde je vhodnost nalezení závažných pochybení obzvláště vhodná.

Literatura

Reference na články v časopisech:

1. Peleg M, Tu S, Bury J, Ciccarese P, Fox J, Greenes RA, et al. Comparing computer-interpretable guideline models: a case-study approach. Journal of the American Medical Informatics Association., roč. 10, č. 1, s. 52-68, led. 2003.
2. L. Ohno-Machado et al., „The GuideLine Interchange Format: A Model for Representing Guidelines", J. Am. Med. Inform. Assoc. JAMIA, roč. 5, č. 4, s. 357, srp. 1998.
3. A. A. Boxwala et al., „GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines", J. Biomed. Inform., roč. 37, č. 3, s. 147–161, čer. 2004.

Reference na technickou specifikaci:

4. M. Peleg, A. Boxwala, S. Tu, D. Wang, O. Ogunyemi, a Q. Zeng, „Guideline Interchange Format 3.5 Technical Specification". InterMed Collaboratory, 04-kvě-2004.
5. Ch. Appleton, QML Pathology document Investigation of Hyponatraemia, Medical-Objects GLIF Editor, pros. 2016

Annotation:

Experts procedural knowledge and model GLIF

The text aims to acquaint the reader with the issue of expert procedural knowledge. This knowledge can be recorded by using of the model GLIF, where when working with real records arise specific problems. Address issues such as the assessment of time in the industry recommendations, and comparison of the real records with the recommendations in the form of non-deterministic of the model GLIF may be an improvement in the processes using expert knowledge. In practice, often not available complete records of the executed procedures, where appropriate, may be some information lost when converting to the structured form of the record. The text presents some of the options to address these problems such as work with time in the model GLIF and comparing the incomplete records using the replace missing values or control of all relevant possibilities according to the available data.

Personalizované odporúčanie využívajúce vizuálne stimuly

Peter Gašpar, Michal Kompan, Mária Bieliková

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

{peter_gaspar, michal.kompan, maria.bielikova}@stuba.sk

Abstrakt. Analýza správania používateľa na Webe je v súčasnosti otvoreným výskumným problémom. Používatelia interagujú s webovými stránkami rôznymi spôsobmi a snahou výskumníkov je tieto interakcie správne interpretovať a použiť pre rôzne úlohy. Interakcie používateľov sú dôležitým vstupom pre odporúčacie systémy, ktoré sa využívajú ako jedna z metód pre zlepšenie používateľského zážitku na Webe. Obsahové odporúčanie využíva charakteristiky položiek pre generovanie personalizovaného odporúčania. Obrázky sa stali populárnym zdrojom informácií v mnohých doménach (napr. nakupovanie), kde ovplyvňujú to, ako sa používateľ rozhoduje. V našej práci využívame vizuálne črty extrahované z obrázkov. Naším cieľom je preskúmať zapojenie obrázkov a vizuálnych stimulov do procesu odporúčania so zámerom skvalitniť existujúce prístupy založené na obsahu.

Kľúčové slová: obsahové odporúčanie, analýza správania používateľa, extrakcia vizuálnych stimulov

1 Úvod

Štúdium správania používateľa na Webe sa v posledným rokoch stalo významným výskumným smerom. Mnohí výskumníci sa snažia analyzovať, ako ľudia interagujú s obsahom, aké sú ich preferencie a čo do najväčšej miery ovplyvňuje ich správanie počas rôznych úloh, ktoré vykonávajú na Webe. Poznanie používateľa a poznanie jeho správania otvára cestu pre skvalitňovanie prístupov na adaptáciu a personalizáciu v rôznych doménach.

Odporúčacie systémy vznikli ako jeden z nástrojov pre zlepšenie použiteľnosti. Ich hlavným cieľom je navrhovať používateľovi potenciálne zaujímavé a atraktívne položky. Ich hlavné využitie nachádzame najmä v procese rozhodovania človeka, a to najmä v prípade, keď je negatívne zasiahnutý problémom informačného preťaženia. S týmto problémom sa stretávame napríklad pri výbere filmu, pri nakupovaní v e-obchode alebo pri výbere hudby [3].

Obsahové odporúčanie využíva ako hlavný zdroj informácie extrahované z textu. Jedným z dôvodom je fakt, že text sa stal hlavným prostriedkom na šírenie informácií, a tak môže ovplyvňovať používateľov pri výbere položiek na Webe. Mnohé štúdie

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 176-180.

však ukázali, že aj obrázky môžu mať vplyv na rozhodovanie a môžu byť užitočné v procese odporúčania.

V tomto príspevku prezentujeme koncept dizertačnej práce. Vychádzame z myšlienky, že obrázky vplývajú na správanie používateľa na Webe, a preto ich zohľadnenie môže pomôcť pri návrhu systémov na odporúčanie obsahu. Naša prvotná analýza sa bližšie zamerala na obsahové odporúčanie obohatené o vizuálne črty. Naše hypotézy sme sa snažili podporiť experimentom, v ktorom sme porovnali dva rôzne prístupy k extrahovaniu vizuálnych čŕt z obrázkov.

2 Obsahové odporúčanie

Taxonómia personalizovaných odporúčacích systémov rozlišuje dva základné prístupy k odporúčaniu: kolaboratívne filtrovanie a obsahové odporúčanie. Kým *kolaboratívne filtrovanie* analyzuje podobnosť medzi správaním viacerých používateľov, *obsahové odporúčanie* skúma aktivitu každého používateľa samostatne a snaží sa mu odporúčať položky na základe tých, ktoré si obľúbil v minulosti. Pri výpočte podobnosti medzi položkami sa využívajú charakteristiky jednotlivých položiek – obsah [3]. Obsah vychádza z vlastností položiek, pričom tieto vlastnosti majú najčastejšie formát textu a závisia od domény, v ktorej sa pohybujeme – napr. v prípade filmov sú to žánre, herci či postavy.

S obsahovým odporúčaním sa najčastejšie stretávame práve v tých doménach, kde text tvorí jadro obsahu, napr. v doméne správ a filmov. Existujú však domény, v ktorých nemusia byť textové opisy postačujúce. Príkladom sú filmy (filmové plagáty), nakupovanie alebo umenie. V týchto prípadoch môže ovplyvňovať rozhodovanie aj výzor položiek.

V práci [8] boli obrázky použité na odporúčanie tovaru. Autori modelovali preferencie používateľa na základe čŕt, ktoré bolo možné identifikovať aj v obrázkoch (napr. farba tovaru). Zhou et al. [9] študovali odporúčač oblečenia, ktorý pre vstupný obrázok (reprezentujúci oblečenie) vedel vygenerovať podobné obrázky, pričom sa zohľadňovali viaceré črty (napr. tvar, farba). Zaujímavým problémom je však aj porovnanie obsahových a vizuálnych čŕt a ich významu pri odporúčaní.

2.1 Akvizícia vizuálnych čŕt

Pri získavaní vizuálnych čŕt z obrázkov môžeme rozlišovať tri základné prístupy: obrázkové deskriptory, farebné charakteristiky a sémantické črty získané z neurónových sietí.

Obrázkové deskriptory sa snažia extrahovať kľúčové body v obrázkoch, čím vznikajú spravidla binárne vektory aplikovateľné na rôzne úlohy, a to obzvlášť výpočet podobnosti obrázkov či rozpoznávanie objektov na obrázkoch. Medzi najznámejšie deskriptory patria SIFT, SURF a BRIEF.

Špecifickým druhom obrázkových deskriptorov sú *farebné charakteristiky,* pri ktorých pracujeme na úrovni jednotlivých pixelov obrázka a analyzujeme jeho farebné zložky. Výstupom tejto analýzy sú črty, ktoré dokážu charakterizovať vizuál

obrázka na základe farieb. Literatúra pritom najčastejšie pracuje s farebným modelom HSV (Hue-Saturation-Value), ktorý považujeme za blízky ľudskému vnímaniu farieb [6]. Tu môžeme skúmať nielen farbu, ale aj ďalšie charakteristiky, akými sú príjemnosť, dominancia a nabudenie vychádzajúce zo psychologických modelov.

Na opis obsahu obrázku využívame *sémantické črty*, pričom v tejto oblasti sa najčastejšie stretávame s konvolučnými neurónovými sieťami. V posledných rokoch vznikli viaceré prístupy a architektúry (napr. Inception-v3 [5] a VGG19 [4]), ktoré umožnili získať črty vhodné pre úlohu porovnania podobnosti obrázkov.

3 Porovnanie obsahových a vizuálnych čŕt v doméne filmov

V našom prvom experimente sme sa zamerali na analýzu filmových žánrov s využitím dátovej množiny MovieLens, ktorá obsahuje filmy a k nim prislúchajúce žánre [1]. Naším cieľom bolo porovnať vizuálne črty s obsahovými a zároveň porovnať dva druhy vizuálnych čŕt. Pre každý film sme pomocou The Movie Database API zvolili hlavný obrázok (plagát filmu), z ktorého sme extrahovali farebné charakteristiky a sémantické črty. Jednotlivé plagáty sme následne usporiadali do skupín podľa toho, do akého žánru prislúchali filmy. V rámci týchto skupín sme následne zisťovali priemernú kosínusovú vzdialenosť a varianciu medzi všetkými vektormi čŕt. Keďže niektorým filmom boli priradené viaceré žánre, umiestnili sme ich aj do viacerých skupín (napr. film majúci žánre komédia a romantický sa nachádzal aj v skupine filmov so žánrom komédia, ale taktiež aj v skupine filmov so žánrom romantický).

Na extrakciu farebných charakteristík sme využili farebný model HSV, pričom sme analyzovali jas, sýtosť a odtieň farieb, ale taktiež aj príjemnosť, dominanciu a nabudenie.

Pre získanie sémantických čŕt sme využili konvolučnú neurónovú sieť Inception-v3 [5] natrénovanú na dátovej množine ImageNet[1]. Inception-v3 bola natrénovaná na úlohu predikcie kľúčových slov opisujúcich obsah obrázka.

Vychádzali sme pritom z práce [2], kde autori využili predposlednú vrstvu tejto natrénovanej siete. Pomocou siete sa pokúsili predikovať kľúčové slová z inej domény (na ktorej sieť nebola natrénovaná), po natrénovaní extrahovali predposlednú vrstvu siete a s použitím SVM klasifikovali produkty v e-obchode (ktoré sa v pôvodnej dátovej množinu ImageNet nenachádzajú).

Tento princíp sme uplatnili aj my a nechali sme konvolučnú neurónovú sieť Inception-v3 predpovedať kľúčové slová pre filmové plagáty. Tie nás však nezaujímali, a preto predposledná vrstva predstavovala vektor čŕt reprezentujúci film.

Tabuľka 1 zobrazuje výsledky našej analýzy pre vybrané filmové žánre. Pri farebných charakteristikách bola pre niektoré žánre priemerná kosínusová vzdialenosť významne nižšia, ako pre iné žánre. Napríklad, ak porovnáme detský žáner s westernom, vidíme výraznejšie rozdiely. Vo všeobecnosti bola kosínusová vzdialenosť relatívne vyššia pri sémantických črtách, a tu sme už podobné rozdiely, ako pri farebných charakteristikách nespozorovali.

[1] http://www.image-net.org

Tabuľka 1. Porovnanie priemernej vzdialenosti (avg_D) a variancie (σ) pre farebné a sémantické črty v rámci jednotlivých filmových žánrov.

Žáner	**Farebná charakteristika**		**Sémantické črty**		**Počet filmov**
	avg_D	σ	avg_D	σ	
Detský	0.09	0.06	0.35	0.07	1 139
Animovaný	0.10	0.07	0.40	0.09	1 027
Dobrodružný	0.11	0.06	0.36	0.07	2 329
Komediálny	0.13	0.07	0.37	0.06	8 374
Fantazijný	0.13	0.08	0.37	0.06	1 412
Mysteriózny	0.17	0.09	0.37	0.07	1 514
Dramatický	0.15	0.09	0.38	0.06	13 344
Dokumentárny	0.21	0.10	0.40	0.07	2 471
Kriminálny	0.22	0.11	0.36	0.06	2 939
Vojnový	0.23	0.13	0.37	0.07	1 194
Westernový	0.33	0.15	0.34	0.07	676

Takéto zistenie nám môže naznačovať, že farebné charakteristiky sú dobrým indikátorom toho, do akého žánru patrí film. K týmto zisteniam sme dospeli aj manuálnou analýzou plagátov, kde sme odhalili, že napríklad plagáty v rámci žánru *Detský* sú si podobné na základe dominantnej farby, ktorá sa pri mnohých plagátoch opakuje. Naopak, samotný obsah (detekovateľný pomocou sémantických čŕt) sa pri týchto plagátoch líšil.

Vo všeobecnosti, výsledky pre sémantické črty naznačujú, že nie sú vhodné na získanie informácie o žánri, avšak by mohli byť vhodné pre akvizíciu ďalších (latentných) charakteristík filmov.

4 Záver

V našej práci skúmame, ako možno využiť vizuálne stimuly extrahované z obrázkov pre skvalitnenie existujúcich prístupov k obsahovému odporúčaniu. V prvotnej fáze sme sa zamerali na experiment, v ktorom sme sa snažili zistiť, či vizuálne črty vedia nahradiť obsahovú črtu. Využili sme na to rozsiahlu dátovú množinu z domény filmov. Výsledky naznačujú, že vizuálne črty majú potenciál pre ďalšie využitie v budúcej práci, avšak je potrebné preskúmať ich aplikáciu v konkrétnom prístupe odporúčania. Plánujeme preto navrhnúť hybridný Top-N odporúčač, ktorý by s pomocou vizuálnych stimulov vedel skvalitniť vygenerované odporúčanie.

Literatúra

1. Harper, F. M., Konstan, J. A.: The MovieLens Datasets: History and Context. ACM Trans. Interact. Intell. Syst. 5 (2015) 19:1 19:19.

2. Kernix Lab: Image classication with a pre-trained deep neural network. https://www.kernix.com/blog/image-classification-with-a-pre-trained-deep-neural-network_p11 [navštívené 25.8.2017]. (2016).
3. Ricci, F., Rokach, L., and Shapira, B.: Recommender Systems Handbook. Springer US, New York, NY, USA, 2nd edition, (2015).
4. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Proc. of The International Conference on Learning Representations (2015).
5. Szegedy, Ch., et al.: Rethinking the Inception Architecture for Computer Vision. CoRR (2015).
6. Tkalcic, M., Tasic, J. F.: Colour spaces: perceptual, historical and applicational background. In: Proc. of The IEEE Region 8 EUROCON 2003. Computer as a Tool., (2003).
7. Valdez, P., Mehrabian, A.: Effects of color on emotions. Journal of experimental psychology: General (1994), 394.
8. Yu, L., Fangjian, H., Shaobing, H., Yiwen, L.: A content-based goods image recommendation system. Multimedia Tools and Applications (2017) 1 15.
9. Zhou, Z., Xu, Y., Zhou, J., Zhang, L.: Interactive Image Search for Clothing Recommendation. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM, New York, NY, USA (2016), 754 756.

Poďakovanie: Tento článok vznikol vďaka čiastočnej podpore projektov *Informačné správanie sa človeka v digitálnom priestore* Agentúry na podporu výskumu a rozvoja, projekt č. APVV-15-0508 a *Inovatívne metódy výučby informatiky vo veľkých skupinách s podporou online vzdelávania* Kultúrnej a edukačnej grantovej agentúry, grant č. KEGA 028STU-4/2017.

Annotation:

Personalized Recommendation Enhanced by Visual Stimuli

User behavior analysis on the Web is still an open research problem. Users interact with Web pages in a various ways and researchers are trying to correctly interpret this behavior. User interactions are an input for the recommender systems that are employed as one of the methods intended to enhance user experience on the Web. Content-based recommendation uses item properties to generate personalized recommendations. Images have become a popular source of information in many domains (e.g., online shopping) where they influence user decision process. In our work, we analyze visual features extracted from images. Our goal is to study how images can be incorporated into the recommendation process.

The agent-based model of the dynamic spectrum access networks based on the bilateral bargaining

Marcel Vološin, Eugen Šlapak, Juraj Gazda

Department of Computers and Informatics, Technical University of Kosice

marcel.volosin@student.tuke.sk,
eugen.slapak@student.tuke.sk juraj.gazda@tuke.sk

Abstract. This paper describes a simple bargaining mechanism that allows autonomous agents to engage in a bargaining decisions taking place in the dynamic spectrum access market. The agents take on the role of operators which aims to purchase the frequency spectra on the wholesale market and in turn, provide the services towards the end-users. In the paper, the wholesale distribution of the resources in the agent-based model is governed by the bilateral bargaining between the operators and the spectrum broker. The aim of such an interaction is to reach agreements (agreement in terms of the negotiated wholesale price and number of contracted channels) through an iterative bargaining. The operators aim to attract the end-users via dynamic retail pricing scheme. The activity of the end-users follows the truncated Gaussian distribution, which mimics the activity of the end-users throughout the daily cycle. The numerical results suggest that the model is capable of capturing the stochastic activity of the end-users, which reflects itself in the variable operator's profit and retail price.

Keywords: agent-based modeling, cognitive networks, electronic markets, MASCEM

1 Introduction

The spectrum trading allows the holders of the certain spectrum licenses to transfer or lease all or part of their rights and obligations under their license to another party. Several countries have implemented spectrum trading, but the trading process is often time-consuming, hence hampering the usage. The UK regulator Ofcom is in the forefront on the spectrum trading arena, allowing spectrum sale and spectrum leasing [1]. The relevance of the spectrum leasing raises its further importance in conjunction with the concepts of the software-defined radio (SDR). SDR enhances the efficiency of the frequency spectrum utilization through the embedded software allowing the terminal to operate in multiple frequency bands using multiple transmission protocols [2]. Spectrum leasing and SDR are technologies coined throughout the paper as the dynamic spectrum access (DSA) strategies.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 181-186.

DSA strategies fundamentally change the traditional telecom model based on "vertical integration". Here the single entity delivers the service, maintains the network and the network infrastructure [3]. Originally, the available services were mainly limited to the telephony, the radio, and the television. However we do witness the convergence of these services nowadays and also the roles of the service provider and the network owner are separated, and the service providers get access to the network and the end customers through the secondary spectrum trading on fair and non-discriminatory conditions [4]. This telecommunication concept is recognized as the open access network.

Most trades today are direct trades between organizations with the regulator as an intermediate giving the final consent to commit the trades. However, in order to facilitate spectrum trades on a shorter time scale an organizational unit such as a band manager could be introduced to mediate between the traders. Furthermore, the organizational units could be introduced to monitor for compliance with the committed trades and that the spectrum is not misused. Overall, an ecosystem is required in order to realize the spectrum micro-trading. Therefore, in this paper, we propose an agent based model of dynamic spectrum access network capable of capturing key network characteristics.

