Zbornik 21. mednarodne multikonference

INFORMACIJSKA DRUŽBA - IS 2018

Zvezek C

Proceedings of the 21st International Multiconference

INFORMATION SOCIETY - IS 2018

Volume C

Odkrivanje znanja in podatkovna skladišča - SiKDD

Data Mining and Data Warehouses - SiKDD

Uredila / Edited by
Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

8.–12. oktober 2018 / 8–12 October 2018
Ljubljana, Slovenia

Zbornik 21. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2018

Zvezek C

Proceedings of the 21st International Multiconference

INFORMATION SOCIETY – IS 2018

Volume C

Odkrivanje znanja in podatkovna skladišča - SiKDD
Data Mining and Data Warehouses - SiKDD

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

8.–12. oktober 2018 / 8–12 October 2018
Ljubljana, Slovenia

Urednika:

Dunja Mladenić
Laboratorij za umetno inteligenco
Institut »Jožef Stefan«, Ljubljana

Marko Grobelnik
Laboratorij za umetno inteligenco
Institut »Jožef Stefan«, Ljubljana

Založnik: Institut »Jožef Stefan«, Ljubljana
Priprava zbornika: Mitja Lasič, Vesna Lasič, Lana Zemljak
Oblikovanje naslovnice: Vesna Lasič

Dostop do e-publikacije:
http://library.ijs.si/Stacks/Proceedings/InformationSociety

Ljubljana, oktober 2018

Informacijska družba
ISSN 2630-371X

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID=31884839
ISBN 978-961-264-137-5 (pdf)

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2018

Multikonferenca Informacijska družba (http://is.ijs.si) je z enaindvajseto zaporedno prireditvijo osrednji srednjeevropski dogodek na področju informacijske družbe, računalništva in informatike. Letošnja prireditev se ponovno odvija na več lokacijah, osrednji dogodki pa so na Institutu »Jožef Stefan«.

Informacijska družba, znanje in umetna inteligenca so še naprej nosilni koncepti človeške civilizacije. Se bo neverjetna rast nadaljevala in nas ponesla v novo civilizacijsko obdobje ali pa se bo rast upočasnila in začela stagnirati? Bosta IKT in zlasti umetna inteligenca omogočila nadaljnji razcvet civilizacije ali pa bodo demografske, družbene, medčloveške in okoljske težave povzročile zadušitev rasti? Čedalje več pokazateljev kaže v oba ekstrema – da prehajamo v naslednje civilizacijsko obdobje, hkrati pa so notranji in zunanji konflikti sodobne družbe čedalje težje obvladljivi.

Letos smo v multikonferenco povezali 11 odličnih neodvisnih konferenc. Predstavljenih bo 215 predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic. Prireditev bodo spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica, ki se ponaša z 42-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2018 sestavljajo naslednje samostojne konference:

- Slovenska konferenca o umetni inteligenci
- Kognitivna znanost
- Odkrivanje znanja in podatkovna skladišča – SiKDD
- Mednarodna konferenca o visokozmogljivi optimizaciji v industriji, HPOI
- Delavnica AS-IT-IC
- Soočanje z demografskimi izzivi
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Delavnica za elektronsko in mobilno zdravje ter pametna mesta
- Vzgoja in izobraževanje v informacijski družbi
- 5. študentska računalniška konferenca
- Mednarodna konferenca o prenosu tehnologij (ITTC)

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija, Slovensko društvo za umetno inteligenco (SLAIS), Slovensko društvo za kognitivne znanosti (DKZ) in druga slovenska nacionalna akademija, Inženirska akademija Slovenije (IAS). V imenu organizatorjev konference se zahvaljujemo združenjem in institucijam, še posebej pa udeležencem za njihove dragocene prispevke in priložnost, da z nami delijo svoje izkušnje o informacijski družbi. Zahvaljujemo se tudi recenzentom za njihovo pomoč pri recenziranju.

V letu 2018 bomo šestič podelili nagrado za življenjske dosežke v čast Donalda Michieja in Alana Turinga. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe bo prejel prof. dr. Saša Divjak. Priznanje za dosežek leta bo pripadlo doc. dr. Marinki Žitnik. Že sedmič podeljujemo nagradi »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono letos prejme padanje državnih sredstev za raziskovalno dejavnost, jagodo pa Yaskawina tovarna robotov v Kočevju. Čestitke nagrajencem!

Mojca Ciglarič, predsednik programskega odbora

Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2018

In its 21st year, the Information Society Multiconference (http://is.ijs.si) remains one of the leading conferences in Central Europe devoted to information society, computer science and informatics. In 2018, it is organized at various locations, with the main events taking place at the Jožef Stefan Institute.

Information society, knowledge and artificial intelligence continue to represent the central pillars of human civilization. Will the pace of progress of information society, knowledge and artificial intelligence continue, thus enabling unseen progress of human civilization, or will the progress stall and even stagnate? Will ICT and AI continue to foster human progress, or will the growth of human, demographic, social and environmental problems stall global progress? Both extremes seem to be playing out to a certain degree – we seem to be transitioning into the next civilization period, while the internal and external conflicts of the contemporary society seem to be on the rise.

The Multiconference runs in parallel sessions with 215 presentations of scientific papers at eleven conferences, many round tables, workshops and award ceremonies. Selected papers will be published in the Informatica journal, which boasts of its 42-year tradition of excellent research publishing.

The Information Society 2018 Multiconference consists of the following conferences:

- Slovenian Conference on Artificial Intelligence
- Cognitive Science
- Data Mining and Data Warehouses - SiKDD
- International Conference on High-Performance Optimization in Industry, HPOI
- AS-IT-IC Workshop
- Facing demographic challenges
- Collaboration, Software and Services in Information Society
- Workshop Electronic and Mobile Health and Smart Cities
- Education in Information Society
- 5th Student Computer Science Research Conference
- International Technology Transfer Conference (ITTC)

The Multiconference is co-organized and supported by several major research institutions and societies, among them ACM Slovenia, i.e. the Slovenian chapter of the ACM, Slovenian Artificial Intelligence Society (SLAIS), Slovenian Society for Cognitive Sciences (DKZ) and the second national engineering academy, the Slovenian Engineering Academy (IAS). On behalf of the conference organizers, we thank all the societies and institutions, and particularly all the participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

For the sixth year, the award for life-long outstanding contributions will be presented in memory of Donald Michie and Alan Turing. The Michie-Turing award will be given to Prof. Saša Divjak for his life-long outstanding contribution to the development and promotion of information society in our country. In addition, an award for current achievements will be given to Assist. Prof. Marinka Žitnik. The information lemon goes to decreased national funding of research. The information strawberry is awarded to the Yaskawa robot factory in Kočevje. Congratulations!

Mojca Ciglarič, Programme Committee Chair

Matjaž Gams, Organizing Committee Chair

KONFERENČNI ODBORI
CONFERENCE COMMITTEES

International Programme Committee

Vladimir Bajic, South Africa
Heiner Benking, Germany
Se Woo Cheon, South Korea
Howie Firth, UK
Olga Fomichova, Russia
Vladimir Fomichov, Russia
Vesna Hljuz Dobric, Croatia
Alfred Inselberg, Israel
Jay Liebowitz, USA
Huan Liu, Singapore
Henz Martin, Germany
Marcin Paprzycki, USA
Karl Pribram, USA
Claude Sammut, Australia
Jiri Wiedermann, Czech Republic
Xindong Wu, USA
Yiming Ye, USA
Ning Zhong, USA
Wray Buntine, Australia
Bezalel Gavish, USA
Gal A. Kaminka, Israel
Mike Bain, Australia
Michela Milano, Italy
Derong Liu, USA
Toby Walsh, Australia

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki
Mitja Lasič
Blaž Mahnič
Jani Bizjak
Tine Kolenik

Programme Committee

Franc Solina, co-chair
Viljan Mahnič, co-chair
Cene Bavec, co-chair
Tomaž Kalin, co-chair
Jozsef Györkös, co-chair
Tadej Bajd
Jaroslav Berce
Mojca Bernik
Marko Bohanec
Ivan Bratko
Andrej Brodnik
Dušan Caf
Saša Divjak
Tomaž Erjavec
Bogdan Filipič
Andrej Gams
Matjaž Gams
Marko Grobelnik
Nikola Guid
Marjan Heričko
Borka Jerman Blažič Džonova
Gorazd Kandus
Urban Kordeš
Marjan Krisper
Andrej Kuščer
Jadran Lenarčič
Borut Likar
Mitja Luštrek
Janez Malačič
Olga Markič
Dunja Mladenič
Franc Novak
Vladislav Rajkovič
Grega Repovš
Ivan Rozman
Niko Schlamberger
Stanko Strmčnik
Jurij Šilc
Jurij Tasič
Denis Trček
Andrej Ule
Tanja Urbančič
Boštjan Vilfan
Baldomir Zajc
Blaž Zupan
Boris Žemva
Leon Žlajpah

KAZALO / TABLE OF CONTENTS

Zbornik 21. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2018

Zvezek C

Proceedings of the 21st International Multiconference

INFORMATION SOCIETY – IS 2018

Volume C

Odkrivanje znanja in podatkovna skladišča - SiKDD
Data Mining and Data Warehouses - SiKDD

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

11. oktober 2018 / 11 October 2018
Ljubljana, Slovenia

PREDGOVOR

Tehnologije, ki se ukvarjajo s podatki so v devetdesetih letih močno napredovale. Iz prve faze, kjer je šlo predvsem za shranjevanje podatkov in kako do njih učinkovito dostopati, se je razvila industrija za izdelavo orodij za delo s podatkovnimi bazami, prišlo je do standardizacije procesov, povpraševalnih jezikov itd. Ko shranjevanje podatkov ni bil več poseben problem, se je pojavila potreba po bolj urejenih podatkovnih bazah, ki bi služile ne le transakcijskem procesiranju ampak tudi analitskim vpogledom v podatke – pojavilo se je t.i. skladiščenje podatkov (data warehousing), ki je postalo standarden del informacijskih sistemov v podjetjih. Paradigma OLAP (On-Line-Analytical-Processing) zahteva od uporabnika, da še vedno sam postavlja sistemu vprašanja in dobiva nanje odgovore in na vizualen način preverja in išče izstopajoče situacije. Ker seveda to ni vedno mogoče, se je pojavila potreba po avtomatski analizi podatkov oz. z drugimi besedami to, da sistem sam pove, kaj bi utegnilo biti zanimivo za uporabnika – to prinašajo tehnike odkrivanja znanja v podatkih (data mining), ki iz obstoječih podatkov skušajo pridobiti novo znanje in tako uporabniku nudijo novo razumevanje dogajanj zajetih v podatkih. Slovenska KDD konferenca pokriva vsebine, ki se ukvarjajo z analizo podatkov in odkrivanjem znanja v podatkih: pristope, orodja, probleme in rešitve.

INTRODUCTION

Data driven technologies have significantly progressed after mid 90's. The first phases were mainly focused on storing and efficiently accessing the data, resulted in the development of industry tools for managing large databases, related standards, supporting querying languages, etc. After the initial period, when the data storage was not a primary problem anymore, the development progressed towards analytical functionalities on how to extract added value from the data; i.e., databases started supporting not only transactions but also analytical processing of the data. At this point, data warehousing with On-Line-Analytical-Processing entered as a usual part of a company's information system portfolio, requiring from the user to set well defined questions about the aggregated views to the data. Data Mining is a technology developed after year 2000, offering automatic data analysis trying to obtain new discoveries from the existing data and enabling a user new insights in the data. In this respect, the Slovenian KDD conference (SiKDD) covers a broad area including Statistical Data Analysis, Data, Text and Multimedia Mining, Semantic Technologies, Link Detection and Link Analysis, Social Network Analysis, Data Warehouses.

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Dunja Mladenić, Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana

Marko Grobelnik, Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana

Preparing multi-modal data for natural language processing

Erik Novak
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
erik.novak@ijs.si

Jasna Urbančič
Jožef Stefan Institute
Ljubljana, Slovenia
jasna.urbancic@ijs.si

Miha Jenko
Jožef Stefan Institute
Ljubljana, Slovenia
miha.jenko@ijs.si

ABSTRACT

In education we can find millions of video, audio and text educational materials in different formats and languages. This variety and multimodality can impose difficulty on both students and teachers since it is hard to find the right materials that match their learning preferences. This paper presents an approach for retrieving and recommending items of different modalities. The main focus is on the retrieving and preprocessing pipeline, while the recommendation engine is based on the k-nearest neighbor method. We focus on educational materials, which can be text, audio or video, but the proposed procedure can be generalized on any type of multi-modal data.

KEYWORDS

Multi-modal data preprocessing, machine learning, feature extraction, recommender system, open educational resources

ACM Reference Format:
Erik Novak, Jasna Urbančič, and Miha Jenko. 2018. Preparing multi-modal data for natural language processing. In *Proceedings of Slovenian KDD Conference (SiKDD'18).* ACM, New York, NY, USA, Article 4, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

There are millions of educational materials that are found in different formats – courses, video lectures, podcasts, simple text documents, etc. Because of its vast variety and multimodality it is difficult for both students and teachers to find the right materials that will match their learning preferences. Some like to read a short scientific papers while others just like to sit back and watch a lecture that can last for hours. Additionally, materials are written in different languages, which is a barrier for people who are not fluent in the language the material is written in. Finding a good approach of providing educational material would help improving their learning experience.

In this paper we present a preprocessing pipeline which is able to process multi-modal data and input it in a common semantic space. The semantic space is based on Wikipedia concepts extracted from the content of the materials. Additionally, we developed a content based recommendation model which uses Wikipedia concepts to find similar items based on the model input. Throughout the paper we focus on educational material but the approach can be generalized to other multi-modal data sets.

The reminder of the paper is structured as follows. In section 2 we go over related work. Next, we present the data preprocessing pipeline which is able to process different types of data – text, video and audio – and describe each component of the pipeline in section 3. A content based recommendation model that uses Wikipedia concepts to compare materials is presented in section 4. Finally, we present future work and conclude the paper in section 5.

2 RELATED WORK

In this section we present the related work which the rest of the paper is based on. We split this section into subsections – multi-modal data preprocessing and recommendation models.

Multi-modal Data Preprocessing. Multi-modal data can be seen as classes of different data types from which we can extract similar features. In the case of educational material the classes are video, audio and text. One of the approaches is to extract text from all class types. In [6] the authors describe a Machine Learning and Language Processing automatic speech recognition system that can convert audio to text in the form of transcripts. The system can also process video files as they are also able to extract audio from it. Their model was able to achieve a 13.3% word error rate on an English test set. These kind of systems are useful for extracting text from audio and video but would need to have a model for each language.

Recommendation models. These models are broadly used in many fields – from recommending videos based on what the user viewed in the past, to providing news articles that the user might be interested in. One of the most used approaches is based on collaborative filtering [16], which finds users that have similar preferences with the target user and recommends items based on their ratings. Recommender systems now do not contain only one algorithm but multiple which return different recommendations.

Authors of [10] discuss about the various algorithms that are used in the Netflix recommender system (top-n video ranker, trending now, continue watching, and video-video similarity), as well as the methods they use to evaluate their system. A high level description of the Youtube recommender system is found in [3]. They developed a candidate generation model and a ranking model using deep learning. Both Netflix and Youtube recommend videos based on users' interaction with them and the users history. To some extent this can be used for educational resources but cannot be generalized on the whole multi-modal data set since we cannot acquire data about users' interaction with, for instance, text.

SiKDD'18, October 2018, Ljubljana, Slovenia

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

A collaborative filtering based recommendation system for the educational sector is presented in [8]. They evaluated educational content using big data analysis techniques and recommended courses to students by using their grades obtained in other subjects. This gives us insight into how recommendations can be used in education but our focus is to recommend educational materials rather than courses. In a sense courses can be viewed as bundles of educational material; thus, our interest is recommending "parts of courses" to the user.

3 DATA PREPROCESSING

In this paper we focus on open educational resources (OER), which are freely accessible, openly licensed text, media, and other digital assets that are useful for teaching, learning and assessing [21]. These are found in different OER repositories maintained by universities, such as MIT OpenCourseWare [12], Università di Bologna [7], Université de Nantes [4] and Universitat Politècnica de València [5], as well as independent repositories such as Videolectures.NET [20], a United Nations award-winning free and open access educational video lectures repository.

For processing the different OER we developed a preprocessing pipeline that can handle each resource type and output metadata used for comparing text, audio and video materials. The pipeline is an extension of the one described in [11]; its architecture is shown in figure 1. What follows are the descriptions of each component in the preprocessing pipeline.

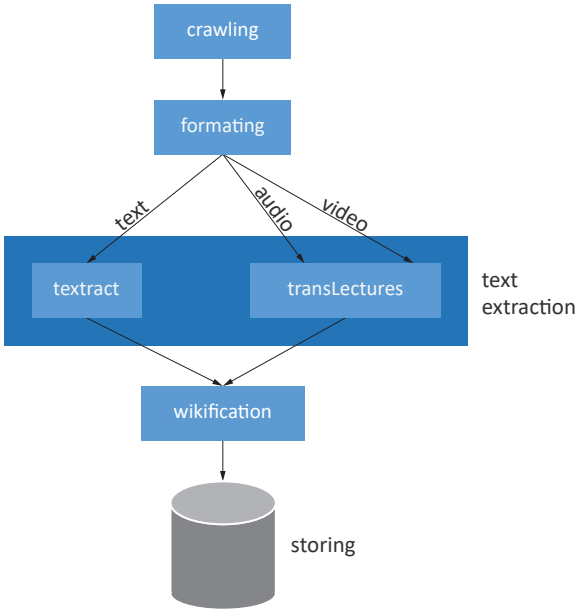


Figure 1: The preprocessing pipeline architecture. It is designed to handle each data type as well as extract features to support multi- and cross-linguality.

Crawling. The first step is to acquire the educational materials. We have targeted four different OER repositories (MIT OpenCourseWare, Università di Bologna, Université de Nantes and Videolectures.NET), for which we used their designated APIs or developed custom crawlers to acquire their resources. For each material we acquired its metadata, such as the materials title, url, type, language in which it is written and its provider. These values are used in the following steps of the pipeline as well as to represent the material in the recommendations.

Formatting. Next, we format the acquired material metadata. We designate which attributes every material needs to have as well as set placeholders for the features extracted in the following steps of the pipeline. By formatting the data we set a schema which makes checking which attributes are missing easy. We do not have a mechanism for handling missing attributes in the current pipeline iteration but we will dedicate time to solve this problem in the future.

Text Extraction. The third step, we extract the content of each material in text form. Since the material can be a text, video or audio file to handled each file type separately.

For text we employed *textract* [1] to extract raw text from the given text documents. The module omits figures and returns the content as text. The extracted text is not perfect - in the case of materials for mathematics it does not know how to represent mathematical equations and symbols. In that case, it replaces the equations with textual noise. Currently we do nothing to handle this problem and use the output as is.

For video and audio we use the subtitles and/or transcriptions to represent the materials content. To do this, we use *transLectures* [18] which generates transcriptions and translations of a given video and audio. The languages it supports are English, Spanish, German and Slovene. The output of the service is in dfxp format [17], a standard for xml caption and subtitles based on timed text markup language, from which we extract the raw text.

Wikification. Next, we send the material through wikification - a process which identifies and links material textual components to the corresponding Wikipedia pages [15]. This is done using Wikifier [2], which returns a list of Wikipedia concepts that are most likely related to the textual input. The web service also supports cross- and multi-linguality which enables extracting and annotating materials in different languages.

Wikifier's input text is limited to 20k characters, because of which longer text cannot be processed as a whole. We split longer text into chunks of at most 10k characters and pass them to Wikifier. Here we are careful not to split the text in the middle of a sentence and if that is not possible, to at least not split any words.

We split the text as follows. First we make a 10k characters long substring of the text. Next, we identify the last character in the substring that signifies the end of a sentence (a period, a question mark, or an exclamation point) and split it at that character. If there is no such character we find the last whitespace in the substring and split it there. In the extreme case where no whitespaces are found we take the substring as is. The substring becomes one chunk of the original text. We repeat the process on the remaining text until it is fully split into chunks.

When we pass these chunks into Wikifier, it returns Wikipedia concepts related to the given chunk. These concepts also contains

the Cosine similarity between the Wikipedia concept page and the given input text. To calculate the similarity between the concept and the whole material we aggregated the concepts by calculating the weighted sum

$$S_k = \sum_{i=1}^{n} \frac{L_i}{L} s_{ki},$$

where S_k is the aggregated Cosine similarity of concept k, n is the number of chunks for which Wikifier returned concept k, L_i is the length of chunk i, L is the length of the materials raw text, and s_{ki} is the Cosine similarity of concept k to chunk i. The weight $\frac{L_i}{L}$ represents the presence of concept k, found in chunk i, in the whole material. The aggregated Wikipedia concepts are stored in the materials metadata attribute.

Data Set Statistics. In the final step, we validate the material attributes and store it in a database. The OER material data set consists of approximately 90k items. The distribution of materials over the four repositories is shown in figure 2.

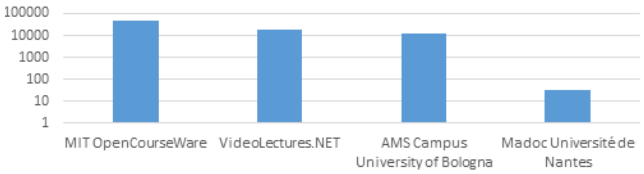


Figure 2: Number of materials per repository crawled in logarithm scale. Most materials come from MIT OpenCourseWare followed by Videolectures.NET.

Some of the repositories offer material in different languages. All repositories together cover 103 languages, however for only 8 languages the count of available materials is larger than 100. The distribution of items over languages is shown in figure 3 where we only show languages with more than 100 items available. Most of the materials is in English, followed by Italian and Slovene. The "Unknown" column shows that for about 6k materials we were not able to extract the language. To acquire this information, we will improve the language extraction method in our preprocessing pipeline.

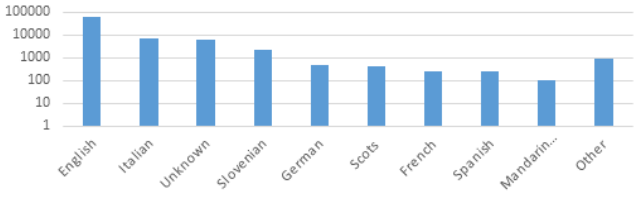


Figure 3: Number of materials per language in logarithm scale. Most of the material is in English, followed by Italian and Slovenian.

As shown in before the preprocessing pipeline is designed to handle different types of material - text, video and audio. Each type can be represented in various file formats, such as pdf and docx for text, wmv and mp4 for video, and mp3 for audio. We visualized the distribution of materials over file types in figure 4, but we only show types with more than 100 items available.

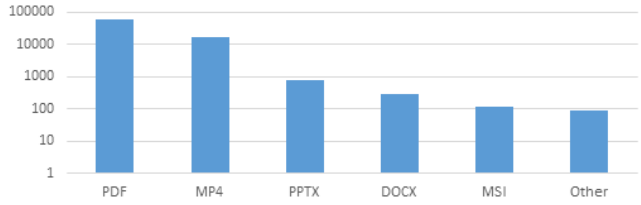


Figure 4: Number of items per file type in logarithm scale. The dominant file type is text (pdf, pptx and docx), followed by video (mp4).

As seen from the figure, the dominant file type is text (pdf, pptx and docx) followed by video (mp4). The msi file type is an installer package file format used by Windows but it can also be a textual document or a presentation. If we generalize the file type distribution over all OER repositories we can conclude that the dominant file type is text. This will be taken into count when improving the preprocessing pipeline and recommendation engine.

4 RECOMMENDER ENGINE

There are different ways of creating recommendations. Some employ users' interests while other are based on collaborative filtering. In this section we present our content based recommendation engine which uses the k-nearest neighbor algorithm [13]. What follows are descriptions of how the model generates recommendations based on the user's input, which can be either the identifier of the OER in the database or a query text.

Material identifier. When the engine receives the material identifier (in our case the url of the material) we first check if the material is in our database. If present, we search for k most similar materials to the one with the given identifier based on the Wikipedia concepts. Each material is represented by a vector of its Wikipedia concepts where each value is the aggregated Cosine similarity of the corresponding Wikipedia concept page to the material. By calculating the Cosine similarity between the materials the engine then selects k materials with the highest similarity score and returns them to the user. Because of the nature of Wikipedia concepts this approach returns materials written in different languages - which helps overcoming the language barrier.

Query text. When the engine receives the query text we search for materials with the most similar raw text using the bag-of-words model. Each material is represented as a bag-of-words vector where each value of the vector is the tf-idf of the corresponding word. The materials are then compared using the Cosine similarity and the engine again returns the k materials that have the highest similarity score. This approach is simple but it is unable to handle multilingual documents. This might be overcome by first sending the query text to Wikifier to get its associated Wikipedia concepts and use them in a similar way as described in the *Material identifier* approach.

4.1 Recommendation Results

The described recommender engine is developed using the QMiner platform [9] and is available at [14]. When the user inputs a text query the system returns recommendations similar to the given text. These are shown as a list where each item contains the title, url, description, provider, language and type of the material. Clicking on an item redirects the user to the selected OER.

We have also discussed with different OER repository owners and found that they would be interested in having the recommendations in their portal. To this end, we have developed a compact recommendation list which can be embedded in a website. The recommendations are generated by providing the material identifier or raw text as query parameters in the embedding url. Figure 5 shows the embed-ready recommendation list.

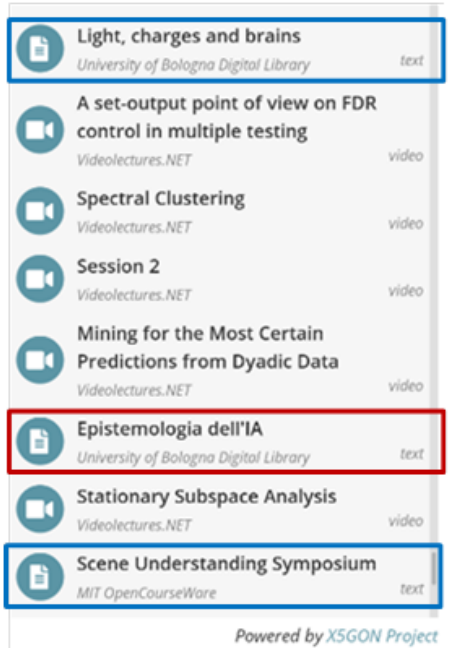


Figure 5: An example of recommended materials for the lecture with the title "Is Deep Learning the New 42?" published on Videolectures.NET [19]. The figure shows cross-lingual, cross-modal, and cross-site recommendations.

The recommendation list consists of the top 100 materials based on the query input. As shown in the figure the recommendation contain materials of different types, are provided by different repositories and written in different languages. We have not yet evaluated the recommendation engine but we intend to do it in the future.

5 FUTURE WORK AND CONCLUSION

In this paper we present the methodology for processing multimodal items and creating a semantic space in which we can compare these items. We acquired a moderately large open educational resources data set, created a semantic space with the use of Wikipedia concepts and developed a basic content based recommendation engine.

In the future we will evaluate the current recommendation engine and use it to compare it with other state-of-the-art. We intend to use A/B testing to optimize the models based on the user's interaction with them. We wish to improve the engine by collecting user activity data to determine what materials are liked by the users, explore different deep learning methods to improve results, and develop new representations and embeddings of the materials.

We also aim to improve the preprocessing pipeline by improving text extraction methods, handle missing material attributes, and adding new feature extraction methods to determine the topic and scientific field of the educational material as well as their quality.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and X5GON European Unions Horizon 2020 project under grant agreement No 761758.

