In-Vehicle Big Data Exploration for Road Maintenance

(Discussion Paper)

Devis Bianchini[1], Valeria De Antonellis[1] and Massimiliano Garda[1]

[1] *University of Brescia, Dept. of Information Engineering*
Via Branze 38, 25123 - Brescia (Italy)

Abstract

Big Data Exploration techniques may benefit from the availability of huge amount of data (e.g., collected from IoT infrastructures) for improving resilience of monitored systems. In this paper, we discuss the application of such techniques in a research project to pursue mobility resilience in Smart Cities applications. Among the aspects to be considered for enabling resilience in mobility, we specifically focus on road maintenance, gathering data streams from vehicles equipped with sensors and designing proper exploration scenarios. Scenarios rely on three precise components as main pillars of the proposed approach: (i) a multi-dimensional model apt to represent the road network and to enable data exploration; (ii) data summarisation techniques, in order to simplify exploration of high data volumes; (iii) a measure of relevance, aimed at attracting the attention of the road maintainers on relevant data only.

Keywords

Multi-dimensional model, data summarisation, big data exploration, smart and resilient mobility.

1. Introduction

In the latest years, the increasing availability of big data has become a key factor in shifting towards a data-centric vision of modern Smart Cities [1]. In particular, the concept of smart mobility, and its impact on the transportation of goods and people, is experiencing radical changes, capitalising on big data generated from sensor networks and IoT devices [2]. Indeed, through such data, issues that can arise may be promptly noticed and tackled, increasing the efficiency of delivered services [3]. For instance, sensor data in vehicles may provide in near real-time valuable information about the quality of the area-wide road surface and may be used by road maintainers to focus monitoring and maintenance activities on urban and public infrastructure, for enhancing mobility resilience. In this landscape, road maintainers should be equipped with valuable tools to gain insights from the data and ensure a safer and more efficient infrastructure. Nevertheless, the variety and volume of collected data call for models, tools and methods for data representation and exploration [4]. To support road maintainers in analysing and assessing surface conditions of roads, in this paper we propose an approach, based on big data exploration techniques consisting in the following three components: (i) a *multi-dimensional model*, apt to represent portions of the road network (based on distinguishing features such as type of road, area/district, mileage extent) and to enable their data-driven

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

devis.bianchini@unibs.it (D. Bianchini); valeria.deantonellis@unibs.it (V. De Antonellis); massimiliano.garda@unibs.it (M. Garda)

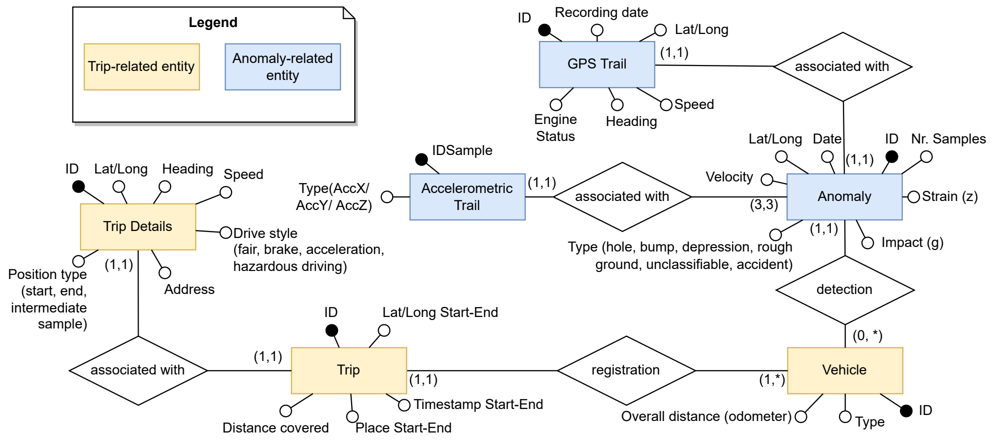

Figure 1: E-R model for monitoring road surface conditions (only a subset of attributes is displayed).

exploration; (ii) *data summarisation techniques*, in order to simplify exploration of high data volumes by extracting snapshots of measures gathered by vehicles, which evolve over time; (iii) a *measure of relevance*, aimed at attracting the road maintainers' attention on portions of the road network in the multi-dimensional model that present substantial changes over time (e.g., to plan corrective actions in the case of road conditions decay). This paper illustrates the application of the proposed approach in the scope of the MoSoRe project (Italian acronym for "Mobilità Sostenibile e Resiliente"), whose aim is to investigate the resilience of mobility systems and infrastructure in the city of Brescia (Italy). Specifically, three exploration scenarios have been devised to assist road maintainers when inspecting road surface conditions. A preliminary presentation of the approach has been provided in [5].