2 Agent-based model

In order to simulate the spectrum trading in the open access spectrum network, we propose the model based on a MASCEM [5] which simulates a behavior of a subjects taking place in the electricity markets. Similarly to the DSA network, MASCEM includes both the wholesale and the retail market where the bilateral negotiation is implemented. Inspired by the electricity markets, our model uses the following price shaping functions on the wholesale market:

$$starting\_offer_{i+1} = starting\_offer_i \pm \delta_{i+1} \tag{1}$$

$$\delta_{i+1} = starting\_offer_i \times \left(\beta + \frac{\Delta_i}{BWavail_i \times \alpha}\right) \tag{2}$$

$$\Delta_i = BWavail_i - BWleased_i. \tag{3}$$

The new starting offer $starting\_offer_{i+1}$ depends on the $starting\_offer_i$ from the previous period's offer. The formula scales the change by a ratio to lease all the available capacity. The amount of change increases with the difference between the amount of the bandwidth (BW) administrator wanted to lease and the amount it actually leased. α and β are price-shaping parameters.

The negotiation itself consists of the time bounded series of rounds during which the operators and the administrator negotiate the contract conditions according to:

$$offer_{i+1} = offer_i \pm \varepsilon. \tag{4}$$

The price-change value is denoted as ε. While the administrator lowers its price with the given ε, the operators do increase their offers with the ε. When negotiating, nei-

ther of the stakeholders exceeds its $limit_price_i$which is set at the beginning of the each series according to:

$$limit\_price_i = starting\_offer_i \pm \vartheta\ , \tag{5}$$

where the ϑ represents the limit offer parameter. When the negotiation successfully finishes, the operators consider the amount of spectrum according to the negotiated price. Higher the final price, lower the amount of the channels will be leased. Due to the different characteristics of the frequency spectrum and the electric energy, it is necessary to define suitable rules that will be applied when the trading takes place on the retail market. The acceptance probability function was adopted from [6] with some slight modifications. The spectrum obtained via bilateral negotiation is sold to the end-users on the retail market for the prices accommodated by the end-users demand. The following equations are used for the per channel retail price calculation:

$$p_i = p_{i,t} + \left(\Psi_{i,t-1} - 0.5\right) \times \mu \tag{6}$$

$$\Psi_{i,t-1}\begin{cases} 1/2 & (BWavail_i = 0) \wedge (S_i = 0) \\ 0 & (BWavail_i > 0) \wedge (S_i = 0) \\ S_i^{idle \to conn}/S_i & (BWavail_i > 0) \wedge (S_i > 0), \end{cases} \tag{7}$$

where $BWavail_i$ represents the amount of the available frequency channels, S_i is the total number of the end user connection attempts towards a particular operator and $S_i^{idle \to conn}$ denotes the number of successful connections.

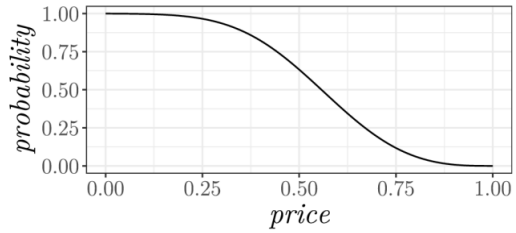


Figure 1 Acceptance probability according to retail price

The behavior of the end-users introduces the stochasticity to the model. During the simulation, their state is being switched between the following three different options: IDLE, ACTIVE and CONNECTED. The state change IDLE → ACTIVE occurs randomly with a probability that changes during daytime on the interval $\langle 0.1; 0.8 \rangle$ with the Gauss-like curve. In the ACTIVE state, the end-users in order to obtain the available frequency channels actively search for the most suitable operator providing the best offer. Each offer is evaluated with the acceptance probability AP by:

$$AP_i = 1 - e^{-c(1-p_i)^\gamma} \tag{8}$$

The higher the value of AP the more probable the connection will be successful. The values of c and γ adjust the end-users sensitivity towards the price p of the i-th operator. Figure 1 shows the end-users sensitivity when the parameters c and γ are set according to Table 1.

3 Simulation Results

The presented data illustrate the key aspects of the model behavior during the period of the five simulated days, each consisting of 500 virtual time units. The model was set according to the Table 1 in order to imitate the real world. The end-users sensitivity towards the utility reflects the signal decay and also affects the overall activity of these users. The values of price-shaping and the negotiation parameters of MASCEM inspired mechanism, as was found, are tightly bounded with the model configuration and were therefore found using a modified model which had an ability to vary the parameter values on fly in order to maximize the profit and the spectrum utilization. The retail market mechanism was set according to [6].

Table 1 Simulation parameters

Parameter	Value	Description
N_{bts}	1	Number of base stations
BW_{total}	200	Total number of available channels
N_{inv}	2	Number of operators
$N_{end-users}$	$\langle 300; 500 \rangle$	Number of end-users
α_{adm}	-0.7	Price-shaping parameter of administrator
β_{adm}	0.02	Price-shaping parameter of administrator
α_{inv}	-3	Price-shaping parameter of operator
β_{inv}	0.1	Price-shaping parameter of operator
p_{min}	0.15	Minimum wholesale price
ε	0.01	Negotiation parameter
P_{act}	$\langle 0.1; 0.8 \rangle$	End-users' activation probability
P_{disc}	1	End-users' disconnection probability
ϑ	0.3	Limit offer parameter
γ	3	End-users' price sensitivity parameter
δ	0.5	End-users' utility sensitivity parameter
μ	0.2	Price coefficient
c	8	Acceptance probability parameter

Figure 2 illustrates the most important characteristic of our agent based model. The presented data were collected from the simulation with the fixed number of the end-users (450 users). In the Negotiation process plot, the price lag can be observed. The situation, when the price increases much slower than the end-users' demand. The consequences of the described lag can be seen in the Operators' profit plot which is higher at the beginning of a day despite a significantly lower demand. After the peak

utilization, the negotiation starts the prices to decay to the preset minimum and the retail price becomes more volatile, due to the smaller number of the active end-users.

To verify the model's stability with a different environment setup, we run multiple simulations with the increasing number of the end-users. The each setup was run multiple times and it consisted of 30 simulated days during which data were collected and each plotted dot represents mean value of 15000 entries.

From the Figure 4, we can conclude that increasing number of end users, results in the increasing spectrum demand and its variance too. Higher demand causes the higher wholesale prices that have the higher variance due to the MASCEMs characteristics of the negotiation process. The mean retail prices are also raised but unlike the wholesale price, more active users result in the less volatile prices.

4 Conclusion

This paper describes the implementation of a spectrum trading mechanism in a NetLogo agent based model inspired by the electricity market model called MASCEM. An administrator owning the spectrum license, the operators providing the services using a rent infrastructure and the end-users benefiting from the operators' competition were included into the model in order to simulate an operation of such network. The presented results show that the bilateral negotiation between the stakeholders, an administrator and the operators, on the wholesale market produces the satisfactory results in capturing the stochastic activity of the end-users whose activity follows the truncated Gaussian distribution in order to mimic a daily cycle. However, the observed price lag makes space for further improvement.

References

1. Grønsund P., et al.: Towards spectrum micro-trading. Future Network & Mobile Summit (FutureNetw), 2012.
2. Arslan, Hüseyin, ed.: Cognitive radio, software defined radio, and adaptive wireless systems. Vol. 10. Berlin: Springer, 2007.
3. Zhang, Ning, et al.: Dynamic spectrum access in multi-channel cognitive radio networks., IEEE Journal on Selected Areas in Communications 32.11, 2014, 2053-2064.
4. Cramton, Peter, and Linda Doyle.: Open access wireless markets., Telecommunications Policy, 2017.
5. Isabel Praça, et al.: MASCEM: A multiagent system that simulates competitive electricity markets., IEEE Intelligent Systems 18.6, 2003: 54-60.
6. Pastirčák J.,et al.: An Agent-Based Economy Model of Real-Time Secondary Market for the Cognitive Radio Networks, Journal of Network and Systems Management. 12.10 2015, 1-17

Acknowledges: This work was supported by the Slovak Research and Development Agency, project number APVV-15-0358, and by European Intergovernmental Framework COST Action CA15140: Improving Applicability of Nature-Inspired Optimisation by Joining Theory and Practice.

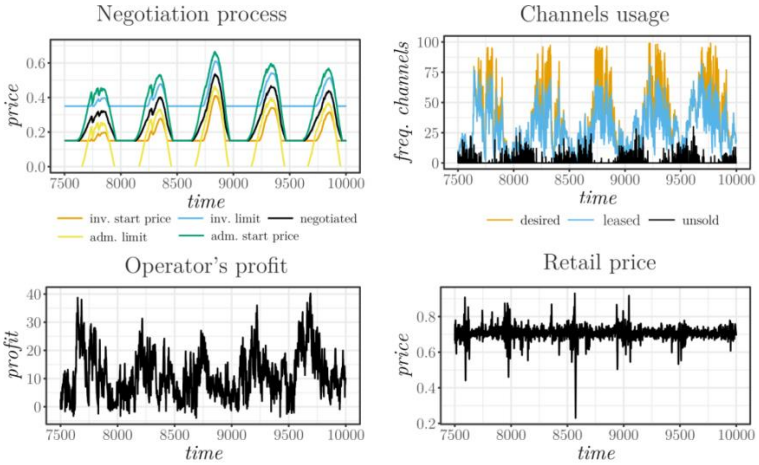


Figure 2 Real-time behavior of spectrum trading model inspired by MASCEM

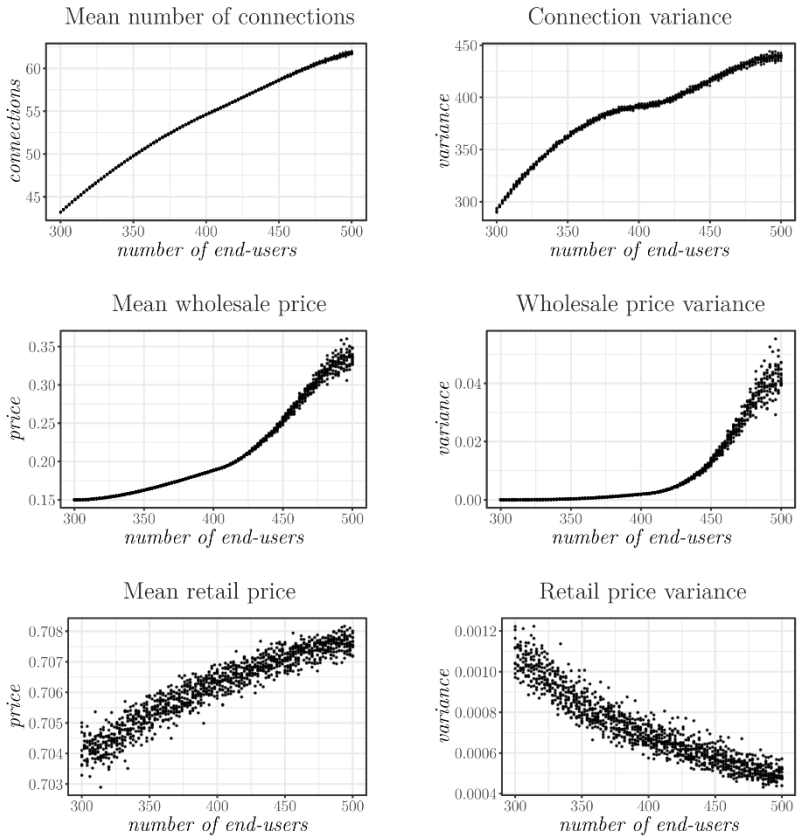


Figure 3 Model behavior with increasing number of end-users

Predikcia úpadku spoločností s ručením obmedzeným využitím metód pre rozpoznanie odľahlých bodov

Peter Gnip[1], Martin Zoričák[2] a Peter Drotár[1]

[1]Katedra počítačov a informatiky, Technická univerzita v Košiciach
[2]Katedra financií, Technická univerzita v Košiciach

Abstrakt. Spoločnosti pôsobiace v rámci obchodného a priemyselného odvetvia sa môžu vplyvom nepriaznivej finančnej situácie, alebo nevhodného obchodovania, dostať do finančných ťažkostí, ktoré neskôr vyústia do celkového úpadku spoločnosti. Analyzovali sme dáta obsahujúce tisíce záznamov spoločností s ručením obmedzeným (s.r.o) pôsobiacich na Slovensku v rôznych odvetviach hospodárstva v období rokov 2013-2016. K nastolenému problému sme pristupovali ako k problému rozpoznania odľahlých hodnôt (outliers), pričom bola použitá metóda podporných vektorov pre detekciu odľahlých bodov (OneClassSVM). Dáta pozostávali z 20 štandardných ekonomických ukazovateľov. V prvotnej analýze sme sa zamerali na predikciu úpadku s.r.o. na základe účtovných údajov z jedného roku a kombináciou dvoch po sebe idúcich rokov. Dosiahnutá presnosť predikcie bola od 60,56% do 77,91 % v závislosti od roku v ktorom sme uvažovali výsledný stav spoločnosti a roku z ktorého boli čerpané ekonomické ukazovatele.

1 Úvod

Predikovanie úpadku firiem je téma, ktorou sa autori zaoberajú posledných vyše 80 rokov. Schopnosť čo najpresnejšie predikovať úpadok firmy na základe údajov z predchádzajúcich období je cenná pre manažment spoločnosti, investorov, veriteľov ako aj banky. Opodstatnenosť témy zdôrazňuje počet publikovaných vedeckých článkov v danej oblasti s využitím veľkého počtu metód. Približne od začiatku 30. rokov do 60. rokov autori uvádzali metódy na základe pomerových ukazovateľov, kde aplikovali len jeden takýto ukazovateľ [1]. Postupne začali vznikať modely s viacerými ukazovateľmi, a postupne bol pridávaný väčší počet pomerových ukazovateľov do modelov, z ktorých sú najznámejšie diela Altmana 1968 s použitím piatich faktorov [2], resp. Boritza a Kennedyho 1995 so štrnástimi faktormi [3]. Frekventovaná metóda bola diskriminačná analýza, ktorú následne začali nahrádzať iné štatistické metódy ako napríklad regresná analýza (logit, probit). V rámci štatistických metód boli aplikované metódy na redukciu dimenzionality, konkrétne faktorová analýza a analýza hlavných komponentov. S rozvojom techniky a prístupom k stále rozsiahlejším databázam, autori aplikovali tzv. inteligentné metódy - metódy strojového učenia a dolovania dát (data mining) v oblasti predikovania úpadkov. Medzi tieto metódy patrí využitie neurónových sietí, rozhodovacích stromov, metódy podporných vekto-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 187-191.

rov (SVM), genetických algoritmov, lineárne programovanie, kvadratické programovanie či data envelopement analysis[4][5]. Problém identifikácie spoločnosti pred bankrotom z pohľadu inteligentných metód spočíva v správnej binárnej klasifikácii firmy do rizikovej (bankrotujúcej) skupiny alebo nerizikovej (nebankrotujúcej) skupiny. Na základe takto postaveného problému je potom možné testovať modely z viacerých hľadísk, ako presne vedia zatriediť firmy do správnej kategórie.

Vzhľadom na charakter problematiky, často dochádza k problému nevyváženej dátovej sady, pričom počet zdravých firiem mnohonásobne presahuje počet bankrotujúcich. Niektoré z predchádzajúcich štúdií [4],[6] boli založené na dátach obsahujúcich niekoľko stoviek firiem, pričom tieto boli rovnomerne rozdelené medzi spoločnosti v úpadku a zdravé spoločnosti. Ako naznačujú dáta, získané zo štandardných výkazov finančnej správy, takéto rozdelenie dát je v praxi nereálne, nakoľko počet spoločností v úpadku predstavuje len zlomok z celkového počtu spoločností. Preto je potrebné k predikcii úpadku pristupovať ako k problému predikcie nevyvážených dát (*imbalanced data*) [7].

V ďalšom bližšie opíšeme použité dáta, metodológiu, ako aj prvotné výsledky, t.j. presnosť s akou dokážeme identifikovať upadajúce spoločnosti.

2 Dáta

Analyzovali sme dáta obsahujúce tisíce záznamov spoločností s ručením obmedzeným (s.r.o) pôsobiacich na Slovensku v rôznych odvetviach hospodárstva v období rokov 2010-2016. Odvetvia, zahrnuté v dátovej sade sú: maloobchod, poľnohospodárstvo, priemysel, stavebníctvo a hotely. V tejto štúdii sa zameriame výhradne na dáta z oblasti priemysel. Ku každej s.r.o sú k dispozícii dáta z účtovných uzávierok za posledné tri roky, pred rokom evaluácie *R* (rok v ktorom sa hodnotil stav spoločnosti zdravá/úpadok). Spoločnosť je hodnotená ako zdravá ak neukončila v danom roku činnosť. Pod upadajúcou/bankrotujúcou spoločnosťou rozumieme spoločnosť, na ktorú bolo v predmetnom roku evaluácie vyhlásené konkurzné konanie. Konkrétne počty s.r.o v úpadku a zdravých spoločností hodnotených v jednotlivých rokoch sú uvedené v Tab. 1. Dáta sú získané z databázy Bisnode Magnusweb, kde je možné exportovať dáta vo formáte *csv,* tieto boli ďalej spracovávané použitím programovacieho jazyka *Python* a jeho modulov.

Tabuľka 1. Charakteristika použitých dát

rok evaluácie *R*	2013	2014	2015	2016
počet zdravých/bankrotujúcich	4094/30	4482/30	5046/28	5959/16

Databáza Bisnode Magnusweb umožňuje priamy export 20 štandardných ekonomických ukazovateľov, ktoré sme využili pre predikciu úpadku. Tieto štandardné ukazovatele pozostávajú menovite z: rentabilita aktív (ROA), rentabilita vlastného imania (ROE), rentabilita tržieb (ROS), okamžitá likvidita (L1), pohotová likvidita (L2), bežná likvidita (L3), doba obratu aktív (DOZ), doba obratu kr. pohľadávok (DOKP),

doba obratu kr. záväzkov (DOKZ), doba obratu zásob (DOZ), celková zadlženosť (CZ), zadlženosť vlastného imania (ZVI), finančná páka (FP), návratnosť investícií (NI), krytie cudzích zdrojov (CZ), finančné krytie cudzích zdrojov (FKCZ), krytie stálych aktív (KSA), zadlženosť voči bankám (ZB), mzdy / tržby (MT), mzdy / pridaná hodnota (MPH). Predmetné ukazovatele obsahujú informáciu o finančných tokoch a majetkových pomeroch v rámci firiem pri zohľadnení veľkostných rozdielov medzi firmami.

