Guest Editorial Special Issue on Knowledge Discovery From Mobility Data for Intelligent Transportation Systems

The recent technological advances on telecommunications create a new reality on mobility sensing. Nowadays, we live in an era where ubiquitous digital devices are able to broadcast rich information about human mobility in real-time and at a high rate. Such fact exponentially increased the availability of large-scale mobility data which has been popularized in the media as the new currency, fueling the future vision of our smart cities that will transform our lives. The reality is that we just began to recognize significant research challenges across a spectrum of topics. Consequently, there is an increasing interest among different research communities (ranging from civil engineering to computer science) and industrial stakeholders on building knowledge discovery pipelines over such data sources. However, such availability also raises privacy issues that must be considered by both industrial and academic stakeholders on using these resources.

The present Special Issue is focused on applications, architectures, and advanced technologies for Knowledge Discovery from Mobility Data that leverage and/or elaborate on related fields (Statistics, Machine Learning, and Artificial Intelligence) to develop and/or deploy Intelligent Transportation Systems. 11 high-quality papers were carefully selected to bring to you innovations to the state-of-the-art in these topics. The approached problems include the traditional Predictive/Descriptive Analytics and Model-Based approaches to non-trivial Clustering, Dimensionality Reduction and Supervised Learning. Applications range across multiple hot topics such as Data-Driven Public Transport Planning and Highway Traffic Control, Predictive Maintenance, Data Quality Evaluation and Foundations of Mobility Mining.

Estimating Inefficiency in Bus Trip Choices From a User Perspective With Schedule, Positioning and Ticketing Data

A citywide methodology to estimate and evaluate Public Transportation Systems inefficiency is proposed, using the 1.8M-inhabitant Brazilian city of Curitiba as a case-study. City's GTFS feed, AVL and AFC systems data are integrated to estimate an origin-destination matrix, which is used as the basis for such analysis. In order to achieve the desired integration at city scale, two map-matching techniques are proposed. The designed techniques prove to increase matching effectiveness, when compared to the state-of-the-art, in scenarios where GTFS specifications define that a bus route has multiple possible trajectories. The case-study evaluation points out that there is significant room for improvement in one third of the trips analyzed.

Efficient Transport Simulation With Restricted Batch-Mode Active Learning

In the context of simulation metamodeling, an active learning algorithm based on Gaussian Processes that gathers the most informative simulation data points in batches, according to both their predictive variance and to the relative distance between them, is proposed. This allows us to explore the simulators' input space with fewer data points and in parallel, and thus in a more efficient way, while avoiding computationally expensive simulation runs at the same time. We also suggest two simple and practical stopping criteria so that the iterative learning process can be fully automated. The results show that the proposed methodology is able to improve the exploration efficiency of the simulation input space in comparison with standard batch-mode active learning procedures.

Spatio-Temporal Profiling of Public Transport Delays Based on Large Scale Vehicle Positioning Data From GPS in Wrocław

A large-scale study of public transport delay data in GPS traces provides an approach for clustering and detecting characteristics of delay changes in the city of Wrocław, Poland. Six delay change profiles are detected using earth mover's distance and hierarchical agglomerative clustering with Vor Hees's linkage method on stop pairs. The profiles divide the public transport infrastructure by the scale of increase and decrease of delay of vehicles which are travelling between two given stops. Provided methods and delay change profiles allow traffic optimization, timetable improvement and detecting parts of the transport network which need improvement.

Traffic Risk Mining From Heterogeneous Road Statistics

Nowadays a large amount of traffic-related data have been obtained. In this paper, we propose a novel framework for mining traffic risk from such heterogeneous data. Traffic risk refers to the possibility that traffic accidents occur. We specifically focus on two issues; 1) predicting the number of accidents at any road and intersection and 2) clustering roads to identify risk factors for risky road clusters. We give a unifying approach to these issues by means of feature-based nonnegative matrix factorization (FNMF). We demonstrate using real traffic data in Tokyo that with our proposed algorithm, we are able to predict traffic risk at any location more accurately and efficiently than existing methods and that a number of clusters of risky roads are identified and characterized by two risk factors. Through this study we open a new research area of traffic risk mining.

Digital Object Identifier 10.1109/TITS.2018.2877063

Mining Smart Card Data for Travellers' Mini Activities

In the context of public transport modelling and simulation, we address the problem of mismatch between simulated transit trips and observed ones. We point to the weakness of the current travel demand modelling process; the trips it generates are over-optimistic and do not reflect the real passenger choices. To explain the deviation of simulated trips from the observed trips, we introduce the notion of mini activities that the travellers do during the trips. We propose to mine the smart card data and identify characteristics that help detect the mini activities. We develop a technique to integrate them into the generated trips and learn such an integration from the trip history and trip planner recommendations. We test our method on the trip dataset collected in Nancy, France. The evaluation results demonstrate a very important reduction of the trip generation error, and a good capacity to cope with new simulation scenarios.

