Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek E

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume E

Izkopavanje znanja in podatkovna skladišča (SiKDD 2015)

Data Mining and Data Warehouses (SiKDD 2015)

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

5. oktober 2015 / October 5th, 2015
Ljubljana, Slovenia

Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek E

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume E

Izkopavanje znanja in podatkovna skladišča

Data Mining and Data Warehouses

Uredila / Edited by

Dunja Mladenič, Marko Grobelnik

5. oktober 2015 / October 5th, 2015
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, Mitja Lasič

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

Ljubljana, oktober 2015

CIP – Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

004.8(082)(0.034.2)

MEDNARODNA multikonferenca Informacijska družba (18 ; 2015 ; Ljubljana)
Izkopavanje znanja in podatkovna skladišča (SiKDD 2015) [Elektronski vir] : zbornik 18. mednarodne multikonference Informacijska družba – IS 2015, 5. oktober 2015, [Ljubljana, Slovenia] : zvezek E = Data mining and data warehouses (SiKDD 2015) : proceedings of the 18th International Multiconference Information Society – IS 2015, October 5th, 2015, Ljubljana, Slovenia : volume E / uredila, edited by Dunja Mladenić, Marko Grobelnik. – El. zbornik. – Ljubljana : Institut Jožef Stefan, 2015

Način dostopa (URL): http://is.ijs.si/zborniki/!%20E%20-%20Izkopavanje%20znanja%20-%20ZBORNIK.pdf

ISBN 978-961-264-086-6 (pdf)
1. Gl. stv. nasl. 2. Vzp. stv. nasl. 3. Dodat. nasl. 4. Mladenić, Dunja
4050939

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2015

Multikonferenca Informacijska družba (http://is.ijs.si) je z osemnajsto zaporedno prireditvijo osrednji srednjeevropski dogodek na področju informacijske družbe, računalništva in informatike. Letošnja prireditev traja tri tedne in poteka na Fakulteti za računalništvo in informatiko in Institutu »Jožef Stefan«.

Informacijska družba, znanje in umetna inteligenca se razvijajo čedalje hitreje. V vse več državah je dovoljena samostojna vožnja inteligentnih avtomobilov, na trgu je moč dobiti čedalje več pogosto prodajanih avtomobilov z avtonomnimi funkcijami kot »lane asist«. Čedalje več pokazateljev kaže, da prehajamo v naslednje civilizacijsko obdobje, hkrati pa so konflikti sodobne družbe čedalje težje razumljivi.

Letos smo v multikonferenco povezali dvanajst odličnih neodvisnih konferenc. Predstavljenih bo okoli 300 referatov v okviru samostojnih konferenc in delavnic, prireditev bodo spremljale okrogle mize in razprave ter posebni dogodki kot svečana podelitev nagrad. Referati so objavljeni v zbornikih multikonference, izbrani prispevki pa bodo izšli tudi v posebnih številkah dveh znanstvenih revij, od katerih je ena Informatica, ki se ponaša z 38-letno tradicijo odlične znanstvene revije.

Multikonferenco Informacijska družba 2015 sestavljajo naslednje samostojne konference:

- Inteligentni sistemi
- Kognitivna znanost
- Izkopavanje znanja in podatkovna skladišča
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Vzgoja in izobraževanje v informacijski družbi
- Soočanje z demografskimi izzivi
- Kognitonika
- Delavnica »SPS EM-zdravje«
- Delavnica »Pametna mesta in skupnosti kot razvojna priložnost Slovenije«
- Druga študentska konferenca s področja računalništva in informatike za doktorske študente
- Druga študentska konferenca s področja računalništva in informatike za vse študente
- Osma mednarodna konferenca o informatiki v šolah: razmere, evolucija in perspektiva.

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija, SLAIS in Inženirska akademija Slovenije. V imenu organizatorjev konference se zahvaljujemo združenjem in inštitucijam, 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 2015 bomo tretjič podelili nagrado za življenjske dosežke v čast Donalda Michija in Alana Turinga. Nagrado Michie-Turing za izjemen življenjski prispevek k razvoju in promociji informacijske družbe bo prejel prof. dr. Jurij Tasič. Priznanje za dosežek leta je pripadlo dr. Domnu Mungosu. Že petič podeljujemo nagradi »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono je dobilo počasno uvajanje informatizacije v slovensko pravosodje, jagodo pa spletna aplikacija »Supervizor«. Čestitke nagrajencem!

Niko Zimic, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2015

In its 18th 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 2015 it is extended over three weeks located at Faculty of computer science and informatics and at the Institute "Jožef Stefan".

The pace of progress of information society, knowledge and artificial intelligence is speeding up. Several countries allow autonomous cars in regular use, major car companies sell cars with lane assist and other intelligent functions. It seems that humanity is approaching another civilization stage. At the same time, society conflicts are growing in numbers and length.

The Multiconference is running in parallel sessions with 300 presentations of scientific papers at twelve conferences, round tables, workshops and award ceremonies. The papers are published in the conference proceedings, and in special issues of two journals. One of them is Informatica with its 38 years of tradition in excellent research publications.

The Information Society 2015 Multiconference consists of the following conferences:

- Intelligent Systems
- Cognitive Science
- Data Mining and Data Warehouses
- Collaboration, Software and Services in Information Society
- Education in Information Society
- Facing Demographic Challenges
- Cognitonics
- SPS EM-Health Workshop
- Workshop »Smart Cities and Communities as a Development Opportunity for Slovenia«
- 2nd Computer Science Student Conference, PhD Students
- 2nd Computer Science Student Conference, Students
- 8th International Conference on Informatics in Schools: Situation, Evolution, and Perspective.

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, SLAIS and the Slovenian Engineering Academy. In the name of the conference organizers we thank all societies and institutions, all participants for their valuable contribution and their interest in this event, and the reviewers for their thorough reviews.

For 2013 and further, the award for life-long outstanding contributions will be delivered in memory of Donald Michie and Alan Turing. The life-long outstanding contribution to development and promotion of information society in our country is awarded to Dr. Jurij Tasič. In addition, a reward for current achievements was pronounced to Dr. Domnu Mungosu. The information strawberry is pronounced to the web application "Supervizor, while the information lemon goes to lack of informatization in the national judicial system. Congratulations!

Niko Zimic, 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, Chicago, USA
Toby Walsh, Australia

Organizing Committee

Matjaž Gams, chair
Mitja Luštrek
Lana Zemljak
Vesna Koricki-Špetič
Mitja Lasič
Robert Blatnik
Mario Konecki
Vedrana Vidulin

Programme Committee

Nikolaj Zimic, chair
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
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

Izkopavanje znanja in podatkovna skladišča / Data Mining and Data Warehouses (SiKDD) 1
PREDGOVOR / FOREWORD 3
The Pursuit of Journalistic News Values Through Text Mining Techniques/ Belyaeva Evgenia, Košmerlj Aljaž, Mladenić Dunja 5
Relating Biological and Clinical Features of Alzheimer's Patients with Predictive Clustering Trees/ Breskvar Martin, Ženko Bernard, Džeroski Sašo 9
Ingredients Matching in Bakery Products/ Eftimov Tome, Koroušić Seljak Barbara 13
Mining Scientific Literature About Ageing to Support Better Understanding and Treatment of Degenerative Diseases/ Gubiani Donatella, Petrič Ingrid, Fabbretti Elsa, Urbančič Tanja 17
Modelling in Energy Related Scenarios/ Kenda Klemen, Škrjanc Maja, Borštnik Andrej 21
Forecasting Sales Based on Card Transactions Data/ Moraru Alexandra, Mladenić Dunja 25
A Topological Data Analysis Approach to the Epidemiology of Influenza/ Pita Costa Joao, Škraba Primož 29
Event Detection in Twitter with an Event Knowledge Base/ Rei Luis, Grobelnik Marko, Mladenić Dunja 33
A Multi-Scale Methodology for Explaining Data Streams/ Stopar Luka 37
Inverted Heuristics in Subgroup Discovery/ Valmarska Anita, Robnik Šikonja Marko, Lavrač Nada 41
Indexing of Large N-Gram Collection/ Zajec Patrik, Grobelnik Marko 45

Indeks avtorjev / Author index 49

Zbornik 18. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2015

Zvezek E

Proceedings of the 18th International Multiconference

INFORMATION SOCIETY – IS 2015

Volume E

Izkopavanje znanja in podatkovna skladišča

Data Mining and Data Warehouses

Uredila / Edited by

Dunja Mladenič, Marko Grobelnik

5. oktober 2015 / October 5th, 2015
Ljubljana, Slovenia

Preface / Predgovor

Data Mining and Data Warehouses (SiKDD 2015)

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.

Odkrivanje znanja in podatkovna skladišča (SiKDD 2015)

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 (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 zakonitosti v podatkih: pristope, orodja, probleme in rešitve.

Editors and Program Chairs / Urednika

- Dunja Mladenić
- Marko Grobelnik

The Pursuit of Journalistic News Values through Text Mining Techniques

Evgenia Belyaeva, Aljaž Košmerlj, Dunja Mladenić

Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
Email addresses: firstname.lastname@ijs.si

Abstract

The paper addresses a problem of pursuit of journalistic news values, more specifically *frequency, threshold and proximity* through various text mining methods is presented. We illustrate how text mining can assists journalistic work by finding ideological, often orthodox news values of different international publishers across the world news that contribute to ubiquitous news bias. Our experiments on selected publishers and on news about *Apple's launch of new iPhone 6* and *Apple Watch* confirm that journalists still follow some of the well known journalistic values.

Keywords: News values, newsworthiness, text mining, Apple

1. INTRODUCTION

Every news outlet has a different agenda for selecting which stories to cover and publish. Mass media have traditionally relied on the so-called news values to evaluate the newsworthiness of a story i.e. what to publish and what to leave out, introduced firstly by Galtung and Ruge (1). News values are certain guidelines to follow in producing a news story, so-called ideological factors in understanding decisions of journalists (2). The more news values are present in a story, the more likely that you will see the story featured in different mass media. It is a known and well-studied problem that old system of news values contributes to the ubiquitous media bias.

In the last years there has been a growing interest to work on the intersection of social and computer sciences (3). Text mining is emerging as a vital tool for social sciences and the trend will most likely increase (4). Due to the abundance of news information and with the advances in text mining, it is now possible to help journalists to process information in every day job and at the same time prove old media theories and discover old biased patterns in news across the world.

Some research has been already done to detecting news bias (5, 6), but very little attention paid to automatic detection of news values (7). We argue that in order to understand and detect automatically news bias, it is first important to explain and try to detect news values.

We make a first attempt to automate the detection of several news values by applying various text mining techniques from selected publishers and when reporting about *Apple Corporation* since it has a great impact on our lives and as any technology it is newsworthy by default. Our goal is to distinguish if the theory of newsworthiness by Galtung and Ruge is a valid approach to predict news selection values and to see some interesting recurring patterns in the news.

2. DATA DESCRIPTION

News articles analysed in this paper were first aggregated and processed by the Event Registry[1] - global media monitoring service that collects and processes articles from more than 100.000 news sources globally in more than 10 languages (8).

We extracted news about the Apple Corporation (***iPhone 6*** and ***Watch*** launch) from 16 selected online outlets during the period of 01.09.2014 – 21.10.2014. The time range corresponds to the announcements of the launch of the two above-mentioned products and the start of sales. The sources under our analysis correspond to the most influential daily news websites, easily accessible, widely read in the following three languages: English **(EN),** German **(DE)** and Spanish **(ES).**

The Publisher	Total Nr. Events	Total Nr. Articles on Apple	Headquarters
The Next Web	1064	1670	Amsterdam
Gizmodo	2007	3911	New York
The Guardian	14299	19997	London
BBC	15582	23852	London
USA Today	7692	13629	Tysons Corner
Wall Street J.	7197	18837	New York
Heise.de	4194	2190	Hannover
Chip online.de	907	1212	Munich
Stern	4194	10092	Hamburg
Die Zeit	3722	5600	Hamburg
Die Welt	14683	30359	Berlin
Der Spiegel	2261	2759	Hamburg
El Mundo	6707	8705	Madrid
ABC.es	7431	10388	Madrid
El Pais	686	979	Madrid
El Dia	6700	12752	Barcelona

Table 1. Publishers and Totals of Events/Article on Apple

[1] http://eventregistry.org

The Table 1 summarizes the total number of events and the total number of articles reporting on Apple collected and analysed per publisher during the above-mentioned period including the information on the headquarters of each publisher.

Important to note that the websites were selected to cover different geographical places (EU plus USA) in order to identify one of the news values, i.e. proximity – geographical or cultural proximity of the event to the source. The core available piece of information for each article for our experiment included the date, the location of the event and the location of the publishers' headquarters as well as size of events (i.e. number of articles about them).

3. MINING NEWS VALUES

Galtung and Ruge originally came up with a taxonomy of 13 news values (1), but due to the space limitations, the goal of this work is to identify the first three news values: *frequency, threshold* and *proximity*. Frequency and threshold are impact criteria, calculated through the number of articles per publisher (frequency) and the number of articles per events (threshold), whereas, the proximity criterion is rather about audience identification and geographical distance.

The following Table outlines Galtung and Ruge's theory of news selection and its news values (1).

News Values	Short Description
Frequency	***Time span of an event***
Threshold	***The size of an event***
Proximity	***Geographical closeness***
Unambiguity	Clarity of the meaning
Meaningfulness	Great value to the audience
Consonance	Conventional expectations
Continuity	Continuous over time
Unexpectedness	Unplanned/Unexpected
Composition	Other pieces of info
Reference to elite nations	Relate to famous nations
Reference to elite people	Relate to famous people
Negativity	Bad news, conflict oriented
Personalisation	Action of individuals

Table 2. News Values by Galtung and Ruge

3.1 Frequency

Frequency as news value refers to the time-span of an event (1). Since Apple has become a new religion of the 21st century, it is newsworthy by default and news about it exists in most outlets around the world. In the experiment, we have analysed the frequency of all articles from the selected publishers mentioning *iPhone* and *Apple Watch* respectively. We are interested in finding trends or particular patterns among publishers during the selected period of time. The Figure 1 summarizes the time distribution (i.e. frequency value) of mentions related to iPhone 6.

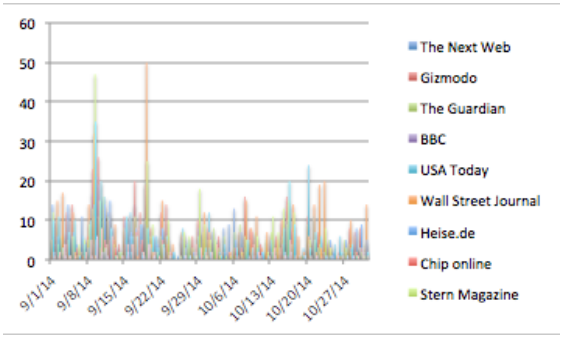


Figure 1. iPhone 6 Frequency Distribution

The frequency measurement experiment indicates that there are two sudden busts in frequency among certain publishers, one corresponding to the announcement of launch *Watch* on the 9th of September 2014 and the second one being the announcement of the *iPhone 6* release on the 19th of September 2014. Both announcements received a much bigger coverage (especially, the following publishers Wall Street Journal, Stern Magazine and die Welt) in respect to the actual start of sales of the products at the end of October.

The frequency distribution of new *Apple Watch* has a similar to *iPhone 6* trend, having, however, less coverage per publisher, per day. The following Fig. 2 outlines the frequency of *Watch* coverage among the selected publishers during 01.09 – 31.10.2014. The two bursts are also visible in the coverage of *Watch*. It can be explained by journalistic standards to include background information to a story i.e. writing about *Watch* a journalist is likely to mention *iPhone* or simple the Apple Corporation.

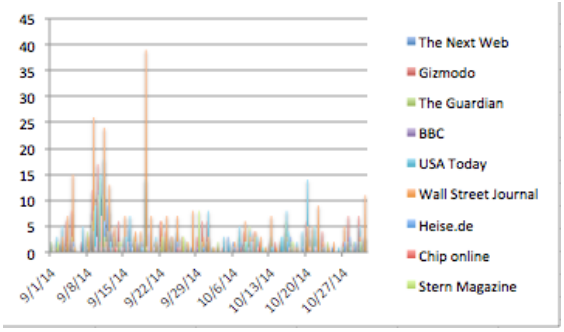


Figure 2. Watch Frequency Distribution

We also measured frequency between specific tech publishers to other international publishers. Our assumption that tech publishers do publish more news with higher frequency on Apple was not confirmed as also seen from Table 1. This partially could be explained by the small size of journalistic teams working for tech publishers in comparison to big media corporations with journalists all over the world like BBC or USA Today.

3.2 Threshold

The threshold criterion often refers to the impact of an event and its effect on the readers i.e. a size needed for an event to become news, e.g. thousands of people buying new *iPhone 6* and not just one person buying it in a small local store.

It is indeed difficult to measure something that should have a larger affect on the readers. We understand that events can meet this threshold value either by being large in absolute terms or by having a higher frequency or an increase in reporting of a topic. In this experiment, we decided to look at the size of events among the selected publishers without limiting our search to reports about Apple in order to take in more data. The main reason we limit ourselves to Apple related stories in frequency analysis is that we can manually show that remarkable and infrequent events like new product launches draw more media attention.

Event is understood as a group of articles that are clustered to report on the same issues in the world (9). Our assumption is that a single article might not be always very informative, but a group of articles on a certain issue, which is picked up by more publishers can form a part of a bigger story with more impact on the readers and match the threshold value. Note that frequency and threshold values are both impact criteria, threshold is more about the size of an event, whereas frequency should be also understood as events unfolding within the production cycle of a news media and will be reported on repeatedly.

Therefor, for the threshold analysis we aim at capturing the size of clusters (number of articles of all publishers in event clusters) and assume to witness a greater number of articles that form an event. To note that news articles are first aggregated by the JSI Newsfeed[2] – real-time stream of articles from more than 1900 RSS-enabled websites in several major world languages, then we process the articles by a linguistic and semantic analysis pipeline that provides semantic annotations. The semantic annotation tool developed within XLike project comprises three main elements: *named entity recognition* based on corresponding Wikipedia pages, *Wikipedia Miner Wikifier* – detecting similar phrases in any document of the same language as Wikipedia articles and cross-lingual semantic analysis that links articles by topics (10).

The data in the following scatterplot shows the average event size per publisher during the same period of time and confirms our hypothesis: the higher the threshold (number of articles per event), the greater the impact of a publisher (i.e. The Guardian, BBC), the more intense and more frequent the coverage about an event is.

If an article is written by an influential publisher other bigger and smaller publishers will most likely pick it up and eventually it will form an event. Interestingly, Spanish and German publishers have a smaller average event size, which could be explained as those publishers are more interested in local events or events within their countries.

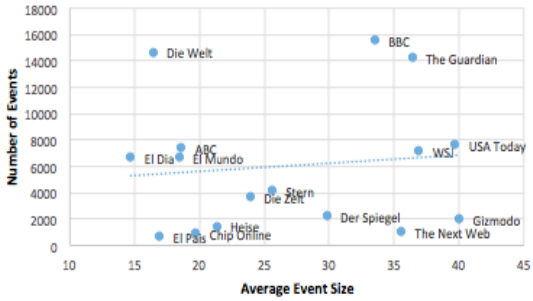


Figure 3. Threshold analysis per publisher

3.3 Proximity

The Proximity value corresponds to physical i.e. geographical or often cultural (in terms of religion or language) closeness of a news story to a media publisher (1). Proximity helps readers to relate to a story on a more personal level. It can change over time and is open to subjective interpretations. However, proximity might also mean an emotional (fear, happiness, pride etc.) trajectory in the audience's eyes, regardless of where it takes place (11).

Our assumption when measuring this journalistic value is that the closer the geographical location of the story to the news publisher is, the more frequent and the more intense (higher threshold) the coverage is.

Event location detection is done automatically in Event Registry in the following way: we first try to identify a dateline in the article (a piece of text at the beginning of every article) that names a location; we assume that this is the location of the event. When the dateline does not appear in the article, we check the event Registry to use the event's location. A classification algorithm that considers all articles belonging to the same event determines it. In some cases, the Event Registry does not determine the location; we try to avoid such cases in our analysis.

The headquarters of each publisher was manually searched for on the official websites of the selected publishers. We did not limit our search to the stories reporting about Apple, since we assume to see some recurring proximity patterns of the selected publishers in spite of a story kind.

Since our system was not able to automatically identify location for all events, we use only a sub-selection of our data for each publisher for which we compute the distance in kilometres and calculate how many of them report from the same country and same city where the publisher is. The following Table 4 outlines the proximity experiment results: Total number of sub-Selection of events where Country/City were detected and total

[2] http://newsfeed.ijs.si

number of events where publishers reported either on the same country or the same city where a publisher has headquarters.

Publisher	Total Nr. Country Sub-Selection	Same country	Total Nr. City Sub-Select.	Same city
The Next Web	178	1	174	1
Gizmodo	***371***	***204***	370	15
Guardian	6563	2261	6510	462
BBC	7105	2909	7039	438
USA Today	***4299***	***2842***	4291	0
WSJ	3091	1194	1074	122
Heise	***586***	***262***	585	5
Chip	211	71	211	1
Stern	***2704***	***1197***	2701	90
Die Zeit	***2505***	***1103***	2504	95
Die Welt	***9185***	***5340***	9182	1248
Der Spiegel	1592	630	1590	55
El Mundo	***4077***	***2269***	4076	861
ABC	***4493***	***2372***	4491	879
El Pais	337	46	337	7
El Dia	***3789***	***2399***	3785	154

Table 4. Geographical proximity analysis per publisher

It has been found that the coverage of most publishers is not local, they do not report on the events close to the headquarters; it can be explained by the fact that the selected publishers are not local publishers and are considered to be the most read outlets in each country and some even in the world. Interesting to note that proximity value was not confirmed for, for example, The Next Web – technology oriented website with headquarters in Amsterdam, Netherlands, reported only once on events from Amsterdam and the Netherlands. Whereas, some selected publishers, mainly Spanish and German outlets, (*in italics*) dedicate more or less half of their attention to the news from the same country, which confirms relatively strong proximity news value. Not surprisingly, The Guardian, BBC and the World Street Journal do not support journalistic proximity value since their geographical scope is scattered around the world.

4. DISCUSSIONS AND FUTURE WORK

We made an initial attempt to automate detection of journalistic values, in particular, *frequency* in the context of Apple news, *threshold* and *proximity* in the context of selected publishers. We believe that using text mining methods is an essential step of interaction between social and computer sciences approaches. This hybrid approach will not only help journalists in their everyday work, but it will also potentially help to identify various ideological patterns or news bias of various global publishers.

Future work will include developing our framework, which will automate the process of assessing newsworthiness of all 12 news values applied to different languages, as well as to different domains like conflicts, natural disasters, political crises etc. By detecting news values through text mining we also aim at confirming still existing ideological patterns, i.e. news bias of different publishers. Research designed more specifically and comprising automation of all values could provide more answers to the problems of out-dated and orthodox news values that keep on contributing to the news bias. To our knowledge, there are no automated systems to compare our approach with, thus, in the future we also plan on conducting several evaluations including manual evaluation to verify our results.

Acknowledgments

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under XLike (ICT-STREP-288342) and XLime (FP7-ICT-611346).

REFERENCES

(1) Galtung, J., Ruge, M. (1965): Structuring and Selecting News. In: Journal of International Piece Studies, In: Journal of International Piece Research, 2 (1), pp. 70-71.

(2) Cotter, C. (2010): News Talk. Investigating the Language of Journalism. Cambridge: Cambridge University Press.

(3) Greening, T. (Ed.) (2000): Computer Science Education in the 21st Century. Springer New York.

(4) Ampofo, L., Collister, S. et al. (2015): Text Mining and social Media: When Quantative Meets Qualitative, and software meets human. In: Halfpenny, P. and Procter, R. (eds.) Innovations in Digital Research Methods. London: Sage.

(5) Ali, O., Flaounas, I., De Bie, T., et al. (2010): "Automating News Content Analysis: An Application to Gender Bias and Readability" JMLR: Workshop and conference Proceedings 11.

(6) Flaounas, I., Turchi, M., et al. (2010) "The Structure of the EU Mediasphere" PLoS ONE. Vol.5, Issue 12.

(7) De Nies, T., D'heer, E., et al. (2012): Bringing Newsworthiness into the 21st Century. In: Proceedings Web of Linked entities Workshop, ISWC. Boston, pp. 106-117.

(8) Leban, G., Fortuna B., Brank J., Grobelnik M., Event Registry – Learning About World Events From News, WWW 2014, pp. 107-111.

(9) Leban, G., Košmerlj, A., Belyaeva, E. et al. (2014): News reporting bias detection prototype. XLike Deliverable D5.3.1.

(10) Carreras, X., Padró, L., et al. (2014): XLike project language analysis services. In: Proceedings of EACL'14: demos, pp. 9-12.

(11) Schults, B. (2005): Broadcast News Producing. Sage Publications, London.

Relating Biological and Clinical Features of Alzheimer's Patients With Predictive Clustering Trees

Martin Breskvar[1,2] martin.breskvar@ijs.si
Bernard Ženko[1] bernard.zenko@ijs.si
Sašo Džeroski[1,2] saso.dzeroski@ijs.si

[1]Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[2]Jožef Stefan Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

ABSTRACT

This paper presents experiments with Predictive Clustering Trees that uncover several subpopulations of the Alzheimer's disease patients. Our experiments are based on previous research that identified the everyday cognition as one of the most important testing domains in the clinical diagnostic process for the Alzheimer's disease. We are investigating which biological features have a role in the progression of the disease by observing behavioral response of the patients and their study partners. Our dataset includes 342 male and 317 female patients from the ADNI database that are described with 243 clinical and biological attributes. The resulting clusters, described in terms of biological features, show behavioral and gender specific differences between clusters of patients with progressed disease. These findings suggest a possibility that the Alzheimer's disease is manifested through different biological pathways.