3 Metodológia

K nastolenému problému sme pristupovali ako k problému rozpoznania odľahlých hodnôt (*outliers*), pričom bola použité metóda podporných vektorov pre detekciu odľahlých bodov (*OneClassSVM*) [8], konkrétne jej implementácia v scikit-learn [9].

Chýbajúce dáta boli imputované doplnením stredných hodnôt vybraného ukazovateľa. Na natrénovanie klasifikátora bolo použitých 80% vzoriek zdravých spoločností. Samotné ukazovatele boli škálované na nulovú strednú hodnotu a jednotkový rozptyl. Výkonnosť klasifikátora bola testovaná na vzorkách upadnutých spoločností a zvyšných 20% vzorkách zdravých spoločností. Tieto experimenty boli opakovane päť krát s náhodným výberom trénovacích dát a výsledky boli spriemernené.

Môže dôjsť k dvom typom chýb, chyba prvého typu je identifikácia bankrotujúcej firmy ako zdravej a chyba druhého typu je, ak sa zdravá firma identifikuje ako bankrotujúca, preto sme okrem celkovej presnosti predikcie hodnotili aj správne pozitívne hodnoty (TPR, z angl. *true positive rate*) a správne negatívne hodnoty (TNR, z angl. *true negative rate*)

4 Numerické výsledky

V prvotnej analýze sme odhadovali hroziaci bankrot s.r.o, pričom boli použité ekonomické ukazovatele z dát z predchádzajúcich troch rokov jednotlivo, t.z. pre s.r.o, bankrotujúcu v roku 2016, boli na predikciu použité dáta z roku 2015 (*R-1*), potom samostatne z roku 2014 (*R-2*) a samostatne z roku 2013 (*R-3*). Obdobne aj pre spoločnosti bankrotujúce v rokoch 2013-2015. Následne bola predikcia realizovaná využitím dát z dvoch po sebe idúcich rokov, t.j. pre s.r.o bankrotujúce v 2016, boli pre predikciu použité najprv dáta z rokov 2014 a 2015 (*R-2 & R-1*) a následne dáta z rokov 2013 a 2014 (*R-3 & R-2*). Výsledky predikcie sú uvedené v Tab. 2.

Z uvedených hodnôt presnosti predikcie môžeme pozorovať, že dáta z rokov bližšie k roku evaluácie sú výpovednejšie. Na základe údajov z obdobia *R-1* bola dosiahnutá presnosť predikcie na úrovni 77,91 % pre rok *R=2015*. To je v súlade s očakávaním, nakoľko môžeme predpokladať, že s blížiacim sa bankrotom sa v účtovných uzávierkach výraznejšie prejavujú finančné problémy.

Na druhej strane, spojením dát z dvoch rokoch sme očakávali aj zvýšenie presnosti predikcie. V tomto smere sa však presnosť výrazne nezmenila. Maximálna presnosť predikcie v tomto prípade dosahovala 72,41 % pre rok evaluácie *R=2015* a dáta z rokov

Tabuľka 2. Presnosť predikcie úpadku v jednotlivých rokoch

	Rok z ktorého pochádzajú dáta použité na predikciu				
Rok evaluácie (*R*)	*R-1*	*R-2*	*R-3*	*R-1 & R-2*	*R-2 & R-3*
	73,58±2,84	**67,21±2,19**	**68,08±1,57**	**71,88±3,37**	**73,51±2,57**
	TNR: 64,52±9,48	TNR: 77,42±8,56	TNR: 56,67±5,67	TNR: 70,97± 6,99	TNR: 53,33± 7,12
	TPR: 69,28± 5,92	TPR: 70,33± 8,59	TPR: 67,84± 4,86	TPR: 68,56± 9,33	TPR: 78,51± 8,29
	74,25±2,07	**66, 55±2,14**	**62,50±1,30**	**67,58±2,80**	**61,91±0,80**
	TNR: 80,65± 6,42	TNR: 41,94± 5,28	TNR: 53,33± 11,18	TNR: 74,19± 6,74	TNR: 43,33± 4.33
	TPR: 77,69± 4,56	TPR: 79,88± 6,59	TPR: 75,00± 10,17	TPR: 79,70± 4,06	TPR: 88,29± 5,07
	77,91±2,64	**72,28±1,05**	**66,41±2,33**	**72,41±2,67**	**69,12±1,41**
	TNR: 73,08± 10,58	TNR: 67,86± 5,71	TNR: 57,14± 6,93	TNR: 61,54± 6,21	TNR: 57,14± 11,52
	TPR: 71,09± 8,22	TPR: 83,18± 4,96	TPR: 78,41± 5, 30	TPR: 81,04± 3,99	TPR: 70,69± 10,39
	67,07±4,08	**60,56±4,52**	**62,80±2,09**	**67,58±3,56**	**68,18±3,06**
	TNR: 14,29±12,14	TNR: 50,00± 14,64	TNR: 68,75± 12,33	TNR: 67,76±12,14	TNR: 62,50± 10,90
	TPR: 95,59±6,98	TPR: 65,67± 9,96	TPR: 65,70± 10,77	TPR: 70,86±8,15	TPR: 69,30± 7,54

2014 a 2013. Dôvodom môžu byť samotné dáta, kde pridaním údajov z ďalšieho roku nie je pridaná nová informácia, iba tá, ktorá je už obsiahnutá v dátach. Ďalšou hypotézou je, že po spojení údajov z dvoch rokov nedochádza k zlepšeniu, nakoľko klasifikátor nie je schopný dáta dostatočne zmapovať. Ďalšie vyšetrovanie bude predmetom ďalšieho výskumu.

Ako môžeme na základe údajov v Tab. 2 pozorovať, klasifikátor dosahuje pomerne odlišné výsledky v jednotlivých rokoch. V presnosti predikcie je rozdiel až 10% v predikcii bankrotu s.r.o., ktoré upadli v roku 2015 a v roku 2016. To naznačuje, Že v prípade databáz, ktoré sú využívané na predikciu úpadku je vhodnejšie analyzovať dáta z viacerých rokov.

5 Záver a ďalšie smerovanie výskumu

V tomto príspevku sme prezentovali prvotné výsledky predikcie úpadku s.r.o. s maximálnou úspešnosťou 77,91%. Výsledok je značne ovplyvnený rokom z ktorého sa dáta čerpajú. Avšak tieto výsledky naznačujú, že úpadok je predikovateľný.

V ďalšej práci sa zameriame na aplikáciu iných metód pre predikciu odľahlých hodnôt, a aj metód pre klasifikáciu s učiteľom (*supervised*), využitie metód pre podvzorkovanie majoritnej triedy a podobne. Ďalšou sľubnou oblasťou je návrh nových

ukazovateľov, ktoré dokážu zachytiť finančnú situáciu výstižnejšie ako tradičné finančné ukazovatele, čo môže značne uľahčiť úlohu klasifikačnému algoritmu.

Poďakovanie: Táto práca bola podporená Agentúrou na podporu výskumu a vývoja projektom číslo APVV-15-0358.

References

1. Bellovary, Jodi L., Don E. Giacomino, and Michael D. Akers. "A review of bankruptcy prediction studies: 1930 to present." Journal of Financial education (2007): 1-42.
2. Altman, Edward I. "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy." The journal of finance 23.4 (1968): 589-609.
3. Boritz, J. Efrim, and Duane B. Kennedy. "Effectiveness of neural network types for prediction of business failure." Expert Systems with Applications 9.4 (1995): 503-512.
4. Zhao, D., Huang, C., Wei, Y., Yu, F., Wang, M., Chen, H.: An Effective Computational Model for Bankruptcy Prediction Using Kernel Extreme Learning Machine Approach. Computational Economics 49(2), 325-341 (2017).
5. Wang, L., Wu, C.: Business failure prediction based on two-stage selective ensemble with manifold learning algorithm and kernel-based fuzzy self-organizing map. Knowledge-Based Systems 121, 99-110 (2017).
6. Pietruszkiewicz, W.: Dynamical systems and nonlinear Kalman filtering applied in classification. Cybernetic Intelligent Systems, 2008. CIS 2008. 7th IEEE International Conference on.London, UK.
7. Zhou, Ligang. "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods." Knowledge-Based Systems 41 (2013): 16-25.
8. Manevitz, L., Yousef, M.: One-Class SVMs for Document Classification. Journal of Machine Learning Research 2, 139-154 (2001).
9. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Detecting Anomalous Trajectories and Traffic Services

Mazen Ismael

Faculty of Information Technology, BUT
Božetěchova 1/2, 66 Brno

Mazen.ismael@vut.cz

Abstract. Among the traffic studies; the importance of detecting anomalous trajectories of vehicles rises to support many services, starting from securing and safety services to the maps and navigation services. The combination of many methods and concepts could offer interesting advantages, and *iBAT* (Isolation-Based Anomalous Trajectory) is one of the advanced frameworks which detect anomalous traffic trajectories. *iBOAT* (Isolation Based Online Trajectory) came after that as a version of *iBAT* able to process online data. Beside of that, using semantic locations could support the navigations studies, and increase the maps' accuracy. In fact, developing the *iBOAT* framework with use semantic ocation could bring out interesting results. The aim of this paper is to present the progress of detecting anomalous driving patterns from GPS trajectories, which will be achieved by using the concept of semantic locations for improving the scene partitioning.

Keywords: Anomalous Trajectory, Semantic Location, Traffic, iBAT/iBOAT

1 Introduction

Nowadays, several methods and frameworks have been proposed to achieve the goal of detecting anomalous trajectories by using *GPS* data, which aim to support securing, safety services, the maps and navigation services, and many other services. Some of these frameworks [1] [2] had good results but still need to improve their methods to achieve satisfactory results. Important preprocessing methods have been used by some frameworks [3] [4] to prepare the backgrounds. Moreover, semantic locations concept could play an efficient action in these methods for improving clustering of the trajectories to divide the background into zones. It is useful to mention here that the techniques will be used for design the framwork are connected to the offline data only. In this paper, we will present the work progress in design our framework, and review the current progress. Beside of that, we will show our implementation of the *iBOAT* framework. Finally, conclusion and future work will be presented.

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 192-197.

2 Related work

In this section; related methods in each phase of our framework will be presented. The first phase is a preprocessing which mainly focuses on preparing the data and by partitioning the scene. In this area, *Zheng, Y* [5] reviewed many algorithms for generating and reducing the storage of the data (Uniform sampling, Douglas-Peucker (DP), Top-down time-ratio (TD-TR), Bellman). Besides that, many filters were reviewed by *Zheng, Y* [5](Mean and Median Filters, Kalman Filter, Particle Filter) to filter the trajectories and ignore the outliers and reduce their effect. In the direction of preparing the background, many frameworks processed the background by dividing the scene to grid view cells similar to how iBAT and iBOAT frameworks processed the scene [1][2]. On the other hand, *Brun, L* [3][4] divided the scene into zones using an algorithm is able to cluster a dataset from the trajectories. In details, the algorithm could summarize as follows: consider the entire scene as one zone and then divide the zone into *L* fixed number of zones by using the distribution of training set. After that, each zone will be represented by using statistical properties (mean, the major axis and covariance matrix).

For the secound phase, studies of detecting anomalous behavior which is related to the traffic are widely common, most of these studies have built their frameworks based on Hidden Markov Model (HMM) *Cai, Y* [6]. While other methods used graphs[4] or built other frameworks similar to *Zhang et al.* [2] when he built iBAT framework and after that iBOAT [1] framework. iBAT/iBOAT frameworks aim to discover anomalous driving patterns from taxis trajectories. In fact, iBOAT improved iBAT to be able to work in online environments and maximized the grid cell size in the step of the scene's preprocessing to ensure the accuracy, by using experimental sizes (250m x 250m). Beside of that, iBOAT used a function to solve the problem of low sampling rate and cells' gaps, by this function the algorithm considers the points of trajectories located in the neighbor cells as normal points. Moreover, iBOAT used an *adaptive working window* and hasPath method which is described in the following algorithm [1]:

1. $\chi \leftarrow \emptyset$// initialization, χ is the set of anomalous points
2. $T_0 \leftarrow T$ // T_0 first trajectory, T set of trajectories
3. $i \leftarrow 0$ // Position in incoming trajectory
4. $w \leftarrow \emptyset$// Adaptive window from t
5. $score(0) \leftarrow 0$
6. **while** the testing trajectory is not completed **do**
7. $i \leftarrow i + 1$
8. $g_i = \rho(p_i)$
9. $w \leftarrow w \cdot g_i$
10. $support(i) = |hasPath(T_{i-1}, w)|/|T_{i-1}|$
    // *hasPath* returns the set of trajectories that contain all of the points in t in the correct order
11. $T_i \leftarrow hasPath(T_{i-1}, w)$ // working set reduced.
12. **if** $support(i) < \theta$ **then** // where θ is threshold.
13. $\chi \leftarrow \chi \cup p_i$
14. $T_i \leftarrow T$ // reset the working set
15. $w \leftarrow g_i$
16. **end if**
17. $score(i) = score(i - 1) + \sigma(support(i)) * dist(p_{i-1}, p_i)$ // $\sigma(x) = 1/1 + e^{\lambda(x-\theta)}$
18. **end while**

3 Proposed framework

The structure of the proposed framework contains two main phases, preprocessing and detecting abnormal trajectories. In this framework, besides of preparing the data, we are looking to develop a method for dividing the scene into zones based on a dataset of trajectories. Semantic locations from the third party should affect the weight of the places, there for, GPS points are near the places have a significant weight will consider as visited the places and could cluster better. This improvement should avoid the outliers (which exist because of inaccuracies in sampling the GPS) and increase the robustness of clustering of the trajectories.

The second phase of this framework aims to detect anomalous trajectories, using the *adaptive working window*. Updating maps could be one of the implementations of this framework, by discovering the new official roads based on the number of anomalous trajectories passing through the same path.

Most of the studies of the related art of the works were finished, and some datasets from many sources have been tested if they could be suitable for this study. After that, an interesting dataset has been selected Berlin Moving Object Data (BerlinMOD[1]) for two days as *Figure 1* shows; the data has been preprocessed and imported in PostgreSQL server. Moreover, an interesting daily updated information about Berlin City offered by Geofabrik – German[2] has been imported to the PostgreSQL server and presented by ArcMap system. And by using filters of the building in ArcMap, and by

[1] http://dna.fernuni-hagen.de/secondo/BerlinMOD/BerlinMOD.html
[2] http://download.geofabrik.de/europe/germany/berlin.html

using filters of the building in ArcMap, museums (as examples of important places which should have more weight) could be separated from other buildings.

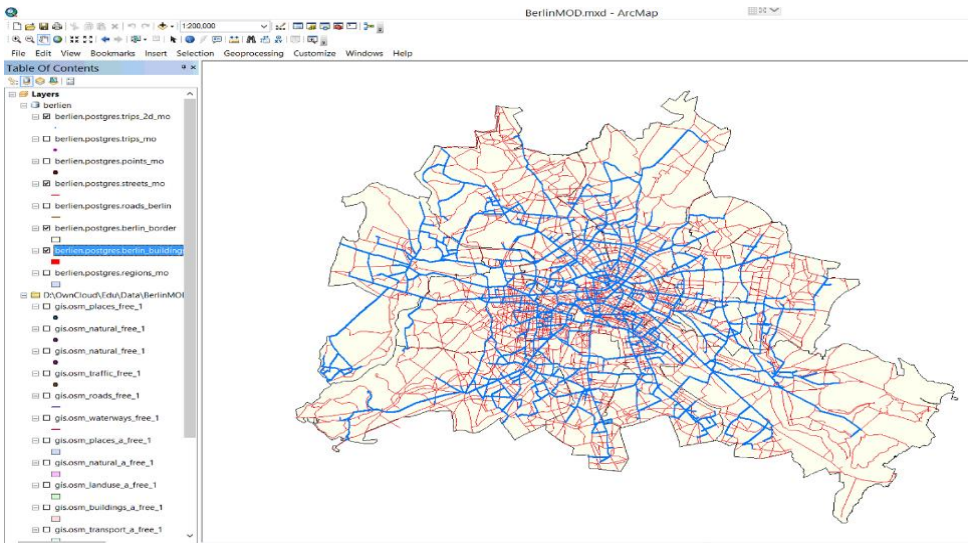


Figure 1 BerlinMOD with Geofabrik data (Berlin)

In addition, iBOAT has been implemented by Koňárek, P [7], this implementation was developed and special functions were added to load and process BerlinMOD dataset which recorded two days from Berlin traffic. *Figure 2*. For that process the data. And according to the trajectories dimensions, the cell size is (37.08 x 37.08) in the metric system, and *theta* threshold is 0.005. Moreover, when trajectory with an *ID* is selected in the left-hand table, the anomalous trajectory is drawn in bold line on the map, and *iBOAT Score* chart draws where the trajectory act anomaly (fixed score values are normal sub-trajectories).

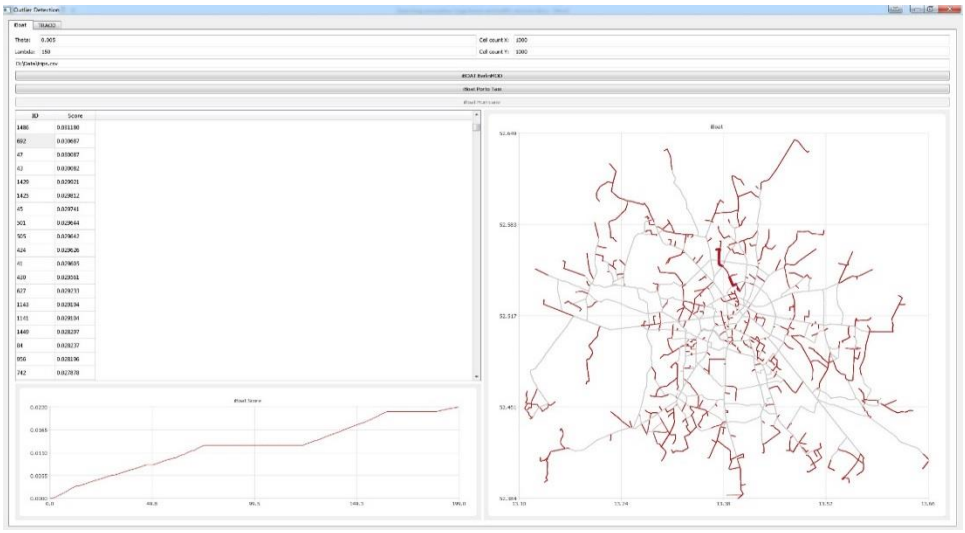


Figure 2 Implementation of BerlinMOD dataset

The big cells' size will allow many trajectories to act anomalously without discovering that by the algorithm. On the other hand, the small cells' size means the algorithm will consume more computing time (more cells objects), and the Pos function will find that the new cell is not neighbor to any of the trajectory's cells and, as a result, the normal points will be detected as abnormal. For that, the cells size should be chosen manually based on the dataset dimensions. Furthermore, the threshold θ should not be big (no anomalous points will be selected) or too small (all the points will be recognized as anomalous trajectories).