REFERENCES

[1] David Bashford. 2018. GitHub - dbashford/textract: node.js module for extracting text from html, pdf, doc, docx, xls, xlsx, csv, pptx, png, jpg, gif, rtf and more! https://github.com/dbashford/textract. Accessed: 2018-09-03.
[2] Janez Brank, Gregor Leban, and Marko Grobelnik. 2017. Annotating documents with relevant Wikipedia concepts. *Proceedings of SiKDD*.
[3] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 191–198.
[4] Université de Nantes. 2018. Plate-forme d'Enseignement de l'Université de Nantes. http://madoc.univ-nantes.fr/. Accessed: 2018-09-03.
[5] Universitat Politècnica de València. 2016. media UPV. https://media.upv.es/#/portal. Accessed: 2018-09-03.
[6] Miguel Ángel del Agua, Adrià Martínez-Villaronga, Santiago Piqueras, Adrià Giménez, Alberto Sanchis, Jorge Civera, and Alfons Juan. 2015. The MLLP ASR Systems for IWSLT 2015. In *Proc. of 12th Intl. Workshop on Spoken Language Translation (IWSLT 2015)*. Da Nang (Vietnam), 39–44. http://workshop2015.iwslt.org/64.php
[7] Università di Bologna. 2018. Universita di Bologna. https://www.unibo.it/it. Accessed: 2018-09-03.
[8] Surabhi Dwivedi and VS Kumari Roshni. 2017. Recommender system for big data in education. In *E-Learning & E-Learning Technologies (ELELTECH), 2017 5th National Conference on*. IEEE, 1–4.
[9] Blaz Fortuna, J Rupnik, J Brank, C Fortuna, V Jovanoski, M Karlovcec, B Kazic, K Kenda, G Leban, A Muhic, et al. 2014. » QMiner: Data Analytics Platform for Processing Streams of Structured and Unstructured Data «, Software Engineering for Machine Learning Workshop. In *Neural Information Processing Systems*.
[10] Carlos A Gomez-Uribe and Neil Hunt. 2016. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2016), 13.
[11] Erik Novak and Inna Novalija. 2017. Connecting Professional Skill Demand with Supply. *Proceedings of SiKDD*.
[12] Massachusetts Institute of Technology. 2018. MIT OpenCourseWare | Free Online Course Materials. https://ocw.mit.edu/index.htm. Accessed: 2018-09-03.
[13] Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia* 4, 2 (2009), 1883.
[14] X5GON Project. 2018. X5GON Platform. https://platform.x5gon.org/search. Accessed: 2018-09-04.
[15] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 1375–1384.
[16] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.
[17] Speechpad. 2018. DFXP (Distribution Format Exchange Profile) | Speechpad. https://www.speechpad.com/captions/dfxp. Accessed: 2018-09-04.
[18] transLectures. 2018. transLectures | transcription and translation of video lectures. http://www.translectures.eu/. Accessed: 2018-09-03.
[19] VideoLectures.NET. 2018. Is Deep Learning the New 42? - Videolectures.NET. http://videolectures.net/kdd2016_broder_deep_learning/. Accessed: 2018-09-03.
[20] VideoLectures.NET. 2018. VideoLectures.NET - VideoLectures.NET. http://videolectures.net/. Accessed: 2018-09-03.
[21] Wikipedia. 2018. Open educational resources - Wikipedia. https://en.wikipedia.org/wiki/Open_educational_resources. Accessed: 2018-09-03.

TOWARDS SMART STATISTICS IN LABOUR MARKET DOMAIN

Inna Novalija
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia

inna.koval@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia

marko.grobelnik@ijs.si

ABSTRACT

In this paper, we present a proposal for developing smart labour market statistics based on streams of enriched textual data and illustrate its application on job vacancies from European countries. We define smart statistics scenarios including demand analysis scenario, skills ontology development scenario and skills ontology evolution scenario. We identify stakeholders – consumers for smart statistics and define the initial set of smart labour market statistical indicators.

KEYWORDS

Smart statistics, labour market, demand analysis.

1. INTRODUCTION

An essential feature of modern economy is the appearance of new skills, such as digital skills. For instance, e-skills lead to the exponential increases in production and consumption of data.

While job profiles vary and are still in the process of being defined, organizations agree that they need the new breed of workers.

Accordingly, the European institutions take major initiatives related to digitalization of labor market, training of new skills and meeting the labour demand.

Historically, the labour market statisticians use standard measures of the labour demand and labour supply based on traditional surveys – job vacancy surveys, wage survey, labour force surveys. The unemployment rate provides information on the supply of persons looking for work in excess of those who are currently employed. Data on employment provide information on the demand for workers that is already met by employers.

The data-driven smart labour market statistics intends to:

- use the available historical job vacancies data,
- use the available real-time job vacancies data,
- use the available real-time and historical dataset of additional data (described below),
- align data sources,
- construct models and obtain novel smart labour market indicators that will complement existing labour market statistics,
- provide a system for delivering results to the users.

The smart labour market statistics approach will combine advanced data processing, modelling and visualization methods in order to develop trusted techniques for job vacancies analysis with respect to defined scenarios – demand analysis, skills ontology development and skills ontology evolution.

2. BACKGROUND

The development of smart labour market statistics touches a number of issues from labour market policies area and would provide contributions to questions related to:

- job creation,
- education and training systems,
- labour market segmentation,
- improving skill supply and productivity.

For instance, the analysis of the available job vacancies could offer an insight into what skills are required in the particular area. Effective trainings based on skills demand could be organized and that would lead into better labour market integration.

A number of stakeholder types will benefit from the development of smart labour market statistics. In particular, the targeted stakeholders are:

- Statisticians from National and European statistical offices who are interested in the application of new technologies for production of the official statistics.
- Individual persons who are searching for new employment opportunities. In particular, individuals are interested in the job vacancies that are compatible with their current skills and in the methods (like trainings) providing the possibilities to obtain new skills in demand.
- Public and private employment agencies interested in up-to-date employees profiles.
- Education and training institutions from different levels and forms of education - general/vocational education, higher education, public/private, initial/ adult education. Educational institutions are interested in relevant skills and topics that should be part of the curriculum programs.
- Ministries of labour/manpower, economy/industry/trade, education, finance, etc. The policy makers, such as ministries, are interested in the overall labour market situation, with respect to location and time, in the labour market segmentation and in the processes of improving supply and productivity.
- Standards development organizations. National or International organizations whose primary activities are developing, coordinating, promulgating, revising, amending, reissuing, interpreting, or otherwise producing technical standards that are intended to address the needs of some

relatively wide base of affected adopters. Interested in new technologies developed in relation to labour market.

- Academic and research institutes. Public and private entities who conduct research in relevant areas. Research institutions are interested in the development of novel methodologies and usage of appearing new data sources.

3. RELATED WORK

The European Data Science Academy (EDSA) [1] was an H2020 EU project that ran between February 2015 and January 2018. The objective of the EDSA project was to deliver the learning tools that are crucially needed to close the skill gap in Data Science in the EU. The EDSA project has developed a virtuous learning production cycle for Data Science, and has:

- Analyzed the sector specific skillsets for data analysts across Europe with results reflected at EDSA demand and supply dashboard;
- Developed modular and adaptable curricula to meet these data science needs; and
- Delivered training supported by multiplatform resources, introducing Learning pathway mechanism that enables effective online training.

EDSA project established a pipeline for job vacancy collecting and analysis that will be reused for the purpose of smart statistics.

An ontology called SARO (Skills and Recruitment Ontology) [2] has been developed to capture important terms and relationships to facilitate the skills analysis. SARO ontology concepts included relevant classes to job vacancy datasets, such as Skill and JobPosting. Examples of instances of class Skill would be skills, such as “Data analysis”, “Java programming language” et al.

ESCO [3] is the multilingual classification of European Skills, Competences, Qualifications and Occupations. It identifies and categorizes skills/competences, qualifications and occupations relevant for the EU labour market and education and training, in 25 European languages. The system provides occupational profiles showing the relationships between occupations, skills/competences and qualifications. For instance, one example of existing ESCO skill is “JavaScript” (with alternative labels “Client-side JavaScript”, "JavaScript 1.7" et al.).

Both SARO and ESCO ontologies are useful for the aim of smart statistics, in particular for skills ontology development and skills ontology evolution scenarios. However, the ontologies usually are manually manipulated, and the methods developed for smart labour market statistics should overcome the difficulties related to this issue. The ontology evolution scenario of smart labour market statistics envisions automatic identification of emerging and decreasing skills from the data perspective.

4. PROBLEM DEFINITION

4.1 DATA SOURCES

The main data sources available for the development of smart labour market statistics are historical and current data about job vacancies in the area of digital technologies and data science around Europe (~5.000.000 job vacancies 2015-2018).

Additional data sources may include:

- Social media data, such as news, Twitter data that might be relevant for labour market.

- Labour supply data (based on user profile analysis).

Open job vacancies can be found using job search services. These services aggregate job vacancies by location, sector, applicant qualifications and skill set or type. One such service is Adzuna [4], a search engine for job ads, which mostly covers English-speaking countries.

For data acquisition and enrichment, dedicated APIs, including Adzuna API, are used, as well as custom web crawlers are developed. The data is formatted to JSON to aid further processing and enrichment. The job vacancy dataset is obtained with respect to trust and privacy regulations, the personal data is not collected.

Job vacancies usually contain the information, such as job position title, job description, company and job location. In such way, job vacations that are constantly crawled/web-scraped present a data stream. The job title and job description are textual data that contain information about skills that employee should have.

On the obtained data wikification - identifying and linking textual components (including skills) to the corresponding Wikipedia pages [5] is performed. This is done using Wikifier [6], which also supports cross and multi-linguality enabling extraction and annotation of relevant information from job vacancies in different languages. The data is tagged with concepts from GeoNames ontology [7]. To job postings where latitude and longitude have been available, GeoNames location uri and location name are added. To the postings where only location name has been available, the coordinates and location uri are added.

The job vacancy data representation level depends on the specific country. For the United Kingdom, France, Germany and the Netherlands there is a substantial collection of job vacancies in the area of digital technologies.

4.2 CONCEPTUAL ARCHITECTURE

The labour market statistics conceptual structure is built upon the following major blocks:

1. Data sources related to different aspects of smart labour market. The main data source aggregates historical and current job vacancies in the area of digital technologies and data science around Europe.

2. Modelling smart labour market statistics takes central part of the smart labour market statistics approach, where the goal is to construct models based on different data sources, updated in business-real-time (as needed or as data sources allow). Models shall bring understanding of the smart labour market statistics domain and shall be used for aggregation, ontology development and ontology evolution.

3. Targeted users are smart statistics consumers. There are several major groups of users (described above). The example users might include statisticians, policy makers, individual users (residents and non-residents), training and educational organizations and other.

4. Finally, applications of smart labour market statistics are multiscale - they can be presented at cross-country level (around

Europe) country level (UK, France, the Netherlands etc.), city/area level and conceptual level (ontology).

Figure 1 illustrates the conceptual architecture diagram for smart labour market statistics.

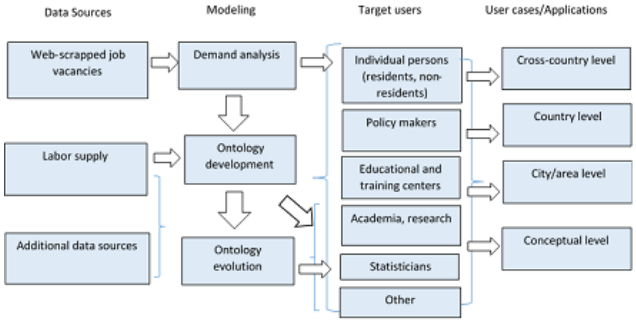


Figure 1: Conceptual Architecture

The key characteristics of the development techniques will include:

- Interpretability and transparency of the models – the aim is, for a model to be able to explain its decision in a human readable manner (vs. black box models, which provide results without explanation).

- Non-stationary modelling techniques are required due to changing data and its statistical properties in time. For instance, the ontology evolution process will be modeled taking to the account the incremental data arriving to the system.

- Multi-resolution nature of the models, having the property to observe the structure of a model on multiple levels of granularity, depending on the application needs.

- Scalability for building models is required due to the nature of incoming data streams.

4.3 SCENARIOS

The smart labour market statistics proposal includes three scenarios - demand analysis scenario, ontology development scenario and ontology evolution scenario described below.

4.3.1 DEMAND ANALYSIS

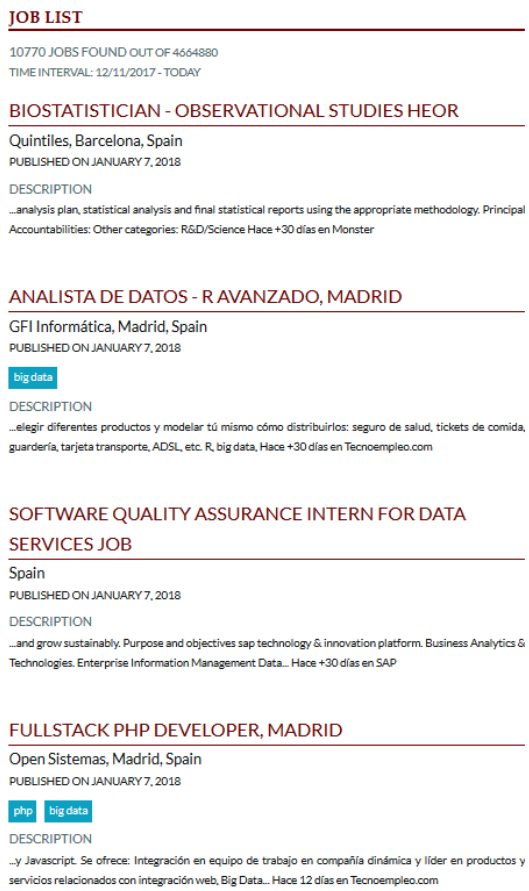
Demand analysis scenario suggests production of statistical indicators based on the available job vacancies using techniques for data preprocessing, semantic annotation, cross-linguality, location identification and aggregation.

Job vacancies in structural and semi-structural form are the input to into the system, while statistics related to overall job demand, job demand with respect to particular location, job demand with respect to particular skill (skill demand) and time frame are the outputs of the system.

Figure 2 presents an example of crawled and processed job vacancies.

4.3.2 SKILLS ONTOLOGY DEVELOPMENT

Ontologies reduce the amount of information overload in the working process by encoding the structure of a specific domain and offering easier access to the information for the users. Gruber [8] states that an ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain. The specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provide meanings for the vocabulary and formal constraints on its coherent use.



JOB LIST

10770 JOBS FOUND OUT OF 4664880
TIME INTERVAL: 12/11/2017 - TODAY

BIOSTATISTICIAN - OBSERVATIONAL STUDIES HEOR

Quintiles, Barcelona, Spain
PUBLISHED ON JANUARY 7, 2018

DESCRIPTION
...analysis plan, statistical analysis and final statistical reports using the appropriate methodology. Principal Accountabilities: Other categories: R&D/Science Hace +30 días en Monster

ANALISTA DE DATOS - R AVANZADO, MADRID

GFI Informática, Madrid, Spain
PUBLISHED ON JANUARY 7, 2018

big data

DESCRIPTION
...elegir diferentes productos y modelar tú mismo cómo distribuirlos: seguro de salud, tickets de comida, guardería, tarjeta transporte, ADSL, etc. R, big data, Hace +30 días en Tecnoempleo.com

SOFTWARE QUALITY ASSURANCE INTERN FOR DATA SERVICES JOB

Spain
PUBLISHED ON JANUARY 7, 2018

DESCRIPTION
...and grow sustainably. Purpose and objectives sap technology & innovation platform. Business Analytics & Technologies. Enterprise Information Management Data... Hace +30 días en SAP

FULLSTACK PHP DEVELOPER, MADRID

Open Sistemas, Madrid, Spain
PUBLISHED ON JANUARY 7, 2018

php big data

DESCRIPTION
...y Javascript. Se ofrece: Integración en equipo de trabajo en compañía dinámica y líder en productos y servicios relacionados con integración web, Big Data... Hace 12 días en Tecnoempleo.com

Figure 2: Example of Job Vacancies Crawled and Processed

Ontologies are often manually developed and maintained, what requires a sufficient user efforts.

In the ontology development scenario an automatic (or semi-automatic) bottom-up process of creating ontology from available job vacancies will be suggested.

The relevant skills (extracted from the job vacancies) will be defined and formalized. Using semantic annotation and cross-linguality techniques for skills extraction based on JSI Wikifier tool [6] will enable the possibility of including the newest available skills “on the market” that are not yet captured in the ontologies, taxonomies and classifications that are manually developed. The input to the ontology development scenario is a set of job vacancies and the output is ontology of skills presenting the domain structure that can be compared to or used for official classifications.

4.3.3 SKILLS ONTOLOGY EVOLUTION

Ontology Evolution is the timely adaptation of an ontology to the arisen changes and the consistent propagation of these

changes to dependent artefacts [9]. Ontology evolution is a process that combines a set of technical and managerial activities and ensures that the ontology continues to meet organizational objectives and users' needs in an efficient and effective way.

Ontology management is the whole set of methods and techniques that is necessary to efficiently use multiple variants of ontologies from possibly different sources for different tasks [10].

Scenario 3 will suggest an automatic (or semi-automatic) ontology evolution process based on the real-time job vacancy stream. With respect to the nature of job vacancy data stream and skills extracted from job it will be possible to see the dynamics of evolving skills – when the new skills (not included into the current ontology versions appear) and how the skills ontology is changing with time.

In particular, it could be possible to observe appearing new skills and suggest them for inclusion into official skills classifications. In addition, it could be visible how fast the ontology changes, which could be the indicator of the technological progress on the relevant market.

For instance, the current version of ESCO classification does not contain "TensorFlow" skill (TensorFlow [11] is an open-source software library for dataflow programming across a range of tasks, appeared in 2015). TensorFlow, which is already present in job vacancies, could be captured during ontology evolution process and suggested as a new concept for official classifications.

5. STATISTICAL INDICATORS

Traditionally the indicators related to labour market have been based on survey responses. The smart labour market statistics proposal introduces a possibility to complement standard statistical indicators, such as job vacancy rate with novel "data inspired" knowledge.

The smart labour market statistics indicators use data sources, previously not covered by official statistics, and in such way complementary to traditional data sources. The smart labour market statistics indicators are based on real-time data streams, which makes possible to obtain not only historical, but also current values for job vacancies that could be used for different purposes, such as nowcasting. In addition, the smart labour market statistics indicators take into the account data cross-lingual and multi-lingual nature of streaming data and can be produced at the multiscale levels – cross-country, country, city (area) levels.

The scenarios described above would result into a number of smart labour market indicators with multiscale options. In particular:

- Up-to date job vacancies statistics on a cross-country/country/city(area) level. Example: job vacancies in UK and France in the last month
- Up-to date skills statistics on a cross-country/country/city(area) level. Example: top 10 skills in UK in the last month
- Up-to date location statistics. Example: top locations for specific skill
- Ontology development statistics. Example: number of concepts in the ontology
- Ontology evolution statistics. Example: emerging skills in the ontology in the last 3 months

Since the data has a streaming nature, different kinds of multiscale and aggregation options can be handled with respect to time parameters.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a proposal for developing smart labour market statistics based on streams of enriched textual data, such as job vacancies from European countries. We define smart statistics scenarios, such as demand analysis scenario, skills ontology development scenario and skills ontology evolution scenario. The future work would include the implementation of the smart labour market scenarios, quality assessment and evaluation of the produced statistical outcomes.

7. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and EDSA European Union Horizon 2020 project under grant agreement No 64393.

8. REFERENCES

[1] EDSA, http://edsa-project.eu (accessed in August, 2018).

[2] Sibarani, Elisa & Scerri, Simon & Mousavi, Najmeh & Auer, Sören. (2016). Ontology-based Skills Demand and Trend Analysis. 10.13140/RG.2.1.3452.8249.

[3] ESCO taxonomy, https://ec.europa.eu/esco/portal (accessed in August 2018).

[4] Adzuna developer page, https://developer.adzuna.com/overview (accessed in August, 2018).

[5] Ratinov, L., Roth, D., Downey, D. and Anderson, M. Local and global algorithms for disambiguation to wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pages 1375–1384. Association for Computational Linguistics, 2011.

[6] JSI Wikifier, http://wikifier.org (accessed in May, 2018).

[7] GeoNames ontology, http://www.geonames.org/ontology/documentation.html (accessed in August, 2018).

[8] Ontology (by Tom Gruber), http://tomgruber.org/writing/ontology-definition-2007.htm (accessed in August, 2018).

[9] M. Klein and D. Fensel, Ontology versioning for the Semantic Web, Proc. International Semantic Web Working Symposium (SWWS), USA, 2001

[10] L. Stojanovic, B. Motik, Ontology evolution with ontology, in: EKAW02 Workshop on Evaluation of Ontology-based Tools (EON2002), CEUR Workshop Proceedings, Sigüenza, vol. 62, 2002, pp. 53–62

[11] TensorFlow, https://en.wikipedia.org/wiki/TensorFlow (accessed in August, 2018).

Relation Tracker - tracking the main entities and their relations through time

M. Besher Massri
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
besher.massri@ijs.si

Inna Novalija
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
inna.koval@ijs.si

Marko Grobelnik
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
marko.grobelnik@ijs.si

ABSTRACT
In this paper, we present Relation Tracker, a tool that tracks main entities [people and organizations] within each topic through time. The main types of relations between the entities are detected and observed in time. The tool provides multiple ways of visualizing this information with different scales and durations. The tool uses events data from Event Registry as a source of information, with the aim of getting holistic insights about the searched topic.

KEYWORDS
Information Retrieval, Visualization, Event Registry, Wikifier, Dmoz Taxonomy

1. INTRODUCTION
Every day, tremendous amounts of news and information are being streamed throughout the Internet, which is requiring the implementation of more tools to aggregate this information. With technology advancement, those tools have been increasing in complexity and options provided. However, there has been a demand for tools that give simple yet holistic summary of the searched topic in order to acquire general insights about it.

Hence, we provide the Relation Tracker tool that tries to achieve this goal; it is based on the data from Event Registry [1], which is a system for real-time collection, annotation and analysis of content published by global news outlets. The tool presented in this paper takes the events and groups them into topics, and within each topic, it provides an interactive graph that shows the main entities of each topic at each time and the main topic of relations between those entities. In addition, a summary information about entities and their relationship is visualized through different graphs to help understand more about the topic.

The remainder of this paper is structured as follows. In section 2, we show the related work done in this area. In section 3, we provide a description of the used data. Section 4 explains the methodology and main challenges that were involved in this work. Next, we explain the visualization features of the tool in section 5. Finally, we conclude the paper and discuss potential future work.

2. RELATED WORK
Similar works have been done in the area of visualizing information extracted from news. We see in [2] a tool for efficient visualization of large amount of articles as a graph of connected entities extracted from articles, enriched with additional contextual information provided as characteristic keywords, for a quick detection of information from the original articles.

Regarding classifying news, we observe in [3] a new technique that uses Deep Learning to increase the accuracy of prediction of online news popularity.
In the paper explaining Event Registry [1], we see how articles from different languages are grouped into events and the main information and characteristics about them are extracted. Additionally, a graphical interface is implemented which allows search for events and visualize the results in multiple ways that together give a holistic view about events.

This work begins with the events as a starting point, and it is one more step on the same path; it groups events further into topics and trends, then it focuses on tracking how some entities are appearing as main entities regarding the selected topic, and how the relationship between them is changing through time.

3. DESCRIPTION OF DATA
We used part of the events from Event Registry as our main source of data. We obtained a dataset of ~ 1.8 million events as a list of JSON files, with event's dates between Jan 2015 and July 2016. Each event consists of general information like title, event date, total article count, etc., and a list of concepts that characterize the event, which is split into entity concepts and non-entity concepts. Entity concepts are people, organizations, and locations related to the event. Whereas non-entity concepts represent abstract terms that define the topic of the event, like technology, education, and investment. Those concepts were extracted using JSI Wikifier [4] which is a service that enables semantic annotation of the textual data in different languages. In addition, each concept has a score that represents the relevancy of that concept to the event.

4. METHODOLOGY

4.1 Clustering and Formatting Data
To group the events into topics, we used K-Means clustering algorithm, where each event is represented as a sparse vector of the non-entity concepts it has, with the weights equal to their scores in that event. The constant number of topics is set experimentally to be 100 clusters, in a balance between mixed clusters and repeated clusters. Each cluster describes a set of events that fall under the same topic, whereas the centroid vector of each cluster represents the main characteristics of it. To name

the clusters, we used category classifier service from Event Registry, which uses Dmoz Taxonomy [5], a multilingual open-content directory of World Wide Web links, that is used to classify texts and webpages into different categories; for each cluster, we formed a text consisting of the components of its centroid vector, taking into account their weights within the vector. The resulted cluster names were ranged from technology and business to refugees and society, and clusters were exported as a JSON file for processing them in the visualization part.

4.2 Choosing the Main Entities

Under any topic, the top entities at each duration of time has to be chosen. At first, the concepts were filtered from outliers like publishers and news agencies. Then, an initial importance value has been set for each concept based on two parameters: the TF-IDF score of concept with respect to each event, and the number of articles each event contains. If we denote the set of events that occur in the interval of time D by E_D, the number of articles that event e contains is A_e, the TF-IDF score of concept c at event e by $S_{c,e}$, then the importance value of each item with respect to the interval D is calculated by the formula:

$$Imp_{init}(c)_D = \sum_{\substack{e \in E_D \\ e \text{ has concept } c}} S_{c,e} \; * \sum_{\substack{e \in E_D \\ e \text{ has concept } c}} A_e$$

The TF-IDF function is used to give importance to the concept based on its relevance to the events, and the number of articles is used to give more importance to the events that have more articles talking about it, and hence, more importance to the concepts that it has. We decided on using the product of summation rather than summation of product because of its computation efficiency while still producing good results. However, to prevent the case where all the chosen entities get nominated because of one or two big events (which results in a bias towards those few events), a modification to the importance value formula has been made by introducing another parameter, which is the links between concepts (whenever two concepts occur in the same event, there is a link between them). Each concept now affects negatively the other concepts it is linked to by an amount equal to the initial importance value divided by the number of neighbors. If we denote the set of neighbors of concept c during the interval of time D by $N_{c,D}$, then the negative importance value is defined by:

$$Imp_{neg}(c)_D = \sum_{c' \in N_{c,D}} \frac{Imp_{init}(c')_D}{|N_{c',D}|}$$

The final score is just the initial importance value minus the negative importance value, which is then used to sort and nominate the top entities.

$$Imp_{final}(c)_D = Imp_{init}(c)_D - Imp_{neg}(c)_D$$

4.3 Detecting the Characteristics of Relationship

The main goal was to model the relationship between any two entities through a vector of words where two entities are collocated. Since the relationship between two entities at any given time is based on the shared events between them, and each event is characterized by a set of concepts, we decided on using those concepts - specifically the abstract or the non-entity concepts - to characterize such relationships. For each pair, we aggregated all the non-entity concepts from the shared events between them, and each one of them was assigned a value based on the number of events it is mentioned in and its score in those events. Those concepts were sorted and ranked depending on their values, and the top ones were chosen as the main features of the relationship. In addition, these values of the concepts were used to rank the shared events and extract the top ones; by giving each event a value equal to the aggregated values (the ones calculated in previous step) of all non-entity concepts it has. To summarize the set of characteristics, we classified them using Dmoz category classifier in a similar way to what we have done in determining the names of the clusters. These categories were used to label the relationship between the entities, indicating the main topic of the shared events between them.

5. VISUALIZING THE RESULTS

To access a topic, a search bar is provided to select among the list of extracted topics from clustering step. Once the user selects a topic, a default date is chosen and a network graph is shown explaining the topic.

5.1 Characteristics of the Main Graph

Since the tool's main goal is to show the top entities and their relations, the network graph is the best choice for this matter. Following that, we have built an interactive network graph that has the following features:

- The main entities within that topic at the selected interval of time are represented by the vertices of the graph.
- The size of the vertices reflects the importance value of each entity, scaled to a suitable ratio to fit in the canvas.
- The colors represent the type of the entity, whether it is a person [red] or an organization [blue].
- The links between the entities represent the existence of shared events in that interval of time between them under that topic, and hence indicating some form of relations. The thickness of the links is proportional to the number of shared events, whereas the labels are the ones calculated in previous section.

Figure 1 presents top companies with relevant relations in July 2015 found among business news.

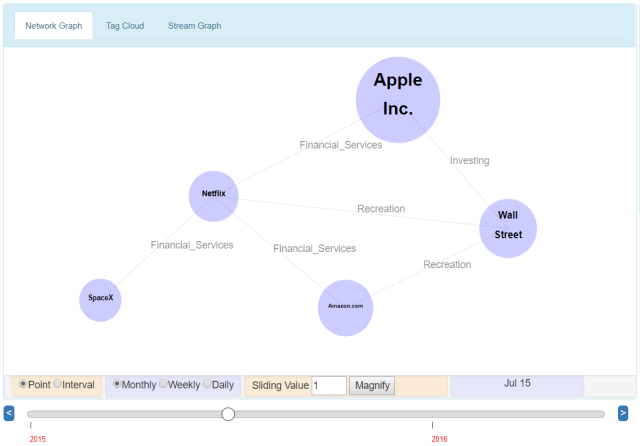


Figure 1: Top companies in July 2015 and their relations under the business topic.

5.2 Main Functionality

As the tool is concerned about tracking the changes with time. The graph is supported with a slide bar that allows the user to choose from the dates where there is at least one event occurred with respect to the selected topic. Different scales for moving dates are also provided; the user can choose to move day by day, week by week, or month by month and see the changes accordingly. In addition, the user can choose a specific interval of time, and track how the entities and their relations are changing when the interval moves slightly with respect to its length. An interval magnifier is also given if the user wants to get a closer look at the changes that happen in a small interval.

An example illustrating that can be seen in Figures 2 and 3. In Figure 2, we see the top 10 entities under the refugee topic in the last two months of 2015. When the interval is moved by 15 days, we notice that some of the entities disappear, like European Commission, indicating that they are no longer among the top 10 entities, whereas "United States House of Representitive" entity emerges and connects to "Barack Obama" and "Republican Party". The change in size indicates the change in the importance value of each one, while Society is the general theme among all labels.