Taxi Demand Forecasting: A HEDGE-Based Tessellation Strategy for Improved Accuracy

This paper compares two widely used spatial tessellation strategies in the context of location-based taxi demand modeling. The dependence of the performance of the tessellation strategy on the spatial distribution of the data, the city geography, and the time of the day is highlighted. Motivated by the lack of a clear winning tessellation strategy, a HEDGE based combining algorithm is proposed to pick the best tessellation strategy at each time step. The proposed strategy performs consistently better than either of the two tessellation strategies across the data sets considered, at multiple time scales, and with different performance metrics.

On Learning From Inaccurate and Incomplete Traffic Flow Data

Today, pervasive sensor networks both collect and broadcast rich digital footprints about the human mobility. However, most of this data often comes in an incomplete and/or inaccurate fashion. In this paper, we propose a Knowledge Discovery Framework to handle such issues in the context of Automatic Incident Detection Systems fed with traffic flow data. This framework operates by firstly removing faulty sensors through a tailored unsupervised learning algorithm. Then, we propose a novel fundamental diagram that discovers the critical density of a given road section/spot on a data-driven fashion that is resistant to both outliers and noise within the input data. Large scale experiments were conducted over traffic flow data provided by a major Asian highway operator. The obtained results illustrate a drastic reduction of the noise within the raw data; it also allows to determine reliable definitions of traffic states on a completely automated way.

Fast and Scalable Big Data Trajectory Clustering for Understanding Urban Mobility

A novel, fast and scalable algorithm, Fast-clusiVAT is proposed to cluster large-scale vehicle trajectories in urban environments. To address the lack of appropriate distance measure between trajectories, a novel Dijkstra based Dynamic Time Warping distance measure, trajDTW is proposed, which is suitable for large numbers of overlapping trajectories in a dense road network, common in major cities around the world. Numerical experiments were conducted on a large-scale taxi trajectory dataset consisting of 3.28 million trajectories obtained from the GPS traces of 15,061 taxis within Singapore over a period of one month. The proposed approach provides insight into urban traffic patterns and how they change with time, particularly for optimizing public transport routes and frequencies.

Learning Low-Dimensional Representation of Bivariate Histogram Data

With more data becoming available in the automotive domain, new methods are needed to automatically extract general representations, suitable not only for known tasks, but also for (similar) ones that can emerge in the future. Finding a low-dimensional representation that can be used for multiple purposes is an important step towards knowledge discovery in aware intelligent transportation systems. This study evaluates several approaches for mapping high-dimensional sensor data into a low-dimensional representation useful for prediction. Original data are two types of bivariate histograms: turbocharger and engine. Low-dimensional representations were evaluated in a supervised fashion by mean equal error rate using random forest classifier on a set of 27 1-vs-Rest detection tasks. Results from unsupervised learning experiments indicate that the most effective way to create low-dimensional representation of the original bivariate histogram is by using autoencoder for an intermediate representation, followed by t-SNE.

Scale-Free Properties of Human Mobility and Applications to Intelligent Transportation Systems

We discuss the scale-free properties of some important human mobility characteristics, namely spatial node density and mobility degree, and show that they exhibit behavior that can be described by a power-law. Based on their power-law characteristics, we derive analytical models for the spatial node density and mobility degree and show that the data generated by the proposed analytical models closely approach empirical data extracted from the real mobility traces. We use the proposed analytical models to build a synthetic mobility regime that is suitable for simulations of intelligent transportation systems. We show, through network simulations, that ad-hoc network routing behavior under our mobility regime closely approximates routing behavior when the corresponding real trace is used.

On Evaluating Floating Car Data Quality for Knowledge Discovery

A new methodology for evaluating the quality of floating car data (FCD) is proposed. It leverages a set of statistical indicators covering multiple dimensions of FCD such as spatio-temporal coverage, accuracy and reliability. These indicators provide a quick and intuitive means to assess the potential 'value' and 'veracity' characteristics of the data.