The paper is organised as follows: in Section 2 a conceptual model for data collected within the MoSoRe project is described; Section 3 presents the ingredients of the big data exploration approach; exploration scenarios for smart mobility are illustrated in Section 4, whereas Section 5 describes implementation and experimental evaluation; related work are discussed in Section 6; finally, Section 7 closes the paper, sketching future research directions.

2. Conceptual data model

In the MoSoRe project, a fleet of commercial vehicles has been equipped with black boxes, gathering data when vehicles transit on specific *road portions* (i.e., delimited sections of roads covered during daily trips). Collected data regards both contextual information (e.g., details of the journey) and measures from gravity acceleration sensors, which can be used to infer road surface conditions. Collected data is conceptually represented through the Entity-Relationship (E-R) diagram in Figure 1 as follows.

Trip-related data. This kind of data regards the journeys accomplished by vehicles during the monitored period (at the time of writing, data is transferred from vehicles on a daily basis). The black box collects the position (in terms of GPS coordinates) of the hosting vehicle at the

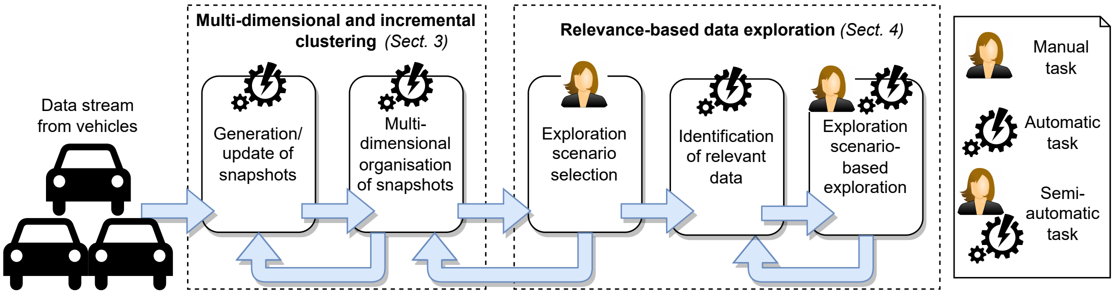
Data stream
from vehicles
Multi-dimensional and incremental clustering (Sect. 3)
Generation/ update of snapshots
Multi-dimensional organisation of snapshots
Relevance-based data exploration (Sect. 4)
Exploration scenario selection
Identification of relevant data
Exploration scenario-based exploration
Manual task
Automatic task
Semi-automatic task

Figure 2: Big data exploration approach overview.

begin/end of the trip, twice per kilometre and when some driving events occur (e.g., burst of speed, hard braking). For privacy preservation issues, only few characteristics of the vehicle are recorded (e.g., the type, to denote either a private or a commercial vehicle), but they cannot be used to infer any information regarding the owner.

Anomaly-related data. This kind of data concerns anomalous events recorded by the black box of a vehicle, which can be: (i) induced by bad road surface conditions or (ii) caused by car accidents. For both the types of anomalous events, the black box collects data from the accelerometers on X-Y-Z axes, with a rate which varies from 200 up to 800 Hz for each trace, depending on the model of the black box. Hence, depending on the frequency, accelerometric traces from different vehicles may have a different number of samples. In the scope of the research project, the big data exploration approach presented in this paper (whose overview is reported in Figure 2) is rooted on the analysis of anomalous events related to road surface conditions; indeed, each black box assigns a probabilistic estimate to the cause that induced the event (either hole, bump, depression, rough ground or undetermined – not assignable to any specific category). Noteworthy, upon anomalous event occurrence, a GPS trail is recorded and the position of the vehicle is sampled.

3. The ingredients of the Big Data Exploration approach

In this section, we briefly report the three building blocks of the big data exploration approach presented in [6], which is fostered in the MoSoRe project to cope with variety, velocity and volume of anomaly-related data, in order to let road maintainers monitor the status of the road surfaces and plan corrective actions.