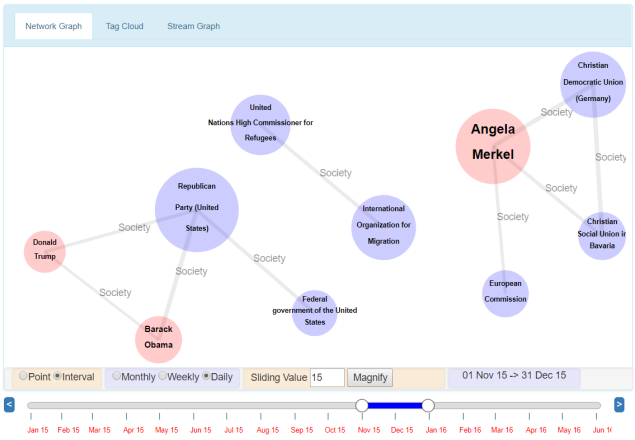


Figure 2: Top entities for the last two months of 2015 under the refugee topic.

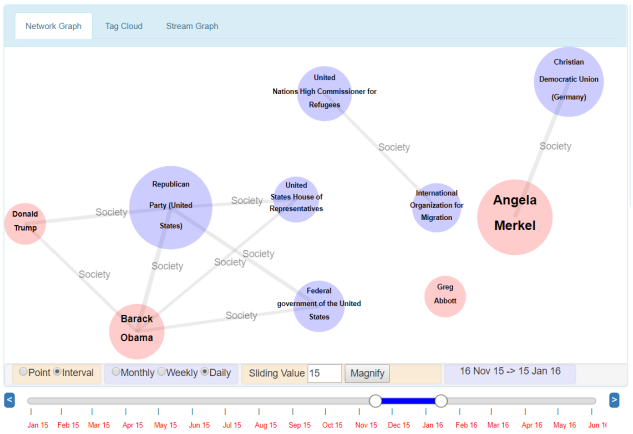


Figure 3: The changes in top entities under the same topic after moving the interval for 15 days.

5.3 Displaying Relation Information

Whenever the user selects a pair of entities, detailed information about their relationship in the selected interval of time is given, such as the number of shared events and articles, along with the top events both concepts were mentioned in. Also, the top shared characteristics that shape the relationship between them at this period is shown and sorted by percentage of importance. As seen in Figure 4; when selecting Jeff Bezos and Elon Musk under the space topic between January and September 2015, we see a list of the top events that involve both of them during this period. We see also that the relationship between them is mainly about sending astronauts by rockets to the international space station, as it can be understood from the top shared characteristics.

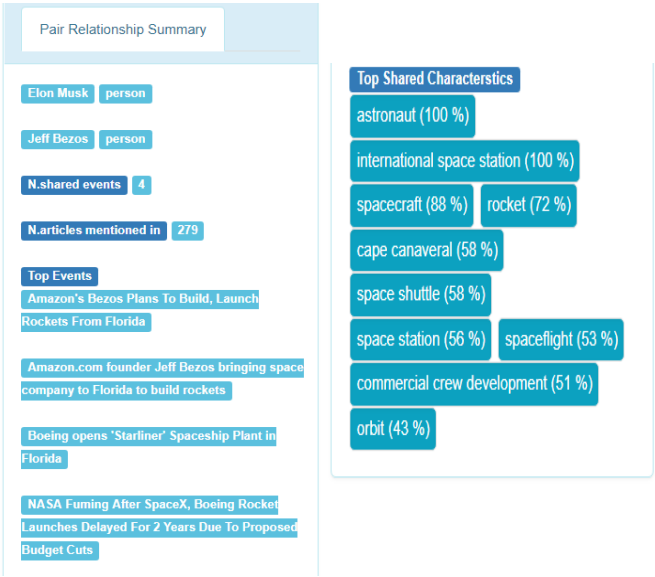


Figure 4: Relationship summary about Jeff Bezos and Elon Musk between January and September 2015 under the Space topic.

To illustrate how the importance of those top features with respect to the relationship is changing through time, a stream graph is used as shown in Figure 5.

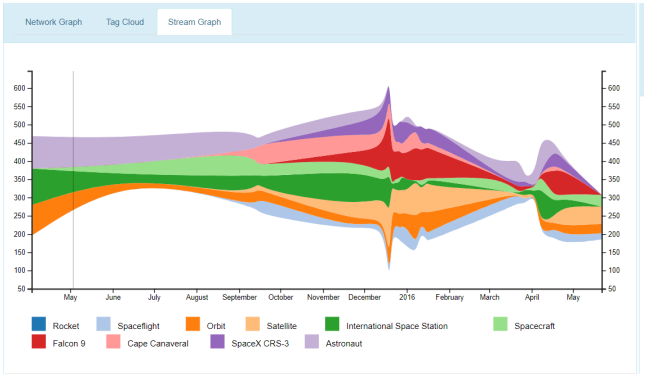


Figure 5: Stream graph showing how the effect of the main features on the relationship between Jeff Bezos and Elon Musk is changing through time.

Finally, the set of all characteristics that affect the relationship is visualized in a tag cloud to give a big picture about it. Figure 6 shows the tag cloud of the same relationship mentioned above.

Figure 6: Tag cloud illustrating a general view about all the characteristics that affects the relationship between Jeff Bezos and Elon Musk under the space topic.

6. CONCLUSION AND FUTURE WORK

In this paper, we provide a tool that uses events data from Event Registry to show the main entities within each topic, and how the characteristics of relationship among them is changing through time. However, there are a couple of limitation to the tool that we want to improve in the future. Although we were able to detect the characterestics of the relationship between entities and how they are changing through time, the main type of relation that we used to label the links were very broad and hence rarely changing- improving the methodology for relation extraction and observation of relations in time will be the subject of future work. In addition, we limited the search space for topics for the 100 topics we obtained from clustering, we would like to generalize the search by enabling searches for any concept or keyword with different options to filter the search.

7. ACKNOWLEDGMENTS

This work was supported by the euBusinessGraph (ICT-732003-IA) project [6].

8. REFERENCES

[1] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. 2014. Event registry: learning about world events from news. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion). ACM, New York, NY, USA, 107-110. DOI: https://doi.org/10.1145/2567948.2577024

[2] Marko Grobelnik and Dunja Mladenić. 2004. Visualization of news articles. Informatica 28.

[3] Sandeep Kaur and Navdeep Kaur Khiva. 2016. Online news classification using Deep Learning Technique. IRJET 03/10 (Oct 2016).

[4] Janez Brank, Gregor Leban and Marko Grobelnik. 2017. Annotating documents with relevant Wikipedia concepts. In Proceedings of siKDD2017. Ljubljana, Slovenia.

[5] Dmoz, open directory project, http://dmoz-odp.org/ (accessed in July, 2018)

[6] euBusinessGraph project, http://eubusinessgraph.eu/ (accessed in July, 2018).

Cross-lingual categorization of news articles

Blaž Novak
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenia
+386 1 477 3778
blaz.novak@ijs.si

ABSTRACT
In this paper we describe the experiments and their results performed with the purpose of creating a model for automatic categorization of news articles into the IPTC taxonomy. We show that cross-lingual categorization is possible using no training data from the target language. We find that both logistic regression and support vector machines are good candidate models, while random forests do not perform acceptably. Furthermore, we show that using Wikipedia-derived annotations provides more information about the target class than using generic word features.

General Terms
Algorithms, Experimentation

Keywords
News, articles, categorization, IPTC, Wikifier, SVM, Logistic regression, Random forests.

1. INTRODUCTION
The JSI Newsfeed [1] system ingests and processes approximately 350.000 news articles published daily around the world, in over 100 languages. The articles are automatically cleaned up and semantically annotated, and finally stored and made available for downstream consumers.

One of the annotation tasks that we would like to perform in the future is to automatically categorize articles into the IPTC "Media Topics" subject taxonomy [2]. IPTC – the International Press Telecommunications Council – provides a standardized taxonomy of roughly 1100 terms, arranged into a 5 level taxonomy, describing subject matters relating to daily news. The vocabulary is accessible in a machine readable format – RDF/XML and RDF/Turtle – at http://cv.iptc.org/newscodes/mediatopic.

There are two relations linking concepts in the vocabulary – the 'broader concept' taxonomical relation, and a 'related concept' sibling relation. The 'related concept' links concepts both to other concepts from the same taxonomy, and directly to external Wikidata [3] entities.

The purpose of this work is to evaluate multiple machine learning algorithms and multiple sets of features with which we could automatically perform the categorization. As we would like to categorize articles in all the languages the Newsfeed system supports, but we only have example articles in English and French, the method needs to be language independent.

2. EXPERIMENTAL SETUP
The dataset that we have access consists of 30364 English and 29440 French articles, each of which is tagged with 1 to 10 categories. We consider each document belonging to all categories that are explicitly stated, and all of their parents. We will compare the performance of model predictions on the same language and in the cross-lingual setting, where we train the model on the entire dataset available for one language, and measure its performance on the other language.

Basic features of the dataset can be seen in the following 2 figures. Figure 1 shows the distribution of number of articles in each category, and Figure 2 shows that most categories contain a roughly even number of articles in both languages, but there are some outliers. We ignored categories with less than 15 examples per language, which resulted in 308 categories.

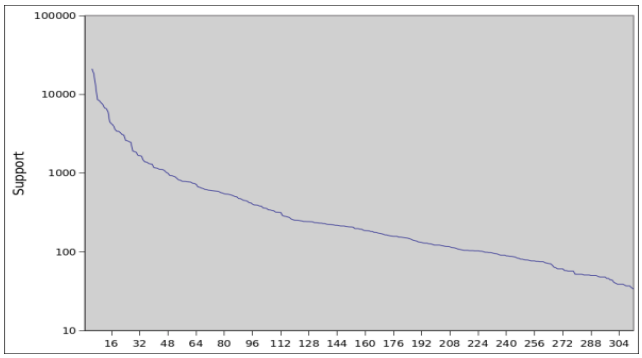


Figure 1. Number of articles in each category. Discrete categories on x axis are ordered by descending number of articles.

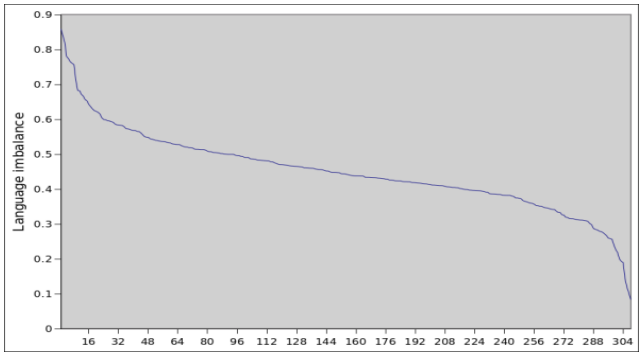


Figure 2. Language imbalance for each category. Discrete categories on x axis are ordered from "mostly English" to "mostly French".

We compare three different machine learning models – random forests, logistic regression (LR), and Support Vector Machines (SVM).

We try two different types of features, and their combinations.

The first kind of a feature set we use is a projection of the bag-of-words representation of the document text into a 500 dimensional vector space. The KCCA [4] method uses an aligned multi-lingual corpus to find such a mapping, that words with similar meanings map to a similar vector, regardless of their language. We represent a document as a sum of all word vectors.

The second set of features we use is the output of the JSI Wikifier [5] system. The Wikifier links each word in a document to a set of Wikipedia pages that might represent the meaning of that word. For each such annotation, we also get a confidence weight.

We consider these annotations as a classical vector space model -- as a bag-of-entities. We use two versions of the TF-IDF [7] scheme: in the first case, we use the number of times an entity annotation is present for any word in a document as the TF (term frequency) factor, and in the second version, we use the sum of annotation weights of an entity across the document. In both cases, we perform L1 normalization of the vector containing TF terms. For IDF terms, we use $\log\left(1+\frac{N}{n}\right)$ where N is the number of all documents and n the number of documents where an annotation was present at least once.

Finally, we use a combination of both KCCA-derived and Wikifier-derived features as the last feature set option.

For model training, we use Pythons scikit-learn [6] software package. In the case of logistic regression, we use L2 penalty, with automatic decision threshold fitting, using the liblinear library backend.

For the SVM model, we use a stochastic gradient descent optimizer. We performed a grid search for the optimal regularization constant C, but since there were no significant accuracy changes, we used the default of 1.0 in all other experiments.

For the random forest model, we used 4 different parameter combinations:

- default – 10 trees, splitting until only one class is in the leaf
- 30 trees, maximum tree depth of 10
- 50 trees, maximum tree depth of 10
- 30 trees, maximu tree depth of 20

In all cases, GINI index was used as the node splitting criterion.

Since the majority of categories only have a small number of documents, we automatically weighed training examples by the inverse of their class frequency. We also performed some experiments without this weighting scheme, but got useless models in all cases except for the couple largest categories.

All reported results are the average of a 3-fold cross-validation.

So far, we only created one-versus-all models for each category independently, and only used the taxonomy information of categories to select all examples from sub-categories when training the more general category.

3. RESULTS

Table 1 shows ROC scores for cross-validation of all three models on four sets of feature combinations, for English and French separately. SVM and logistic regression are comparable in behavior and promising, while the random forest model performs significantly worse. "Wiki-W" denotes the weighted version of Wikifier annotations, and "Wiki-K" the combination of KCCA-derived features and Wikifier annotations. Every second line in the table is the standard deviation of the result when averaged across all categories.

Table 1. ROC scores by model and feature type, cross-validation

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.75	0.71	0.96	0.95	0.95	0.94
(stdev)	0.11	0.11	0.04	0.04	0.05	0.04
Wiki	0.70	0.70	0.95	0.95	0.94	0.94
(stdev)	0.12	0.12	0.04	0.04	0.05	0.04
Wiki-W	0.71	0.71	0.95	0.95	0.94	0.94
(stdev)	0.12	0.11	0.04	0.04	0.05	0.04
Wiki+K	0.71	0.69	**0.97**	**0.96**	0.96	0.95
(stdev)	0.12	0.11	0.03	0.03	0.03	0.04

Looking at the feature selections, we see almost no significant difference -- both kinds of features -- KCCA and Wikipedia annotations have useful predictive value. The combination of both feature types slightly improves the ROC score.

Table 2 shows F1 cross-validation scores of all three models. Logistic regression scores much higher than SVM here, possibly indicating that the SVM model would benefit from a post-processing step of optimizing the decision threshold on a separate training set.

Table 2. F1 scores by model and feature type, cross-validation

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.16	0.12	0.30	0.25	0.20	0.18
(stdev)	0.21	0.18	0.21	0.20	0.21	0.19
Wiki	0.07	0.07	0.41	0.44	0.25	0.29
(stdev)	0.15	0.15	0.21	0.21	0.22	0.22
Wiki-W	0.08	0.08	0.40	0.43	0.24	0.28
(stdev)	0.17	0.17	0.21	0.21	0.21	0.22
Wiki+K	0.09	0.07	**0.44**	**0.46**	0.27	0.30
(stdev)	0.16	0.15	0.21	0.21	0.22	0.22

The combination of both feature sets performs significantly better than either alone, with generic word-based features providing the least amount of information.

The feature usefulness changes when looking at cross-lingual classification performance. Table 3 shows the ROC score for all three models, when the model trained on English is used to predict categories of French articles, and vice versa. Decision trees give essentially a random result, and SVM scores somewhat higher than logistic regression.

Table 3. ROC scores - cross-lingual classification

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR

KCCA	0.50	0.50	0.50	0.50	0.50	0.51
(stdev)	0.00	0.00	0.01	0.03	0.04	0.08
Wiki	0.51	0.51	0.76	0.80	0.81	0.84
(stdev)	0.04	0.04	0.12	0.11	0.11	0.10
Wiki-W	0.51	0.52	0.78	0.82	**0.82**	**0.84**
(stdev)	0.04	0.05	0.11	0.10	0.10	0.10
Wiki+K	0.50	0.50	0.57	0.70	0.66	0.81
(stdev)	0.01	0.01	0.10	0.13	0.14	0.12

The biggest change here is the influence of KCCA cross-lingual word embedding: by itself it provides no informative value, as indicated by ROC value of 0.5 in all cases, and it even reduces the performance of the combined Wikifier + KCCA model.

In the Table 4, F1 scores from the same experiment are shown. Logistic regression still has a big advantage over SVM, as in the same-language categorization setting. The change from previous experiments is the influence of weighting of Wikipedia features -- it increases the performance of all models.

Table 4. F1 scores - cross-lingual classification

	Rand. Forest		Log. Reg.		SVM	
	EN	FR	EN	FR	EN	FR
KCCA	0.00	0.00	0.00	0.01	0.00	0.02
	0.02	0.02	0.02	0.06	0.01	0.06
Wiki	0.03	0.04	0.48	0.44	0.30	0.26
	0.10	0.11	0.21	0.20	0.22	0.22
Wiki-W	0.03	0.05	**0.49**	**0.44**	0.29	0.26
	0.11	0.13	0.20	0.21	0.22	0.22
Wiki+K	0.00	0.00	0.18	0.40	0.20	0.23
	0.04	0.04	0.22	0.22	0.19	0.21

An interesting observation is that the performance of the cross-lingual model is occasionally higher than that of the baseline cross-validation experiment. This anomaly however disappears for categories with large amount of positive training examples. It also disappears if we reduce the amount of training examples in the cross-lingual experiment by 1/3 – the effect seems to be caused by cross-validation reducing the training dataset size.

KCCA cross-lingual word embedding feature generation used here was tested in other experiments and systems and gives a useful feature set for comparison of documents across languages, so its negative impact on the performance of these models needs to be investigated in the future.

As the weighted Wikipedia feature set appears to be the best for the stated goal of cross-lingual article categorization, the results of next experiments are shown only for it, but we performed the same experiments on all other combinations, and the results broadly follow the conclusions from the previous section.

The following figures show correlation of testing and cross-lingual performance of logistic regression and SVM models. Both F1 score and area under ROC curve are shown for each of 308 categories in the experiment, since they provide complementary information. As the figures show, there is a good agreement between the cross-validation and the cross-lingual classification performance, giving us an ability to estimate cross-lingual performance based on the cross-validation score in the production environment. The difference between distributions for French and English language models is consistent with the class imbalance for each of the categories.

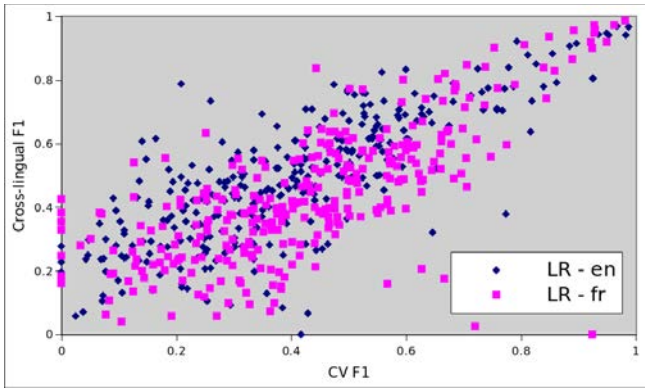


Figure 3. F1 score correlation for logistic regression

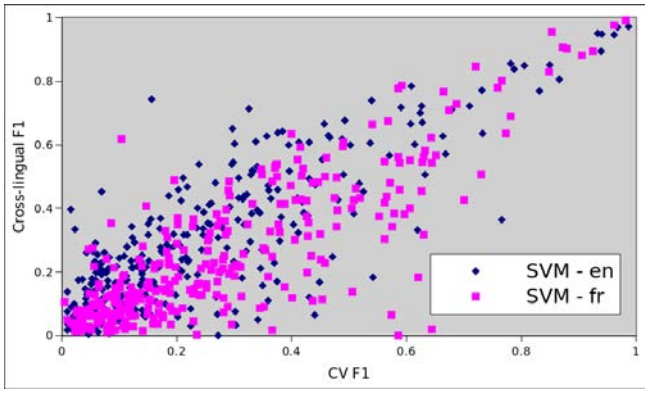


Figure 4. F1 score correlation for SVM

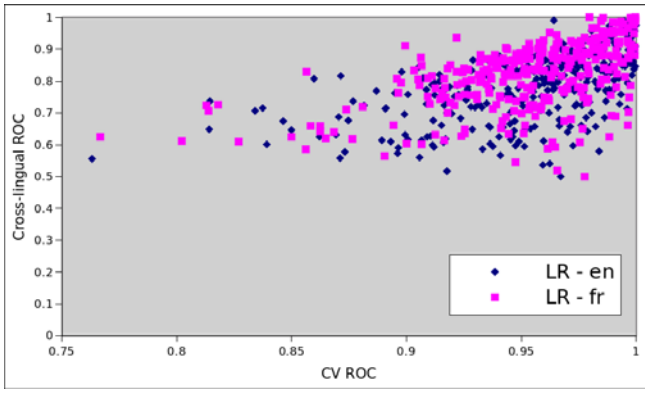


Figure 5. ROC score correlation for logistic regression

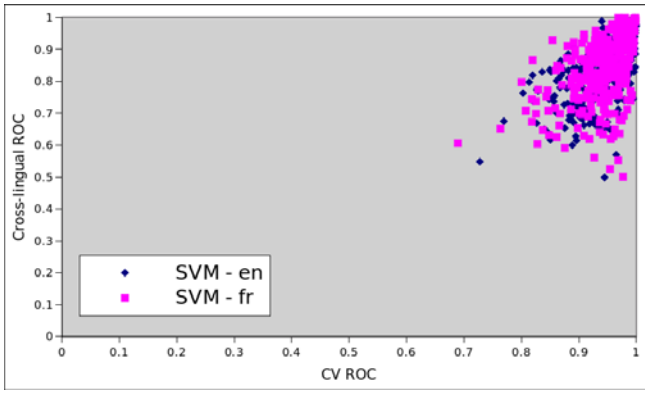


Figure 6. ROC score correlation for SVM

The SVM model seems to have a more consistent behavior, so we will use it in the final application instead of logistic regression.

Figures 7 through 10 show the F1 and ROC score behavior of logistic regression and SVM models for cross-validation and cross-lingual classification with regard to the number of positive examples in the category, separately for English and French language. While the SVM model underperforms on the F1 metric on average, it produces a better ranking of documents with respect to a category, as seen on ROC plots, especially for smaller categories. This further indicates the need for decision threshold tuning in the SVM model before we use its predictions.

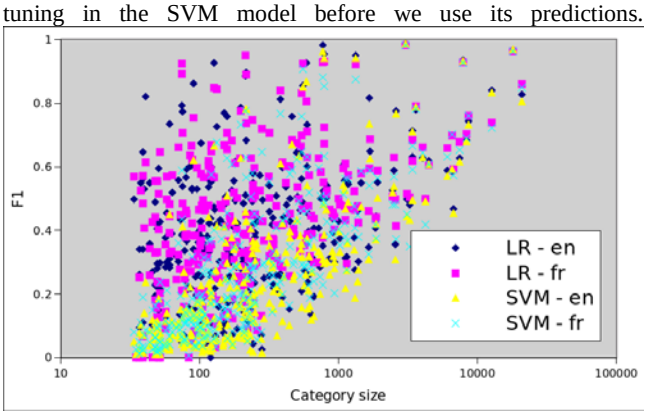


Figure 7. F1 score with respect to category size, cross-validation

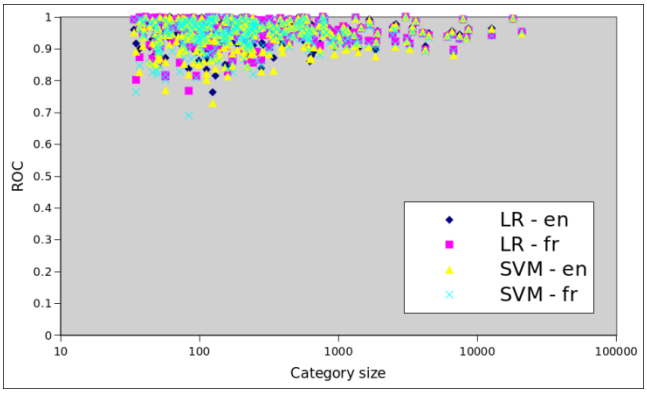


Figure 8. ROC score with respect to category size, cross-validation

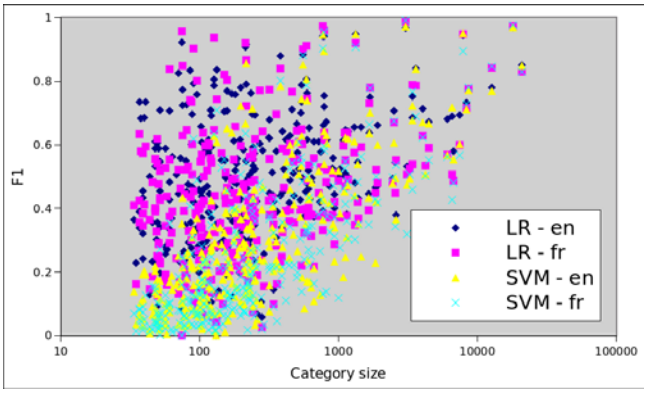


Figure 9. F1 score with respect to category size, cross-lingual prediction

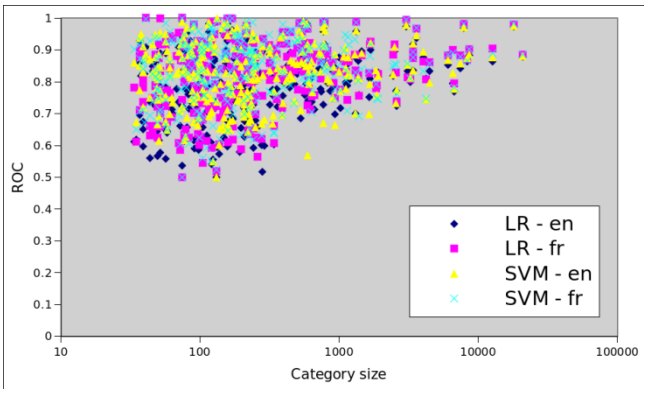


Figure 10 ROC score with respect to category size, cross-lingual prediction

As expected, classification performance of all models improves with the number of training examples, but in cases of small categories, it appears that some are much easier to learn than others.

4. CONCLUSIONS AND FUTURE WORK

We found that using a logistic regression model with weighted Wikifier annotations gives us a good enough result to use IPTC category tags as inputs for further machine processing in the Newsfeed pipeline. Before we can use this categorization for human consumption, we need to investigate automatic tuning of SVM decision thresholds on this problem, and add an additional filtering layer that takes into consideration interactions between categories beyond the sub/super-class relation. Additionally, the negative effect of KCCA-derived features for cross-lingual annotation needs to be examined.

5. ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency as well as the euBusinessGraph (ICT-732003-IA) and EW-Shopp (ICT-732590-IA) projects.

6. REFERENCES

[1] Trampuš M., Novak B., “The Internals Of An Aggregated Web News Feed” Proceedings of 15th Multiconference on Information Society 2012 (IS-2012).

[2] https://iptc.org/standards/media-topics/

[3] https://www.wikidata.org/wiki/Wikidata:Main_Page

[4] Rupnik, J., Muhič, A., Škraba, P. “Cross-lingual document retrieval through hub languages”. NIPS 2012, Neural Information Processing Systems Worshop, 2012

[5] Brank J., Leban G. and Grobelnik M. “Semantic Annotation of Documents Based on Wikipedia Concepts”. Informatica, 42(1): 2018.

[6] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. “Scikit-learn: Machine Learning in Python”. Journal of Machine Learning Research, 12. 2011, pp. 2825-2830.

[7] K. Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". Journal of Documentation, 28 (1). 1972

Transporation mode detection using random forest

Jasna Urbančič
Artificial Intelligence Laboratory,
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana,
Slovenia
jasna.urbancic@ijs.si

Veljko Pejović
Faculty of Computer and Information science,
University of Ljubljana
Večna pot 113, 1000 Ljubljana
Slovenia
veljko.pejovic@fri.uni-lj.si

Dunja Mladenić
Artificial Intelligence Laboratory,
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana,
Slovenia
dunja.mladenic@ijs.si

ABSTRACT

This paper addresses transportation mode detection for a mobile phone user using machine learning and based on mobile phone sensor data. We describe our approach to data collection, preprocessing and feature extraction. We evaluate our approach using random forest classification with focus on feature selection. We show that with feature selection we can significantly improve classification scores.

1. INTRODUCTION

In the recent years we have witnessed a drastic increase in sensing and computational resources that are built in mobile phones. Most of modern cell phones are equipped with a set of sensors containing triaxial accelerometer, magnetometer, and gyroscope, in addition to having a Global Positioning System (GPS). Smart phone operating system APIs offer activity detection modules that can distinguish between different human activities, for example: being still, walking, running, cycling or driving in a vehicle [2, 3]. However, APIs cannot distinguish between driving in different kind of vehicles, for example driving a car or traveling by bus or by train. Recognizing different kind of transportation, also known as transportation mode detection, is crucial for mobility studies, for routing purposes in urban areas where public transportation is often available, for facilitating the users to move towards more environmentally sustainable forms of transportation [1], or to inspire them to exercise more.