4 Conclusion and future work

This paper described the most important issues related to the presented framework designed to detect traffic anomalous trajectories. In our framework, the development will affect all the levels to solve the problems of the outliers which exist because the low sampling rates and will change the results to be closer to the lifestyle.

Finally, information extracted from the web for rating the attraction places in Berlin will be used in the next step to affect the zoning process. Moreover, update the maps, or detecting the importance of locations based on detecting anomalous trajectories, all of that should be implemented by this framework.

References:

1. Chen, C., Zhang, D., Castro, P. S., Li, N., Sun, L., Li, S., & Wang, Z. (2013). iBOAT: Isolation-based online anomalous trajectory detection. IEEE Transactions on Intelligent Transportation Systems, 14(2), 806–818. https://doi.org/10.1109/TITS.2013.2238531

2. Zhang, D., Li, N., Zhou, Z.-H., Chen, C., Sun, L., & Li, S. (2011). iBAT: Detecting anomalous taxi trajectories from GPS traces. Proceedings of the 13th International Conference on Ubiquitous Computing, 99–108. https://doi.org/10.1145/2030112.2030127
3. Brun, L., Capellini, B., Saggese, A., & Vento, M. (2014). Detection of anomalous driving behaviors by unsupervised learning of graphs. 11th IEEE International Conference on Advanced Video and Signal-Based Surveillance, AVSS 2014, 405–410. https://doi.org/10.1109/AVSS.2014.6918702
4. Brun, L., Saggese, A., & Vento, M. (2014). Dynamic Scene Understanding for Behavior Analysis Based on String Kernels, 24(10), 1669–1681.
5. Zheng, Y. (2011). Compute with spatial trajectories, ISBN: 9781461416289. Chapter 1.
6. Cai, Y., Wang, H., Chen, X., & Jiang, H. (2015). Trajectory-based anomalous behavior detection for intelligent traffic surveillance. Iet Intelligent Transport Systems, 9(8), 810–816. https://doi.org/10.1049/iet-its.2014.0238
7. Koňárek, P., (2017). Graduation thesis "Mining Anomalous Behavior in Trajectory Data." Faculty of Information Technology, BUT.

Acknowledgments: The research is supported by project "ICT tools, methods, and technologies to support smart cities" FIT-S-17-3964. And data used in the research was provided by Secondo BerlinMOD and Geofabrik – German for "Geo factory" projects.

Applying Trusted Knowledge in Evaluation Phase of Data Mining

Viktor Nekvapil

Department of Information and Knowledge Engineering, FIS VSE Praha
nam. W. Churchilla 4
130 67 Praha 3

viktor.nekvapil@vse.cz

Abstract. New concept of Trusted Knowledge (TK) is introduced. Trusted Knowledge are data from trusted organizations such as ministries, statistical offices and so on which can replace domain expert in the evaluation phase of the data mining task. The approach called "A/TK-formulas" enables to filter out resulting patterns which are consequences of Trusted Knowledge and thus enables user to concentrate on interesting ones. Conversely, user can request to show only resulting patterns which are consequences of TK to see which of them are in line with TK. The third option enables to request patterns which are in contradiction to the TK. Further new features of Trusted Knowledge framework are introduced in this paper – Trusted Knowledge for mining histograms and Trusted Knowledge hints.

Key words: Trusted Knowledge, evaluation of data mining.

1 Introduction

The approach presented in this paper incorporates additional knowledge in the evaluation phase of data mining but avoids lengthy and complex task of building a belief system of the user (see e.g. [4], [7], more recently in [2]). The idea is to enhance user's domain knowledge using available trusted sources of data – that is data from trusted organisations such as statistical offices, ministries and so on. I refer to this knowledge as ***Trusted Knowledge***. The Trusted Knowledge Framework has been introduced in [3]. In this paper, new features are presented.

The concept of Trusted Knowledge is inspired by FOFRADAR framework [5]. FOFRADAR is based on a logical calculus of association rules. The interpretation is based on mapping important items of knowledge to the sets of association rules which can be considered as their consequences. Important items of knowledge are expressed using a simple mutual influence among attributes. These are predefined relationships of attributes which are used to determine whether the association rule can be seen as a consequence of the item of knowledge or not. For example, the simple mutual influence (SI-formula) *Income* ↑↑ *Loan* means: "*if Income increases, then Loan increases as well*". The set of atomic consequences of this SI-formula can be expressed by the

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 198-203.

following union: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan}$, saying that "*if Income is high, then Loan is high or if Income is medium then Loan is medium or if Income is high then Loan is high*". Based on the *levels* in the union, it is possible to say whether the resulting rule is a consequence of the defined *SI-formula* or not. This feature is used in the proposed framework and further developed, as obvious in the following sections.

2 Trusted Knowledge

I define Trusted Knowledge as follows: ***Trusted Knowledge*** (TK) is the data from trusted sources which can be connected to the results of a data mining task and are used in the evaluation phase of the data mining task to help with the understanding of the results. Trusted Knowledge can be seen as special case of domain knowledge.

Trusted Knowledge is obtained from a trusted organisation. An example of such knowledge is average and median income per district in the Czech Republic obtained from Czech Statistical Office [1].

Measure of Trusted Knowledge (measure of TK) is a formalised piece of Trusted Knowledge. An example of the measure of TK is depicted in **Table II**. Basic feature of measure of TK is its close connection to the results of a data mining task (resulting patterns). I use association rules as an example. Geographical dimension (locality) is used as a *connecting element* between measure of TK and resulting patterns. *An average income in District X* as a measure of TK and *The loan amount taken by a client in District X* as an attribute from analysed data can be examples of such a connection.

To distinguish between data and *Trusted Knowledge*, I use term *attribute* for variables derived from analysed data and *measure of TK* for variables used as *Trusted Knowledge*. Note that both measure of TK and the attribute connected via *connecting element* are ordinal.

Levels of measures of TK enables us to easily compare attributes and measures of TK. The way how domain experts evaluate the found patterns is commonly expressed by easily interpretable phrases saying for example *"Income is low"*, *"Amount is high"* and so on. Recall the set of atomic consequences of SI-formula *Income* $\uparrow\uparrow$ *Loan*: $Low_{Income} \times Low_{Loan} \cup Medium_{Income} \times Medium_{Loan} \cup High_{Income} \times High_{Loan.}$ Now we have to define, what means for example *"Income is low"* (that is to define the level Low_{Income}).

Expert-based approach means that domain expert decides which category is assigned to each *level*. ***Rank-based approach*** is the newly proposed way of automatic definition of *levels*. Categories of a particular attribute or measure of TK are sorted from the lowest to the highest. Then, we assign rank to each of the category according to the value of attribute or measure of TK. Last step comprises of assigning *Level(l)* to each rank. For example, consider the categories of attribute *Loan_amount* depicted in **Table I**. Based on the rankings of the categories, we can assign respective categories to levels. Having the levels of attributes and measures of TK defined, we can compare levels and draw consequences based on values of the levels. This is further elaborated upon in section 3.

Table I: Levels for attribute Loan amount

Loan_amount	Rank	Level
<0; 100000)	1	Very low
<100000; 150000)	2	Very low
<150000 ;200000)	3	Low
…	…	…
<500000; 550000)	8	High
<550000; 650000)	9	Very high
<650000; 2600000>	0	Very high

Table II: Levels for measure of TK Income

District	Income	Rank	Level
Hlavni mesto Praha	35 115	1	Very high
Stredocesky kraj	27 345	2	Very high
Jihomoravsky kraj	26 116	3	Very high/High
Plzensky kraj	26 026	4	High
…	…	…	…
Pardubicky kraj	24 067	12	Low/Very low
Zlinsky kraj	23 873	13	Very low
Karlovarsky kraj	22 707	14	Very low

2.1 Applying Trusted Knowledge

One of the possible solutions of the automatic formulation of conclusions using domain knowledge is presented in the FOFRADAR framework, as described above.

Using the measures of TK, we can define mutual influence between an attribute and measure of TK. I call this mutual influence ***Attribute / Trusted Knowledge-formula* (*A/TK-formula*)**. The principle of A/TK-formula is the same as for SI-formulas in FOFRADAR, but instead of one of the attributes, measure of TK is used in the mutual influence.

The proposed framework works as follows: After the results are obtained, Trusted Knowledge repository is queried for A/TK-formulas which are available and are relevant for the resulting patterns. Afterwards, A/TK-formulas can be applied. In [3], two ways how the consequences of A/TK-formulas can be applied are presented:

1. to obtain patterns which are consequences of A/TK-formula – this way is useful when the user wants to know which resulting patterns are in line with the overall knowledge (trusted knowledge);
2. to filter out patterns which are consequences of A/TK-formula – this way the user can filter out resulting patterns which are in line with Trusted Knowledge and concentrate on patterns which are not consequences of TK;

In this paper, I introduce the third possible way how the consequences of A/TK-formulas can be applied:

3. to obtain patterns which are in contradiction to the A/TK-formula – this way the user can obtain only rules which are in contradiction to the A/TK formula and concentrate on this resulting patterns (exceptions).

As an example, let us discuss the A/TK-formula *Income* ↑↑ *Loan amount*. *Income* is a measure of TK. Using *rank-based approach*, it is possible to assign values to respective levels, as shown in **Table II**. The categories of the attribute *Loan amount* can be

assigned to the levels, as depicted in **Table I**. Then the set of consequences of the A/TK-formula *Income* ↑↑ *Loan amount* is defined by the following union:

Very low$_{INCOME}$ × *Very low*$_{LOAN}$ ∪ *Low*$_{INCOME}$ × *Low*$_{LOAN}$ ∪ *Medium*$_{INCOME}$ × *Medium*$_{LOAN}$ ∪ *High*$_{INCOME}$ × *High*$_{LOAN}$ ∪ *Very high*$_{INCOME}$ × *Very high*$_{LOAN}$.

To obtain patterns which are in contradiction to the A/TK-formula *Loan amount* ↑↑ *Income* (way 3 above), I modify the union in the following way: *Very low*$_{INCOME}$ × *Very high*$_{LOAN}$ ∪ *Low*$_{INCOME}$ × *High*$_{LOAN}$ ∪ *High*$_{INCOME}$ × *Low*$_{LOAN}$ ∪ *Very high*$_{INCOME}$ × *Very Low*$_{LOAN}$

Note that medium Levels of *Income* and *Loan amount (Medium*$_{INCOME,}$ *Medium*$_{LOAN}$*)* are not present in the union, because they appear together in consequences of A/TK-formula *Loan amount* ↑↑ *Income* and thus they cannot be part of contradictions of A/TK-formula *Loan* amount ↑↑ *Income*.

Main difference between way 2 and 3 is the fact that in 3, we explicitly obtain rules which are in contradiction with the A/TK-formula while in way 2, we obtain rules which are not consequences of A/TK-formula (meaning that also rules with no relation to the A/TK-formula are present).

As an example, let us use the resulting association rule: `District (Zlinsky kraj) -> Loan amount (<100000; 150000)`. Level of *Loan amount* is '*very low'*, connecting element *District* with value *Zlinsky kraj* is used to link the rule to the measures of TK *Income*. If one looks at the level of the measure *Income*, it is *very low* according to **Table II**. So we can conclude that this rule is consequence of the A/TK-formula *Income* ↑↑ *Loan amount* and is not a contradiction of A/TK-formula *Income* ↑↑ *Loan amount.* We can determine the relationship of each rule to the three ways mentioned above.

Further newly defined features of Trusted Knowledge framework include Trusted Knowledge for mining histograms and Trusted Knowledge hints.

Trusted Knowledge for mining histograms

Data mining with histograms has been introduced in [6] using the CF-Miner procedure of the LISp-Miner system. In a simplified manner, the task is to find 'interesting' histograms. Each histogram *Hsg* is in a form *Hsg*(*Attribute*, *Condition*, *Data Matrix*, *Abs/Rel*), where *Condition* is Boolean attribute which each row of the *Data Matrix* must satisfy and *Abs/Rel* states whether the frequencies for *Attribute* are absolute or relative (relative to the overall data matrix without the *Condition*). Furthermore, interestingness measure ≈ called *CF-quantifier* is used to find interesting histograms. For example, a *CF-quantifier* $\approx^{U}_{100,6}$ defines that histogram is interesting if it has at least 100 of objects satisfying *Condition* and there are 6 steps up. That means that 6 consecutive categories of *Attribute* has higher frequency than the previous category.

In [6], domain knowledge is used to filter out resulting histograms which are consequence of defined SI-formula for the *CF-quantifier* ≈ in a similar manner as mentioned above in the FOFRADAR framework. Firstly, a set of atomic consequences of SI-formula for the *CF-quantifier* ≈ needs to be defined. For example, we can define atomic consequences of SI-formula *Price of flat* ↑↑ *Loan amount* for a CF-quantifier $\approx^{U}_{100,6}$ as a set of histograms $\approx^{U}_{100,6}$ *Price of flat / Loan amount(α)* satisfying that level α of *Condition Loan amount* is HIGH or VERY HIGH and the *Attribute Price of*

flat has 7 categories (to ensure that there are only steps up). Levels of *Loan amount* are defined as stated in **Table I**. For example, resulting histogram $\approx^{U}_{100,6}$ *Price of flat / Loan amount (<550000; 650000)* is an atomic consequence of SI-formula *Price of flat* ↑↑ *Loan amount* for *CF-quantifier* $\approx^{U}_{100,6}$. The consequences can be then used for evaluation of the found histograms (for example, filter out histograms which are consequences of SI-formula). The business interpretation of the SI-formula *Price of flat* ↑↑ *Loan* is that if level α of *Loan amount* is HIGH or VERY HIGH, then the level of *Price of flat* will probably also be HIGH or VERY HIGH. Furthermore, agreed consequences are defined in [6], which I do not further discuss here.

Using A/TK-formulas of the Trusted Knowledge framework defined above, we can proceed analogously as in [6] as follows. Let us define atomic consequences of the A/TK-formula *Loan amount* ↑↑ *Income* for *CF-quantifier* $\approx^{U}_{100,9}$ as a set of histograms $\approx^{U}_{100,9}$ *Loan amount / District(α)*, level α being VERY HIGH or HIGH. Then, the resulting histogram $\approx^{U}_{100,9}$ *Loan amount / District(Hlavni mesto Praha)* of relative frequencies is an atomic consequence because the *District Hlavni mesto Praha* in Trusted Knowledge Repository has the level of measure of TK *Income* '*very high*' (see **Table II**). Ways how to apply the consequences of A/TK-formulas for histograms are the same as for association rules mentioned above and are now studied in detail.

Moreover, there could be more flexible ways of applying *CF-quantifiers*. One issue of the *CF-quantifiers* is the fact that they are 'strict' in a sense that if there is one category in a histogram that breaks the overall trend in histogram, the histogram will not be considered as interesting and will not be in resulting histograms. For example, if an attribute in histogram has 6 categories, all but one satisfying the steps up quantifier but the 3rd and 4th category does not satisfy the *CF-quantifier* steps up (but only slightly), the *CF-quantifier* will not be satisfied while from the business perspective, the histogram has the upwards tendency and thus is interesting. Ways how to overcome this issue will be further studied.

Trusted Knowledge hints

Another way how to exploit Trusted Knowledge is to compare a measure of TK and attribute in the analysed data in case they have similar content. For example, it is possible to compare average *Income* presented in the analysed data to the average *Income* as a Trusted Knowledge aggregated to districts. In case the analysed data has sufficient number of objects in each of the groups (groups according to geographical dimension), we can get additional knowledge. For example, we can get following information: The average income of clients in data in district Praha is lower than the average income of people in district Praha of the whole population (as Trusted Knowledge). This can bring us to the deeper investigation of the origin of the data. For example, we can say that in general, affluent clients do not take a consumer loan. If we have a data of clients who took a consumer loan, we can derive that their income will be below the average income in district Praha.

3 Conclusions and future work

First experiments has shown that using all three ways how the consequences of A/TK-formulas can be applied significantly reduce the amount of work user needs to evaluate the resulting rules. This helps the user to concentrate on rules which are interesting from the user's perspective. New features of Trusted Knowledge framework were introduced: Trusted Knowledge for mining histograms and Trusted Knowledge hints. Both features brings new ways of applying Trusted Knowledge.

Another way how to elaborate upon the Trusted Knowledge framework is to study the situation when one pattern is supported by more than one A/TK-formula. Furthermore, different sources of Trusted Knowledge could be evaluated according to their trustworthiness. For example, one source is better than another one from the user's perspective and this information could be further incorporated into the Trusted Knowledge framework. Both features will be further studied.

References

1. Czech Statistical Office (CSO), 2015. Výsledky sčítání lidu, domů a bytů 2011 (Census 2011 – in Czech) [online]. https://www.czso.cz/csu/czso/otevrena_data_pro_vysledky_scitani_lidu_domu_a_bytu_2011_-sldb_2011- Last modified on 14 th April 2015.
2. De Bie, T., 2013. Subjective interestingness in exploratory data mining. In Advances in Intelligent Data Analysis XII: 12th International Symposium, IDA 2013, London, UK, October 17-19, 2013.
3. Nekvapil, V. 2017. Data Mining with Trusted Knowledge. FedCSIS Conference, Prague. 3-6 September 2017. Accepted for publication.
4. Padmanabhan, B., Tuzhilin, A., 1998. A belief-driven method for discovering unexpected patterns. In Proc. of the 4th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 94-100, 1998.
5. Rauch, J., 2015. Formal Framework for Data Mining with Association Rules and Domain Knowledge – Overview of an Approach. Fundamenta Informaticae, 137 No 2, pp. 1–47
6. Rauch, Jan, Šimůnek, Milan. Data Mining with Histograms – A Case Study. In: *Foundations of Intelligent Systems* [online]. Lyon, 21.10.2015 – 23.10.2015. Cham : Springer International Publishing, 2015, s. 3–8. ISBN 978-3-319-25251-3. DOI: 10.1007/978-3-319-25252-0.
7. Silberschatz, A., Tuzhilin, A., 1995. On subjective measures of interestingness in knowledge discovery. In Proc. of the 1st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 275-281, 1995.

Acknowledgment: The work described here has been supported by the internal grant agency of the University of Economics, Prague under project IGA 29/2016.