1. INTRODUCTION

Alzheimer's disease (AD) is a form of dementia, which represents a large portion of all dementias. It is a neurodegenerative disease affecting many aspects of the patients life, including physical, psychological and social wellbeing. This inevitably leads to severe decrease of life quality. Currently about 47.5 million people worldwide suffer from dementia,[1] and its incidence is expected to triple by the year 2050.

In order to diagnose AD with certainty, a histopathologic examination has to be conducted, which is the main reason why in practice AD diagnosis is mainly based on clinical criteria that can be subjective. Finding links between the clinical and biological characteristics of the disease is therefore an important research topic: its advancement could potentially improve the understanding of the disease pathophysiology and enable its detection at earlier stages.

In this work, we address the problem of finding possible connections between biological and clinical features of AD patients with the use of Predictive Clustering Trees (PCTs). Our goal is not to provide a model for diagnosing the disease, but rather to cluster patients into homogeneous groups that share biological features. This way we should be able to investigate the traits of the grouped patients in more detail. One of the most distinctive properties of PCTs is their ability to learn models for predicting structured or complex variables, e.g., vectors, time-series or hierarchies. By using PCTs, we were able to construct clusters homogeneous in respect of several clinical variables simultaneously and not just a single one as with standard decision trees. We use a dataset of Alzheimer's patients obtained from the ADNI database[2].

The remainder of the paper is structured as follows. Section 2 presents the dataset, methodology and the experimental design. Section 3 describes the results. Finally, in Section 4 we analyze the results and present our conclusions.

2. DATA AND METHODOLOGY

2.1 The Data

All data used comes from Alzheimer's Disease Neuroimaging Initiative (ADNI) database[2]. ADNI is an international observational study of healthy, cognitively normal elders, people with mild cognitive impairment (MCI) and people with Alzheimer's disease. It collects a wide range of clinical and biological data for each patient at multiple time points. We used the ADNIMERGE table, which is a joined dataset from multiple ADNI data collection domains.

The dataset includes information on 659 patients (342 male, 317 female). Each patient is described with 56 biological and 187 clinical attributes. Some numerical values have been transformed in order to make them more linear. Out of 243 attributes, 74 contain missing values.

Biological attributes include ABETA peptides, APOE4 genetic variations, intracerebral volume (ICV), results from many laboratory measurements like glucose and protein levels, red and white blood cell counts, MRI volumetric data, (Ventricles, Hippocampus, WholeBrain, Entorhinal gyrus,

[1]Source: World Health Organization (march 2015).

[2]The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. More information can be found at http://www.adni-info.org and http://adni.loni.usc.edu.

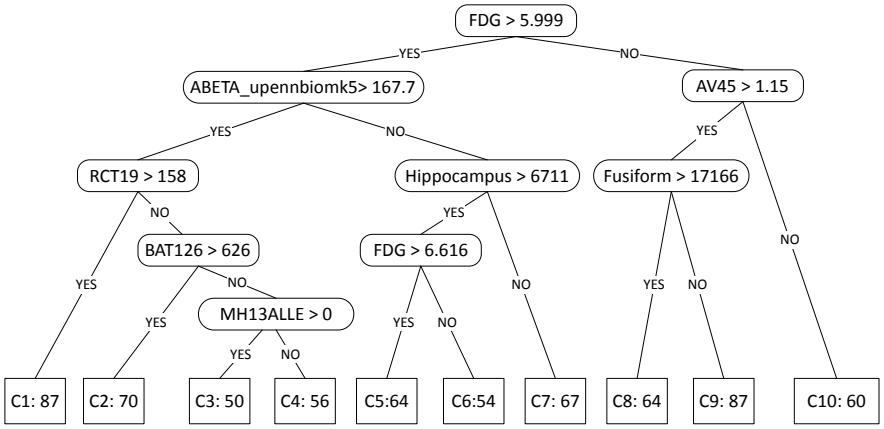


Figure 1: Predictive Clustering Tree, showing 10 clusters (cluster IDs and numbers of patients in each cluster).

Fusiform gyrus, Middle temporal gyrus), TAU and PTAU proteins, and PET imaging results (FDG-PET and AV45).

Clinical attributes include Alzheimer's Disease Assessment Scale (ADAS13), Mini Mental State Examination (MMSE), Rey Auditory Verbal Learning Test (RAVLT), which is divided into several different stages (immediate, learning, forgetting and percantage of forgetting), Functional Assessment Questionnaire (FAQ), Montreal Cognitive Assessment (MOCA) and Everyday Cognition, which consists of questions that are answered by patients themselves (ECogPt) and their study partners[3] (ECogSP). Again, this cognitive evaluation consists of several domains (Memory, Language, Organization, Planning, Visuospatial abilities, Divided attention and Total score). Also Neuropsychiatric Inventory Examination, Neurological Exam, Modified Hachinski Ischemia Scale, Geriatric Depression Scale, Baseline symptoms (nausea, vomiting, diarrhea, sweating, etc.), Clinical Dementia Rating (CDR), Medical History, patient gender and handedness have been included.

The diagnosis (DX) that has been given by the physician at the first examination is included in the data. The possible values for the DX attribute are Cognitively Normal (CN), Significant Memory Concern (SMC), Early Mild Cognitive Impairment (EMCI), Late Mild Cognitive Impairment (LMCI) and Alzheimer's Disease (AD). The diagnosis distribution is the following: CN=173, SMC=94, EMCI=148, LMCI=134, AD=110.

We are using only the baseline data (i.e., data gathered when patients enrolled in the ADNI study and have been examined and tested for the first time).

2.2 Experimental design

In our study we are especially interested in the everyday cognition of patients, therefore we will give a brief overview of the everyday cognition, as it is understood and evaluated within the ADNI database. Everyday cognition (ECog) is a questionnaire, that requires cooperation of both patient and his or her study partner. It assesses the patient's capability to perform normal, everyday tasks. Patients and their corresponding study partners must individually compare the patient's current activity levels and capabilities with levels from 10 years prior the examination. The domains of memory, language and executive functioning are assessed. Answers are evaluated on a 5 point scale: (1) no change or performing better, (2) occasionally performs worse, (3) consistently performs worse, (4) performs much worse, (5) does not know. According to Farias et. al.[4], everyday cognition shows promise as a tool for measuring general and domain-specific everyday functions in the elderly. We have decided to design our experiment on that assumption and we aim to connect existing biological and clinical features in order to observe differences of predicted values between clusters.

We have used Predictive Clustering Trees for the task of multi-target prediction. Our targets were all the ECog components and the diagnosis itself. The descriptive space was defined by all the laboratory measurements, neuropathology, medical history and gender. We have included medical history in the descriptive space because we wanted to observe whether pre-existing conditions such as alergies play a role in the disease progression. Additionally we included gender, because according to Barnes et. al.[1] gender specific differences do exist. We have pre-pruned our clustering tree with the constraint of minimum 50 examples per leaf.

2.3 Predictive Clustering Trees

The concept of predictive clustering was introduced in 1998 by H. Blockeel [2] and can be seen as a generalization of supervised and unsupervised learning. Even though predictive modeling and clustering are usually viewed as two separate tasks, they are connected by the methods that partition the instance space into subsets. We can also consider these methods to be clustering methods. An example of such methods are decision trees.

If we consider a decision tree in the predictive clustering paradigm, the tree is a hierarchy of clusters. We refer to those trees as predictive clustering trees (PCTs). An obvious benefit of PCTs is that they, in addition to predictions, also provide symbolic descriptions of the clusters. Each node in the clustering tree represents a cluster and has a

[3]Each patient must have a study partner, a person who is in frequent contact with the patient, provides information about the patient and is able to independently evaluate the patient's functioning.

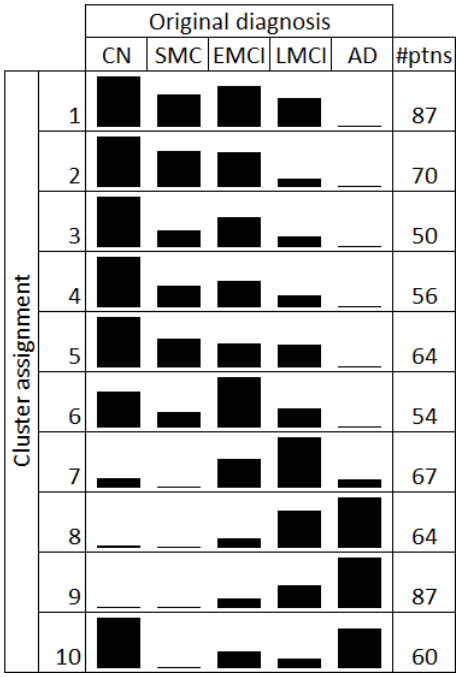


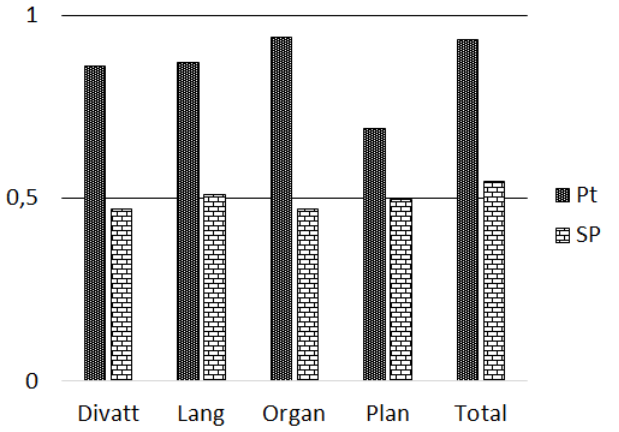
Figure 2: Normalized distribution of original diagnoses with respect to the clusters modeled by the PCT in Figure 1.

symbolic description (except for the root node) in the form of a conjunction of conditions on the path from the root node to the selected cluster node. In case of the PCT in Figure 1, the examples in the root node are split according to condition $FDG > 5.999$. Examples, whose value of the FDG attribute is greater than the value 5.999 will go to the left branch, the others to the right branch. On the next level of the clustering tree, nodes $AV45 > 1.15$ and $ABETA\_upennbiomk5 > 167.7$ are now split again iteratively until we reach leaf nodes $C1$. Examples in cluster $C1$, for example, are those that correspond to the condition: $FDG > 5.999$ & $ABETA\_upennbiomk5 > 167.7$ & $RCT19 > 158$. PCTs support multi-target predictions which means we can learn a model with respect to not only one target variable but many variables simultaneously. This gives us the tool needed to predict complex structures that can also be interconnected.

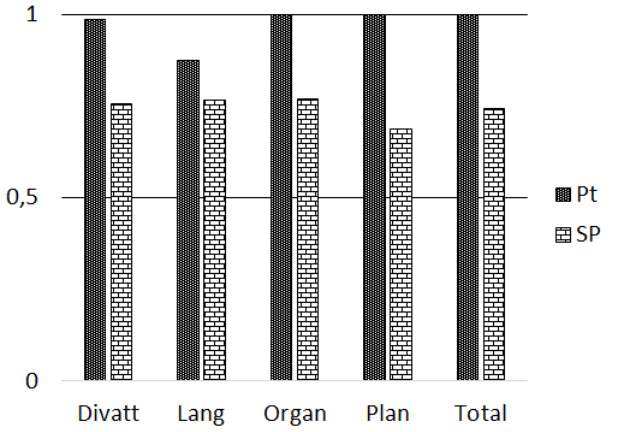
Several different predictive clustering methods [3, 5, 6, 7, 8, 9] are implemented in the software package CLUS (available at http://sourceforge.net/projects/clus/).

3. RESULTS

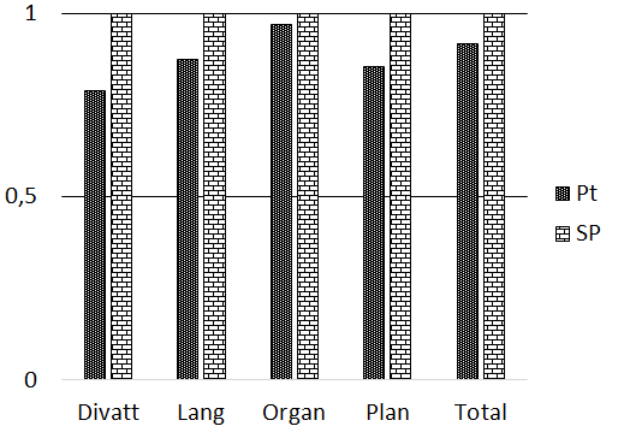
The result of our analysis is the PCT presented in Figure 1. We have investigated all ten clusters in the leaf nodes and Figure 2 shows relative distribution of original diagnoses (DX) in all the clusters. Clusters 1 to 6 are relatively diverse and we can state that the presence of Alzheimer's patients in these clusters is unlikely. With the exception of cluster 6, cognitively normal patients are dominant. Cluster 6 also contains patients in the early stage of the disease (EMCI) as well as some LMCI patients. We have focused our attention on clusters 7, 8 and 9 because they mainly consist of patients in the stages of late MCI or already developed AD.



(a) Patients evaluate their cognition as worse than 10 years ago. Study partners evaluate the same behavior as approximately half as bad.

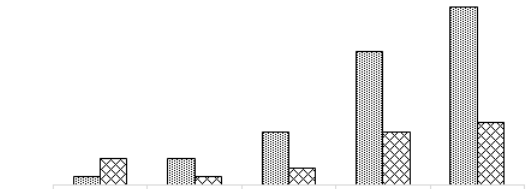


(b) Patients evaluate their cognition as much worse than patients in cluster 7. Study partners also evaluate it worse.



(c) Patients evaluate their behavior milder as their study partners, which consistently give the worst scores.

Figure 3: Normalized Everyday cognition (ECog) predictions for clusters 7 (3a), 8 (3b) and 9 (3c).



	CN	SMC	EMCI	LMCI	AD
M	1	3	6	15	20
F	3	1	2	6	7

Figure 4: Gender difference in cluster 8.

We have examined the profiles of predicted ECog features. The normalized predictions are shown in Figure 3. Cluster 10 is interesting in the sense that it includes two extremes, healthy patients and heavily affected patients. We assume that this cluster should be further split into two more homogeneous clusters. The exploration of this cluster is planned for further work.

Patients in cluster 7 (Fig.3a) evaluate their cognition as worse than 10 years ago. Their study partners evaluate the same behavior as approximately half as bad. The majority of patients have early and late MCI and the predictions for this cluster correspond to the distribution in Figure 2 quite well.

In cluster 8 (Fig.3b), where the majority classes are AD and LMCI, patients evaluate their behavior worse than those in cluster 7. Study partners in this cluster see the situation worse than study partners in cluster 7. In both clusters 7 and 8 the patients always evaluate their behavior worse than their study partners.

In cluster 9 (Fig.3c) we observe a change in this perception. Study partners evaluate the patients' behavior worse than the patients themselves. We see that the study partners consistently give worst scores for every testing sub-domain, which differs from evaluations in clusters 7 and 8. We can speculate that this observation is a direct result of the disease progression and medication. Given the fact, that clinical depression is very common with AD patients, it is possible, that the switch in perception is simply caused by medication for easing depression. On the other hand it could indicate a new disease signature, where patients have a different view on the world that manifests in different cognition. A more pessimistic explanation could be, that the ECog test is not suitable for this kind of analysis, since it only uses subjective scores, where evaluations from study partners cannot be considered as ground truth but a general direction of the functional direction of the patients. We have also analyzed the gender distribution within clusters 7, 8 and 9. We discovered that cluster 7 is gender balanced. Cluster 8 contains more male patients and this dominance is exhibited for all diagnoses as shown in Figure 4. Cluster 9 on the other side contains more women. Specifically, differences occur in classes EMCI and LMCI.

In addition to identifying a cluster of severely affected males and establishing a difference of perception between the patients and their study partners, we have also identified some important features that show potential for discovering specialized clusters. Our results show that AV45, FDG, hippocampal and fusiform volumes and ABETA_upennbiomk5 play an important role in the description of our clusters. As we already mentioned in Section 2.2, we have pre-pruned our clustering tree. The unpruned tree reveals additional important features such as the volume of entorhinal cortex, several laboratory measurements, including glucose level, PTAU_upennbiomk5, and white blood cell count.

4. CONCLUSIONS

This work presents an application of predictive clustering trees to the problem of discovering connections between biological and clinical features of patients with Alzheimer's disease. The result is a PCT with ten clusters, three of which are interesting. We have analyzed all three and discovered interesting indications that biological features have an impact on the observed clinical behavior of the patients.

We have also discovered gender specific differences, as we have initially expected in the design of the experiment. We have identified several biological features that might be connected with the Alzheimer's disease progression. The results are promising and in line with other studies, but additional research will need to be conducted in order to further validate the current results presented here.

5. ACKNOWLEDGMENTS

We would like to acknowledge the support of the Slovenian Research Agency (through a young researcher grant to MB and the programme grant Knowledge Technologies) and the European Commission (through the projects MAESTRA – Learning from Massive, Incompletely annotated, and Structured Data - grant FP7-ICT-612944) and HBP – The Human Brain Project - grant FP7-ICT-604102).

6. REFERENCES

[1] L. L. Barnes, R. S. Wilson, J. L. Bienias, J. A. Schneider, D. A. Evans, and D. A. Bennett. Sex differences in the clinical manifestations of alzheimer disease pathology. *Archives of General Psychiatry*, 62(6):685–691, 2005.

[2] H. Blockeel. *Top-Down Induction of Clustering Trees*. PhD thesis, Katholieke Universiteit Leuven, Department of Computer Science, 1998.

[3] H. Blockeel and J. Struyf. Efficient algorithms for decision tree cross-validation. *The Journal of Machine Learning Research*, 3:621–650, 2003.

[4] S. T. Farias, D. Mungas, B. R. Reed, D. Cahn-Weiner, W. Jagust, K. Baynes, and C. DeCarli. The measurement of everyday cognition (ecog): scale development and psychometric properties. *Neuropsychology*, 22(4):531, 2008.

[5] D. Kocev, C. Vens, J. Struyf, and S. Džeroski. Ensembles of multi-objective decision trees. *Machine Learning: ECML 2007*, pages 624–631, 2007.

[6] I. Slavkov, V. Gjorgjioski, J. Struyf, and S. Džeroski. Finding explained groups of time-course gene expression profiles with predictive clustering trees. *Molecular BioSystems*, 6(4):729–740, 2010.

[7] J. Struyf and S. Džeroski. *Constraint based induction of multi-objective regression trees*. Springer, 2006.

[8] C. Vens, J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.

[9] B. Ženko. *Learning Predictive Clustering Rules*. PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 2007.

Ingredients matching in bakery products

Tome Eftimov [1,2], and Barbara Koroušič Seljak [1]
[1] Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000, Ljubljana, Slovenia
[2] Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000, Ljubljana, Slovenia
{tome.eftimov, barbara.korousic}@ijs.si

ABSTRACT
In this paper, we present the analytical results of the ingredients matching in bakery products. We collected recipes from a free recipes web site and the main goal was to find association rules between the recipes' ingredients. For this purpose we applied an *Apriori algorithm* and various visualization techniques to represent the discovered association rules. The paper covers: data extraction, data preprocessing, association rules and visualization of the results during this work.

Keywords
association rules, text mining, ingredients matching

1. INTRODUCTION
The aim of the analysis presented in this paper was to find potentially interesting and relevant relations between the recipes' ingredients.
As our target data, we selected bakery recipes in English and focused on exploring relations between ingredients that occur in the bakery recipes.
First, we collected the data from a free Internet data source [1]. Afterwards, we preprocessed it in the form needed for the analysis. Then we looked for association rules and finished by representing discovered results and possible future work.

2. DATA
The data we used is a collection of 1,900 bakery recipes written in English, and we collected it using HTML parser to extract the information from a free recipes web site [1].
We considered the names of the ingredients for each recipe, while the quantity-unit pair associated with the ingredient was ignored as our goal was analysing only the relations between the ingredients.
Before the analysis, we preprocessed our target data. Because the data contained many adjectives that are associated with the cooking process (e.g. sliced, mashed), we removed them. We also located synonyms that appear in the data (e.g. pumpkin puree, pumpkin) and mapped them in the form required for the analysis. After cleaning the data, the preprocessed data was transformed into a document-text matrix and after that into a transactional matrix that is the form needed for our analysis. At the end, our transformed data contained 1,900 transactions (rows) and for each transaction we needed to consider the presence of 542 ingredients (columns).

For the cleaning process and the mapping of the synonyms we applied some regular expressions using the R programming language. The summary of the basic statistics of our data shows that the data set is rather sparse with a density just above 1.65%. The ingredient "salt" is the most popular and the average transaction contains less than 9 ingredients.

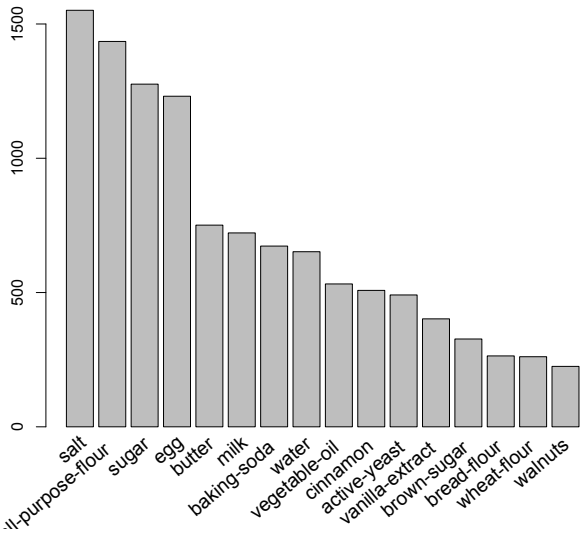


Figure 1: The most frequently used ingredients

In Figure 1, we can see that the ingredients "all-purpose flour", "egg", "salt" and "sugar" are most frequently used and because the probability of the presence of these ingredients in a bread recipe is very high, we rejected them for the analysis and focused upon the relations between other ingredients. After excluding the above mentioned most frequently used ingredients, our data set contained 1,900 transactions, each having 538 ingredients. The data set is rather sparse with density just above 1.13% and the average transaction contains less than 7 ingredients.

3. METHODS
Finding potentially interesting and relevant relations between the ingredients in bakery products is a task of the descriptive data mining method, known as the association rules mining [6]. In our case, having an association rule

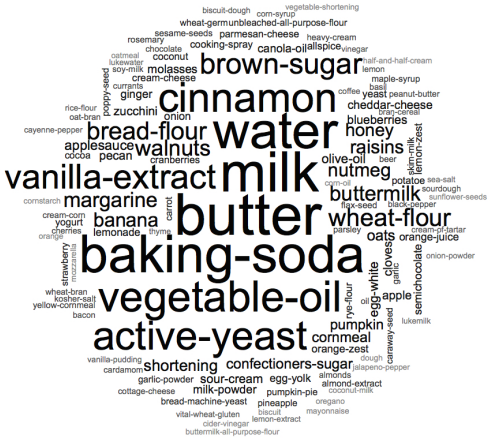


Figure 2: A wordcloud of the ingredients

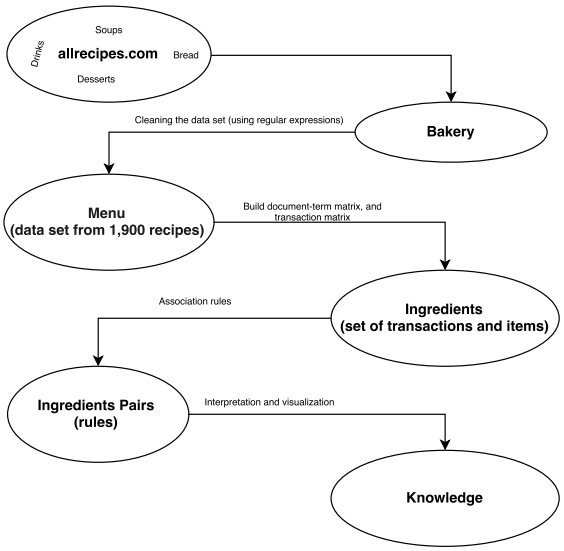


Figure 3: The knowledge discovery process

X→Y, where X and Y are sets of ingredients, the intuitive meaning of such a rule is that a recipe that contains all ingredients from X also tends to contain all ingredients from Y. The sets of ingredients X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule, respectively.

Because usually the number of such rules is huge, the space of all possible association rules needs to be reduced and for this purpose two criteria are used, support and confidence of the association rule.

Support of an association rule is the ratio of the number of recipes that have true values for all ingredients in X and Y and the number of recipes in our database. The confidence is the ratio between the number of recipes that have true values for all ingredients in X and Y and the number of recipes that have true values for all ingredients in X. Another measure that we used is a lift which tells us how many more times the ingredients in X and Y occur together then it would be expected if the sets of ingredients (X and Y) were statistically independent.

The whole knowledge discovery process is represented in Figure 3.

4. EVALUATION

There are several association rules algorithms and in our analysis we used the basic algorithm known as Apriori [3] and its implementation from the package "arules" in R [5].

After we imported the data into R, we used the Apriori algorithm to find the association rules and we tried it out for different values of the minimum support and minimum confidence. At the end, we decided to fix the support on 0.005, which means that at minimum 10 recipes will contain the ingredient and the confidence on 0.75. The number of discovered rules using these parameters is 1,235. Because some rules are redundant, which provide little or no extra information when some other rules are in the result, we pruned them and at the end we have 594 rules. The top 15 rules with respect to the lift measure are given in Table 1.

Because the number of the discovered association rules is huge and it is not recommended to go through all of them, we used some visualization techniques, which are implemented in the R's package "arulesViz" [4]. For visualization of our result we used graph-based visualization, parallel coordinates plots and grouped matrix-based visualization.

