Zbornik 19. mednarodne multikonference

INFORMACIJSKA DRUŽBA - IS 2016

Zvezek D

Proceedings of the 19th International Multiconference

INFORMATION SOCIETY - IS 2016

Volume D

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

Uredila / Edited by
Dunja Mladenić, Marko Grobelnik

http://is.ijs.si

10. oktober 2016 / 10 October 2016
Ljubljana, Slovenia

Zbornik 19. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2016

Zvezek D

Proceedings of the 19th International Multiconference

INFORMATION SOCIETY – IS 2016

Volume D

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

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

10. oktober 2016 / 10 October 2016
Ljubljana, Slovenia

Urednika:

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

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

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

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

Ljubljana, oktober 2016

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

004.8(082)(0.034.2)

MEDNARODNA multikonferenca Informacijska družba (19 ; 2016 ; Ljubljana)
Izkopavanje znanja in podatkovna skladišča (SiKDD) [Elektronski vir] : zbornik 19. mednarodne multikonference Informacijska družba - IS 2016, 10. oktober 2016, [Ljubljana, Slovenija] : zvezek D = Data mining and data warehouses (SiKDD) : proceedings of the 19th International Multiconference Information Society - IS 2016, 10 October 2016, Ljubljana, Slovenia : volume D / uredila, edited by Dunja Mladenić, Marko Grobelnik. - El. zbornik. - Ljubljana : Institut Jožef Stefan, 2016

Način dostopa (URL):
http://library.ijs.si/Stacks/Proceedings/InformationSociety/2016/IS2016_Volume_D%20-%20SiKDD.pdf

ISBN 978-961-264-100-9 (pdf)
1. Gl. stv. nasl. 2. Vzp. stv. nasl. 3. Mladenić, Dunja
29872935

PREDGOVOR MULTIKONFERENCI INFORMACIJSKA DRUŽBA 2016

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

Informacijska družba, znanje in umetna inteligenca so spet na razpotju tako same zase kot glede vpliva na človeški razvoj. Se bo eksponentna rast elektronike po Moorovem zakonu nadaljevala ali stagnirala? Bo umetna inteligenca nadaljevala svoj neverjetni razvoj in premagovala ljudi na čedalje več področjih in s tem omogočila razcvet civilizacije, ali pa bo eksponentna rast prebivalstva zlasti v Afriki povzročila zadušitev rasti? Čedalje več pokazateljev kaže v oba ekstrema – da prehajamo v naslednje civilizacijsko obdobje, hkrati pa so planetarni konflikti sodobne družbe čedalje težje obvladljivi.

Letos smo v multikonferenco povezali dvanajst odličnih neodvisnih konferenc. Predstavljenih bo okoli 200 predstavitev, povzetkov in referatov v okviru samostojnih konferenc in delavnic. Prireditev bodo spremljale okrogle mize in razprave ter posebni dogodki, kot je svečana podelitev nagrad. Izbrani prispevki bodo izšli tudi v posebni številki revije Informatica, ki se ponaša z 39-letno tradicijo odlične znanstvene revije. Naslednje leto bo torej konferenca praznovala 20 let in revija 40 let, kar je za področje informacijske družbe častitljiv dosežek.

Multikonferenco Informacijska družba 2016 sestavljajo naslednje samostojne konference:

- 25-letnica prve internetne povezave v Sloveniji
- Slovenska konferenca o umetni inteligenci
- Kognitivna znanost
- Izkopavanje znanja in podatkovna skladišča
- Sodelovanje, programska oprema in storitve v informacijski družbi
- Vzgoja in izobraževanje v informacijski družbi
- Delavnica »EM-zdravje«
- Delavnica »E-heritage«
- Tretja študentska računalniška konferenca
- Računalništvo in informatika: včeraj za jutri
- Interakcija človek-računalnik v informacijski družbi
- Uporabno teoretično računalništvo (MATCOS 2016).

Soorganizatorji in podporniki konference so različne raziskovalne institucije in združenja, med njimi tudi ACM Slovenija, SLAIS, DKZ in druga slovenska nacionalna akademija, Inženirska akademija Slovenije (IAS). 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 2016 bomo četrtič 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. Tomaž Pisanski. Priznanje za dosežek leta bo pripadlo prof. dr. Blažu Zupanu. Že šestič podeljujemo nagradi »informacijska limona« in »informacijska jagoda« za najbolj (ne)uspešne poteze v zvezi z informacijsko družbo. Limono je dobilo ponovno padanje Slovenije na lestvicah informacijske družbe, jagodo pa informacijska podpora Pediatrične klinike. Čestitke nagrajencem!

Bojan Orel, predsednik programskega odbora
Matjaž Gams, predsednik organizacijskega odbora

FOREWORD - INFORMATION SOCIETY 2016

In its 19th 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 2016 it is organized at various locations, with the main events at the Jožef Stefan Institute.

The pace of progress of information society, knowledge and artificial intelligence is speeding up, but it seems we are again at a turning point. Will the progress of electronics continue according to the Moore's law or will it start stagnating? Will AI continue to outperform humans at more and more activities and in this way enable the predicted unseen human progress, or will the growth of human population in particular in Africa cause global decline? Both extremes seem more and more likely – fantastic human progress and planetary decline caused by humans destroying our environment and each other.

The Multiconference is running in parallel sessions with 200 presentations of scientific papers at twelve conferences, round tables, workshops and award ceremonies. Selected papers will be published in the Informatica journal, which has 39 years of tradition of excellent research publication. Next year, the conference will celebrate 20 years and the journal 40 years – a remarkable achievement.

The Information Society 2016 Multiconference consists of the following conferences:

- 25th Anniversary of First Internet Connection in Slovenia
- Slovenian Conference on Artificial Intelligence
- Cognitive Science
- Data Mining and Data Warehouses
- Collaboration, Software and Services in Information Society
- Education in Information Society
- Workshop Electronic and Mobile Health
- Workshop »E-heritage«
- 3st Student Computer Science Research Conference
- Computer Science and Informatics: Yesterday for Tomorrow
- Human-Computer Interaction in Information Society
- Middle-European Conference on Applied Theoretical Computer Science (Matcos 2016)

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

For the fourth year, the award for life-long outstanding contributions will be delivered in memory of Donald Michie and Alan Turing. The Michie-Turing award will be given to Prof. Tomaž Pisanski for his life-long outstanding contribution to the development and promotion of information society in our country. In addition, an award for current achievements will be given to Prof. Blaž Zupan. The information lemon goes to another fall in the Slovenian international ratings on information society, while the information strawberry is awarded for the information system at the Pediatric Clinic. Congratulations!

Bojan Orel, 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
Mitja Lasič
Robert Blatnik
Aleš Tavčar
Blaž Mahnič
Jure Šorn
Mario Konecki

Programme Committee

Bojan Orel, chair
Nikolaj Zimic, co-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

Zbornik 19. mednarodne multikonference

INFORMACIJSKA DRUŽBA – IS 2016

Zvezek D

Proceedings of the 19th International Multiconference

INFORMATION SOCIETY – IS 2016

Volume D

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

Uredila / Edited by

Dunja Mladenić, Marko Grobelnik

10. oktober 2016 / 10 October 2016
Ljubljana, Slovenia

PREDGOVOR

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

FOREWORD

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.

Dunja Mladenić, Marko Grobelnik

PROGRAMSKI ODBOR / PROGRAMME COMMITTEE

Dunja Mladenić
Marko Grobelnik

Near Real-Time Transportation Mode Detection Based on Accelerometer Readings

Jasna Urbančič, Luka Bradeško, Matej Senožetnik
Jožef Stefan Institute
and
Jožef Stefan International Postgraduate School
Jamova cesta 39
1000 Ljubljana, Slovenia

ABSTRACT

This paper describes a method for automatic transportation mode detection based on smartphone sensors. Our approach is designed to work in real-time as it only requires 5s of sensor readings for the detection. Because we used accelerometer instead of GPS signal it uses less battery power and is therefore more user and phone-friendly. For the mode detection we use multiple support vector machine models which enable us distinguishing between multiple modes (bus, train, car). Before the classification, raw measurements are preprocessed in order to cancel out the constant acceleration that is caused by the force of gravity. The results of the paper are promising and are based on the collected training data from approximately 20 hours of driving on trains and public buses in Ljubljana.

Keywords

activity recognition, support vector machine classification, accelerometer

1. INTRODUCTION

Nowadays most smartphones have built-in sensors that measure motion, acceleration, orientation, and various other environmental conditions with quite high precision and sampling frequency. This can be used with great success in everyday challenges, for example tracking and routing applications. It has been proven that smartphone sensors are useful in monitoring three-dimensional device movement and positioning [1], and also user's activity detection, which is also in the domain of this paper.

Mobile operating system developers are aware of such applications, therefore their APIs include activity recognition packages. However, they detect only a few modes - still, walking, running, cycling, and in vehicle. Such coarse-grained classification is not enough for tracking and routing purposes, specially in use-cases for urban environments, where public transportation with buses and trains can be a good alternative to private vehicles.

As the main smartphone APIs already support fine-grained classification of non-motorized forms of transportation [2, 3], we focused on distinguishing means of motorized transport, specifically cars and buses, as the majority of passenger traffic in Ljubljana represent cars and buses. Trains and motorbikes are not that common, whereas subway and tram infrastructures do not exist. Our goal is to recognize each mode of transportation in near real time while mobile phone users are traveling.

2. RELATED WORK

Ever since smartphones appeared and gained accessibility there has been a lot of research activity for their usage in user activity recognition and transportation mode detection. While the first attempts to recognize user activity were done before smartphones, the real effort in that direction started with the development of mobile phones having built-in sensors [7]. Besides GPS sensors, also GSM triangulation and local area wireless technology (Wi-Fi) can be employed for the purpose of transportation mode detection. However its accuracy is relatively low compared to GPS, therefore we deem these out of scope of this paper[8].

Latest state of the art research is focused on transportation mode detection based on GPS signal and/or accelerometer data. Approaches that rely solely on GPS trajectories require GPS signal of high-quality, whereas phones GPS receiver is generally severely shielded during daily activities [10]. This may occur during travel underground, inside stations, or even when user is not sufficiently close to a window when traveling in a vehicle [5], and results in loss of positioning information. Another known issue when using GPS signal on mobile device is high power consumption [5], which is especially not pleasing in the case of longer commutes. Both of these two issues suggest that the accelerometer sensor is more appropriate for activity detection.

Another advantage of using accelerometer data over GPS signal is that it does not require additional external data. Many researches using GPS data used external data, such as GIS data on bus stations, bus routes and railway lines [9].

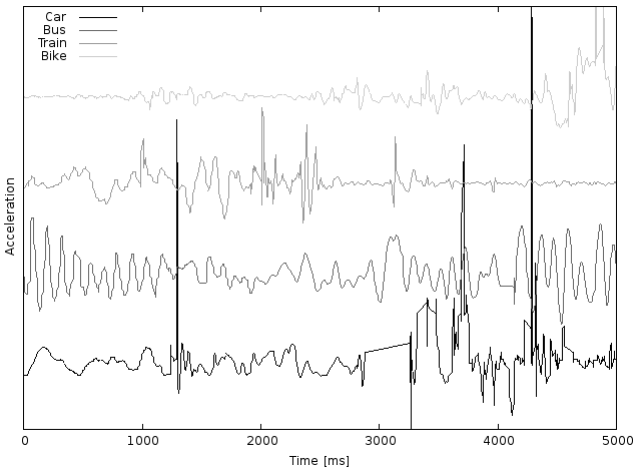

Figure 1: Amplitudes of raw accelerometer data for different means of motorized transportation.

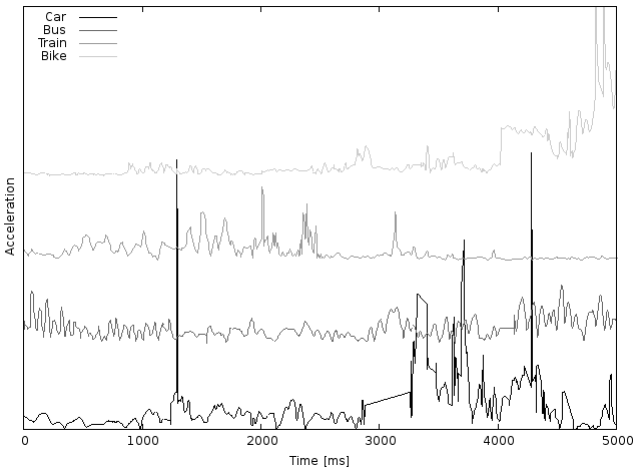

Figure 2: Amplitudes of preprocessed accelerometer data for different means of motorized transportation.

3. TRANSPORTATION MODE DETECTION

For the purpose of collecting accelerometer data we extended the GPS tracking mobile application with the accelerometer measuring ability. The phone sensor measures acceleration forces in m/s^2 for all of the three physical axes (x, y, z). The sampling rate is 100Hz (1 measurement every 10 ms). To increase the diversity of the training data-set, measurements were acquired in multiple ways:

- Person is collecting data while traveling by the car and stops the collection at the destination.
- Person is collecting data while traveling by the bus and stops the collection on exit.
- Person is traveling by the train and is collecting data until the arrival to the final destination
- Person is collecting data while driving a motorbike and stops the collection at the destination.

We collected approximately 20 hours of travelling measurements, with the travel modes distributed as follows: $car - 57\%$, $bus - 32\%$, $train - 11\%$, $motorbike - 0.1\%$. Amplitudes of the raw accelerometer data are shown in Figure 1.

3.1 Preprocessing

In order to reduce the computation time and to have faster response time for real-time classification, we split the recorded accelerometer signal into smaller pieces that do not exceed 5s timewise. This enables us to work on chunks of record that span only through 5s or less. This additionally preserves battery life, saves space, and reduces usage of mobile data.

Acceleration measurements are correlated with the orientation of the phone in 3D space, as gravity is measured together with the dynamic acceleration caused by phone movements. Thus we have to be able to separate the constant acceleration caused by the gravity and the dynamic part of the acceleration.

The gravity estimation algorithm works as follows: for a chosen sampling interval (in our case 1s), obtain an estimate of the gravity component on each axis by averaging all measurements in the interval on that axis [6]. After obtaining the estimates, we subtracted the gravity component from all of the entries on corresponding axis in given time interval. Through this we obtained only the dynamic acceleration component of the signal. Amplitudes of dynamic accelerometer signal are shown in Figure 2.

3.2 Classification

After preprocessing the accelerometer readings we extracted features for the classification process. We used mean, variance, skewness, 5th, and 95th percentile of acceleration data on all three axes. We also split the acceleration into positive part, which indicates that the velocity of movement in that direction is increasing, and negative part, which indicates that the velocity is decreasing, and calculated the same statistics on these two parts.

We used support vector machine (SVM) classifier as it was previously successfully used in similar work [8]. The implementation was SVM classifier (SVC) from QMiner package. QMiner is an open source analytics platform for performing large scale data analysis written in C++ and exposed via a Javascript API [4].

First we focused on the binary classification of car and bus transportation versus the rest. We trained binary classifiers for each of the labels in one against the rest manner. That means that examples labled with this particular label represented positive examples, whereas all other examples, regardless of class represented negative examples. However, we also did one against one classification for each pair of labels. That means that examples of one class were marked as positive, examples of another class were marked as negative, and the rest of the learning set was filtered out. Later, we extended this to support multi-class classification. For multi-class classification we used binary models and combined their predictions based on the distance between the

Table 1: Table of all extracted features.

Acceleration data	Features
X axis	Mean (Total, Acceleration, Deceleration)
Y axis	Standard deviation (Total, Acceleration, Deceleration)
Z axis	Skewness (Total, Acceleration, Deceleration)
Amplitude	5th percentile (Total, Acceleration, Deceleration)
	95th percentile (Total, Acceleration, Deceleration)
Total number of features	60

Table 2: Classification accuracy, precision, recall, and F1 score for binary classifiers of different transportation modes as results of 10-fold cross validation.

	Accuracy	Precision	Recall	F1 score
Car	0.855	0.852	0.910	0.880
Bus	0.720	0.620	0.694	0.655
Train	0.876	0.726	0.671	0.697

separating hyper plane and the test sample.

4. EVALUATION

Evaluation section is divided into two parts. In the first one are presented the results of experiments with one against the rest classification, whereas in the second part we discuss the results of one against one classification. In both parts we considered two different scenarios. In the first one we tested if our approach can recognize a specific transportation mode from all the others (one-vs-all) or if SVC can distinguish between two specific modes of transportation (one-vs-one). In the second, we used previously obtained binary models to classify to three(3) different classes. Our main focus was recognizing traveling by cars and buses as these represent majority of the passenger traffic in Ljubljana and therefore the majority of our training data.

We measured performance of the models with classification accuracy, precision and recall. Furthermore, we estimated a harmonic mean of precision and recall with the F1 score. We used 10-fold cross validation to tune the parameters of each binary model.

4.1 One against all

Binary classification (one-vs-all) was done for each of the three classes (car, bus and train). Classification accuracy, precision and recall for the most suitable values of parameters are listed in Table 2.

We got the best results (accuracy, precision and recall) for car travel detection. There was some drop in the performance of bus travel detection, and even bigger drop for the train detection. We assume that the main cause for performance drop is smaller training data set for bus and train. We plan to resolve this with additional data-set collection as part of the future work.

For the multi-class classification we used binary models from Table 2. We mapped results into 4 classes (car, bus, train and UC - unable to classify). If according to the binary classification, an instance belongs to none of the classes or more than one, we label it as UC (unable to classify). The plan behind UC is, that we ask the application providing accelerometer data for new sample, which can help us reclassify into the proper class. The results in Table 3 and Table 4 are not surprising as the majority of cars is classified as cars and also most of misclassified cars are instances that belong to either none or more than one class. In contrast to cars, proportions of correctly classified buses and trains are smaller than the proportions of UC for these two classes, which shows that our approach to combining predictions for multiple classifiers might not be the best.