Multi-Dimensional Model. The proposed Multi-Dimensional Model (MDM) is grounded on the following two main pillars: (i) *dimensions* and (ii) *exploration facets*. A dimension d_i is an entity representing a single aspect of a road portion (e.g., belonging city area, if it is part of either an urban or suburban road) defined on domain $Dom(d_i)$. A combination of different dimension instances is apt to identify a specific road portion and constitutes a facet $\phi_j = \{v_{d_i}, \ldots, v_{d_n}\}$ where $v_{d_i} \in Dom(d_i)$. We denote with Φ the set of all facets, representing road portions with homogeneous characteristics. An example of facet $\phi_1 \in \Phi$ may contain instances of the following four dimensions: {*RoadType*, *SpeedLimit*, *District*, *MileageExtension*}. For ϕ_1, *RoadType*

= Urban and *District* = District1 are sample dimension instances. The MDM is leveraged to organise the measures associated with physical quantities recorded by black boxes on vehicles, the latter referred to as *features*. Measures are conceived as a stream or a time series and are associated with specific road portions.

Clustering-based data summarisation. Once focusing the attention on a specific facet, the evolution over time of the stream of records from a single black box can be used to ascertain whether road surface conditions are diverging from reference values. To obtain an effective representation of the temporal evolution of a road surface condition, data summarisation based on an incremental clustering algorithm is applied. In particular, the clustering algorithm, which is based on the CluStream [7] algorithm, takes as input the stream of measures related to the observed features. At a given time t, the algorithm produces as output a set of syntheses $\mathcal{S}(t)$, where each synthesis corresponds to a cluster of records, starting from records collected from timestamp $t - \Delta t$ to timestamp t and built on top of the previous set of syntheses $\mathcal{S}(t - \Delta t)$, for a given road portion. Roughly speaking, syntheses conceptually represent a specific status of road surface conditions. A set of syntheses at a given timestamp t corresponds to a *snapshot* $SN_i(t)$, a data structure defined as the following tuple $\langle \mathcal{S}_i(t), F_{SN_i}, \phi_i, \epsilon_{SN_i}, ID_{\epsilon_{SN_i}} \rangle$, where: (i) $\mathcal{S}_i(t)$ is a set of syntheses generated at time t, (ii) F_{SN_i} is the set of the monitored features; (iii) ϕ_i is the facet that identifies the road portion; (iv) ϵ_{SN_i} is the type of anomalous event the snapshot refers to (i.e., either hole, bump, rough ground, depression, undetermined event); (v) $ID_{\epsilon_{SN_i}}$ is the identifier of the anomalous event as assigned by the black box of a vehicle.

Identification of relevant data. Given two snapshots $SN_i(t_1)$ and $SN_j(t_2)$ (with $i \neq j$ and $t_2 > t_1$), changes between syntheses in the two snapshots are apt to identify *relevant data*, which can be proposed to road maintainers to start the exploration from. In particular, the measure of relevance is based on the notion of *distance* between the syntheses sets $\mathcal{S}_i(t_1) \in SN_i(t_1)$ and $\mathcal{S}_j(t_2) \in SN_j(t_2)$, obtained by combining different factors to detect movements of syntheses, expansion/contraction and changes in density (i.e., the difference in the number of measures aggregated by the syntheses with respect to their hyper-volume). Snapshots for which the distance value from a *reference snapshot* $\overline{SN}(t_0)$ falls within an interval $[val_{min}, val_{max}]$ are highlighted as *relevant*. At the moment, the val_{min} and val_{max} are predefined thresholds set by road maintainers. For instance, different domain experts may set different threshold intervals to highlight relevant snapshots complying with their different expertise and goals (e.g., since road surface repair interventions imply a huge economic expense, a road maintainer may be forced to limit the intervention on a subset of anomalous events, selecting them according to the relevance value). Focusing on the set of syntheses of a relevant snapshot, it is possible to check what are the syntheses that changed over time (namely, appeared, merged or removed), which contributed to make that snapshot relevant.

4. Exploration scenarios for Smart Mobility

The big data exploration techniques illustrated in the previous section are being applied, in the MoSoRe project, to implement three exploration scenarios, targeted to assist road maintainers when inspecting snapshots of measures related to road surface conditions.