In Figure 4, we present the graph-based visualization with ingredients and rules as vertices for our top 10 rules with respect to the lift measure. Here the rules are the vertices, the size of the vertex is the support of the rule, while the color of the vertex is the lift of the rule. We can see how the rules are composed of individual ingredients and how they share ingredients. For example, we can see that if the recipe contains "garlic powder" and "milk" also tends to contain "cheddar-cheese". The graph-based visualization is an efficient technique to represent analytical results to people who are unfamiliar with data mining as from the graph they can see the relation between ingredients.

Another visualization suitable for people without knowledge on data mining is the parallel coordinate plot. In Figure 6, we present the parallel coordinate plot of our top 30 rules with respect to the lift. The width of the arrows gives the support and the intensity of the color presents the confidence. On the x-axis are represented the position in the rule, i.e., first ingredient, second ingredient, etc., while the arrow is used for the consequent.

In Figure 7, we have presented the grouped matrix-based visualization using a ballon plot with antecedent groups as columns and consequents as rows. The color of the ballon is the aggregated lift in the group, while the size of the ballon is the aggregated support. The aggregated lift is decreasing top down and from left to right, and the most interesting group is on the top left corner. The group of most interesting rules contains 5 rules, which contain "caraway seed" and 3 other ingredients in the antecedent and "rye flour" in the consequent. Another interesting group contains 2 rules,

	LHS	RHS	support	confidence	lift
1	{bread-flour, caraway-seed}	{rye-flour}	0.006	0.928	45238
2	{active-yeast, caraway-seed}	{rye-flour}	0.008	0.888	43304
3	{caraway-seed, water}	{rye-flour}	0.008	0.888	43304
4	{cranberries, orange-juice}	{orange-zest}	0.005	0.846	30333
5	{orange-juice, walnuts}	{orange-zest}	0.005	0.833	29874
6	{baking-soda, cinnamon, molasses}	{ginger}	0.006	0.800	24126
7	{garlic-powder, milk}	{cheddar-cheese}	0.005	0.769	21181
8	{cream-cheese, milk, vanilla-extract}	{confectioners-sugar}	0.005	0.909	16142
9	{baking-soda, cinnamon, nutmeg, water}	{pumpkin}	0.007	0.823	14901
10	{baking-soda, nutmeg, water}	{pumpkin}	0.007	0.789	14285
11	{butter, cream-cheese, milk}	{confectioners-sugar}	0.005	0.785	13951
12	{cinnamon, pumpkin-pie}	{pumpkin}	0.005	0.769	13919
13	{allspice, water}	{pumpkin}	0.005	0.769	13919
14	{pumpkin-pie, vegetable-oil}	{pumpkin}	0.006	0.764	13837
15	{bread-flour, butter, water, wheat-flour}	{honey}	0.005	0.833	10021

Table 1: The top 15 rules

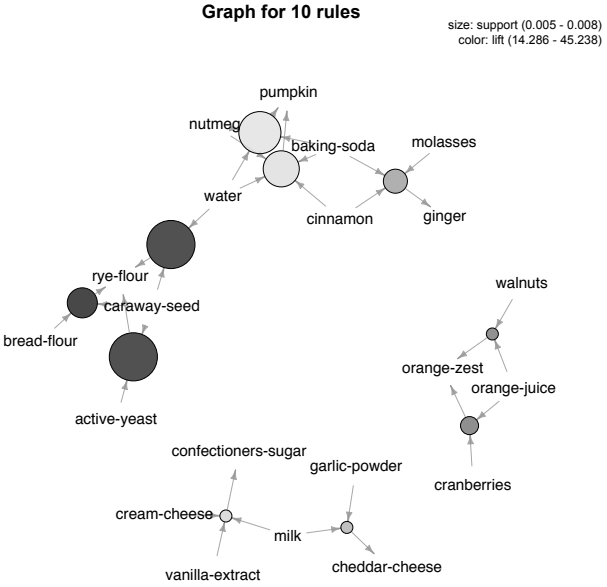


Figure 4: Graph-based visualization with ingredients and rules as vertices for top 10 rules

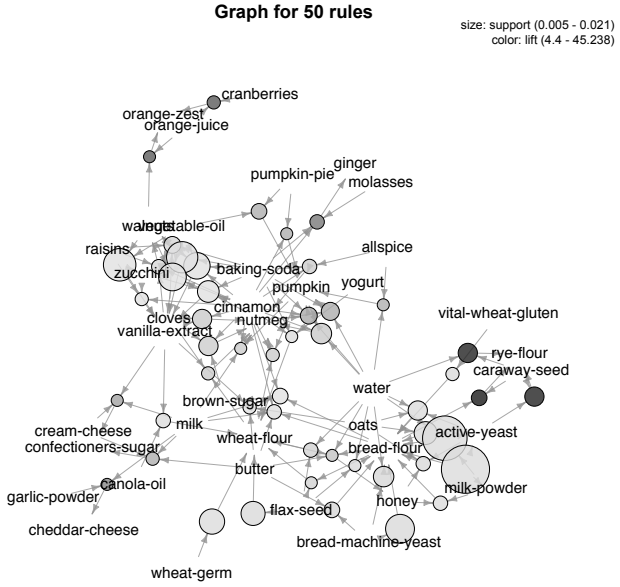


Figure 5: Graph-based visualization with ingredients and rules as vertices for top 50 rules

which contain "orange-juice" and 2 other ingredients in the antecedent and "orange-zest" in the consequent.

5. CONCLUSION

We analyzed 1,900 bakery recipes and found some interesting relations between the ingredients of the recipes. Some of the discovered rules are intuitively known, for example if the recipe contains "yeast" also tends to contain "water", if the recipe contains "apple" also tends to contain "cinnamon". We also found some unexpected combinations of the ingredients that occur in bakery recipes, for example the recipe that contains "baking-soda", "cinnamon" and "molasses" also tends to contain "ginger", the recipe that contains "baking soda", "nutmeg" and "water" also tends to contain "pumpkin". This analysis allows us to see how the ingredients are combined in bakery recipes. The information is very important for food compilers who need to collect analytical data for food items frequently used in national dietary surveys based on foods and recipes.

In the future, we would like to analyze these combinations in order to determine the nutritional properties for different values of quantity-unit pair for each ingredient and to discover for which values of the quantity-unit pair of each ingredient in the combination is good in the meaning of healthy diet. Also, to compare these relations with the relations provided by Foodpairing® that suggests, for one ingredient, those ingredients that create tasteful combinations with the given ingredient [2].

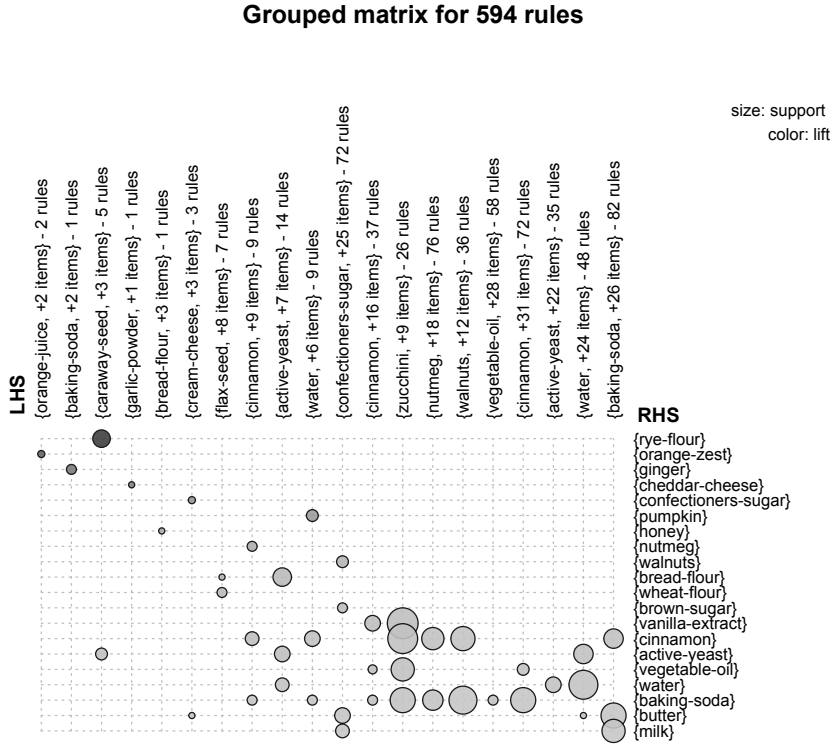


Figure 7: Grouped matrix-based visualization

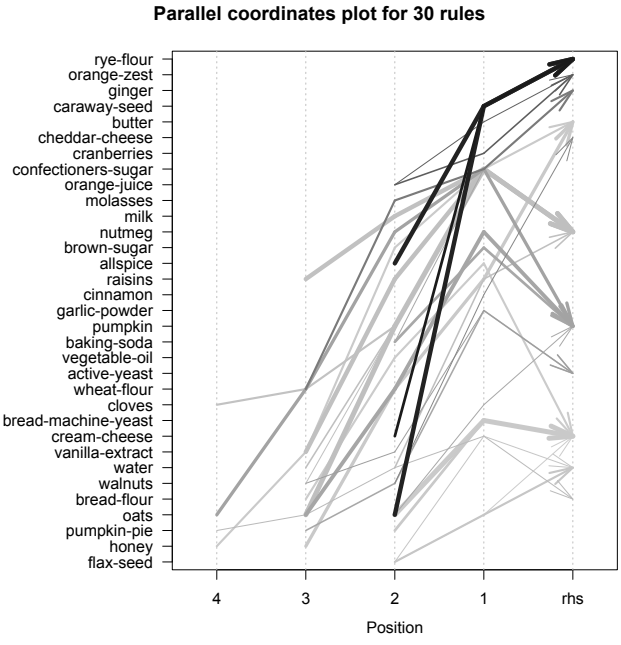


Figure 6: Parallel coordinate plot for top 30 rules

Acknowledgments

This work was supported by the project ISO-FOOD, which received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 621329 (2014-2019).

References

[1] Data source. http://allrecipes.com/Recipes/Bread/Main.aspx. Accessed: 2014-10-30.

[2] Foodpairing. https://www.foodpairing.com. Accessed: 2015-09-10.

[3] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[4] Hahsler, M., and Chelluboina, S. arulesviz: Visualizing association rules and frequent itemsets. *R package version 0.1-5* (2012).

[5] Hahsler, M., Grün, B., and Hornik, K. Introduction to arules–mining association rules and frequent item sets. *SIGKDD Explor* (2007).

[6] Tan, P.-N., and Kumar, V. Chapter 6. association analysis: Basic concepts and algorithms. *Introduction to Data Mining. Addison-Wesley. ISBN 321321367* (2005).

Mining scientific literature about ageing to support better understanding and treatment of degenerative diseases

Donatella Gubiani[1], Ingrid Petrič[1], Elsa Fabbretti[1] and Tanja Urbančič[1,2]

[1] University of Nova Gorica
Vipavska 13, Rožna Dolina
5000 Nova Gorica, Slovenia

[2] Jožef Stefan Institute
Jamova 39
1000 Ljubljana, Slovenia

{donatella.gubiani, ingrid.petric, elsa.fabbretti, tanja.urbancic}@ung.si

ABSTRACT

In this paper we demonstrate how literature mining can support experts in biomedicine on their way towards new discoveries. This is very important in complex, not yet sufficiently understood domains, where connections between different sub-specialities and fields of expertise have to be connected to fully understand the phenomena involved. As a case study, we present our preliminary literature mining work in the domain of ageing. The results confirm very recent discoveries about connections between diet and degenerative diseases, and indicate some concrete directions for further research needed to reveal the connections between microbiota and Alzheimer disease.

1. INTRODUCTION

Due to the rapid growing of scientific literature and the difficulties in knowledge integration between different scientific fields, IT represents a useful support to solve complex interdisciplinary questions. It has been demonstrated that connections and valuable hypotheses can be generated by linking findings across scientific literature with the use of literature mining methods and tools [5].

The first proposal of discovery based on different literatures was given by Swanson was given by Swanson [22]. He proposed the in silico model ABC that performs a search for new indirect relations between two disjoint sets of literature. Later, two main approaches have developed [25]. The first one, namely the closed discovery process, focuses on the test of a starting hypothesis: given two starting domains *a* and *c*, and their corresponding literatures A and C, the process extracts the common terms *b*, appearing in both literatures and representing potential bridges between the domains. Differently, the open discovery process is characterized by the absence of advance specification of target concepts: starting from a specific domain *c* and the corresponding literature C, the candidates for *a* are the result of the discovery process. Some methods combine both approaches. An example is the RaJoLink method [18] that suggests candidates for *a* based on exploration of rare terms in the literature on *c*.

Literature mining has been successfully applied in the field of biomedicine. Detailed survey of the earlier work is found in [11] and [14]. Recently, Zhang et al. [26] presented an application of their literature mining tool in retrieval of comorbidities for asthma in children and adults. Oh and Deasy [19] used literature mining to investigate chemoresistance-related genes and pathways of multiple cancer types. Their comprehensive survey and analysis provide a systems biology-based overview of the underlying mechanisms of chemoresistance. Rajpal et al. [21] applied literature mining for understanding disease associations in drug discovery. They showed how a literature mining system can be used to predict emerging trends at a relatively early stage of obesity and psoriasis by analysing the literature-identified genes for genetic associations, druggability, and biological pathways. Cameron et al. [7] also implemented a literature mining method for discovering informative and potentially unknown associations between biomedical concepts. Given a pair of concepts, their method automatically generates a ranked list of subgraphs that capture multifaceted complex associations between biomedical concepts.

In our study we applied literature mining to the domain of ageing. Ageing is an urgent health priority, with social and economical implications, that requires interdisciplinary biomedical research investments to validate multi-system intervention strategies and early diagnostic and prognostic biomarkers, as well as containment tools. While single-cell mechanisms of ageing processes are studied since several years, limited knowledge is available on the changes occurring at tissue, organ and system levels leading to progression of complex chronic age-related disorders, such as cardiovascular or neuronal diseases and cancer. On the top of genetic and individual predisposition, lifestyle environment and diet strongly contribute to occurrence of ageing and age-related diseases, although link among these factors is difficult to predict.

The main contribution of the paper is the proof of principle of in silico approaches to integrate the available literature concerning "ageing". Starting from the investigation of the existing scientific literature with ontologies and exploiting the method RaJoLink, we uncover candidate hypotheses for discoveries from terms, appearing rarely in scientific literature on this topic. As these methods indicated interesting connections between dietary issues and degenerative diseases in our preliminary studies, we focus on these particular aspects.

The paper is organized as follows. Section 2 describes the general methodology used in our work. In Section 3, we describe the investigated context exploiting ontologies and, in Section 4, we summarize different steps and results by applying the method

RaJoLink. Finally, conclusions and future works are drawn in Section 5.

2. METODOLOGY

PubMed [23] is a medical bibliographic database that contains more than 24 millions citations from years '40 to the present. Starting from the selected citations, the first step of our work consisted of a systematic analysis of the considered domain using ontologies (Subsection 2.1). Then, using an open discovery process allowed to detect new hypotheses (Subsection 2.2). Since background knowledge has an important role in guiding the discovery process, we used also expert assessments on relevant scientific issues.

2.1 Ontologies and OntoGen

An ontology is a data model that represents a domain. It is used to reason about it, about its main elements and the relationships between them. Several tools have been developed to support users to construct ontologies. One of these tools is OntoGen [13], a semi-automatic and data driven ontology editor focusing on editing of topic ontologies. Starting from a set of text-documents, it combines text-mining techniques with an efficient user interface to support the construction of ontologies including concept hierarchy visualization, concepts management with suggestions based on unsupervised and supervised machine learning methods, and concept visualization (an example in Figure 1).

2.2 Open discovery process and RaJoLink

RaJoLink [18] is an associative approach to identify and connect information across different contexts (literatures). It explores the role of rare terms in the texts that are not typical for the study domain. RaJoLink involves three main steps:

- ***Ra***re: the literature about a specific problem *A* (the domain under investigation) is examined in order to identify interesting terms that rarely appear in the selected documents (*R*);
- ***Jo***int: disjoint sets of documents about the selected rare terms are inspected to detect interesting joint terms $c_1,\ldots c_n$ that appear in their intersection;
- ***Link***ing: implementing the closed discovery, the last step links terms that bridge the gap between the starting literature *A* and the literature *C*.

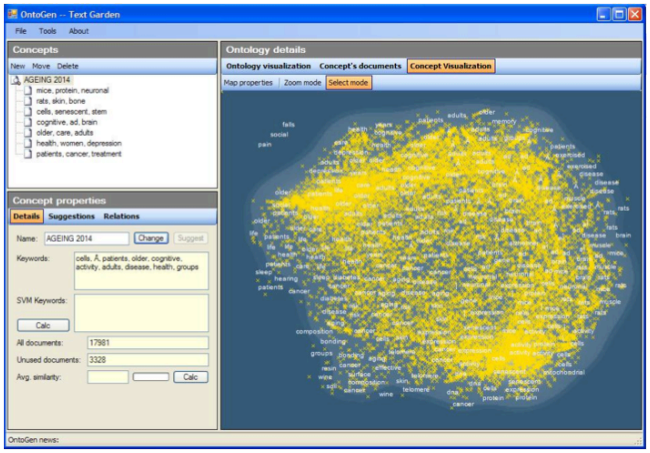


Figure 1: Concept visualization for literature about ageing in 2014.

3. ANALYSIS OF THE DOMAIN

At present, PubMed contains more than 300.000 citations about "ageing" (and its synonymous terms). When considered only the year 2014, more than 19.000 citations were found and Figure 1 shows the corresponding concept visualization. In it, we can view that recent investigations focus on different concepts, some of them connected with nutrition and food.

To validate our in silico method, we focused on important new findings, namely the link between ageing and nutrition, recently expanded by the use of high throughput -omics technologies (metagenomics, lipidomics, metabolomics, etc) and high power analysis of 'big data projects' [12]. Our work focused in a subset of 4.839 citations (with abstracts) obtained combining two properties: the conjunction of the term "ageing" with the term "food" and the most recent years (from year 2009 to 2014). Using OntoGen system, we created the two-level ontology shown in Figure 2. The first level distinguished four main data clusters. For the cluster related to the diet, at the second level we got a group of papers, related to brain.

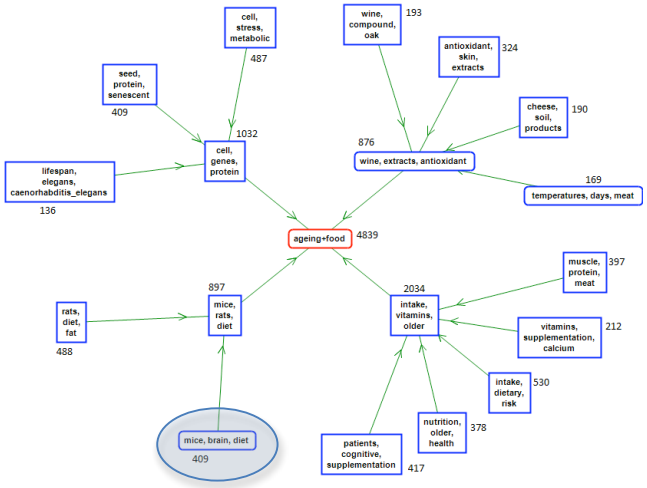


Figure 2: A two level ontology that captures the view of the domain "ageing and food".

The built ontology, obtained from the data for the last five years, revealed that "brain ageing" is a significantly reported keyword. This is consistent with the urgent need of understanding neurodegenerative diseases in elderly population. Following this route, we focused on the "Brain - Gut Axis" as described in [8] and briefly summarized at the beginning of the next section.

4. SEARCHING FOR CONNECTIONS EXPLANAINING "BRAIN-GUT AXIS"

The gut–brain axis integrates neural, hormonal and immunological signalling between the gut and the brain [8]. Bidirectional relation between these organs is mediated by neuro-active molecules, which are produced by the gut nervous system. In addition, gut microbiota biodiversity allows appropriate brain chemistry and can influence BDNF levels, learning and behavior. The diet and individual genetic and lifestyle factors influence gut microflora in their effective transformation of food in active micronutrients and essential enzymatic co-factors [4]. Interdisciplinary studies that integrated new sequencing approaches, with proteomic and metabolomic studies, have demonstrated the impact of the gut function on a wide range of health and disease processes, including chronic neuronal

disorders. A likely possibility is that personalized diet and nutrition supplements (such as probiotics) might limit chronic disease progression, cancer and ageing. These findings suggest the urgency to use IT tools to integrate data from different sectors (human health, biomedicine, microbiology, nutrition, etc) to provide guidelines for evidence based interdisciplinary research. The ability to predict a particular drug's pharmacokinetics and a given patients population's response to drugs via interfering on gut microbiome and diet, is one of the largest impacts of this field.

In the following subsection, the application of RaJoLink method and its results in this domain are described.

4.1 Rare

There is a vast interest about microbes that colonize the human gut (collectively referred as microbiota) and our health is the focus of a growing number of research initiatives. Starting from the most recent literature concerning "microbiota" (last 1000 papers with abstract available in PubMed in early June 2015), we applied the first step of RaJoLink method.

From the wide amount of rare terms detected by RaJoLink tool, we identified 3 rare terms "BDNF", "homocysteine" and "ubiquitin". While BDNF is a marker of learning abilities and brain well function [20], homocysteine is associated to ischemic brain damage [10] and ubiquitin refers to occurrence of protein post-translational modification mechanisms involved in protein quality control, a mechanism lost in neurodegenerative diseases, where large stack of unfolded unsoluble protein assembly is found [24]. Link of these terms with nutrition is subject of study, specifically associated to ageing diseases. In particular, the benefit of food supplements in endogenous BDNF levels is subject of clinical trials [9, 2].

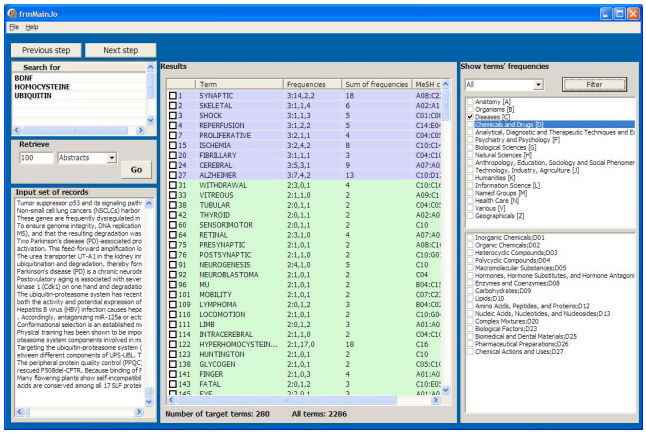


Figure 3: RaJoLink: from rare terms to candidate hypotheses.

4.2 Joint

The three rare terms "BDNF", "homocysteine" and "ubiquitin" have been the starting point for the second step in RaJoLink method. As visible in Figure 3, several joint terms have been detected. Checking their frequency, the first three terms have been "synaptic" (18 occurrences), "Alzheimer" (13 occurrences), and "cerebral" (9 occurrences).

In PubMed, we verified the connections between different literatures. In particular, Figure 4 focuses on "Alzheimer" literature: the (gut) "microbiota" and the "Alzheimer" literatures had very weak direct connection (Figure 5) but, considering joint terms and related literatures, indirect connections were found.

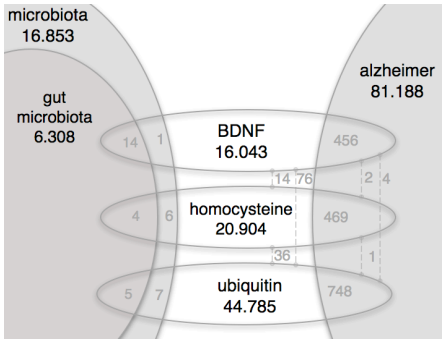


Figure 4: Analysis of connections between the scientific literature on "microbiota" and the scientific literature on Alzheimer through rare terms (PubMed: 31 August 2015).

4.3 Linking

Recent metagenomic and metabolomic screenings performed in the population, has proven the association of personal diet to individual gut microbiome profile in healthy and pathological ageing, rapidly expanding the impact of nutrition science [17]. In the next step, we performed a more detailed analysis between the scientific literatures on join terms. In particular, we focused our attention on the link between microbiota and Alzheimer literatures.

On today's date (August 2015), if we perform a search about combined literature ("microbiota" and "alzheimer"), we obtain only 9 results, 6 focusing on "gut" (Figure 5). They are published from 2013 to 2015: some of them later than papers used as input for our analysis. Significant part of these references suggests connection between the two topics being through diabetes and obesity (for gut microbiota [1, 16, 3, 6]). As described in [3], the composition of gut microbiota contributes to the development of diabetes Types 1, 2 and 3 by complex interactions of genetic and several environmental factors. Moreover, insulin resistance in the brain has been associated with Alzheimer's disease.

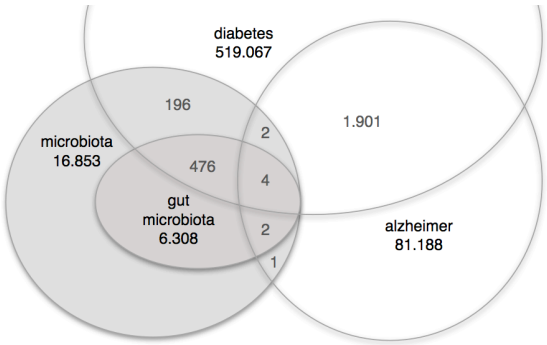


Figure 5: Connections between the scientific literature on microbiota and the scientific literature on Alzheimer (PubMed: 31 August 2015).

5. CONCLUSIONS AND FUTURE WORK

Presented literature mining methodologies are general and can be applied in different fields to guide discovery processes, providing that there is a good coverage of documents available on-line. The process is more efficient if there is a field expert available to cooperate since his or her guidance in selecting rare or joint terms might significantly fasten the convergence towards novel and interesting results. We applied literature mining to study ageing

context and we explored the role of rare terms that are not typical in the literature regarding “ageing” and “food”, to detect interesting, not yet fully understood connections, showing promising directions for further research concerning as dietary issues and degenerative diseases.