Učenie s prenosom medzi prirodzenými jazykmi

Matúš Pikuliak, Marián Šimko a Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva,
Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

matus.pikuliak@stuba.sk

Abstrakt. Hlboké učenie sa aktuálne javí ako veľmi perspektívny prístup k mnohým úlohám spracovania prirodzeného jazyka. Tento úspech sa však zatiaľ prejavuje najmä pri jazykoch, ktoré majú dostatočné množstvo zdrojov na natrénovanie komplexných neurónových sietí. Menšie jazyky s menším množstvom zdrojov majú problém tieto nové techniky využiť a priepasť medzi nimi a "bohatými" jazykmi sa tak prehlbuje. V našej práci sa venujeme tomu, ako túto priepasť zmenšiť pomocou prenosu naučenej informácie z jazyka do jazyka. Hlavnou myšlienkou je trénovanie hlbokých neurónových sietí v multilingválnom režime tak, aby sa model naučil využívať znalosti z jedného jazyka aj pre vstupy z iných. Navrhli a vykonali sme experiment s prenosom informácie o sentimente slov pomocou zdieľaného priestoru distribučných vektorov. V experimente sme dosiahli výsledky porovnateľné s nemeckými manuálne vytvorenými lexikónmi sentimentu, pričom sme však nepoužili žiadne nemecké dáta týkajúce sa sentimentu.

Kľúčové slova: spracovanie prirodzeného jazyka, učenie s prenosom, multilingválne učenie, umelé neurónové siete.

1 Úvod

Umelé neurónové siete v posledných rokoch výrazným spôsobom zasiahli do spracovania prirodzeného jazyka. Ukázalo sa, že rozličné architektúry neurónových sietí dokážu veľmi dobre spracovať veľké objemy textových dát a naučiť sa z nich prínosné informácie. V krátkom čase pomocou nich výskumníci dokázali prekonať existujúce riešenia pre veľké množstvo úloh, vrátane tých najpokročilejších, ako napríklad strojový preklad [3], analýza sentimentu [10] či syntaktická analýza [13]. Výhodou týchto neurónových prístupov sú malé nároky na doménovú expertízu. Zdá sa, že strojové učenie sa tu konečne dokáže naučiť všetko, čo potrebuje na úspešné vyriešenie úlohy, bez toho, aby výskumníci museli ručne navrhovať vysoký počet čŕt. Črty sa neurónové siete učia extrahovať samé a môžeme teda viac všeobecne hovoriť o učení sa reprezentácií, resp. učení sa čŕt (angl. *representation learning*) [1].

Tieto zaujímavé pokroky sú podmienené rastúcim množstvom textových dát, ktoré ako ľudstvo vytvárame. Najmä s nástupom Internetu sa zlepšili možnosti tvorby

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 204-208.

a zdieľania rozličných textov, ako napr. recenzií, príspevkov na sociálnych sieťach, blogov či článkov. Aj keď tento rast je globálny, niekoľko svetových jazykov (najmä angličtina) ho pocítilo omnoho viac. Len v týchto jazykoch existujú dostatočne veľké datasety, na ktorých sa danú súčasné modely často postavené na hlbokom učení (angl. *deep learning*) trénovať.

Jedným z možných riešení je pokúsiť sa prenášať informácie naprieč jazykmi. V strojovom učení poznáme koncept učenia s prenosom (angl. *transfer learning*), kedy sa snažíme model natrénovaný na určitej doméne či úlohe použiť v iných doménach či na iných úlohách [8]. Prenos medzi rozličnými jazykmi môžeme chápať ako jeden z druhov takéhoto učenia a v tejto práci sa venujeme práve jemu.

2 Existujúce prístupy

Pri analýze existujúcich prístupov ku učeniu s prenosom medzi jazykmi sme identifikovali 3 lingvistické úrovne, na ktorých môže učenie medzi jazykmi prebiehať. Ide o (1) morfologickú úroveň, (2) úroveň slov a (3) úroveň viet. Postupne tieto úrovne predstavíme.

Morfologická úroveň. Na morfologickej úrovni prebieha učenie najmä medzi lingvisticky podobnými jazykmi, ktoré teda do istej miery zdieľajú podobnú morfológiu a slovnú zásobu (napr. Slovenčina-čeština, španielčina-portugalčina). Takéto učenie sa dá aplikovať pri modeloch, ktoré pracujú na úrovni znakov alebo iných častí slov. Predpokladáme, že znalosti získané o morfológii jedného jazyka sú aplikovateľné aj pre iné jazyky. Takéto učenie bolo použité napr. pri rozpoznávaní pomenovaných entít [2] alebo pri značkovaní viet [12].

Úroveň slov. V spracovaní prirodzeného jazyka sa stali populárne tzv. vektory latentných čŕt slov (angl. *word embeddings*) [6]. Ide o reprezentácie slov vytvorené v skrytých vrstvách neurónových sietí, ktoré zachytávajú sémantiku slov. Takto vznikajú aj multilingválne reprezentácie, kde sú do jedného vektorového priestoru premietané slová z viacerých jazykov [11]. Toto potom môže byť použité pre prenos informácie medzi jazykmi. Náš experiment patrí do tejto kategórie.

Úroveň viet. Posledná a v istom zmysle najnáročnejšia úroveň je úroveň viet. Na tejto úrovni sa výskumníci snažia vytvárať reprezentácie viet tak, aby zachovávali ich sémantickú podobnosť naprieč jazykmi. Táto úroveň je najmenej preskúmaná, riešenia existujú napríklad pre strojový preklad [3] alebo analýzu sentimentu [14].

3 Experiment s analýzou sentimentu

Na základe analýzy tejto oblasti sme navrhli náš vlastný experiment, ktorého úlohou je preniesť informáciu o sentimente slov z jedného jazyka do druhého. Na analýzu sentimentu sa často používajú tzv. lexikóny sentimentu. V podstate ide o zoznam ohodnotených slov, kde hodnota značí či je dané slovo pozitívne alebo negatívne ladené. Zostavovanie takýchto lexikónov je prácna úloha a automatické prístupy ani zďaleka nedosahujú potrebnú kvalitu [9]. Našou základnou myšlienkou je využiť informáciu z kvalitného cudzojazyčného lexikónu. Pre účely experimentu sme sa

zamerali na anglické lexikóny, ktoré chceme zužitkovať pre analýzu sentimentu v nemeckom jazyku.

Pre prenos sentimentu používame tzv. multilingválne vektory latentných čŕt [11]. Ide o vektorový priestor, v ktorom sú reprezentované slová z viacerých jazykov (v našom prípade z angličtiny a nemčiny) tak, že je zachovaná ich sémantická podobnosť. Anglické slova *good* a *fine* sú v takomto priestore blízko pri sebe. Taktiež sa však v ich blízkosti nachádzajú aj nemecké *gut* a *fein*. Nad takýmto multilingválnym priestorom potom trénujeme model sentimentu, ktorý sa naučí klasifikovať slová na pozitívne, negatívne a neutrálne.

Tento model je jednoduchá neurónová sieť (viacvrstvový perceptrón) s jednou skrytou vrstvou, ktorá sa snaží z vektorovej reprezentácie slova vyrátať jeho hodnotenie. Vstupom do tejto neurónovej siete sú predpripravené vektory slov, zatiaľ čo výstupom je vektor s 3 položkami, pričom každá zodpovedá jednej triede sentimentu. Ako trénovaciu množinu pre tento model použijeme kvalitný anglický lexikón sentimentu. Na výsledný vektor sa aplikuje softmaxová funkcia. Ako stratovú funkciu používame cross entropy funkciu, trénovanie prebieha s použitím bežného algoritmu Stochastic Gradient Descent. Natrénovaný model potom môžeme použiť pre klasifikáciu sentimentu nemeckých slov, keďže ich máme k dispozícii ich reprezentácie zo zdieľaného anglicko-nemeckého vektorového priestoru slov.

Pri experimentoch sme vyskúšali použiť viacero alternatívnych architektúr – testovali sme vplyv počtu skrytých vrstiev aj ich šírky. Napokon sa ako najlepšie riešenie ukázala sieť s jednou skrytou vrstvou s dĺžkou 10. Takáto neurónová sieť dokázala prekonať lineárnu aj logistickú regresiu. Myslíme si, že je spôsobené nelinearitou neurónových sietí. Sentiment vo vektorovom priestore totižto nemusí byť lineárne separovateľný. Správanie slov a ich vlastností v takomto priestore môže byť jedným zo smerov ďalšieho výskumu.

Takýto model sa naučí na anglických slovách odhadovať ich sentiment. Keďže je však zachovaná sémantická príbuznosť, dokážeme tento model rovno použiť aj na ohodnocovanie nemeckých slov. My sme model použili na ohodnotenie každého nemeckého slova v priestore. Získaný lexikón sentimentu pre nemecký jazyk sme potom použili v štandardnej klasifikácii na analýzu sentimentu nemeckých viet. Z každej vety sme vyextrahovali určité črty (prítomnosť slov, emotikonov, štylistických prvkov) a navyše sme pridali niekoľko čŕt extrahovaných pomocou lexikónu (počet pozitívnych/negatívnych slov, priemerná hodnota sentimentu). Nad výsledným vektorom čŕt sme natrénovali SVM klasifikátor.

Vo výsledku náš plne automatizovaný prístup presvedčivo predbehol iné automatizované prístupy. Ako nemecké lexikóny sme použili tie vygenerované v [9]. Automaticky vytvorený lexikón TKM napríklad dosiahol súhrnné makro F1 skóre 0,75 pre binárnu, resp. 0,59 pre ternárnu klasifikáciu. Náš prístup dosiahol najlepšie skóre s anglickým lexikónom sentimentu NRC [7]: 0,79, resp. 0,61. Tento náš výsledok je porovnateľný s manuálne vytvorenými state-of-the-art lexikónmi pre nemecký jazyk. Najlepší z nich, GPC, dosahuje výsledky 0,80, resp. 0,62. Vidíme, že sa s pomerne jednoduchou metódou dokážeme priblížiť kvalitným, manuálne vytvoreným lexikónom. Potrebovali sme pritom len paralelný anglicko-nemecký korpus a anglický lexikón sentimentu. Namiesto rátania reprezentácii slov v multilingválnom priestore

z paralelného korpusu sme v našich experimentoch použili už hotový anglicko-nemecký vektorový priestor [5], ktorý obsahuje 95,195 nemeckých a 40,054 anglických slov. Dĺžka vektorov je 512.

4 Záver

Učenie s prenosom medzi jazykmi má potenciál priniesť kvalitné riešenia pre jazyky, ktoré nemajú dostatok zdrojov. Tento smer je obzvlášť aktuálny teraz, keď v spracovaní prirodzeného jazyka sa do popredia dostáva hlboké učenie. Reprezentácie, ktoré sa pri hlbokom učení vytvárajú, sú totižto obzvlášť vhodné na učenie s prenosom. V tejto práci sme predstavili problematiku učenia s prenosom medzi jazykmi a uviedli sme aj tri úrovne, na ktorých takéto učenie môže prebiehať. Na úrovni slov sme ho aplikovali aj pri našom experimente, kedy sa nám podarilo automaticky zostrojiť kvalitný lexikón sentimentu pre nemecký jazyk.

V ďalšej práci chceme nadviazať na tento experiment a vylepšiť metódu prenosu informácie medzi slovami. Taktiež sa chceme venovať prenosu s učením na úrovni viet, ktoré predpokladá vytváranie reprezentácií s pokročilými metódami hlbokého učenia. Pri tomto učení sa chceme zamerať aj na interpretáciu natrénovaných modelov a lepšie pochopenie toho, čo sa vlastne umelé neurónové siete pri spracovaní prirodzeného jazyka učia.

Literatúra

1. Bengio, Y., Courville, A., Vincent, P.: Representation Learning: A Review and New Perspectives. IEEE Trans. on Pattern Analysis & Machine Intelligence 35.8 (2013), 1798 1828.
2. Gillick, D., et al.: Multilingual Language Processing From Bytes. In: Proc. of the 2016 Conf. of the North American Chapter of the ACL. ACL, San Diego (2016), 1296 1306.
3. Johnson, M., et al.: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. arXiv preprint arXiv:1611.04558 (2016).
4. Levy, O., Sogaard, A., Goldberg, Y.: A Strong Baseline for Learning Cross-Lingual Word Embeddings from Sentence Alignments. In: Proc. of the 15th Conf. of the European Chapter of the ACL 1. ACL, Valencia (2017), 765 774.
5. Luong, T., Pham, H., Manning, C.D.: Bilingual word representations with monolingual quality in mind. In Proc. of the 1st Workshop on Vector Space Modeling for NLP. ACL, Berlin (2015), 151 159.
6. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems 26, (2013), 3111-3119.
7. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word–emotion association lexicon. Computational Intelligence 29.3 (2013), 436 465.
8. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22.10 (2010), 1345 1359.
9. Sidarenka, U., Stede, M.: Generating Sentiment Lexicons for German Twitter. In: Proc. of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media. The COLING 2016 Organizing Committee, Osaka (2016), 80 90.

10. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing.ACL, Seattle (2013), 1631 1642.
11. Upadhyay, S., et al.: Cross-lingual Models of Word Embeddings: An Empirical Comparison. In: Proc. of the 54th Annual Meeting of the ACL 1. ACL, Berlin (2016), 1661 1670.
12. Yang, Z., Salakhutdinov, Z., Cohen,WW.: Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks. In: 5th Int. Conf. on Learning Representations. (2017).
13. Zhou, H., et al.: A Neural Probabilistic Structured-Prediction Model for Transition-Based Dependency Parsing. In: Proc. of the 53rd Annual Meeting of the ACL 1. ACL, Beijing (2015), 1213 1222.
14. Zhou, X., Wan, X., Xiao, J.: Cross-Lingual Sentiment Classification with Bilingual Document Representation Learning. In: Proc. of the 54th Annual Meeting of the ACL 1. ACL, Berlin (2016), 1403 1412.

Poďakovanie: Tento článok vznikol vďaka čiastočnej podpore projektov VEGA 1/0646/15 a KEGA 009STU-4/2014.

Annotation:

Transfer Learning between Natural Languages

Deep learning is currently a popular approach to natural language processing. It is however very data-demanding and languages without proper resources can not utilize it fully. Transfer learning between languages tackles this problem as it transfer trained models from one language to another. We discuss such learning here and we also present our own experiment concerned with word-level induction of sentiment from English to German.

Smerom k automatickej detekcii problémov s použiteľnosťou prostredníctvom sledovania pohľadu

Martin Svrček a Mária Bieliková

Ústav informatiky, informačných systémov a softvérového inžinierstva,
Fakulta informatiky a informačných technológií, Slovenská technická univerzita v Bratislave
Ilkovičova 2, 842 16 Bratislava

{martin.svrcek, maria.bielikova}@stuba.sk

Abstrakt. Rozličné anomálie používateľského rozhrania aplikácií a webových stránok sú často nazývané problémy použiteľnosti. V súčastnosti existuje v tejto oblasti mnoho štúdií, ktoré sa zaoberajú detekciou problémov s použiteľnosťou softvéru. Avšak, existujúce metódy sa snažia automatizovať tento proces alebo jeho časti použitím výlučne používateľských aktivít vo forme klikov myši alebo vstupu z klávesnice pri kvantitatívnych štúdiách z online správania sa používateľov. Naším cieľom je analyzovať pohľad používateľa s cieľom odhalenia takých pohybov očí, ktoré môžu byť použité pre detekciu problémov s použiteľnosťou. Opisujeme používateľskú štúdiu so sledovaním pohľadu, ktorú sme vykonali. Porovnávame metriky pohľadu medzi účastníkmi, ktorí pracovali s webovou stránkou bez zavedených problémov s použiteľnosťou a účastníkmi, ktorí pracovali s rovnakou webovou stránku avšak so zavedenými problémami s použiteľnosťou. Cieľom bolo identifikovať tie metriky pohľadu, ktoré dokážu najlepšie odhaliť tieto problémy. Odhalené metriky môžeme použiť pre ďalšiu analýzu, kde sa chceme pozerať na pohľad ako na sekvenciu pohybov očí. Takéto sekvencie sú v porovnaní so štandardnými prístupmi schopné odhaliť komplexnejšie vzory v dátach o používaní webu, čo umožní lepšiu identifikáciu problémov s použiteľnosťou.

Kľúčové slová: problémy použiteľnosti, používateľský zážitok, UX, sledovanie pohľadu, sekvencie pohybov očí, metriky pohľadu

1 Úvod

V súčasnosti je používateľský zážitok (angl. user experience, UX) veľmi dôležitou a populárnou oblasťou. Vyjadruje pocity používateľa pri používaní systému a zároveň významne ovplyvňuje samotnú použiteľnosť. V kontexte Webu existuje množstvo prác, ktoré sú súčasťou výskumu v oblasti použiteľnosti [3][7][8][9][11]. Z dôvodu širokého uplatnenia použiteľnosti v mnohých disciplínach neexistuje jedna všeobecná definícia vyhovujúca všetkým týmto oblastiam. Jednou z najvhodnejších definícií je práve ISO definícia použiteľnosti, ktorá hovorí, že použiteľnosť sú vnemy a odozvy osôb, ktoré vznikajú pri používaní produktu, systému alebo služby [11].

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 209-214.

Ak však nevieme používať produkt, ak nemôžeme nájsť položky, ktoré hľadáme alebo iba nevieme dosiahnuť naše ciele tak budú tieto vnemy skôr negatívne. V tomto prípade môžeme hovoriť o problémoch s použiteľnosťou, ktoré nám zabraňujú v dokončení našej aktivity alebo robia túto aktivitu náročnejšou a zdĺhavejšou. Samotný problém použiteľnosti je aspekt systému, ktorý spôsobuje, že tento systém je nepríjemný, neefektívny na použitie alebo dokonca úplne zabraňuje jeho použitiu v kontexte dosahovania našich cieľov [10].

Hlavnou úlohou štúdií použiteľnosti webových stránok alebo aplikácií je detegovať tieto problémy. Podobné štúdie zvyčajne vykonávajú UX experti, ktorí odhaľujú tieto problémy ich manuálnym vyhodnocovaním, čo je často pomerne zdĺhavý proces.

Cieľom viacerých prác v tejto oblasti je automatizovať tento proces. V súčasnosti už existuje viacero podobných prístupov avšak tieto používajú iba záznamy v podobe klikov myši alebo vstupu z klávesnice [4][6][5][7][12]. Avšak my veríme, že tento proces detekcie problémov s použiteľnosťou môže byť zlepšený zavedením metrík sledovania pohľadu čo dokazujú aj niektoré práce v tejto oblasti [2].

Naším cieľom je v prvom kroku analyzovať pohľad používateľa a odhaliť charakteristické pohyby očí v kontexte metrík pohľadu, ktoré dokážu odhaľovať konkrétne problémy s použiteľnosťou a zlepšiť tak proces ich identifikácie. Ako súčasť našej práce sme vytvorili zoznam 53 problémov s použiteľnosťou na základe dostupnej literatúry, keďže takto ucelený zoznam nebol dostupný.