Table 3: Confusion matrix for classification for 3 classes with 10-fold cross-validation.

True\ Pred.	Car	Bus	Train	UC
Car	0.818	0.012	0.008	0.162
Bus	0.198	0.219	0.072	0.511
Train	0.118	0.042	0.344	0.496

Table 4: Classification accuracy, precision, recall, and F1 score for classification with 3 classifiers with 10-fold cross-validation.

	Accuracy	Precision	Recall	F1 score
Car	0.823	0.826	0.818	0.827
Bus	0.725	0.844	0.219	0.347
Train	0.859	0.683	0.181	0.286
Average	0.803	0.784	0.401	0.535

4.2 One against one

We did similarly for one-vs-one binary classification. Results of this are shown in Table 5, which shows that cars are very well distinguishable from trains and vice versa. Buses are less distinguishable from cars and trains, however the accuracy and F1 score of all the classifications are still above 0.8.

We used these six binary classifiers for multi-class classification. Confusion matrix and accuracy, precision, recall and F1 score are listed in Tables 6 and 7. Tables show that classification accuracy, precision, recall and F1 score are higher than in case of one against all classification. Confusion ma-

Table 5: Accuracy / F1 score for one-vs-one binary classification. Rows represent positive examples, whereas columns are negative examples.

	Car	Bus	Train
Car		0.848/0.889	0.943/0.962
Bus	0.856/0.832		0.858/0.896
Train	0.934/0.883	0.815/0.760	

Table 6: Confusion matrix for classification for 3 classes with 10-fold cross-validation.

True\ Pred.	Car	Bus	Train
Car	0.893	0.088	0.019
Bus	0.282	0.573	0.145
Train	0.167	0.162	0.671

Table 7: Classification accuracy, precision, recall, and F1 score with 10-fold cross-validation.

	Accuracy	Precision	Recall	F1 score
Car	0.825	0.785	0.893	0.835
Bus	0.787	0.724	0.573	0.640
Train	0.886	0.672	0.671	0.671
Average	0.832	0.727	0.712	0.716

trix shows that nearly 90% of cars is classified correctly, whereas for buses and trains the percentage of correctly classified instances drops to 57% and 67% respectively. It also shows that trains are equally likely to be miss-classified as cars and the buses. In contrast to trains, buses are more often miss-classified as cars than trains. In comparison with one-vs-all multi-class classification, values of recall and F1 score are much higher, whereas accuracy and precision for these two approaches are comparable.

5. CONCLUSION

In the presented work we showed that it is possible to detect transportation mode using support vector machine, with short readings of accelerometer signal. This proves that near real-time activity detection with fine-grained motorized transportation modality is possible.

Nonetheless, there are still some issues for the future work. First one is regarding unbalanced data set as we will have to collect more data, especially for train and motorbike detection. Secondly, our task will be improving binary classification regarding buses as this class was most often misclassified. However, this might be also caused by our policy for combining multiple binary classification results, which is also something we will have to work on.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the ICT program of the EC under project OPTIMUM (H2020-MG-636160).

7. REFERENCES

[1] Sensors overview. `https://developer.android.com/guide/topics/sensors/sensors_overview.html`, 2016. [Online; accessed 25-August-2016].

[2] `ActivityRecognition`. `https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognition`, 2016. [Online; accessed 25-August-2016].

[3] `CMMotionActivity`. `https://developer.apple.com/library/ios/documentation/CoreMotion/Reference/CMMotionActivity_class/index.html#//apple_ref/occ/cl/CMMotionActivity`, 2016. [Online; accessed 25-August-2016].

[4] 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. In *Software Engineering for Machine Learning Workshop, Neural Information Processing Systems*, 2014.

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

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

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

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

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

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

Application of Advanced Analytical Techniques in Support of Socio-economic Impact Assessment of Innovation Funding Programmes

Katja Berčič
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana
Slovenia
katja.bercic@ijs.si

Gabriella Cattaneo
IDC Italia
Viale Monza 14
20127 Milano
Italy
gcattaneo@idc.com

Mario Karlovčec
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana
Slovenia
mario.karlovcec@ijs.si

Flavio Fuart
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana
Slovenia
flavio.fuart@ijs.si

Gaber Cerle
Jožef Stefan Institute
Jamova cesta 39
1000 Ljubljana
Slovenia
gaber.cerle@ijs.si

ABSTRACT

This paper gives an introduction to the European Programme for Internet Innovation and the needs for advanced analytical techniques in assessment of the socio-economic impact of funded projects. A technical overview and architecture of developed IT tools follows. Broadly, two set of tools were explored. (1) An on-line assessment environment that provides programme managers and companies with feedback about their potential socio-economic impact. For this set of tools the emphasis is on visualisation techniques, therefore a detailed rationale and scientific background for those is provided with usage examples. (2) Statistical modules for identification of funding approaches that, potentially, provide best results ("best practices") have been developed and applied to available data. Through one use-case we have shown that the proposed IT system is useful in measuring the impact of innovation funding programmes, detecting good management practices and giving support to funded projects. The whole system has been published on a public repository with an open source license, giving opportunity to re-use, customize and integrate the reporting system into other impact assessment environments.

Categories and Subject Descriptors

H.4.0 [**Information Systems Applications**]: General—*reporting, statistical analysis, visualisation, networks*

General Terms

Algorithms

Keywords

impact assessment, network visualisation, statistical methods

1. INTRODUCTION

The Future Internet Public-Private Partnership (FI-PPP) is the European programme for Internet innovation. Phase Three of the FI-PPP funding is targeting more that 1000 entrepreneurs, start-ups or SMEs in an attempt to multiply the uptake and impact of the technologies developed in previous phases[1]. Under that framework the Future Internet Impact Assurance (FI-MPACT)[2] Support Action was funded in order to collect and assess the qualitative and quantitative evidence of the potential socio-economic impact of the programme by measuring and projecting market sector economic potential, stakeholder take-up and technological impact of Phase III SME Accelerator projects to 2020. Accelerator projects aim at investing in the strongest start-ups ("subgrantees") across Europe and create an ecosystem where entrepreneurs and business incubators meet.

Based on Key Performance Indicators (KPI) elaborated in the Impact Assessment Guidebook [3], a set of **Impact Assessment** tools were developed to collect empirical data from subgrantees and other interested initiatives. In this context, developed IT tools provide an automated assessment system, driven by a set of KPIs, allowing Accelerator projects, other start-ups and entrepreneurs to measure the potential impact compared to industry standards and the global community of FI-PPP projects. By benchmarking their progress in relation to different business processes, they can identify areas where improvements are needed and measure progress.

Furthermore, a set of scripts for statistical data processing were combined in **Accelerator Benchmarking** tools

[1]https://www.fi-ppp.eu/
[2]http://cordis.europa.eu/project/rcn/191426_en.html

to support the identification of accelerators best practices. This quantitative analysis was used to support, guide and strengthen the expert judgement of qualitative indicators of best performing accelerators.

In this paper we present the developed IT tools [2], their usage in practice, some results and possible use of these tools in future.

2. IMPACT ASSESSMENT TOOL

The Impact Assessment tool[3] was focused around funded subgrantees to facilitate mapping of this portfolio, to contribute to the overall impact assessment of the FI-PPP Phase 3 and assist in forecasting the potential impact of this intervention up to 2020. The Self-Assessment tool is open to all interested parties and respondents can undertake the survey at different stages to measure their progress.

The Impact/Self-Assessment tool provides a start-up sanity check, by calculating KPIs for Innovation, Market Focus, Feasibility and Market Needs. Instant feedback is provided by benchmarking respondent's scores with average scores of his/her peers, or a group of most successful peers.

2.1 Key Performance Indicators Benchmarking

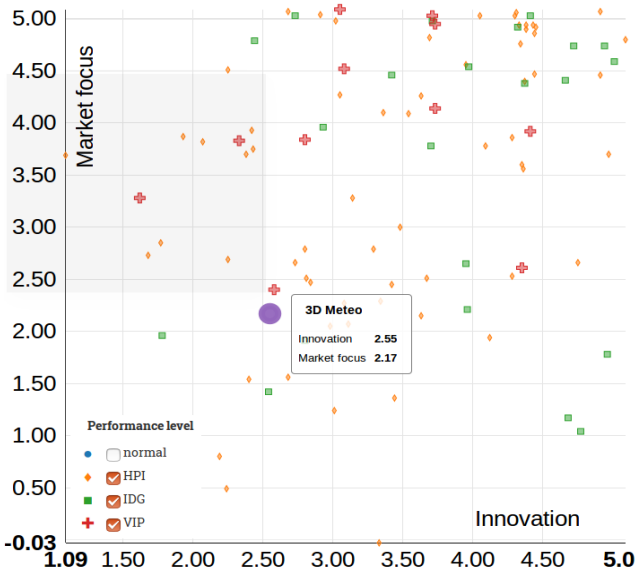

Figure 1: Performance graph based on market focus and innovation variables for a selected project represented with a large violet circle. Different shapes correspond to performance levels: rhombus for Hight Performance Initiatives, squares and crosses for a small selection promising initiatives that received further funding for dissemination activities.

This dynamic benchmarking against other respondents gives entrepreneurs and their mentors a tool to monitor progress and identify areas where additional support is required. The tool is implemented as a interactive scatter plot that positions projects in two-dimensional space defined with performance indicators. This enables comparison of projects with different performance levels against different pairs of indicators. Performance indicators supported by the tool are: innovation, market focus, feasibility, market needs and mattermark growth. Figure 1 gives an example of performance graph where the selected project is compared with other high performance projects, based on market focus and innovation variables.

2.2 Multi-attribute Based Similarity Tool

In order to provide both the sub-grantees and accelerators better insight in project's performance, a tool that shows the selected project's position in relation to other projects was created. In this tool, similarities among projects are determined using a wide range of attributes that describe projects. This gives users a chance to form their own judgement about "good" or "bad" similarities. The tool is implemented using several well-known data and network analysis methods, and visualized as a network graph. The visualization gives sub grantees and their mentors/reviewers insight into how they compete with other projects and ideas, identify possible similarities, find opportunity windows or search for possible partners. The tool is implemented as an open API using QMiner [5] data analytics platform for processing large-scale real-time streams containing structured and unstructured data. The approach used for implementing multi-attribute based similarity tool consists of following steps: (1) Importing data and feature extraction; (2) Computing the main similarity graph; (3) Generating custom graph for selected project; and (4) Visualization.

In the first step, project data is imported into the system and features are extracted from the data. Prior to importing, data is transformed into JSON file format. The file contains a configuration part that determines which attributes of the data will be used as features in further analysis. This makes the system flexible for feature-set changes. The feature-set consists of numerical, categorical and textual data obtained from questionnaires. Feature extractor for textual data applies English stop-words removal in the tokenization phase and normalizes the word frequencies using TF-IDF weighting.

The main similarity graph shows project similarities based on multiple attributes in form of a two dimensional network graph. This graph is the basis for constructing other custom graphs where a particular selected project is in the focus. The graph is constructed in two steps. First, the multidimensional vector representation of projects obtained in the feature extraction phase is transformed into two dimensional representation by using Multidimensional scaling (MDS) [7] method. A similar approach of dimensionality reduction was used in [4] for visualization of a text document corpus. In the second step of graph construction, the two dimensional projections of projects are connected using Delaunay triangulation [9].

A custom graph is such a graph where one project is in the main focus. This can be a project selected from the list of already imported projects. It can also be generated for a new project that is in the preparation phase, to see

[3]https://github.com/JozefStefanInstitute/fi-impact

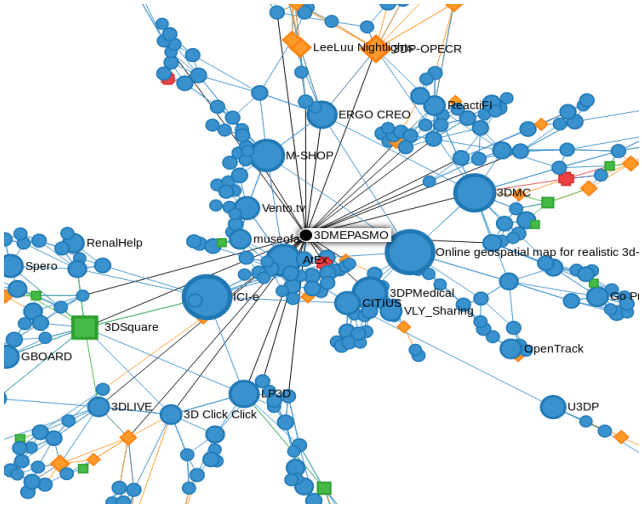

Figure 2: Multi-attribute based similarity graph for an arbitrary project. The selected project is represented with a black dot in centre of the figure.

where is it positioned in relation to other projects. A custom graph is generated by first extracting features from the selected project using the same feature extractors that were used for computing the main similarity graph. In second step, distance between each of the imported project and the selected project is computed (either cosine or euclidean distance can be used). Based on user selected threshold, an edge is created between more similar projects, while less similar projects are excluded from the graph. Finally, projects selected for the custom graph are connected based on the main similarity graph.

The final step of the similarity tool is graph visualization. The graph is visualized using SigmaJS[4] library and a plug-in that implements ForceAtlas2 [6] algorithm. The layouting algorithm mimics the physical attraction and repulsion forces, placing well connected and similar project closer on the two dimensional plane. An example of a similarity graph for a project is given in Figure 2.

3. ACCELERATOR BENCHMARKING/BEST PRACTICES REPORTING TOOLS

Accelerator Benchmarking/Best Practices Reporting Tools are written in *R* [8] programming language and software environment for statistical computing and graphics. In FI-IMPACT, several *R* modules[5] have been developed to automatically produce statistical reports about Accelerator Benchmarking and Best Practices based on FI-IMPACT databases.

The databases are imported and merged into an R data frame. The tool then produces correlation heatmaps, practice scores with accompanying mean difference t-test results and FI PPP phase projects connectivity information. Most of the results are produced in final-form CSV files. The tool also outputs several sets of diagrams (boxplots for all categorical indicators, CDF diagrams for practice scores and score histograms). All the diagrams are produced in two versions - with all sub-grantees and per-accelerator. Results are grouped into three main categories, as described below. The diagram in Figure 3 shows the structure of the tool.

The *degree of connectivity* tables are a result of analysis of FI-PPP project partnerships. They contain counts of different types of connections. Every pair of projects can have more than one connection, where each such connection between two projects in the three FI-PPP phases represents a partner that appeared in both.

Correlation analysis is performed on all numerical accelerator properties and all numerical subgrantee performance indicators by correlating each possible pair. The results include a heatmap of correlations, a table of correlations and a corresponding table with sizes of samples on which the correlations were computed.

Best practices identification produces notched box plots for categorical accelerator properties, CDF plots for binary accelerator properties, correlation outputs for numerical accelerator properties (heatmap and tables) and histograms for performance indicators. These can then provide insight into which practices coincided with better subgrantee performance.

We define a good practice as an activity performed by one or more of the Accelerators' consortia according to their acceleration plans, which based on objective evidence, is shown to have contributed to the good performance of subgrantees. Here, good performance of the subgrantees is primarily their market success (measured in terms of positive dynamics of revenue growth and customer growth); their ability to convince potential investors and collect additional funding ("traction"); if they are not yet on the market, their market readiness (measured by FI-IMPACT's KPIs scores).

The full report about the usage of developed tools, combined with qualitative data based on face-to-face interviews with Accelerators coordinators and interpretation of results is given in [1]. Some insights were drawn from the results computed by the benchmarking tool, while some of the conclusions formulated with other methods were confirmed by it [1]. For example the geographical scope of the partnerships did not seem to make a major difference, while the presence of professional accelerators had a strong positive influence on most subgrantee performance indicators.

There were several results computed in the benchmarking tool that did not provide any meaningful insights, but did confirm that the data exhibits some expected properties. For example, unsurprisingly, the benchmarking tool showed a strong positive correlation between investment per subgrantee and KPIs measuring feasibility and innovation.

4. CONCLUSIONS

We have shown how a set of IT tools and computational methods can be used to assess the socio-economic impact of a big investment programme, such as FI-PPP European programme for Internet innovation. On one hand, the implementation followed a strict methodology used for bench-

[4]http://sigmajs.org/
[5]https://github.com/JozefStefanInstitute/FI-Impact-R-analysis

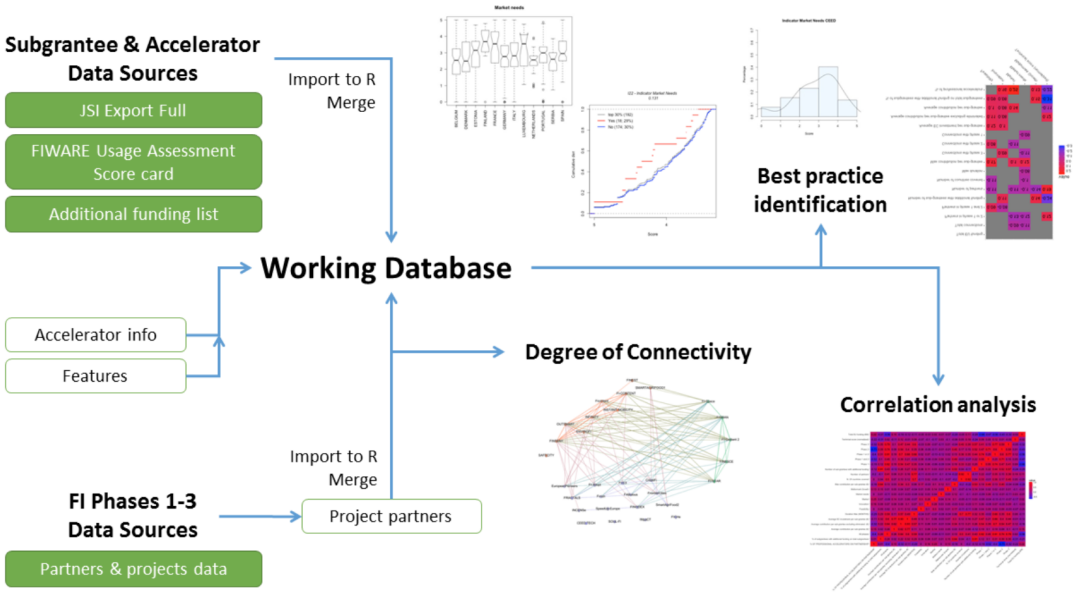

Figure 3: Benchmarking tool inputs and outputs

marking through questionnaires and calculation of respective KPIs, on the other hand, more exploratory techniques were used for statistical analysis and visualisation.