The increasing rate of data generation across all scientific fields provides new opportunities for data-driven research, with the potential to inspire new scientific trends. With the use of high throughput technologies, not only in the genetic field, but also in the metabolomic, nutrition and microbiology fields, the exploitation of 'big data' enable us to face the challenge of in silico integratomic analysis to find causative relations that might stimulate re-evaluation of existing knowledge as well as identify new data-driven medicinal chemistry and drug discovery processes [15]. In the filed of nutrition and diet, in particular, data integration analysis is often complicated by many confounding (lifestyle) factors. Validation of robust in silico tools is therefore highly desired for exploiting previous research in new interdisciplinary applications.

6. ACKNOWLEDGMENTS

This work was performed within the Creative Core project (AHA-MOMENT), partially supported by the Ministry of Education, Science and Sport, Republic of Slovenia, and European Regional Development Fund.

7. REFERENCES

[1] M. Alam, Q. Alam, M. Kamal, A. Abuzenadah, and A. Haque. A possible link of gut microbiota alteration in type 2 diabetes and Alzheimer’s disease pathogenicity: an update. *CNS Neurol Disord Drug Targets*, 13(3):383–90, 2014.

[2] J. E. Beilharz, J. Maniam, and M. J. Morris. Diet-induced cognitive deficits: The role of fat and sugar, potential mechanisms and nutritional interventions. *Nutrients*. 7(8):6719-38, 2015.

[3] P. Bekkering, I. Jafri, F. van Overveld, and G. Rijkers. The intricate association between gut microbiota and development of type 1, type 2 and type 3 diabetes. *Expert Rev Clin Immunol.*, 9(11):1031–41, 2013.

[4] J. Bienenstock, W. Kunze, and P. Forsythe. Microbiota and the gut-brain axis. *Nutr Rev.* 73 Suppl 1:28-31, 2015.

[5] P. Bruza and M. Weeber. *Literature-based Discovery*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[6] R. Buchet, J. Millán, and D. Magne. Multisystemic functions of alkaline phosphatases. *Methods Mol Biol.*, 1053:27–51, 2013.

[7] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. *Journal of Biomedical Informatics*, 54:141-57, 2015.

[8] S. M. Collins, M. Surette, and P. Bercik. The interplay between the intestinal microbiota and the brain. *Nat Rev Microbiol*., 10(11):735-42, 2012.

[9] A. D. Dangour, P. J. Whitehouse, K. Rafferty, S. A. Mitchell, L. Smith, S. Hawkesworth, and B. Vellas. B-vitamins and fatty acids in the prevention and treatment of Alzheimer's disease and dementia: a systematic review. *Journal Alzheimers Dis.* 22(1):205-24, 2010.

[10] G. Douaud, H. Refsum, C. A. de Jager, R. Jacoby, T. E. Nichols, S. M. Smith, and A.D. Smith. Preventing Alzheimer's disease-related gray matter atrophy by B-vitamin treatment. Proc Natl Acad Sci U S A, 110(23):9523-8, 2013.

[11] R. A.-A. Erhardt, R. Schneider, and C. Blaschke. Status of text-mining techniques applied to biomedical text. *Drug Discov Today*, 11 (7/8), 315-325, 2006.

[12] A. R. Ferguson, J. L. Nielson, M. H. Cragin, A. E. Bandrowski, and M. E. Martone. Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nat Neurosci.*, 17(11):1442-7, 2014.

[13] B. Fortuna, D. Mladenić, and M. Grobelnik. Semi-automatic construction of topic ontologies. In *Proc. of Joint Int. Workshops on Semantics, Web and Mining*, 121–131, 2006.

[14] L. Jensen, J. Saric, and P. Bork. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*, 7:119-129, 2006.

[15] S. J. Lusher, R. McGuire, R. C. van Schaik, C. D. Nicholson, J. de Vlieg. Data-driven medicinal chemistry in the era of big data. *Drug Discov Today*. 19(7):859-68, 2014.

[16] M. Naseer, F. Bibi, M. Alqahtani, A. Chaudhary, E. Azhar, M. Kamal, and M. Yasir. Role of gut microbiota in obesity, type 2 diabetes and alzheimer’s disease. *CNS Neurol Disord Drug Targets*, 13(2):305–11, 2014.

[17] J. K. Nicholson, E. Holmes, J. Kinross, R. Burcelin, G. Gibson, W. Jia, and S. Pettersson. Host-gut microbiota metabolic interactions. *Science*. 336(6086):1262-7, 2012.

[18] I. Petrič, T. Urbančič, B. Cestnik, and M. Macedoni-Lukšič. Literature mining method rajolink for uncovering relations between biomedical concepts. *Journal of Biomedical Informatics*, 42(2):219–227, 2009.

[19] J. H. Oh, and J. O. Deasy. A literature mining-based approach for identification of cellular pathways associated with chemoresistance in cancer. *Brief Bioinform.*, pii: bbv053, 2015.

[20] S. L. Patterson. Immune dysregulation and cognitive vulnerability in the aging brain: Interactions of microglia, IL-1β, BDNF and synaptic plasticity. Neuropharmacology. 96(Pt A):11-8, 2015.

[21] D. K. Rajpal, X. A. Qu, J. M. Freudenberg, and V. Kumar. Mining emerging biomedical literature for understanding disease associations in drug discovery. *Methods Mol Biol.*, 1159:171-206, 2014.

[22] D. R. Swanson. Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association*, 78(1):29–37, 1990.

[23] U.S. National Library of Medicine. PubMed. url: http://www.ncbi.nlm.nih.gov/pubmed

[24] D. Vilchez, I. Saez, and A. Dillin. The role of protein clearance mechanisms in organismal ageing and age-related diseases. *Nat Commun.*, 5:5659, 2014.

[25] M. Weeber, R. Klein, and L. T. W. de Jong-van den Berg. Using concepts in literature-based discovery: Simulating Swanson’s Raynaud-fish oil and migraine-magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(7):548–557, 2000.

[26] Y. Zhang, I. N. Sarkar, and E. S. Chen. PubMedMiner: Mining and Visualizing MeSH-based Associations in PubMed AMIA. Annu Symp Proc., 1990–1999, 2014.

Modelling in Energy Related Scenarios

Klemen Kenda
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
klemen.kenda@ijs.si

Maja Škrjanc
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
maja.skrjanc@ijs.si

Andrej Borštnik
Jozef Stefan Institute
Jamova ulica 39
Ljubljana, Slovenia
andrej-borstnik@hotmail.com

ABSTRACT
Fusing heterogeneous multivariate data in stream mining scenarios is a demanding task. Successful fusion requires a well-thought approach. We propose the use of a stream processing engine (SPE) that enables implementation of all the needed methods and ensures almost real-time responsiveness of the system.

In the paper we propose an infrastructure that is able to receive data from various heterogeneous sources (static properties, weather data and forecasts, other forecasts, and primarily sensor data). In the implementation of the proposed infrastructure we address issues related to the heterogeneous nature of the data, like different frequency, different update interval, and different nature of the data. The pipeline was used to prepare stream prediction models for five different energy-related use cases, which include public buildings, a thermal plant production, university campus buildings, and EPEX energy spot market prices alongside the total traded energy.

Keywords
Data fusion, modelling, prediction, data streams, sensor data, sensor networks, regression models, QMiner.

1. INTRODUCTION
Nature of obtaining data (wide availability, vast amounts) has changed the paradigm of modelling nowadays. It is fairly easy to measure certain phenomena with a continuous stream of measurements and it is even easier to add various open data to the set of modelling features.

Most of the systems are working in (almost) real time, which favours the streaming setup for modelling and predicting. Many of the classical prediction methods have already been ported to the streaming scenario. However in our work we have tackled a demanding technical challenge of fusing heterogeneous multivariate data sources to prepare valid feature vectors for modelling.

In this paper we are addressing methods for predicting energy-related phenomena in public buildings, energy markets, and at energy provides.

The paper presents an overview of potential additional data sources for the problem in question, showcases a suggested set of features for certain cases in energy related modelling, it provides an evaluation of different prediction methods, suggests an architecture for the multimodal stream modelling data fusion, and finally presents results from four different use cases processed within the platform.

2. FEATURES AND FEATURE VECTORS
Accuracy of prediction models is usually more dependent on the features used than on the modelling method chosen. Extensive analysis of five energy related use cases [2] has lead us to the following set of features with specific properties: sensor features, forecasts, and static properties. Table 1 depicts an example of a full feature vector for energy consumption modelling of the National Technical University of Athens (NTUA) campus building.

Table 1. Full feature set for campus building (NTUA) use case

Type	Feature			
	Name	***UoM[a]***	***Value[b]***	***Aggr.[c]***
Sensor	current_l1	A	X(0)	
	current_l2	A	X(0)	
	curent_l3	A	0	
	energy_a	kWh	0, -1h, -1d	
	demand_a	MW	0	yes
	demand_r	kvar	0	
Weather	temperature	°C		yes
	wind speed	m/s		yes
	wind direction	°		yes
	Visibility	km		yes
	Humidity	%		yes
	Pressure	mbar		yes
	cloud cover	%		yes
Weather forecast	temperature	°C	t	
	wind speed	m/s	t	
	wind direction	°	t	
	cloud cover	%	t	
	Humidity	%	t	
Static properties	weekday		t	
	dayOfWeek		t	
	month		t	
	working day		t	

Type	Feature			
	Name	*UoM*[a]	*Value*[b]	*Aggr.*[c]
	working hour		t	
	holiday		t	
	day before holiday		t	
	day after holiday		t	

[a] Unit of measurement

[b] Value, expressed with relative time (0 = current timestamp, -1h je timestamp 1 hour ago; t denotes the timestamp of prediction)

[c] Configuration of aggregates is much more complex, further details can be found in [1]

The sensor data can be understood as the most fundamental streaming data. In an ideal case it is arriving to the SPE in (almost) real time as a conservative data stream (where measurements are ordered by a timestamp). Often transport systems implement different kinds of buffering, which means that the data is coming either with a delay, or even in chunks of multiple measurements. In a streaming scenario it is important that we are able to handle any exceptions and ensure that stream mining methods are fed feature vectors only when they include the most recent data.

Forecast data represents different kind of predictions, most commonly weather predictions. Forecasted data can also be classified as a stream, but with different properties. Forecasts get updated regularly. For example weather forecasts are updated every few hours and the system needs to be able to update the time series.

Static properties data is relatively easy to handle, as it can (in most cases) be calculated "a priori". Such data includes features like time of day, week, day of year, day of week, holidays, working days, weekends, moon phase etc. The data is similar to sensor data in the sense that it does not need to be updated and to prediction data in the sense that models usually refer to the future (and not current) values.

3. HANDLING MULTI-MODAL DATA

The implemented system has already been described in detail in [1]. In this contribution we will only describe the outline of the work done on the technical part and will rather focus on the results.

The system is built on top of the QMiner open-source platform [3]. We implemented two different systems: a data system and a modelling system. In the current setup we are running one instance of a data system (which collects data, orders it by time, handles properties and static data, and distributes it) and multiple instances of modelling systems (which merge separate data streams, resample them, create feature vectors, and model).

In the whole pipeline a number of components have been implemented: store generators, data adapters, aggregators, a time sync component, a load manager, a receiver, merger, a resampler, a meta-merger, and a semi-automated modeller.

The final result is a (single point) configurable system that is able to learn on the past data and give almost real-time predictions for various phenomena.

4. RESULTS

The proposed methodology follows the on-line learning paradigm. All the evaluations were done on real-time, or a simulated stream of real, data.

The following methods have been implemented in the platform:

- Linear regression (LR)
- Support Vector Machine Regression (SVMR)
- Neural networks (NN)
- Moving average multiple models (MA)
- Hoeffding trees (HT)

Most of the methods were adjusted to work in a stream mining scenario, except SVMR, which uses repetitive learning.

Properties of the predicted values have enabled us to use well known evaluation metrics. We have computed the mean error (ME), the root mean squared error (RMSE), and the R^2 measure to determine the best possible model.

Feature sets in the results below are denoted by:

- AR – auto-regressive features
- F – weather forecasts
- P – static properties
- S – additional sensor data
- ALL –a full feature vector

Modelling demand in the energy related scenarios seemed to be quite unified for all the studied use cases. The customer is usually interested in an energy profile for the next 24-36 hour period at around 12:00 each day.

The data has a distinct daily period and the first modelling decision was to build 24 models for the task – each predicting for a specific hour of the day.

An interesting observation was that the weather data (current) never improved the accuracy of predictions. Weather data from available global web services also seems to contribute little to the prediction model. Historic weather forecasts from the web service used are very accurate (service provides the latest – short term - prediction for a location), which means that some bias of the longer term weather predictions might be lost. How this effects the modelling has not been studied.

4.1 Public Building (CSI)

Public building in Turin offered 2 years of data for the learning phase and 1.3 years of data for evaluation. The total number of features was 48. We were trying to predict building electricity consumption (cooling excluded).

All of the methods behaved quite well on this data set, but SVMR yielded the best results. All the methods in this case have been significantly better than the best base-line method (moving average over the last week). Results are shown in Table 2.

Table 2. Results from public building (CSI) use case

Method-feature set (parameters)	Error Measure		
	ME	*RMSE*	R^2
SVMR-ARFP (eps=0.015)	-2,74	16,50	0,84

Method-feature set (parameters)	Error Measure		
	ME	*RMSE*	R^2
SVMR-ARP (eps=0.05)	-2,51	17,23	0,83
LR-ARFP	-3,24	17,96	0,81
LR-ARP	-3,46	18,19	0,81
SVMR-ALL (eps=0.05)	-1,96	18,67	0,80
LR-ARSFP	-0,78	19,54	0,78
LR-ARSP	-0,81	19,74	0,77
NN-ALL (6,lr=0.02)	0,32	19,90	0,77
HT-ARSFP	-2,69	20,02	0,77
MA (7)	0,01	30,89	0,44

In Figure 1 an example of prediction vs. measurements is depicted. This is a normal example form the validation part of the data set. We can see that the model is unexpectedly good with an exception of Mondays, where something that could not be modelled by the feature set appeared.

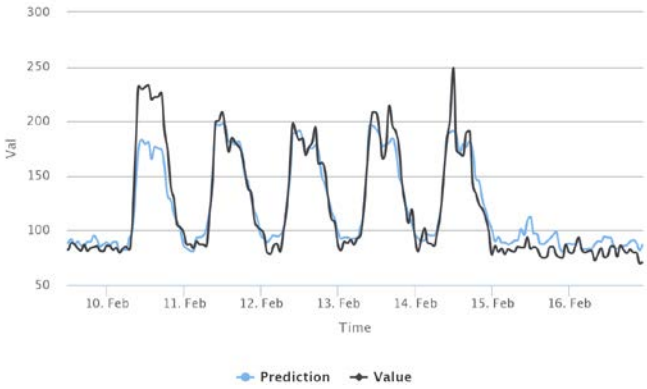


Figure 1. Prediction for a selected Turin public building, for a week in February 2015.

Further drill-down of the weights of the LR model has shown that the most significant features were the 1(one) week aggregates of auto-regressive and other sensor features. From the weather forecast feature temperature was surprisingly not among the most significant features, but cloud cover (solar radiation) and humidity were. Additional features such as day/hour classification (weekend, holiday, day after holiday, working hours, and heating season) were utilized the most.

4.2 University Campus Building (NTUA)

University campus of NTUA offers 5 years of valid data, which was divided into 3 years for learning and 2 years for evaluation. We are modelling average power demand for a selected building.

Results of the tests on this dataset have shown that the features provided for modelling are unable to handle all the dynamics of the system. Parts of the test data have been modelled quite well (see Figure 2), but the model did not handle the other parts well (see Figure 3). This might be a good indicator of possibly faulty, or simply missing data in the feature set.

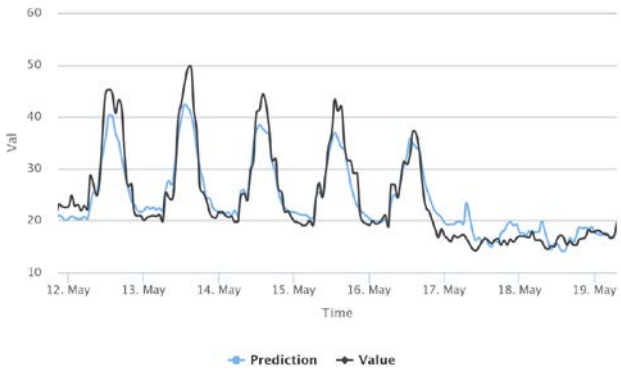


Figure 2. Model works well at some point.

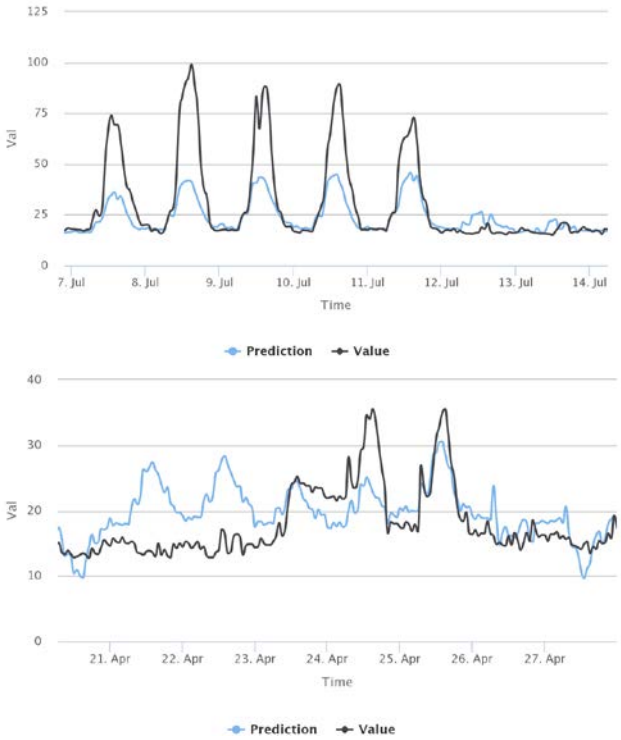


Figure 3. Unhandled exceptions in the modelling.

4.3 Energy Prices in Energy Spot Market (EPEX)

Data for the energy spot market has been scraped from the EPEX spot market web pages and streamed into the platform. At the testing phase there were 3 years of data available for learning and 1.4 years for evaluation. Energy prices and total trading energy for Germany were used in the experiments.

Energy prices strongly depend on the production of energy from the alternative energy sources. The production costs for such energy are usually very low or equal to zero, but the energy grid is not yet prepared for such irregular intake of energy. Excessive production of energy form alternate sources therefore results in lowered prices (sometimes even negative prices).

Feature vectors have therefore included data from 6 different weather stations across Germany, especially the wind data (speed and bearing) and cloud cover were expected to be the most important features.

Table 3. Results from energy prices (EPEX) use case

Method-feature set (parameters)	Error Measure		
	ME	***RMSE***	***R^2***
LR-ARSFP	-0,53	8,59	0,71
LR-ALL	-0,28	8,64	0,70
SVMR-ALL (c=0,037, eps=0,034)	1,01	8,94	0,63
LR-ARFS	-0,22	10,29	0,58
HT-ARSFP	-2,29	13,41	0,29

According to Table 3 the safest methods behave best. Weight analysis of the LR showed that the most important features were energy prices in the previous days, total traded energy averages for 1 week, 1 month, and minimum/maximum total traded energies for previous week.

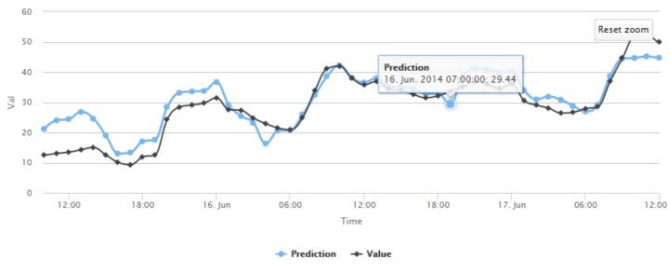


Figure 4. Prediction for EPEX use case for energy prices.

From the weather data it was interesting to see that wind bearing was the most dominant feature (as it was much more weighted than wind speed). Cloud cover has not contributed significantly to the behaviour of the models.

4.4 Thermal Plant (IREN)

1.6 years of data for thermal plant in Reggio nell'Emilia were available. 1.1 year was used in the learning phase and 0.5 years for testing. There were 43 features in the dataset.

During the experiments we were unable to satisfactory model part of the data and therefore some of the measures in Table 4 are distorted. The results for most of the data set are however very good, as can be seen in Figure 5.

Table 4. Results from thermal plant (IREN) use case

Method-feature set (parameters)	Error Measure		
	ME	***RMSE***	***R^2***
LR-ALL	-1,27	17,41	0,80
LR-AR	-0,08	17,94	0,79
MA (4)	-0,70	17,99	0,79
NN (4-6-3, lr=0.04)	-0,10	18,65	0,77
SVMR (c=0.03, e=0.02)	0,19	19,25	0,75

The weight analysis of the LR model shows significant contributions from most of the features.

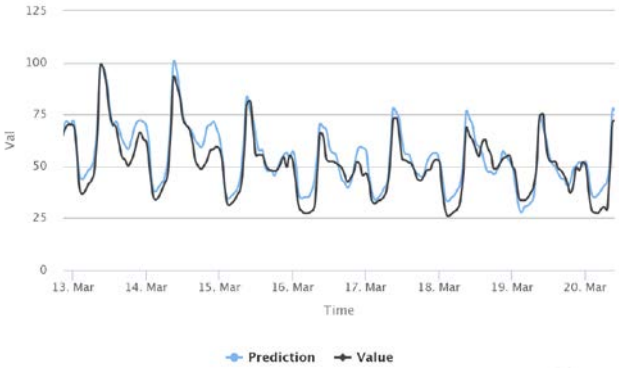


Figure 5. IREN use case prediction example.

5. CONCLUSIONS

In this paper we have presented models developed in the energy related scenarios using stream mining methods. We have developed a stack of components that are able to handle sensor data, forecasts, and static properties in a stream mining scenario. The platform has enabled us to provide streaming models for different energy-related phenomena. With the platform we are able to handle heterogeneous data from different independent data sources, such as different sensor systems, web services, or static flat files.

Accuracy of the developed models is mostly very good; however there are periods in the evaluation sets that we were sometimes unable to handle, which indicates insufficient feature sets.

The described developed models are currently in use.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under NRG4Cast (FP7-ICT-600074).

REFERENCES

[1] K. Kenda, M. Škrjanc and A. Borštnik. *Modelling of the Complex Data Space.* Information, Intelligence, Systems, Applications, Corfu, July, 2015.

[2] K. Kenda et al. *Modelling of the Complex Data Space*, NRG4CAST project deliverable D3.1, Ljubljana, November, 2014.

[3] B. Fortuna, J. Rupnik, J. Brank, C. Fortuna, V. Jovanoski and M. Karlovcec. *QMiner: Data Analytics Platform for Processing Streams of Structured and Unstructured Data*, Software Engineering for Machine Learning Workshop, Neural Information Processing Systems, 2014.

[4] K. Kenda, L. Stopar, M. Grobelnik. *Multilevel Approach to Sensor Streams Analysis*, Discovery Science, Bled, October, 2014.

Forecasting sales based on card transactions data

Alexandra Moraru, Dunja Mladenić
Jozef Stefan Institute and Jožef Stefan International Postgraduate School,
Jamova 39, 1000 Ljubljana , Slovenia
firstname.lastname@ijs.si

ABSTRACT

Smart cities are an important topic in today's research problems, with high impact in many domains from economy to transportation, health and living style. The problem addressed in this paper is that of sales forecasting for a specific category of products. We present the results of three regression algorithms, applies on real live data, for predicting the cumulative hourly sales of petrol. The prediction is made for three short term intervals, of 1, 4, and 8 hours into the future. A study has also been conducted in order to identify the amount of historical data required for optimal results.

Keywords

Data mining, sale forecasting, regression algorithms, smart cities.

1. INTRODUCTION

Smart cities are an important topic in today's research problems and can be defined as complex problems, combining multimodal data from several sources. The high level requirements for making a city smarter, as envision by IBM in the larger Smarter planet program [1], refer to the collaboration and coordination between city agencies managing different domains (e.g. water management, transportation, buildings, etc.) in order to be able to optimize the limited resources and to efficiently and effectively deliver city services. Moreover, different technologies may enable smarter cities, such as: communication channels (e-mail, instant messaging, etc.), business rules, data sharing (data models, accessibility) and integration of different sources of data [4]. In another study [3] the classification of cities as smart is made based on 6 criteria: economy, people, governance, mobility, environment and living. The problem presented in this paper relates to the first criterion, the smart economy, with indirect relation to transportation and living, as the main objective is that of petrol sales forecasting. [2]

The task of forecasting employs a set of methods and tools for making predictions of the future, relying on past and present data and analysis of trends. A well-known example is the prediction of a target variable at a specific time in the future. Sales forecasting is important in business management and decision making. Short-term and long-term sales forecasting can be affected by many factors, including economic up or downturns, changing trends and fashion, season, etc.

A typical research problem is sales forecast for a particular business or industry. In this paper we present the results of cumulative sales forecasting, where the amount we are predicting refers to a whole category of products, over a larger geographical area, independent of the particular businesses performance. More specifically, we predict the petrol sales in a city, given historical card transactions. Individual card transactions are aggregated on an hourly basis and used for short term prediction of 3 different intervals: 1, 4 and 8 hours into the future. Three regression algorithms are applied on a real live datatset, and their performance is evaluated for different forecasting.

The rest of the paper is structured as follows. Section 2 describes the data used in our experiments. Section 3 describes the methods and algorithm used, while Section 4 presents the results. Finally we conclude the paper.