In this paper we discuss the use of random forest in transportation mode detection based on accelerometer signal. We focus on

1. feature extraction, and
2. feature analysis to determine the most meaningful features for this specific problem and the choice of classifier.

Our main contribution is feature analysis, which revealed the impact of each feature to the classification scores.

2. RELATED WORK

While the first attempts to recognize user activity were initiated before smart phones, the real effort in that direction begun with the development of mobile phones having built-in sensors [10], including GPS and accelerometer sensors. There are still some studies that use custom loggers to collect the data [11, 17] or use dedicated devices as well as smart phones [5]. Although GSM triangulation and local area wireless technology (Wi-Fi) can be employed for the purpose of transportation mode detection, their accuracy is relatively low compared to GPS [11], so latest state of the art research is focused on transportation mode detection based on GPS tracks and/or accelerometer data.

Machine learning approaches for transportation mode detection often rely on statistical, time-based, frequency-based, peak-based and segment-based [8] features, however in most cases statistical features and features based in frequency are used [10, 11, 16]. Feature domains are described in Table 1. Statistical, time-based, and spectral attributes are computed on a level of a time frame that usually covers a few seconds, whereas peak-based features are calculated from peaks in acceleration or deceleration. On the other hand, segment-based features are computed on the recordings of the whole trip, which means that they cover much larger scale. Statistical, time-based, and spectral features are able to capture the characteristics of high-frequency motion caused by user's physical movement, vehicle's engine and contact between wheels and surface. Peak-based features capture movement with lower frequencies, such as acceleration and breaking periods, which are essential for distinguishing different motorized modalities. Additionally, segment-based features describe patterns of such acceleration and deceleration periods [8].

Machine learning methods that are most commonly used in accelerometer based modality detection include support vector machines, decision trees and random forests, however naive Bayes, Bayesian networks and neural networks have been used as well [11, 12]. Often these classifiers are used in an ensemble [16]. The majority of algorithms additionally use Adaptive Boosting or Hidden Markov Model to improve the performance of the methods mentioned above [16, 8, 11, 10]. In the last years, deep learning has also been used [6, 14].

Accelerometer-only approach where only statistical features have been used reported 99.8% classification accuracy, however users were instructed to keep the devices fixed position during a trip. Furthermore, only 0.7% of data was labeled as train [11]. State of the art approach to accelerometer-only

Domain	Description
Statistical	These features are include mean, standard deviation, variance, median, minimum, maximum, range, interquartile range, skewness, kurtosis, root mean square.
Time	Time-based features include integral and double integral of signal over time, which corresponds to speed gained and distance traveled during that recording. Other time-based features are for example auto-correlation, zero crossings and mean crossings rate.
Frequency	Frequency-based features include spectral energy, spectral entropy, spectrum peak position, wavelet entropy and wavelet coefficients. These can be computed on whole spectrum or only on specific parts, for example spectral energy bellow 50Hz. Spectrum is usually computed using fast Fourier transform, whereas wavelet is a result of the Wavelet transformation. Entropy measures are based on the information entropy of the spectrum or wavelet [7].
Peak	Peak-based features use horizontal acceleration projection to characterize acceleration and deceleration periods. These features include volume, intensity, length, skewness and kurtosis.
Segment	Segment-based include peak frequency, stationary duration, variance of peak features, and stationary frequency. The latter two are similar to velocity change rate and stopping rate used by [17]. Segment-based features are computed on a larger scale than statistical, time-based or frequency-based features.

Table 1: Feature domains and their descriptions adopted from [8].

transportation mode detection relies on long accelerometer samples. It uses features from all five domains for classification with AdaBoost with decision trees as a weak classifier and achieves 80.1% precision and 82.1% recall [8].

The performance of transportation mode detection systems depends on the effectiveness of handcrafted features designed by the researchers, researcher's experience in the field affects the results. Thus, there have been approaches that use deep learning methods, such as autoencoder or convolutional neural network, to learn the features used for classification. Using a combination of handcrafted and deep features for classification with deep neural network resulted in 74.1% classification accuracy [15].

3. PROPOSED APPROACH

Work flow of the proposed approach is visualized in Figure 1. The first task is data collection. To collect data we use *NextPin* mobile library [4] developed by the Artificial Intelligence Laboratory at Jožef Stefan Institute. Library is embedded into two free mobile applications. The first one is *OPTIMUM Intelligent Mobility* [1]. *OPTIMUM Intelligent Mobility* is a multimodal routing application for three European cities — Birmingham, Ljubljana, and Vienna. The second one is *Mobility patterns* [4]. *Mobility patterns* is essentially a travel journal. Both applications send five second long accelerometer samples every time OS's native activity recognition modules, Google's ActivityRecognition API [2] for Android and Apple's CMMotionActivity API [3], detect that the user is traveling in a vehicle. We use that accelerometer samples for fine-grained classification of motorized means of transportation.

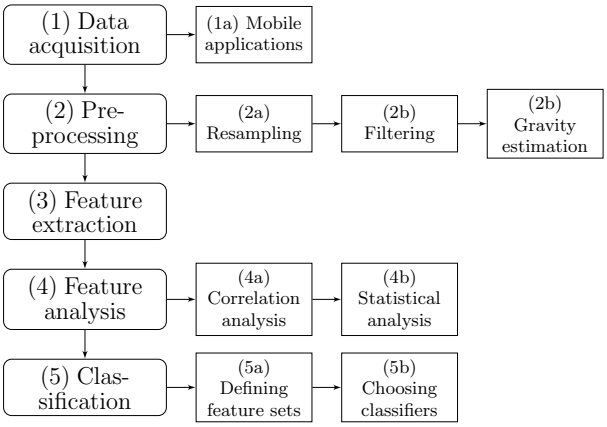


Figure 1: Detailed work flow diagram of the proposed approach. We stacked general, high-level tasks common in other approaches vertically, whereas subtasks specific to our approach are pictured horizontally.

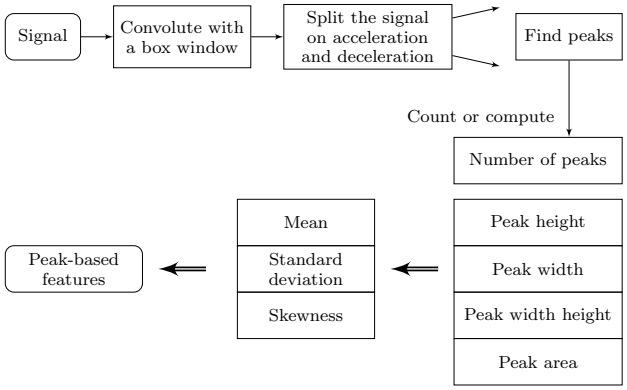


Figure 2: Work flows for extraction of peak-based features.

We collect five second samples of sensor data and resample them to sampling frequency 100 Hz in the preprocessing phase. Resampling ensures us that our samples all contain 500 measurements. The most prominent problem we face in preprocessing concerns the correlation of acceleration measurements with the orientation of the phone in the three dimensional space. Practically this means that gravity is measured together with the dynamic acceleration caused by phone movements. To eliminate gravity from the samples we perform gravity estimation on raw accelerometer signal. We follow an approach proposed by Mizell [9]. Gravity estimation splits the acceleration to static and dynamic component. Static component represents the constant acceleration, caused by the natural force of gravity, whereas dynamic component is a result of user's motion. Furthermore, using this approach we are able to calculate vertical and horizontal components of acceleration.

We use preprocessed signal to extract features for classification. Features are divided into five domains, based on their meaning and method of computation. We have listed the domains in Table 1. We extract features from three domains — statistical, frequency, and peak. We extract statistical features (maximal absolute value, mean, standard deviation, skewness, 5th percentile, and 95th percentile) from dynamic acceleration and its amplitude, horizontal acceleration and

Set	Accele.	Features	Size
D-S	Dynamic	Statistical	54
D-SF	Dynamic	Statistical, Frequency	94
D-SFP	Dynamic	Statistical, Frequency, Peak	172
H-S	Horizontal	Statistical	54
H-SF	Horizontal	Statistical, Frequency	94
H-SFP	Horizontal	Statistical, Frequency, Peak	172
ALL			376

Table 2: Predefined feature sets used for classification.

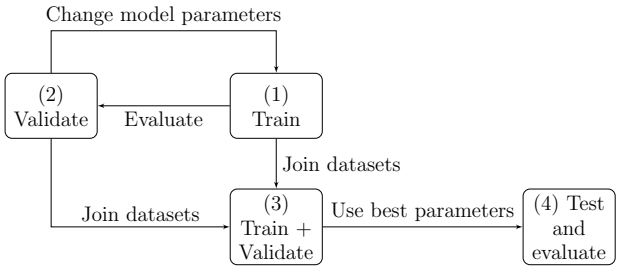


Figure 3: Schema of evaluation scenario.

its amplitude, amplitude of raw acceleration, and amplitude of vertical acceleration. From the same signals we extract frequency-based features using fast Fourier transformation. As frequency-based features we use the power spectrum of the signal aggregated in 5 Hz bins. Pipeline for extraction of peak-based features from dynamic and horizontal in acceleration is pictured in Figure 2. To extract peak-based features we first smooth out the signal with convolution with a box window and split it into moments of acceleration and moments of deceleration. Then we find peaks and compute peak heights, peak widths, peak width heights, and peak areas. As there is usually more than one peak we aggregate these values using mean, standard deviation, and skewness. All together we extract 376 features. We organize features into seven predefined feature sets we use for classification. Feature sets are listed in Table 2.

To evaluate the capabilities and performance of the proposed approach, we divide our dataset in 3 subsets — train, validation, and test set — based on the date the samples were recorded on. By doing so we avoided using in this domain methodologically questionable random assignment of samples collected during the same trip to different subsets. The reason why we did not apply cross-validation is similar. Using samples from the same trip in train and test set would result in significantly higher evaluation scores. For

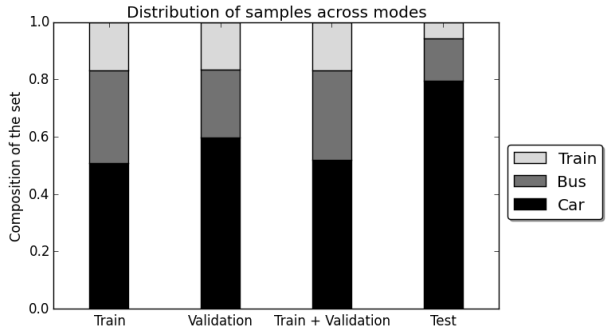


Figure 4: Distribution of modes in train, validation, and test set. We also added joint train and validation set, which we use to train the final model.

Feature set	CA	RE	PR	F1
D-S	0.48	0.41	0.39	0.37
D-SF	0.60	0.41	0.41	0.39
D-SFP	0.46	0.39	0.40	0.35
H-S	0.64	0.40	0.43	0.41
H-SF	0.53	0.39	0.43	0.36
H-SFP	0.50	0.37	0.40	0.34
ALL	0.47	0.35	0.40	0.33

Table 3: Classification metrics for classification with random forest on predefined feature sets.

the training set we use the data from [13], whereas validation and test sets were obtained during Optimum pilot testing in 2018. During validation step we are trying to maximize F1 score as our data set is imbalanced. We visualized the evaluation scenario in Figure 3, while the composition of the sets in pictured in Figure 4.

4. RESULTS

We trained random forest classifier on the predefined feature sets from Table 2. Classification metrics we report on include classification accuracy (CA), recall (RE), precision (PS) and F1 score (F1) Results are listed in Table 3. Table 3 shows that we achieved the highest F1 score of 0.41 using H-S feature set. This feature set consists of statistical features calculated on the horizontal acceleration vector. Classification accuracy in that case is also high, compared to other feature sets. The peak features seems to be the source of noise in the data, as using peak features in combination with the other features sets decreases the performance, for example F1 drops from 0.39 for D-SF to 0.35 for D-SFP.

F1 score and classification for dynamic acceleration increase when we add frequency-based features, whereas these two measures decrease in case of similar action for horizontal acceleration. This offers two possible interpretations. One is that frequency-based features of dynamic acceleration carry more information compared to frequency-based features of horizontal acceleration. The second one is that statistical features of horizontal acceleration are much better than statistical features from dynamic acceleration.

We noticed that smaller feature sets generally perform better than larger so we focused on feature selection. We initially train the model with all features and evaluate it on validation set. Then we remove each feature one by one, train the model, evaluate it on the validation set and compare all F1 scores. We eliminate the feature with the highest F1 score, as this means that when the model was trained without that feature if performed better than when the eliminated feature was included. We repeat this procedure until the feature set consists of one feature. Similarly, we do feature addition — we start with an empty feature set and gradually add features one by one.

Using the described process of forward feature selection and backward feature elimination we selected two feature sets that performed the best — in case of forward selection the best feature set has 10 features, whereas feature set produced with backward elimination has 28 features. Feature set obtained by forward selection mostly contains statistical features, followed by peak-based. Only one frequency-based features appears in that set. Additionally, features are in vast majority extracted from dynamic acceleration. On the other hand feature set obtained by backward elim-

Feature set	CA	RE	PR	F1
Forward selection (10)	0.70	0.50	0.47	0.48
Backward elimination (28)	0.73	0.50	0.48	0.49

Table 4: Classification metrics for classification with the selected features in feature selection.

Forward selection

T \P	Car	Bus	Train
Car	0.78	0.27	0.05
Bus	0.51	0.40	0.09
Train	0.47	0.21	0.32

Backward elimination

T \P	Car	Bus	Train
Car	0.83	0.12	0.05
Bus	0.55	0.35	0.10
Train	0.45	0.23	0.32

Table 5: Confusion matrix for classification with the selected features in feature selection.

ination contains more peak-based features than statistical, again only one frequency-based feature appears. Dynamic acceleration and horizontal acceleration appear in similar proportions. We evaluated the models trained with that feature sets against the test set. Results are listed in Table 4. Both feature sets achieve better F1 scores than any previous feature sets. Confusion matrix in Table 5 reveals what are the differences between these two feature sets. We can see that in case of eliminating features there is less cars missclassified as buses and more buses missclassified as cars. Classification of trains is fairly consistent. For buses and trains the largest part of samples is still missclassified as cars.

5. CONCLUSIONS

We showed that while transportation mode with random forest is possible, careful feature selection is necessary. Using feature selection we were able to improve classification scores for at least 0.04, in some cases even over 0.10. Although classification scores improved, most of non-car samples were still misclassified as cars. We observed that even though peak-based features did not perform as well in predefined feature sets, they appeared consistently among selected features in feature selection. That does not hold for frequency-based feature only one feature appeared among selected features. For the future work we suggest maximization of another classification score as we focused on maximization of F1 score.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency under project *Integration of mobile devices into survey research in social sciences: Development of a comprehensive methodological approach* (J5-8233), and the ICT program of the EC under project *OPTIMUM* (H2020-MG-636160).

7. REFERENCES

[1] Optimum project - European Union's Horizon 2020 research and innovation programme under grant agreement No 636160-2. `http://www.optimumproject.eu/`, 2017. [Online; accessed 4-November-2017].

[2] `ActivityRecognition`. `https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognition`, 2018. [Online; accessed 31-August-2018].

[3] `CMMotionActivity`. `https://developer.apple.com/library/ios/documentation/CoreMotion/Reference/CMMotionActivity_class/index.html#//apple_ref/occ/cl/CMMotionActivity`, 2018. [Online; accessed 31-August-2018].

[4] L. Bradeško, Z. Herga, M. Senožetnik, T. Šubic, and J. Urbančič. Optimum project: Geospatial data analysis for sustainable mobility. In *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining Project Showcase Track*. ACM, 2018. `http://www.kdd.org/kdd2018/files/project-showcase/KDD18_paper_1797.pdf`.

[5] K.-Y. Chen, R. C. Shah, J. Huang, and L. Nachman. Mago: Mode of transport inference using the hall-effect magnetic sensor and accelerometer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):8, 2017.

[6] S.-H. Fang, Y.-X. Fei, Z. Xu, and Y. Tsao. Learning transportation modes from smartphone sensors based on deep neural network. *IEEE Sensors Journal*, 17(18):6111–6118, 2017.

[7] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing*, 14(7):645–662, 2010.

[8] S. Hemminki, P. Nurmi, and S. Tarkoma. Accelerometer-based transportation mode detection on smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, page 13. ACM, 2013.

[9] D. Mizell. Using gravity to estimate accelerometer orientation. In *Proc. 7th IEEE Int. Symposium on Wearable Computers (ISWC 2003)*, page 252. Citeseer, 2003.

[10] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[11] M. A. Shafique and E. Hato. Use of acceleration data for transportation mode prediction. *Transportation*, 42(1):163–188, 2015.

[12] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM, 2011.

[13] J. Urbančič, L. Bradeško, and M. Senožetnik. Near real-time transportation mode detection based on accelerometer readings. In *Information Society, Data Mining and Data Warehouses SiKDD*, 2016.

[14] T. H. Vu, L. Dung, and J.-C. Wang. Transportation mode detection on mobile devices using recurrent nets. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 392–396. ACM, 2016.

[15] H. Wang, G. Liu, J. Duan, and L. Zhang. Detecting transportation modes using deep neural network. *IEICE TRANSACTIONS on Information and Systems*, 100(5):1132–1135, 2017.

[16] P. Widhalm, P. Nitsche, and N. Brändie. Transport mode detection with realistic smartphone sensor data. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 573–576. IEEE, 2012.

[17] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

FSADA, an anomaly detection approach

A modern, cloud-based approach to anomaly-detection, capable of monitoring complex IT systems

Viktor Jovanoski
Jozef Stefan International Postgraduate School
Jamova 39
Ljubljana, Slovenia
viktor@carvic.si

Jan Rupnik
Jozef Stefan Institute
Jamova 39
Ljubljana, Slovenia
jan.rupnik@ijs.si

ABSTRACT

Modern IT systems are becoming increasingly complex and inter-connected, spanning over a range of computing devices. As software systems are being split into modules and services, coupled with an increasing parallelization, detecting and managing anomalies in such environments is hard. In practice, certain localized areas and subsystems provide strong monitoring support, but cross-system error-correlation, root-cause analysis and prediction are an elusive target.

We propose a general approach to what we call *Full-spectrum anomaly detection* - an architecture that is able to detect local anomalies on data from various sources as well as creating high-level alerts utilizing background knowledge, historical data and forecast models. The methodology can be implemented either completely or partially.

Keywords

Anomaly detection, Outlier detection, Infrastructure monitoring, Cloud

1. INTRODUCTION

Modern IT systems need several key capabilities, apart from tracking and directing the underlying businesses. They need to manage errors and failures - predict them in advance, detect them in their early stages, help limit the scope of the damage and mitigate their consequences. All this is achieved by analyzing past and current data and detecting outliers in it.

Anomaly detection must happen in near-real time, while simultaneously analyzing potentially thousands of data points per second. Incoming data that such a system can monitor is very diverse. Data can come in different shapes (numeric, discrete or text), in regular time intervals or sporadically, in huge volumes or just a few data points per day. Designing a system that can cope with such diverse situations can be challenging.

Another important aspect is *"actionability"* of the reported anomalies. When human operator is presented with a new alert, the message as to what is wrong needs to be clear. The operator must be able to immediately start addressing the problem. Sometimes all we need is a different presentation of the result, but most often the easy-to-describe algorithms and models are used - e.g. linear regression or nearest neighbour.

This high velocity of data (volume and rate) makes some of the algorithms less usable in such scenarios - specifically batch processing that requires random access to all past data is not desired. Ideally, we would only use *streaming algorithms* - algorithms that live on the stream of incoming data, where each data point is processed only once and then discarded.

The contribution of this paper is a hollistic approach to anomaly detection system that clearly defines different parts and stages of the processing, including active learning as a crucial part of the processing loop. The design addresses modern challenges in IT system monitoring and is suitable for cloud deployment.

2. ANOMALY-DETECTION

The most common definition of an anomaly is *a data point that is significantly different from the majority of other data points.* See [2] for a detailed explanation. This definition is strictly analytical. But most often the users define it within the scope of their operation, such as finding abnormal engine performance in order to prevent catastrofic failure, flagging unexpected delays in manufacturing pipeline in order to prevent shipment bottlenecks, detecting unusual user behavior that indicates intrusion and identifying market sectors that exibit unusual trends to detect fraud and tax evasion.

The anomaly-detection process is thus heavily influenced by the target domain. It also needs process-specific way of measuring the detection efficiency. For instance, in classification problems we can use several established measures such as *accuracy*, *recall*, *precision* or $F1$. In anomaly detection, on the other hand, we often don't have classes to work with

and secondly, we need strong user feedback to evaluate our results. Sometimes anomaly detection looks more like a forecasting and optimization problem. We measure how much the current state of a complex system is different from the optimal or predicted value.

2.1 Actionability

It is not sufficient for algorithms to just detect unusual patterns. Human operators that get notified about them must clearly understand the detected problems and be able to act upon them - we call this property of alerts *actionability*. For instance, it is not enough to report "the euclidian distance between multi-dimensional vectors of regularized input values is too big" - end-users will have no clue about what is wrong here. Instead, the system should report something like "The average processing time of customer orders is well above its usual values. This situation will very likely result in a significant drop of daily productivity." Some algorithms produce models that are easier to translate into human language than others. This feature needs to be taken into account when an anomaly-detection system is being implemented.

2.2 Modern challenges

In the era of big data there are many systems that produce data and a lot of the generated data can be used to monitor, maintain and improve the target system. The data volumes are staggering and need to be addressed properly within the system implementation.

Users expect results to be available as soon as possible - within hours, sometimes even minutes or seconds. In cases where automated response in possible, this time-frame shortens to miliseconds (e.g. stock trading, network intrusion).

Current systems for anomaly detection are developed as addons to the existing systems for collecting and processing data. This makes sense, since they developed organically, during the usage by the competent users, which identified areas that require advanced monitoring. We belive this provides necessary business validation of anomaly detection systems. However, there are limitations of such approach.

- Data that is available in one part of the system might **not be available** in another part, where anomaly-detection could greatly benefit from it.
- **Data volume** could prove to be too big for effective anomaly detection analysis, because needed resources might not be available (e.g. computing power is needed for main processing and anomaly detection should not interfere with it).
- Anomaly detection has **local scope** as it only processes data from one part of the system. The alerts are thus not aware of the potential problems in other parts of the system, so resolving issues takes longer and involves more people from several departments to coordinate during problem escalations.
- There is no systematic way of collecting the **user feedback** that would guide and improve the anomaly detection process.

3. THE SYSTEM ARCHITECTURE

To create a system that is able to ingest such huge amount of different data streams, detect anomalies in them and present user with a manageable amount of actionable alerts we propose a reference architecture of such system (figure 1). The acronym **FSADA** stands for *Full-Spectrum Anomaly Detection Architecture*, is based on the Kappa architecture [5] and comprises the components described below.

- **Storage module** contains historical data (raw and derived), background knowledge as well as generated alerts and incidents.
- **Stream-processing module** performs incoming-data pre-processing, as well as signal- and incident-detection.
- **Batch processing module** calculates aggregations, pattern discovery as well as background knowledge refresh.
- **User-interface module** (commonly abbreviated as GUI) displays raw-data, generated alerts along with feedback and active learning support.

3.1 Terminology

From now on we will be using the following terminology:

anomalies - any kind of abnormal behavior inside the system, regardless of the effect on the system performance.

signals - low-level anomalies that have been detected on single data-stream.

incidents - complex anomaly or a group of them with major impact on the system. Its time duration is usually limited to several minutes or hours. They are closely related to the way users perceive the system problems and outages.

alerts - an anomaly that is reported to the user, self-contained with explanation and basic context.

3.2 Storage module

The system needs to store several types of data that perform different functions. Each part of the storage layer can be located in separate system that best matches the requirements.

Measurement data represents raw values that were observed and processed in order to monitor the system. This data is strictly speaking not necessary when our algorithms are designed to work on a stream, but they are required for batch algorithms, for model retraining, active learning and for ad-hoc analytics. **Generated signals and incidents** are stored, additionally processed and viewed by the user. The storage needs to support flexible format of alerts, since each one of them is ideally an independent chunk of data that can be visualized without additional data retrieval. The algorithms can use **domain knowledge** to guide their execution. To facilitate this, the data needs to be stored in a storage system that provides fast searching, in order to be used in stream processing steps for enrichment, routing and aggregation. The algorithms inside the system create and

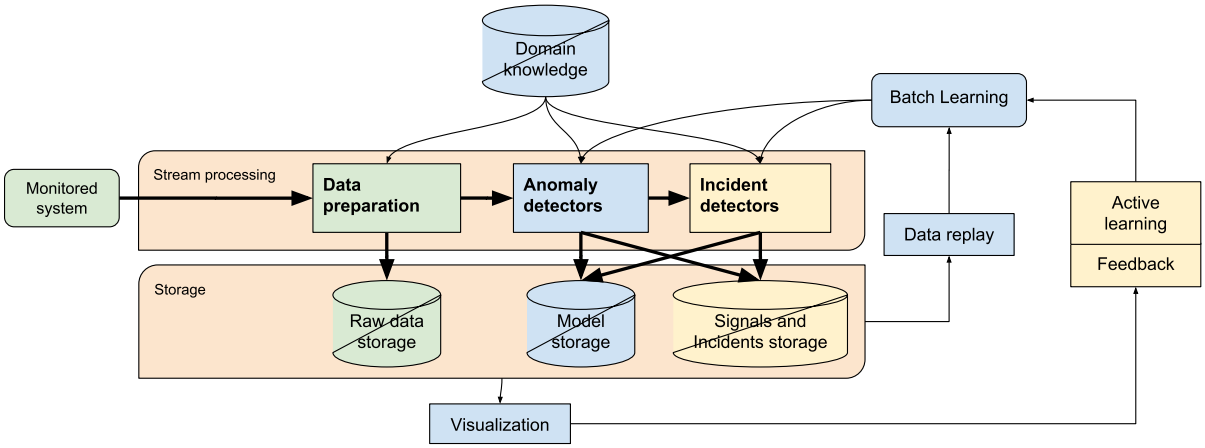


Figure 1: The big picture - displays the main building layers such as stream processing and storage, as well as the flow of the data between different components of the system.

update their **models** all the time, so this part of the storage needs to support reliable and robust storing of possibly large binary files.

3.3 Stream processing module

This module contains the most important part of the system - the components that transport the data, run the processing and generate alerts.

3.3.1 Incoming data pre-processing

Incoming data that the system analyses arrives at different volumes and speeds (*high-velocity*), as well as in many different types and formats. This data needs to be pre-processed before it reaches any anomaly detection algorithms.

Coping with such high-volume data stream requires special technologies. Streaming solutions such as Apache Kafka [4] have been developed and battle tested for processing millions of data records per second in a distributed manner. This step needs to perform several functions.

Data formatting and enrichment - transform messages from the input format into a common format that is accepted by the internal algorithms. Also, additional data fields can be attached, based on background knowledge.

Data aggregation - sometimes we want to measure characteristics of all the data within some time intervals (e.g. average speed in the last 10 minutes).

Data routing - send the transformed and aggregated data to relevant receivers. There may be several listeners for the same type of input data.

3.3.2 Signal detectors

When data is ready for processing, it is routed to signal detectors. They operate on a single data stream, often only on a small partition of it - e.g. single stock, group of related stocks. They handle huge data volumes, so they need to be fast, using very little resources. To achieve great flexibility regarding input data a dynamic allocation of such processors is required. This enables handling of previously unseen data partitions as well as scalability in parallel processing.

These anomalies (signals) have simple models and consequently alert explanations. But they are local in nature - their scope is most often very limited. They also operate on single-data stream, so they don't take into account the anomalies in *"the neighbourhood"*. To overcome this deficiencies, we propose the third step of stream processing, to which signals should be sent.

3.3.3 Incident detectors

This stage of the processing receives signals from different parts of the system, performing scoring of their importance, combining them into *incidents* and thus achieving several advantages.

The scoring algorithm provides option to assign user-guided *subjective importance* to signals - e.g. two statistically equally important anomalies can have completely difference perceived value to the user. This step can also can correlate data across data-streams, a step that is hard to achieve and that proves to be very valuable. Given data from differnt parts of the system it can create more complex constructs that better evaluate the impact of the current problem on the whole system.

This **level of abstraction** is the main access point for end-users - it more closely follows their way of addressing system malfunctions (e.g "if module A breaks, it will have impact on modules B and C, but module D should remain unaffected").

3.3.4 Background knowledge

To help guide the algorithms during the signal detection we can provide additional background knowledge in different forms, such as metadata, manual thresholds and rules, graphs and other structures. All this data can be used to perform various enhancements of basic algorithms, such as creation of **additional features** in data pre-processing, **up- and down-voting** of results (e.g. estimated impact of detected anomaly), **pruning of search space** in optimization steps, **estimation of affected entities** for given anomaly

or **support for complex simulations** when current performance is measured against historical values. These rules and metadata can be acquired by analyzing historical data as well as collecting knowledge from end-user, e.g. manual feedback/sign-off and active learning.