While it is true that many of the results of the statistical analysis implemented in the Accelerator benchmarking tools did not show statistically significant new results, it is also true that the results displayed most of the expected correlations. The statistical analysis explains little of the variations in performance of the subgrantees. In our opinion, the main weakness of the statistical correlation analysis was the very limited available data on actual market performance. However, weak signals, combined with with the results of the qualitative interviews, point to the positive role of professional accelerators within consortia, and positive impacts of practices such as workshops, matchmaking and providing gateways to further funding.

Impact Assessment and Similarity Graph benchmarking tools have been deployed and made available to over 700 subgrantees (projects), their mentors an coordinators. The tool was deemed very useful for long-term monitoring of sub-grantees performance. In this respect, the FI-IMPACT project consortium is drawing plans to assure long-term sustainability of the deployed IT systems.

5. ACKNOWLEDGEMENTS

This work was supported by the the ICT Programme of the EC under FI-IMPACT (FP7-ICT-FI-632840).

6. REFERENCES

[1] FI-IMPACT Project. Analysis of accelerators' good practices annex 8.5 to deliverable d2.4 update of impact assesment and forecast, 2016. Annex 8.5 of D2.4 Available from http://www.fi-impact.eu/page/docdownload/477/ [accessed 12 September 2016].

[2] FI-IMPACT Project. Fi-impact online assessment environment, 2016. Deliverable D4.3 Available from http://www.fi-impact.eu/page/docdownload/479/ [accessed 12 September 2016].

[3] FI-IMPACT Project. Impact assessment guidebook, v2, 2016. Deliverable D2.1.v2 Available from http://www.fi-impact.eu/page/docdownload/474/ [accessed 12 September 2016].

[4] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica (Slovenia)*, 29(4):497–504, 2005.

[5] B. Fortuna, J. Rupnik, J. Brank, C. Fortuna, V. Jovanoski, M. Karlovcec, B. Kazic, K. Kenda, G. Leban, D. Mladenic, A. Muhic, B. Novak, E. Novak, J. Novljan, M. Papler, L. Rei, B. Sovdat, and L. Stopar. Qminer: Data analytics platform for processing streams of structured and unstructured data. In *Software Engineering for Machine Learning Workshop, Neural Information Processing Systems*, 2014.

[6] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*, 9(6):1–12, 06 2014.

[7] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[9] P. Su and R. L. S. Drysdale. A comparison of sequential delaunay triangulation algorithms. In *Proceedings of the Eleventh Annual Symposium on Computational Geometry*, SCG '95, pages 61–70, New York, NY, USA, 1995. ACM.

Information flow between news articles: Slovene media case study

Jan Chołoniewski
Center of Excellence for Complex Systems Research, Faculty of Physics, Warsaw University of Technology, Koszykowa 75, PL-00662, Warsaw, Poland
choloniewski@if.pw.edu.pl

Gregor Leban
Artificial Intelligence Laboratory, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
gregor.leban@ijs.si

Sebastijan Maček, Aljoša Rehar
Slovenska Tiskovna Agencija, Tivolska 48, 1000 Ljubljana, Slovenia
{sm,ar}@sta.si

ABSTRACT

We present results of a study on usage of text similarity measures based on co-occurrence of words and phrases to classify a relation between a pair of news articles (i.e. no relation, both based on a common source, one based on the other). For each Slovenian article written in Slovene and published online on 27th June 2016, we found the most similar release from the Slovenian Press Agency (STA) database to obtain a list of candidate article-source pairs. Four experts from STA were asked to score the pairs, and their annotations were used to train classifiers and evaluate their accuracy.

1. INTRODUCTION

Propagating, exchanging, organizing and processing information are important parts of human social interactions on both micro- [1] and macro-level [2]. After years of local and nationwide scale, newspapers, press agencies and news outlets started to operate at global level using Internet. An easy and open access to their releases is desirable for news consumers and can be tracked e.g. with website traffic statistics or in social media. Article reuse by other publishers (authorized or not) is however not that straight-forward.

Combining natural language processing methods with data gathered in the Internet allows to quantify and measure social information processing phenomena [3, 4, 5, 6]. The advancements in NLP might serve all types of text-based media (in particular online news outlets) to provide tools to track spreading of their texts.

A tool that automatically finds articles based on a given article might be useful for news outlets and press agencies to track usage of their releases and to find cases of plagiarism or unauthorized use. Moreover, it might be applied to large scale news spreading studies [3]. A software-assisted plagiarism detection is a well-known problem in an information retrieval field [7], and using text similarity-based methods is one of the most popular approaches [8]. To the best of our knowledge, the following paper is the first published study of plagiarism detection in Slovene media supported by professional press agency workers.

The aim of the presented work is to check if text comparison methods based on co-occurrence of phrases can be successfully applied to determine a relation between two articles. Possible relations that we want to determine are (a) there is no relation, (b) they share a common source, or (c) one is based on the other one. To find the most efficient way to do that, we calculated cosine similarity of "bag of n-grams" representations of articles from Slovene media published on one day with releases from Slovenska Tiskovna Agencija (STA; Slovenian Press Agency) to preselect the most similar release to each article, asked experts to annotate the candidate pairs, and compared results for different thresholds and n-grams with the annotations.

The rest of the paper is structured as follows: in Section 2 we highlight a process of obtaining experimental data (candidate source release matching, expert annotation study), in Section 3 we describe applied methods and benchmark parameters, then in Section 4 we present results of classification study in two simplified cases; Section 5 contains a discussion and possible improvements, and Section 6 sums up the research.

2. DATA

Data for the study consist of randomly selected 469 articles out of 895 published on 27th June 2016 from 62 Slovene online news outlets as tracked by the EventRegistry [9]. For each article, we have found the most similar one in the STA releases database in terms of cosine similarity of two-, three- and four-word phrases ({2,3,4}-grams) occurrence vectors with TF-IDF weighting (see section 3 for details). A histogram of obtained similarities is presented in Figure 1. About 75% (354 out of 469) of candidate pairs obtained a cosine similarity below 0.1, and 10% (47 out of 469) over 0.9.

The pairs were scored by four experts from STA (A1, A2, A3, A4). They were asked to mark each pair with one of the following scores:

- NF – the proper source release has not been found despite it is present in the STA database,
- NC – the proper source release has not been found and it is not present in the STA database,
- DS – the proper source release has been found (although it might be one of many sources of the article),

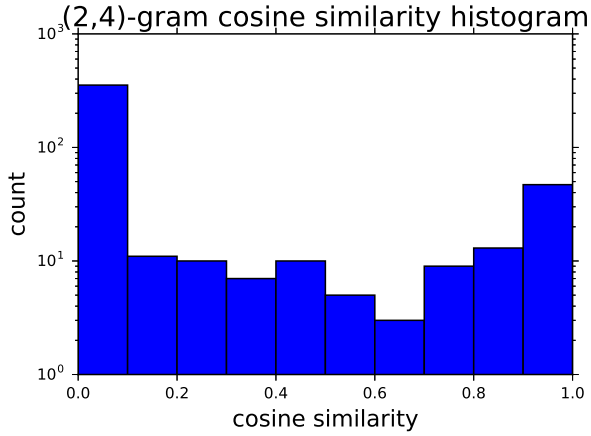

Figure 1: A histogram of {2,3,4}-gram cosine similarities of candidate pairs with a logarithmic Y-axis.

- IDS – the article and the proposed source release are both based on the same third party source.

In cases where the source was not found (NF), the annotators provided a link to the proper source release.

In Table 1, we present basic statistics of the annotations given by experts. We considered two methods of simplifying the annotations. The first one (A), merges DS and IDS marks to discriminate between two classes – a given pair contains pieces of the same information or is unrelated. The second one (B), merges IDS with NC – the algorithm's task is to check if one text is directly based on the other one.

person	total	NF	NC	DS	IDS
A1	469	3	315	98	53
A2	469	2	358	97	12
A3	95	0	61	23	11
A4	95	0	70	20	5

Table 1: Basic statistics of raw candidate release-article pairs annotations by the STA experts. total – a number of annotated pairs; NF – source not detected despite the source release is in the STA archive; NC – no source release in the STA archive; DS – one article is a direct source of the other; IDS – both documents based on the same third source article.

In Table 2, percentages of agreement among annotators are being presented for (a) raw annotations, (b) simplification A and (c) simplification B (see above).

The annotators were sometimes non-unanimous when both articles in a pair had a common source (compare Table 2a and 2b, mean agreement = 87%). They were more consistent when a release was a source of a given article (compare Table 2a and 2c, mean agreement = 96%).

Additionally, because of score inconsistencies, the final list has been prepared after discussing problematic cases.

	A1	A2	A3	A4
A1	100%	87%	86%	87%
A2	87%	100%	89%	89%
A3	86%	89%	100%	83%
A4	87%	89%	83%	100%

(a) Raw annotations

	A1	A2	A3	A4
A1	100%	88%	86%	88%
A2	88%	100%	91%	89%
A3	86%	91%	100%	84%
A4	88%	89%	84%	100%

(b) Simplified A – DS and IDS merged

	A1	A2	A3	A4
A1	100%	96%	99%	95%
A2	96%	100%	99%	96%
A3	99%	99%	100%	95%
A4	95%	96%	95%	100%

(c) Simplified B – IDS and NC merged

Table 2: Agreement among annotators.

3. METHODS

Articles and releases were mapped to "bag of n-grams" representations. Additionally, n-gram counts were transformed using term frequency-inverted document frequency (TF-IDF) weighting trained on a corpus of 5,000 randomly selected Slovene articles stored in the EventRegistry published during two weeks preceding the analyzed day. Terms which occurred in more than 25% of documents were discarded. Laplace smoothing with $\alpha = 1$ was applied to include terms which were not present in the corpus.

For $n = 1, ..., 5$, weighed term vectors of Slovene articles from 27th June 2016 were compared with all vectors of STA releases published between 20th and 27th June 2016 to find candidate source releases. For each n, we tested classifiers with a threshold from 0.00 to 1.00 with steps of 0.01 to find the threshold for which the method achieves the highest F1-score for A and B simplifications separately. The releases were compared with the list created using preselected pairs and experts' comments. A source release for a given article and a given threshold was considered as correctly found, when it matched the one annotated by human and cosine similarity score was above the threshold. A given article was considered correctly classified if a source release was correctly found or if the article was correctly marked as not having a source release in the STA database.

Parameters used to score the classification were accuracy, recall, precision, and F1-score. We used following definitions:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FN}$$

$$F1 = 2\frac{recall \times precision}{recall + precision}$$

where TP – number of articles with a correctly found source, TN – number of articles correctly marked as not having source in the STA database, FP – number of articles incorrectly marked as having source in the database, FN – number of articles incorrectly marked as not having source in the database. Cases when articles had incorrectly found source were counted separately as *errors*.

Each annotator could have scored differently each article-source pair thus the mean values and standard deviations of parameters were calculated when considering lists of annotations separately.

4. RESULTS

For each n value, we have found a threshold which maximized mean F1-score over all annotators. Results are shown in Table 3a for the simplification A and in Table 3b for the simplification B.

n	threshold	acc	σ_{acc}	F1	σ_{F1}	errors
1	0.29	0.90	0.02	0.83	0.04	18
2	0.09	0.91	0.02	0.84	0.03	4
{2,3,4}	0.06	0.91	0.01	0.84	0.02	3
3	0.05	0.91	0.01	0.84	0.02	4
4	0.04	0.90	0.02	0.83	0.02	3
5	0.03	0.90	0.02	0.83	0.02	6

(a) Simplified A – direct and indirect relations merged

n	threshold	acc	σ_{acc}	F1	σ_{F1}	errors
1	0.56	0.95	0.01	0.88	0.03	4
2	0.46	0.96	0.01	0.90	0.04	1
{2,3,4}	0.27	0.96	0.01	0.90	0.03	1
3	0.25	0.96	0.01	0.90	0.03	1
4	0.22	0.96	0.01	0.90	0.03	1
5	0.13	0.95	0.01	0.89	0.02	2

(b) Simplified B – indirect relations and lacks of relation merged

Table 3: Thresholds resulting with the best F1 for different ns. acc – mean accuracy, σ_{acc} – standard deviation of accuracy, F1 – mean F1-score, σ_{F1} – standard deviation of F1-score, errors – mean number of incorrectly found sources.

The results for the simplification A are satisfying when compared to the agreement among annotators. There were no significant difference between specific $n > 1$ but for $n = 1$ there were as many as 18 errors. The results for the simplification B are comparable with an agreement among annotators (see Table 2) and the classifiers could not find a correct source only in one case in which article was mainly based on some other release and only partially on the detected one. Again, there is very little difference among different n-grams which might suggest that in most cases articles use similar phrasing as the source release and the method is efficient.

In Figures 2 and 3, we show a histogram of cosine similarities and a stacked bar plot showing fraction of each score in the final list in each cosine similarity bin for $n = 1$ and $n = 3$ (respectively). The cosine similarities are not dramatically more separated in any of the cases but using $n = 1$ leads to significantly higher number of errors, and using $n = 5$ – to a slight increase of number of errors.

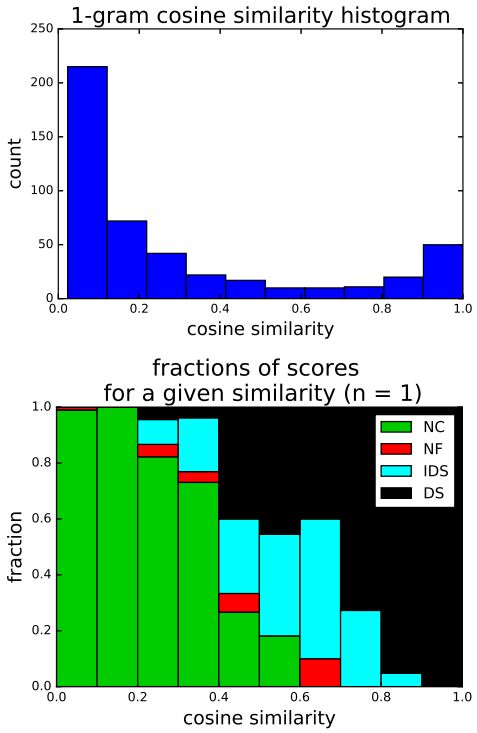

Figure 2: (top) A histogram of n-gram cosine similarities and (bottom) a fraction of each score in each similarity bin (see Section 2 for abbreviation expansions) for $n = 1$.

5. DISCUSSION

For most values of n, over 85% of candidate pairs had extreme cosine similarity values (below 0.1 or over 0.9). Two articles with cosine similarity equal to 1 are duplicates while the articles with cosine similarity equal to 0 are completely unrelated. Similarities between those values are not that clear to interpret. Obtaining more pairs with intermediate values would make results for boundary cases more reliable. After closer examination, very similar pairs which were marked as unrelated turned out to be annotators' mistakes. On the other hand, in the opposite cases (pairs with low cosine similarity but marked as related) the analyzed articles were rewritten; using lemmatization might be sufficient to identify them as similar.

Using different ns did not cause significant changes of accuracies and F1-scores of classifiers in both simplified cases but $n > 1$ allows to correctly find more sources than $n = 1$. In most $n = 1$ errors, the algorithm pointed at some more general release about a given topic.

We considered three types of relations between text pairs – lack of relation, common source, and direct sourcing (one based on the other). For the first and the last types of relation, it was usually possible to distinguish between them

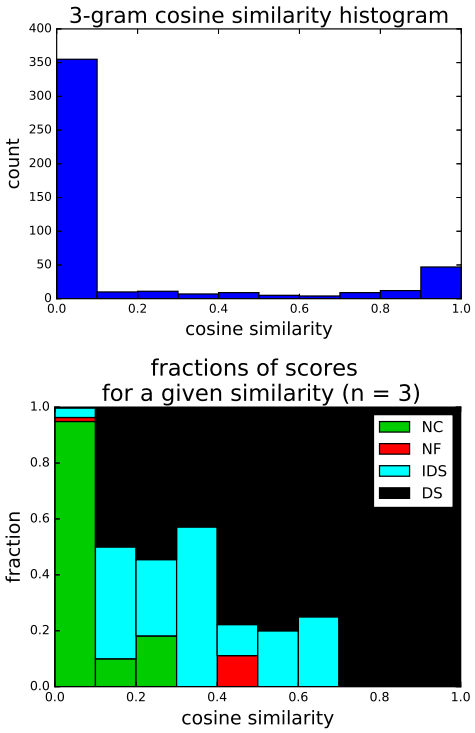

Figure 3: (top) A histogram of cosine similarities and (bottom) a fraction of each score in each similarity bin (see Section 2 for abbreviation expansions) for $n = 3$.

but in the proposed way it was not possible to accurately identify when two articles had a common source.