LUIS MOREIRA-MATIAS
Kreditech Holding SSL
Hamburg, Germany
e-mail:
luis.moreira.matias@gmail.com

JOÃO GAMA
University of Porto
Porto, Portugal
e-mail: jgama@fep.up.pt

RAHUL NAIR
IBM Research Ireland
Dublin, Ireland
e-mail: rahul.nair@ie.ibm.com

CRISTINA OLAVERRI MONREAL
Johannes Kepler Universität Linz
Linz, Austria
e-mail: cristina.olaverri-monreal@jku.at

ROBERTO TRASARTI
KDD Lab–ISTI CNR
Pisa, Italy
e-mail: roberto.trasarti@isti.cnr.it

Luis Moreira-Matias received the M.Sc. degree in informatics engineering and the Ph.D. degree in machine learning from the University of Porto in 2009 and 2015, respectively. He was a recipient of the International Data Mining Competition held at the Research Summer School, TU Dortmund, in 2012. He served for the Program Committee of multiple high-impact research venues, such as KDD, AAAI, IEEE TKDE, ESWA, ECML/PKDD, and KAIS, among others.

From 2014 to 2018, he was a Senior Researcher with NEC Laboratories Europe, leading research lines in AutoML, with applications to transport and retail. He is currently the Head of data science at Kreditech, one of the largest European Fintechs. He has authored over 40 high-impact peer-reviewed publications. He was invited to give keynotes around the globe, in locations ranging from Brisbane, Australia, to Las Palmas, Spain.

João Gama received the Ph.D. degree in computer science from the University of Porto in 2000. He joined the School of Economics, where he holds the position of Associate Professor. He is also a Senior Researcher at LIAAD, a group belonging to INESC Porto. He has worked in projects and has authored papers in areas related to machine learning, data streams, and adaptive learning systems and in applications ranging from energy, economics, telco, and transportation, among others. Recently, he has authored a book *Knowledge Discovery from Data Streams*. He is a member of the editorial board of multiple international journals in his areas of expertise and has authored over 250 DBLP-listed papers.

Cristina Olaverri Monreal received the master's degree in computational linguistics, computer science, and phonetics from the Ludwig-Maximilians University, Munich, in 2002, and the Ph.D. degree in cooperation with BMW in 2006.

After working several years in different European countries and the U.S., both within the industry and academia, she is holding the BMVIT Endowed Professorship and the Chair for sustainable transport logistics 4.0 at Johannes Kepler University Linz, Linz, Austria. Her research aims at studying solutions for an efficient and effective transportation focusing on minimizing the barrier between users and road systems and applying wireless communication and sensing technologies. Her current research interests include automated driving, multi-functional systems for in-vehicle information, overall efficiency of user and system utilization, and driver behavior; simulation tools, and research concerning intelligent transportation systems (ITS). She was the General Chair of the IEEE ICVES 2017 Conference, the Chair of the Technical Activities Committee on Human Factors in the ITS Society, and the Vice-President of Educational Activities. She has been recently recognized for her dedicated contribution to continuing education in the field of ITS with the 2017 IEEE Educational Activities Board Meritorious Achievement Award in Continuing Education. In addition, she serves as an Associate Editor and as an Editorial Board Member of several journals in the field, including the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and the IEEE *Intelligent Transportation Systems Magazine*.

Rahul Nair received the Ph.D. degree from the University of Maryland, College Park, in 2010. He was with the Center for Advanced Transportation Technology, where he developed performance management systems based on probe data for the Federal Highway Administration and Maryland State departments. He has been a Research Staff Member at IBM Research Ireland, Dublin, since 2012. He has co-authored over 40 scientific publications. He focuses on mobility applications, multi-modal transportation systems, and innovative urban mobility solutions. He received the Best In-Use Paper Award for traffic diagnosis at the International Semantic Web Conference, the Best Development Insight Prize for the use of telecom data to optimize public transport in 2013, and the Best Doctoral Thesis at the Network Modeling Committee of the Transport Research Board of the National Academies.

He has previously organized the ITS Workshop in Dublin in 2013, co-located with the European ITS Conference, and organized sessions at the Transportation Research Board as part of the Network Modeling Committee.

Roberto Trasarti received the Ph.D. degree in 2010. His Ph.D. thesis is entitled "Mastering the Spatio-Temporal Knowledge Discovery Process." He has been a Researcher with the KDD Laboratory–ISTI CNR, Italy, since 2012. His main research topics are data mining, spatio-temporal data analysis, artificial intelligence, automatic reasoning, and parallel computation and privacy.

He has more than 50 publications with 1155 citations (most cited paper "Wherenext: a location predictor on trajectory pattern mining" with 365 citations) and with an H-index of 15 (source: Google Scholar). He is also a Work Package Leader in a project called SoBigData (H2020 no. 654024) with the objective of building an e-infrastructure able to create a research community for researchers. In particular, he is also responsible for the "city of citizens" exploratory focused on smart cities and mobility data analysis.

Dr. Roberto has already organized several workshops, such as DAMASCA and DATAMOD, and serves on the Program Committee of the International Conferences, such as ECML/PKDD, ICDM, and KDD.