3.4 Improving actionability

The system modules presented so far are mostly established components that are used also in normal processing steps of modern, cloud-based systems (see [1]). We propose that they should be upgraded with the following functionalities in order to achieve the goal of high-quality actionable alerts, empowering users to manage their complex systems in the most efficient way.

3.4.1 Feedback

Historical incidents are very valuable for learning of informative features that aid detection of anomalies. They are also used for calibrating scoring algorithm that assigns relevance scores to generated signals and incidents. It is common for the organization to require every major detected incident to be manually signed off - a *relevance tag* (e.g. high-relevant, semi-relevant, not-relevant, noise, new-normal) has to be assigned to it by the operators. These tags are used for training of incident-classification algorithms, but we can also construct a more complex setting where a form of backtracking is used to *calibrate* signal detectors.

3.4.2 Active learning

The *active learning* approach [3] can be used to make the manual classification effort more efficient. The system provides untagged examples/incidents where the criteria function returns the value that is the closest to the decision boundary. The user then manually classifies the incident and the classification model is re-trained with this new data. By guiding users in this way the system requires relatively small number of steps to cover the search space and obtain good learning examples.

Our proposed approach incorporates this continuous activity in several areas. GUI module should contain appropriate pages where user can enter his feedback and active-learning input. Storage module contains alerts historical data that can be used for re-training of incident detectors. Storage module also contains old and new incident-detector models that can be picked up automatically by the processing module.

4. VALIDATION AND DISCUSSION

Based on our extensive experience with practical anomaly detection implementation, we identified several new requirements for these systems. The presented approach satisfies them by supporting big-data real-time analytics on one side and actionability via active-learning support on the other.

The system architecture is deployable to cloud environment by design. We also employ modern streaming and storage technologies for transporting and storing of different input data and alerts.

We observed that users appreciate our notion of incidents - a grouping of alerts that occur in certain small time interval. Users feel comfortable with seeing the big picture (an indcident) in then drill down into specific data (individual signal). They reported this feature enables them to cut down time for understanding the problem by an order of magnitude (from hours to minutes).

Active-learning component was well received, as it made manual work more efficient. The users noticed how the algorithm was choosing more and more complex learning examples for manual classification. This helped them feel productive and engadged. They also reported positive impact of active learning on their problem understanding, as they were presented with some potentially problematic situations that went unnoticed in the past.

Based on above observations were conclude that our proposed approach has positively impact on the organization, both for technologies as well as human operators. Additional ideas that were collected from users are listed under future work.

5. CONCLUSIONS AND FUTURE WORK

The focus of our future work is on several advanced scenarios where a lot of added value for users is expected, mixing anomaly detection, optimization and simulation. Main gains are expected to come from feedback collection and active learning. Apart from monitoring IT systems, the target domains are also manufacturing and smart cities. We also collected some features that users commonly inquired about:

- **Root-cause analysis** - when a major incident occurs, many parts of the system get affected. To resolve issues as quickly as possible, the operators should be pointed to the origin of the problem. The anomaly detection system should thus have a capability to point to the first signal with high-impact on the final relevance score.
- **Predictions** - The goal for all monitoring systems is to detect problems as soon as possible. The system must that not only be able to detect signals, but also forecast them, based on past behavior. In order to do that, the algorithms require more metadata and structure to properly model inter-dependencies between signals. Mere observation is much easier than simulation of a complex system with many moving parts. But it is possible and is what users expect from a *modern AI-based system*. Our future reserach will be oriented towards providing and efficiently integrating these functionalities into our anomaly-detection approach.

6. REFERENCES

[1] Anodot anomaly detection system. `http://www.anodot.com`, 2018.

[2] C. C. Aggarwal. *Outlier Analysis*. Springer New York, New York, New York, 2013.

[3] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *CoRR*, cs.AI/9603104, 1996.

[4] N. Garg. *Apache Kafka*. Packt Publishing, 2013.

[5] J. Lin. The lambda and the kappa. *IEEE Internet Computing*, 21(5):60–66, 2017.

Predicting customers at risk with machine learning

David Gojo
Jožef Stefan International Postgraduate School,
Jamova 39, 1000 Ljubljana, Slovenia
david.gojo@ijs.si

Darko Dujič
Ceneje d.o.o.,
Jožef Stefan International Postgraduate School,
Štukljeva cesta 40, 1000 Ljubljana, Slovenia
darko.dujic@ceneje.si

ABSTRACT

Today's market landscape is becoming increasingly competitive as more advanced methods are used to understand customer's behavior. One of key techniques are churn mitigation tactics which are aimed at understanding which customers are at risk to leave the service provider and how to prevent this departure. This paper presents analyzes accounts renewal rates and uses easily applicable models to predict which accounts will be decreasing spend at the time when they are due to renew their existing contract based on number of attributes. Key questions it tries to explore is if customer behavioral or customer characteristic data (or combination of both) is better at predicting accounts that will renew at lower than renewal target amount (churn rate).

Categories and Subject Descriptors

F.2.1 **[Numerical Algorithms and Problems]:** Data mining, Structured prediction

General Terms

Algorithms, Management, Measurement, Documentation, Performance

Keywords

Data Mining, Analysis, Churn prediction.

1. INTRODUCTION

The main issue of business is how to make educated decision with support of analysis that dissect complex decisions on addressable problems using measurements and algorithms. Where there are many disciplines are researching methodological and operational aspects of decision making, at the main level, we distinguish between decision sciences and decision systems [1]. With increasing number of companies trying to use machine learning to assist in their decision-making process we examined how decision science can be supplemented by applying machine learning models to the company's customer data. We partnered with the medium sized B2B business operating in Europe and Africa with the aim to help them better understand their 'customers at risk' segment of clients.

To this end we developed two easily applicable performance algorithms designed to highlight customers at risk and company can address to mitigate their risk of leaving as clients.

The paper has the following structure: in section 2 we are presenting related work to the area recorded historically. Next, data acquisition is explained in section 3 followed by results acquired from the tested algorithms in section 4. We then conclude our observations in section 5.

2. RELATED WORK

Improvements in tracking technology have enabled data driven industries to analyze data and create insights previously unavailable to the business. Data mining techniques have evolved to now support the prediction of behavior of customers such risk of leaving due to the attributes that are trackable [2]. The use of data mining methods has been widely advocated as machine learning algorithms, such as random-forest approaches have several advantages over traditional explanatory statistical modeling [3].

Lack of predefined hypothesis makes algorithms excel in these tasks as it is making it less likely to overlook predictor variables or potential interactions that would otherwise be labelled unexpected [4]. Models are often labelled as business intelligence models aimed at finding customers that are about to switch to competitors or leave the business [5].

Key classifications are observed in work related to churn that we will use in our data set for review [6]:

- Behavioral data - will consist of attributes that we have historically observe that play a role in whether the account will renew or not: product utilization, activity levels of the account, number of successful actions in the account and number of upsells done ahead of renewal.
- Characteristic attributes - will consist of size of the account in terms of spend, number of members in the company, number of active users of the products in the company, payment method and how they renew the contract, geography and what level of support the product is given (number of sales visits and interactions with the customer).

3. DATA ACQUISITION

3.1 Data understanding

Working with the customer we have arranged a set of interviews with the leadership to better understand their business and challenges they are experiencing together with ambitions they have in the business. After the interview rounds we focused on reviewing 2 hypotheses flagged in the examination process:

- What is driving churn numbers: behavior of the customers or better structure of the base?

- Does acquisition of new accounts represent a risk in churn number with historic observation of accounts renewing lower / not renewing in their first-year renewal?

3.2 Data pre-processing

The data we used derives from company's internal customer bookings and customer databases we consolidated. As customers are on yearly renewals we have taken the renewal and the data on the account before the renewal as the key building block for analysis.

3.3 Feature engineering

We enriched the data using SQL joins on the customer numbers to include key characteristics of accounts, tenure of the client, products utilization information, behavior of the customer before the renewal and their usage of the product.

In terms of regional split of the market the dataset consists of 4 key geo and segment regions in Europe and Africa:

- Medium-business segment
- UK & Ireland market
- Europe Enterprise segment
- Eastern Europe, Middle-East and Africa

Through feature engineering and reviewing descriptive statistics key attributes we nominalized are 11.

For the machine learning purposes for the calls we have selected 3 possible outcomes related to the outcome of customer spend after it's renewal:

- Customer_Renew (Not renew, Partial renew, Full renew)

3.4 Data Set Statistics

We selected bookings period from 2016 to end of 2017 including 23,043 instances in above selected renewal of 12,872 accounts. The attributes that were nominalized are listed below:

- (nom) Has main product – has product 1
- (nom) Has_assisting_product – has product 2
- (nom) Has_media_product – product 3
- (nom) Account_potential – size and potential of the account
- (nom) Is_Auto_Renew – auto renewal option enabled
- (nom) First_renewal – is the client renewing first time
- (nom) Upsold_Before_renewal – was there an upsell before
- (nom) JS_Utilization – utilization of product 2 - indicator
- (nom) Score_Engagement – engagement of the recruiter
- (nom) LRI_Score – savviness of the user of the product

4. RESULTS

4.1 Brief description of the methods used

Where multiple algorithms were used during the testing due to important feature that the result needed to have at least one interpretable model, so we went in the direction of nominalizing attributes and decided to use J-48 model and Random forest classifier on the data set.

J48. Decision trees C4.5 (J48 in Weka) algorithm: deals with continuous attributes as observed in the related work.

Where the method is classification-only the main machine learning method applied is J48 pruned tree or WEKA-J48 machine learning method. Tree tries to partition the data set into subsets by evaluating the normalized information gain from choosing a descriptor for splitting the data. The training process stops when the resulting nodes contain instances of single classes or if no descriptor can be found that would result to the information gain.

Random Forest. We assume that the user knows about the construction of single classification trees. Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest) [7]. Both methods were applied to imported dataset numerous times with continuous testing of parameters to improve performance.

4.2 Application of J48

Working with Weka on the dataset of the customer we tried to launch the model to tune the parameters. Key modifications:

- "10-fold cross validation" used to improve accuracy
- Minimum number of objects moved to 100

As Figure 2 shows this reduced the number of leaves to 16 which was something comprehendible enough.

Summary of results below:

```
=== Summary ===

Correctly Classified Instances       16789               72.8626 %
Incorrectly Classified Instances      6253               27.1374 %
Kappa statistic                          0.249
Mean absolute error                      0.2759
Root mean squared error                  0.3716
Relative absolute error                 88.6325 %
Root relative squared error             94.189  %
Total Number of Instances            23042
```

Figure 1: J-48 model error estimates

4.3 Application of Random forest

We ran several tests on Random forest vs Random trees. When tuning parameters Random tree tended to not respond well to pruning so Random forest was a preferred option. Like J48, application with key modifications was focused on validation and additionally on setting maximum depth of the random forest:

- "10-fold cross validation"
- Max. depth set at 3

Summary of results below:

```
=== Summary ===

Correctly Classified Instances       16635               72.1943 %
Incorrectly Classified Instances      6407               27.8057 %
Kappa statistic                          0.1852
Mean absolute error                      0.28
Root mean squared error                  0.3685
Relative absolute error                 89.9598 %
Root relative squared error             93.401  %
Total Number of Instances            23042
```

Figure 2: Random forest model error estimates

4.4 Comparisons of models

Overall the J48 model has surprisingly 0.7pp points higher Classification Accuracy than the Random forest model.

Validation Measures	J48	Random Forest
Classification Accuracy	72.9%	72.2%
Mean absolute error	0.276	0.280

Table 1. Baseline benchmark validation measures

Key observation analyzing the data was that neither model was predicting any partially churned accounts after we forced their trees to be pruned.

J48 predictions:

```
a    b      c   <-- classified as
0  2745   285 |   a = PARTIAL_RENEW
0  1528   789 |   b = FULL_RENEW
0  2434  1504 |   c = NOT_RENEWED
```

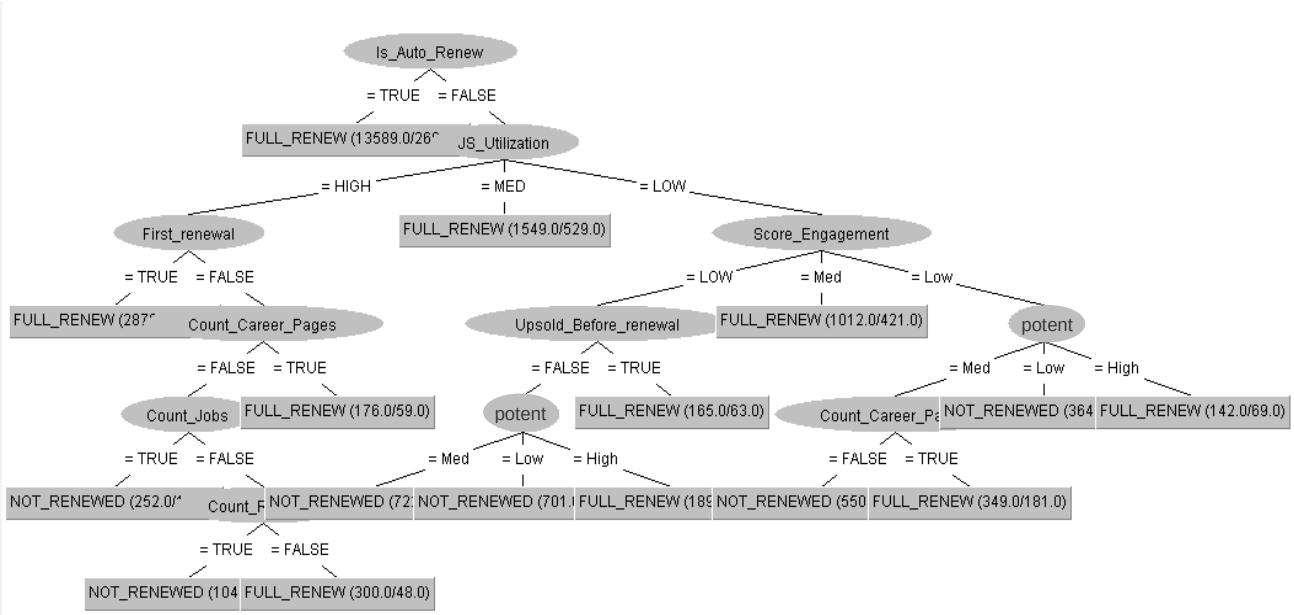


Figure 3: The J48 decision tree

Random forest predictions:

```
a    b      c   <-- classified as
0  2857   173 |   a = PARTIAL_RENEW
0 15591   483 |   b = FULL_RENEW
0  2894  1044 |   c = NOT_RENEWED
```

Even though Random forest has a lower classification accuracy J48 offers significantly higher interpretability with tree pruning offering valuable insights, short description below and discussed in evaluation of models.

J48 provided a significantly better interpretability and classification accuracy than the Random forest or any test on the Random tree model. Some additional tests were done on Naïve Bayes model and J48 was superior in the results. Key area it accelerated was in predicting accounts that will not renew. Where the precision is just above 38% this is almost double comparing to Random forest model.

3 key takeaways observed that the company found the most insightful were:

- One of the new features designed by the product team that encouraged the auto-renew of their clients played the most important at predicting the renewal rate
- Customer behavior is a better signal for probability of renewal vs general account characteristics
- Account potential which is the predictor of account potential and size plays the role only after product utilization and engagement of the account with our products

5. CONCLUSION AND FUTURE WORK

For the acceleration of performance, the decision tree is of paramount importance and value. Insight that Auto renew as a feature is one of the key predictors if the account will renew fully is truly remarkable based on the simplicity of the models and how easily applicable they are.

Applications of this models will be of great foundation for driving the discussion on different account features and metrics. This is especially true as it is tackling one of the key challenges observed in hypothesis as in how important ‘account potential’ is for the account ahead of the renewal.

Observing the attributes added into the analysis scope and optimizing them for the J48 we were able to get valuable insight which account characteristics vs account behaviors ahead of the renewal are the best predictors for the account to renew at the full amount.

6. REFERENCES

[1] M. Bohanec, Decision Making: A Computer-Science and Information-Technology Viewpoint, vol. 7, 2009, pp. 22-37.

[2] A. Rodan, A. Fayyoumi, H. Faris, J. Alsakran and O. Al-Kadi, "Negative correlation learning for customer churn prediction: a comparison study.," *TheScientificWorldJournal*, vol. 2015, p. 473283, 23 3 2015.

[3] A. K. Waljee, P. D. R. Higgins and A. G. Singal, "A Primer on Predictive Models," *Clinical and Translational Gastroenterology*, vol. 5, no. 1, pp. e44-e44, 2 1 2014.

[4] Y. Zhao, B. Li, X. Li, W. Liu and S. Ren, "Customer Churn Prediction Using Improved One-Class Support Vector Machine," Springer, Berlin, Heidelberg, 2005, pp. 300-306.

[5] M. Óskarsdóttir, B. Baesens and J. Vanthienen, "Profit-Based Model Selection for Customer Retention Using Individual Customer Lifetime Values," *Big Data*, vol. 6, no. 1, pp. 53-65, 3 2018.

[6] S. Kim, D. Choi, E. Lee and W. Rhee, "Churn prediction of mobile and online casual games using play log data," *PLOS ONE*, vol. 12, no. 7, p. e0180735, 5 7 2017.

[7] J. Hadden, A. Tiwari, R. Roy and D. Ruta, "Computer assisted customer churn management: State-of-the-art and future trends," *Computers & Operations Research*, vol. 34, no. 10, pp. 2902-2917, 10 2007.

[8] A. K. Meher, J. Wilson and R. Prashanth, "Towards a Large Scale Practical Churn Model for Prepaid Mobile Markets," Springer, Cham, 2017, pp. 93-106.

[9] M. Ballings, D. Van den Poel and E. Verhagen, "Improving Customer Churn Prediction by Data Augmentation Using Pictorial Stimulus-Choice Data," Springer, Berlin, Heidelberg, 2012, pp. 217-226.

Text mining MEDLINE to support public health

João Pita Costa, Luka Stopar, Flavio Fuart, Marko Grobelnik
Jožef Stefan Institute, Ljubljana
Quintelligence, Ljubljana, Slovenia

Raghu Santanam, Chenlu Sun
Arizona State University, USA

Paul Carlin
South Eastern Health and Social Care Trust, UK

Michaela Black, Jonathan Wallace
Ulster University, UK

ABSTRACT

Today's society is data rich and information driven, with access to numerous data sources available that have the potential to provide new insights into areas such as disease prevention, personalised medicine and data driven policy decisions. This paper describes and demonstrates the use of text mining tools developed to support public health institutions to complement their data with other accessible open data sources, optimize analysis and gain insight when examining policy. In particular we focus on the exploration of MEDLINE, the biggest structured open dataset of biomedical knowledge. In MEDLINE we utilize its terminology for indexing and cataloguing biomedical information – MeSH – to maximize the efficacy of the dataset.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Measurement, Performance, Health.

Keywords

Big Data, Public Health, Healthcare, Text Mining, Machine Learning, MEDLINE, MeSH Headings.

1. MEANINGFUL BIG DATA TOOLS TO SUPPORT PUBLIC HEALTH

The Meaningful Integration of Data, Analytics and Service [MIDAS], Horizon 2020 (H2020) project [1] is developing a big data platform that facilitates the utilisation of healthcare data beyond existing isolated systems, making that data amenable to enrichment with open and social data. This solution aims to enable evidence-based health policy decision-making, leading to significant improvements in healthcare and quality of life for all citizens. Policy makers will have the capability to perform data-driven evaluations of the efficiency and effectiveness of proposed policies in terms of expenditure, delivery, wellbeing, and health and socio-economic inequalities, thus improving current policy risk stratification, formulation, implementation and evaluation. MIDAS enables the integration of heterogeneous data sources, provides privacy-preserving analytics, forecasting tools and visualisation modules of actionable information (see the dashboard of the first prototype in Figure 1). The integration of open data is fundamental to the participatory nature of the project and core ideology, that heterogeneity brings insight and value to analysis. This will democratize, to some extent, the contribution to the results of MIDAS. Moreover, it enables the MIDAS user to profit from the often powerful information that exists in these open datasets. A point in case is MEDLINE, the scientific biomedical knowledge base, made publicly available through PubMed. The set of tools described in this demonstration paper focuses on this large open dataset, and the exploration of its structured data.

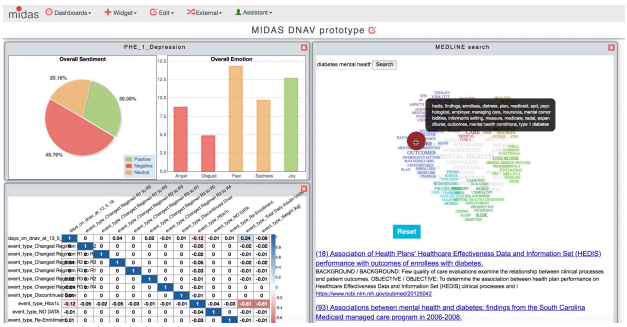


Figure 1. MIDAS platform dashboard, composed of visualisation modules customized to the public health data sourced in each governamental institution, and combined with open data.

2. THE MEDLINE BIOMEDICAL OPEN DATA SET AND IT'S CHALLENGES.

2.1. MEDLINE DATASET.

With the accelerating use of big data, and the analytics and visualization of this information being used to positively affect the daily life of people worldwide, health professionals require more and more efficient and effective technologies to bring added value to the information outputs when planning and delivering care. The day-to-day growth of online knowledge requires that the high quality information sources are complete, high quality and accessible. A particular example of this is the PubMed system, which allows access to the state-of-the-art in medical research. This tool is frequently used to gain an overview of a certain topic using several filters, tags and advanced search options. PubMed has been freely available since 1997, providing access to references and abstracts on life sciences and biomedical topics. MEDLINE is the underlying open database [7], maintained by the United States National Library of Medicine (NLM) at the National Institutes of Health (NIH). It includes citations from more than 5200 journals worldwide journals in approximately 40 languages (about 60 languages in older journals). It stores structured information on more than 27 million records dating from 1946 to the present. About 500,000 new records are added each year. 17.2 million of these records are listed with their abstracts, and 16.9 million articles have links to full-text, of which 5.9 million articles have full-text available for free online use. In particular, it includes 443.218 full-text articles with the key-words string "public health".

2.2. MEDLINE STRUCTURE.

The MEDLINE dataset includes a comprehensive controlled vocabulary – the *Medical Subject Headings* (MeSH) – that

delivers a functional system of indexing journal articles and books in the life sciences. It has proven very useful in the search of specific topics in medical research, which is particularly useful for researchers conducting initial literature reviews before engaging in particular research tasks. Humans annotate most of the articles in MEDLINE with MeSH Heading descriptors. These descriptors permit the user to explore a certain biomedical related topic, which relies on curated information made available by the NIH. MeSH is composed of 16 major categories (covering anatomical terms, diseases, drugs, etc) that further subdivide from the most general to the most specific in up to 13 hierarchical depth levels.

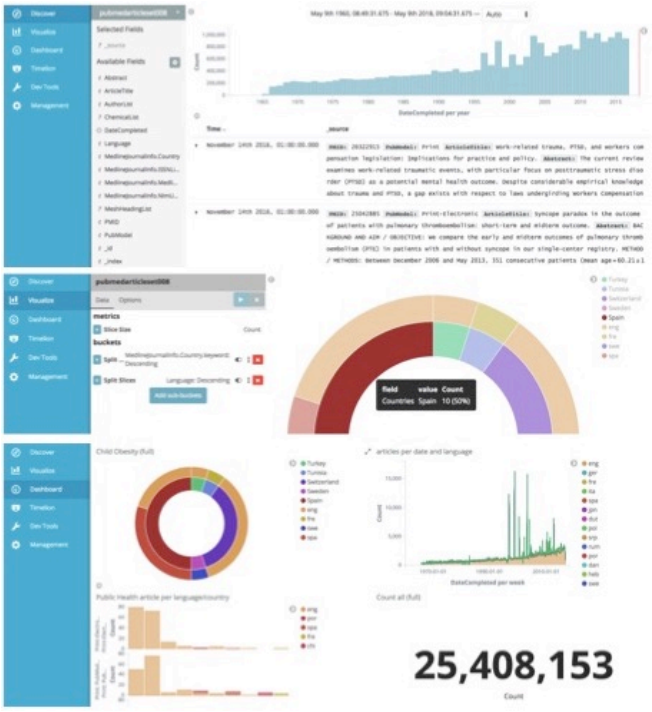


Figure 2. *MEDLINE data visualisation tool enabling exploration of that open dataset in its full potential, based on data representations easy to understand and to communicate. It provides an interactive public instance that can be managed at the dashboard management tool (below) for which the visualisation modules are constructed (in the center) based on the queries made to the MEDLINE dataset (above).*

2.3. MEDLINE INDEX.

This paper demonstrates the interactive data visualisation text-mining tools that enable the user to extract meaningful information from MEDLINE. To do that we are using the underlying ontology-like structure MeSH. MEDLINE data, together with the MeSH annotation, that is indexed with ElasticSearch and made available to data analytics and visualisation tools. This will be discussed in more detail in the next section.

The manipulation and visualization of such a complete data source brings challenges, particularly in the efficient search, review and presentation when choosing appropriate scientific knowledge. The manipulation and visualisation of complex text data is an important step in extracting meaningful information from a dataset such as MEDLINE. Although powerful, the online search engine provided by the NLM does not provide suitable tools for in-depth analysis and the emergence of scientific information. As one of the main goals of MIDAS is to experiment with advanced visualisation techniques in support of public health policy making, a suitable MIDAS PubMed repository had to be developed. This repository has to allow exploration of a wide range of different visualisation techniques in order to evaluate their applicability to policy-making tasks within the policy cycle. Therefore, there was a need for a selection of a powerful, semi-structured text index, that would allow free text searches, but also allow the creation of complex queries based on available metadata. An obvious choice is elasticSearch, which combines features provided by NoSQL databases with standard full text indexes, as it is based on the Apache Lucene Index. The main design challenge when choosing this particular toolset was that querying based on arrays or parent-child relations are not supported, meaning that for complex use-cases different indexes based on the same source dataset have to be created. Nevertheless, excellent results, particularly with regards to the area of performance have been obtained.

2.4. MEDLINE DASHBOARD.

One of the identified needs motivating this work is assuring the availability of a dynamic dashboard that permits the user to explore data visualisation modules, representing the queries to the MEDLINE dataset through pie charts, bar charts, etc [5]. The dashboard that we made available (in Figure 2) feeds on that dataset through the elasticSearch index earlier discussed. It is composed of several interactive visualisation modules that utilises the mouse hover when interacting and provide information through pop-up messages on several aspects of the data based on particular queries of interest (e.g. a pie chart representing the “public health” citations that talk about “childhood obesity” during a selected period of time; or a bar chart showing different concepts included in the articles related to “mental health” in Finnish scientific journals). The MEDLINE dataset is mostly in the English language but includes a significant volume of translated abstracts of scientific articles that were written in several other languages. The open source data visualisation Kibana is a plugin to elasticSearch that supports the described dashboard. Thus, it is the data visualisation dashboard of choice for elasticSearch-based indexes, such as the one we present here. It is used in the context of MIDAS for fast prototyping and support of part of MIDAS use-cases. While the dashboard itself serves the less technical user to explore the data available (over a subset of the data generated by a topic of interest), other options are available that permit more control of the data by the data scientists at a more operational level. These are: (i) the management dashboard, where the technical user can perform the appropriate subsampling based on the topics of interest and the required advanced options over the available data features; and (ii) the visual modules creator permitting the technical user to easily create new interactive visualisation modules. Moreover, this tool enables one to query large datasets and produce different types of visualisation modules that can be later integrated into customized dashboards. The flexibility of such dashboards permits the user to profit from data visualisations that feed on his/her preferences, previously set up as filters to the dataset. The MIDAS data visualisation tools permit the user to explore the potential of the MEDLINE dataset, based on pie charts and other representations that are easy to comprehend, interact with, and to communicate. It also enables a public instance based on a particular query to the dataset, which includes different types of data visualisation modules that can later integrate a customised dashboard, designed in agreement with the workflows and preferences of the end-user. This live dashboard can easily be

integrated through an iframe in any website, not showing the customization settings but maintaining the interaction capability and the real-time update. *It permits a complete base solution to further explorer the MEDLNE index and the associated dataset [6].*

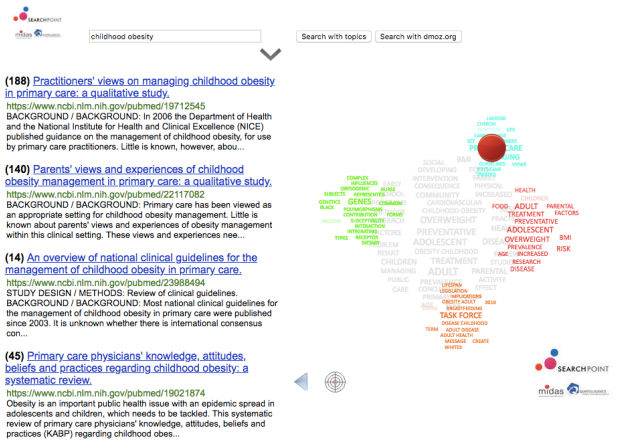


Figure 4. A screenshot of the MEDLINE SearchPoint, with groups of keywords (on the right) extracted from the search results, represented by different colors, and on the left the reindexed search results themselves with the number that they appear in the original index *[6]*.