The method we used has not created a completely clear separation between considered relation types. In the further work the approach could be improved with lemmatization, mapping to WordNet synsets, discarding proper nouns, or a proper treatment of quotations.

It is also important to take into account that the experts were able to discriminate pairs because of their domain-specific knowledge. Nevertheless, even highly trained individuals scored some pairs differently. In many cases, there can be more then one source release of an article or an article might be based only partially on a given release.

An important future work will include use of cross-lingual techniques (e.g. [10]) to compute similarities and detect plagiarism in news articles in different languages.

6. CONCLUSIONS

We have presented a case study of estimating usage of STA releases by Slovene news outlets. We applied "bag on n-grams" representations of articles and releases with TF-IDF weighting, and compared them pairwise using cosine similarity. Detected candidate "article-source release" pairs were annotated by experts.

We compared results of automatic source detection with the annotations, and as expected found that articles have higher cosine similarity to releases when they are directly based on them, and can be detected with about 96% accuracy. A discrimination between not related and related pairs was possible with a 90% accuracy.

The results might be useful for a broader use although a partial supervision in boundary cases would be required. We suspect that lemmatization, proper quotations filtering and discarding proper nouns might result in achieving higher accuracies. Using cross-lingual similarity measures would be another interesting modification.

7. ACKNOWLEDGMENTS

We thank Aleš Pečnik for a technical support, and Tjaša Doljak, Naum Dretnik and Sabina Zonta from STA for annotations. This research has received funding as *RENOIR* Project from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691152. JCh has been also supported by Ministry of Science and Higher Education (Poland), grant No. 34/H2020/2016.

8. REFERENCES

[1] M.N. Bechtoldt, C.K.W. De Dreu, B.A. Nijstad, and H.-S. Choi. Motivated information processing, social tuning, and group creativity. *J. Pers. Soc. Psychol.*, 99(4):622–637, 2010.

[2] O. Oh, M. Agrawal, and H.R. Rao. Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quart.*, 37(2):407–426, 2013.

[3] K. Lerman. Social information processing in news aggregation. *IEEE Internet Comput.*, 11(6):16–28, 2007.

[4] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec. QUOTUS: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 798–808, 2015.

[5] A. Chmiel, J. Sienkiewicz, M. Thelwall, G. Paltoglou, K. Buckley, A. Kappas, and J.A. Hołyst. Collective emotions online and their influence on community life. *PLoS ONE*, 6(7), 2011.

[6] J. Chołoniewski, J. Sienkiewicz, J. Hołyst, and M. Thelwall. The role of emotional variables in the classification and prediction of collective social dynamics. *Acta. Phys. Pol. A*, 127(3):A21–A28, 2015.

[7] A Parker and JO Hamblen. Computer algorithms for plagiarism detection. *IEEE T. Educ.*, 32(2):94–99, 1989.

[8] D. Metzler, Y. Bernstein, W.B. Croft, A. Moffat, and J. Zobel. Similarity measures for tracking information flow. In *Proceedings of International Conference on Information and Knowledge Management*, pages 517–524, 2005.

[9] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 107–110, 2014.

[10] http://xling.ijs.si.

DATA ANALYTICS IN AQUACULTURE

Joao Pita Costa and Matjaž Rihtar
Artificial Intelligence Laboratory
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
joao.pitacosta@ijs.si, matjaz.rihtar@ijs.si

ABSTRACT

The specific challenges in aquaculture today reveal needs and problems that must be addressed appropriately and in sync with the most recent optimization methods. It is now the time to bring the techniques of aquaculture to a new level of development and understanding. In that, one must consider the state of the art methods of statistics and data mining that permit a deeper insight into the aquaculture reality through the collected datasets, either from daily data or from sampling to sampling data. This must also be tuned to the expert knowledge of the fish farmers, their procedures and technology in use today. In this paper we review the state of the art of data analytics methodology in aquaculture, the data available deriving from the procedures characteristic to this business, and propose mathematical models that permit a deeper insight on the data. We also address the data unknowns and strategies developed that will contribute to the success of the business, leading to discover valuable information from the data that can be made usable, relevant and actionable.

Categories and Subject Descriptors

E.3 Data Structures; I.2 Artificial Intelligence; I.6 Simulation and Modelling

General Terms

Algorithms, Data Science, Aquaculture

Keywords

Aquaculture, data analytics and visualization,

1. INTRODUCTION

Modern research and commercial aquaculture operations have begun to adopt new technologies, including computer control systems. Aquafarmers realize that by controlling the environmental conditions and system inputs (e.g. water, oxygen, temperature, feed rate and stocking density), physiological rates of cultured species and final process outputs (e.g. ammonia, pH and growth) can be regulated [2]. These are exactly the kinds of practical measurements that will allow commercial aquaculture facilities to optimize their efficiency by reducing labor and utility costs. Anticipated benefits for aquaculture process control and artificial intelligence systems are: increased process efficiency; reduced energy and water losses; reduced labor costs; reduced stress and disease; improved accounting; improved understanding of the process.

The technologies and implementation of the technologies necessary for the development of computer intelligent management systems come in a wide variety [8] and enhanced commercial aquaculture production [3]. Today's artificial intelligence (AI) systems offer the aquaculturist a proven methodology for implementing management systems that are both intuitive and inferential. The major factors to consider in the design and purchase of process control and artificial intelligence software are functionality/intuitiveness, compatibility, flexibility, upgrade path, hardware requirements and cost. Of these, intuitiveness and compatibility are the most important. The software must be intuitive to the user or they will not use the system. Regarding compatibility, the manufacturer should be congruent with open architecture designs so that the chosen software is interchangeable with other software products.

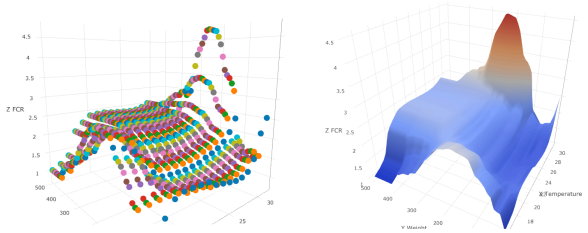

Figure 1. Dynamical plots developed for the project aquaSmart, available through a public interface where the fish farmers can upload their data and do a preliminary analysis and visualization.

The models presented in this paper were developed in the context of the EU project aquaSmart [1]. This project aims enhancing the innovation capacity within the aquaculture sector, by helping companies to transform captured data into knowledge and use this knowledge to dramatically improve performance. In particular, the tools constructed in that context (illustrated in Figure 1) serve the aquafarmers to evaluate feed performance, considering important factors such as the water temperature and average fish weight, but also underlying factors such as the oxygen level.

2. UNIQUE CHALLENGES

It is well known that the production in aquaculture has specific features and objectives associated with it. When talking about the adaptation of existing technology, the features important to the production in aquaculture come from weather prediction. These are the oxygen levels and water temperature, which are very specific to this activity. The tasks in fish farming carry several uncertainties – often expressed by measurements or even evaluations – that permit further optimization [9]. A classic example is the aim for a better control on the food loss and food quality. A contribution of data mining in this context would be of interest to the aquafarming industry, saving or relocating resources.

An important variable that remains undetermined during the complete production pipeline is the exact number of fish. A margin of up to 10% of number of fries is added to the initial production at time *t=0* due to uncertainty of number of deaths in the transport. That means that we already have a maximum of 10% more fish than our estimations (assuming that no fries die during transport or adaptation at *t=0*). Other than that we can only

have less fish than we estimated due to the lost fish because of unknown reasons. This is already an open problem at the level of the bounds for total amount of harvested fish and the description of best-case scenario and worst-case scenario. This represents a big lack of knowledge about production. In fact, the unknown number of fish until the end of the production is important for the amount of food given and, consequently, for the resources spent.

Feed composition has also a large impact on the growth of animals, particularly marine fish. Quantitative dynamic models exist to predict the growth and body composition of marine fish for a given feed composition over a timespan of several months [7]. The model takes into consideration the effects of environmental factors, particularly temperature, on growth, and it incorporates detailed kinetics describing the main metabolic processes (protein, lipid, and central metabolism) known to play major roles in growth and body composition. That showed that multiscale models in biology can yield reasonable and useful results. The model predictions are reliable over several timescales and in the presence of strong temperature fluctuations, which are crucial factors for modeling marine organism growth.

3. UNKNOWNS IN THE DATA

It is curious that the underlying problems with the data unknowns in aquaculture represent a problem of large dimensions for the industry of aquaculture, in which the production is straightforward. In fact, it is not known at any time in production, the exact number of fish in production, and therefore it is not possible to calculate with exactness the amount of food needed to support an appropriate growth. Furthermore, there are many conditionings in the progress of the production that must be taken into account and are hard to measure with the existing and available technology. In that, it is important to describe some of the features of the data including an assessment on its quality and measures to overcome obstacles to the analysis.

The input and output variables of the dataset are classified as: numerical and categorical. Numerical variables can be: continuous measured quantities expressed as a float (e.g. 'av. weight'); discrete quantities expressed as an integer (e.g. 'number of fish'). Categorical variables can be: regular categorical data including non-ordered classes (e.g. species Bream/Bass); or ordinal classes that can be ordered in levels (e.g. estimations poor/fair/good). From the variables that can be measured it is important to distinguish between: (i) variables that do not change over time, often identifying population attributes (e.g. identifications such as 'year' or 'hatchery'); (ii) variables that can change over time but do not change within a sampling period (e.g. 'batch'); (iii) variables that change daily, taken into account when samplings occur (e.g. 'average weight').

Table 1. Classification of values according to time dependence.

change in time?	**direct**	**calculated**	**derived**
yes	water temp.	FCR, SFR	av. weight
no	identification	av. weight at t=0	hatchery

Essentially we have four types of input data according to the impact they assure: (1) identification data that permits the fish farmer to manage the production and correctly identify the fish; (2) Daily data that is provided by the fish farmers resulting from their everyday data input (e.g. 'date', 'av. wt.', 'actual feed', etc.); (3) Sampling data, collected at predetermined points of the fish growth timeline, to confirm the model values and make the appropriate adjustments; (4) Life To Date (LTD) cumulative data that is calculated from the time when the fish enters the net as a fry to the date of data collection, and will last until the date of the harvest.

The identification data in input (1) is rather unspecific, as we cannot at this date in time identify the fish one by one as it is done in other animal farming such as cows and pigs. The data in this input category is distinguished between the group of production indicating localization - *Unit* - and the individual production series of fish - *Batch*. There is no further distinction in the identification. Batch has to go with Unit. Aquafarmers may have different batches in one unit or fish from one batch in many units.

The daily data in input (2) is recorded by the aquafarmers on a daily basis. These data columns follow the development of the fish since day one when it enters as a fry. The data inputted mostly follows one batch of fish from the beginning till the end of the production. One input data can have several units but, for purposes of the algorithms used, we consider only the time spent in one unit. For some of the algorithms used, the data is split this way (some data tables don't have values in the column 'harvest') with clear input/output within one unit.

The sampling data in input (3) serves the aquafarmer to improve/fix his/her initial Feed Conversion Ratio (FCR) model with real data. This includes features that can be learned by a specific set of data. Those features will later be important for the algorithms. They often correspond to columns with potential effect on the end result. Also, they can influence the production (e.g. 'feeder'). The software will adapt to data and will try to do the analysis and prediction from the available data. Note that the input will also include data columns unknown to the system and optional to the aquafarmer. We cannot predict the relevance of the data on those columns (neither their nature) but will consider them in the overall global analytics.

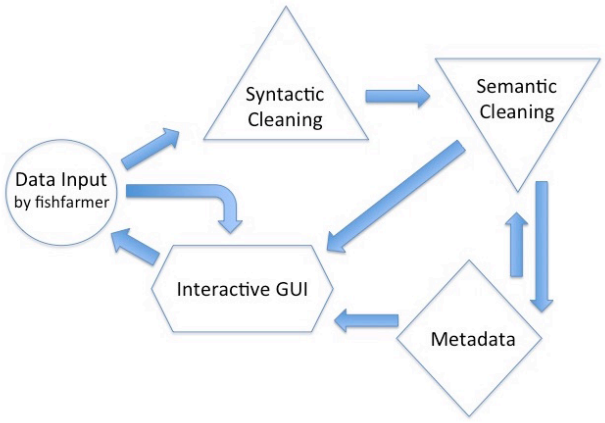

Figure 2. The proposed data cleaning process for aquaculture data, including the update of the metadata in the system and user interaction.

The daily data, the sampling data and the LTD data in inputs 2, 3 and 4 fall into three categories: (i) Direct values, that correspond to the direct observation of the aquafarmers on either variables values including small errors measured in the field (e.g. sampling measures such as average weight) or precise values provided by external sources (e.g. water temperature or oxygen level); (ii) Calculated values, that are dependent of a number of other observed values (e.g. LTD values calculated from the daily data); (iii) Derived values – values deriving from previously available

data (e.g. FCR calculated from the table, given average weight and water temperature).

The original data provided by the aquafarmers has variances/holes and is not precise because it is not measured automatically but instead entered by human hand (with some exceptions such as 'temperature'). Sometimes it is not entered for 1 or 2 days due to the bad weather, which complicates the access to the measurements and to the units themselves (sometimes this adds up to 4 days without entries). Sometimes this is due to intentional fasting to readjust features and in that case the data measurements stay the same as the ones in the previous fields, just before fasting takes place. The major discrepancies should be pushed to the user as a compromise. If the data is missing up to a certain threshold, the data will be sent back to the user in order to be inputted once again after appropriate corrections. The options for the missing data problem are to consider it as an error and report it to the user requesting the missing data, or consider the average from the missing data in the sense of interpolation on a fixed mesh grid.

4. DATA ANALYTICS IN AQUACULTURE

Mathematical modeling aims to describe the different aspects of the real world, their interaction, and their dynamics through mathematics. It constitutes the third pillar of science and engineering, achieving the fulfillment of the two more traditional disciplines, which are theoretical analysis and experimentation [4]. Nowadays, mathematical modeling has a key role also in aquaculture. In the following section we will present an overview of that. Growth and reproductive modeling of wild and captive species is essential to understand how much of food resources an organism must consume, and how changes to the resources in an ecosystem alter the population sizes [6].

The FCR is an important performance indicator to estimate the growth of the fish. It is widely used by the aquaculture fish farmers in pair with the Specific Feeding Ratio (SFR). Its importance follows from the fact that 70% of the production costs in aquaculture are assigned to the food given to the fish during growth. Some of it will fall through the net and some will be spared. The optimization of the feeding of the fish can carry great benefits to the economic development of the fish farms.

Specifically, the FCR permits the aquafarmer to determine how efficiently a fish is converting feed into new tissue, defined as growth [10]. Recall that the FCR is a ratio that does not have any units provided by the formula:

$$\text{FCR} = \text{dry weight of feed consumed/wet weight of gain}$$

while the feed conversion efficiency (FCE) is expressed as a percentage as follows:

$$\text{FCE} = 1/\text{FCR} \times 100$$

There seems to be some controversy among aquatic animal nutritionists as to which is the proper parameter to measure, but in aquaSmart we used FCR (exposing here FCE for completion). Moreover, the FCR and FCE are based on dry weight of feed and fish gain, as the water in dry pelleted feed is not considered to be significant. A typical feed pellet contains about 10% moisture that will only slightly improve the FCR and FCE.

The FCR table allows the fish farmer to assess the amount of food to give to the fish according to their average weight and the temperature of the water. Each farm has its own FCR table. This is an opportunity to create our own table/model by tweaking the numbers accordingly. Also specifying the influence of sexual maturity and the lack of oxygen, which are done by hand/intuition, have features to take in consideration by the math model. The FCR models in this paper consider only temperature and average weight.

Each aquaculture entity draws an appropriate FCR table to that batch of fish. Higher temperature leads to lower energy spent and faster growth, and consequently to a lower FCR. As the fish gets bigger, he needs more food to increase his biomass in percentage, and thus the FCR grows higher with the increase of the average weight. The quality of the food and the size of the pellet size are not considered at this point. At high temperatures (above 30 degrees in the case of bream and bass) low oxygen leads to low conversion to biomass. This is one of the hidden variables in the model, which should be considered separately at a later stage. One of the possibilities would be to penalize the FCR tables for the lack of oxygen. The other variable is the high reproduction of the fish in low temperatures and high average weight, which highly affects the growth of the fish. Recall that the Economic FCR is the real FCR index following from the quotient between food given to the fish and the fish biomass. When the temperature is too high or too low we should ignore the data that is filled in with zeros and considered empirical data.

In the following we present the plots of the models for the three fish farms in AquaSmart. It includes 3 fish farms.

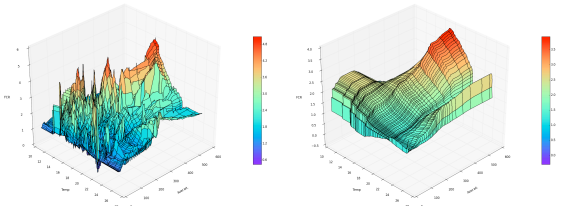

Figure 3. Company A: Real data (on the left) and FCR model (on the right) for the bream production.

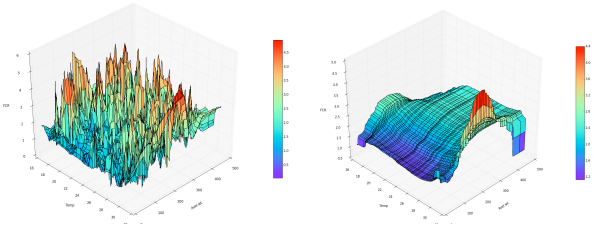

Figure 4. Company B: Real data (on the left) and FCR model (on the right) for the bream production.