2 Automatická identifikácia problémov s použiteľnosťou

Kliknutia myši, skrolovanie alebo vstup z klávesnice dokážu poskytnúť množstvo dôležitých informácií pre identifikáciu anomálií v správaní sa používateľa. Avšak, ak hovoríme o vizuálnej pozornosti, tak neexistuje žiadna lepšia metóda ako sledovanie pohľadu [12]. Z dostupnej literatúry je možné vidieť, že existuje viacero rôznych prístupov k automatickej detekcii problémov s použiteľnosťou využívajúce reverzné inžinierstvo, strom úloh alebo práce vychádzajúce z pachov zdrojového kódu.

V týchto prácach vidíme rozdiel aj v kontexte sledovaných problémov s použiteľnosťou, kedy sa každá z prác zameriava na odlišné problémy. V prvej štúdii [6] sa autori zamerali na automatickú detekciu 16 problémov s použiteľnosťou ako sú napríklad: element nie je samo opisný, zavádzajúci odkaz, stránka bez odozvy, vzdialený obsah.

Ďalšie práce využívali odlišné prístupy pre detekciu. Snažili sa napríklad detegovať problémy s použiteľnosťou na základe pachov zdrojového kódu [1]. Pri ich práci využívali celkovo 6 problémov s použiteľnosťou, ktoré boli viazané na implementačné aspekty systému. Výhodou takýchto prístupov vychádzajúcich priamo zo zdrojového kódu je určite lepšie prepojenie problémov s použiteľnosťou a samotnej implementácie. Toto nám umožňuje a zjednodušuje automatickú refaktorizáciu odhalených problémov.

Niektoré z prác na druhej strane vychádzali z konceptu reverzného inžinierstva [12], kedy bol základom pre detekciu znovu zdrojový kód a rovnako aj rôzne modely pomáhajúce samotnej detekcii. Ďalším veľmi zaujímavým prístupom je využitie

stromov úloh [7]. V rámci tohto prístupu sú jednotlivé akcie používateľov transformované do stromu úloh. V ďalšom kroku je tento strom skenovaný s cieľom odhalenia určitých vzorov prepojených na problémy s použiteľnosťou.

Okrem podobných prístupov využívajúcich klasické akcie používateľov je veľmi zaujímavé aj sledovanie pohľadu [2]. V rámci tohto prístupu autori vychádzajú z predpokladu, že experti svojimi odhadmi poskytujú iba určitý prehľad a nevedia určiť presné prepojenie medzi metrikami a problémami s použiteľnosťou.

Vo vykonanom experimente [2] boli zozbierané dáta z pozorovania ako aj zo sledovania pohľadu od 19 účastníkov a zisťovali sa korelácie medzi týmito dvoma zdrojmi dát. Autori na konci poskytujú sumarizáciu sledovaných problémov s použiteľnosťou a s nimi súvisiacich metrík sledovania pohľadu. Avšak aj na základe ich výstupov je zrejmé, že kombinácia rôznych vzorov pohľadu musí byť skúmaná v kontexte správania sa používateľov viac do hĺbky.

3 Potenciál sledovania pohľadu pre detekciu problémov s použiteľnosťou

V rámci cieľu našej práce bolo potrebné preskúmať potenciál sledovania pohľadu pre detekciu problémov s použiteľnosťou a preto sme sa rozhodli vykonať experiment so sledovaním pohľadu. Experimentu sa zúčastnilo 10 ľudí. Išlo o 8 mužov a 2 ženy. Všetci účastníci boli vysokoškolskí študenti. Pre účely testovania sme využili webovú stránku obsahujúcu novinové články. Na tejto stránke mali účastníci za úlohu vykonať nasledujúce tri úlohy:

- Úloha 1: Registrácia sa do systému
- Úloha 2: Nájdenie špecifického článku
- Úloha 3: Aktualizácia informácií v profile

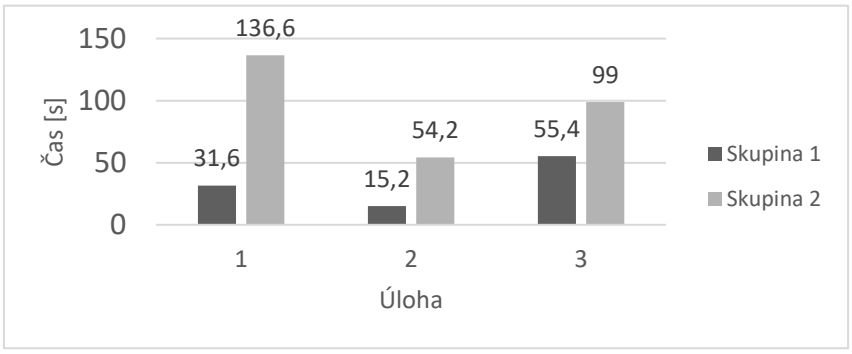
Počas celého experimentu sme zaznamenávali ich pohľad prostredníctvom zariadenia na sledovanie pohľadu s frekvenciou 300 Hz, ktorý je dostupný vo Výskumnom centre používateľského zážitku a interakcie (http://uxi.sk) na Fakulte informatiky a informačných technológií STU v Bratislave.

Pre účely porovnania výsledkov vykonania úloh sme sa rozhodli rozdeliť účastníkov experimentu do dvoch skupín. Skupina 1 pracovala s webovou stránkou, ktoré neobsahovala problémy s použiteľnosťou. Skupina 2 pracovala s webovou stránkou, ktoré obsahovala problémy s použiteľnosťou. V rámci nášho experimentu sme vybrali 8 problémov, a tieto sme manuálne zaviedli do našej webovej stránky:

- Dlhý registračný formulár
- Krátke zadávacie pole
- Nedôležité zadávacie pole
- Navigácia obsahujúca príliš veľa položiek
- Mätúce názvy položiek navigácie
- Zlá pôvodná hodnota
- Chýbajúce vysvetlenie chýb

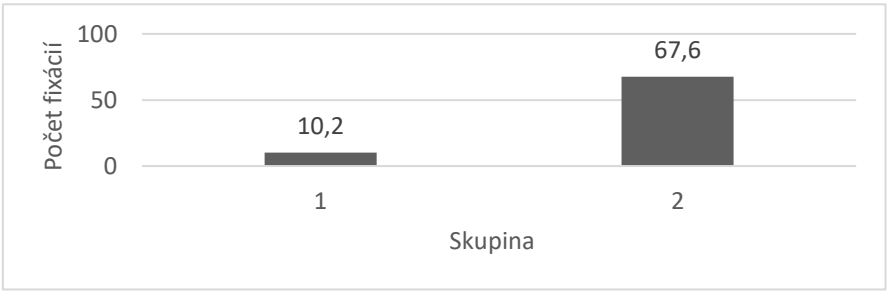
– Chýbajúca klientská validácia

Týchto 8 problémov sme vybrali s ohľadom na možnosť porovnania správania sa používateľov pri odlišných problémoch s použiteľnosťou. Našou hypotézou bolo, že existujú konkrétne metriky pohľadu, ktoré dokážu identifikovať konkrétne problémy s použiteľnosťou. Na základe analýzy získaných dát z experimentov sme odhalili viaceré odlišnosti. Na obrázku 1 je vidieť, že výrazný rozdiel bol pri všetkých úlohách v rámci dĺžky trvania úloh. Pri úlohe 1 bol tento rozdiel najväčší z dôvodu náročného formulára na vyplnenie.



Obrázok 1: Priemerný čas trvania úloh

Rovnaký vzor však môžeme pozorovať aj na metrikách pohľadu. Napríklad pri úlohe 2 (Obrázok 02) sa ukázal veľký rozdiel medzi počtom fixácií pohľadu, kde skupina so zavedenými problémami s použiteľnosťou potrebovala sledovať položky navigácie viackrát (priemerne 10.2 fixácií pre skupinu 1 vs. 67.6 fixácií pre skupinu 2). Avšak metrika dĺžky fixácií sa neukázala ako vhodná pre identifikovanie problému súvisiaceho s navigáciou, keďže rozdiel medzi dĺžkou fixácií bol iba minimálny.



Obrázok 2: Počet fixácií pri úlohe 2

Podobné výsledky sme dosiahli aj pri ostatných problémoch, a preto môžeme usudzovať, že existuje predpoklad, že rôzne problémy s použiteľnosťou je možné odhaliť

odlišnými metrikami. Avšak pre kvalitnejšiu validáciu týchto dosiahnutých výsledkov by sme potrebovali vykonať experiment s väčším počtom účastníkov.

4 Zhodnotenie

Prítomnosť zvolených problémov s použiteľnosťou v našom experimente bola overovaná na základe troch metrík (celková doba fixácie, počet fixácií a čas splnenia úlohy). Experimenty ukázali, že samotná dĺžka fixácií nie je vhodná pre odhaľovanie zvolených problémov. Avšak tieto výsledky sú iba prvými výstupmi v rámci našej práce. V kontexte ďalšej práce sa ukázalo ako odlišné sú jednotlivé problémy, a preto sa chceme zamerať na konkrétnu problematiku (napr. navigácia) a preskúmať iba problémy súvisiace s touto oblasťou.

Z našich pozorovaní ako aj z pozorovaní iných prác [2] je zrejmé, že samotné metriky pohľadu a ich vplyv musí byť skúmaný viac do hĺbky. Rovnako si uvedomujeme, že problémy s použiteľnosťou nie sú prepojené iba na jednu jednoduchú metriku pohľadu ale na špecifickejšie sekvencie týchto vzorov. Tu sa dostávame k sekvenciám pohybu očí (angl. scanpath) a ich rôznym interpretáciám.

V rámci tejto problematiky existuje viacero možných reprezentácií týchto sekvencií (počet fixácií ,dĺžka fixácie, sakády, atď.). Práve experiment, ktorý sme vykonali nám môže poskytnúť základné informácie potrebné pre ďalší výskum v kontexte voľby vhodnej reprezentácie pre konkrétny problém s použiteľnosťou.

Sekvencie pohybov očí poskytujú pohľad na dáta zo sledovania pohľadu spolu s kontextom, v ktorom sa vykonali. Preto práve odhaľovanie problémov s použiteľnosťou môže byť výrazne obohatené, keďže tieto problémy môžu byť často súčasťou určitého reťazca aktivít. Veríme, že odhaľovanie týchto reťazcov a vzorov aktivít zlepší presnosť identifikácie problémov s použiteľnosťou softvéru.

Literatúra

1. Almeida, D., Campos, J. C., Saraiva, J., & Silva, J. C. (2015). Towards catalog of usability smells. *ACM Symposium on Applied Computing* (s. 175-181). ACM.
2. Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. *People and Computers: HCI... but not as we know it. 1*, s. 119-128. British Computer Society.
3. Falkowska, J., Sobecki, J., & Pietrzak, M. (2016). Eye tracking Usability Testing enhanced with EEG Analysis. *International Conference of Design, User Experience, and Usability* (s. 399-411). Springer International Publishing.
4. Ganesh, S. G., Sharma, T., & Suryanarayana, G. (2013). Towards a Principle-based Classification of Structural Design Smells. *Journal of Object Technology*, 1-1.
5. Grigera, J., Garrido, A., & Rivero, J. M. (2014). A tool for detecting bad usability smells in an automatic way. *International Conference on Web Engineering.* Springer International Publishing.
6. Grigera, J., Garrido, A., Rivero, J. M., & Rossi, G. (2017). Automatic detection of usability smells in web applications. *International Journal of Human-Computer Studies*, 129-148.

7. Harms, P., & Grabowski, J. (2014). Usage-based automatic detection of usability smells. *Human-Centred Software Engineering* (s. 217-234). Springer Berlin Heidelberg.
8. Hertzum, M. (2016). Usability testin: too early? too much talking? too many problems? *Journal of Usability Studies*, 83-88.
9. Hlaváč, P. (2016). Impact of characteristics of individuals on evaluating the quantitative studies. *UMAP*.
10. Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, 246-266.
11. Law, E. C., Roto, V., Hassenzahl, M., Vermeeren, A. P., & Kort, J. (2009). Understanding, scoping and defining user experience: a survey approach. *Human factors in computing systems* (s. 719-728). ACM.
12. Silva, J. C., Campos, J. C., Saraiva, J., & Silva, J. L. (2014). An approach for graphical user interface external bad smells detection. In *New Perspectives in Information Systems and Technologies* (Zv. 2, s. 199-205). Springer International Publishing.
13. Xu, P., Sugano, Y., & Bulling, A. (2016). Spatio-temporal modeling and prediction of visual attention in graphical user interfaces. *Human Factors in Computing Systems* (s. 3299-3310). ACM.

Poďakovanie: Tento článok vznikol vďaka čiastočnej podpore projektu APVV-15-0508 a vďaka podpore v rámci projektu „Rozvoj výskumnej infraštruktúry STU", projekt č. 003STU-2-3/2016 zo zdrojov štátneho rozpočtu prostredníctvom dotácie z Ministerstva školstva, vedy, výskumu a športu Slovenskej republiky.

Annotation:

Towards Automatic Detection of Usability Problems Using Eye-Tracking

Various anomalies of the user interface design of applications and web pages are also called usability problems. There is a number of studies in this field, which deal with detection of so called problems. However, existing methods try to automate this detection process using solely user activity in the form of clicks or keyboard input. Our aim is to analyze gaze of the users to determine the eye movement metrics that can be used in future work to detect usability problems. We performed user study with eye-tracking. First group of participants worked with a web page that did not contain usability problems and the second group worked with the same web page but with usability problems. We compared each gaze metrics between these two groups in order to identify gaze characteristics which can reveal usability problems. These characteristics can be than used as a start point for next analysis, where we want to look at the gaze as a sequence of eye movement. These sequences are called scanpaths and in comparison with standard approach, they can reveal more complex patterns of user behavior in order to identification of usability problems.

Použitie spracovaných záznamov reči pacientov pre určenie štádia Parkinsonovej choroby

Michal Vadovský, Ján Paralič

Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00 Košice, Slovenská republika

michal.vadovsky@tuke.sk, jan.paralic@tuke.sk

Abstrakt. Lekársky postup diagnostikovania určitej choroby u pacientov je časovo zdĺhavý a veľmi náročný. Metódy dolovania v dátach môžu tento proces urýchliť a pomôcť tak lekárom pri rozhodovaní v zložitých situáciách. V prípade Parkinsonovej choroby (PCH) je najväčším problémom diagnostika prvotného štádia, pretože symptómy nie sú tak jednoznačné a ľahko pozorovateľné. Preto sme sa v tomto článku zamerali na určenie štádia PCH z dát zaznamenávajúcich rečové signály pacientov pomocou rozhodovacích stromov (C4.5, C5.0, CART). S cieľom zlepšenia klasifikačných modelov sme použili aj metódy RandomForest, Bagging a Boosting. Odhad presnosti modelov bol realizovaný použitím k-násobnej krížovej validácie a validácie s vynechaním jedného záznamu (Leave-one-out). Okrem toho sme vykonali aj experimenty s odstránením kolinearity v dátach vypočítaním inflačného faktoru rozptylu (VIF) za účelom zvýšenia presností modelov.

Kľúčové slová: štádium Parkinsonovej choroby, reč, dolovanie v dátach

1 Úvod

Parkinsonova choroba (PCH) [1] je veľmi vážne neurologické ochorenie, na ktoré dodnes neexistuje žiaden liek. Hlavnou príčinou vzniku ochorenia je odumieranie nervových buniek, ktoré produkujú v mozgu dôležitú chemickú látku s názvom dopamín [2]. Medzi prvotné príznaky u ľudí trpiacich PCH patrí stuhnutie svalstva, problémy s rečou (dysfónia), pohybom alebo písaním (dysgrafia).

Pre meranie štádia ochorenia sa vytvorila jednotná škála hodnotenia PCH s názvom UPDRS (Unified Parkinson´s Disease Rating Scale) [3], ktorej celkové skóre je získané vyhodnotením dotazníka skladajúceho sa zo štyroch častí: I. myslenie, správanie a nálada; II. aktivity bežného života, III. vyšetrovanie hybnosti; IV. komplikácie liečby v poslednom týždni. Na základe výsledkov UPDRS je možné pomocou modifikovanej stupnice štádia (podľa Hoehnonovej a Yahra) rozdeliť pacientov do 8 štádií PCH: 0 (bez príznakov ochorenia), 1 (jednostranné príznaky), 1.5 (jednostranné a axiálne postihnutie), 2 (obojstranné postihnutie bez poruchy rovnováhy), 2.5 (obojstranné postihnutie s miernou poruchou rovnováhy, schopnosť vyrov-

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 215-220.

nať postoj), 3 (mierne až stredné obojstranné postihnutie, sebestačný), 4 (ťažká nespôsobilosť, schopný chodiť alebo stáť bez pomoci), 5 (odkázaný na vozík alebo pripútaný k lôžku, vstávanie s pomocou).

V našom článku sme sa zamerali na určenie štádia PCH pomocou transformovaných dát zo zvukových záznamov reči. Diagnostika PCH v skorom štádiu je veľmi náročná, pretože príznaky sú nejednoznačné a ťažšie rozpoznateľné. Preto sme sa v tomto článku snažili zistiť, akú najvyššiu presnosť dokážeme získať pomocou modelov klasifikujúcich záznamy pacientov do 8 rôznych tried. V prípade získania vysokých presností by mohli byť tieto modely implementované do systémov na podporu diagnostiky PCH pre lekárov.

2 Prehľad súčasného stavu

Keďže je PCH veľmi časté ochorenie a stále na ňu neexistuje liek, tak sa množstvo výskumníkov zameriava na diagnostiku tohto ochorenia priamo z prvotných symptómov, napr. reč alebo písmo. V publikácii [4] sa G. Yadav s kol. zameriavali na reč pacientov, kde pre vytvorenie modelov pre klasifikáciu do dvoch tried (1 – pacient s PCH, 0 – zdravý pacient) využil 3 metódy dolovania v dátach. Pre ich porovnanie a vyhodnotenie použili 10-násobnú krížovú validáciu, pričom najlepšie výsledky dosiahli pomocou SVM (76%), nasledovala metóda rozhodovacích stromov (75%) a logistická regresia (64%). A. Tsanas s kol. [5] sa venovali predikovaniu numerickej hodnoty UPDRS (0-176). Zozbierané dáta obsahovali taktiež rečové signály pacientov a okrem celkovej hodnoty UPDRS sa zamerali aj na predikciu škály hybných (motorických) funkcií pacienta (0-108). Použili pritom metódy troch lineárnych regresií a jednej nelineárnej regresie. Podľa dosiahnutých výsledkov dokázali predikovať motorické hodnoty UPDRS približne v rozmedzí 6 bodov a celkové UPDRS v rozsahu 7.5 bodov. Tieto výsledky odrážajú najlepší odhad chyby predikcie pri 1000 spusteniach 10 násobnej krížovej validácie. Konečné predpovede hodnôt pomocou modelov sú veľmi blízke lekárskym pozorovaniam na klinike.