4. MEDLINE SEARCHPOINT.

The efficient visualisation of complex data is today an important step in obtaining the research questions that describe the problem that is extracted from that data. The MEDLINE SearchPoint is an interactive exploratory tool refocused from the proprietary open source technology *SearchPoint* [8] (available at searchpoint.ijs.si) to support health professionals in the search for appropriate biomedical knowledge. It exhibits the clustered keywords of a query, after searching for a topic. When we use indexing services (such as standard search engines) to search for information across a huge amount of text documents – the MEDLINE index described in Section 2 being an example – we usually receive the answer as a list sorted by a relevance criteria defined by the search engine. The answer we get is biased by definition. Even by refining the query further, a time consuming process, we can never be confident about the quality of the result. This interactive visual tool helps us in identifying the information we are looking for by reordering the positioning of the search results according to subtopics extracted from the results of the original search by the user. For example, when we enter a search term 'childhood obesity, the system performs an elasticSearch search over the MEDLINE dataset, extracts groups of keywords that best describe different subgroups of results (these are most relevant, and not most frequent terms). This tool gives us an overview of the content of the retrieved documents (e.g. we see groups of results about prevention, pregnancy, treatments, etc.) represented by: (i) a numbered list of 10 MEDLINE articles with a short description extracted from the first part of the abstract; (ii) a word-cloud representing the k-means clusters of topics in the articles that include the searched keywords; (iii) a pointer that can be moved through the word-cloud and that will change the priority of the listed articles. The word-cloud in (ii) is done by taking a set of MEDLINE documents S and transforming them into vectors using TF-IDF, where each dimension represents the "frequency" of one particular word. For example, lets say that we have document D_1: "psoriasis is bad" and document D_2: "psoriasis is good". This could be transformed as $D_1 = (1, 1, 1, 0)$ and $D_2 = (1, 1, 0, 1)$. Then the documents are clustered into k groups $G_1, G_2, ..., G_k$ using the k-means algorithm. For each group we compute the "average" document (centroid), which is the representative of the group. The most frequent words in the "average" document are drawn in the word cloud - the central grey word cloud is the "average" of all the documents in S. We can calculate how similar a particular document d is to a group G_i by calculating the cosine of the angle between the vector representation of d and the "average" document (centroid) of the group G_i. By dragging the red SearchPoint ball over the word-groups, we provide the relevance criteria to the search result, thus bringing to the top results the articles we are most interested in (see Figure 4). When that ball is moved, for each document, we calculate the similarity to each of the word-groups and combine it with the distance between the ball and the group. The result is used as the ranking weight where the document with the highest cumulative weight is ranked first. When having the mouse over the word-clouds we get a combination of the most frequent words shown in the tag clouds that change based on how close the ball is to a particular group. After getting to a position with the SearchPoint over the word cloud highlighting "primary care", a qualitative study in primary care on childhood obesity that occupied the position 188 is now in the first position. The user can read its title and first lines of abstract, and when clicking on it, the system opens the article in the browser at its PubMed URL location.

3. MeSH CLASSIFIER

This rich data structure in the MEDLINE open set is annotated by human hand (although assisted by semi-automated NIH tools) and therefore is not available in the most recent citations. However, in the context of MIDAS we made available an automated classifier based on [2] that is able to suggest the categories of any health related free text. It learns over the part of the MEDLINE dataset that is already annotated with MeSH, and is be able to suggest categories to the submitted text snippets. These snippets can be abstracts that do not yet include MESH classification, medical summary records or even health related news articles. To do that we use a nearest centroid classifier [3] constructed from the abstracts from the MEDLINE dataset and their associated MeSH headings. Each document is embedded in a vector space as a feature vector of TF-IDF weights. For each category, a centroid is computed by averaging the embeddings of all the documents in that category. For higher levels of the MeSH structure, we also include all the documents from descendant nodes when computing the centroid. To classify a document, the classifier first computes its embedding and then assigns the document to one or more categories whose centroids are most similar to the document's embedding. We measure the similarity as the cosine of the angle between the embeddings. Preliminary analysis shows promising results. For instance when classifying the first paragraph of the Wikipedia page for "childhood obesity", excluding the keyword "childhood obesity" from the text, the classifier returns the following MeSH headings:

"Diseases/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Body Weight/Overweight",
"Diseases/Pathological Conditions, Signs and Symptoms/Signs and Symptoms/Body Weight/Overweight/Obesity".

The demonstrator version of the MeSH classifier is already available through a web app, as well as through a REST API

using a POST call, and is at the moment under qualitative evaluation. This is being done together with health professionals with years of practical experience in using MeSH themselves through PubMed. In addition, we aim to further explore the potential of the developed classifier in several public health related contexts including non classified scientific articles of three types: (i) review articles; (ii) clinical studies; and (iii) standard medical articles. The potential impact of this technology will also include electronic health records and the monitoring health related news sources. We believe that his approach will address an identified recurrent need of health departments to enhance the biomedical knowledge, and motivate a step change in health monitoring.

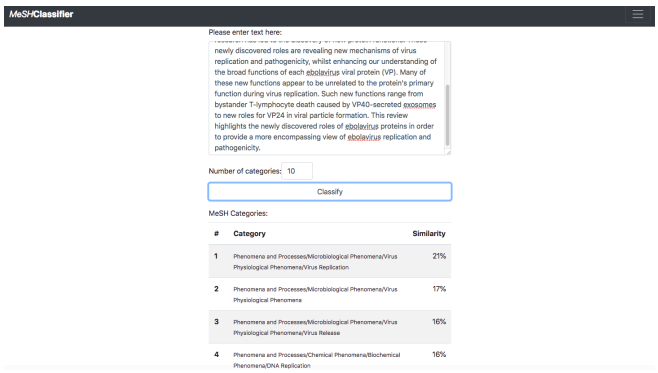


Figure 3. *A screenshot of the web app to the MEDLINE classifier, when requesting the automated MeSH annotation of a scientific review article abstract extracted from PubMed (in the body of text above) and the results as MeSH headings descriptors including their tree path in the MeSH ontology-like structure (center), their similarity measure (right) and their positioning in the classification (left).*

5. CONCLUSION AND FUTURE WORK

To further extend the usability of the MEDLINE SearchPoint, we are developing a data visualisation tool that permits the user to plot the top results mostly related with a topic of interest, as explored with SearchPoint. Based on the choice of a time window and a certain topic, such as "mental health", the user is able to view the clustered MEDLINE documents, rolled over the plot or click to view the plotted points, each of which represents an article in PubMed. This will be done through multidimensional scaling, plotting the articles in the subsample using cosine text similarity. The difficulties to plot large datasets using these methods, and the lack of potential in the outcomes of that heavy computation, provided a focus for the team to only plot the first hundred results of the explorations done within MEDLINE SearchPoint. With this extended tool the healthcare professional will be able to: (i) explore a certain area of research with the aim of a more accessible scientific review, in identifying the evidence base for a medical study or a diagnostic decision; (ii) identify areas of dense scientific research corresponding to searchable topics (e.g. the evaluation of the coverage of certain rare diseases that need more biomedical research, or the identification of alternative research paths to overpopulated but inefficient research); and (iii) exploration of the research topic over time windows that enable filtering to avoid known unreliable results.

In line with this work we have been developing research to contribute to the smart automation of the production of biomedical review articles. This collaborative research lead by Raghu T. Santanam at Arizona State University, aims to provide a wide knowledge over a restricted topic over the wider knowledge available at MEDLINE. We utilize the deep learning algorithm Doc2vec [4] to create similarity measures between articles in our corpus. In that we built a balanced test dataset and three different representations of the corpus, and compared the performance between them. The implementation currently builds a matrix of similarity scores for each article in the corpus. In the next steps, we will compare similarity of documents from our implementation against the baseline for a randomly chosen set of articles in the corpus.

The further development of the MeSH classifier will consider the feedback of the usability of health professionals working in partner institutions, profiting of their years of experience with MEDLINE and MeSH itself, to tune the system to ensure the best usability in the domain. Furthermore, we will use the outcomes of the final version of this classifier to label health related news with the MeSH Headings descriptors, potentiating a new approach on the processing and monitoring of population health, population awareness of certain diseases, and the general public acceptance of public health decisions through news.

ACKNOWLEDGMENTS

We thank the support of the European Commission on the H2020 project MIDAS (G.A. nr. 727721).

REFERENCES

[1] B. Cleland et al (2018). Insights into Antidepressant Prescribing Using Open Health Data, Big Data Research, doi.org/10.1016/j.bdr.2018.02.002
[2] L. Henderson, Lachlan (2009). Automated text classification in the dmoz hierarchy. TR.
[3] C. Manninget al (2008), "Introduction to Information Retrieval," Cambridge Univ. Press, 2008, pp. 269-273.
[4] T. Mikolov et al (2013). Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781.
[5] J. Pita Costa et al (2017). Text mining open datasets to support public health. In *Conf. Proceedings of* WITS 2017.
[6] J. Pita Costa et al (2018). MIDAS MEDLINE Toolset Demo. http://midas.quintelligence.com (accessed in 28-8-2018).
[7] F. B. Rogers, (1963). Medical subject headings. *Bull Med Libr Assoc*. **51**: 114–6.
[8] L. Stopar, B. Fortuna and M. Grobelnik (2012). Newssearch: Search and dynamic re-ranking over news corpora. In Conf. Proceedings of SiKDD2012.

Crop classification using PerceptiveSentinel

Filip Koprivec
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
filip.koprivec@ijs.si

Matej Čerin
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
matej.cerin@ijs.si

Klemen Kenda
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39, 1000 Ljubljana, Slovenia
klemen.kenda@ijs.si

ABSTRACT
Efficient and accurate classification of land cover and land usage can be utilized in many different ways: ranging from natural resource management, agriculture support to legal and economic processes support. In this article, we present an implementation of land cover classification using the PerceptiveSentinel platform. Apart from using base 13 bands, only minor feature engineering was performed and different classification methods were explored. We report an F_1 and accuracy score (80-90%) in range of state of the art when using pixel-wise classification and even comparable to time series based land cover classification.

Keywords
remote sensing, earth observation, machine learning, classification

1. INTRODUCTION
Specific aspects of earth observation (EO) data (huge amount of data, widespread usage, many different problem settings etc.), coupled with the recent launch of ESA Sentinel mission that provides a huge volume of data relatively frequently (every 5-10 days), present an environment that is suitable for current machine learning approaches.

Efficient and accurate land cover classification can provide an important tool for coping with current climate change trends. Classification of crops, their location and potentially their yield prediction provide various interested parties with information on crop resistance, adapting to changes in temperature and water level changes. Along with direct help, accurate crop classification tools can be used in a variety of other programs, from government based subsidies to various insurance schemes.

Along with previously highly promising features of EO data, data acquisition and processing pose some specific challenges. Satellite acquired data is highly prone to missing data due to various reasons; mostly cloud coverage, (cloud) shadows, atmospheric refraction due to changes in atmospheric conditions. Additionally, correct training data, either for classification or regression, is hard to come by, must be relatively recent, and regularly updated due to changes in land use. Furthermore, correct labels and crop values are almost impossible to verify and usually self-reported, which often means that quality of training data is not perfect. Valero et al. [13] raise the problem of incorrect (or partially correct) data labels, which will become apparent when interpreting results.

Another class of problems is posed by the spatial resolution of images. Since satellite images provided by the ESA Sentinel-2 mission have a resolution of $10\,\mathrm{m} \times 10\,\mathrm{m}$ on most granular bands and $60\,\mathrm{m} \times 60\,\mathrm{m}$ on bands used for atmospheric correction, land cover irregularities falling in this order of magnitude might not be detected and correctly learned on. This problem is especially prevalent in smaller and more diverse regions, where individual fields are smaller and land usage is more fragmented.

The current state of the art land classification focuses heavily on the temporal dimension of acquired data [1], [13], [14]. The time-based analysis offers clear advantages since it considers growth cycles of sample crops, enables continuous classification etc., and generally produces better results, with reported F_1 scores for crop/no-crop classification in a range from 0.80-0.93 [14]. One major drawback of time-based classification is the problem of missing data. In our test drive scenario, 70% of images are heavily obscured by clouds [9], a problem which removes a lot of the advantages of time-based classification and demands major compensations with missing data imputation.

In this paper, we present a possible approach on a land cover classification of single time image acquired using the PerceptiveSentinel [1] platform, using multiple classification methods for tulip field classification in Den Helder, Netherlands.

2. PERCEPTIVESENTINEL PLATFORM
2.1 Data
Data used in this article is provided by ESA Sentinel-2 mission. The Sentinel-2 mission comprises a constellation of two polar-orbiting satellites placed in the same orbit, phased at 180° to each other [2]. Sentinel-2A was launched on 23$^{\mathrm{rd}}$ June 2015, while the second satellite was launched on 7$^{\mathrm{th}}$ March 2017. Revisit time for equator is 10 days for each satellite, so since the launch of the second satellite, each data point is sampled at least every 5 days (a bit more frequently when away from the equator).

Each satellite collects data in 13 different wavelength bands presented in figure 1, with varying granularity. Data obtained for each pixel is firstly preprocessed by ESA where

[1] http://www.perceptivesentinel.eu/

atmospheric reflectance and earth surface shadows are corrected [4].

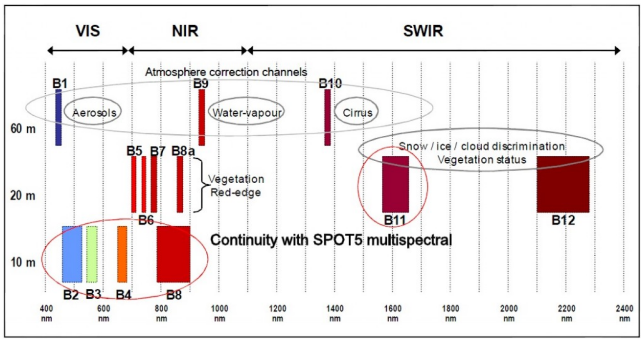


Figure 1: Sentinel 2 spectral bands [12]

2.2 Data Acquisition

Satellites provide around 1TB of raw data per day, which is provided for free on Amazon AWS. Images are then processed and indexed by Sinergise and subsequently provided for free along with their SentinelHub [11] library. As part of the PerceptiveSentinel project, a sample platform was developed on top of SH library, which eases data acquisition, cloud detection and further data analysis on acquired data.

The whole dataset currently consists of images captured from the end of June 2015 till August 2018. Images are available for use in a few hours after being taken. Since working with data for the whole world is impractical, smaller geographical regions are usually queried and analyzed on their own. One important aspect when analyzing larger regions that must be taken care of is the fact that EO data is acquired in swaths with the width of approximately 290 km [3]. Since the swaths overlap a bit, care must be taken when sampling larger areas (in a size of small state), as the area might be chopped into a few irregular tiles covering only part of an area of interest.

Corrected images are available using the SentinelHub library. PerceptiveSentinel platform provides an easy to use framework that combines satellite data acquisition, subsequent cloud detection enables an easy way to pipeline machine learning framework. They also provide an easy way to integrate (vectorized or rasterized) geopedia layers as a source of ground truth for classification.

2.3 Data Preprocessing

Most of the preprocessing is already done by ESA (atmospheric reflectance, projection . . .). The data is mostly clean and presented as a pixel array with dimensions H×W×B, where W and H are image dimensions (in our case 589 and 590) and B is number of bands selected (in our case 13, but we may individually preconfigure the Sentinelhub library to return variable number of bands and even custom calculations based on other bands).

When preprocessing we only need to consider one part of problematic data, namely clouded parts of images. ESA provides some sort of cloud detection, but our experiments proved it unsatisfactory, so we used the s2cloudless library developed by Sinergise for this task [10].

3. METHODOLOGY

3.1 Sample Data

For purpose of this article, a sample patch of fields in Den Helder, Netherlands, with coordinates: (4.7104, 52.8991), (4.7983, 52.9521) was analyzed. Three different datasets were considered: tulip fields in year 2016 and 2017 and arable land in 2017. For each of these datasets, the first observed date with no detected clouds was selected and binary classification (tulips vs no-tulips and arable vs non-arable land) was performed on the image from that date. The date selection was based on the fact that tulips' blooms are most apparent during late April and beginning of May [9].

3.2 Feature Vectors

Three additional earth observation indices that were used as features are presented in Table 1 as suggested by [8].

Name	Formula
NDVI	$\frac{B08 - B04}{B08 + B04}$
EVI	$\frac{2.5(B08 - B04)}{(B08 + 6B04 - 7.5B02 + 1)}$
SAVI	$(1 + 0.5)\frac{B08 - B04}{B08 + B04 + 0.5}$

Table 1: Additional indices

For each selected image, all 13 Sentinel2-bands were considered as feature vectors for each pixel, in the second experiment, additional land cover based classification indices from Table 1 were added.

3.3 Experiment

The experiment was conducted in the Den Helder region to asses the effectiveness of classification and improvement with additional features. The same region is also used as a test drive location for the PerceptiveSentintel project. One important characteristic to keep in mind is the fact that classification classes are not uniformly distributed. Tulip fields constitute 17.1% and 17.7% of all pixels in the year 2016 and 2017 respectively, while arable land accounts for 64% of pixels in 2017 data set. Care must, therefore, be taken when assessing the predictive power of a model.

For each dataset, multiple classification algorithms were tested on base band feature vectors and feature vectors enriched with calculated indices from Table 1. Experiments were carried out using python library `scikit-learn` and default parameters were used for each type of classifier. For each data set and each classifier (Ada Boost, Logistic regression, Random Forest, Multilayer perceptron, Gradient Boosting, Nearest neighbors and Naive Bayes), 3-fold cross-validation was performed. Folds were generated on a non-shuffled dataset with balanced classes ratios.

4. RESULTS

Results of selected classifiers are presented in Tables 2–4 (ind column indicates additional indices as features) are comparable with results from related works [5], [6] which report

accuracy results from 60-80%, although our experimental dataset was quite small and homogeneous, which might offer some advantage over larger plots of land.

Alg.	Ind	Prec	Rec	Acc	F_1	T
Logistic Regression	No	**0.895**	0.551	0.912	0.682	2.8
	Yes	0.877	0.564	0.912	0.686	3.6
Decision Tree	No	0.640	0.697	0.881	0.667	7.9
	Yes	0.629	0.698	0.878	0.662	11.3
Random Forest	No	0.870	0.675	**0.927**	0.760	15.0
	Yes	0.867	0.680	**0.927**	0.762	21.7
ML Perceptron	No	0.875	0.720	**0.935**	**0.790**	184.2
	Yes	0.835	0.740	**0.931**	**0.784**	241.3
Gradient Boosting	No	0.878	0.657	0.926	0.751	84.8
	Yes	0.856	0.657	0.923	0.743	120.6
Naive Bayes	No	0.343	**0.800**	0.704	0.480	0.4
	Yes	0.316	**0.808**	0.669	0.454	0.6

Table 2: Tulip fields in 2016 results

Alg.	Ind	Prec	Rec	Acc	F_1	T
Logistic Regression	No	0.514	0.561	0.829	0.537	2.8
	Yes	0.545	0.615	0.841	0.578	4.0
Decision Tree	No	0.574	0.633	0.852	0.602	7.3
	Yes	0.565	0.634	0.849	0.598	11.2
Random Forest	No	0.786	0.599	0.900	0.680	13.8
	Yes	**0.788**	0.604	0.901	0.683	20.5
ML Perceptron	No	**0.790**	0.673	**0.911**	0.727	375.9
	Yes	0.780	0.693	**0.911**	**0.734**	419.8
Gradient Boosting	No	0.786	0.613	0.902	0.689	84.4
	Yes	0.785	0.614	0.902	0.689	120.3
Naive Bayes	No	0.330	**0.861**	0.666	0.477	0.4
	Yes	0.318	**0.858**	0.649	0.464	0.6

Table 3: Tulip fields in 2017 results

For each test, precision, recall, accuracy, and F_1 score were reported along with the timing of the whole process. As seen from the tables, multilayer perceptron achieved best results when comparing F_1 score across all data sets, but its training was considerably slower than all other classification methods (in fact, it's training time was comparable to all other classification times combined). Multilayer perceptron was followed closely by random forest, which achieved just marginally worse results, but was way less expensive to train and evaluate, while still retaining score that was higher or comparable with related works.

Adding additional features to feature vector did not significantly improve classification score and has in some cases even hampered performance while having a significant impact on the training time.

Using classifier trained on 2016 tulips data and predicting data in 2017 yielded an F_1 score of 0.57, while classifier trained on 2017 data yielded an F_1 score of 0.67 on 2016 data, indicating the robustness of the classifier.

Graphical representation of classification errors can be seen in Figure 2 and 3 which show true positive (TP) pixels in purple color, false positive (FP) in blue color and false negative (FN) in red. Looking at the images it can easily be

Alg.	Ind	Prec	Rec	Acc	F_1	T
Logistic Regression	No	0.853	0.913	0.843	0.882	2.7
	Yes	0.854	0.908	0.841	0.880	3.2
Decision Tree	No	0.878	0.868	0.837	0.873	9.6
	Yes	0.885	0.868	0.842	0.876	14.5
Random Forest	No	0.928	0.889	0.884	0.908	17.3
	Yes	**0.934**	0.891	0.889	0.912	26.3
ML Perceptron	No	**0.929**	**0.932**	**0.911**	**0.931**	522.4
	Yes	0.926	**0.940**	**0.913**	**0.933**	586.2
Gradient Boosting	No	0.899	0.921	0.883	0.910	82.6
	Yes	0.905	0.926	0.890	0.915	118.4
Naive Bayes	No	0.823	0.830	0.776	0.827	0.4
	Yes	0.814	0.806	0.757	0.810	0.6

Table 4: Arable land in 2017 results

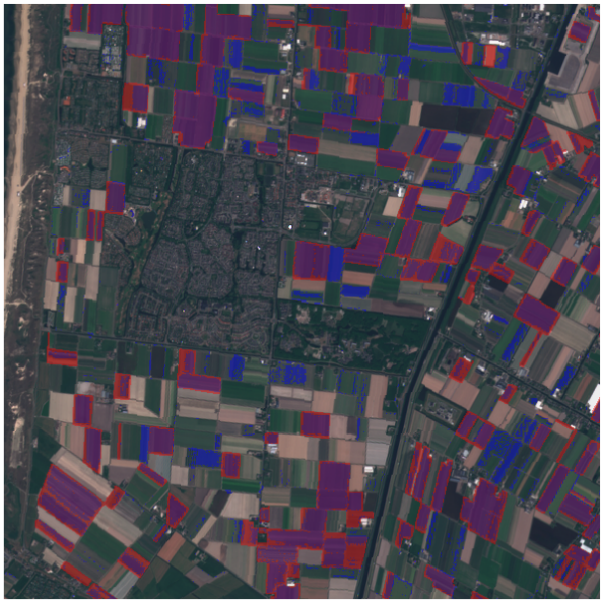


Figure 2: Graphical representation of errors in ML perceptron classification of tulip fields in 2017 (TP in purple, FP in blue, FN in red)

seen, that classification produced quite satisfactory results. An important thing to notice when inspecting Figure 2 is that the true positive pixels were classified in densely packed groups with clear and sharp edges, which correspond nicely to field boundaries seen with the naked eye (both RF and Gradient boosting decision trees produced visually very similar results). This might suggest that algorithms have detected another culture similar to tulips and classified it as tulips (or conversely, that the ground truth might not be that accurate). A lot of FN pixels can also be spotted on field boundaries, which may correspond to different quality or mixing of different plant cultures near field boundaries.

Similarly, observing results of arable land classification, one immediately notices small (false positive) blue patches and some red patches. Most notably, a long blue line is spotted in the left part of the image (near the sea), which may indicate some sort of wild culture near the sea that was not

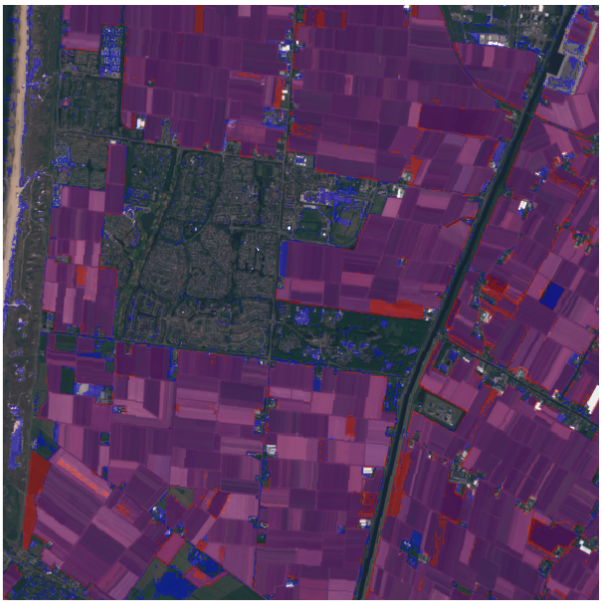


Figure 3: Graphical representation of errors in ML perceptron classification of arable land in 2017 (TP in purple, FP in blue, FN in red)

included in the original mask. Further manual observation of misclassified red patch in the middle of arable land suggests that this field is barren (easily seen in Figure 2) and possibly wrongly classified as arable land in training data.

5. CONCLUSIONS

In our work, we have examined the use of different classification methods and additional features for land cover classification problem on a single image. Our results are comparable with results from the related literature. We propose that classification strength and adaptability be further improved by considering time series and stream aggregates for each pixel as researched in [14] [7]. Additionally, pixels might be grouped together into logical objects to enable object (field) level classification as proposed by [13].

Furthermore, results have shown, that correct ground truth mask is essential for good classification performance. As seen from our results, even seemingly correct labels might miss some cultures or classify empty straits of land as crops.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT program of the EC under project PerceptiveSentinel (H2020-EO-776115). The authors would like to thank Sinergise for it's contribution to sentinelhub and cloudless library along with all help with data analysis.

7. REFERENCES

[1] Belgiu, M., and Csillik, O. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment 204* (2018), 509 – 523.

[2] ESA. Satellite constellation / Sentinel-2 / Copernicus / Observing the Earth / Our Activities / ESA. `https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Sentinel-2/Satellite_constellation`. Accessed 13 August 2018.

[3] ESA. Sentinel-2 - Missions - Resolution and Swath - Sentinel Handbook. `https://sentinel.esa.int/web/sentinel/missions/sentinel-2/instrument-payload/resolution-and-swath`. Accessed 13 August 2018.

[4] ESA. User Guides - Sentinel-2 MSI - Level-2 Processing - Sentinel Online. `https://sentinel.esa.int/web/sentinel/user-guides/sentinel-2-msi/processing-levels/level-2`. Accessed 13 August 2018.

[5] Guida-Johnson, B., and Zuleta, G. A. Land-use land-cover change and ecosystem loss in the espinal ecoregion, argentina. *Agriculture, Ecosystems & Environment 181* (2013), 31 – 40.

[6] Gutiérrez-Vélez, V. H., and DeFries, R. Annual multi-resolution detection of land cover conversion to oil palm in the peruvian amazon. *Remote Sensing of Environment 129* (2013), 154 – 167.

[7] Gómez, C., White, J. C., and Wulder, M. A. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing 116* (2016), 55 – 72.

[8] Jiang, Z., Huete, A. R., Didan, K., and Miura, T. Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment 112*, 10 (2008), 3833 – 3845.

[9] Kenda, K., Kažič, B., Čerin, M., Koprivec, F., Bogataj, M., and Mladenić, D. D4.1 Stream Learning Baseline Document. Reported 30. April 2018.

[10] Sinergise. sentinel-hub/sentinel2-cloud-detector: Sentinel Hub Cloud Detector for Sentinel-2 images in Python. `https://github.com/sentinel-hub/sentinel2-cloud-detector`. Accessed 14 August 2018.

[11] Sinergise. sentinel-hub/sentinelhub-py: Download and process satellite imagery in Python scripts using Sentinel Hub services. `https://github.com/sentinel-hub/sentinelhub-py`. Accessed 14 August 2018.

[12] Spaceflight 101. Sentinel-2 Spacecraft Overview. `http://spaceflight101.com/copernicus/wp-content/uploads/sites/35/2015/09/8723482_orig-1024x538.jpg`. Accessed 14 Aug. 2018.

[13] Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P., and Koetz, B. Production of a dynamic cropland mask by processing remote sensing image series at high temporal and spatial resolutions. *Remote Sensing 8(1)* (2016), 55.