SiKDD'15, October 1–2, 2015, Ljubljana, Slovenia.

2. DATA DESCRIPTION AND PREPROCESSING

The data used in our experimentation consists of individual card transactions over a period of 13 days. The information available is the time of the transaction, the amount of euros spent and the category of goods purchased. We focus on overall petrol purchases in a city and describe this data in detail.

The aggregated amount (sum) of euros spent on petrol every hour is illustrated in Figure 1. Based on this data, our main objective is to predict the hourly consumption for different time intervals in the future, respectively 1, 4, and 8 hours. A second objective is to analyze how much historical data is optimal for our prediction.

A training instance consists of the current hour of transaction and several aggregated amount of euros spent every hour in the past, for various time intervals. For example, if we would like to predict the amount of euros spent in the next hour, using 4 hours of historical data, our training instance consists of 6 attributes: current hour, euros spent this hour, euros spent 1, 2, 3 and 4 hours ago, where the class (target attribute) is euros spent this hour.

The dataset created for experimentation consists of 303 instances, which contain up to 26 attributes (one attribute is the prediction hour, the rest are hourly aggregates), for the experiments where 24 hours of history are used. All attributes are numeric and there are no missing values. The statistic values of 1 hour aggregates euros spent on petrol, for minimum, maximum, mean and standard deviation are presented in Table 1.

Table 1. Statistic measures of 1 hour aggregate consumption for petrol

Statistic	Value
Minimum	0
Maximum	430982.18
Mean	81529.56
Standard Deviation	83710.81

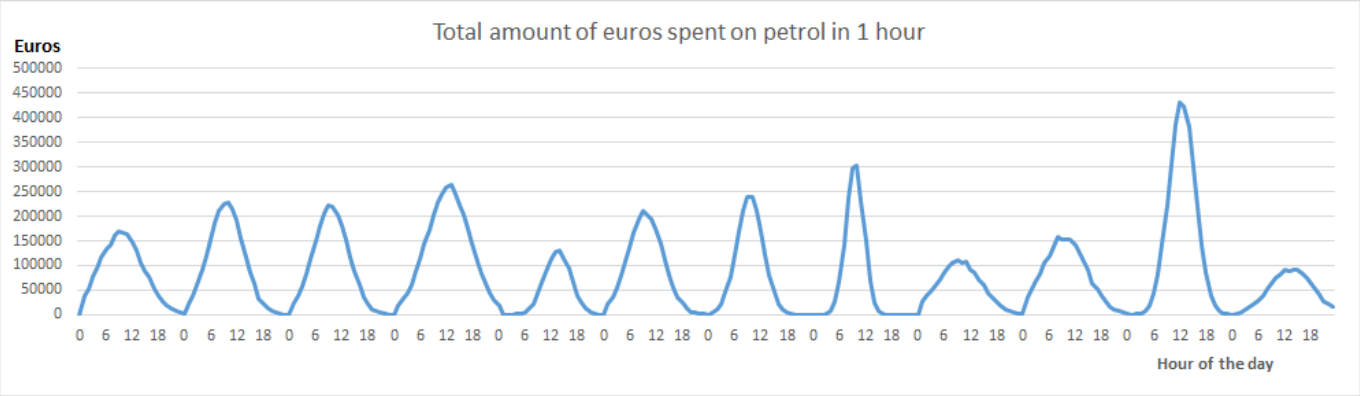


Figure 1 Hourly amount of euros spent on petrol

3. LEARNING METHODS

As our objective is to predict the amount of euros spent in 1 hour for the selected purchase category (petrol), the problem can be formulated as a regression problem. As several regression algorithms are available, we took into account the simplicity, the model understandability and the best reported performances reported, before selecting the algorithms for experimentation. Therefore, for our experiments we have selected three regression algorithms: linear regression, regression tree (M5P model tree and rules) and support vector machine for regression (SMOreg). All algorithm implementation have been selected from Weka toolkit [5].

The target variable to be predicted is the amount of euros that will be sent in 1 hour, 4 hours and 8 hours into the future. We conducted several experiments, providing between 4 and 24 hours of historical data (in 4 hour increments). The performance results of the algorithms are reported in Table 2.

For evaluation we have used separate training and test set, as we consider it to be closer to real live situation, compared to cross-fold validation. The split percentage is 66%, and the order of split is preserved, meaning that first 200 instances are used for training and remaining 103 instances are used for test.

Table 2 Evaluation for prediction of total amount of euros spent in 1 hour for transactions in the petrol category. The two evaluation measures reported are correlation coefficient and relative absolute error. The predictions are made for 1 hour, 4 hours and 8 hours into the future, while the number of hours in the past, used for training the models, are varied between 4 and 24 hours. The algorithms evaluated are support vector machine for regression (SMOreg), regression tree (M5P) and linear regression (LinearReg).

	predicting 1 hour ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.9938	0.9942	0.994	0.9926	0.99	0.9853	10.4874	10.3596	10.9284	12.3995	13.5481	16.786
M5P	0.9623	0.9785	0.983	0.973	0.9652	0.9378	22.5781	20.2358	19.1709	22.8912	23.6852	29.8357
LinearReg	0.5223	0.706	0.5565	0.5237	0.5237	0.8437	80.1759	71.7612	83.5529	85.5285	85.5285	57.3464
	predicting 4 hours ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.7753	0.8165	0.8011	0.748	0.7168	0.6811	58.5435	56.5327	60.8278	67.2914	63.6155	76.0387
M5P	0.7696	0.709	0.6801	0.7346	0.6276	0.6107	60.492	63.7024	73.8301	68.9027	76.6383	88.384
LinearReg	0.7362	0.6628	0.6212	0.5785	0.6327	0.6107	66.589	77.3408	83.7837	90.6461	74.2057	88.384
	predicting 8 hours ahead											
	Correlation coefficient						*Relative absolute error*					
	4h	*8h*	*12h*	*16h*	*20h*	*24h*	*4h*	*8h*	*12h*	*16h*	*20h*	*24h*
SMOreg	0.4957	0.5594	0.5294	0.4599	0.4485	0.4232	100.2677	94.3854	96.3161	96.5257	97.1401	108.3501
M5P	0.5637	0.5939	0.6238	0.4883	0.4919	0.4892	82.5027	88.767	88.5318	103.6142	98.4414	99.3843
LinearReg	0.277	0.5528	0.5131	0.497	0.4527	0.4619	104.3004	98.7528	103.1733	92.3864	101.7982	103.167

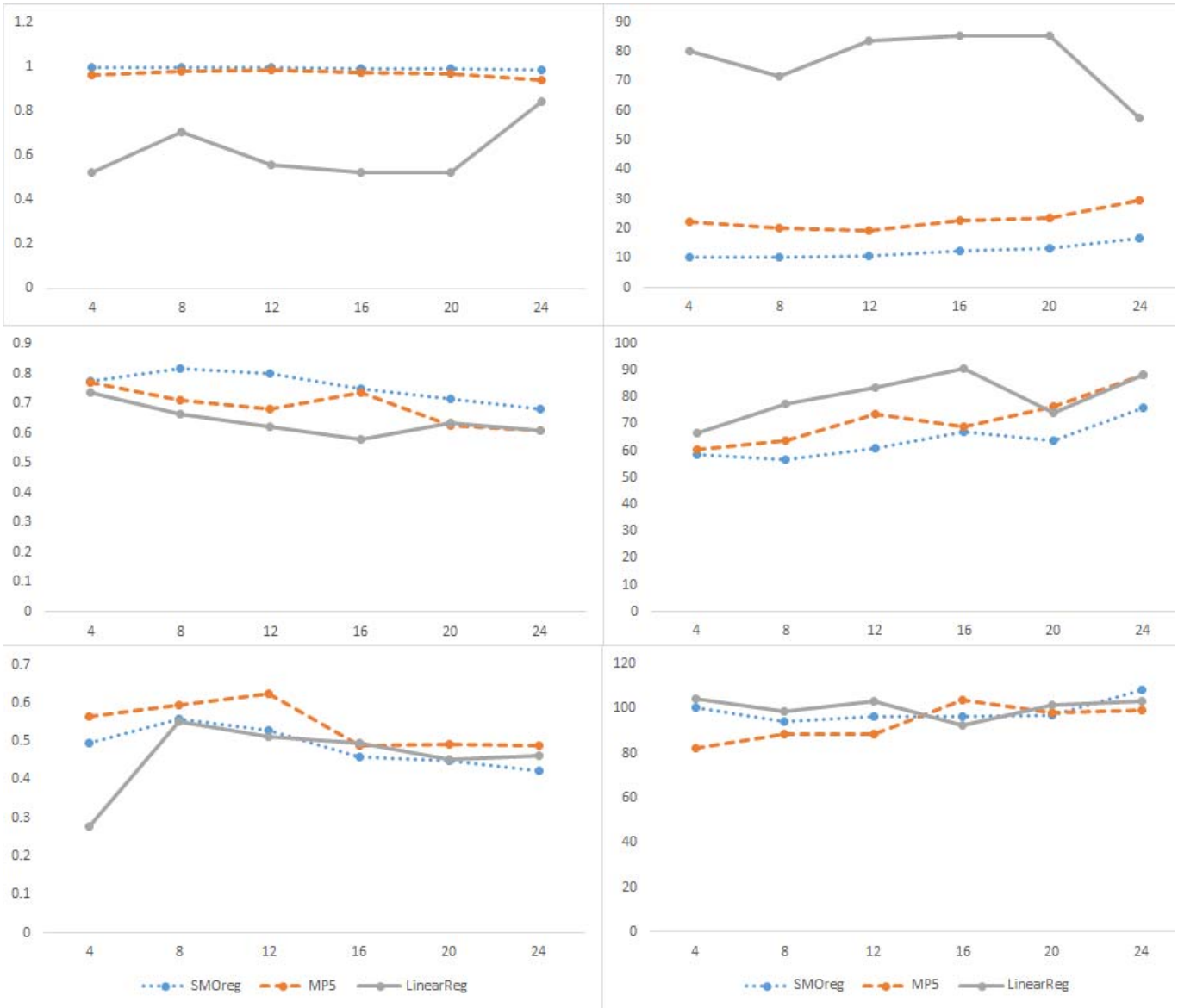


Figure 2 Algorithms performance for different amount of historical data used, reported by the correlation coefficient (left side) and relative absolute error (right side). The top graphs are for 1 hour ahead prediction, middle graphs for 4 hours ahead prediction and bottom graph for 8 hours ahead prediction. On the Y axis are the values of the evaluation measured and on the X axis is the amount of historical data used for building the models (from 4 to 24 hours, in 4 hours increments)

4. RESULTS AND DISCUSSION

The algorithms have been tested for different amount of historical data used in building the prediction model, in order to find the optimal amount of historical data needed and to identify the more robust algorithms for our problem.

The performance of the learning algorithms has been measured in terms of correlation coefficient and relative absolute error.

From the analysis illustrated in Figure 2, several observations can be made:

- As expected, the algorithms performed best when the amount of euros spend predicted is for a shorter time interval in the future.
- SMOreg algorithms performed the best in most cases, closely followed by M5P.
- Shortest prediction interval, that of 1 hour ahead, does not benefit from more the 12 hours of historical data
- Linear regression presented high variability to the amount of historical data provided, counter intuitive to what one would expect. More specifically, it can be observed the when given more the 8 hours of historical data the performance of the algorithm decreases.
- When the prediction interval is larger (8 hours into the future) none of the algorithm reported very good performance, however, it can be noticed that M5P is slightly superior to the rest.

The actual and predicted values for 1 hour petrol consumption are illustrated in Figure 3. The best performing model has been selected for each of the 3 category of prediction: 1, 4, and 8 hours. The first two graphs illustrate the results using SMOreg algorithm for 1 and 4 hours ahead prediction and the third graph illustrated the results for 8 hours ahead prediction using M5P.

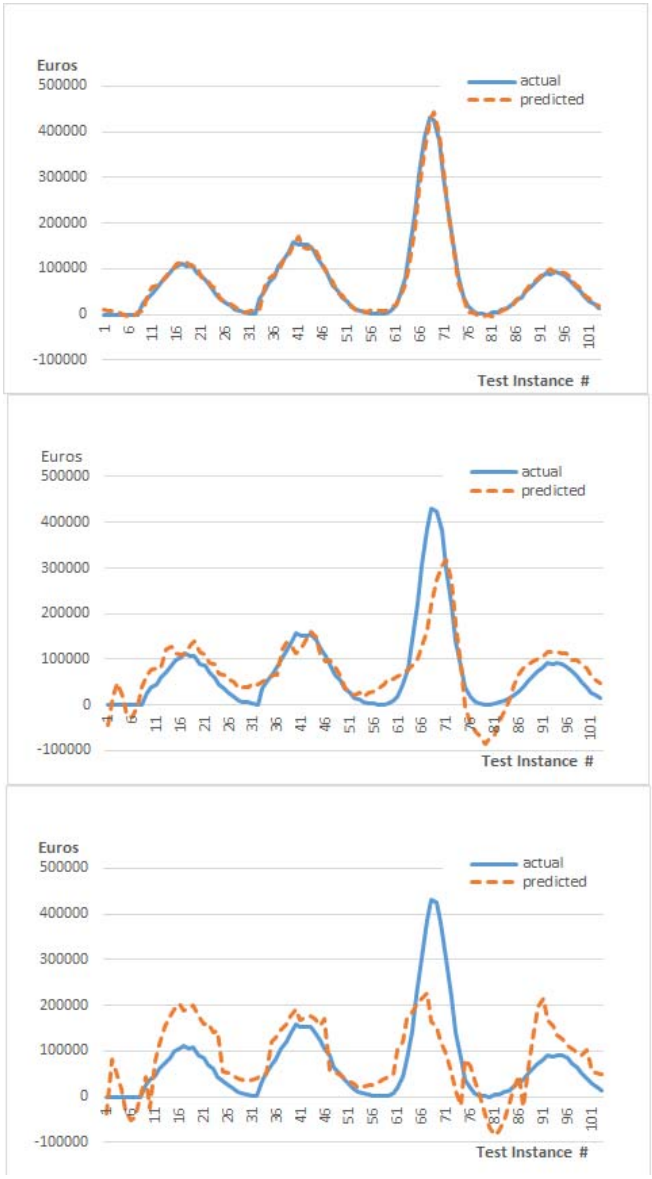


Figure 3 Actual and predicted values of 1 hour petrol consumption. From top to bottom: 1 hour in the future using SMOreg, 4 hour in the future using SMOreg, 8 hours in the future using M5P

5. CONCLUSIONS

In this paper we have reported the results of our study of predicting the amount of euros spent in one hour for a specific purchasing category, using different amounts of historical data. The machine learning algorithm used in the experiments are linear regression, regression tree and support vector machine for regression. The performance measures reported are correlation coefficient and relative absolute error. The results for 1 hour ahead prediction have been very good, while 4 and 8 hours ahead prediction only satisfactory.

Possible improvements for last two cases could be obtained if more than 13 days of data is available. Future work can be conducted in the direction of linking this dataset to other data for the time and geographical region, such as popular event, holidays, transportation.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under PlanetData (ICT-NoE-257641) and NRG4Cast (FP7-ICT-600074).

7. REFERENCES

[1] A Smarter Planet: *http://www.ibm.com/smarterplanet*.

[2] Berry, M.J. a. and Linoff, G.S. 2004. *Data mining techniques: for marketing, sales, and customer relationship management*.

[3] Giffinger, R. et al. 2007. *Smart cities Ranking of European medium-sized cities*.

[4] Wang, Q. et al. 2010. Smarter City: The Event Driven Realization of City-Wide Collaboration. *2010 International Conference on Management of e-Commerce and e-Government*. (Oct. 2010), 195–199.

[5] Witten, I.H. et al. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers.

A TOPOLOGICAL DATA ANALYSIS APPROACH TO THE EPIDEMIOLOGY OF INFLUENZA

Joao Pita Costa and Primož Škraba
Artificial Intelligence Laboratory, Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144; fax: +386 1 4773905
e-mail: joao.pitacosta@ijs.si

ABSTRACT

Influenzanet is a system to monitor the activity of influenza-like-illness [ILI] with the aid of internet volunteers. Topological data analysis [TDA] examines the structure of data and contributes to the development of medicine, studying properties of a continuous space by the analysis of a discrete sample of it. Using TDA we analyze the topology of Influenzanet data identifying noise and distinguishing higher dimension features. This is done both in terms of the overall structure of a disease as well as its evolution. It provides a way to test agreement at a global scale arising from standard local models. We also compare this qualitative method to other quantitative methods such as Fourier analysis or dynamical time warping [DTW].

1 INTRODUCTION

Topological data analysis [TDA] provides us with the topological features that describe the structure of a given point cloud. It infers high-dimensional structure from low-dimensional representations and studies properties of a continuous space by the analysis of a discrete sample of it, assembling discrete points into global structure. The basic technique encodes topological features of a given point cloud by diagrams representing the lifetime of those topological features. A good introduction to topological data analysis can be found in [1]. Recently, these topological methods on data have seen a relevant application to the study of the influenza virus as described in [2].

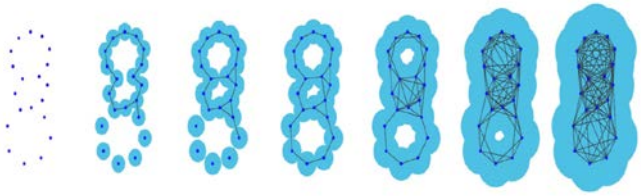


Figure 1. Topological data analysis: the filtration of a simplicial complex of a given pointcloud according to the growing radious of balls centered in the input data points.

The system *Influenzanet* monitors online the activity of *influenza-like-illness* [ILI] with the aid of volunteers via the internet. It has been operational for more than 10 years, and at the EU level since 2008. Influenzanet obtains its data directly from the population, contrasting with the traditional system of sentinel networks of mainly primary care physicians. Influenzanet is a fast and flexible monitoring system whose uniformity allows for direct comparison of ILI rates between countries [5].

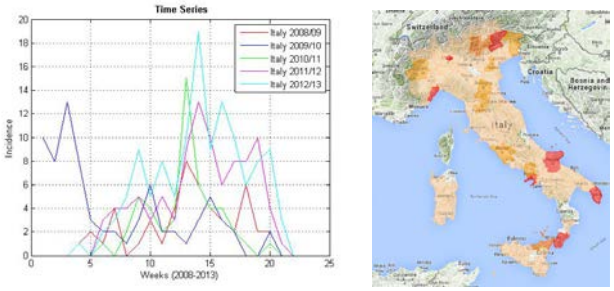


Figure 2. Influenzanet: the time-series for the incidence of influenza in Italy during the flu seasons of 2008-2013 (on the left); a screenshot of the influenzanet system in Italy, taken in May 2015 (on the right).

Our goal with this project is to analyze the Influenzanet data using persistence, identifying topological features relevant to the epidemiological study. To do so, we identify data noise, distinguish higher dimension features and look at the overall structure of the disease as well as its evolution during the flu season in Portugal and Italy. In particular, this provides a way to test agreement at a global scale arising from standard local models.

2 TOPOLOGICAL ANALYSIS OF EPIDEMIOLOGICAL DATA.

The *Mahalanobis distance* is a measure of the distance between a point *P* and a distribution *D*, widely used in cluster analysis and classification techniques. When considering this metric on the space while using TDA, we get a perspective of that space under different scales, where small features will eventually disappear. We have used in [8] several techniques to preprocess the input data, including subsampling and colliding data points that are closer than a

given parameter. In particular, we embed the data in higher dimensions, compute persistence, and look for outliers.

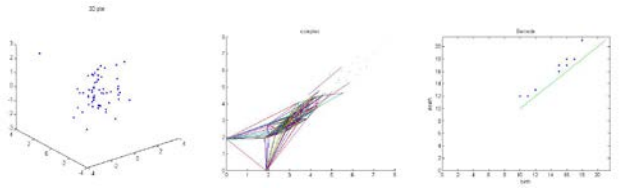


Figure 3. The pipeline for the computation of topological data analysis for the time series of Italy 2009/10: the given pointcloud of the input data (on the left); the Viatoris-Rips complex approximating the space of the pointcloud (in the center); the correspondent persistence diagram encoding the lifetime of the persistent topological features (on the right).

The analyzed data lists the number of active participants and the number of ILI onsets, for three different ILI case definitions of the Influenzanet in Italy for every week in years of the Influenza seasons from 2010/11 to 2012/13. Based on this data we have used several algorithms to preprocess it, prior the construction of the Vietoris-Rips complex that corresponds to the given data. This method permits us to encode the qualitative features of that data into a persistence diagram.

The images in Figure 3 show the cloud of input data points, the corresponding simplicial complex, and persistence diagram for dimension 1. These topological tools complement the information obtained by classical data analysis. The computation of the persistence diagrams is done via Vietoris-Rips complexes using *Perseus*, the open source persistent homology software [4]. The input structure is given as a symmetric distance matrix where the entries come from pairwise distances between points in a given point cloud. In the figures below we can see three steps of the construction of the Vietoris-Rips complex that will provide us with the persistence diagram encoding the topological information of the Influenzanet data.

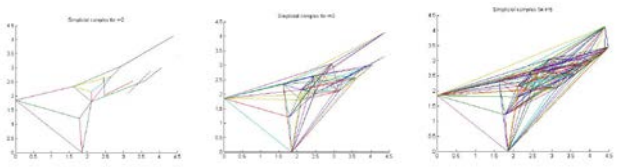


Figure 5. The filtration of the simplicial complex at several levels varying according a parameter *r* for the input time series of Italy in the flu season of 2009/2010: $r = 2$ (on the left); $r = 3$ (in the center); $r = 5$ (on the right).

3. QUANTITATIVE AND QUALITATIVE ANALYSIS OF INFLUENZA.

Fourier analysis is widely used to identify patterns in a time series. We used the time series of the incidence of influenza in Portugal and Italy for the flu seasons of 2008-2013. In Figure 4 we can see the plot of the two time series and their correspondent Fourier transform.

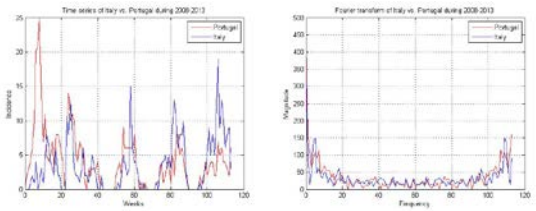


Figure 4. Comparing the flu seasons of Portugal and Italy during 2008-2013: the time series (on the left); the Fourier transform (on the right).

We computed in [9] the Fourier transform for each pair of time series *(country, year)* to compare the flu seasons of Portugal and Italy. In that work we compared the quantitative methods of Fourier analysis with the qualitative methods of TDA. In Figure 7 the reader can see an extension of the results of this comparison with highlighted biggest and smallest values.

When comparing two time series that may vary in time or speed it is usual to apply the algorithm *dynamic time warping* [DTW] measuring the similarity between those temporal sequences. In this study we compared each pair of time series *(country, year)* obtaining the respective measure that can be seen in the table of Figure 7.

The usage of TDA for the analysis of time series was explored in [6] towards the quantification of periodicity and identification of periodic signals in gene expression in [7]. We also use TDA to analyze the input time series data, following an approach developed specifically for Influenza. Barcodes and correspondent persistence diagrams seen as multi-scale signatures encode the lifetime of topological features within pairs of numbers representing birth and death times. We have computed a persistence diagram for each time series *(country, year)* embedded in higher dimensions. As shown by the persistence diagrams below, the distinguishable features are seen in dimension 1.

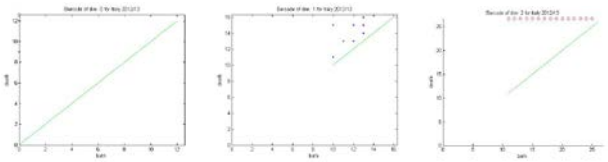


Figure 6. The persistence diagrams for the input time series of Italy in the flu season of 2009/2010: dimension 0 (on the left); dimension 1 (in the center); dimension 2 (on the right). The red circles mean that the lifetime of the considered features does not end.

Persistence landscapes are techniques of TDA that permit us to measure the pairwise distance between persistence diagrams at several different levels. The distance value between these two persistence diagrams in the tables of Figure 7 was calculated using the *persistence landscapes*

toolbox [3] to compute the distance between diagrams considering different norms.

The following tables represent the comparison between the Fourier analysis, dynamical time warping and topological analysis of the incidence of influenza in Italy and Portugal for the flu seasons of 2008-2013.

Fourier		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	**2,2518**	1,523	2,0147	1,1977	0,95957
	2009	1,523	1,0536	1,3576	0,86667	0,74338
	2010	2,0147	1,3576	1,1635	0,70203	0,67165
	2011	1,1977	0,86667	0,70203	0,67071	0,6352
	2012	0,95957	0,74338	0,67165	0,6352	**0,61559**

DTW		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	89	80	20	15	15
	2009	106	32	58	60	58
	2010	75	88	**13**	23	23
	2011	60	92	16	28	35
	2012	44	**111**	35	48	55

TDA		Portugal				
		2008	2009	2010	2011	2012
Italy	2008	2,91548	2,85774	2,25462	2,81366	2,51661
	2009	2,32737	2,27303	1,73205	1,63299	1,75594
	2010	**0,288675**	0,408248	1,22474	1,32288	0,957427
	2011	0,957427	1	1,35401	1,52753	1,32288
	2012	4,09268	4,04145	3,37886	3,7305	**3,58236**

Figure 7. Comparing the flu seasons of Portugal and Italy during 2008-2013: the distance tables for the Fourier analysis (on the top), the dynamic time warping (on the center), and the topological data analysis (on the bottom).

When comparing the distances obtained by Fourier analysis, DTW and TDA we can see that these three methods look at different features of the data.