1) Analysis of the evolution over time of an anomalous event. In this exploration scenario, the road maintainer analyses a sequence of snapshots related to a single anomalous event, for the monitored road section $\widetilde{\phi}$. For the same road section, a reference snapshot, taken under normal conditions (i.e., not referred to an anomalous event) is considered. The goal of this scenario is to compare the evolution of syntheses in the sequence of snapshots against the reference snapshot. As a result, it is possible to understand the triggering causes of the anomalous event, focusing only on relevant snapshots and, for such snapshots, inspecting the evolution of syntheses that changed over time. Starting from the dimensions in $\widetilde{\phi}$, and considering the Multi-Dimensional Model, a road maintainer may apply the renowned OLAP operators (e.g., roll-up, drill-down) to discover the *Area* of the city with the highest percentage of relevant snapshots. Once found, he/she can then narrow down the exploration with a drill-down operator, enabling the inspection of anomalous events at *District* level.
2) Comparison of anomalous events of the same type. In this exploration scenario, the snapshots considered for exploration are the ones belonging to the monitored road section $\widetilde{\phi}$, having a type $\widetilde{\epsilon}$ (e.g., hole), but related to different events. Considering a single reference snapshot corresponding to a critical event (e.g., occurred in the past on the monitored road section), the goal of this scenario is to determine which snapshot reflects the highest critical situation. To this aim, the distance values between analysed snapshots and the reference critical one are exploited to establish an order from the most to the least relevant. Hence, road maintainers may exploit the organisation of road portions from the MDM, to identify where the most severe anomalous events of a certain type occurred and then, taking one of such anomalous events, they can resort to the scenario (1) to analyse the temporal evolution of snapshots associated with the event.
3) Classification of an undetermined event. The goal of this scenario is to determine the similarity of an undetermined event with respect to the known typologies (in the MoSoRe project: hole, bump, depression and rough ground), thus performing a classification task. In this respect, four reference snapshots are considered, one for each of the aforementioned types. Snapshots to be considered for analysis are the ones of an undetermined event for a monitored road section $\widetilde{\phi}$. Classification is accomplished by calculating the distances of snapshots considered for analysis from each of the four reference snapshots, focusing only on the relevant ones. The lowest distance corresponds to the highest similarity, which is used to properly classify the undetermined event. Similarly to the former scenarios, road maintainers may focus on road portions with a considerable rate of undetermined events occurrence to start the exploration from. Through this scenario, they can firstly ascertain the classification of such events, and then they can assess their severity, resorting to the scenario (2).

5. Implementation and preliminary validation

Architecture. In Figure 3, the architecture implementing the approach is reported. Clustering-based data summarisation modules have been implemented in R relying on the streamMOA library. Mobility data is made available by the vehicles black boxes provider on an FTP repository

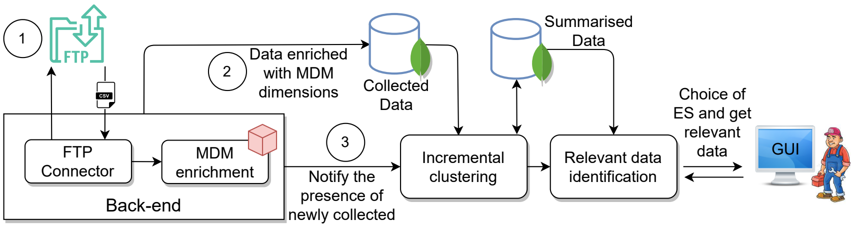
1
FTP
CSV
2
Data enriched with MDM dimensions
Collected Data
Summarised Data
FTP Connector
MDM enrichment
Back-end
3
Notify the presence of newly collected
Incremental clustering
Relevant data identification
Choice of ES and get relevant data
GUI

Figure 3: Architecture overview.