[14] Waldner, F., Canto, G. S., and Defourny, P. Automated annual cropland mapping using knowledge-based temporal features. *ISPRS Journal of Photogrammetry and Remote Sensing 110* (2015), 1 – 13.

Towards a semantic repository of data mining and machine learning datasets

Ana Kostovska
Jožef Stefan IPS &
Jožef Stefan Institute
Ljubljana, Slovenia
ana.kostovska@ijs.si

Sašo Džeroski
Jožef Stefan Institute &
Jožef Stefan IPS
Ljubljana, Slovenija
saso.dzeroski@ijs.si

Panče Panov
Jožef Stefan Institute & Jožef
Stefan IPS
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

With the exponential growth of data in all areas of our lives, there is an increasing need of developing new approaches for effective data management. Namely, in the field of Data Mining (DM) and Knowledge Discovery in Databases (KDD), scientists often invest a lot of time and resources for collecting data that has already been acquired. In that context, by publishing open and FAIR (Findable, Accessible, Interoperable, Reusable) data, researchers could reuse data that was previously collected, preprocessed and stored. Motivated by this, we conducted extensive review on current approaches, data repositories and semantic technologies used for annotation, storage and querying of datasets for the domain of machine learning (ML) and data mining. Finally, we identify the limitations of the existing repositories of datasets and propose a design of a semantic data repository that adheres to FAIR principles for data management and stewardship.

1. INTRODUCTION

One of the main use of data is in the process of knowledge discovery, where scientist employ ML and DM methods and try to solve various real-life problems from diverse fields, from systems biology and medicine, to ecology and environmental sciences. In order to obtain their objectives, they need high-quality data. The quality of the data is crucial to a DM project's success. Ultimately, no level of algorithmic sophistication can make up for low-quality data. On the other hand, progress in science is best achieved by reproducing, reusing and improving someone else's work. Unfortunately, datasets are not easily obtained, and even if they are, they come with limited reusability and interoprability.

A key-aspect in advancing research is making data open and **FAIR**. FAIR are four principles that have been recently introduced to support and promote good data management and stewardship [17]. Data must be easily findable (**Findability**) by both humans and machines. This means data should be semantically annotated with rich metadata and all the resources must be uniquely identified. The metadata should always be accessible (**Accessibility**) by standardized communication protocols such as HTTP(S) or FTP, even when the data itself is not. Data and metadata from different data sources can be automatically combined (**Interoperabiliy**). To do so, the benefits of formal vocabularies and ontologies should be exploited. Data and metadata is released with provenance details and data usage licence, so that humans and machines know whether data can be replicated and reused or not (**Reusabiliy**).

The benefits of publishing FAIR data are manifold. It speeds up the process of knowledge discovery and reduces the consumption of resources. When the FAIR-compliant data at hand does not contain all the information needed it can be easily integrated with data from external sources and boost the overall KDD performance [12].

Semantic data annotation, being very powerful technique, is massively used in some domains, i.e. medicine, but it is sill in the early phases in the domain of data mining and machine learning. To the best of our knowledge, there are no semantic data repositories that adhere to the FAIR principles. We recognize the ultimate benefits of having one and we are going in depths of the research covering semantic data annotation, ontology usage, storing and querying of data.

2. BACKGROUND AND RELATED WORK

The Semantic Web (Web 3.0) is an extension of the World Wide Web in which information is given semantic meaning, enabling machines to process that information. The aim of the Semantic Web initiative is to enhance web resources with highly structured metadata, known as semantic annotations. When one resource is semantically annotated, it becomes a source of information that is easy to interpret, combine and reuse by the computers [13]. In order to achieve this, the Semantic Web uses the concept of Linked Data. Linked data is build upon standard web technologies [7] including HTTP, RDF, RDFS, URIs, Ontologies, etc.

For uniquely identifying resources across the whole Linked Data, each resource is given a **Unified Resource Identifier (URI).** The resources are then enriched with terms from controlled vocabularies, taxonomies, thesauruses, and ontologies. The standard metadata model used for logical organization of data is called **Resource Description Framework (RDF).** Its basic unit of information is the triplet compiled from a subject, a predicate, and an object. These three components define the concepts and relations, the building blocks of an ontology.

In the context of computer science, **ontology** is "an explicit formal specifications of the concepts and relations among them that can exist in a given domain" [3]. As computational artifacts, they provide the basis for sharing meaning both at machine and human level. When creating an ontology, there are multiple languages to choose from. **RDF Shema (RDFS)** is ontology language with small expressive power. It provides mechanisms for creating simple taxonomies of

concepts and relations. Another commonly used ontology language is the **Web Ontology Language (OWL).** OWL supports creation of all ontology components: concepts, instances, properties (or relations). Finally, **SPARQL**[1] is standard, semantic query language used for querying fast-growing private or public collections of structured data on the Web or data stored in RDF format.

There are diiferent technologies for storing data and metadata. The most broadly used are **relational databases**, digital databases based on the relational model of data organized in tables, forming entity-relational model. Another approach that became popular with the appearance of Big Data are **NoSQL** databases [5], which are flexible databases that do not use relational model. **Triplestores** are specific type of NoSQL databases, that store triples instead of relational data. Triplestores use URIs and can be queried over trillions of records, which makes them very applicable.

Data in an information system can reside in different heterogeneous data sources, both internal and external to the organization. In this setting, the relevant data from the diverse sources should be integrated. Accessing disparate data sources has been a difficult challenge for data analysts to achieve in modern information systems, and an active research area. **OBDA** [1, 11] is much longed-for method that addresses this problem. It is a new paradigm, based on a three-level architecture constituted of the ontology, the data sources, and the mappings between the two (see **Figure 1**). With this approach, OBDA provides data structure description, as well as semantic description of the concepts in the domain of interest and roles between them.

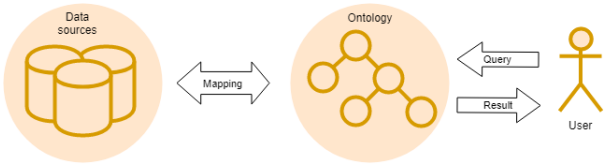


Figure 1. The OBDA architecture

In the context of semantic ML data repository, we group ontologies in three categories, i.e., ontologies for describing machine learning and data mining, ontologies for provenance information, and domain ontologies. **OntoDM** ontology describes the domain of data mining. It is composed of three sub-ontologies: OntoDT [10] - generic ontology for representation of knowledge about datatypes; OntoDM-core [8] - ontology of core data mining entities (e.g., data, DM task, generalizations, DM algorithms, implementations of algorithms, DM software); OntoDM-KDD [9] - ontology for representing the knowledge discovery process following CRISP-DM process model. **The Data Mining OPtimization Ontology (DMOP)** [6] has been designed to support automation at various choice points of the DM process, i.e., choosing algorithms, models, parameters. **The PROV Ontology (PROV-O)**[2] **and Dublin Core vocabulary** [16] facilitate the discovery of electronic resources by providing a base for describing provenance information about resources.

[1] https://www.w3.org/TR/rdf-sparql-query/
[2] https://www.w3.org/TR/prov-o/

There are numerous repositories of ML datasets available online. The UCI repository[3] is the most popular repository of ML datasets. Each dataset is annotated with several descriptors such as dataset characteristics, attribute characteristics, associated task, number of instances, number of attributes, missing values, area, etc. Similarly, Kaggle Datasets[4], Knowledge Extraction based on Evolutionary Learning (KEEL), and Penn Machine Learning Benchmarks (PMLB)[5] are well-known dataset repository that provide users with data querying based on the descriptors attached to the datasets. OpenML[6] is an open source platform designed with the purpose of easing the collaboration of researchers within the machine learning community [14]. Researchers can share datasets, workflows and experiments in such a way that they can be found and reused by others. When the data format of the datasets is supported by the platform, the datasets are annotated with measurable characteristics [15]. These annotations are saved as textual descriptors and are used for searching through the repository.

In contrast to the above mentioned repositories, there are frameworks in other domains that offer advanced techniques for describing, storing and querying datasets. One cutting-edge framework in the domain of neuroscience is **Neuroscience Information Framework(NIF)** [4]. Its core objective is to create a semantic search engine that benefits from semantic indexes when querying distributed resources by keywords. **The Gene Ontology Annotation (GOA)**, is a database that provides high-quality annotations of genome data [2]. The annotations are based on GO, a vocabulary that defines concepts related to gene functions and relation among them. Large part of the annotations are generated electronically by converting existing knowledge from the data to GO terms. Electronic annotations are associated with high-level ontology terms. The process of generating more specific annotations can hardly be automated with the current technologies, therefore it is done manually.

3. CRITICAL ASSESSMENT

In this section, we conduct critical assessment of the current research based on the review presented in the previous section.

Semantic Web technologies. The whole stack of semantic technologies provide ways of making the content readable by machines. The metadata that describes the content can be used not only to disregard useless information, but also for merging results to provide a more constructed answer. A major drawback of this process of giving data a semantic meaning is that it is time consuming and requires great amount of resources, thus people sometimes feel unmotivated to do it. Another point to make is that semantic annotations cannot solve the ambiguities of the real world.

Technologies for storing data and metadata. The data in relational databases is stored in a very structured way, making them a good choose for applications that relay

[3] https://archive.ics.uci.edu/ml/
[4] https://www.kaggle.com/datasets
[5] https://github.com/EpistasisLab/penn-ml-benchmarks
[6] https://www.openml.org/

on heavy data analysis. Moreover the referential integrity guarantees that transactions are processed reliably. While relational databases are a suitable choice for some applications, they have difficulties dealing with large amounts of data. On the other hand, NoSQL databases were designed primarily for big data and can be run on cluster architectures. Non-relational databases store unstructured data, with no logical schema. They are flexible, but this comes with the price of potentially inconsistent data.

Describing data and metadata. OntoDM is an ontology that describes the domain of DM, ML and KDD with a great level of detail. Because it covers a wide area, some parts would be irrelevant for our application. DMOP is ontology built with the special use case of optimizing the DM process. Nevertheless, both of them can be used for describing ML and DM datasets. DC vocabulary and PROV-O define a wide range of provenance terms, therefore both of them can be employed in the provenance metadata generation.

Repositories of machine learning datasets. The UCI repository offers a wide range of datasets, but they are not available through a uniform format or API. Although it also provides data descriptors for searching the data, a major setback is that none of the descriptors is based on any vocabulary or ontology, which certainly limits interoperability. Kaggle Datasets, KEEL, PMLB also provide similar meta annotations, but they all lack semantic interpretability. Another shortcoming of the UCI repository, KEEL and PMLB is that they don't allow uploading new datasets. All datasets stored in the OpenML repository can be downloaded in CSV or ARFF format. The annotation are based on Exposé ontology, and they can be downloaded in JSON, XML or RDF format. A major weakness of this repository is that annotations are not stored, but they are calculated on-the-fly and can not be used for semantic inference.

Frameworks for describing, storing and querying domain datasets. The NIF framework is very progressive in terms of semantic annotation, storing, and querying. Its advantages come from providing domain experts with the ability to contribute to the ontology development, by adding new terms through the use of Interlex. It has a powerful search engine, and it follows the OBDA paradigm. Heterogeneous data is stored in its original format. The user defined, keyword query is mapped to ontological terms to find synonyms, and then translated to a query relevant to the individual data store. With respect to the genomics domain, GOA database is favourable because of its high-quality annotations. Curators put extreme efforts in generating manual annotations. To speed up the query execution it uses the Solr document store. Another superiority of GOA database is that it provides advanced filtering of the annotations, for downloading customized annotation sets. The deficiency of NIF and GOA database is that they are not able to query and access the annotations in RDF format, which is an emerging standard for representing semantic information

4. PROPOSAL FOR SEMANTIC REPOSITORY OF DM/ML DATASETS

In this section, we propose three possible architecture designs of the semantic data repository for the domain of ML and DM. The proposals are based on the critical review of the approaches and technologies. Each of the proposed architectures has positive and negative sides, so there will be trade-off when choosing one.

The common part of the three designs is that DM and ML datasets will be annotated through a semantic annotation engine. The semantic query engine will receive SPARQL query as input, and it will bring back results in form of set of RDF triples. There will be SPARQL endpoint through which users can specify the query used as input in the semantic query engine. Another open possibility is to enable users to query data and metadata by simply writing keywords. Later, the system itself generates SPARQL query based on those keywords. The anotation schema used by the semantic annotation engine will be based on three different types of ontologies such as ontologies for DM and ML (e.g., OntoDT, OntoDM-core, Onto-KDD, DMOP), domain ontologies, and ontologies and schemes for describing provenance information (e.g., Dublin Core ontology, PROV-O). Part of the annotations will be generated automatically, e.g., annotations related to datatypes, while others will be semi-automatically because they require concept mapping, e.g., annotations based on domain ontologies.

We plan to build a web-based user interface that will enable users to search and query both datasets and metadata annotations. Users will be given a chance of uploading new datasets in CSV or ARFF fromat. Besides the dataset, users will be expected to specify some additional information about it such as data mining task they plan to execute on the data, domain, provenance information, descriptions of the attributes, etc. Since the whole process of semantic annotation can't be automatic, when new dataset is uploaded, it won't be immediately available on the site. First it must be curated, and only when the complete set of metadata annotations is generated, the metadata will be published online. The dataset itself will be released under clear data usage licence.

The three architectural designs differ in the way of storing the datasets. The metadata annotations will be RDF triplets and they will be stored in triplestore that optimizes physical storage. Next, we briefly explain the differences between storing the datasets and what are the effects on querying.

Proposal I. The simplest approach of storing a dataset would be to store it in RDF format in the same triplestore as the metadata. The datasets from their original format, will be converted to RDF triples. Having only one triplestore will ease querying, but it will require more storage capacity (see Figure 2).

Proposal II. The second option is to store the datasets in a relational database and the metadata in RDF triplestore. Datasets from CSV or ARFF format will be translated into a relational database. Here, querying becomes more complicated, for which we will need a federated query engine. A federated query engine allows simultaneous search on multiple data sources. A user makes a single query request, which is distributed across the management systems participating in the federation and translated to a query written in a language relevant to the individual system. We will have two data stores, one for the data itself and one for the metadata. For querying the two data stores, we will still use the same

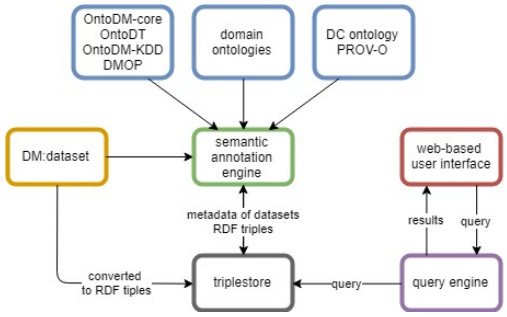


Figure 2. Architectural design I

RDF query language, SPARQL. In order to query the relational database with SPARQL, it will be mapped to virtual RDF graph (see Figure 3).

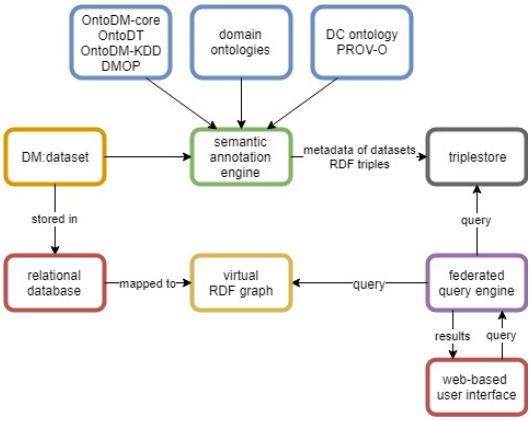


Figure 3. Architectural design II

Proposal III. Instead of mapping the relational database to virtual RDF graph, we can use the OBDA methodology and federated querying to use a combination of SQL queries and SPARQL queries. Metadata will be queried with SPARQL queries, but for the datasets, they will be mapped to SQL queries. The integrated results are brought back to the user (see Figure 4).

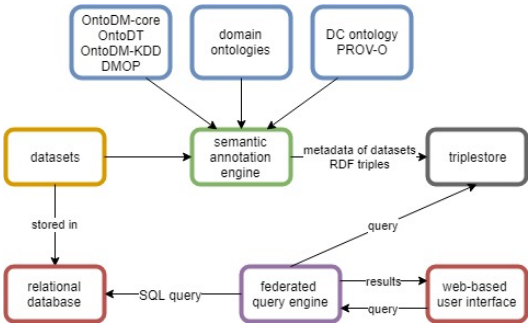


Figure 4. Architectural design III

5. CONCLUSION

We have conducted a literature overview of research being done in the field of semantic annotation, storage, and querying of ML and DM datasets. We also examined specific implementations of frameworks in the domain of neuroscience and genomics. Taking into consideration the critical assessment of the current state-of-the-art we will construct semantic data repository for ML and DM datasets. The semantic repository would be utilized for easy access of semantically rich annotated datasets and semantic inference. This, will improve the reproducibility and reusability in ML and DM research area. Moreover, annotating the datasets with domain ontologies will facilitate the process of understanding the analyzed data. As of now, we have three proposed architectural designs for the semantic data repository that differ in the way of storing the datasets. We will either store both data and metadata in a triplestore, or we will have multiple data stores which will require usage of tools and methods from the ontology based data access paradigm.

Acknowledgements

The authors would like to acknowledge the support of the Slovenian Research Agency through the projects J2-9230, N2-0056 and L2-7509 and the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia through its scholarship program.

6. REFERENCES

[1] Mihaela A Bornea et al. Building an efficient rdf store over a relational database. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 121–132. ACM, 2013.

[2] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic acids research*, 43(D1):D1049–D1056, 2014.

[3] Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

[4] Amarnath Gupta et al. Federated access to heterogeneous information resources in the neuroscience information framework (nif). *Neuroinformatics*, 6(3):205–217, 2008.

[5] Jing Han et al. Survey on nosql database. In *Pervasive computing and applications (ICPCA), 2011 6th international conference on*, pages 363–366. IEEE, 2011.

[6] C Maria Keet et al. The data mining optimization ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:43–53, 2015.

[7] Brian Matthews. Semantic web technologies. *E-learning*, 6(6):8, 2005.

[8] Panče et al. Panov. Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28(5-6):1222–1265, 2014.

[9] Panče Panov et al. Ontodm-kdd: ontology for representing the knowledge discovery process. In *International Conference on Discovery Science*, pages 126–140. Springer, 2013.

[10] Panče Panov et al. Generic ontology of datatypes. *Information Sciences*, 329:900–920, 2016.

[11] Antonella Poggi et al. Linking data to ontologies. In *Journal on data semantics X*, pages 133–173. Springer, 2008.

[12] Petar Ristoski and Heiko Paulheim. Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36:1–22, 2016.

[13] Gerd Stumme et al. Semantic web mining: State of the art and future directions. *Web semantics: Science, services and agents on the world wide web*, 4(2):124–143, 2006.

[14] Jan N Van Rijn et al. Openml: A collaborative science platform. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 645–649. Springer, 2013.

[15] Joaquin Vanschoren et al. Taking machine learning research online with openml. In *Proceedings of the 4th International Conference on Big Data, Streams and Heterogeneous Source Mining*, pages 1–4. JMLR. org, 2015.

[16] Stuart Weibel. The dublin core: a simple content description model for electronic resources. *Bulletin of the Association for Information Science and Technology*, 24(1):9–11, 1997.

[17] Mark D Wilkinson et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.

Towards a semantic store of data mining models and experiments

Ilin Tolovski
Jožef Stefan International Postgraduate School & Jožef Stefan Institute
Ljubljana, Slovenia
ilin.tolovski@ijs.si

Sašo Džeroski
Jožef Stefan Institute & Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
saso.dzeroski@ijs.si

Panče Panov
Jožef Stefan Institute & Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

Semantic annotation provides machine readable structure to the stored data. We can use this structure to perform semantic querying, based on explicitly and implicitly derived information. In this paper, we focus on the approaches in semantic annotation, storage and querying in the context of data mining models and experiments. Having semantically annotated data mining models and experiments with terms from domain ontologies and vocabularies will enable researchers to verify, reproduce, and reuse the produced artefacts and with that improve the current research. Here, we first provide an overview of state-of-the-art approaches in the area of semantic web, data mining domain ontologies and vocabularies, experiment databases, representation of data mining models and experiments, and annotation frameworks. Next, we critically discuss the presented state-of-the-art. Furthermore, we sketch our proposal for an ontology-based system for semantic annotation, storage, and querying of data mining models and experiments. Finally, we conclude the paper with a summary and future work.

1. INTRODUCTION

Storing big amounts of data from a specific domain comes in hand with several challenges, one of them being to semantically represent and describe the stored data. Semantic representation enables us to infer new knowledge based on the one that we assert, i.e. the description and annotation of the data. This can be done by providing semantic annotations of the data with terms originating from a vocabulary or ontology describing the domain at hand. In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined [15]. Ontologies provide more detailed description of a domain, first by organizing the classes into a taxonomy, and further on by defining relations between classes. With semantic annotation we attach meaning to the data, we can infer new knowledge, and perform queries on the data.

Data mining and machine learning experiments are conducted with faster pace than ever before, in various settings and domains. In the usual practice of conducting data mining experiments, almost none of the settings are recorded, nor the produced models are stored. These predicaments make for a research that is hard to verify, reproduce and upgrade. This is also in line with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, introduced by Wilkinson et al. [9]. Implementing these principles for the annotation, storing, and querying of data mining models and experiments will provide a solid ground for researchers interested in reproducing and reusing the results from the previous research on which they can build and improve.

In the literature, there exist some approaches that address some of these problems. In both ontology engineering and data mining community, there are approaches that aim towards describing the data mining domain, as described in Section 2. Furthermore, Vanschoren et al. [5] developed the OpenML system, a machine learning experiment database for storing various segments of a machine learning experiment such as datasets, flows (algorithms), runs, and completed tasks.

In other domains, such as life sciences, storing annotated data about experiments and their results is a common practice. This is mostly due to the fact that the experiments are more expensive to conduct, and require specific preparations. From the perspective of annotation frameworks, there are significant advances in these domains, such as The Center for Expanded Data Annotation and Retrieval (CEDAR) workbench [8] , and the OpenTox framework [11].

This paper is organized as follows. First, we make an overview of the state-of-the-art approaches in annotating, storing, and querying of models and experiments. Next, we critically assess these approaches and sketch a proposal for a system for annotating, storing and querying data mining models and experiments. Finally, we provide a summary and discuss the possible approaches for further work.

2. BACKGROUND AND RELATED WORK

The state-of-the-art in semantic annotation of data mining models and experiments provides very diverse research, ranging from domain-specific data mining ontologies, experiment databases, to new languages for deploying annotations in unified format. Here, we provide an introduction to the state-of-the-art in semantic web, ontologies and vocabularies, representations of data mining models and experiments, experiment databases, and annotation frameworks.

Semantic technologies. The Semantic Web is defined as an extension of the current web in which information is

given well-defined meaning, enabling computers and people to work in cooperation [14]. The stack of technologies consists of multiple layers, however, in this paper we will focus on the ones essential for our scope of research. Resource Description Framework (RDF) represents a metadata data model for the Semantic Web, where the core unit of information is presented as a triple. A triple describes the subject by its relationship, which is what the predicate resembles, with the object. RDF files are stored in triple store (typically organized as relational or NoSQL databases [12]), on which we can perform semantic queries, by using querying languages such as SPARQL. Finally, ontology languages, such as Resource Description Framework Schema (RDFS) and Ontology Web Language (OWL), are formal languages used to construct ontologies. RDFS provides the basis for all ontology languages, defining basic constructs and relations, while OWL is far more expressive enabling us to define classes, properties, and instances.

Ontologies & vocabularies. Currently, there are several ontologies that describe the data mining domain. These include the OntoDM ontology [16], DMOP ontology [7], Expose [4], KDDOnto [1], and KD ontology [10]. MEX [2] is an interoperable vocabulary for annotating data mining models and experiments with metadata. In addition there have been developments in formalism for representing scientific experiments in general, such as the EXPO ontology [6].

Representation of models. With the constant development of new environments for developing data mining software, it is necessary to have a unified representation of the constructed data mining models and the conducted experiments. The first open standard was the Predictive Model Markup Language (PMML). For a period of time it provided transparent and intuitive representation of data mining models and experiments. However, due to the fast growth in the development of new data mining methods, PMML was unable to follow the pace and extend its more and more complicated specification. Its successor, the Portable Format for Analytics (PFA), was developed having the PMML's drawbacks as guidelines for improvement.

Experiment and model databases. Storing already conducted experiments in a well structured and transparent manner is essential for researchers to have available, verifiable, and reproducible results. An experiment database is designed to store large number of experiments, with detailed information on their environmental setup, the datasets, algorithms and their parameter settings, evaluation procedure, and the obtained results [3]. The state-of-the-art in storing setups and results is abundant with approaches and solutions in different domains. For example, OpenML[1] is the biggest machine learning repository of data mining datasets, tasks, flows, and runs, the BioModels[2] repository stores more than 8000 experiments and models from the domains of systems biology, and ModelDB[3] is an online repository for storing computational neuroscience models.

Annotation frameworks. When it comes to frameworks for (semi) automatically or manually annotating data, there are several solutions that exist outside of the data mining domain, which provide innovative approaches and good foundation for development in the direction of creating a software to enable ontology-based semantic annotation of models and experiments, their storage and querying. The CEDAR Workbench [13] provides an intuitive interface for creating templates and metadata annotation with concepts defined in the ontologies available at BioPortal[4]. On the other hand, OpenTox [11] represents domain specific framework that provides unified representation of the predictive modelling in the domain of toxicology.

3. CRITICAL ASSESSMENT

In this section, we will critically assess the presented state-of-the-art in Section 2 in the context of semantic annotation, storage and querying of data mining models and experiments.

The state-of-the-art in *ontology design* for data mining provides well documented research with various ontologies that thoroughly describe the domain from different aspects and can be used in various applications. OntoDM provides unified framework of top level data mining entities. Building on this, it describes the domain in great detail, containing definitions for each part of the data mining process. Because of the wide reach, it lacks a particular use case scenario. On the other hand, this same property makes this ontology suitable for wide range of applications where there is a need of describing a part of the domain.

Ontologies like EXPO and Exposé have a essential meaning in the research since the first one describes a very wide and important interdisciplinary domain, while the latter uses it as a base for defining a specific sub-domain. DMOP ontology describes the process of algorithm and model selection in the context of semantic meta mining. Both the KD ontology and KDDOnto describe the knowledge discovery process in the context of constructing knowledge discovery workflows. They differ mainly in the key concepts on which they were built. At the same time, the MEX vocabulary provides a lightweight framework for automating the metadata generation. Since it is tied with Java environment, it provides a library which only uses the MEX API and can also be implemented in other programming languages.

All in all, the current state of the art in ontologies for data mining provides a good foundation for development of applications which will be based on one or several of these ontologies. Given the wide of coverage they can be easily be combined in a manner to suit the application at hand.

In the area of *descriptive languages for data mining models and experiments*, one can see the path of progress in research. PMML was the first, ground-breaking, XML-based descriptive language. However, with the expansion of the data mining domain, several weaknesses of PMML emerged. The language was not extensible, users could not create chains of models, and it was not compatible with the distributed data processing platforms. Therefore, the same community started working on a new, extensible, portable

[1] `https://www.openml.org/`
[2] `http://www.ebi.ac.uk/biomodels/`
[3] `https://senselab.med.yale.edu/modeldb/`
[4] `https://bioportal.bioontology.org/`

language. Since its inception, the PFA format was intended to fill the small gaps that PMML had. Made up of analytic primitives, implemented in Python and Scala, it provides the users with more customizable framework, where they can create custom models, model chains, and implement them in a parallelized setting.

Storing and annotating experiments is of great significance in multiple scenarios. First, in domains where conducting the experiment is not a trivial task, i.e. the physical or financial conditions challenge the process, there needs to be a database where the setup and the findings of the experiment will be saved. For example, in BioModels.net we have two groups of experiments: Manually curated with structured metadata, and experiments without structure. The main drawback with this type of storage is the need for manual curation of the metadata. It is repetitive, time consuming task for which there is a strong need to be automated.

In the domain of neuroscience, ModelDB provides an online service for storing and searching computational neuroscience models. In this database, alongside the files that constitute the models, researchers also need to upload the code that defines the complete specification of the attributes of the biological system represented in the model, together with files that describe the purpose and application of the model. Therefore, researchers can search the database for models with specific applications describing biological systems.

OpenML provides a good framework for storing and annotating data mining datasets, experimental setups and runs, as well as algorithms. One particular drawback of OpenML is that it does not store the actual models that are produced from each experimental run, and one can not query the models. Furthermore, it's founded on relational-database which can not provide execution of semantic queries.

All in all, these three examples show significant advances in storing and annotating models and experiments. However, there is also a significant room for improvement in the direction of storing the models and experiments into NoSQL databases that are better suited for this task.