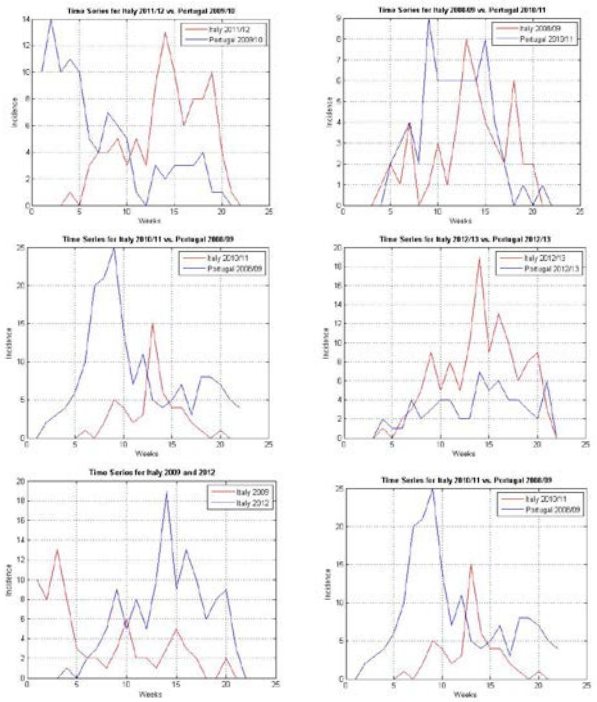


Figure 8. Comparing the flu seasons of Portugal and Italy during 2008-2013: selected plots of time-series to compare the results in the Fourier analysis, the topological data analysis and the dynamical time warping.

The plots in Figure 8 represent time-series for selected flu seasons from 2008 to 2013. They serve us to compare the different data analysis methods used in this study.

When comparing the distances between Italy 2011/12 and Portugal 2009/08, the Fourier provides us with a high value of 92, while DTW analysis has a low value of 0,86667. On the other hand, for the flu seasons of Italy 2008/09 and Portugal 2010/11, the DTW has a low value of 20 while the Fourier analysis has a relatively high value of 2,0147. In the first case the monotony of the curves match, although the periodicity not being close. The second case shows two high peaks for Italy 2008/09 against one for Portugal 2010/11 explaining the low level of DTW.

To compare the quantitative Fourier analysis with the qualitative analysis of TDA we look at the flu seasons of Italy 2010/11 and Portugal 2008/09 where TDA achieved the low value of 0.288675 and the Fourier analysis reached the high value of 2,0147. On the other hand, the flu seasons of Italy and Portugal in 2012/13 reach a high TDA value of 3,58236 and a low Fourier value of 0,61559. The first case shows a big difference of peaks which does not happen in the second case where the periodicity is lower, implying the lower level for the Fourier analysis.

Finally, the comparison between TDA and DTW points us to the flu seasons Italy 2008/09 and Portugal 2012/13, where TDA reached a high value of 2,51661 (due to the higher similarity of peaks) and DTW reached a low value of 15 (describing the different behavior of the curves); and the flu seasons of Italy 2010/11 and Portugal 2008/09, where TDA achieved a low level of 0.288675 (with great difference of peaks as pointed out earlier) and DTW achieved the high value 75 (pointing out the similar behavior of the curves).

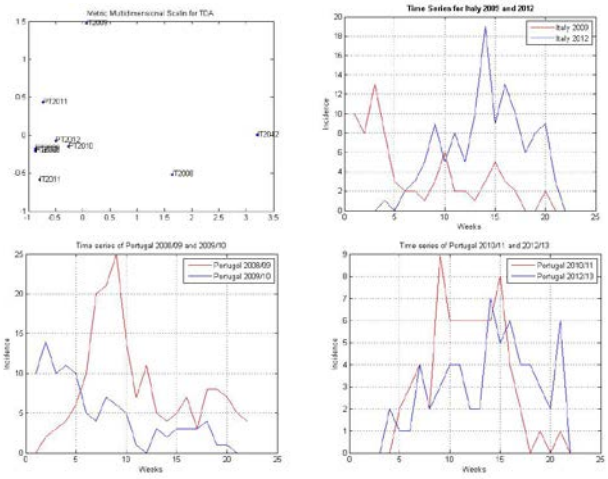


Figure 9. Comparing the flu seasons of Italy and Portugal during 2008-2013 using metric multidimensional scalling (on the upper left) to identify:

the outlier flu seasons of Italy 2009/10 and 2012/13, with time series ploted for analysis and interpretation (on the upper right); the close flu seasons of Portugal 2008/09 and 2009/10 (on the lower left); and the flu seasons of Portugal 2010/11 and 2012/13, close to the diagonal (on the lower right).

We used multidimensional scaling as in Figure 9 to identify outliers for each of the three methods within the flu seasons analyzed in this study. TDA provides a qualitative analysis of the time series of the incidence of influenza, looking in particular at the peaks and dramatic changes. In that perspective, the time series of Italy 2009/10 and 2012/13 plotted in Figure 9 describe very different flu seasons with very different peaks. On the other hand, the flu seasons of Portugal 2008/09 and 2009/10 are identified being very close with very similar peaks, although the behavior of the curve being different. The knowledge on secondary attack rates in the influenza season is of importance to access the severity of the seasonal epidemics of the virus, estimated recently with information extracted from social media in [10]. Here lies a strong point of TDA where it can provide relevant contribution complementing other methods.

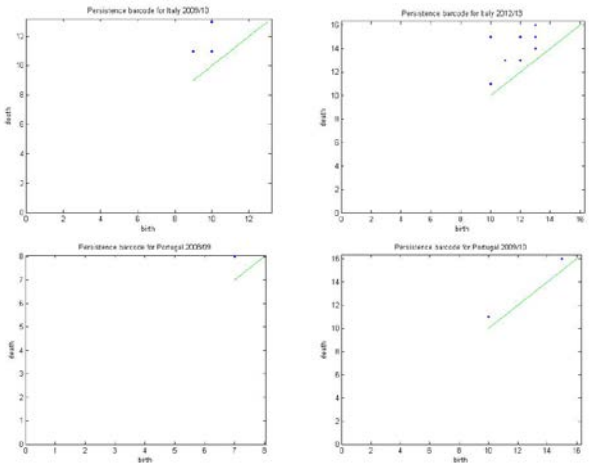


Figure 10. Comparing the flu seasons using persistence diagrams for dimension 1 for: Italy 2009/10 (on the upper left), Italy 2012/13 (on the upper right), Portugal 2008/09 (on the lower left), and Portugal 2009/10 (on the lower right), identified as particular cases in Figure 9.

The persistence diagrams of Figure 10, correspondent to the identified flu seasons of Italy 2009/10 and Italy 2012/13, and Portugal 2008/09 and 2009/10. They encode the lifetimes of the topological features of the curves of the time series of those seasons. Persistence diagrams are a clear and practical tool that allows us the detection of outliers and to capture the qualitative features of the dynamics of the system. These ideas provide a new approach to the analysis of the seasons in the epidemiology of Influenza.

4. CONCLUSION AND FUTURE WORK

The study of Epidemiology is a great source of problems relating to nonlinear systems, large scale data and development of more accurate models, where TDA can contribute, providing high dimension techniques for medical data analysis. In this study we showed how they can be used to analyze the incidence for different ILI case definitions, contributing to a better understanding of the features distinguished by those definitions. The information provided by quantitative methods such as DTW or the Fourier analysis of time series can be complemented by the topological analysis of that data. The examples considered in Figure 8 show that these methods do not express the same information about the development of the epidemics during the flu season. The knowledge provided by each of these methods complements the knowledge coming from the other methods and can be put together in a global information map. Further research considers the analysis of the impact of the qualitative aspects of TDA for modeling and prediction of the current Influenza season. We will also use state of the art artificial intelligence methods to learn metrics more appropriate to the input time series data aiming, to grasp a better understanding of the severity of the epidemics both in past seasons and during the ongoing season.

ACKNOWLEDGMENTS

The authors would like to acknowledge that his work was funded by the EU project TOPOSYS (FP7-ICT-318493).

REFERENCES

[1] G. Carlsson (2009). Topology and data. Bulletin of the American Mathematical Society 46.2 : 255-308.

[2] J. M. Chan, G. Carlsson and R. Rabadan (2013). Topology of viral evolution. Proceedings of the National Academy of Sciences 110.46: 18566-18571.

[3] P. Dlotko (2014). Persistence Landscapes Toolbox (www.math.upenn.edu/ dlotko).

[4] V. Nanda (2014). Perseus (www.sas.upenn.edu/˜ vnanda/perseus).

[5] D. Paolotti et al (2014). Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection 20.1: 17-21.

[6] J. A. Perea and J. Harer (2013). Sliding windows and persistence: An application of topological methods to signal analysis. Foundations of Computational Mathematics 15.3: 799-838.

[7] J. A. Perea et al (2015). SW1PerS: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data. BMC bioinformatics 16.1: 257.

[8] J. Pita Costa and P. Škraba (2014). A topological data analysis approach to epidemiology. European Conference of Complexity Science 2014.

[9] J. Pita Costa and P. Škraba (2015). Topological epidemiological data analysis. ACML Health 2015.

[10] E. Yom-Tov et al (2015). Estimating the Secondary Attack Rate and Serial Interval of Influenza-like Illnesses using Social Media. Influenza and other respiratory viruses. DOI:10.1111/irv.12321

Event Detection in Twitter With an Event Knowledge Base

Luis Rei, Marko Grobelnik and Dunja Mladenić
Jožef Stefan Institute and Jožef Stefan Postgraduate School
Jamova cesta 39, 1000 Ljubljana, Slovenia
Tel: + 386 14773528; fax: + 386 14773851
e-mail: {luis.rei, marko.grobelnik, dunja.mladenic}@ijs.si

ABSTRACT

We present two methods for event detection in *Twitter* using an event knowledge base. The knowledge base used contains world events reported in the media that we identify as multi-lingual clusters of mainstream news stories. Given this fact, we reduce the problem of event detection to matching tweets to mainstream news stories. The first method consists of using URLs to mainstream news sites present in tweets and in the knowledge base. We use this method to build a supervised corpus of tweets and then create and evaluate a supervised classifier as our second method. Experimental evaluation on real-world data shows that the proposed methods perfom well on our dataset.

1 INTRODUCTION

Twitter is a microblogging social network service with 316 million monthly active users who together generate an average of 500 million messages called tweets per day as of June 2015[1]. *Twitter* users publish tweets about any topic and in any language they choose with a limit of 140 characters per tweet. Because of its large number of active users, it's massive volume of data and the fact that most published tweets are publicly accessible as opposed to other social networks where messages are traditionally restricted to friends, *Twitter* is often used for research.

The definition of event as been subject to academic discussion with different authors adopting slightly different definitions. It is generally agreed upon that an event should be defined as a real-word occurrence over a specific period of time and in a specific location [1]. We adopt that definition and restrict this work to *significant events* as defined in [1] where an event is significant if it may be discussed in traditional media.

The problem of event detection in streams has often been tackled using stream clustering and topic modelling techniques [2]. Stream clustering is the approach used by Event Registry[2] [3]. Social Media streams in general and *Twitter* in particular pose a bigger challenge to traditional event detection techniques: 1) high volume 2) a high degree of non-relevant messages *("meaningless bables")* [4] 3) reduced context for textual based methods as social media messages are usually much shorter than tradional news articles, with *Twitter* messages being limited to 140 characters. Our approach instead relies on the existence of an event knowledge base. Event Registry is one such knowledge base, automatically created from news articles retrieved by newsfeed [5] which collects content from more than 100,000 news sources worldwide with between 100,000 and 150,000 news articles collected daily. Events in Event Registry consist of a multi-lingual cluster [6] of news articles as well as information extracted from them such as named entities, categories and keywords. Since most of the topics discussed on Twitter are also mainstream news [7] and this also corresponds to our definition of significant, the choice of knowledge base seems optimal. Furthermore, the multi-lingual nature of the event information helps us to create mostly language independent multi-lingual methods. Finally, once we match a tweet to an event we can immediately obtain more context for the event. The obvious downside is that we can only detect events already present in the knowledge base.

In Section 2 we present our URL based matching strategy while in Section 3 we present our content based supervised classifier method. Section 4 contains the details of our dataset. Section 5 contains the results of our supervised classifier and Section 6 our final remarks.

2 URL MATCHING

In URL based matching we look for URLs in tweets and compare them to URLs in our knowledge base from Event Registry. If a tweet contains a URL that matches the URL of an article in an event, we can say that the tweet is related to that article and thus to the event that contains the article.

This task is made slightly more challenging than simple string matching by the fact that the relationship between an article and a URL is often one-to-many. The most visible case is when URL shorteners are employed, where a different shorter domain is used in conjunction with a short code which then redirects to the longer URL[3]. Another case is when the URL for the article changes and the old URL redirects to a new one using the HTTP response status code 301 Moved Permanently [8]. It is also common for URLs to contain tracking query strings such as

[1] Twitter, https://about.twitter.com/company
[2] Event Registry: http://eventregistry.org

[3] For example, *http://alj.am/1OAQ9K9* directs the user who clicks on that link to *http://america.aljazeera.com/articles/2015/3/29/nashvilles-boom-pricing-out-middle-and-lower-class.html*.

http://www.example.org/1?utm_campaign=ex or query strings that specify viewing options such as *http://www.example.org/1?page=1*. Furthermore publishing software often makes the same content available under different URLs based on the category structure, e.g. *http://example.com/politics/new-economic-policy/* and *http://example.com/economy/new-economic-policy/* . As a final example, several string level differences can be configured to be ignored by web server software to allow users to reach the right content even when they do not enter the exact string into their browsers, e.g. http://www.example.com/article can be the point to the same article as http://www.example.com/article/ .

To avoid being penalized in search engines rankings for content duplication, many publishers implement either HTTP redirection or the canonical link element [9]. The canonical link element commonly referred to as the canonical tag, is an HTML *<link>* element with the attribute *rel="canonical"* that can be inserted into the <head> section of an article (or any web page) e.g. <link rel="canonical" href="http://example.org/article/new-economic-policy/" />. It is also possible for the canonical link to be present in the HTTP headers instead of the HTML source. It is also common for publishers to implement the Open Graph Protocol [10] which allows for better integration with Facebook and requires a *<meta>* tag with the property *og:url* containing the article's canonical URL.

For each URL in newsfeed and in a tweet we make a request to it, taking note of any redirection, analyzing the headers and processing the HTML response body to attempt to obtain some of the most likely alternative URLs for the content. Each article is associated with a list of URLs used to reference it and each tweet is associated with a list of URLs mentioned in it. We use these lists to match the two. The order of precedence used is:

1. canonical tag/header;
2. open graph *og:url* property;
3. redirection;
4. the original URL;
5. the original URL with or without a trainling slash depending on wether it was not originally present or if it was.

We have opted at present not to address other issues with URLs such as removing query parameters and other URL normalization techniques that can introduce false positives.

3 CONTENT MATCHING

Not all tweets that refer to an event will include a URL to a news story about the event. Thus a different strategy is necessary to match those tweets to events. This is accomplished based on textual similarity between the tweet and events.

Each event within Event Registry contains a cluster of news articles, we select the medoid news article for each language in the event, i.e. the most representative news article for a given language, and use its text to compare to the tweet in that language. If the event does not have a news article in the tweet's language, it is not considered for matching with that tweet.

The problem of matching an article and a tweet is treated as a supervised binary classification problem where given an article and a tweet our classifier must answer if they 'match' or not.

3.1 Preprocessing and Feature Extraction

Text in articles and tweets is preprocessed similarly. Each document is preprocessed according to the following steps:

1. converted to lower case;
2. all URLs are removed;
3. all non-alphanumeric characters are removed (including punctuation and the hashtag symbol);
4. all characters are converted to their Unicode normal form [11];
5. the text is tokenized based on whitespaces;
6. stopwords are removed.

All tweets which after this preprocessing have less than 4 tokens are discarded. Once a document has been preprocessed, we generate its unigrams, bigrams trigrams and quadgrams i.e. its n-grams where $n \in [1, 4]$. For news articles, the title and the body are processed separately.

For each article-tweet pair and for each $n \in [1, 4]$ we generate a similarity vector containing different measures of similarity between their n-grams [12]:

1. the Jaccard similarity between the title of the article and the tweet;
2. the number of common terms between the tweet and the body of the article multiplied by logarithm of the number of terms in tweet;
3. the Jaccard similarity between the body of the article and the tweet;
4. the cosine similarity between the body of the article and the tweet.

3.2 Classifier

We used a linear Support Vector Machine (SVM) as our binary classifier and then performed a random 50-50 split on the dataset into development and test subsets. Using precision as our scoring function, we performed parameter tuning using grid search with 5 fold cross evaluation on the development set for the penalty parameter (C) and class weight hyper-parameters, arriving at C=10 and a positive class weight 0.6 (negative class weight was kept fixed at 1). The positive class weight value multiplies C, since it is lower than 1 it allows the SVM to learn a decision function that makes more misclassifications of positive examples. In particular, more false negatives.

3.3 Language Dependencies

While this classifer is mostly language independent, a few aspects are note. Chief among them is the whitespace based tokenization which while we can expect to work equally well across European languages, will not work at all for Asian languages that do not use whitespaces.

The next, more subtle problem is Unicode normalization. We can expect it to perform better in language in which this normalization corresponds to the way people write on social media than in languages where it does not. For example, in languages which use graphical accents, this normalization step removes them and uses simply the corresponding vowel or consonant. It is common for social media users to also eschew the use of graphical accents. In French and Portuguese, for example, this normalization step matches social media users exactly. However in German, Umlaute are instead commonly replaced in social media with the corresponding vowel followed by an "e". So ä is replaced by ae, ö by oe. This does not match Unicode normalization and thus we can expect worse results from this performing this step in German than we would in Portuguese.

The final language dependency is stopword removal. We rely on the availability of stopword lists compiled by language experts. These may not be available for all languages and/or their quality can vary. It is theoretically possible to generate these lists automatically however we have not taken this step or performed any comparison.

4 DATASET

In order to treat our problem as a supervised classification problem we must first create a supervised dataset. The URL matching described in Section 2 was used on historical data to create the positive examples dataset. The negative examples are generated by pairing tweets that have been matched by this method to a specific event with a different event. We discarded from the dataset any article-tweet pair with a zero similarity vector in both positive and negative examples except for 1 in the negative examples. The number negative examples generated matches the number of positive examples used *i.e.* the dataset is balanced. The total number of examples in the dataset we generated was 32,372 tweet-event pairs.

This dataset generation process supports our goal of obtaining a high precision classifier: the classifier is trained almost exclusively with the hard cases: negative examples that share some similarity with the article. In practice these are actually an extremely small minority of all possible negative examples, since most tweets do not share any similarity with a given article. It also underlies the fact that by relying on simple textual similarity between a tweet and a single article in an event ensures that many true positives are disregarded since they will also have a 0 similarity vector.

4 RESULTS

Our classifier obtained an AUC score of 0.91 on our dataset. The Precision-Recall curve is shown in Figure 1*: Precision Recall Curve* and the Precision-Recall vs Threshold curve is shown in Figure 2*: Precision-Recall vs Threshold*.

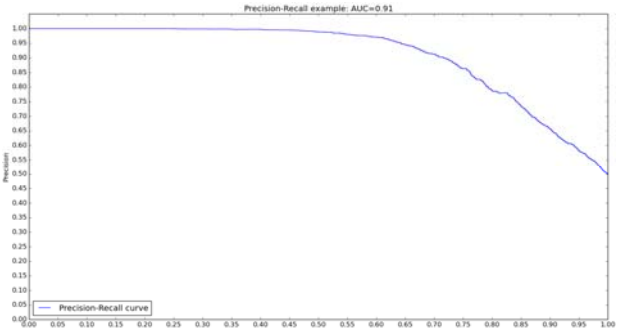


Figure 1: Precision Recall Curve

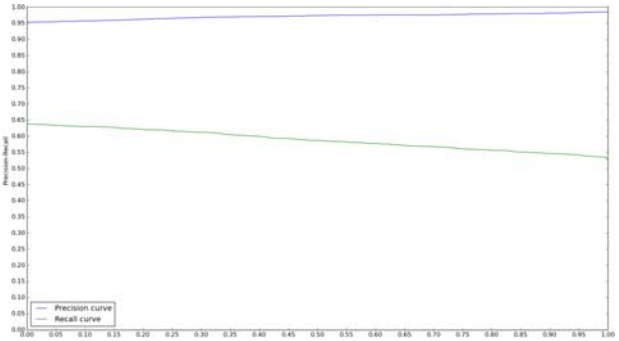


Figure 2: Precision-Recall vs Threshold

We can see that if we chose a threshold near 1, our classifier has nearly 100% precision while lowering our classifiers recall to nearly 55%. The procedure for generating our dataset introduces a huge bias into our evaluation: in reality, we will have many more false negatives since many true positives have zero similarity vectors and will thus become false negatives (those cases were discarded in our dataset). Real recall can be expected to be much lower than the recall on our test dataset. We can however expect that this will be partially offset by an expected redundancy in social media messages and a large volume of messages regarding events. The emphasis on precision over recall in our work is also understandable in the context of future applications. The end use of such a classifier is likely to be either to directly show tweets to end users of a web site or to give a social dimension to the analysis of events. In either case, the loss of tweets from a sample seems preferable to either showing the wrong tweets to an end user or to reduce the accuracy of social media analysis with respect to events under analysis.

6 FINAL REMARKS

Event Registry adds between 5000 and 40000 events to its database every day. Considering also the daily volume of tweets, even if we consider only the public twitter stream which contains only 1% of all tweets, we are looking at an estimated lower bound of 25M daily possible tweet-event pairs. This number becomes considerably more problematic if we add a reasonable window of 6 days around the

publishing of a tweet when considering which events to match it to. While URL matching is computationally cheap and a classification algorithm can also be considered computationally cheap, feature extraction is not so cheap. Thus, running a classifier against event-tweet pairs in practice should be restricted to a subset of all possible event-tweet pairs that are considered good candidates. Fortunately, since the classifier relies exclusively on textual similarity, we can rely on decades of research and development in Information Retrievel and databases to provide a set of good candidates efficiently.

We consider the biggest contributions of this work to be the use of a knowledge base for event detection in social media and the introduction of a fully automated technique for generating a supervised event detection dataset.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under Symphony (FP7-ICT- 611875).

References

[1] A. J. McMinn, Y. Moshfeghi and J. M. Jose, "Building a large-scale corpus for evaluating event detection on twitter," in *Proceedings of the 22nd ACM international conference on Conference on information \& knowledge management*, 2013.

[2] C. C. Aggarwal and K. Subbian, "Event Detection in Social Streams," *SDM,* vol. 12, pp. 624--635, 2012.

[3] G. Leban, B. Fortuna, J. Brank and M. Grobelnik, "Event registry: Learning about world events from news," in *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, 2014.

[4] J. Weng and . B.-S. Lee, "Event Detection in Twitter," in *ICWSM*, 2011.

[5] M. N. B. Trampuš, "Internals Of An Aggregated Web News Feed," in *SiKDD*, Ljubljana, Slovenia, 2012.

[6] J. Rupnik, A. Muhic and P. Skraba, "Multilingual Document Retrieval Through Hub Languages," in *Proceedings of the Fifteenth International Multiconference Information Society*, 2012.

[7] H. Kwak, C. Lee, H. Park and S. Moon, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World wide web*, 2010.

[8] R. G. J. M. J. F. H. M. L. B.-L. T. Fielding, "Request for Comments: 2616: Hypertext Transfer Protocol -- HTTP/1.1," Network Working Group, The Internet Society, 1999. [Online]. Available: https://tools.ietf.org/html/rfc2616. [Accessed 2015 March 10].

[9] M. K. J. Ohye, "Request for Comments: 6596," Internet Engineering Task Force (IETF), April 2012. [Online]. Available: http://tools.ietf.org/html/rfc6596. [Accessed 15 March 2015].

[10] Facebook, "The Open Graph Protocol," 20 October 2014. [Online]. Available: http://ogp.me/. [Accessed 10 March 2015].

[11] M. Davis and K. Whistler, "Unicode Standard Annex 15: Unicode Normalization Forms," The Unicode Consortium, 2015.

[12] P. Saleiro, L. Rei, A. Pasquali, C. Soares, J. Teixeira, F. Pinto, M. Nozari, C. Felix and P. Strecht, "POPSTAR at RepLab 2013: Name ambiguity resolution on Twitter," in *CLEF 2013 Eval. Labs and Workshop Online Working Notes*, 2013.

A MULTI-SCALE METHODOLOGY FOR EXPLAINING DATA STREAMS

Luka Stopar
Jozef Stefan Institute and Jožef Stefan International Postgraduate School,
amova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 477-53-61
e-mail: luka.stopar@ijs.si

ABSTRACT

This paper presents a novel, multi-scale, framework, for the simultaneous analysis of multiple data streams, called StreamStory. The framework models the data streams as a hierarchical Markovian model by automatically learning states and transitios, and aggregating them into a hierarchy of Markov chains. This approach aims to compensate the gap between low-level streaming observations and high-level output/alerts which provide a value for higher levels of streaming data analysis, like inference and prediction, and provides ground for qualitative interpretation of the data.

1 INTRODUCTION

Sensory systems typically operate in cycles with a continuously time-varying component. Such systems can be characterized by a set of states, along with the associated state transitions. These states, on a high level, may include a "day" state, "night" state or maybe states with high and low productivity. For example when a pilot of an aircraft wishes to change the aircrafts heading, they will put the aircraft into state "banking turn" by lowering one aileron and raising the other, causing the aircraft to perform a circular arc. After some time, the wings of the aircraft will be brought level by an opposing motion of the ailerons and the aircraft will go into the "level" state.

Such high-level states can further be split into lower-level states, giving us a multi-resolution view of the system, allowing us to observe the system on multiple aggregation levels. For example, a "banking turn" state can be split by the aircrafts roll and angular velocity, resulting in perhaps three states: "initiate turn", "full turn" and "end turn".

