QROWD: Because Big Data Integration is Humanly Possible

Eddy Maddalena
University of Southampton
Southampton, UK
e.maddalena@soton.ac.uk

Luis-Daniel Ibáñez
University of Southampton
Southampton, UK
l.d.ibanez@soton.ac.uk

Elena Simperl
University of Southampton
Southampton, UK
e.simperl@soton.ac.uk

Mattia Zeni
University of Trento
Trento, Italy
mattia.zeni@disi.unitn.it

Enrico Bignotti
University of Trento
Trento, Italy
enrico.bignotti@unitn.it

Fausto Giunchiglia
University of Trento
Trento, Italy
Jilin University
Changchun, China
fausto@disi.unitn.it

Claus Stadler
University of Leipzig
Leipzig, Germany
cstadler@informatik.uni-leipzig.de

Patrick Westphal
University of Leipzig
Leipzig, Germany
patrick.westphal@informatik.uni-leipzig.de

Luís P. F. Garcia
University of Leipzig
Leipzig, Germany
garcia@informatik.uni-leipzig.de

Jens Lehmann
University of Bonn
Bonn, Germany
Fraunhofer IAIS, Germany
Sankt Augustin, Germany
jens.lehmann@cs.uni-bonn.de

ABSTRACT

We present QROWD, a project funded by the Horizon 2020 research programme, which aims at offering socio-technical solution to cross-sectorial Big Data integration in a European urban Smart Transportation context through a hybrid architecture for Big Data integration and analytics.

CCS CONCEPTS

• **Human-centered computing → User centered design**; **Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Classification and regression trees**;

KEYWORDS

Big Data Integration

ACM Reference Format:
Eddy Maddalena, Luis-Daniel Ibáñez, Elena Simperl, Mattia Zeni, Enrico Bignotti, Fausto Giunchiglia, Claus Stadler, Patrick Westphal, Luís P. F. Garcia, and Jens Lehmann. . QROWD: Because Big Data Integration is Humanly Possible. In *Proceedings of (KDD 2018).* ACM, New York, NY, USA, 2 pages.

KDD 2018, 19 - 23 August 2018, London, United Kingdom

1 THE QROWD PROJECT

Almost 75% of European Citizens live in urban areas [1]. Traffic congestion represents a cost of €100 billion per year for the community [1]. Furthermore, the road transport impacts on air pollution with almost the 40% of the total CO_2 emissions and the 70% of the transport emissions. Transport represents a pivotal sector for the EU-28 countries, since it involves almost the 4.4% of the total Gross Value Added (GVA), €560 billions and more than 9 millions of employees. To improve transport and mobility, EU spent €7.9 billion in the period over 2007-2013, and expects to spend a minimum of €12 billion in the period over 2014-2020[2]. QROWD is a project funded with € 3.5 million by the EU Horizon 2020 research and innovation programme (grant agreement n.732194)[3]. The project, that has begun in December 2016 and will terminate in November 2019, is part of the Big Data PPP Value Public-Private Partnership.

The project consortium is comprised by three academic partners (University of Southampton, University of Trento, InfAI Leipzig), two large companies (ATOS and TomTom), one SME (AI4BD) and one public sector partner (Municipality of Trento)

2 USE CASES

The project counts two business cases that are then divided into 11 use cases. The identification of the use cases was driven by the Municipality of Trento (Italy), which aims to reduce urban traffic

[1] http://ec.europa.eu/transport/themes/urban/urban_mobility/index_en.htm
[2] http://civitas.eu/eu-funding
[3] https://cordis.europa.eu/project/rcn/206181_en.html

and thus the CO2 emissions. The use cases cover multiple aspects such us parking, tourism, data collections and mapping. Of particular importance for this abstract is (i) the *The Modal split* use case that is a fundamental formal metric for understanding how citizens use various means of transport. Modal spit allows to compute the percentage of travelers using a particular type of transportation for everyday travel. Some of the others use cases are: (ii) *Parking Availability* to compute the probability to find a parking spot for four/two-wheeled vehicles; (iii) *Completing mobility infrastructure information* which gives information about mobility infrastructure retrieved through spatial crowdsourcing.