Finally, in the context of annotation tools the CEDAR Workbench and the OpenTox Framework provide a good insight in annotation frameworks. CEDAR enables the user to execute the annotations in modular manner by creating templates and adding elements to them. After curating the annotations, they can export the schemas either in JSON, JSON-LD, or RDF file. OpenTox [11] is also based on ontology terms and represents a complete framework that describes the predictive process in toxicology, starting with toxicity structures and ending with the predictive modelling.

4. A PROPOSAL FOR SEMANTIC STORE OF MODELS AND EXPERIMENTS

After analysing the previous and current research, we can conclude that despite the great achievements, there is a wide area for improvement in which we will contribute in the upcoming period by developing an ontology-based framework for storage and annotation of data mining models and experiments. In order to annotate a data mining experiment, we need to have complete information about the conditions in which that experiment was conducted. Namely, we need to have an annotated dataset, annotation of the algorithm and its parameter settings for the specific run of the experiment. Since one experiment usually consists of multiple algorithm runs we annotate each run separately, as well as each of the results from each of them. For annotating the results, we use the definitions of the performance metrics formalized in the data mining ontologies. A sketched example of the proposed solution is shown in Figure 1.

The proposed system for ontology-based annotation, storage, and querying of data mining experiments and models will consist of several components. The users will interact with the system through an user interface enabling them to run experiments on a data mining software, which will export models and experiment setups to a semantic annotation engine. For example, for testing purposes we plan to use CLUS[5] software for predictive clustering and structured output prediction, which generates different types of models and addresses different data mining tasks.

In the semantic annotation engine, the data mining models and experiments will be annotated with terms from the extended OntoDM ontology and then stored in a database. Once stored, the users will be able to semantically query the models and experiments in order to infer new knowledge. This will be done through a querying engine based on the SPARQL language, accessible through a user interface.

In order to perform annotation, we will extend the existing OntoDM ontology by adding a number of new terms, linking it to other domain ontologies, such as Exposé and EXPO. Linking OntoDM to these ontologies will extend the domain of OntoDM towards connecting the data mining entities that it already covers with new entities that describe the experimental setup and principles. With this we will obtain a schema for annotation of data mining models and experiments. The schema will then be used to annotate the data mining models and experiments through a semantic annotation engine. The engine will have to read the models and experiments from a data mining software system, annotate them with terms from developed schema and produce an RDF representation of the annotated data.

Furthermore, the RDF graphs will be stored in a triple store database. Since the data mining models and experiments differ a lot in their structure, we have yet to decide on the type of database in which we will store them. The data stored in this way is set for performing semantic queries on top of it. Therefore, we will develop a SPARQL-based querying enigne so the users can perform predefined or custom semantic queries on top of the storage base.

Finally, the format of the results is another point where we need to decide whether the results will be presented as RDF graphs, or in a different format (such as JSON) that is easier to interpret. This software package along with the storage will then be added as a module to the CLUS software, developed at the Department of Knowledge Technologies.

[5] `http://sourceforge.net/projects/clus`

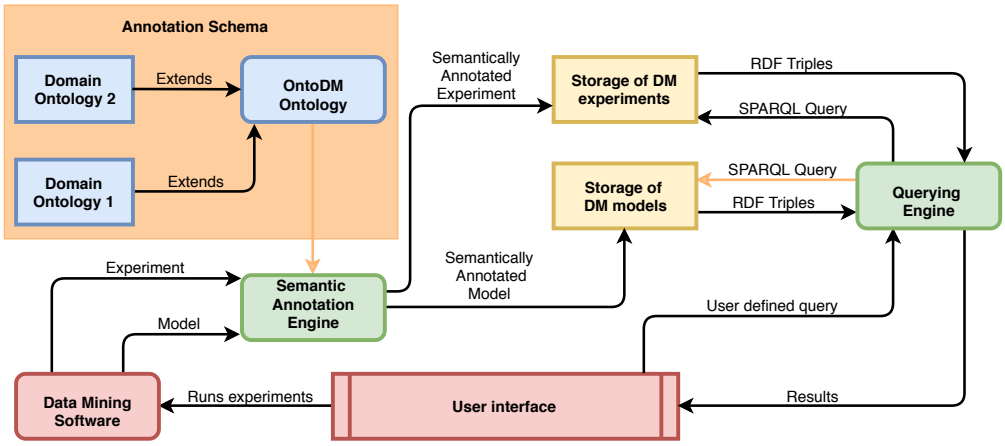


Figure 1. Schema of the proposed solution

5. CONCLUSION & FURTHER WORK

In this paper, we presented the state-of-the-art in annotation, storage and querying in the light of designing a semantic store of data mining models and experiments. We first gave an overview of semantic web technologies, such as RDF, SPARQL, RDFS, and OWL that provide a complete foundation for annotation and querying of data.

Furthermore, we critically reviewed the state-of-the-art ontologies and vocabularies for describing the domain of data mining provide detailed description of the domain of data mining and machine learning (OntoDM, Expose, KD Ontology, DMOP and KDDOnto, MEX). Next, we focused on experiment databases as repositories where the experiment datasets, setups, algorithm parameter settings, and the results are available for the performed experiments in various domains. Furthermore, we saw that annotation frameworks provide environments for (semi) automatically or manually annotating data, by discussing two frameworks from the domains of biomedicine and toxicology in order to analyze best practices present in those domains.

Finally, given the performed analysis of the state-of-the-art, we outlined our proposal for an ontology-based framework for annotation, storage, and querying of data mining models and experiments. The proposed framework consists of an annotation schema, a semantic annotation engine, and storage for data mining models and experiments with a querying engine, all of which will be controlled from an user interface. It will allow users to semantically query their data mining models and experiments in order to infer new knowledge.

In the future, we plan to adapt this framework for the needs of research groups or companies that conduct high volume of data mining experiments, enabling them to obtain a queryable knowledge base consisting of annotated metadadata for all experiments and produced models. This will enable them to reuse existing models on new data for testing purposes, infer knowledge based on past experimental results, all while saving time and computational resources.

Acknowledgements

The authors would like to acknowledge the support of the Slovenian Research Agency through the projects J2-9230, N2-0056 and L2-7509 and the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia through its scholarship program.

6. REFERENCES

[1] Claudia Diamantini et al. KDDOnto: An ontology for discovery and composition of kdd algorithms. *Towards Service-Oriented Knowledge Discovery (SoKD'09)*, pages 13–24, 2009.

[2] Diego Esteves et al. MEX Vocabulary: a lightweight interchange format for machine learning experiments. In *Proceedings of the 11th International Conference on Semantic Systems*, pages 169–176. ACM, 2015.

[3] Hendrick Blockheel et al. Experiment databases: Towards an improved experimental methodology in machine learning. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 6–17. Springer, 2007.

[4] Joaqin Vanschoren et al. Exposé: An ontology for data mining experiments. In *Towards service-oriented knowledge discovery (SoKD-2010)*, pages 31–46, 2010.

[5] Joaqin Vanschoren et al. Taking machine learning research online with OpenML. In *Proceedings of the 4th International Conference on Big Data, Streams and Heterogeneous Source Mining*, pages 1–4. JMLR. org, 2015.

[6] Larisa N Soldatova et al. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006.

[7] Maria C Keet et al. The Data Mining OPtimization Ontology. *Web Semantics: Science, Services and Agents on the World Wide Web*, 32:43–53, 2015.

[8] Mark A Musen et al. The Center for Expanded Data Annotation and Retrieval. *Journal of the American Medical Informatics Association*, 22:1148–1152, 2015.

[9] Mark D. Wilkinson et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 2016.

[10] Monika Záková et al. Automating knowledge discovery workflow composition through ontology-based planning. *IEEE Transactions on Automation Science and Engineering*, 8:253–264, 2011.

[11] Olga Tcheremenskaia et al. OpenTox predictive toxicology framework: toxicological ontology and semantic media wiki-based openToxipedia. In *Journal of biomedical semantics*, page S7, 2012.

[12] Olivier Curé et al. *RDF database systems: triples storage and SPARQL query processing*. Morgan Kaufmann, 2014.

[13] Rafael S Gonçalves et al. The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that Describe Scientific Experiments. In *International Semantic Web Conference*, pages 103–110. Springer, 2017.

[14] Tim Berners-Lee et al. The semantic web. *Scientific American*, 284:34–43, 2001.

[15] Tom Gruber. Ontology. *Encyclopedia of database systems*, pages 1963–1965, 2009.

[16] Panče Panov. *A Modular Ontology of Data Mining*. PhD thesis, Jožef Stefan IPS, Ljubljana, Slovenia, 2012.

A Graph-based prediction model with applications

[Extended Abstract]

András London*
University of Szeged, Institute of Informatics
Poznan University of Economics, Department of Operations Research

József Németh
University of Szeged, Institute of Informatics

Miklós Krész
InnoRenew CoE
University of Primorska, IAM
University of Szeged, Institute of Applied Sciences

ABSTRACT

We present a new model for probabilistic forecasting using graph-based rating method. We provide a "forward-looking" type graph-based approach and apply it to predict football game outcomes by simply using the historical game results data of the investigated competition. The assumption of our model is that the rating of the teams after a game day correctly reflects the actual relative performance of them. We consider that the smaller the changing of the rating vector – contains the ratings of each team – after a certain outcome in an upcoming single game, the higher the probability of that outcome. Performing experiments on European football championships data, we can observe that the model performs well in general and outperforms some of the advanced versions of the widely-used Bradley-Terry model in many cases in terms of predictive accuracy. Although the application we present here is special, we note that our method can be applied to forecast general graph processes.

Categories and Subject Descriptors

I.6 [**Simulation and Modeling**]: Applications; I.2 [**Artificial Intelligence**]: Learning

1. INTRODUCTION

The problem of assigning scores to a set of individuals based on their pairwise comparisons appears in many areas and activities. For example in sports, players or teams are ranked according to the outcomes of games that they played; the impact of scientific publications can be measured using the relations among their citations. Web search engines rank websites based on their hyperlink structure. The centrality of individuals in social systems can also be evaluated according to their social relations. Ranking of individuals based on the underlying graph that models their bilateral relations has become the central ingredient of Google's search engine and later it appeared in many areas from social network analysis to optimization in technical networks (e.g. road and electric networks) [16].

Making predictions in general, and especially in sports as well, is a difficult task. The predictions generally appear in the form of betting odds, that, in the case of "fixed odds", provide a fairly acceptable source of expert's predictions regarding sport games outcomes [21]. Thanks to the increasing quantity of available data the statistical ranking, rating and prediction methods have become more dominant in sports in the last decade. A key question is that how accurate these evaluations are, more concretely, the outcomes of the upcoming games how accurately can be predicted based on the statistics, ratings and forecasting models in hand.

Statistics-based forecasting models are used to predict the outcome of games based on some relevant information of the competing teams and/or players of the teams. A detailed survey of the scientific literature of rating and forecasting methods in sports is beyond the scope of this paper, we refer only some important and recent results in the topic. For some papers with detailed literature overview and sport applications of the the celebrated *Bradley-Terry model* [3], see e.g. [5, 7, 24]). Other popular approach is the Poisson goal-distribution based analysis. For some references, see for instance [10, 15, 20]. In these models the goals scored by the playing teams follow a Poisson distribution with parameter that is a function of attack and defense "rate" of the respective teams. A large family of prediction models only consider the game results win, loss (and tie) and usually uses some probit regression model, for instance [11] and [13]. More recently, well-known data mining techniques, like artificial neural networks, decision trees and support vector machines have also become very popular; some references - without being exhaustive - see e.g [8, 9, 14, 18].Based on the huge literature it can be concluded that the prediction accuracy strongly depends on the investigated sport and the feature set of the machine learning algorithms used. A notable part of prediction models based on the historical data of game results use the methodology of ranking and rating. Some recent articles in the topic are e.g. [2, 6, 12, 17, 23]. Specifically highlighting [2] the authors analyzed the he predictive power of eight sports ranking methods using only win-loss and score difference data of American major sports. They found that the least squares and random walker meth-

*Corresponding author, email: london@inf.u-szeged.hu

ods have significantly better predictive accuracy than other methods. Moreover, utilizing score-differential data are usually more predictive than those using only win-loss data.

In contrast to those techniques that use the actual respective strength of the two competing teams, we provide a graph-based and forward-looking type approach. The assumption of our model is that if a rating of the teams after a game day correctly reflects the actual relative performance, then the smaller the change in that rating after a certain result occurs (in an upcoming single game) the higher the probability of that event occur.

The structure of this paper is follows. After presenting the classical approaches ("Betting Odds" and "The Bradley-Terry Model"), our new model is introduced. Then in Sec. 3 we present our preliminary experimental results, and finally in Sec. 4 we conclude and discuss some possible research directions.

2. MODELS

Let $V = (1, \ldots, n)$ be the set of n *teams* (or players) and let R be the number of *game days* in a competition among the teams in V. A *rating* is a function $\phi^r : V \to \mathbb{R}^n$ that assigns a score to each team after each game day r ($r = 1, \ldots, R$). This is considered as the quantitative "strength" of the teams. A *ranking* $\sigma^r : V \to V$, after game day r, is an ordering of the teams that is simply obtained by sorting the teams according to the rating ϕ^r. Using the game results data set, one can define a directed multigraph (i.e. a graph where multiple links are allowed), where nodes represent teams, while links between them represent outcomes of games they played. The links are directed and each of them is going from the loser team to the winning team. If ties are also considered they can be represented by two directed links with opposite directions and half weight. An edge weighting can be naturally considered if the final scores of the games are given

2.1 Betting Odds

Bookmakers determine *betting odds* for the games according to their expectations of outcome probabilities. Here we deal with fixed odds, means that they do not vary over time depending on the betting volumes. These "fixed-odds" represent the predictions of bookmakers [21]. The meaning of the betting odds for an upcoming game is the following: Assume that the betting odds between team i and team j are odds(i) and odds(j), respectively. It means that if one bets \$1 to i's win and it comes out, he wins odds(i) dollars, while if j wins, then the bettor loses his \$1. We can calculate the probabilities of the respective events as

$$\Pr(i \text{ beats } j) = \frac{1/\text{odds}(i)}{1/\text{odds}(i) + 1/\text{odds}(j)}$$

and

$$\Pr(j \text{ beats } i) = \frac{1/\text{odds}(j)}{1/\text{odds}(i) + 1/\text{odds}(j)}.$$

We should note here that odds provided by betting agencies do not represent the true chances (as imagined by the bookmaker) that the event will or will not occur, but are the amount that the bookmaker will pay out on a winning bet. The odds include a profit margin meaning that the payout to a successful bettor is less than that represented by the true chance of the event occurring. This means mathematically that $1/\text{odds}(i) + 1/\text{odds}(j)$ is more than one. This profit expected by the agency is known as the "over-round on the book".

2.2 The Bradley-Terry Model

The *Bradley-Terry model* [3] is a widely-used method to assign probabilities to the possible outcomes when a set of n individuals are repeatedly compared with each other in pairs. For two elements i and j, the probability that i beats j defined as

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \pi_j},$$

where $\pi_i > 0$ is a parameter associated to each individual $i = 1, \ldots, n$, representing the overall skill, or "intrinsic strength" of it. Equivalently, π_i/π_j represents the odds in favor i beats j, therefore this is a "proportional-odds model". Suppose that i and j played N_{ij} games against each other with i winning W_{ij} of them, and all games are considered to be independent. The likelihood is given by

$$L(\pi_i, \ldots, \pi_n) = \prod_{i<j} \left[\frac{\pi_i}{\pi_i + \pi_j}\right]^{W_{ij}} \left[\frac{\pi_j}{\pi_i + \pi_j}\right]^{N_{ij} - W_{ij}}.$$

Then the log-likelihood is

$$\begin{aligned}\ell(\pi_i, \ldots, \pi_n) &= \sum_{1 \le i \ne j \le n} \left[W_{ij} \log \pi_i - W_{ij} \log(\pi_i - \pi_j)\right] \\ &= \sum_{i=1}^{n} W_{ij} \log \pi_i - \sum_{1 \le i < j \le n} N_{ij} \log(\pi_i + \pi_j)\end{aligned}$$

which need to be maximized.

One possible derivation of the model assumes team i produces an unobserved score S_i, no matter which is the opposing team, with the cumulative distribution function

$$S_i \sim F_i(s) = \exp[-e^{-(s - \log \pi_i)}].$$

It follows that distribution of the difference $S_i - S_j$ follows a logistic distribution function

$$S_i - S_j \sim F_{ij}(s) = \frac{1}{1 + e^{-(s - (\log \pi_i - \log \pi_j))}},$$

which implies that

$$\begin{aligned}\Pr(S_i > S_j) = \Pr(S_i - S_j > 0) &= 1 - \frac{1}{1 + e^{\log \pi_i - \log \pi_j}} \\ &= \frac{\pi_i}{\pi_i + \pi_j}.\end{aligned}$$

Extension with Home advantage and Tie. A natural extension of the Bradley-Terry model with "home-field advantage", according to [1], say, is to calculate the probabilities as

$$\Pr(i \text{ beats } j) = \begin{cases} \frac{\theta \pi_i}{\theta \pi_i + \pi_j}, & \text{if } i \text{ is at home} \\ \frac{\pi_i}{\pi_i + \theta \pi_j}, & \text{if } j \text{ is at home} \end{cases}$$

where $\theta > 0$ measures the strength of the home-field advantage (or disadvantage). Considering also a tie as a possible

final result of a game, the following calculations, proposed in [22], can be used :

$$\Pr(i \text{ beats } j) = \frac{\pi_i}{\pi_i + \alpha\pi_j},$$

$$\Pr(i \text{ ties } j) = \frac{(\alpha^2 - 1)\pi_i\pi_j}{(\pi_i + \alpha\pi_j)(\alpha\pi_i + \pi_j)}$$

where $\alpha > 1$. Combining them is straightforward. In our experiments, we used the Matlab implementations found at `http://www.stats.ox.ac.uk/~caron/code/bayesbt/` using the *expectation maximization* algorithm, described in detail in [7].

2.3 Rating-based Model with Learning

Our new model is designed as follows. We will use the term "game day" in each case when at least one match is played on the given day. For any game day in which we make a forecast, we consider the results matrix that contains all the results of the previous $T = 40$ game days. For the 40 game days time window, the entries of the results matrix S are defined as $S_{ij} = \#\{$scores team home-i achieved against team away-j$\}$. To take into account the home-field effect, for each team i we distinguish team home-i and team away-i. Thus, we define a $2n \times 2n$ results matrix, which, in fact, describes a bipartite graph where each team appears both in the home team side and the away team side of the graph. For rating the teams, a time-dependent PageRank method is used. The PageRank scores are calculated according the time-dependent PageRank equation

$$\phi = \mathbf{\Pi} = \frac{\lambda}{N}[I - (1-\lambda)S_{mod}^t(\mathbf{1}\mathbb{1}^t)^{-1}]^{-1}\mathbb{1}, \qquad (1)$$

defined in [19]. The damping factor is $\lambda = 0.1$, while we may multiply each entry of S with the exponential function 0.98^α to consider time-dependency and obtaining S_{mod}, where α denotes the number of game days elapsed since a given result occurred (and stored in S). Note, that a home team and an away team PageRank values are calculated for each team. We would like to establish a connection between team home-i and team away-i using the assumption that home-i is not weaker than away-i. In our implementation we assumed that home-i had a win $2:1$ against away-i to give a positive bias for home-i at the beginning. In our experiments this setup performed well, but it was not optimized precisely.

Using the above-defined results matrix S and the PageRank rating vector ϕ, we assign probabilities to the outcomes {home team win, tie, away team win} of an upcoming game in game day r between home-i and away-j as follows. Before the game day in which we make the forecast, let the calculated PageRank rating vector be $\phi_{40}^{r-1}(V)$. We use δ_{xy}^r to measure how the rating vector of the teams changes if the result of an upcoming game between teams i and j is $x : y$, where $x, y = 0, 1, \ldots$ are the scores achieved by team i and team j, respectively[1]. We define δ_{xy}^r as the Euclidean distance between $\phi_{40}^{r-1}(V)$ and $\phi_{40}^r(V)$ that is the rating vector for the new results matrix obtained by adding x to S_{ij} and y to $S_{n+j,i}$. In the results graph interpretation this simply means that an edge from node away-j to node home-i with weight x and an edge from node home-i to node away-j with weight y are added to the graph, respectively. Our assumption is that if an outcome $x : y$ has a high probability and it occurs, then it causes a small change in the PageRank vector; hence δ_{xy} will be small. To simplify the notations let $\{\delta_1, \ldots, \delta_m\}$ be the distance values obtained by considering different results $\{E_1, \ldots, E_m\}$ of the upcoming game between i and j. The goal now is to calculate the probability that a certain result occurs if $\{\delta_1, \ldots, \delta_m\}$ is given. To do this, we use the following simple statistics-based machine learning method. Let $f^+()$ be the probability density function of δ_i random variable where the event (game result) E_i occurred. In our implementation $E_i \in \{0:0, 1:0, 1:1, \ldots, 5:5\}$, assuming that the probability of other results equals 0. Similarly, let $f^-()$ be the probability density function of δ_i random variable in which case the event (game result) E_i did not occur. To approximate the $f^+()$ and $f^-()$ functions, for each game we use the training data set contains all results and related δ_i $(i = 1, \ldots, m)$ values of the preceding $T = 40$ game days of the considered game. In our experiments, the gamma distribution (and its density function) turned out to be a fairly good approximate for $f^+(\delta)$ and $f^-(\delta)$.

Assuming that $\delta_1, \ldots, \delta_m$ are independent, using the Bayes theorem and the law of total probability, we can calculate that

$$\Pr(E_i|\{\delta_1, \ldots, \delta_m\}) = \frac{f^+(\delta_i)\prod_{k\neq i} f^-(\delta_k)}{\sum_\ell f^+(\delta_\ell)\prod_{k\neq l} f^-(\delta_\ell)}.$$

We should note here that in this way we assign probabilities to concrete game final results, which is another novelty of our model. Then, for the upcoming game between i and j, the outcome probability of the event "i beats j" is calculated as

$$\Pr(i \text{ beats } j) = \sum_{\substack{k:\ E_k \text{ encodes a result} \\ \text{of team-i win}}} \Pr(E_k|\{\delta_1, \ldots, \delta_m\}),$$

where we sum over those E_k results for which i beats j (i.e. 1:0, 2:0, 2:1, 3:0, 3:1, etc.). The probabilities $\Pr(i \text{ ties } j)$ and $\Pr(j \text{ beats } i)$ can be calculated in a similar way.

3. EXPERIMENTAL RESULTS

To measure the accuracy of the forecasting we calculate the mean squared error, which is often called *Brier scoring rule* in the forecasting literature [4]. The Brier score measures the mean squared difference between the predicted probability assigned to the possible outcomes for event E and the actual outcome o_E. Suppose that for a single game g, between i and j, the forecast is $\mathbf{p}^g = (p_w^g, p_t^g, p_l^g)$ contains the probabilities of i wins, the game is a tie and i loses, respectively. Let the actual outcome of the game be $\mathbf{o}^g = (o_w^g, o_t^g, o_l^g)$, where exactly one element is 1, the other two are 0. Noting that the number of games played (and predicted) is N, BS is defined as

$$BS = \frac{1}{N}\sum_{g=1}^{N} ||\mathbf{p}^g - \mathbf{o}^g||_2^2$$
$$= \frac{1}{N}\sum_{g=1}^{N}[(p_w^g - o_w^g)^2 + (p_t^g - o_t^g)^2 + (p_l^g - o_l^g)^2].$$

[1]We should note here that if the result is $0:0$, then $x = y = 1/2$ is used.

The best score achievable is 0. In the case of three possible outcomes (win, lost, tie) we can easily see that the forecast $\mathbf{p}^g = (1/3, 1/3, 1/3)$ (for each game g and any N) gives accuracy $BS = 2/3 = 0.666$. We consider this value as a worst-case benchmark. One question of our investigation is that how better BS values can be achieved using our method, and how close we can get to the betting agencies' fairly good predictions.

The data set we used contained all final results of given seasons of some football leagues, listed in the first two column of Table 1. We tested our method as it was described in Sec. 2.3. We start predicting games starting from the 41th game day; for each single game predictions are made using the results of the previous 40 game day before that game. The Brier scores were calculated using all predictions we made. Our initial results are summarized in Table 1. To calculate the betting odds probabilities we used the betting odds provided by bet365 bookmaker available at `http://www.football-data.co.uk/`. We could see that these predictions gave the best accuracy score (BS) in each case. We highlighted the values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01. Although we can see that slightly more than half of the cases the Bradley-Terry model gives a better accuracy, the results are still promising considering the fact that the parameters of our method and the implementation are far from being optimized.

4. CONCLUSIONS

We presented a new model for probabilistic forecasting in sports, based on rating methods, that simply use the historical game results data of the given sport competition. We provided a forward-looking type graph based approach. The assumption of our model is that the rating of the teams after a game day is correctly reflects their current relative performance. We consider that the smaller the changing in the rating vector after a certain result occurs in an upcoming single game, the higher the probability that this event will occur. Performing experiments on results data sets of European football championships, we observed that this model performed well in general in terms of predictive accuracy. However, we should note here, that parameter fine tuning and optimizing certain parts of our implementation are tasks of future work.

We emphasize, that our methodology can be also useful to compare different rating methods by measuring that which one reflects better the actual strength (rating) of the teams according to our interpretation. Finally we should add that the model is general and may be used to investigate such graph processes where the number of nodes is fixed and edges are changing over time; moreover it also has a potential to link prediction.

5. ACKNOWLEDGMENTS

This work was partially supported by the National Research, Development and Innovation Office - NKFIH, SNN117879.

Miklós Krész acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program.

6. REFERENCES

[1] A. Agresti. *Categorical data analysis*. John Wiley & Sons, New York, 1996.

[2] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2):187–202, 2013.

[3] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952.

[4] G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[5] K. Butler and J. T. Whelan. The existence of maximum likelihood estimates in the Bradley-Terry model and its extensions. *arXiv preprint math/0412232*, 2004.

[6] T. Callaghan, P. J. Mucha, and M. A. Porter. Random walker ranking for NCAA division IA football. *American Mathematical Monthly*, 114(9):761–777, 2007.

[7] F. Caron and A. Doucet. Efficient bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

[8] A. C. Constantinou, N. E. Fenton, and M. Neil. Pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322–339, 2012.

[9] D. Delen, D. Cogdell, and N. Kasap. A comparative analysis of data mining methods in predicting NCAA bowl outcomes. *International Journal of Forecasting*, 28(2):543–552, 2012.

[10] M. J. Dixon and P. F. Pope. The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20(4):697–711, 2004.

[11] D. Forrest, J. Goddard, and R. Simmons. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3):551–564, 2005.

[12] R. Gill and J. Keating. Assessing methods for college football rankings. *Journal of Quantitative Analysis in Sports*, 5(2), 2009.

[13] J. Goddard and I. Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.

[14] A. Joseph, N. E. Fenton, and M. Neil. Predicting football results using bayesian nets and other machine learning techniques. *Knowledge-Based Systems*, 19(7):544–553, 2006.

[15] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.

[16] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[17] J. Lasek, Z. Szlávik, and S. Bhulai. The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46, 2013.

[18] C. K. Leung and K. W. Joseph. Sports data mining: Predicting results for the college football games. *Procedia Computer Science*, 35:710–719, 2014.

[19] A. London, J. Németh, and T. Németh. Time-dependent network algorithm for ranking in sports. *Acta Cybernetica*, 21(3):495–506, 2014.

[20] M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.

Table 1: Accuracy results on football data sets. The values where the difference between the Bradley-Terry method and the PageRank method was higher than 0.01 are shown in bold.

League	Season	Betting odds error	Bradley-Terry error	PageRank method error
Premier League	2011/12	0.58934	0.60864	**0.59653**
	2012/13	0.56461	0.59744	**0.58166**
	2013/14	0.54191	**0.55572**	0.59406
	2014/15	0.55740	0.60126	0.60966
Bundesliga	2011/12	0.58945	0.59994	**0.59097**
	2012/13	0.57448	0.59794	**0.58622**
	2013/14	0.55724	**0.57803**	0.60125
	2014/15	0.57268	0.60349	0.60604
La Liga	2011/12	0.54598	**0.57837**	0.58736
	2012/13	0.56417	**0.58916**	0.60205
	2013/14	0.57908	**0.58016**	0.60473
	2014/15	0.52317	0.55888	0.56172

[21] P. F. Pope and D. A. Peel. Information, prices and efficiency in a fixed-odds betting market. *Economica*, pages 323–341, 1989.

[22] P. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

[23] J. A. Trono. Rating/ranking systems, post-season bowl games, and 'the spread'. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.

[24] C. Wang and M. L. Vandebroek. A model based ranking system for soccer teams. *Research report, available at SSRN 2273471*, 2013.

Indeks avtorjev / Author index

Konferenca / Conference

Uredila / Edited by

Odkrivanje znanja in podatkovna skladišča - SiKDD / Data Mining and Data Warehouses - SiKDD

Dunja Mladenić, Marko Grobelnik