3 Pochopenie a príprava dát

Dáta, s ktorými sme pracovali sú voľne dostupné na UCI Machine Learning Repository [6] a skladajú sa z celej rady biomedicínskych hlasových meraní od 31 osôb (z toho 23 s PCH). Spolu bolo k dispozícií 195 záznamov (riadkov), pretože každý pacient mal v dátach viacero záznamov, ktoré boli brané nezávisle od seba. Reč pacienta bola transformovaná do 23 atribútov, ako napríklad: priemerná, maximálna a minimálna vokálna frekvencia, miera variability vo frekvencii (atribúty skupiny Jitter), miera variability v amplitúde (atribúty skupiny Shimmer), merania pomeru hluku a tónových zložiek v hlase (NHR, HNR) a množstvo ďalších.

K danému dátovému setu sme následne pripojili ďalšie atribúty, ktoré sme našli vo vedeckom článku [7]. Obsahovali dodatočné informácie o pohlaví, veku, štádiu PCH (8 úrovní) a počte rokov pacienta od kedy mu bola táto choroba diagnostikovaná. Po pridaní atribútu *Stage* sme odstránili atribút *status*, ktorý podával informáciu o tom, či

pacient trpí PCH (1) alebo nie (0). Dáta o pacientoch s najhorším 5. štádiom sme nemali k dispozícii. Po celkovej úprave dát (spojenie datasetov, odstránenie chýbajúcich hodnôt) sme pracovali so 189 riadkami a 25 stĺpcami.

4 Modelovanie

Pre vytváranie klasifikačných modelov sme si hlavne kvôli jednoduchej interpretácie pre lekárov zvolili iba metódu rozhodovacích stromov a jej algoritmy C4.5, C5.0 a CART, ktoré dosiahli z viacerých algoritmov najvyššie presnosti. Pre vyhodnotenie modelov sme použili metódy 10-násobnej krížovej validácie (**10 – CV**) a validácie kde sa vynecháva jeden záznam (*Leave One Out* – **LOO**). Pri LOO sa jedná o podobný spôsob vyhodnocovania ako v prípade *k*-násobnej krížovej validácie, ale klasifikátor sa buduje na *n*-1 záznamoch v dátovej množine a testuje sa len na 1 zázname. Tento proces sa následne opakuje *n*-krát. Pre vytvorenie modelov sme najprv vybrali všetky atribúty a neskôr sme očistili dáta od kolinearity (2 alebo viac atribútov sú navzájom silne závislé), ktorá môže zhoršiť presnosť modelov [8]. Namiesto vytvorenia korelačnej matice je lepším spôsobom pre posúdenie kolinearity vypočítať inflačný faktor rozptylu (**VIF**) pre každý atribút. Najmenšia možná hodnota VIF je 1, čo predstavuje úplnú absenciu kolinearity. Ako pravidlo platí, že ak hodnota VIF presahuje 5 alebo 10, tak hovoríme o problematickom množstve kolinearity. Odstránili sme preto atribúty s hodnotou VIF väčšou ako 5 a ich počet sa zredukoval z 25 na 13.

Tab. 1. Výsledky model při 10 – CV a LOO – CV

Výber atribútov	**10 - CV**			**LOO - CV**		
	CART	*C4.5*	*C5.0*	*CART*	*C4.5*	*C5.0*
Všetky atribúty	69,71%	79,88%	83,54%	71,96%	80,42%	81,48%
VIF < 5	67,66%	85,15%	**86,2%**	72,47%	80,95%	82,01%

Z výsledkov v Tab. 1 si môžeme všimnúť, že odstránenie atribútov s vysokou kolinearitou nám zabezpečilo vyššie presnosti modelov skoro vo všetkých prípadoch. Výnimkou bol algoritmus CART, pri ktorom sa presnosť zmenšila zo 69,71% na 67,66%, avšak tieto presnosti sú v porovnaní s algoritmami C4.5 a C5.0 aj tak podstatne nižšie. Najvyššie presnosti vo všetkých prípadoch dosiahol algoritmus C5.0, pričom pri odstránení atribútov s vysokou kolinearitou a použití 10 – CV sme dosiahli presnosť na úrovni **86,2%** pre klasifikáciu pacientov do 7 tried.

S cieľom vylepšiť presnosti vytvorených modelov sme sa rozhodli použiť metódy **RandomForest**, **Bagging** a **Boosting**, ktoré používajú stromy ako stavebné bloky na vytvorenie silnejších predikčných modelov. RandomForest vytvára viacero rozhodovacích stromov, kde v každom strome pri výbere testovacieho atribútu je braných do úvahy *m* náhodne vybraných atribútov z ich celkového počtu *p*. Výsledná klasifikácia do triedy je zvolená hlasovaním všetkých vygenerovaných stromov. Ak sa pri danom uzle berú do úvahy všetky atribúty *p,* vtedy hovoríme o baggingu. Podobným spôsobom funguje aj boosting, avšak každý rozhodovací strom berie do úvahy aj informá-

ciu z predchádzajúceho stromu [9]. Záznamom, ktoré boli v predchádzajúcom strome klasifikované nesprávne je v ďalšej iterácii priradená väčšia váha, vďaka čomu bude pri ďalšej iterácii kladený na tieto záznamy väčší dôraz. V publikácii [10] autori uvádzajú, že s rastúcim počtom vygenerovaných stromov sa zvyšuje už len výpočtová záťaž a rozdiely v presnostiach sú už veľmi malé. Ich analýza 29 datasetov ukázala, že pri vygenerovaní 128 stromov už nie je významný rozdiel v presnosti ako pri 256, 512, 1024, 2048 a 4096 stromoch. Preto sme počet rozhodovacích stromov nastavili na 50, 100 a 150 a pre výpočet presností sme si zvolili opäť 10 – násobnú krížovú validáciu. Pri pokuse orezať vygenerované rozhodovacie stromy sa úspešnosti zmenšili. Rovnaký problém nastal aj pri obmedzení maximálnej hĺbky rozhodovacieho stromu, preto sme tento vstupný parameter nemenili a nechali sme ho nastavený na predvolenej (defaultnej) hodnote.

Tab. 2. Výsledky metód RandomForest, Bagging a Boosting

Počet stromov	RandomForest	Bagging	Boosting
m = 50	87,25%	77,78%	93,65%
m = 100	86,73%	77,78%	95,24%
m = 150	86,73%	78,31%	**95,77%**

Vo výsledkoch v Tab. 2 vidíme, že najvyššie presnosti dosiahla jednoznačne metóda Boosting a to pri vygenerovaní 150 stromov (95,77%). S rastúcim počtom stromov presnosť klasifikácie rástla pri metódach Bagging aj Boosting, no naopak pre RandomForest trošku klesla. Algoritmus C5.0 pri 10 – CV dosiahol lepšie presnosti ako Bagging v tomto prípade a porovnateľné s metódou RandomForest.

Pre metódu Boosting, ktorá dosiahla najvyššiu presnosť na úrovni 95.77% sme si zobrazili na Obr.1 aj kontingenčnú tabuľku, ktorá porovnáva modelom predikované hodnoty atribútu Stage s tými, ktoré boli poznané a dané v testovacej množine.

	Observed Class						
Predicted Class	0	1	1.5	2	2.5	3	4
0	47	0	0	2	1	0	0
1	0	15	0	0	0	1	0
1.5	0	0	19	0	0	0	0
2	1	1	0	28	0	0	0
2.5	0	2	0	0	42	0	0
3	0	0	0	0	0	23	0
4	0	0	0	0	0	0	7

Obr. 1. Kontingenčná tabuľka pre metódu Boosting

Keďže pre vyhodnotenie modelov sme použili 10-násobnú krížovú validáciu, testovanie prebiehalo na 10 rôznych testovacích množinách. Každý prvok v kontingenčnej tabuľke na Obr. 1 je vypočítaný ako súčet zo všetkých získaných kontingenčných tabuliek, pomocou ktorých je jasné vidieť, pri akej predikcií došlo najčastejšie k chybe. Hlavným cieľom je maximalizovať hodnoty v matici na hlavnej diagonále, čo predstavuje správnu predikciu daného štádia Parkinsonovej choroby. Z celkového

počtu 189 záznamov dokázal model správne predikovať v 181 prípadoch, čo predstavuje 95.77% presnosť. Pre jednotlivé štádia sme dosiahli tieto presnosti (v zátvorke je vyjadrený pomer správne klasifikovaných záznamov ku všetkým záznamov pre dané štádium PCH): štádium 0 (47/48) – 97.92%, štádium 1 (15/18) = 83.33%, štádium 1.5 (19/19) = 100%, štádium 2 (28/30) = 93.33%, štádium 2.5 (42/43) = 97.67%, štádium 3 (23/24) = 95.83%, štádium 4 (7/7) = 100%.

Môžeme si všimnúť, že so 100% presnosťou dokázal model predikovať štádium 1.5 a 4. Naopak najnižšiu presnosť na úrovni 83.33% dosiahlo prvé štádium PCH, kde model vedel v 15tich záznamoch určiť správne štádium a pri 3 záznamoch došlo k chybe (v 1 prípade predikoval štádium 2 a v dvoch prípadoch štádium 2.5). V končenom dôsledku dosiahol model skoro pri všetkých štádiách úspešnosť nad 93%, jedinou výnimkou bolo štádium 1, kedy model dosiahol iba 83.33%.

5 Záver a budúca práca

V tomto článku sme sa zamerali na určenie štádia pacientov s PCH z ich reči pomocou metód dolovania v dátach. Už pri prvom experimente a odstránení kolinearity v dátach sme pomocou rozhodovacieho stromu a algoritmu C5.0 dosiahli presnosť na úrovni 86,2% (pred odstránením kolinearity – 83,54%). Použitím metódy Boosting, ktorá vytvorí viacero rozhodovacích stromov, sme našu presnosť dokázali zvýšiť až na 95,77% (pri $m = 150$), čo je vzhľadom na klasifikáciu záznamov až do 7 tried vysoká úspešnosť. Napr. v publikácii [4] pri binárnej klasifikácii (1 – pacient s PCH, 0 – zdravý pacient) s rovnakými dátami dosiahli autori najvyššiu presnosť len 76% použitím SVM. Rovnako binárnu klasifikáciu pacientov sme robili aj v našej predchádzajúcej publikácii [11] a najlepší výsledok na úrovni 91,43% sme dosiahli použitím algoritmu C4.5. Aj keď sa jednalo o zložitejšiu klasifikáciu pacientov (7 tried) v porovnaní s binárnou klasifikáciou (2 triedy), napriek tomu sme dosiahli vyššiu presnosť modelu pomocou metódy Boosting.

V budúcej práci by sme sa chceli zamerať na spracovanie hovorenej reči do rovnakých atribútov, aby bolo možno vytvoriť aplikáciu, kde si ľudia nahrajú svoju reč a dokážu sa testovať. Ďalej by sme sa chceli zamerať taktiež na dáta, ktoré sme získali od spoločnosti *mPower: Mobile Parkinson Disease Study*. Zaznamenávajú demografické údaje pacientov, ale aj údaje o ich hlase, chôdzi, pamäti a klikaní na obrazovku mobilu. Vďaka týmto dátam by sme mohli rozšíriť náš výskum na viaceré oblasti a príznaky PCH.

Poďakovanie. Táto publikácia vznikla vďaka podpore Vedeckej grantovej agentúry MŠVVaŠ SR a SAV projekt č. 1/0493/16 a Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR, projekt č. 025TUKE-4/2015.

Literatúra

1. De Lau, L. M., Bretler, M. M.: Epidemiology of Parkinson's disease, In: The Lancet Neurology, vol. 5, no. 6 (2006), pp. 525 – 535.

2. Cnockaert, L., et al.: Low-frequency vocal modulations in vowels produced by Parkinsonian subjects. In: Speech Comunication, vol. 50, no. 4 (2008), pp. 288-300.
3. Fans, S., et al.: Members of the UPDRS Development Committee Unified Parkinson's Disease Rating Scale. In: Recent developments in Parkinson's disease, vol. 2 (1987), pp. 153-163.
4. Yadav, G., et al.: Predication of Parkinson's disease using data mining methods: a comparative analysis of tree, statistical, and support vector machine classifiers. In: Indian Journal of Medical Sciences, vol. 65, no. 6 (2011), pp. 231-242.
5. Tsanas, A., et al.: Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. In: IEEE Transactions on Biomedical Engineering, vol. 57, no. 4 (2010), pp. 884-893.
6. UCI Machine Learning repository: Center for Machine Learning and Intelligent Systems – Parkinsons Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Parkinsons.
7. Little, M. A., et al.: Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. In: IEEE Transactions on Biomedical Engineering, vol. 56, no. 4 (2009), pp. 1015-1022.
8. James, G., et al.: An Introduction to Statistical Learning. Springer-Verlag New York, 2013.
9. Schapire, R., et al.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Proceedings of the Second European Conference on Computational Learning Theory, vol. 904 (1995), pp. 23-37.
10. Oshiro, T. M., et al.: How Many Trees in a Random Forest? In: Machine Learning and Data Mining in Pattern Recognition, vol. 7376 (2012), pp. 154-168.
11. Vadovský, M., Paralič, J.: Predikcia Parkinsonovej choroby pomocou signálov reči použitím metód dolovania v dátach. In: WIKT & DaZ 2016, Bratislava: STU, 2016. s. 329-333. ISBN: 978-80-227-4619-9.

Annotation:

Utilizing processed records of patients' speech in determining the stage of Parkinson´s disease

The medical procedures for disease diagnostics are significantly demanding and time-consuming. Data mining methods can accelerate this process and assist doctors in making decisions in complex situations. In case of Parkinson´s disease (PCH), the diagnostics of the initial disease stage is the primary issue, since the symptoms are not so unambiguous and easily observable. Therefore, this article is focused on determining the actual stage of PCH based on the data recording signals of patient´s speech using decision trees (C4.5, C5.0 and CART). Methods such as RandomForest, Bagging and Boosting were also employed to improve the existing classification models. Estimation of model accuracy was achieved by using k-fold cross-validation and validation with omission of one record (Leave-one-out). In addition, experiments were also performed to remove collinearity in data by computing the Variance inflation factor (VIF) in order to increase the accuracy of the models.

Analýza dát za účelom zlepšenia konkrétneho procesu logistickej firmy

Miroslava Muchová, Ján Paralič

Fakulta elektrotechniky a informatiky
Technická univerzita v Košiciach
Letná 9/B, 042 00 Košice, Slovenská republika

{miroslava.muchova, jan.paralic}@tuke.sk

Abstrakt. Systém pre podporu rozhodovania v riadení logistiky môže poskytnúť manažérom logistických spoločnosti cenné informácie potrebné pre uskutočnenie efektívnych rozhodnutí. Jednou z kľúčových a často riešených úloh v logistickej firme nie je len otázka nízkej spotreby paliva ale aj rozhodnutie akého vodiča priradí na akú trasu. Cieľom článku je aplikovať techniky prediktívneho dolovania dát a analyzovať štýl jazdy vodičov ako aj určiť kombináciu faktorov vplývajúcich na spotrebu paliva. Predkladaný článok sa okrem iného zaoberá aj problematikou analýzy dát pre zlepšenie rozhodovania vo vybranom logistickom procese – výber vodičov na plánované dodacie trasy. Vykonaný dotazníkový prieskum ukazuje, že v logistických a prepravných spoločnostiach chýba systém, ktorý by uľahčil prideľovanie vodičov na dodaciu trasu.

Kľúčové slová: analýza dát, priemerná spotreba paliva, systém pre podporu rozhodovania, Naive Bayes

1 Úvod

V poslednej dobe rôzne výskumné štúdie poukázali na výhody použitia veľkých dátových metód v oblasti logistiky a riadenia dodávateľského reťazca. Tan, K. H. a kolektív navrhli analýzu infraštruktúry veľkých dát na posilnenie schopnosti v oblasti inovácií dodávateľského reťazca [6]. Çakıcı prišiel s nápadom použiť RFID údaje na prepracovanie optimálnej politiky zásob [1]. Zhong prišiel s návrhom ako informácie z veľkých dát môžu byť použité na efektívne plánovanie logistiky a plánovanie výroby [8]. Dutta a Bose predstavili riadenie veľkých objemov dát pomocou logistických sietí [2]. Waller a Fawcett argumentujú, že používanie údajov, prediktívnej analýzy môže pomôcť manažérom logistiky splniť interné potreby a prispôsobiť sa zmenám v dodávateľskom reťazci. Jednoducho povedané, nasadenie analýzy dát pre logistiku a riadenie dodávateľského reťazca by malo zvýšiť pridanú hodnotu pre zákazníkov, pričom integrovaný výrobný a distribučný proces v celom dodávateľskom reťazci zahŕňa aj výrobcov, dodávateľov, maloobchodníkov, poskytovateľov logistických služieb, ktoré sú podporované použitím veľkého množstva informácií [7].

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 221-226.

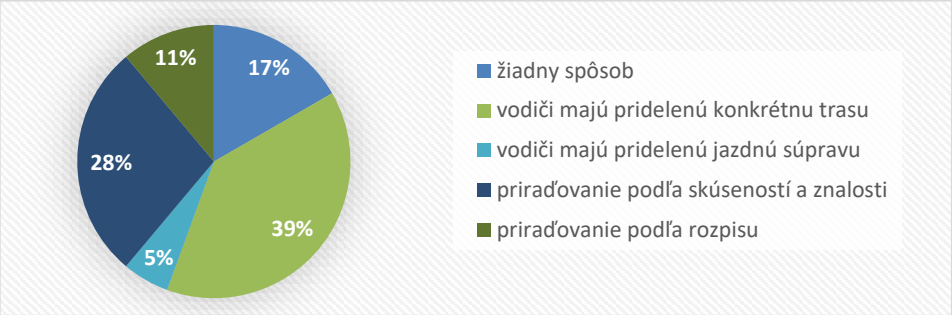
V tomto článku sa preto chceme zamerať na verifikáciu tvrdenia o prínose analýzy dát za účelom zlepšenia konkrétneho firemného procesu – výberu vodičov na plánované dodacie trasy v konkrétnej logistickej firme. Okrem iného, cieľom článku je aplikovať techniky prediktívneho dolovania dát, vďaka ktorým môžeme určiť hlavné faktory, ktoré ovplyvňujú priemernú spotrebu paliva ako aj identifikovať najvhodnejšie jazdné postupy a štýly vodičov.