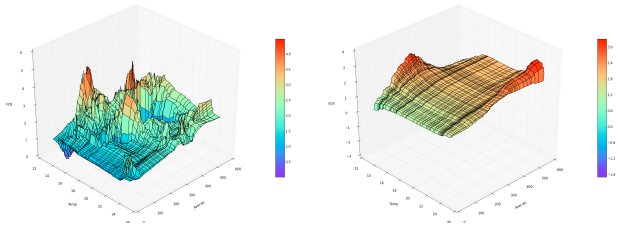

Figure 5. Company C: Real data (on the left) and FCR model (on the right) for the bream production.

The model (on the right) produced based on the sample data (on the left) serves as a base of comparison with the historical data provided by a particular fish farm. Thus, with the new real data getting in our system, the fish farmer can compare it with the

model and make an evaluation on the progress of the production. These models complement and confirm the expert knowledge: the high values on the right correspond to high fish reproduction in cold water temperatures and high average weight values. On the other hand, high temperatures represent low levels of oxygen which request higher feeding rate to maintain and increase the growth rate.

The big number of peaks in the real data, plotted on the left, correspond to the real values. Typically the input data can be seen within a grid. The following images show the grid view of both the real data (on the left) and the FCR model (on the right) for the company C.

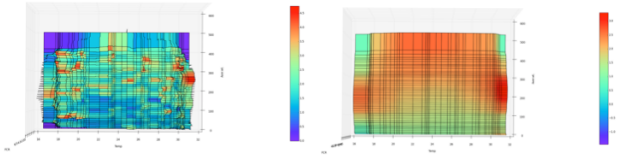

Figure 6. The grid view of both the real data (on the left) and the FCR model (on the right) for the company C.

We then use least squares method to interpolate the missing values including all non-peak values as those interpolated values. It does so by approximate the solution of overdetermined systems. The average weight must be represented using specific values that are important in the fish production decision making, and eventually distinct from fish farm to fish farm. Thus we consider a second interpolation to produce a final FCR table that is consistent with the systems in use by the fish farms. The nearest neighbours algorithm is used here to find the values outside the area [5]. That permits us to consider the complete table of measurements in line with the sample data available and the missing values calculated for the area inside the region.

5. CONCLUSIONS

The challenges of aquaculture for data analytics are very specific in the field and must be addressed with the appropriate methodology and technology, in tune with the expertise of the fish farmers. The uncertainty of measures, such as the number of fish until the time of harvest, derives in variances that do not permit a complete accuracy of some of the calculations. This is particularly important to some of the available tools to monitor the business, such as the feed conversion rate tables in use by the fish farmers to optimize the production costs.

The mathematical models developed in the aquaSmart project and discussed in this paper aim to contribute to the improvement of the aquaculture procedures, providing a deeper insight on the information retained in the collected data, using state-of-the-art methods of data mining in line with the expert knowledge of the field transferred to the metadata in the data store.

Moreover, the statistical analysis of the results permit a clearer visualization of the important features in the data that can boost the production and optimize the processes related to it. That will enable classification and forecast based on the analytics of the available data.

In that, future work includes the production of guidelines validated by end-users in order to facilitate the application of further advanced learning methods in aquaculture.

ACKNOWLEDGMENTS

The authors would like to acknowledge that his work was funded by the project AquaSmart, co-funded by the European Commission with the agreement number 644715 (H2020 Programme). We would also like to thank to Primož Škraba, Luka Stopar, Dunja Mladenić, Pedro Gomes de Oliveira, Ruben Costa, Gerasimos Antzoulatos, Nir Tzohari, John McLaughlin and Gary McManus.

REFERENCES

[1] aquaSmart Consortium, The aquaSmart Project [Online]. URL: www.aquasmartdata.eu. [accessed in 22.8.2016].

[2] Bhujel, R. C. (2011). Statistics for Aquaculture. John Wiley & Sons.

[3] Beveridge M (2004); Cage Aquaculture; Third Edition, Oxford, UK.

[4] N. R. Draper and H. Smith, Applied regression analysis, 3th edition. New York: Wiley, 1998.

[5] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Elsevier.

[6] Dobson, A. J., & Barnett, A. (2008). An Introduction to Generalized Linear Models, Third Edition. CRC Press.

[7] Bar, N. S., & Radde, N. (2009). Long-term prediction of fish growth under varying ambient temperature using a multiscale dynamic model. BMC Systems Biology, 3(1).

[8] Lee P.G. 2000. Process control and artificial intelligence software for aquaculture. Aquacultural Engineering, 23, 13-36.

[9] Rizzo, G., and Spagnolo, M. 1996. A Model for the Optimal Management of Sea Bass Dicentrarchus Labrax Aquaculture. Mar. Resour. Econ. 11: 267–286.

[10] Stickney, R. R. (2007). Aquaculture: An introduction. (C. Publishing, Ed.). CABI Publishing.

Modeling Probability of Default and Credit Limits

Zala Herga
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
zala.herga@ijs.si

Jan Rupnik
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
jan.rupnik@ijs.si

Primoz Škraba
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
primoz.skraba@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Jamova 39
Ljubljana, Slovenija
blaz.fortuna@ijs.si

ABSTRACT

Creditors carry the risk of their clients not meeting their debt obligations. In the literature, these events are often referred to as *default events*. These can be modeled for each company through a *probability of default* (PD). Measures can be taken to limit the default risk: in this paper we focused on credit limit. Firstly, we predict PD of a company using a logistic regression model, based on publicly available financial data. Secondly, we effectively find an optimal portfolio under risk aversion constraints and show how variation of inputs affects the results.

Categories and Subject Descriptors

Mathematics of computing [**Mathematical optimization**]: [Linear programming, Convex optimization]; Computing methodologies [**Machine learning**]: [Supervised learning by regression]

Keywords

PD model, logit, credit limit model, portfolio optimization, linear programming, risk management

1. INTRODUCTION

Payment defaults represent a key default risk (also credit risk) to creditors. Creditors can limit their risk by either insuring their claims or taking preventative measures before extending a credit. Standard tools to measure default risk include different kinds of credit ratings.

Our goal was to create a model that predicts a company's *probability of default (PD)* and provides credit limit suggestions based on the computed PD. One of our constraints was that the underlying PD model be simple and easy to understand. For credit limits, we implemented a linear programming based approach [5] which provides portfolio optimization with risk aversion constraints.

This paper presents the workflow and the methodology that we employed to build a portfolio credit allocation model. Due to privacy concerns it does not include any expiremental results and does not discuss any concrete results or aspects of real data.

The paper is organized as follows. Section 2 provides an overview of related work. In Section 3 the data that is used in modeling is described. Section 4 first describes the approach and computation of the PD model and then presents the results. Section 5 provides a short theoretical introduction to portfolio optimization and then presents our computation and results. Section 6 concludes the paper.

2. RELATED WORK

The Altman Z-score [1] is a widely used credit-scoring model. It is a linear combination of five commonly used financial indicators and it predicts company's degree of PD. Both [6] and [8] argue that Altman Z-score and distance-to-default ([7]) are not appropriate to use in the context of small businesses. The authors in [6] predict PD using delinquency data on French small businesses. They propose a scoring model with an accuracy ratio based solely on information about the past payment behavior of corporations. Similarly, [8] forecasts distress in European SME portfolios. They estimate the PD using a multi-period logit model. They found that the larger the SMEs, the less vulnerable they are to the macroeconomic situation. They also show that SMEs across Europe are sensitive to the same firm-specific factors.

[2] examine the accuracy of a default forecasting model based on Merton's bond pricing model [7] and show that it does not produce a sufficient statistic for the PD. [11] compare the predictive accuracy of PD among six data mining methods. They also present a novel "sorting smoothing method" for estimating the real PD. Using a simple linear regression, they show that artificial neural networks produce the best forecasting model. [3] used the Merton model to show that, on contrary to what theory suggests, the difference in returns between high and low PD stock is negative and that

returns almost monotonically decrease as the PD increases. However, they found a positive relationship systematic default risk exposure and returns.

On the problem of portfolio optimization, [9] showed that *Conditional Value at Risk* minimization with a minimum expected return can be computed using linear programming techniques. [5] built on this idea and showed that alternatively, one can maximize returns while not allowing large risks.

3. DATA

The dataset that we used covers several thousand companies from several European countries. Data for each company consists of two parts: financial data and trading data.

Financial data corresponds to publicly available data - balance sheets and income statements of a company.

Trading data consists of private information on trades between our data provider and his clients. It contains monthly data about the sum of trades, outstanding debts, disputed claims and delayed payments. The data is available for years between January 2010 and June 2016.

4. PD MODEL

All creditors carry the risk of their clients not paying the bills. We can be quite certain that some clients will not pay, however, we have difficulties identifying those clients. Hence, our first task was to compute the probability of default for each client company. The end model should also be simple and easy (intuitive) for domain experts to understand.

Default can be defined in multiple ways, considering the available data and how strict we want our model to be. Do we consider it a default if the client is only one day late on payments? Or do we let them be 30, 45, 60 days late before taking action? Are we going to consider a client defaulted if he owes 10€? What if the client didn't pay one bill but has been paying all the bills after? We are required to make a judgment call and choose a definition that meets the creditor's needs best.

As soon as a client defaulted we removed it from the dataset. Meaning, each client can only default once and after that we assume there was no more trading done with them.

There were some companies that were already defaulted at the beginning of the time-span of our dataset. We removed these companies from the dataset since they provide no useful information. We also filtered out clients with low sales since their financial data can be very unreliable and their impact on our model is big compared to the trading volume that they generate.

Due to the constraint that the model should be simple and easy to understand we chose to model the PD with logistic regression.

From the available financial data we calculated 45 financial indicators for each company that cover aspects of solvency, liquidity, debt, profitability and operative effectiveness status of a company. We transformed each financial indicator vector into a feature vector by binning and assigning *Weight of Evidence (WOE)* [10] to it. The idea is as follows: create n bins in range from min to max indicator value and assign each company to the corresponding bin. Then count the number of 'bad' (defaulted) and 'good' companies in each bin. Then assign WOE to a bin as

$$\log \frac{\mathbb{P}(company = good)}{\mathbb{P}(company = bad)}.$$

Since WOE scores will be used as inputs to a linear model, they should be a monotonous function over bins, meaning, the higher the financial indicator value the better the company is (if WOE is increasing) or the higher the indicator the worse the company is (if WOE is decreasing). In Figure 1 we show an example of binning and WOE transformation on simulated data.

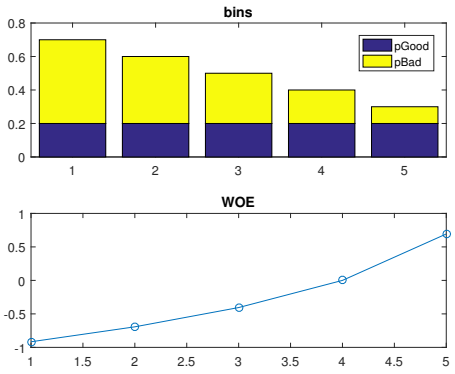

Figure 1: Bins and WOE on simulated data. The greater the value of the simulated data the greater the evidence that a company is 'good'.

Thus, we obtained 40 out of 45 features. Another feature was the size of the company (based on income) and four of them were country features (dummy variables - one for each country).

As described above, we mapped "raw" features into WOE features:

$$(x_1, x_2, ...) \rightarrow (woe(x_1), woe(x_2)..)$$

PD of a company is then predicted through logistic distribution function:

$$F(x) = \frac{1}{1 + \exp^{-(\beta_0 + \beta_1 \cdot woe(x_1) + \beta_2 \cdot woe(x_2) + ... + \beta_n \cdot woe(x_n))}}$$

where β_i denotes linear regression coefficients.

4.1 Computation

Since we have the response variable observations on monthly level, we interpolated features to obtain the same frequency of explanatory variables. We also took into account offset of the financial and trading data: financial data is only made available sometime around June each year for the previous year (e.g. in June 2016 we only know financial statuses of 2015).

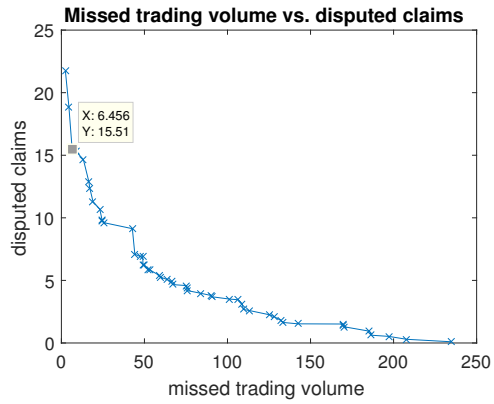

Figure 2: Missed trading volume vs. disputed claims on simulated data. The info box marks 6%-level: if trading was stopped with the worst 6% companies, approx. 7.5 could be saved on disputed claims and $6.5 \cdot margin$ profit would be lost on trading volume. The figure is included for illustrative purposes and is not based on real data for privacy concerns.

We then filtered out features that:

- had high percentage of missing values (We kept threshold as a parameter. in this paper, threshold of 0.7 was used.)
- were highly correlated (In this paper, threshold of 0.9 was used).

After that, 27 features were left in our feature set.

We used Matlab for computation. The data was trained on 42 months and tested on 19 months. Two models were trained: standard logistic model and stepwise logistic model. In stepwise regression, explanatory variables are added to a model by an automatic procedure [4]. Regression coefficients are estimated by maximum likelihood estimation.

4.2 Results

Test data consisted of 20% of the data (not used in training). There were 16 features chosen in stepwise model. Both of the models have 0 p-values and return very similar results in predicted values.

We evaluated models by comparing the amount of disputed claims (true negatives) to the amount of missed trading volume (false positives) given ceased trading with companies with PD exceeding some threshold (Figure 2). Note, that expenses based on missed trading volume cannot be directly compared to disputed claims; disputed claims are a direct expense, whereas missed trading volume number consist largely of expenses (that a company in that case did not have). Hence, one needs to multiply the trading volume with company's (average) margin to obtain actual opportunity costs.

5. CREDIT LIMITS MODEL

Naturally, a question arises once we identify risky clients: how to handle them? Client should have set a credit limit, but how to set the limit? If the limit is to high, the client might not be able to pay the bills, but if the limit is too low, profit is lost on trading volume. The model that we created is inspired by [5], is based on PD calculation presented in first part of this paper and takes into account the level of risk that creditor is willing to take.

Let us introduce some standard financial risk-related terms. *Value at Risk* (VaR) is an upper percentile of loss distribution. Probability level is denoted as α. E.g. $95\% - VaR$ of $1,000,000€$ means that there is a 0.05 probability that loss will exceed $1,000,000€$. *Conditional Value at Risk* (CVaR) is the conditional expected loss under the condition that it exceeds VaR. CVaR at $\alpha = 95\%$ level is the expected loss in the 5% worst scenarios. ω denotes the maximum allowed CVaR of the portfolio at level α.

We will denote loss associated with the portfolio x and random vector y (with density $p(y)$) as $f(x, y)$; $CVaR$ will be noted as $\phi_\alpha(x)$, which is given by

$$\phi_\alpha(x) = (1-\alpha)^{-1} \int_{f(x,y)>VaR_\alpha} f(x,y)p(y)dy.$$

It has been established [9] that $\phi_\alpha(x)$ can be computed by minimizing the following function:

$$F_\alpha(x,\zeta) = \zeta + (1-\alpha)^{-1} \int_{y\in\mathbb{R}^n} \max\{f(x,y)-\zeta, 0\}p(y)dy,$$

and that the value ζ which attains the minimum is equal to VaR_α.

Finding credit allocations $x \in X$ that maximize the expected profit under $CVaR$ is equivalent to the following optimization problem [5]:

$$\begin{aligned} \min_{x\in X, \zeta\in\mathbb{R}} \quad & -R(x) \\ \text{subject to} \quad & F_\alpha(x,\zeta) \le \omega \end{aligned} \tag{1}$$

where $R(x)$ is the expected profit and the set X is given by a set of box constraints (lower and upper bounds on each component of x).

5.1 Computation

By using the PDs from the first part of the paper, we can simulate the default events and compute several random scenarios for our portfolio. Since each company is assigned a probability of default we can generate random scenarios (where certain companies default) over the full portfolio by sampling from independent (with different weight) Bernoulli random variables.

By generating a set of sample scenario vectors $y_1, \ldots, y_J$ with their corresponding probabilities $\pi_1, \ldots, \pi_J$ we can approximate $F(x, y)$ by a finite sum:

$$\tilde{F_\alpha}(x,\zeta) = \zeta + (1-\alpha)^{-1} \sum_{j=1}^{J} \pi_j \max\{f(x,y)-\zeta, 0\}$$

Using

$$z_j \ge f(x,y_j) - \zeta, \quad z_j \ge 0, \quad j = 1,\ldots,J, \quad \zeta \in \mathbb{R}$$

the constraints in (1) can be reduced to a system of linear constraints:

$$\zeta + (1-\alpha)^{-1} \sum_{j=1}^{J} \pi_j z_j \leq \omega \tag{2}$$

$$f(x, y_j) - \zeta - z_j \leq 0, \quad z_j \geq 0, \quad j = 1, \ldots, J, \quad \zeta \in \mathbb{R} \tag{3}$$

In our case, the expected profit is

$$R(x) = (1 - pd) \cdot x \cdot margin - pd \cdot x.$$

As for PD modeling, we used Matlab for computation. We combined left-side part of constraints from (2) and (3) in a matrix A; the right-hand side was combined in a vector b. We also added additional constraints on x, that are specific to our problem: we set an upper and lower bound (ub and lb correspondingly) to the credit limit.