StreamStory models the monitored system as a hierarchical Markovian process by automatically learning the typical lowest-level states and transitions, and aggregating them to obtain a hierarchy of Markovian processes. Such a model allows users to observe the monitored system in a unique way and provide a understanding of its dynamics.

Furthermore, we divide the input streams into two sets: observation set and control set. The observation set of paramters are the paramters that tell us the state of the system and, we assume, cannot influence its dynamics. These are parameters that users cannot directly manipulate, like aircraft tilt. They are used to identify, and aggregate, low-level states, detect outliers (anomalies) and determine the current state of the system. In contrast, users can directly manipulate parameters in the control set. These are parameters like the angle of each aileron, and may directly influence the behavior (observation set) and performance of the system. For example when an operator in a steel factory sets the cooling temperature to a high value, the product will take longer to go from state "extremely hot" to state "warm". As such, control parameters may also influence the occurrence, and expected time, of undesired states that may be associated with some undesired event. Our approach uses control parameters to model state transitions, allowing us to observe the dynamics with respect to the current configuration and gives the user insight into the expected bynamics before changing the configuration.

To implement the framework, we subdivide it into four components: (1) state identifier, (2) hierarchy modeler, (3) state assistant and (4) transition modeler, each responsible for its own subtask and explained in the next section.

The remainder of this paper is structured as follows. Section 2 presents on overview of the systems architecture. The user interface is presented in section 3 and, finally, we provide conclusions and acknowledgements in sections 4 and 5 respectively.

2 STREAMSTORY ARCHITECTURE

This section presents the architecture of the StreamStory system. StreamStory operates in two modes: offline and online. In offline mode the system consumes a batch of the data streams and learns its hierarchical Markovian representation (identifies states, constructs a hierarchy and models transitions). Once the model is learned, it can be applied to the data streams in real-time and offers prediction and anomaly detection services. The component also includes a web-based user interface (UI), where the users can explore and interact with the model.

We continue our discussion with a picture of the architecture in Figure 1.

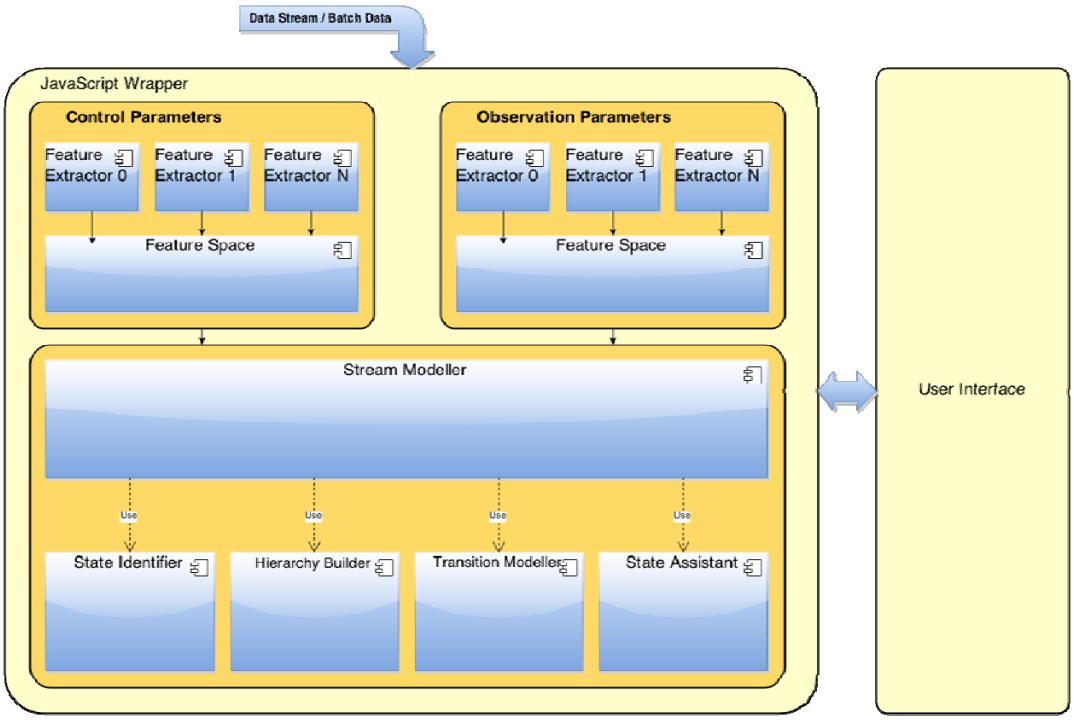


Figure 1: StreamStory architecture.

As shown, the component is a JavaScript wrapper arount two feature spaces (one for each set of parameters) and a stream modeler component, and interacts with the user interface using RESTful web services.

The two feature spaces are responsible for the transformation of the data into (and out of) feature vectors, later used by the algorithms. Each feature space consists of several feature extractors (one for each parameter) which are responsible for transforming a single parameter into a form suitable for machine learning algorithms. The responsibility of the feature space is then to concatenate the ouputs of all the feature extractors into a single feature vector.

The third component is the stream modeler, which is the core component of the system. It is responsible for state identification and assistance, modeling the hierarchy, prediction and anomaly detection. The modeler delegates these tasks to four components: state identifier, hierarchy builder, state assistant and transition modeler, which are explained in the following subsections.

State Identifier

The first of these components is the state identifier. In offline mode it is responsible for the identification and construction of the lowest level states. Once these states are constructed it computes and stores their statistics for further use in the user interface (UI) and detection of anomalies. In online mode, the state identifier is used to identify the current state of the system, from the feature vectors, and for low-level anomaly detection.

To identify the states, the state identifier uses the DPMeans algorithm [1]. The algorithm takes one input parameter λ, which represents the maximum goemetrical radius of each state and is used to control the number of output states. The algorithm is very similar to K-Means [2]. It starts by randomly selecting one of the feature vectors as the initial centroid. In each iteration, it then assigns the feature vectors to their nearest centroid, but unlike K-Means if the vector is outside the radius of all the centroids, it is used to form a new centroid.

When used in online mode, the state identifier is responsible for identifying the state to which the current feature vector belongs. This is done by assigning the feature vector to the state with the nearest centroid. If the feature vector falls outside the radious of all the states, it is marked as an anomaly.

Hierarchy Modeler

Once the low-level states are identified, they are aggregated into a hierarchy. This is precisely the task of the hierarchy modeler. The hierarchy modeler consumer a set of states (centroids), aggregates them into a hierarchy and stores it in two arrays. The first array encodes the topology of the hierarchy, by storing at index *i* the index of *i*-ths parent. The second array stores the hight (level) of each state.

To compute the topology, the hierarchy modeler uses one of several agglomerative clustering strategies: single link, complete link or average link [3].

In online mode the hierarchy modeler is responsible for two tasks: (1) given the current lowest level state, finding its

parent states on specific levels and (2) given a state in the middle of the topology, finding all its lowest level successors later used by the transition modeler.

State Assistant

The state assistant is responsible for assisting the users in identifying the meaning of states. For example, when the user clicks on a state in the UI, the state assistant highlights the attributes that are most typical for the state. This is achieved by extracting weights of individual features from a logistic regression model [4] learned by classifying feature vectors of one state (positive label) against feature vectors of all the other states (negative labels). To balance the class distribution, the component samples the larger set.

Transition Modeler

The final component in our framework is the transition modeler. As the name suggests, the transition modeler models transitions between states. It does so by using a continuous time Markov Chain framework [5]. On the lowest level, the model is defined with a transition rate matrix Q, where the element at posisiton (i,j) represents the rate of going from state i to state j.

To be able to model the hierarchical dynamics, we need to be able to compute the transition rate matrix for Q_l on each level l of the hierarchy. The transition modeler only stores the lowest-level transition matrix and uses it to construct higher level matrices on the fly. To achieve this, it needs to know which low-level states to aggregate. It gets this information in the form of state-sets $\{S_i\}_i$, where each state-set corresponds to an aggregated state on level l. It then computes the transition rate matrix on level l using the following formula:

$$Q_{S_i S_j} = \frac{\sum_{k \in S_i} \pi_k \sum_{h \in S_j} Q_{kh}}{\sum_{k \in S_i} \pi_k}$$

Where $\pi = (\pi_k)_k$ is the stationary distribution of the lowest level Markov chain and is found as a normalized non-trivial solution to the following system of equations: $\pi Q = 0$.

3 USER INTERFACE AND USER INTERACTION

This section presents the user interface (UI) of the StreamStory system. The UI is designed to allow the user to observe the dynamics of the monitored system, as well as its current state, and allow the user to configure the underlying model, prediction and anomaly detection services. It visualizes the monitored system as a hierarchy of states, along with associated transitions, and offers several services that allow the user to identify the meaning of states as well as a messaging service which displays notifications about predictions and anomalies.

Figure 2 shows a screenshot of our web-based user interface. The UI consists of four main components: visualization component, state information component, notifications component and model configuration component. These will be explained in the next subsections.

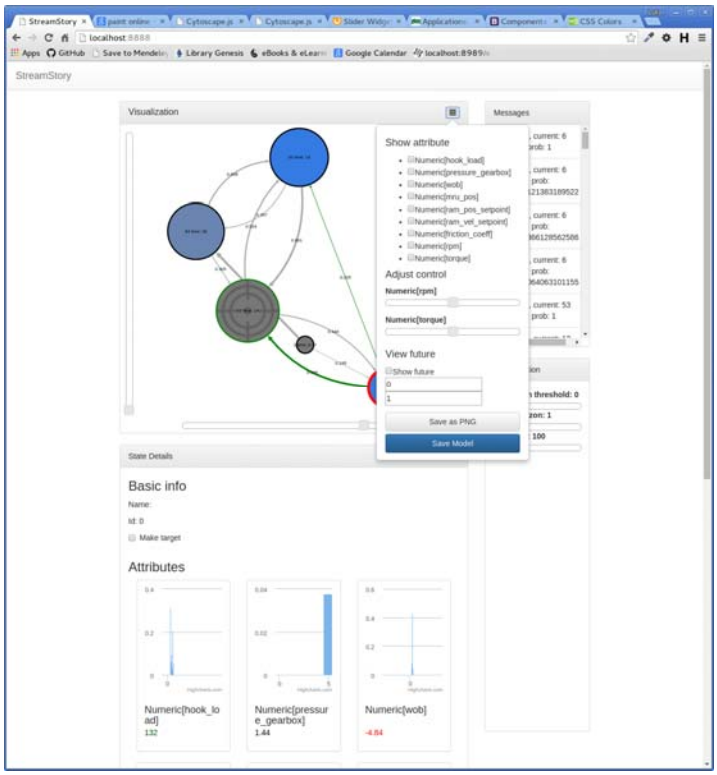


Figure 2: StreamStories web-based user interface .

Visualization Component

The visualization component is the main component of the UI, as it visualizes, and allows the end-users to explore and interact with the model.

States and transitions are represented by a circles and arrows respectively. The size of a circle is proportional to the fraction of time that the system spends in the associated state. Each state displays its identifier (or name if already defined) and the average time spent in the state after arrival. The current and previous states are highlighted with a green and red border respectively, and the most likely future states have a blue background, with the blue component proportional to the probability of jumping into the state. The states with a target in the background are "target" states and the state with the bold bolder is the selected state and its details are shown in the state details component, described in the next subsection. When the system changes states, the UI is automatically updated through web sockets. The thickness of each arrow is proportional to the probability of the corresponding transition, also displayed in the middle of the arrow.

When first opening the UI, the model is shown at the top level with only 2-3 states. The user can use the scroll function to scroll into lower-level states. When scrolling up/down the hierarchy, the states are automatically merged/split.

A popup menu in the top right corner of the component allows the user to:

- Select a target feature: When selecting a target feature, all the states are coloured proportionally to the mean value of the feature in that state. For example, when selecting ambient temperature, the states with higher temperature will become greener than states with lower temperature.
- Observe state probabilities at future times: When moving the spinner at the bottom of the menu states get colored proportionally to the probability of the system being in that state at the appropriate future/past time.
- Simulate control paramters: Using the first group of sliders, the user can simulate a change in parameters in the control set. When adjusting one of the parameters, the transition probabilities are recalculated along with the corresponding holding times and the component automatically redrawn.

State Details Component

As the name suggests, the state details component shows detailed information about the selected state. It provides basic information, like name and id, as well as more detailed information like: the average values of parameters in the state along with their distribution (shown as histograms) and the most typical parameters for the state, which are highlighted (red or green) according to their relevance.

The state details component also allows the user to mark the state as a "target". When a state is marked as a target, notifications about the expected arrival times are displayed in the notifications component presented next.

Notifications Component

The notifications component displays messages to the end-user. These messages include information about the detected anomalies and predictions of arrival into target states. When detailed information about the message is available, the message is clickable and, upon clicking, its details are shown in a popup window.

Model Configuration Component

The model configuration component allows the user to modify the parameters of the underlying model. In the current version, the user can adjust parameters corresponding to the prediction of target states. These include the prediction probability threshold, and the time horizon usied in the calculation of the probability.

4 CONCLUSION

In this paper we presented a novel system for modeling and visualizing data streams called StreamStory. The StreamStory system integrates several machine learning algorithms to model the incoming data streams as a hierarchical Markovian process. As such the system supports several functionalities, including: future state extrapolation, anomaly detection and allows the users to uniquely interprate the stream structure. The system includes a web-based user interface which allows the interaction and exploration of the model.

5 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT Programme of the EC under project ProaSense (FP7-ICT-2013-10-612329).

REFERENCES

[1] K. Brian and M. I. Jordan, "Revisiting k-means: New Algorithms via Bayesian Nonparametrics," in *Preceedings of the 29th International Conference on Machine Learning*, 2012.

[2] I. H. Witten, E. Frank and M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Burlington: Elsevier Inc., 2011.

[3] F. Murtagh and P. Contreras, "Algorithms for Hierarchical Clustering: An Overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 2, no. 1, pp. 86-97, 2012.

[4] A. J. Dobson, An Introduction to Generalized Linear Models, Boca Raton: Chapman & Hall/CRC, 2002.

[5] J. Norris, Markov Chains, Cambridge: Cambridge University Press, 1997.

Inverted Heuristics in Subgroup Discovery

Anita Valmarska
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39
Ljubljana, Slovenia
anita.valmarska@ijs.si

Marko Robnik-Šikonja
Faculty of Computer and Information Science
Večna pot 113
Ljubljana, Slovenia
marko.robnik@fri.uni-lj.si

Nada Lavrač
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Jamova 39
Ljubljana, Slovenia
nada.lavrac@ijs.si

ABSTRACT

In rule learning, rules are typically induced in two phases, rule refinement and rule selection. It was recently argued that the usage of two separate heuristics for each phase—in particular using the so-called inverted heuristic in the refinement phase—produces longer rules with comparable classification accuracy. In this paper we test the utility of inverted heuristics in the context of subgroup discovery. For this purpose we developed a DoubleBeam subgroup discovery algorithm that allows for combining various heuristics for rule refinement and selection. The algorithm was experimentally evaluated on 20 UCI datasets using 10-fold double-loop cross validation. The experimental results suggest that a variant of the DoubleBeam algorithm using a specific combination of refinement and selection heuristics generates longer rules without compromising rule quality. However, the DoubleBeam algorithm using inverted heuristics does not outperform the standard CN2-SD and SD algorithms.

1. INTRODUCTION

Rule learning is one of the earliest machine learning techniques and has been used in numerous applications [5]. It is a symbolic data analysis technique whose aim is to discover comprehensible patterns or models of data [10]. The key advantage of rule learning compared to other statistical learning techniques is its inherent simplicity and human comprehensible output models and patterns.

Symbolic data analysis techniques can be divided into two categories. Techniques for *predictive induction* produce models, typically induced from class labeled data, which are used to predict the class of previously unseen examples. The second category consists of techniques for *descriptive induction*, where the aim is to find comprehensible patterns, typically induced from unlabeled data. There are also descriptive induction techniques that discover patterns in the form of rules from labeled data, which are referred to as *supervised descriptive rule discovery* approaches [10]. Typical representatives of these techniques are contrast set mining (CSM) [2], emerging pattern mining (EPM) [4], and subgroup discovery (SD) [9, 16].

The task of subgroup discovery is to find interesting subgroups in the population i.e., subgroups that have a significantly different class distribution than the entire population. The result of subgroup discovery is a set of individual rules, where the rule consequence is a class value. The main difference between learning of classification rules and subgroup discovery is that the latter induces only individual rules of interest, revealing interesting properties of groups of instances, and not necessarily forming a rule set covering the entire problem space, which is required for classification.

An important characteristic of subgroup discovery task is a combination of predictive and descriptive induction. It provides short and understandable descriptions of subgroups regarding the property of interest. This feature of subgroup discovery has inspired many researchers to investigate new methods that will be more effective in finding more interesting patterns in the data. Most subgroup discovery approaches build on classification algorithms, e.g., EXPLORA [9], MIDOS [16], SD [6], CN2-SD [13], and RSD [14], or on the algorithms for association rule learning, e.g., APRIORI-SD [8], SD-MAP [1], and Merge-SD [7].

In rule learning, during the process of rule construction, conditions that optimize a certain heuristic are added. Typically, the heuristics are used in two different phases of the process: (i) to evaluate *rule refinements*, i.e., to select which of the refinements of the current rule will be further explored, and (ii) for *rule selection*, i.e., to decide which of the refinements that have been explored is added to the rule set. Stecher et al. [15] proposed using separate heuristics for each of the two rule construction phases. In the rule refinement phase they proposed to use the *inverted heuristics*, i.e., the heuristics whose isometrics are rotated around the base rule. These heuristics are used to evaluate the relative gain obtained by the refinement of the current rule.

In this paper we test the utility of inverted heuristics in the context of subgroup discovery. For this purpose we developed a DoubleBeam subgroup discovery algorithm that allows for combining various heuristics for rule refinement and selection. The algorithm was experimentally evaluated on 20 UCI datasets using 10-fold double-loop cross validation.

This paper is organized as follows. In Section 2 we present the findings of Stecher et al. about the use of inverted heuristics in the rule learning process. Section 3 presents the DoubleBeam subgroup discovery algorithm. In Section 4 we describe the data sets used, followed by the empirical evaluation and the obtained results. Finally, in Section 5 we present our conclusions and ideas for further work.

2. INVERTED HEURISTICS

Rule learning algorithms rely on heuristic measures to determine the quality of induced rule. Stecher et al. [15] propose distinction between rule refinement and rule selection heuristics in inductive rule learning. They argue that the nature of the separate-and-conquer rule learning algorithms opens up a possibility to use two different heuristics

in two fundamental steps in the process of rule learning - rule refinement and rule selection. They show in the coverage space why it is beneficial to separate the evaluation of candidates for rule refinement and the selection of rules for the final theory. The rule refinement step in a top-down search requires *inverted heuristics*, which, in principle, results in better rules. Such heuristics evaluate rules from the point of the current base rule, instead of the empty rule. They adapt three standard heuristics with slightly different but related properties:

- ***Precision:***

$$h_{prec}(p,n) = \frac{p}{p+n}; \tag{1}$$

- ***Laplace:***

$$h_{lap}(p,n) = \frac{p+1}{p+n+2}; \tag{2}$$

- ***m-estimate:***

$$h_{m-est}(p,n,m) = \frac{p+m\cdot\frac{P}{P+N}}{p+n+m}. \tag{3}$$

Parameters p and n denote the number of positive and negative examples in a potential subgroup, respectively, and regarding the target class of interest.

For the purpose of rule refinement an inverted heuristic is used. Isometrics of inverted heuristics do not rotate around the origin, but rotate around the base rule. Representations of the inverted heuristics in the coverage space reveal the following relationship with the basic heuristics:

$$h^{'}(p,n) = h(N-n, P-p) \tag{4}$$

where parameters P and N denote the number of positive and negative examples in the data set with respect to the target class, and dependent on the predecessor rule. Consequently, the inverted heuristics have the following forms:

- ***Inverted precision:***

$$h^{'}_{prec}(p,n) = \frac{N-n}{(P+N)-(p+n)}; \tag{5}$$

- ***Inverted Laplace:***

$$h^{'}_{lap}(p,n) = \frac{N-n+1}{(P+N)-(p+n+2)}; \tag{6}$$

- ***Inverted m-estimate:***

$$h^{'}_{m-est}(p,n,m) = \frac{N-n+m\cdot\frac{P}{P+N}}{(P+N)-(p+n+m)}. \tag{7}$$

Overall, in [15] the combination of Laplace heuristic h_{lap} in the rule selection step and inverted Laplace heuristic $h^{'}_{lap}$ in rule refinement step outperformed other combinations in terms of average classification accuracy. An interesting side conclusion from [15] is that the usage of inverted heuristics in the rule refinement produces on average longer rules.

The tendency of inverted heuristics to find longer descriptions and no additional parameters make the separation of rule refinement and rule selection an appealing research approach in the domain of subgroup discovery, therefore, we investigated the use of inverted heuristics in subgroup discovery. For that purpose we implemented a new DoubleBeam algorithm for subgroup discovery which implements the usage of separate refinement and selection heuristics with beam search.

3. DOUBLEBEAM ALGORITHM

We implemented a DoubleBeam algorithm for subgroup discovery. This algorithm consists of two beams, refinement and selection beam. Upon initialization, each beam is filled with the best features according to their refinement and selection quality. The algorithm then enters a loop, where it first refines the elements from the refinement beam with features from the dataset. In each step, rules from the refinement beam are refined by adding features to existing rules. Newly produced rules are added to the refinement beam if their refinement quality exceeds the refinement quality of existing rules in the refinement beam. Newly produced rules are then evaluated according to their selection quality. Selection beam is updated with newly induced rules whose selection quality is better than the selection quality of rules already in the beam. The algorithm exits the loop and stops when there are no changes in the selection beam. The DoubleBeam algorithm is outlined in Algorithm 1.

Input : $E = P \cup N$ (E is the training set. $|E|$ is the training set size, P are the positive (class) examples, N are negative (non-target) examples), $TargetClass$
Output : $subgroups$
Parameters : $min\_support$, $refinementBeamWidth$, $selectionBeamWidth$, $refinement\_heuristics$, $selection\_heuristics$

$CandidateList \leftarrow$ all feature values or intervals
for *each candidate in CandidateList* **do**
  evaluate candidate with $refinement\_quality$
  evaluate candidate with $selection\_quality$
end
sort $CandidateList$ according to the $refinement\_quality$
for $i = 0$ *to* $refinementBeamWidth$) **do**
  $RefinementBeam(i) \leftarrow CandidateList(i)$
end
sort $CandidateList$ according to the $selection\_quality$
for $i = 0$ *to* $selectionBeamWidth$) **do**
  $SelectionBeam(i) \leftarrow CandidateList(i)$
end
do
  $RefinementCandidates \leftarrow$ **refine** $RefinementBeam$ with $CandidateList$
  update $RefinementBeam$ with $RefinementCandidates$ using $refinement\_quality$
  update $SelectionBeam$ with $RefinementCandidates$ using $selection\_quality$
while *while there are changes in SelectionBeam*;
return $SelectionBeam$

Algorithm 1: DoubleBeam algorithm

4. EXPERIMENTAL RESULTS

The DoubleBeam algorithm was implemented in the ClowdFlows platform [12]. For the purpose of our evaluation, we used the following combinations of refinement and selection heuristics: $(h^{'}_{lap}, h_{lap})$, $(h^{'}_{prec}, h_{prec})$, $(h^{'}_{m-est}, h_{m-est})$, $(h^{'}_{g}, h_{g})$, and (h_{g}, h_{g}), (named DB-ILL (DoubleBeam-Inverted Laplace, Laplace), DB-IPP (DoubleBeam-Inverted precision,

precision), DB-IMM (DoubleBeam-Inverted m-estimate, m-estimate), DB-IGG (DoubleBeam-Inverted generalization-quotient, generalization quotient), and DB-GG (DoubleBeam-generalization quotient, generalization quotient) respectively). The h_g heuristic is the generalization quotient proposed in [6], while $h_g^{'}$ is its inverted variant. The generalization quotient is a heuristic used in the SD algorithm. The SD algorithm and the algorithms CN2-SD and APRIORI-SD were already implemented in this platform.

4.1 Experimental setting

We use the same 20 UCI classification data sets as [15] to compare three state-of-the-art subgroup discovery algorithms (SD, CN2-SD, and APRIORI-SD) and the DoubleBeam algorithm with five combinations of refinement and selection heuristics (DB-ILL, DB-IPP, DB-IMM, DB-IGG, and DB-GG).

The comparison is performed in 10-fold double-loop cross validation on each dataset. For each algorithm, a grid of possible parameter values was set beforehand. The value of *min_sup* is set to 0.01. Each learning set (10 learning sets) was additionally split into training and test data. For each algorithm, models were built using the training data and its parameters from the grid. Parameters maximizing the value of unusualness of the produced subgroups were then chosen for building a model using the learning set. Unusualness is a measure which was presented in [13] and defined as:

$$WRAcc(Class \leftarrow Cond) = p(Cond) \cdot (p(Class|Cond) - p(Class)). \quad (8)$$

We use the subgroup discovery evaluation function implemented in Orange by Kralj et al. [11]. The function calculates the following measures: *coverage*, *support*, *size*, *complexity*, *significance*, *unusualness* i.e., WRACC, *classification accuracy*, and *AUC*.

4.2 Results

The WRACC values obtained in the experiments are shown in Table 1. These values are averaged over all the classes for every particular dataset. The values for the APRIORI-SD algorithm tested on the horse-colic dataset are missing as the algorithm did not converge in period over 5 days. For the datasets where the WRACC values for the APRIORI-SD algorithm are 0.000 the algorithm returned over $10,000,000$ itemsets and did not finish properly. According to the obtained results, the CN2-SD and the SD algorithm have the best average ranks, and the Apriori-SD algorithm performs the worst.