2 Dotazníkový prieskum

Hlavným cieľom výskumu je vytvoriť systém pre podporu rozhodovania pre výber vodiča na konkrétnu trasu. Z toho dôvodu bol vykonaný prieskum medzi logistickými spoločnosťami. Vykonaný dotazníkový prieskum ukazuje, že v logistických a prepravných spoločnostiach chýba systém, ktorý by uľahčil prideľovanie vodičov na dodaciu trasu.

Prostredníctvom mailu bolo oslovených viac ako 300 logistických a prepravných spoločnosti. Jednou z otázok bolo aj, či by si vedeli predstaviť systém, ktorý rozhodne akého vodiča priradí na akú trasu, 10% opýtaných si takýto systém nevedia predstaviť, 40% by takýto systém nevyužili a 50% spoločnosti by takýto systém vedeli využiť a v súčasnosti jeden takýto systém testujú a plánujú ho nasadiť. Okrem toho sme prostredníctvom dotazníkov zistili, že viac ako 88% spoločností analyzuje nejakým spôsobom jazdný štýl vodiča, pričom asi 46% na analýzu využíva program.

Na nasledujúcom obrázku môžeme vidieť zloženie odpovedí na ďalšiu položenú otázku: Akým spôsobom priradzujete vodičov na konkrétnu trasu?



Obr. 1. Odpovede na otázku získané prostredníctvom dotazníkového prieskumu

3 Analýza zvolenej množiny dát

Dáta, ktoré máme k dispozícii pochádzajú z prepravnej spoločnosti. Jedná sa o firmu, ktorej predmetom činnosti podľa obchodného registra je výkon povolania prevádzkovateľa nákladnej cestnej dopravy. K dispozícii boli dáta o deviatich vozidlách (z toho sú už tri vyradené) a o pätnástich vodičoch.

Firma na komunikáciu s vozidlami využíva systém Dynafleet Online – Volvo Truck Corporation [5]. Informácie sťahované z vozidiel sú ukladané do databázy. Pre účely našej práce sme mali k dispozícii dva výkazy a to Výkaz hodnotenia spotreby paliva, ktorý obsahoval 26 atribútov, ako napr. meno vodiča, dátum, priemerná rýchlosť, priemerná spotreba paliva a ďalšie atribúty, ktoré podávajú informáciu o celkovom hodnotení jazdného štýlu vodiča – predvídanie, tempomat, voľný dojazd, úsporná a neúsporná jazda atď. Druhým výkazom bol Výkaz sledovania a obsahoval 24 atribútov, napr. čas udalosti, prejdená vzdialenosť, palivo, miesto, hmotnosť nákladu a podobne. Všetky dáta boli uložené vo formáte excelovských tabuliek.

3.1 Príprava dát

Pre potreby našich analýz, sme na základe dátumu obidva dátové súbory zlúčili a informácie o všetkých vozidlách sme spojili do jedného súboru. Získali sme tak 49 atribútov a vyše 2900 záznamov. Odstránili sme nepodstatné atribúty (napr. atribút Úplné zastavenie mal rovnaké hodnoty ako atribút Voľnobeh, atribúty I-Shift v A, I-Shift v M, I-Shift v P a Zaťaženie motora vo viac ako 98% obsahovali rovnaké hodnoty). V ďalšom kroku sme si atribút Miesto rozdelili na Štát, Mesto, Oblasť a Ulicu a odstránili sme tie záznamy, ktoré neobsahovali informáciu o konkrétom meste. Po tejto úprave nám zostalo viac ako 2800 záznamov.

Atribúty, ktoré podávajú informáciu o celkovom hodnotení jazdného štýlu vodiča (Celkové hodnotenie, Predvídanie, Úsporná a Neúsporná jazda, Najvyšší prevodový stupeň, Využitie motora a prevodovky...) boli transformované na intervaly, presne tak, ako to vyhodnocuje systém Dynafleet:

- 80 – 100 = Dobrý výkon,
- 60 – 79 = Priemer,
- 0 – 59 = Potenciál k zlepšeniu.

Ďalším krokom bola diskretizácia na rovnakú hĺbku intervalu, ktorá rozdelila atribút Priemerná spotreba paliva na nasledovne:

- pod 26,3 l/100km
- 26,4 – 31,5 l/100 km
- nad 31,6 l/100 km

3.2 Návrh modelu

Návrh modelu bol realizovaný pomocou softvérového nástroja RapidMiner a programu Visual Studio. Jednalo sa o model Naivného Bayesovského klasifikátora, pomocou ktorého sme určili, ktoré atribúty majú najväčší vplyv na spotrebu paliva. Dáta sme rozdelili na trénovaciu a testovaciu množinu v pomere 80:20. Cieľovým atribútom bola Priemerná spotreba paliva. Celková úspešnosť vytvoreného modelu je 67,93%.

V nasledujúcej tabuľke sú zobrazené najvyššie pravdepodobnosti dosiahnutia nízkej spotreby pre najzaujímavejšie atribúty. Napr. ak vodič bude efektívne využívať tempomat, tak s pravdepodobnosťou 91,8% dosiahne nízku spotrebu paliva.

Tab. 1. Kľúčové faktory vplývajúce na priemernú spotrebu paliva

Attribute	Parameter	do 26.3 ↓
Úsporná jazda	value=Dobrý výkon	0.957
Tempomat	value=Dobrý výkon	0.918
Využitie motora a prevodovky	value=Dobrý výkon	0.908
Najvyšší prevodový stupeň	value=Dobrý výkon	0.907
Voľný dojazd	value=Dobrý výkon	0.765
Prispôsobenie rýchlosti	value=Dobrý výkon	0.743

Vodič ovláda napríklad rýchlosť, zrýchlenie, brzdenie, otáčky motora alebo zaradenie rýchlostného stupňa. Vyhodnotením zozbieraných údajov môže spoločnosť usmernením vodičov alebo zistením prípadných nedostatkov predchádzať rôznym neželaným situáciám a taktiež môže zefektívniť ich jazdu. Napríklad efektívnym používaním tempomatu môže vodič dosiahnuť nižšiu spotrebu a dospieť tak k zníženiu celkových nákladov.

Na základe týchto analýz a analýz uvedených v [3] [4] môžeme zostaviť model dodacej trasy, ktorá nám dokáže s určitou pravdepodobnosťou povedať, akú budú mať priemernú spotrebu jednotlivý vodiči. Napríklad, na trase Trelleborg (Švédsko) – Rostok (Nemecko), Vodič_I spotrebuje viac ako 31,5 l/100km, Vodič_J spotrebuje do 26,3 l/100 km.

Na základe vykonaných experimentov sme získali poznatky týkajúce sa úspory spotreby paliva. Odporúčania, ktoré by viedli k zníženiu spotreby paliva sú nasledovné:

- zlepšiť neúspornú jazdu – častejšie využívať motorovú brzdu,
- zlepšiť úspornú jazdu – predstavuje dobu, kedy je vozidlo v pohybe a vodič má nohu na plyne,
- minimalizovať použitie brzdového pedálu,
- efektívnejšie využívať tempomat.

4 Záver

Vďaka technikám prediktívneho dolovania dát, sme určili hlavné faktory, ktoré ovplyvňujú priemernú spotrebu paliva. Na základe týchto výsledkov sme navrhli odporúčania pre vodičov, ktoré by viedli k zníženiu spotreby paliva. Následne sme uká-

zali ako takéto výsledky môžu ovplyvniť rozhodovanie o priradení vodiča na plánovanú dodaciu trasu s dôrazom na nízku spotrebu paliva.

Ďalší výskum bude preto zameraný na rozšírenie vytvoreného modelu o ďalšie faktory, ktoré pomôžu pri rozhodovaní o výbere vodiča. Na základe požiadaviek majiteľa firmy navrhneme model rozhodovania využívajúceho navrhnuté modely vodiča a dodacej trasy.

Literatúra

1. Çakıcı, Ö.E., Groenevelt, H., Seidmann, A.: Using RFID for the Management of Pharmaceutical Inventory—System Optimization and Shrinkage Control. In: Decision Support Systems. 2011. pp. 842-852, ISSN: 0167-9236.
2. Dutta, D., Bose, I.: Managing a Big Data Project: The Case of Ramco Cements Limited. In: International Journal of Production Economics. 2015. pp. 293-306, ISSN: 0925-5273.
3. Muchová, M., Paralič, J.: Analýza dát za účelom zlepšenia konkrétneho firemného procesu logistickej firmy. In: WIKT&DaZ 2016: Proceedings from Conference. Bratislava: STU, (2016), pp. 299-304, ISBN 978-80-227-4619-9
4. Muchová, M: Big data analysis in selected logistics process. In: SCYR 2016: Proceedings from Conference. Košice: TU, (2016), pp. 55-57. ISBN 978-80-553-2566-8
5. Systém na komunikáciu s vozidlami Dynafleet Online – Volvo Truck Corporation, dostupné na internete: http://www.volvotrucks.com/trucks/dynafleet-help/splash/Pages/splash.aspx
6. [Tan, K.H, Zhan, Y.Z., Ji, G.J., Ye, F., Chang, C.T. Harvesting Big Data to Enhance Supply Chain Innovation Capabilities: An Analytic Infrastructure Based on Deduction Graph. In: International Journal of Production Economics. 2015. pp. 223 – 233, ISSN: 0925-5273.
7. Waller, M. A., Fawcett S. E.: Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. In: Journal of Business Logistics. 2013. pp. 77-84, ISSN: 2158-1592.
8. Zhong, R.Y., Huang, G.Q., Lan, S., Dai, Q.Y., Chen, X., Zhang, T.: A Big Data Approach for Logistics Trajectory Discovery from RFID-Enabled Production Data. In: International Journal of Production Economics. 2015. pp. 260-272, ISSN: 0925-5273.

Poďakovanie:
Táto práca bola podporovaná Agentúrou na podporu výskumu a vývoja na základe zmluvy č. APVV-16-0213 a Kultúrnou a edukačnou grantovou agentúrou MŠVVaŠ SR, projekt č. 025TUKE-4/2015.

Annotation:

Decision support system in the logistics can provide managers of logistics companies valuable information needed to make effective decisions. One of the key and often solved tasks in logistics company are not only the question of low fuel economy but also decisions about assignment of truck drivers to planned routes. The aim of the article is to identify the combination of factors that have a key effect on fuel consumption. This paper focuses on data analysis to improve decision-making in the selected logistics process - the selection of truck drivers for planned routes.

A questionnaire survey shows, that in logistics companies and transport companies often miss information system that would facilitate assigning truck drivers on planned routes.

Information Extraction from the Web by Matching Visual Presentation Patterns

Matej Minarik, Radek Burget

Brno University of Technology, Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czech Republic

iminarikma@fit.vutbr.cz, burgetr@fit.vutbr.cz

Abstract. There is a large amount of data available on the Web. Data are often represented as text, enriched with tables, lists, images or other visual structures. These data are usually coded in HTML without any additional semantics, which makes them nigh impossible to automatically process and extract. There are approaches based on top-down document segmentation according to visual information and layout. We present a bottom-up approach which starts with the smallest consistent elements and matches the visual relationships among these elements to a pre-defined ontological structure of extracted records. This method considers not only the visual attributes of a particular segment, but also its position amongst other segments.

Keywords: web data integration, information extraction, structured record extraction, page segmentation, content classification, ontology mapping

1 Introduction

There is a great amount of documents on the World Wide Web. These documents are usually HTML documents with lots of data encoded in HTML tags. There are some generic tags available, such as *div* or *span* with no semantics at all. HTML 5 [4] introduced new tags with a generic semantics, such as *article*, *header* or *footer*. Additionally, RDFa annotations [6] (among other possibilities) allow expressing the semantics of the individual document elements by mapping to ontological concepts or properties.

Unfortunately, large amounts of data available on the World Wide Web are still only accessible through plain HTML documents with no semantic annotations at all. When we want to machine-process the contained data, it is necessary to identify the contained data records and recognize the semantics of the individual data fields. For some data sources with a regular structure (such as Wikipedia), we may create specific extraction templates. However, Web is a heterogeneous place with many different sources of data represented in countless ways.

Many methods have been developed recently for identifying and extracting structured data records from plain HTML documents [5, 7, 9]. Most of them are based on a

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 227-231.

top-down approach: first, the document is pre-processed in order to locate the most probable regions of interest (called data sections [9] or data regions [5]). Then, the individual data records are found in the regions based on the detection of repeating structures in the model by frequency measures [7] or visual pattern detection [1, 9].

In order to avoid the complex and often unreliable process of data region identification, we have recently proposed an opposite bottom-up approach [2]. Considering that the expected structure of the records to be extracted is usually known in advance, we try to find a best match among this expected structure and the presentation patterns in the whole document starting from the smallest atomic elements. In this paper, we propose possible future development that aims to bridge the gap between the plain HTML documents and the semantic web.

2 Method Overview

The key idea of the method presented in [2] is to identify visual presentation patterns that are used to represent the data we are trying to extract and then to extract these data using discovered patterns.

This method assumes only the visual representation of the documents, so it may be applied not only on HTML documents, but on PDFs and other visual formats as well. Since we are working with the visual information only, we need to create a uniform representation of the source document. This representation is a set of *visual areas* and it is the first step of the process.

The next step is the initial *tagging* step. In the initial tagging, we assign one or multiple *tags* to each discovered visual area. The idea is to identify all the visual areas in the document that *could* possibly correspond to a particular piece of information (e.g. a personal name, date or e-mail address). This tagging may be based on a variety of approaches starting with simple regular expressions and ending with a kind of *NER* classifier. We admit that there will be incorrect tags and even multiple tags assigned for some visual areas.

After the initial tagging step, we need to disambiguate the tags. This method assumes that all the data records are presented in a visually consistent way in the source documents. This allows introducing presentation constraints on the data records. The disambiguation task itself consists of finding matching record presentation and layout which meets visual constraints and has a great support in source documents. This method defines four visual relations:

- $(a_1, a_2) \in R_{side}$ when a_1 and a_2 are on the same line and a_2 is placed to the right of a_1 with no element between them
- $(a_1, a_2) \in R_{after}$ when a_1 and a_2 are on the same line and a_2 is anywhere to the right of a_1 (there may be elements between them)

- $(a_1, a_2) \in R_{under}$ when a_1 and a_2 are roughly in the same column, a_2 is placed below a_1 with no element between them; additionally, the vertical space is not larger than *0.8 em*
- $(a_1, a_2) \in R_{below}$ when a_1 and a_2 are roughly in the same column, a_2 is placed anywhere below a_1

We choose the most supported relation by trying to cover as many tagged visual areas as possible.

The last step is the *record extraction* itself. The methods results in a set of matches which identify the visual areas that contain data we are interested in. By simply concatenating these areas we can obtain the text content.

3 Further Development Directions

The presented method is based on quite simple visual relationships discovered in the documents and it expects small and regular data records being present in the documents. For adapting the methods for more complex documents and data records, several ways of extensions may be considered.

3.1 Additional Visual Relationships

For more complex documents, we would like to consider more ways the relationships among elements may be presented in the documents [3]. We may try to represent more high-level relations, such as *heading-subheading*, or *paragraph* which is basically a couple of lines separated from other paragraphs in some visual way (usually with a newline or a starting tab). Other interesting relation might be the same row or same column inside a table. Same row might tell us that those values are referring to the same entity. Same column usually represents different values of some metric. On top of that we might try to identify grouped columns, which is frequent for timetables.

3.2 Advanced Tagging

DBPedia might be used for improving the tagging step for example by integrating DBPedia Spotlight, which is a tool for automatically annotating mentions of DBpedia resources in text. This tool is able to identify entity mentions and select the best candidate based on the user-provided configuration. Other than that, we may introduce some NLP tasks as a tagging enhancement. The obvious one is a *named entity recognition* tagger.

3.3 Semantic Relationship Representation

In our recent work [8], we proposed using RDF for representing the visual model of the processed documents. Similarly to this approach, we would like to represent even the discovered visual relationships using a RDF graph. The expected benefits are greater efficiency in discovering frequent visual patterns (for example using SPARQL queries) and the possibility to directly map the identified data fields to ontological properties.

4 Conclusion

In this paper, we have discussed a method of structured record extraction from web documents. The method is purely vision-based and unlike most existing approaches, it does not rely on a complex document pre-processing steps for identifying the data regions in the documents. Instead, it uses an opposite approach that marks all the possible (even incorrectly identified) occurences of the given information in the documents and later, the data records are identified by finding the best match between the expected record structure and the visual presentation patterns discovered in the document. We have proposed possible extensions of this method that are expected to allow processing more complex documents and integrate the extracted results with semantic web resources.

References

1. N. Anderson and J. Hong. Visually extracting data records from query result pages. In Web Technologies and Applications: 15th Asia-Pacific Web Conference, APWeb 2013, Sydney, Australia, April 4-6, 2013. Proceedings, pages 392–403, Berlin, Heidelberg, 2013. Springer.
2. R. Burget. Information extraction from web documents based on visual and semantic relationship alignment. In *NLP&DBpedia 2016*. Springer International Publishing, 2017. *To appear*.
3. R. Burget and P. Smrz. Extracting visually presented element relationships from web documents. *International Journal of Cognitive Informatics and Natural Intelligence*, 2013(2):13–29, 2013.
4. S. Faulkner, I. Hickson, S. Pfeiffer, E. D. Navara, R. Berjon, T. O'Connor, and T. Leithead. HTML5. W3C recommendation, W3C, Oct. 2014.
5. P. L. Goh, J. L. Hong, E. X. Tan, and W. W. Goh. Region based data extraction. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 1196–1200, May 2012.
6. I. Herman, S. McCarron, M. Birbeck, and B. Adida. RDFa core 1.1 - third edition. W3C recommendation, W3C, Mar. 2015.
7. J. L. Hong, E.-G. Siew, and S. Egerton. Information extraction for search engines using fast heuristic techniques. *Data Knowl. Eng.*, 69(2):169–196, Feb. 2010.
8. M. Milicka and R. Burget. Multi-aspect document content analysis using ontological modelling. In *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2014)*, pages 9–12. Vydavateĺstvo STU, 2014.

9. D. Weng, J. Hong, and D. A. Bell. Automatically annotating structured web data using a SVM-based multiclass classifier. In *Web Information Systems Engineering – WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I*, pages 115–124, Cham, 2014. Springer International Publishing.

Acknowledgement: This work was supported by the BUT FIT grant FIT-S-17-39

Rejstřík autorů

Josef Steinberger, Martin Zíma, Dalibor Fiala, Martin Dostal, Michal Nykl (editoři)

Data a znalosti 2017, sborník konference

1. vydání

Určeno pro účastníky konference DATA A ZNALOSTI 2017
elektronické vydání
248 stran
Vydává Západočeská univerzita v Plzni, 2017
Sborník je distribuován v elektronické podobě

ISBN 978-80-261-0720-0