We then have to solve linear program:

$$\min_x -R(x) \quad \text{such that} \quad \begin{cases} Ax \leq b \\ lb \leq x \leq ub \end{cases}$$

5.2 Results

Our method provides an optimal portfolio based on α, ω, margin, PDs, credit limit upper- and lower bounds. By optimal portfolio we mean monthly credit limit for each company, which takes value between the provided credit limit bounds. In Figure 3 we present some scatter-plots based on results; the default values used for these graphs are $\alpha = 0.95$, $lb = 0$, $ub = max\ trading\ volume$ (based on historical data) and $margin = 0.01$. Most companies get either zero or maximum credit approved; there are only some companies the get part of the max credit approved. In 3b we decreased ω by a factor of 10. This is a lot stricter constraint and consequently there are more companies that get zero credit approved and many companies that get only a part of max credit approved. In 3c we increased margin from 0.01 to 0.1. In practice this means, that the credit-giver is making profit on trading volume. In 3d, we moved credit lower bound from zero to $0.1 \cdot ub$. This makes the PD threshold stricter.

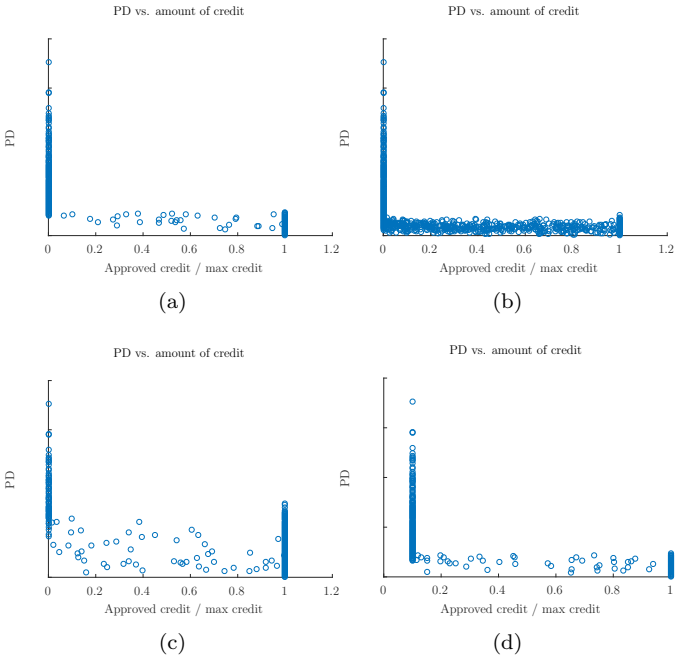

Figure 3: Each circle in these scatter-plots represents one company. The x-axis represents the relative amount of approved credit. Y-axis represents predicted PD of a company. In 3a default values are used, while Figure (b) is based on reducing ω by a factor of 10, figure (c) is based on increasing the margin by a factor of 10 and figure (d) is based on lifting the lower bound to 10% of the upper bound. The scales of PDs are omitted on the graphs due to privacy concerns.

6. CONCLUSION

We presented a logit model based on *weight of evidence* features to predict a company's *probability of default*. Standard and stepwise methods were used to train the data. Both methods provide similar results.

In second part of the paper we introduce an efficient portfolio optimization technique that was used to determine credit limits for creditor's clients. We presented the results and showed how variation of inputs impacts the results.

7. REFERENCES

[1] E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[2] S. T. Bharath and T. Shumway. Forecasting default with the kmv-merton model. In *AFA 2006 Boston Meetings Paper*, 2004.

[3] S. Ferreira Filipe, T. Grammatikos, and D. Michala. Pricing default risk: The good, the bad, and the anomaly. *Theoharry and Michala, Dimitra, Pricing Default Risk: The Good, The Bad, and The Anomaly (March 2015)*, 2015.

[4] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

[5] P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

[6] A. Marouani. Predicting default probability using delinquency: The case of french small businesses. *Available at SSRN 2395803*, 2014.

[7] R. C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, 29(2):449–470, 1974.

[8] D. Michala, T. Grammatikos, S. F. Filipe, et al. Forecasting distress in european sme portfolios. *EIF Research & Market Analysis Working Paper*, 17, 2013.

[9] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[10] D. Sharma. Evidence in favor of weight of evidence and binning transformations for predictive modeling. *available at: http://ssrn.com/abstract=1925510*, September 2011.

[11] I.-C. Yeh and C.-h. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

Big Data Analysis Combining Website Visit Logs with User Segments and Website Content

Matic Kladnik
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
matic.kladnik@ijs.si

Blaž Fortuna
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenia
blaz.fortuna@ijs.si

Pat Moore
Bloomberg LP
New York, USA
pmoore26@bloomberg.net

ABSTRACT
This paper provides three use-cases of combining website visit logs with user segment data, website content and stock volume data. The use-cases provide concrete examples showcasing how to derive insights into behavior of users on the website. We also present how to efficiently derive required data using MapReduce technique, and implement the use-cases using QMiner data analytics platform.

Categories and Subject Descriptors
Information Systems → World Wide Web → Web Mining

General Terms
Algorithms, Management

Keywords
Web log analysis, Big data, MapReduce, Hadoop, Hive, QMiner

1. INTRODUCTION
Visitors on modern websites leave behind large amounts of traces in form of a web server log, query logs, pixel tracking logs, etc. Website owners analyze these traces to better understand how visitors navigate their website, identify their interests, and optimize advertising opportunities. Most typical use-cases are covered by end-user tools such as Google Analytics, ComScore, Omniture, etc. More sophisticated use-cases, which require custom or ad-hoc processing of the data, are covered by big data frameworks such as MapReduce and its implementations such as Apache Hadoop. However, standard tools provide little or no support for analysis that requires combining visit logs with content presented on the website. In this paper we present three use-cases which combine big data frameworks and text mining approaches with the goal of deeper understanding of visitors and their habits.

Paper is organized as follows. First we provide a quick overview of the input data, and outline how we prepare it for the use-cases using MapReduce framework. We conclude by outlining three use-cases applied to data from a large news website: comparing user segments, identifying user segments by example article and correlating company news visit logs with stock volume.

2. INPUT DATA

2.1 Website Visit Logs
Visit logs provide log of page views on a specific website. Each page view is described by the time, user ID (stored as a first-party cookie), page URL, referral URL, user agent, etc. Such information is typically collected using a pixel tracking mechanism and is commonly used by all standard web analytics tools. In use-cases we use ComScore as the data provider,

2.2 User Segment Data
Web ecosystem allows for deeper annotation of users based on their behavior around the web. For example, visit to a car dealership website can be noted by a third-party data provider, which has an agreement and some tracking mechanism installed on the dealership's website. Such *user segment* data is further aggregated and distributed by data providers such as Krux. Connecting visitors to our website with external database of user segments can provide a much richer understanding of our audience and, as we will see in the first use-case, allows for some interesting analysis. Example of user segments covered include estimated household income, gender, estimated home net worth and others.

2.3 Website Content
Content of the pages on our website provide additional source of data we can use in the analysis when joined with visit logs and user segment data. Content can be represented on different levels of granularity: words, entities, topics, etc.

In our use-cases we collect the content by crawling the URLs mentioned in the visit logs and extracting their content by removing boilerplate. Each page is processed using a standard NLP pipeline [1], providing us with list of topics, named entities and tickers (stocks from companies).

2.4 Stock Volume Data
Our last use-case combines visit logs with market data. We use price and trading volume data provided by Kibot. In this paper we will be using data in hourly intervals, meaning that trade volume values are accumulated by hours.

3. PROCESSING
We process input data using MapReduce parallelization paradigm. MapReduce processes the parts of data separately in the *map* phase and later joins the partial results in the *reduce* phase. This allows us to easily split processing into multiple computing units that can be executed in parallel. In this paper we use Apache Hadoop [2] as MapReduce implementation and Apache Hive [3], which allows us to execute SQL-like queries on Hadoop.

3.1 Aggregating Visit Logs and Segment Data
First we address the task of joining visit logs with user segment data using a cookie shared by both datasets. Join can be executed using the query presented in Figure 1. The query joins three tables: a table with the visit logs, a table linking user IDs from visit logs and segment IDs from segment data, and a table providing segment description.

We tested the query on one week of visit logs on several cluster compositions and the results are presented in Table 1. We can see how the performance of the cluster improves when adding additional instances. However, when going from 4 to 6 instances, we can see a smaller deduction in time taken. That is due to the last

reduce process taking similar amount of time with any number of instances.

Table 1. Performance analysis on segment query

Instances	Duration	Improvement
2	3104 seconds	--
4	1763 seconds	x1.76
6	1427 seconds	x2.18

```
INSERT INTO TABLE visitors_segids
SELECT seg_userid, segid
FROM visit_logs t1, segment_data t2
WHERE t1.seg_userid = t2.seg_userid;

INSERT INTO TABLE visitors_segvals
SELECT seg_userid, segid, segment_title
FROM visitors_segids t1, sikdd_segments t2
WHERE t1.segid = t2.segid;
```

Figure 1. The query creates mapping between user IDs and segment names by joining visit logs and data segments.

```
INSERT INTO TABLE user
SELECT userid, collect_set(url)
FROM sikdd_visit_logs
GROUP BY userid;
```

Figure 2. Aggregating visited URLs by users IDs.

3.2 Aggregating Visit Logs and Content

We now address the task of aggregating visited pages by user. We achieve this by simply grouping visit logs around user ID as show in Figure 2 with the following example query.

Results of the query are inserted into a previously created table. We are calling the *collect_set* function, which returns an array of targeted column values when grouping the results. This way we get a column with distinct user ID values and all connected URLs. In Table 2 we compare performance on this query when for several cluster compositions and can see linear improvement as we increase the number of instances.

In Table 3 we compare this with the performance of the cluster on internal Hive table, as opposed to ad-hoc table pulled from some external source (e.g. Amazon S3). As we can see, the performance can be boosted by reading data from an internal table. When using an internal table, the data is stored locally on the instance. External tables are useful for reading data from an external source (e.g. AWS S3 service), which has to support the HDFS (Hadoop File System). External tables can still be stored on the local HDFS of the instance. Copying data from an external source to an internal table comes in handy when a certain table is used frequently, otherwise the difference in time taken is overturned by the amount of time needed to copy data from an external data source to an internal table. There is also a question of resources availability as moving data to the local instance can fill a lot of the cluster's available storage space. In some cases, the data is simply too large to be moved to the cluster's storage.

As we can see in Table 1, Table 2 and Table 3, adding additional task instances to the cluster can greatly affect the performance. However, at some point performance is improved less effectively when adding additional task instances. This depends on the amount of data, number of files the data is stored in and types of task instances. The last reduce task usually takes some time to complete and is processed on a single instance, which is why it's time taken to process cannot be improved.

Table 2. Performance on visit logs data in external table

Instances	Duration	Improvement
2	713 seconds	--
4	446 seconds	x1.60
6	344 seconds	X2.07

Table 3. Performance on visit logs data in internal table

Instances	Time taken	Improvement
2	275 seconds	--
4	265 seconds	x1.04
6	234 seconds	x1.18

4. USE CASES

In this section we present three use-cases that combine visit logs, segment data, content and stock volume data. Use-cases were implemented in QMiner [4], a data analytics platform for processing large-scale real-time streams containing structured and unstructured data.

4.1 Comparing User Segments

The goal of this task is to compare behavior of different user segments on the website. Segments are provided by User Segment data, which we joined with visit logs and website content. The use-case is prototyped as a web app using QMiner and Node.JS.

Figure 3 shows the interface where the user can select which segments of users should be compared. Query is specified as a collection of segments defining a subset of website visitors. The second group can be either a complement of the first group, or can also be defined through another list of segments.

In the example we compare users identified as engineers versus users identified as administrators. We can see results for this query in Figure 4. Report shows some overall statistics and the odds rations that are significant for the first group. The output can be directly transformed into instructions for an ad server.

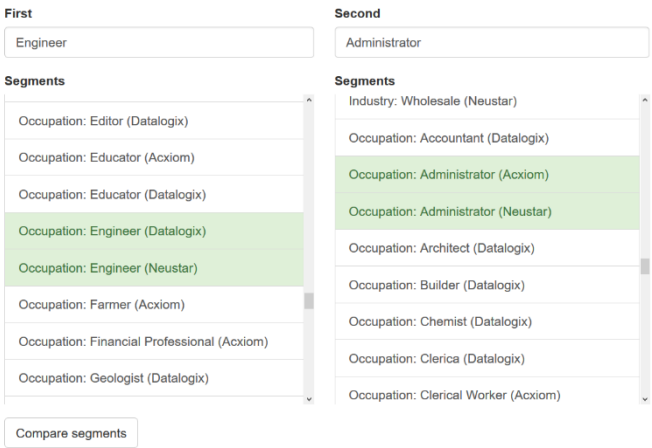

Figure 3. Interface for selecting which segments we want to compare. E.g. engineer vs. administrator

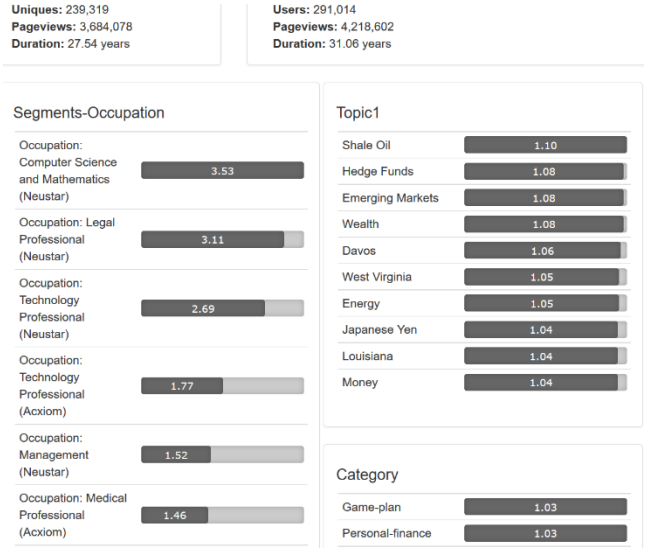

Figure 4. Segment query results

4.2 Analyzing Interests for Topics

The goal of this task is to understand who is interested in specific topics. Topic can be either selected from a predefined list of categories (e.g. Advertisement), or defined via a document. For example, we can search for articles on Product Innovation, which is not included in the predefined list, by using Wikipedia page on product innovation as a query.

Result of such as query is a subset of pages from the website that fit the given topic. We can then join this by using mapping from Section 3.2 to obtain list of visitors that read this content, and, by using mapping from Section 3.1, also their segments. Note that these joins can be executed on one month of data in QMiner in real-near time.

Result is several reports. First, we can generate same kind of report as shown in Figure 4, by contrasting readers of identified pages with the rest of the website's visitors. Second, we can aggregate visited pages and their content in order to identify top topics, people and keywords, as presented in Figure 5 and Figure 6.

4.3 Comparing traffic and stock volume data

As the last use-case we will check for correlation between the stock volume data and a combination of visits logs and content pages. All examples will be focused on AAPL (Apple Inc.) and we will look at the data aggregated by hour.

We start by creating two time-series from the visit logs: (a) number of visits to AAPL quote page, and (b) number of visits to articles related to Apple. First time-series we can obtain directly from visit logs. For the second, we have to first identify significant mentions of entity Apple in the news articles and check their visit statistics. We compere these with a trade volume of AAPL stock. The intuition being, that more news there is about a specific company, the more trading there will be with their stock.

First we compare quote page visits with the trading volume. As we can see in Figure 8, stock volume values fluctuate in similar patterns and time intervals as the number of requests values. Most of the trends can be explained by daily patterns, where most of trading happens during US trading hours, when also website traffic spikes.

If we compare article visits with the trading volume, we can see a clear discrepancy with articles generating more traffic over first few days and the stock having higher trade volume over last few days.

Figure 5. Results for querying on Advertisement input text

Figure 6. Word cloud with top keywords from pages discussing the topic of Advertisement.

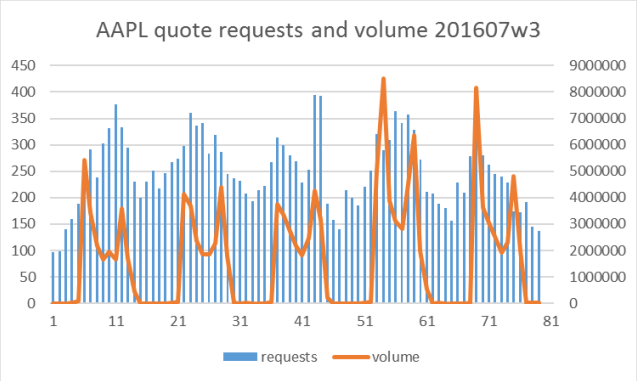

Figure 7. AAPL quote requests and volume

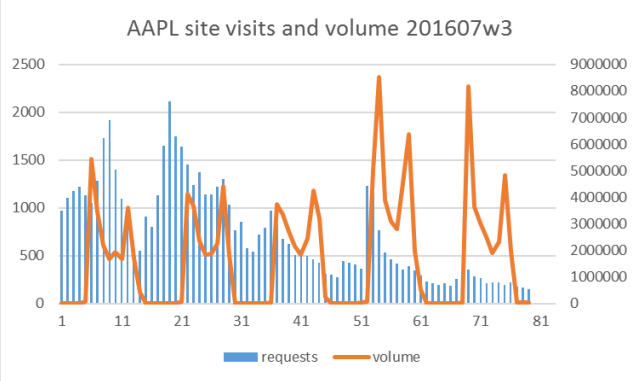

Figure 8. AAPL site visits and volume

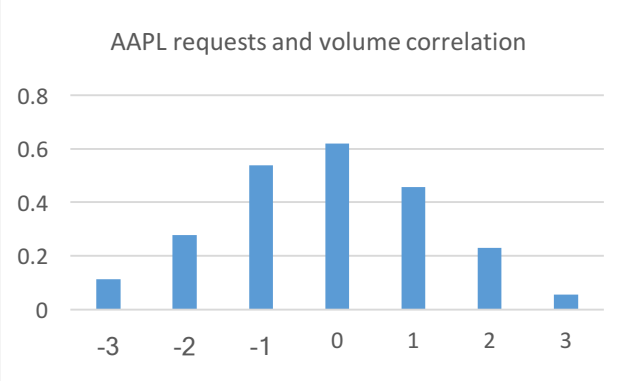

Figure 9. AAPL requests and stock volume correlation

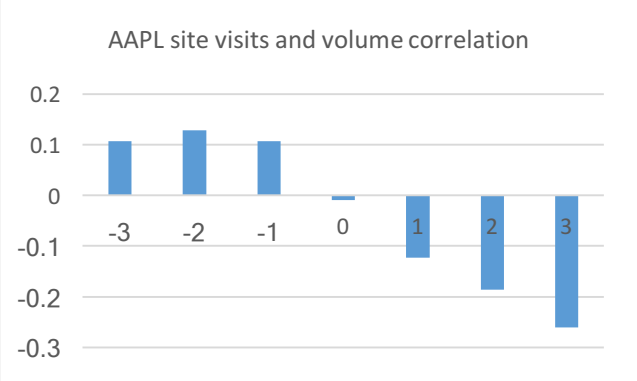

Figure 10. AAPL site visits and stock volume correlation

We can also check the correlation between the trade volumes and the visit data. Figure 9 shows us correlation between quote requests and stock volume. The x-axis represents the offset of request time based on the volume values time, measured in hours. For example, when the offset is -1, we compares quote request data at (h – 1) hour with stock volume data at h hour. The biggest correlation between number of quote requests and stock volume values is within the same hour (no offset). In this case, correlation coefficient is significant as the precise value is 0.62. Whereas if we take the number of quote requests 1 hour in advance of the volume values, we still get a significant correlation coefficient of 0.46, or even greater (0.54) if we take quote requests from 1-hour prior of the volume values. The correlation coefficients fall more, the higher the offset we take between stock volume data values and quote request values. In any case we can observe that correlation between both data is significantly high.