For comparison between methods we use the methodology proposed by Demšar [3]. We operate under the null-hypothesis that all the algorithms are equivalent. Two algorithms differ significantly if the difference between their average ranks is larger than the value of the critical difference.

The results of the Nemenyi test for the average values of WRACC are shown in Figure 1. Average ranks of algorithms are written in parentheses. The critical value is 2.35. It is evident that the CN2-SD algorithm produces the most interesting subgroups, which are statistically more unusual than the ones produced by the DoubleBeam algorithm with the combinations $(h_{lap}^{'}, h_{lap})$, $(h_{prec}^{'}, h_{prec})$, $(h_{m-est}^{'}, h_{m-est})$, and the APRIORI-SD algorithm. There are no statistically significant differences between the CN2-SD algorithm, the SD algorithm, and the DoubleBeam algorithm with the combinations (h_g, h_g) and $(h_g^{'}, h_g)$.

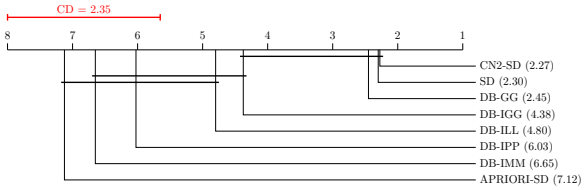


Figure 1: Nemenyi test on WRACC values with a significance level of 0.05.

The results of the Nemenyi test for the average rule size are shown in Figure 2. The DoubleBeam algorithm with the combination $(h_{prec}^{'}, h_{prec})$ produces subgroups which are on average described by the longest rules. The DB-IPP algorithm generates subgroups described by rules that are statistically longer only than the ones produced by the DB-IGG algorithm. There is no statistical evidence that the DB-IPP algorithm produces longer rules than other evaluated algorithms. These results do not confirm that the DoubleBeam algorithm with inverted refinement heuristic produces statistically longer rules than other subgroup discovery algorithms.

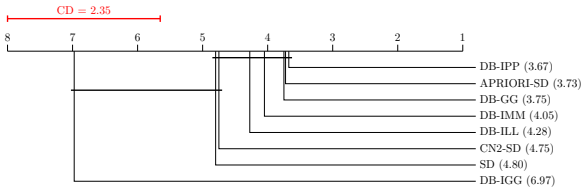


Figure 2: Nemenyi test on ranking of average rule sizes (note that larger rules produce lower rankings) with a significance level of 0.05.

5. CONCLUSIONS

The experiments indicate that subgroup describing rules created using inverted heuristics used in [15] as rule refinement heuristics in subgroup discovery are significantly less interesting than the subgroups induced by the CN2-SD algorithm, the SD algorithm, and the DB-GG algorithm. There is no significant difference of the unusualness of the subgroups induced by the CN2-SD algorithm, the SD algorithm, the DB-GG algorithm, and the DB-IGG algorithm. However, it has to be mentioned that the CN2-SD algorithm uses WRACC as its heuristics for building subgroups.

The results also suggest that when the combination $(h_{prec}^{'}, h_{prec})$ of heuristics is used, the obtained rules tend to have more rule conditions than the rules built by the other state-of-the-art algorithms for subgroup discovery. However, this difference is not statistically significant. The longer rules created by the algorithms using inverted heuristics used in [15] are more specific, thus subgroups contain lower number of examples and this decreases the unusualness of the subgroups. Considering the evaluation results, we can conclude that the DoubleBeam algorithm which uses the combination (h_g, h_g) as refinement and selection heuristics can be

Table 1: Ten-fold double-loop cross validation WRACC results for subgroup discovery. Best values are written in bold.

Datasets	SD	APRIORI-SD	CN2-SD	DB-ILL	DB-IPP	DB-IMM	DB-GG	DB-IGG
breast-cancer	**0.045**	0.003	0.024	0.015	0.010	0.011	**0.045**	0.041
car	0.029	0.006	**0.031**	0.021	0.021	0.003	0.028	0.028
contact-lenses	0.080	0.031	0.066	0.036	0.023	0.000	**0.081**	**0.081**
futebol	**0.017**	0.004	0.009	0.003	0.002	0.000	0.001	0.000
glass	**0.050**	0.017	0.047	0.029	0.026	0.019	0.045	0.038
hepatitis	0.046	0.000	**0.061**	0.041	0.016	0.031	0.060	0.043
horse-colic	**0.131**		0.071	0.045	0.022	0.054	**0.131**	0.084
hypothyroid	0.025	0.000	0.031	0.019	0.009	0.007	**0.032**	0.024
idh	0.088	0.053	**0.094**	0.078	0.068	0.000	0.081	0.088
ionosphere	0.115	0.000	0.111	0.084	0.049	0.062	**0.120**	0.101
iris	0.162	0.119	**0.197**	0.148	0.098	0.045	0.165	0.169
labor	0.074	0.018	**0.106**	0.082	0.021	0.029	0.090	0.063
lymphography	**0.058**	0.015	0.046	0.045	0.042	0.038	0.056	0.036
monk3	**0.068**	0.014	0.065	0.041	0.024	0.007	**0.068**	0.058
mushroom	0.128	0.000	**0.163**	0.147	0.025	0.023	0.146	0.122
primary-tumor	**0.018**	0.004	0.009	0.007	0.009	0.004	0.011	0.008
soybean	0.033	0.000	**0.037**	0.023	0.023	0.009		0.028
tic-tac-toe	**0.053**	0.007	0.021	0.017	0.011	0.007	**0.053**	**0.053**
vote	0.184	0.052	0.201	0.157	0.076	0.120	**0.188**	0.164
zoo	0.083	0.000	**0.096**	0.041	0.020	0.025	0.086	0.058

a good choice for subgroup discovery. It induces subgroups that are comparable to the subgroups induced by the CN2-SD algorithm and the SD algorithm in terms of their unusualness. The subgroups induced by the DB-GG algorithm are on average described by longer rules (see Figure 2).

No additional parameters required with inverted heuristics and the obtained results regarding the average rule length make the proposed approach an interesting research direction. In future we want to focus on the reasons why the rules induced by the DB-ILL algorithm, the DB-IPP algorithm and the DB-IMM algorithm are less interesting than the ones produced by the standard subgroup discovery algorithms and implement an approach which will solve this issue. We also want to research the influence of inverted heuristics in other state-of-the-art subgroup discovery algorithms.

6. REFERENCES

[1] M. Atzmüller and F. Puppe. SD-map - A fast algorithm for exhaustive subgroup discovery. In *Proceedings of Knowledge Discovery in Databases, PKDD 2006*, pages 6–17, 2006.

[2] S. D. Bay and M. J. Pazzani. Detecting group differences: Mining contrast sets. *Data Mining and Knowledge Discovery*, 5(3):213–246, 2001.

[3] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[4] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999*, pages 43–52, 1999.

[5] J. Fürnkranz, D. Gamberger, and N. Lavrač. *Foundations of Rule Learning*. Springer, 2012.

[6] D. Gamberger and N. Lavrač. Expert-guided subgroup discovery: Methodology and application. *Journal of Artificial Intelligence Research*, 17:501–527, 2002.

[7] H. Grosskreutz and S. Rüping. On subgroup discovery in numerical domains. *Data Mining and Knowledge Discovery*, 19(2):210–226, 2009.

[8] B. Kavšek and N. Lavrač. APRIORI-SD: adapting association rule learning to subgroup discovery. *Applied Artificial Intelligence*, 20(7):543–583, 2006.

[9] W. Klösgen. Explora: A multipattern and multistrategy discovery assistant. In *Advances in Knowledge Discovery and Data Mining*, pages 249–271. 1996.

[10] P. Kralj Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10:377–403, 2009.

[11] P. Kralj Novak, N. Lavrač, B. Zupan, and D. Gamberger. Experimental comparison of three subgroup discovery algorithms: Analysing brain ischemia data. In *Proceedings of the 8th International Multiconference Information Society*, pages 220–223, 2005.

[12] J. Kranjc, V. Podpečan, and N. Lavrač. ClowdFlows: A cloud based scientific workflow platform. In *Proceedings of Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012*, pages 816–819, 2012.

[13] N. Lavrač, B. Kavšek, P. A. Flach, and L. Todorovski. Subgroup discovery with CN2-SD. *Journal of Machine Learning Research*, 5:153–188, 2004.

[14] N. Lavrač, F. Železný, and P. A. Flach. RSD: relational subgroup discovery through first-order feature construction. In *Proceedings of the 12th International Conference on Inductive Logic Programming*, pages 149–165, 2002.

[15] J. Stecher, F. Janssen, and J. Fürnkranz. Separating rule refinement and rule selection heuristics in inductive rule learning. In *Proceedings of Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014*, pages 114–129, 2014.

[16] S. Wrobel. An algorithm for multi-relational discovery of subgroups. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD 1997*, pages 78–87, 1997.

INDEXING OF LARGE N-GRAM COLLECTION

Patrik Zajec, Marko Grobelnik
Artificial Intelligence Laboratory,
Jožef Stefan Institute
Jamova cesta 39, 1000 Ljubljana, Slovenia
e-mail: patrik.zajec@ijs.si, marko.grobelnik@ijs.si

ABSTRACT

This paper presents an efficient indexing technique suitable for indexing large n-gram collections with an emphasis on full wildcard query support and speed efficiency. Further we used this technique in building the n-gram search engine, on top of Google's Web 1T 5-gram collection, whose advantages are interactive querying and fast result retrieval with tradeoff on higher memory consumption.

1 INTRODUCTION

In the statistic NLP part of a research project we are using Web 1T 5-gram Version 1 corpus contributed by Google Inc. [1]. The corpus contains English word n-grams with length ranging from unigrams (single word) to five-grams. It contains around 4 billion n-grams and takes around 90 GB of memory.

Due to the corpus size we need an efficient indexing method which enables interactive full wildcard querying support on the data. Query time is our main priority that is why the index is kept in primary memory to eliminate disk IO operations. We have built our own system which is running on a server machine with 512 GB of RAM and has a constant memory usage of 280 GB.

2 RELATED WORK

There is a variety of different approaches and systems for storing and querying large n-gram collections. One class of approaches deals with an efficient representation of data in terms of memory usage while it usually offers only basic membership query.

There are approaches that seek for a tradeoff between memory footprint, speed efficiency and query expressiveness. Usually these are the combinations of fast in memory indexes of data stored on disk, where frequent disk access reduces the performance. Intuitive approach to achieve rich querying capabilities is to store the collection in a relational database and then index the data in desired manner [2], but here speed and memory efficiency suffers.

The work presented in [3] is an approach with in memory B+ tree index to enable wildcard query support but with limitation that wildcards in the query are continuous. Paper [4] describes an offline query pre-processing approach where queries are stored in nested hash table structure and answers are calculated with a single pass through the corpus. [5] summarizes various approaches and presents an efficient storing architecture with n-grams stored in an enhanced prefix tree. It also supports wildcard queries but not as efficiently as our system does.
Main advantages of our system are its **interactivity** (queries are answered online), **full wildcard query support** and very **fast query result retrieval** as searching for n-grams matching a given pattern is fast and straight-forward procedure.

3 INDEXING APPROACH

A. DEFINITION

N-gram is a contiguous sequence of n items. Define the number of items in the sequence as a degree of n-gram – for n-gram Z we mark its degree as $deg(Z)$. As in indexed corpus n-grams are sequences of words, let $Z[i]$ represent the word that appears at position i in the n-gram Z.

Wildcard query is represented with wildcard query pattern which is defined as n-gram Qp of a constant degree D. In our case D equals 5 which is a maximum degree of n-gram from the corpus. For each $i = 1 \ldots D$, $Qp[i]$ is either a word or a wildcard represented as '*' character. Furthermore we define two functions: $isWord(S[i])$ which is **true** if there is a word at position i in given n-gram S and $isWildcard(S[i])$ which is **true** if there is a wildcard at position i in S. We say that n-gram X is a member of the **result set** R of a given wildcard query (n-gram X is said to be a **match**) if for each $i = 1 \ldots D$ the following holds: $isWord(Qp[i])$ and $equals(Qp[i],X[i])$ or $isWildcard(Qp[i])$ – example is shown on figure 1. If $deg(X) < D$ then every such i where $deg(X) < i$ is ignored when determining if X is a match or not.

We define a set of supported wildcard queries Q as a set which contains only wildcard queries our system currently supports – we will need this definition in the next paragraphs.

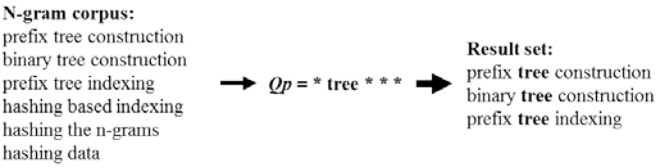


Figure 1: Result set of wildcard query where *Qp=* tree * * ** on the example n-gram corpus.

B. METHOD WITH RESTRICTION ON QUERIES

The naïve way of building the result set for an arbitrary wildcard query on the corpus would be to iterate through the whole corpus and for each n-gram check whether it is a match. But as the corpus is quite large we need a better approach.

Let's try to solve a simple problem where our set *Q* contains just queries where *Qp* has no wildcards. Obviously *R* will contain at most one n-gram as this is in fact a membership query. To speed up answering to that type of queries let's store the n-grams in a **prefix tree** [6] as shown in the figure 2. Answering a query is then just a matter of traversing this prefix tree from top to bottom.

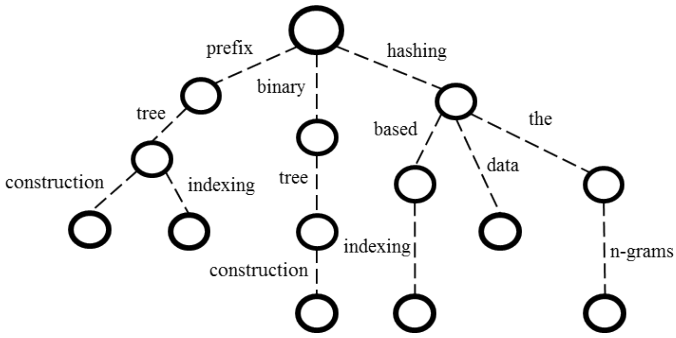


Figure 2: Example corpus stored in prefix tree.

Let's extend our set *Q* with those queries where *Qp* can contain wildcards but with some restrictions. We will split the set of positions from *Qp* in two sets according to a property whether there is a word or a wildcard on a specified position. We call these sets **word positions** *W(Qp)* and **wildcard positions** *C(Qp)*. *W(Qp)* is defined as $\{i;\ 1 \le i \le D$ *and isWord(Qp[i]) }*, *C(Qp)* is defined as $\{i;\ 1 \le i \le D$ *and isWildcard(Qp[i]) }*. Define a **simple query** as query where in its pattern *Qp* all the word positions appear earlier than any of the wildcard ones. In other words *max(W(Qp)) < min(C(Qp)).*

The fact that all the n-grams are stored in prefix tree allows us to use simple tree traversal as we did before until we hit a wildcard in pattern *Qp*. We can realize that all of the n-grams present in the sub-tree rooted at node where we are currently located are exclusively contained in the result set *R*. We can therefore easily generalize searching procedure and extend our set *Q*. But *Q* still contains just a small part of all possible wildcard query patterns.

C. EXTENDING THE SET OF QUERY PATTERNS

Define a **permutation of a n-gram** as a permutation of words in its sequence – we will write permutations in **one-line notation**. Let's permute all the n-grams from the corpus according to some specified permutation π. We then execute an arbitrary simple query on this modified corpus and get a set of results *R'*. If we permute members of *R'* back to their original form we find that *R'* contains same n-grams as if we would execute a query with pattern *Qp'* over original corpus (corpus where n-grams are not permuted). We can get the pattern *Qp'* just by permuting pattern *Qp* according to permutation $inv(\pi)$ where $inv(\pi)$ is inverse of permutation π. Of course pattern *Qp'* does not necessarily correspond to a query from the simple query set *Q*. That is why we cannot execute it in the same efficient way as we can execute simple queries. But we get an important intuition.

If we start with non-simple query pattern *Qc* and transform it to simple query pattern *Qs* with some permutation σ and then make a query on a collection in which all of the n-grams from the corpus are permuted by σ, then querying with pattern *Qs* is possible in the simple tree traversal way as presented before; the results in set *R* are same as they would be if we made a query with pattern *Qc* on the original collection (the results from *R* just have to be permuted with $inv(\sigma)$ to get the original version of resulting n-grams). So we have a procedure to efficiently find the result set for all kinds of wildcard queries. Each query just has to be transformed to simple query by transforming its pattern *Qp* to simple one with some permutation σ and all the n-grams from the corpus have to be permuted by the same permutation σ. On matching stage we can act as if we are dealing with a simple query and efficiently find the set *R* whose members have to be permuted with $inv(\sigma)$ to get the resulting n-grams back in their original form.

D. TRANSFORMING NON-SIMPLE PATTERN TO SIMPLE ONE

Finding σ to transform *Qc* to *Qs* is an easy task. What makes the query pattern non-simple is the fact that there exists a wildcard position which is later followed by a word position. Permutation σ just has to map all the words in front of the wildcards. The catch is that if we want to use an efficient matching method also all the n-grams have to be permuted by σ and stored in an **auxiliary collection**. That is why we have to find the minimum set of permutations *M* so that for each possible wildcard query pattern there exists a permutation from *M* that transforms it to simple pattern. Size of *M* has to be as small as possible because each of permutations form *M* will have a corresponding auxiliary collection of all the n-grams from the corpus permuted according to it. The fact that actual order of words in permuted pattern is irrelevant as long as they are all in front of the wildcards helps us to significantly decrease the upper bound on the number of permutations as there is usually more than one candidate permutation σ that satisfies this transformation property. More formally we have to find such set of permutations M, that for each possible query pattern *Qp* there exists such a permutation σ from *M* that

im(σ(W(Qp))) = {1,…,|W(Qp)|} and *im(σ(C(Qp))) = {|W(Qp)| + 1,…,D}*. It's quite obvious that *|M|* can be bounded by the number of different sets *W(Qp)*. Since size of the set *W(Qp)* is limited with maximum degree of a pattern *Qp* (in our case *D = 5*) we only have 2^5 different possible sets. The upper bound on *|M|* is therefore 32 but can be improved even further.

E. MINIMUM SET OF PERMUTATIONS

Let *M(d)* represent a subset of *M* containing minimum number of permutations required to transform any wildcard query pattern *Qp* where *|W(Qp)| = d* to simple one. We will build the set M inductively on *d*.

For *d = 1* clearly *|M(d)| = D* as for each possible word position a permutation that maps it to 1 is needed.When constructing *M(d)* where *d > 1* we can find for each possible set *W(Qp)* of size *d* the permutation from *M(d – 1)* that maps *d – 1* elements of its elements to *{1,…,d – 1}* and then extend this permutation to map the remaining element to *d*.

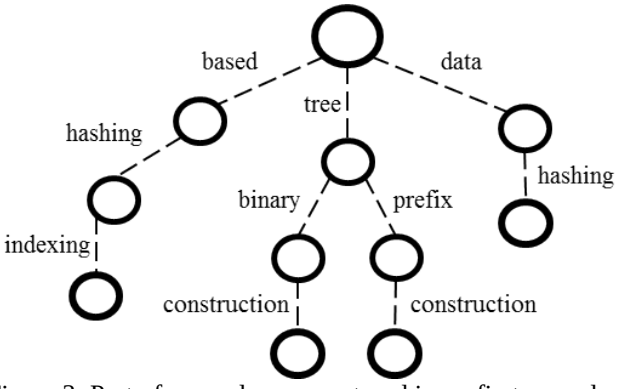


Figure 3: Part of example corpus stored in prefix tree and permuted by *σ = (2 1 3)*.

We can notice that the size of *M(d)* is the same as the number of different possible sets *W(Qp)* of size *d* which equals to *C(D,d)* (where *C(n,k)* represents a combination). As *C(n,k)* is maximal when k equals *n / 2* (integer division) we can set the lower bound on size of *|M|* to *C(D,D / 2)* as there are *C(D,D / 2)* permutations in set *M(D / 2)*. But can we always construct *M* so that its size actually matches the lower bound? It turns out we can, as long as we reuse (and extend) all the permutations from *M(d – 1)* when constructing *M(d)* – where *|M(d – 1)| ≤ |M(d)|*. As all the permutations from *M(d – 1)* are also included (in the extended form) in *M(d)* there always exists a permutation from *M(d)* which transform a pattern with *|W(Qp)| = d - 1* to simple one. It implies that permutations from the set *M(D / 2)* are enough to correctly transform all possible *W(Qp)* where *|W(Qp)| ≤ D / 2*. We can combine the permutations from *M(D / 2)* with those from *M(d)* where *D / 2 < d* to construct the minimum set *M* with *C(D,D / 2)* elements.
For indexing our corpus the minimum set of permutations contains exactly 10 different permutations (in appendix), each of which has its corresponding auxiliary collection with n-grams permuted according to it. Therefore our index is 10 times as large as the corpus.

4 INDEX CONSTRUCTION

N-grams from the corpus are stored in 10 different collections. Each collection has a corresponding permutation according to which n-grams stored in it are permuted. Collections are implemented as prefix trees so that finding the result set can be done in a simple and efficient way as described before. Another advantage of prefix tree is that it can be fairly efficiently compressed once it is built. We are building collections sequentially; building procedure is divided in two stages. In first stage n-grams from the corpus are permuted and sequentially inserted in prefix tree which is stored in primary memory. When built, prefix tree takes about 240 GB of memory as it is not efficiently represented for a sake of faster insertion. On the second stage tree is compressed and in the end it takes only about 30 GB of memory. As compression we mean that nodes, edges and words are represented in a more succinct way which slightly reduces speed efficiency when traversing the tree. At the end the whole index takes 280 GB and is loaded in primary memory to achieve fastest result retrieval. Index construction took around 24 hours on a server machine with 512 GB of RAM.

5 EFFICIENCY

When the query pattern *Qp* is known it is transformed to simple query pattern *Qs* by the right permutation. Then the collection (prefix tree) which corresponds to this permutation is identified. Once we have the right prefix tree finding the result set is just a matter of simple tree traversal. Results can be served all at once or by iteratively calling *getNextResult()* function. While the first match to given query pattern is found instantaneously due to straight-forward matching procedure, the system is capable to retrieve "just" 10^5 matches per second. The slowdown of its performance lies in succinct representation which results in slower tree traversal due to the complex structure of data.

6 CONCLUSION

We have described a novel indexing schema whose main advantages are full wildcard query support, interactive querying and speed efficiency. Even though it is currently a static indexing approach it can be extended to a dynamic version quite easily. Its usage is not limited just on word n-grams.

REFERENCES

[1] Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram corpus version 1. Technical report, Google Research.

[2] S. Evert. 2010. Google web 1t 5-grams made easy (but not for the computer). In Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop, WAC-6 '10, pages 32–40.

[3] Ceylan, H., and Mihalcea, R. 2011. An efficient indexer for large n-gram corpora. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), System Demonstrations, Portland, OR, USA, pp.

103–8. Stroudsburg, PA: Association for Computational Linguistics.
[4] T. Hawker, M. Gardiner, and A. Bennetts. 2007. Practical queries of a massive n-gram database. In Proceedings of the Australasian Language Technology Workshop 2007, pages 40–48, Melbourne, Australia.
[5] Michael Flor. 2013. A fast and flexible architecture for very large word n-gram datasets. Natural Language Engineering, 19(1), 61-93.
[6] Fredkin, E. 1960. Trie memory. Communications of the ACM 3(9): 490–9.

APPENDIX

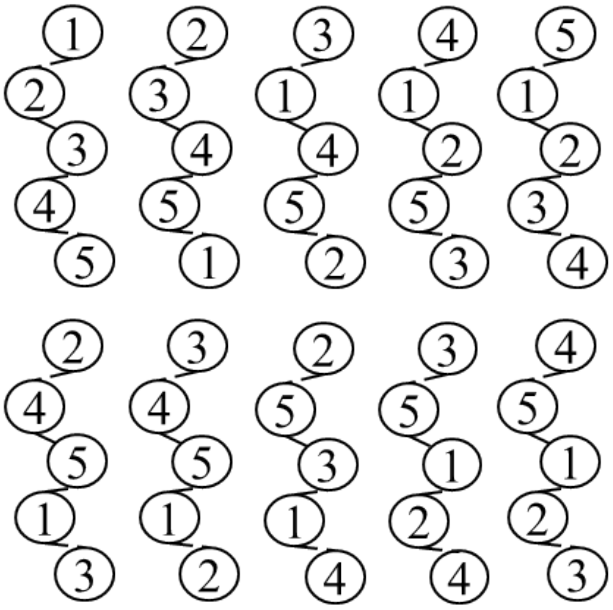


Figure 4: Minimum set of permutations for indexing n-grams with degree up to 5 used in our index. Permutations are written in one-line notation.

Indeks avtorjev / Author index

Belyaeva Evgenia 5
Borštnik Andrej 21
Breskvar Martin 9
Džeroski Sašo 9
Eftimov Tome 13
Fabbretti Elsa 17
Grobelnik Marko 33, 45
Gubiani Donatella 17
Kenda Klemen 21
Koroušić Seljak Barbara 13
Košmerlj Aljaž 5
Lavrač Nada 41
Mladenić Dunja 5, 25, 33
Moraru Alexandra 25
Petrič Ingrid 17
Pita Costa Joao 29
Rei Luis 33
Robnik Šikonja Marko 41
Škraba Primož 29
Škrjanc Maja 21
Stopar Luka 37
Urbančič Tanja 17
Valmarska Anita 41
Zajec Patrik 45
Ženko Bernard 9

Konferenca / Conference

Uredili / Edited by

Izkopavanje znanja in podatkovna skladišča (SiKDD 2015) / Data Mining and Data Warehouses (SiKDD 2015)

Dunja Mladenić, Marko Grobelnik