3 INNOVATION TO SHOW

Modal split has a strategic importance for municipalities. It is usually estimated by using paper or telephone surveys, that are expensive to run and scale, limiting their utility. Previous efforts have looked at how to implement travel surveys into citizen's mobile phones, leveraging the device as a data collection instrument and as a mean to interact with the citizen [2]. The general workflow of such approaches goes as follows: (i) a citizen installs a mobile application that collects sensor data about her movements; (ii) state of the art machine learning models are applied on collected training data to decompose citizen's daily traces into trips, and assign to each trip the transport mode label with highest confidence; (iii) assigned labels are sent back to citizens through an interface, in order for her to validate if the machine assigned labels are correct or not. Corrections are then input to machine learning models to improve them in a reinforcement learning; (iv) eventually, the machine learns to decompose and classify the trips with a high enough confidence, requiring only occasional human confirmation.

However, all approaches rely on the previous collection of sufficient training data, using volunteers or paid citizens to provide high quality labels of their trips during a certain period of time. However, this has three shortcomings: (a) it is not always possible to engage beforehand a sufficient amount of citizens to produce training data covering all possible transportation modes; (b) dynamic variables like weather and traffic affect input data, if classifiers are trained on volunteers that move along a traffic-free route, when applying the model to a citizen that moves along a traffic-heavy route, errors might appear, requiring a higher amount of interaction with citizens; (c) a way to manage the interaction with citizens in this context becomes critical, both to improve the quality of provided labels and to avoid drop-out and disruption of the only person that can correctly label data.

4 DESCRIPTION OF DEMONSTRATION

In QROWD, we propose to overcome these issues by proposing a hybrid pipeline that enables collection of labeled data on-the-go and drives the interaction with the citizen to maximize the accuracy of obtained labels without compromising citizen engagement. We first use a non-supervised method for detecting trip segments, together with questions specifically designed to reduce disruption and increase engagement. Citizens' engagement with questions and data collection is monitored to warn users that provide noisy or incomplete data and suggest changes in question format. The engagement value is also of interest to municipalities to propose further collaborations to citizens (e.g., spatial crowdsourcing tasks). Collected data can be then used to train more advanced models, which may require different types of questions to be tuned. Questions can then be personalised according to previously collected data. When a new citizen that joins a subsequent survey provides valuable data (and question answers) not considered in the previous model, it can be used for further reinforcement learning. Benchmark of the system will be done by measuring the accuracy increase over successive runs and the drop-out rate of citizens.

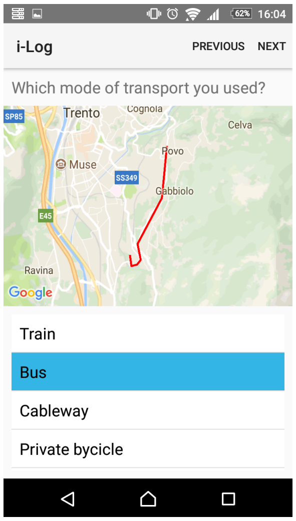
Which mode of transport you used?

Figure 1: Example of question

We will demonstrate the application we developed to collect data and interact with citizens, based on the i-Log system [3]. Then we will describe the journey of a daily data trace of a citizen: (i) How is first processed to check for noise and do a first segmentation, (ii) how questions about generated trips are created and how to choose when they are sent, (iii) how do questions look on the phone, (iv) fast forward to a situation when enough data is available to train a more complex model, then, show same as (i-iii).

We believe our project fits the KDD ecosystem as it provides tools for optimising the relationship between humans and the algorithms that mine and extract knowledge from their data, en route to implementing human-centric processes and services. We also believe that our results might be of use for researchers in the area of question-answering systems within KDD. Finally, we would like to benefit from the interaction with projects more into the technical side of KDD to generalise our findings to data flows beyond the transport domain.

REFERENCES

[1] Eurostat. 2015. *EU transport in figures: statistical pocketbook.* Number 2015. Office for official publications of the European communities.

[2] Philippe Nitsche, Peter Widhalm, Simon Breuss, Norbert BrÃďndle, and Peter Maurer. 2014. Supporting large-scale travel surveys with smartphones âĂŞ A practical approach. *Transportation Research Part C: Emerging Technologies* (2014).

[3] Mattia Zeni, Ilya Zaihrayeu, and Fausto Giunchiglia. 2014. Multi-device activity logging. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication.* ACM, 299–302.