Correlation coefficients on Figure 10 tell us that there is some slight correlation between the numbere of visits to sites that mention Apple Inc. (AAPL) and the volume of AAPL stock on stock exchanges when taking visits prior to the volume movement into account. Similar to Figure 9, the x-axis on Figure 10 represents offset time of site visits, based on the time in stock volume data. There is a small jump in correlation, when taking site visit values that are offset by 2 hours into the past, compared to the stock volume data time. Altough, even at that point the correlation coefficient is 0.13. There is a negative correlation when taking site visit values from the hours following the time of stock market volume values. We can conclude that the correlation between site qoute requests and stock volume data is much stronger than correlation between site visits and stock volume data. However, we can also make an observation that we only get a positive correlation between site visits and stock volume when taking the numbers of site visits that occurred prior to stock volume into account.

5. REFERENCES

[1] Štajner, Tadej, et. al. A service oriented framework for natural language text enrichment. Informatica (Ljublj.), 2010, vol. 34, no. 3, 307-313

[2] Apache Hadoop: http://hadoop.apache.org

[3] Apache Hive: https://hive.apache.org

[4] Fortuna, Blaz, et al. QMiner – Data Analytics Platform for Processing Streams of Structured and Unstructured Data. Software Engineering for Machine Learning Workshop, Neural Information Processing Systems 2014

Visual and Statistical Analysis of VideoLectures.NET

Erik Novak
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
+386 31 272 332
erik.novak@ijs.si

Inna Novalija
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
+386 1 4773144
inna.koval@ijs.si

ABSTRACT
This paper presents learning analytics tools for visual and statistical analysis of data from a portal of video lectures. While learning analytics methods traditionally deal with measurement, collection and analysis of data about learners with an aim of improving the learning process, our solutions are targeted at viewers of VideoLectures.NET. The novel VideoLectures Learning Analytics Dashboard and Lecture Landscape tools allow observing, searching and analyzing viewer behavior and, at the same time, efficiently present the information from the portal to the viewers.

General Terms
Learning Analytics, Measurement, Performance, Design.

Keywords
visualization, analytics, data mining, VideoLectures.NET

1. INTRODUCTION
VideoLectures.NET is a free and open access educational video lectures repository. There are over 20.000 lectures by distinguished scholars and scientists at the most important and prominent events like conferences, summer schools, workshops and science promotional events.

Lectures on the VideoLectures.NET portal are categorized into various categories, such as Arts, Biology, Business, Computer Science, Social Sciences, Technology etc.

In this paper we present tools for visual and statistical analysis of VideoLectures.NET portal data – VideoLectures Learning Analytics Dashboard and Lecture Landscape. While the goal of the VideoLectures Learning Analytics Dashboard is to aggregate, harmonize and analyze event data using various data analysis techniques, the Lecture Landscape visualizes the information about videos, categories and authors in an efficient and user-friendly way.

2. RELATED WORK
Tomas & Cook [14] state that visual analysis should be employed as a means to reveal patterns and trends within data, to develop more intuitive perception and to help with in-depth analysis. Conde et al. [8] present a learning analytics dashboard and its application in real-world case studies. Conceptual framework is developed by Bakharia et al. [7].

In our work we adopt both approaches to learning analytics – visual and statistical analysis, which allows having a deep and more intuitive understanding of the obtained results.

3. LEARNING ANALYTICS FOR VIDEOLECTURES
VideoLectures Learning Analytics Dashboard [1] is a tool developed for the analysis of viewer behavior and detecting which lectures are interesting for the users. We present the results of the preliminary work and the functionalities of the dashboard.

Data Processing. The data considered in the analysis is a set of log files from VideoLectures portal that contain raw events on the portal from September 2012 until December 2015. We have processed 11.3 GB of log files that included the ID, timestamp, session, log entry, lecture and other information such as event type, IP address, location (if present) etc.

With the processed raw log files, we established main event types that appeared in the raw logs: *view* (the user accessed the lecture webpage), *download* (the user downloaded presentation, video etc. from the lecture webpage) and *search* (the user performed a search at the portal).

In addition to the raw log files, we have also collected the log files dedicated to the behavior of particular user while watching particular lectures which we call ranges log files. These present the actions of the user while watching videos like moving forward, moving backwards on the player, skipping some video section etc.

Analysis of Log Files. In order to analyze user behavior at VideoLectures.NET portal, we have utilized and developed a set of data analysis techniques. These were used in the development of VideoLectures Learning Analytics Dashboard, which contains both statistical analysis and visual exploration features.

The analysis has been performed from four perspectives: the aggregated perspective for all lectures, perspective of singular lecture, aggregated perspective of all viewers and perspective of singular viewer.

In addition, we have developed a set of metrics for viewers and lectures that provide us an insight into the user behaviour on the VideoLectures.NET portal.

The ***lecture*** metrics measures the number of views and viewers the lecture has, the average and standard deviation of time the lecture was watched (in minutes and percentage) and the average and standard deviation of moves going forwards and backwards through the lecture video (in minutes and percentage).

The ***viewer*** metrics measures the number of lecture views the viewer made, the average and standard deviation of time spent watching (in minutes and percentage) and the average and standard deviation of moves going forwards and backwards through the lecture video (in minutes and percentage).

First we have analyzed a set of 1000 most popular (by views) videos relevant to the Data Science category. Those videos produced 7055427 views and 3020090 downloads in the period from September 2012 until December 2015. The number of visitors for these videos was 1045860. Moreover, there were 866912 searches at the portal. We have then analyzed all of the log files for which the statistics can be found on the Learning Analytics Dashboard.

3.1 VideoLectures Learning Analytics Dashboard

Our interest is not only to make statistical analysis of the viewer behavior but also to have visual exploration features that enables us to see the traffic and activities made on the VideoLectures portal. For this we created graphs that show the number of views, downloads and searches the viewers make. An example can be seen in Figure 1 which shows the view trends.

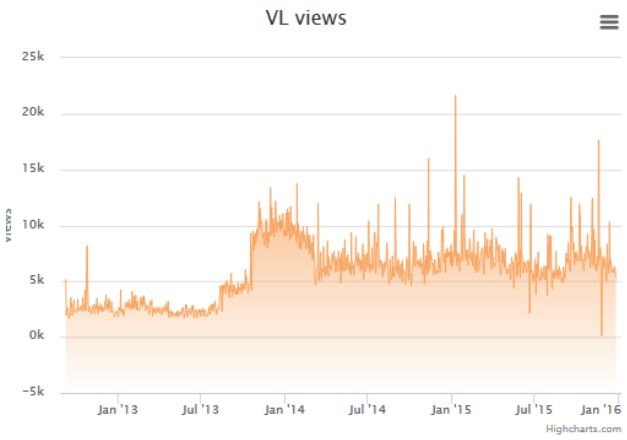

Figure 1. VideoLectures views trend. It shows the total number of lecture views through September 2012 until December 2015.

Statistics for a particular lecture are also available. On the *per lecture* tab one can search for the desired lecture and get its title, description, the measures returned by the lecture metrics and a graph showing the lectures activity. Figure 2 shows the lecture information and measures of "Deep Learning in Natural Langauge Processing" lecture.

The overall statistics of the viewers can also be found under the *all viewers* tab. It shows the distribution of the viewers through countries (see Figure 3) and other statistics measured with the viewer metrics.

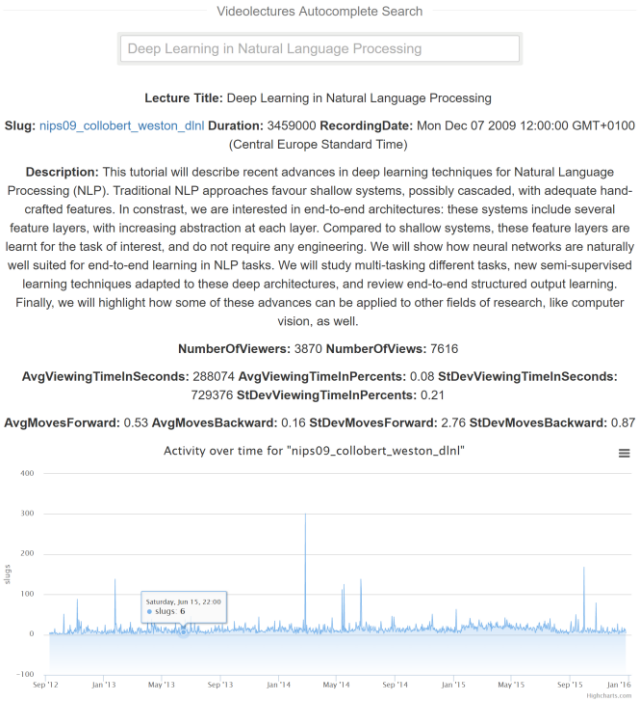

Figure 2. Lecture information for "Deep Learning in Natural Language Processing" lecture. Displays the basic lecture information, measures and graph of its activity.

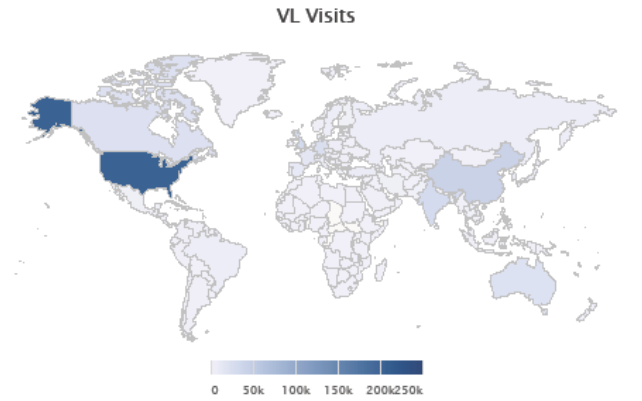

Figure 3. Viewer distribution through countries. Most viewers come from the United States.

In order to implement VideoLectures Learning Analytics Dashboard, we used QMiner [12] for processing data, Node.js [3] for creating the server and dashboard and Highcharts [4] for graphical implementation and dynamics support.

Using the interactive VideoLectures Learning Analytics Dashboard, it is possible to check what are the lectures of broad interest at the VideoLectures.NET portal, how the users of the portal behave through time, where the users come from and how they react at the specific videos.

4. VIDEOLECTURES EXPLORER

Our interest is not only to analyze the viewer behavior but also to see the similarities between lectures. For this we developed VideoLectures Explorer [2], a tool for exploring the lectures published on VideoLectures.NET. The tool enables the user to search through the lectures and find similarities between them, e.g., to find lectures of a specific category or presenter of interest.

Here we explain the method used for visualizing lecture similarities and present the tools functionalities.

4.1 Data Acquisition

The database used for the visualization and basic statistical analysis contains data from all lectures found on VideoLectues.NET. We have acquired data for 23224 lectures, keynotes, interviews, events etc. For each lecture the database contains the lectures title and description, the name of the presenter and his affiliation with city and country, the lectures publication date, video duration, its parent event, number of views and the scientific categories the lecture belongs to. The scientific categories have a hierarchical structure decided by the VideoLectures development team which we use in the visualization process. The database was constructed using the VideoLectures API [5].

4.2 Methodology

Our objective is to draw the lectures into a two-dimensional vector space to allow plotting on the computer screen, where the similarities of the lectures are maintained. The method we used has been described in [11]. Here is its quick summary:

Using the bag-of-words [13] representation, we represent the lectures as vectors in a high-dimensional vector space, where each dimension corresponds to one word from the vocabulary. These are then used to construct the term-document matrix. Using Latent Semantic Indexing [10] we merge the dimensions associated with the terms that have similar meaning and get the most suitable set of words to describe the corresponding lectures. After that we use Multidimensional Scaling [14] to reduce the dimensionality of the original multidimensional vectors and map them onto two dimensions where the distances between vectors are preserved as well as possible. Once we have the two-dimensional coordinates we can draw the landscape using the preferred visualization library.

The features used for representing the lectures similarities are its title, description, categories and parent event. We used the algorithm described in section 4.2 to calculate the lectures coordinates and draw the landscape using d3.js visualization library [6].

4.3 Explorer Functionality

Each lecture is presented by a point and size mapping to the number of views. Similarity between lectures is mapped to distance between points; more similar lectures are brought closer together. Hovering over a point brings up a tooltip containing information about the lecture: its title and description, the name of the presenters his affiliation, the language in which the lecture was presented, which scientific categories it belongs to, its duration, when it was published and the number of views since the last database update. Any data attribute for which values are not available is omitted from the tooltip. Landmarks show areas populated with lectures that are majorly of the same category. The user can zoom in and out of the landscape to enable a more detailed look of the lectures. An example of the landscape can be seen in Figure 4, which shows the of the *machine learning* lecture landscape.

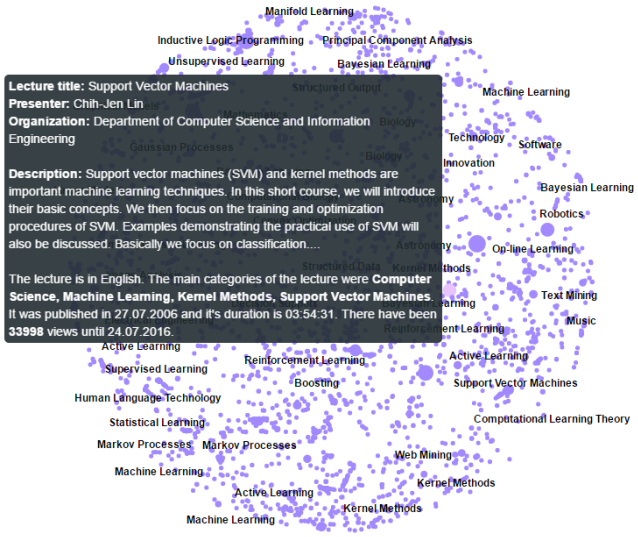

Figure 4. The landscape of machine learning lectures. created using the "Machine Learning" keyword. Lectures that are more similar are brought closer together.

Clicking on the point shows the lectures information. It shows the lectures title and presenter, if the lecture's public and enabled and a link to the lecture video located on VideoLectures.NET. Clicking on the lecture link opens the corresponding lecture video on the portal.

When the landscape is generated the dashboard also shows an additional information window (see Figure 5). This is in two parts: the first part is the query information, containing the names of presenters, organizations and categories, the minimum and maximum number of views, organization location and the language the user used to query the data.

The second part contains the basic statistics about the queried data, the number of lectures in the queried data, the total number of lecture views and scientific categories with the number of their occurrences in the queried data. Clicking on the category in the dashboard automatically queries the data using the category name, and the information window is updated along the landscape view.

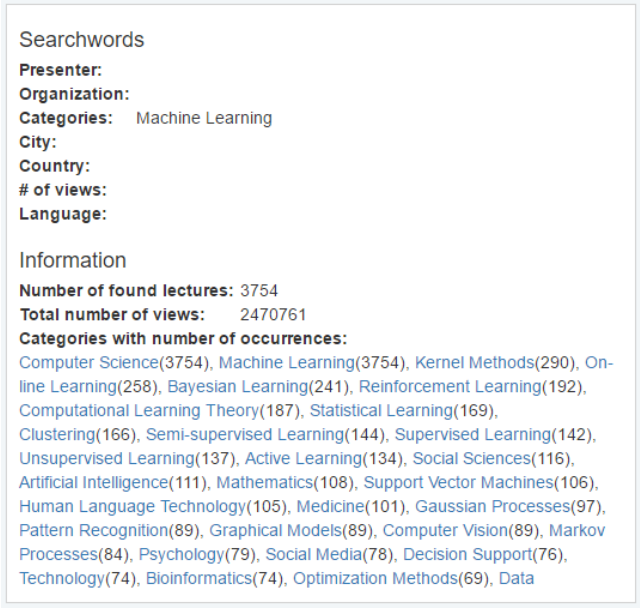

Figure 5. The additional information window created using the "Machine Learning" keyword. It contains the query keywords and the overall statistics about the queried lectures.

5. CONCLUSION AND FUTURE WORK

In this paper we presented learning analytics tools for visual and statistical analysis of data from VideoLectures.NET portal. While learning analytics methods traditionally deal with measurement, collection and analysis of data about learners with an aim of improving the learning process, our solutions are targeted at viewers of VideoLectures. The novel VideoLectures Learning Analytics Dashboard and Explorer tools allow observing, searching and analyzing viewer behavior and, at the same time, efficiently present the information from the portal to the viewers.

The future work will include the development of more efficient learning analytics suggestions for VideoLectures.NET portal and providing the recommendations on how to increase the viewer engagement into the portal.

6. ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and EDSA EU H2020 project (Contract no: H2020-ICT-643937).

7. REFERENCES

[1] Videolectures learning analytics | dashboard, http://learninganalytics.videolectures.net, accessed: 2016-09-09.

[2] Videolectures learning analytics | landscape, http://explore.videolectures.net, accessed: 2016-09-09.

[3] Node.js, https://nodejs.org/en/, accessed: 2016-09-09.

[4] Interactive javascript chart for your website | highcharts, http://www.highcharts.com/, accessed: 2016-09-09.

[5] Swagger ui, http://videolectures.net/site/api/docs/, accessed: 2016-09-09.

[6] D3.js – data-driven documents, http://d3js.org/, accessed: 2016-09-09.

[7] A. Bakharia, L. Corrin, Linda, P. de Barba, G. Kennedy, D. Gašević, R. Mulder, D/ Williams, S. Dawson and L. Lockyer. A conceptual framework linking learning design with learning analytics. *In Proc., Sixth International Conference on Learning Analytics & Knowledge*, ACM, pages 329-338, 2016.

[8] M. Á Conde, F. J. García-Peñalvo, D. A. Gómez-Aguilar and R. Therón. Exploring Software Engineering Subjects by Using Visual Learning Analytics Techniques, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 10(4), pages 242-252, 2015.

[9] T. F. Cox and M. A. Cox. *Multidimensional Scaling*. CRC press, New York, 2000.

[10] S. T. Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188-230, January 2004.

[11] B. Fortuna, D. Mladenić and M. Grobelnik. Visualization of temporal semantic spaces. In *Semantic knowledge management*, pages 155-169, Springer, 2009.

[12] B. Fortuna, J. Rupnik, C. Fortuna, M. Grobelnik, V. Jovanoski, M. Karlovcec, B. Kazic, K. Kenda, G. Leban, J. Novljan, M. Papler, L. Rei, B. Sovdat, L. Stopar and A. Muhic. QMiner – Data analytics platform for processing streams of structured and unstructured data. *Software Engineering for Machine Learning Workshop, Neural Information Processing Systems*, 2014.

[13] G. Salton. Developments in automatic text retrieval. *Science*, 253(5023):974-979, August 1991.

[14] J.J. Thomas and K.A.Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press, Los Alamitos, 2005.

Spatio-temporal clustering methods

Matej Senožetnik, Luka Bradeško, Blaž Kažič, Dunja Mladenić, Tine Šubic
Jožef Stefan Institute
and
Jožef Stefan International Postgraduate School
Jamova cesta 39
1000 Ljubljana, Slovenia

ABSTRACT
Tracking a person, an animal, or a vehicle generates a vast amount of spatio-temporal data, that has to be processed and analyzed differently from ordinary data generally used in knowledge discovery. This paper presents existing spatio-temporal clustering algorithms, suitable for such data and compares their running time, noise sensitivity, quality of results and the ability to discover clusters according to non-spatial, spatial and temporal values of the objects.

Keywords
clustering, spatial-temporal data, density-based algorithms

1. INTRODUCTION
Due to emerging field of ICT and rapid development of sensor technologies, a lot of spatio-temporal data has been collected in the past few years. By processing and enriching raw spatio-temporal data we aim at extracting semantic information, which is a basic requirement for the comprehension and later usage of this data. Normally, the very first step of this process is clustering raw GPS coordinates into more distinct groups of points, which already have some semantics, such as whether the points belong to trajectory or stationary point (i.e., stay point).

This paper aims to investigate different methods, used for clustering spatio-temporal data, generated by mobile phones, by collecting timestamped GPS coordinates of the phones' location. By clustering the collected coordinates, we obtain so called stay points (also referred to as points of interest or stop points [15]), which are the points in space where a moving object has stayed within a certain distance threshold for a longer period of time [16]. When the stay points are calculated, we can process the data further, to calculate the most frequently visited locations (i.e., frequent locations), which is the fundamental building block for further advanced analytics use-cases, such as next location prediction [6].

In general, clustering methods are separated into following categories:

- **Partitioning methods** - classify data into k groups or partitions,
- **Hierarchical methods** - hierarchically decompose given datasets,
- **Density-based methods** - are based on cluster density, where clusters stop growing when neighbourhood density stops exceeding a given threshold,
- **Model-based methods** - definition copied from [1]: "In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points."

The most appropriate methods for clustering spatio-temporal data are density-based methods as they regard clusters as dense regions of objects in a data space that are separated by regions of low density [7]. Thus we will focus on density based methods and their application to cluster GPS coordinates.

2. DENSITY-BASED ALGORITHMS
Table 1 shows some of the more common density-based algorithms, used for the detection of stay points. For a successful detection of stay points, it is important that the clustering algorithm uses temporal information alongside bare spatial data. As indicated in Table 1, most of the algorithms do make use of it, except for DBSCAN. Quality of the algorithm also depends on noise sensitivity (the less sensitivity the algorithm has, the better it is), as cellphones' GPS coordinates are frequently noisy due to connection glitches (for instance, when moving through forests, staying inside buildings or in bad weather conditions). Some algorithms return only clusters such as DBSCAN, ST-DBSCAN and OPTICS, but CB-SMoT, SMoT and SPD return a stay point (also referred to as stops) or a path (also refered to as moves).

DBSCAN has an overall average running time $O(n\ log\ n)$, and the worst case run time complexity is $O(n^2)$. Running time depends on parameter choice and version of implementation. OPTICS has similar time complexity but it is 1.6 seconds slower than DBSCAN. ST-DBSCAN has the same running time as DBSCAN.

Algorithm name	Spatio temporal	Noise sensitivity	Returning stay points/path
DBSCAN	✗	✗	✗
ST-DBSCAN	✓	✗	✗
SMoT	✓	✓	✓
CB-SMoT	✓	✗	✓
SPD	✓	✓	✓
OPTICS	✓	✗	✗

Table 1: The most common density-based algorithms, used for the detection of stay points [14].

2.1 Density Based Spatial Clustering of Application with Noise (DBSCAN)

DBSCAN [5] is a density-based clustering algorithm which identifies arbitrary-shaped objects and detects noise in a given dataset. The algorithm starts with the first point in the dataset and detects all neighboring points within a given distance. If the total number of these neighboring points exceeds a certain threshold, all of them have to be treated as part of a new cluster. The algorithm then iteratively collects the neighboring points within a given distance of the core points. The process is repeated until all of the points have been processed.

DBSCAN's advantages are that it robustly detects outliers, only needs two parameters (Eps and MinPts), is appropriate for large datasets and data input order does not interfere with the results [12].

Numerous research studies have extended DBSCAN, such as in the example of GDBSCAN [13], which is a generalization of the original DBSCAN. GDBSCAN can cluster point objects as well as spatially extended objects. Another one of these extended algorithms is DJ-Cluster [17], used for discovering personal gazetteers. It attaches semantic information to clusters and requires a list of points of interest. Also extension is ST-DBSCAN which is described bellow.

ST-DBSCAN [4] is another algorithm that is based on DBSCAN. It is making use of its ability to discover clusters with various shapes, while improving some of the weak points of the original algorithm. It adds temporal data to the clustering results, identifies noisy objects if there are various densities of the input data, and more accurately differentiates adjacent clusters.

2.2 Stops and Moves of Trajectory (SMoT)

SMoT [2] algorithm divides data into two sets called *moves* and *stops*. The easiest way to understand how SMOT works is by reviewing the example shown in Figure 1. There are three potential stop candidates with geometries Rc_1, Rc_2 and Rc_3 and with information about minimum time duration for each of them. From the figure, we can observe that the point p_0 is not inside any of these geometries, therefore it is a candidate for *move* dataset. The next few points are inside the first stop candidate (Rc_1) and also exceed the minimum time duration which is specified for every geometry. In candidate Rc_2, the point duration does not exceed minimum threshold, therefore Rc_2 is not a stop.

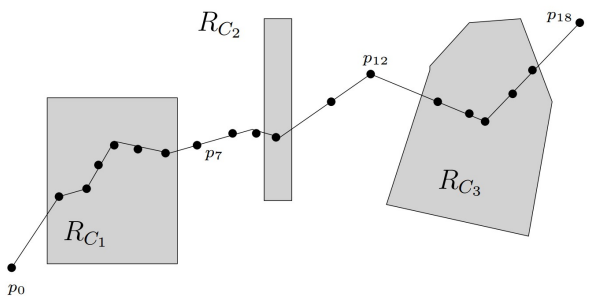

Figure 1: Example of trajectory points with three possible candidate stops [2].

2.3 Clustering-Based Stops and Moves of Trajectory (CB-SMoT)

CB-SMoT [11] algorithm is an alternative to the algorithm SMoT and addresses one of its main drawbacks - the incapability to detect stay points that are not predefined by user. It uses clustering methods to automatically detect stay points. The idea behind this method is, that when we move around points of interest (such as museums, monuments, night-clubs, etc.), we move slower than when we are traveling from one place to another. In the first steps, the potential stops (the slower parts of a trajectory) are identified using the variation of the DBSCAN algorithm which considers one-dimensional trajectories and their speed. In the second step, the algorithm identifies the location of the potential stops (clusters) which were found in the first step. The authors report [11] that their algorithm discovers less incorrect stops, compared to SMoT algorithm.

2.4 Stay Point Detection (SPD)

The SPD [9] algorithm works by detecting whether the observed entity has spent more than 30 minutes within a radius of 100 meters. If this happens, the region is detected as a stay point. Both threshold time and distance parameters are adjustable and are chosen depending on the use-case and accuracy of raw data.

The main advantages of SPD is its need for predefined structures. It is not computationally demanding, but it is sensitive to the accuracy of data usually generated by GPS receiver. Namely, for different accuracies of GPS it returns slightly different positions for the same location. It is also sensitive to noise, but this can be partially reduced by adjusting the parameters of the algorithm.

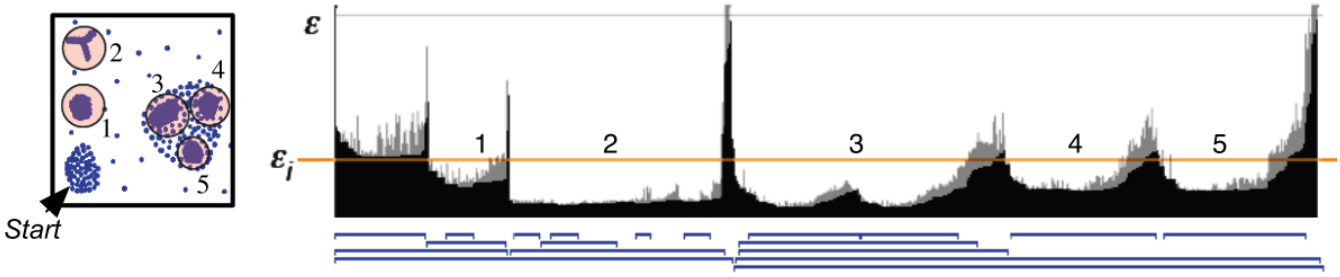

Figure 2: OPTICS result is called reachability plot [10].

2.5 Ordering Points to Identify the Clustering Structure (OPTICS)

OPTICS [3] is an algorithm which is used for finding density-based clusters in spatio-temporal data. Though it works in a similar way as DBSCAN, OPTICS improves on DBSCAN's biggest weakness, the failure to detect clusters when density of the data varies.

In Figure 2, we can observe that OPTICS algorithm generates an easily visualized ordering of points, which can be used to extract partitions and hierarchical clusters [10]. Zheng et al. [16] used OPTICS for clustering stay points. In their article, they present the idea that through the use of a statistically significant number of user-generated GPS trajectories, the correlation between personal geographical locations implicit in a person's location history enables the provision of valuable services, such as personalized recommendations and target specific sales promotions.

2.6 Other algorithms

There exist many other different approaches on dealing with spatial and spatio-temporal data. Some of these are developed as extensions (for example, DB-SMoT [12]) to well known algorithms, while others take an entirely new approach (TRACLUS [8]). It is important to note that other data can also be used alongside coordinates and timestamps.

3. DISCUSSION

The goal of this discussion is to find an algorithm capable of clustering user's raw historical data of locations, as tracked by a mobile phone. Referring back to Table 1, we want an algorithm which has the following properties:

- is able to cluster spatio-temporal data;
- is noise insensitive;
- returns stay points and a paths.

As we had already stated in the beginning of this paper, besides spatial data, temporal information is one of the most important additional information for stay point detection algorithms, which enables better stay point detection performance and empowers additional time related capabilities, such as detecting time spent at each stay point. By using an algorithm that clusters data only by spatial information, we lose part of the useful information (in our case, the time spent on a location), as well as the order of stay points on the timeline, thus making us unable to do further analysis, such as future location predictions and plotting frequency graphs. Due to this shortcoming and also different cluster densities, algorithms such as DBSCAN, or its improved version DJ-cluster, are not appropriate for our use case.

In [14], Sander et al. exposed that the problem of DB-SMoT algorithm is that the quantile function requires a priori knowledge of the proportion between points inside potential stops and total points in the dataset. This proportion varies among datasets since users sometimes spend a whole day inside a stop (i.e., the proportion is one), while on different occasions they might be visiting more than ten stops (i.e., the proportion is much smaller than one). Due to this, we need and online algorithm that can function without prior knowledge of stay point to path ratio. The SMoT algorithm uses predefined regions which can be a problem if we want to detect some stay points outside of those zones (such as outdoors). Algorithm SPD, SMoT and CB-SMoT are all sensitive to noise, but have the advantage of returning stay points ordered by timestamps. With other algorithms, such as DBSCAN and OPTICS, we need to find the right sequence ourselves independently of the algorithm. Given our requirements, we propose CB-SMoT, SMoT and SPD as the most suitable for clustering spatio-temporal data collected from a mobile phone.

An efficient algorithm must be insensitive to noise. Currently, GPS has reception problems in narrow valleys, forests or inside buildings, so it's important to use different data sources (for example, Wi-Fi networks, activity recognition) for accurate stay point detection. Modern mobile phones nowadays also provide rudimentary activity recognition, which we can use to compare our own results against. If the activity recognition system is not detecting movement (the person is standing still) and our collected GPS coordinates are consistently far apart, we may be facing a problem with GPS accuracy. In such cases, additional data sources should be considered for more accurate measurement.

To the best of our knowledge, up to this date SPD algorithm is considered as the best solution for detection of stay points, but suffers from detecting false stay points and paths. This problem can be alleviated by running multiple iterations of the algorithm on resulting dataset. This is already outside of the scope of this paper and will be described in the separate paper which is currently under preparation.

4. ACKNOWLEDGMENT

This work was supported by the Slovenian Research Agency and the ICT program of the EC under project OPTIMUM (H2020-MG-636160).

5. REFERENCES

[1] Clustering analysis. `http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm`.

[2] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems*, GIS '07, pages 22:1–22:8, New York, NY, USA, 2007. ACM.

[3] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod Record*, pages 49–60, 1999.

[4] Derya Birant and Alp Kut. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. pages 226–231. AAAI Press, 1996.

[6] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[7] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[8] Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. Trajectory clustering: A partition-and-group framework. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 593–604, New York, NY, USA, 2007. ACM.

[9] Quannan Li, Yu Zheng, Xing Xie, Yukun Chen, Wenyu Liu, and Wei-Ying Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 34:1–34:10, New York, NY, USA, 2008. ACM.

[10] Chris Mueller. Data Preparation. 2005.

[11] Andrey Tietbohl Palma, Vania Bogorny, Bart Kuijpers, and Luis Otavio Alvares. A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 863–868, New York, NY, USA, 2008. ACM.

[12] Parya Pasha and Zadeh Monajjemi. a Clustering-Based Approach for Enriching Trajectories With Semantic Information Using Vgi Sources a Clustering-Based Approach for Enriching Trajectories With Semantic Information Using Vgi Sources. 2013.

[13] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Min. Knowl. Discov.*, 2(2):169–194, June 1998.

[14] Khoa a Tran, Sean J Barbeau, and Miguel a Labrador. Bacchelors: Automatic Identification of Points of Interest in Global Navigation Satellite System Data : A Spatial Temporal Approach Categories and Subject Descriptors. (January):33–42, 2013.

[15] Y. Zheng and X. Xie. Learning location correlation from gps trajectories. In *2010 Eleventh International Conference on Mobile Data Management*, pages 27–32, May 2010.

[16] Yu Zheng. Trajectory Data Mining: An Overview. *ACM Trans. On Intelligent Systems and Technology*, 6(3):1–41, 2015.

[17] Changqing Zhou, Dan Frankowski, Pamela Ludford, Shashi Shekhar, and Loren Terveen. Discovering personal gazetteers: An interactive clustering approach. In *Proceedings of the 12th Annual ACM International Workshop on Geographic Information Systems*, GIS '04, pages 266–273, New York, NY, USA, 2004. ACM.

Indeks avtorjev / Author index

Konferenca / Conference

Uredili / Edited by

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

Dunja Mladenić, Marko Grobelnik