(in the form of CSV files). The Back-end modules: (i) listen for new files containing measures; (ii) store the measures, automatically enriched with the dimensions of the MDM into a MongoDB database (Collected Data) as JSON documents organised into collections (a collection contains data related to a day). Summarised Data, obtained through the incremental clustering algorithm, relies on MongoDB technology as well. The numbers on the arrows in Figure 3 denote the interaction flow between modules. Once CSV files containing anomalous events, vehicles trips and accelerometric trails are available on the FTP repository, they are retrieved by the Back-end (1), where measures are associated with the dimensions of the MDM, before being stored within MongoDB (2). Then, the Incremental Clustering module is notified about the presence of available data to process from the Collected Data store (3). The output of the Incremental Clustering module is stored within the Summarised Data store and then sent to the Relevance Data Identification module, which is in charge of: (i) identifying relevant snapshots according to the exploration scenario selected; (ii) sending relevant snapshots to be displayed on a GUI.

Preliminary validation. Experimental evaluation aims at: (i) testing the processing time, to prove that data summarisation and the measure of relevance can be efficiently computed, thus facing high acquisition rates; (ii) assessing the quality of relevance evaluation, as the correlation between high relevance values in correspondence to time instants where variations of the collected data occurred. Currently, the average number of daily anomalous events is 400, producing $\approx 3 \cdot 10^6$ records of accelerometric measures. For the experiments, we considered a stream of measures composed of 14160 samples. Regarding processing time, we performed tests varying the width of the time window Δt, retaining the latest collected data to be processed by the clustering algorithm. Figure 4(a) shows the average time required by the incremental clustering algorithm and relevance evaluation to process a single record of measures for different Δt values. Lower Δt values demand more time to process data. Indeed, every time data summarisation and relevance evaluation are performed, some initialisation operations have to be executed (e.g., access to the set of syntheses previously computed). Therefore, to ensure lower processing time, the frequency of clustering execution and relevance evaluation must be reduced, that is, Δt value must be increased. Instead, higher Δt values indicate that clustering execution and relevance evaluation could be performed far from time instants where important variations occurred, thus reducing the quality of data relevance evaluation. This is evident in Figure 4(b), where two different Δt values have been used for demonstration. The rationale is to adaptively increase/decrease Δt value according to the distance of relevant syntheses from warning and error thresholds for the observed features, depending on the road portion, since they correspond to potentially critical situations that must be monitored at finer granularity.

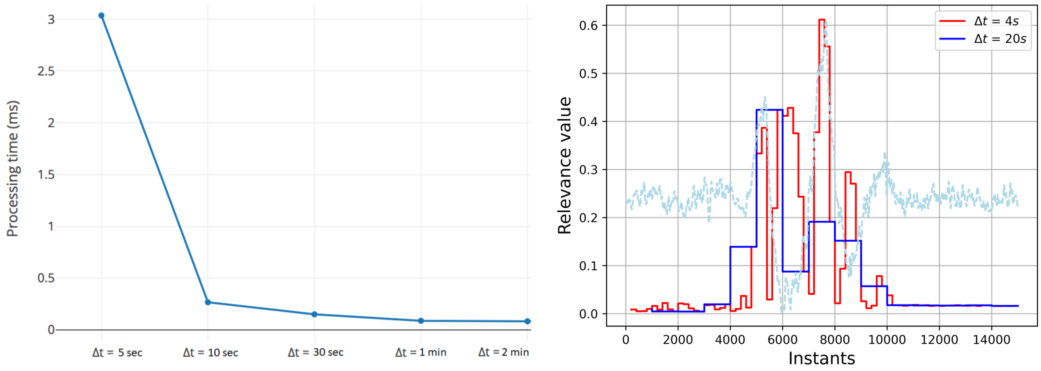

Figure 4: (a) Processing time for each collected measure. (b) Relevance evaluation for different Δt.

6. Related Work

In the literature, several research efforts proposed the adoption of comprehensive solutions for big data exploration, to improve the resilience of Smart City mobility. Authors in [8] propose a framework for analysing road accident data; therein, after data preprocessing, a clustering algorithm is applied and association rules are mined to find possible underlying patterns in the data set. With similar intents, the work in [9] combines IoT and big data to devise the Pavement Managements System (PMS), a road maintenance management structure composed of pavement detection and 3D modelling, data analysis and decision support. It also illustrates use cases for two main actors, the road maintenance company and a technical firm that offers smart solutions for road maintenance. In [10], a city traffic state assessment system is implemented using a big data cloud infrastructure, assuring high scalability, to host clustering methods and find areas of jam. Leveraging the recent advances in the field of computer vision and big data computing, authors in [11] developed a scalable framework for image-based monitoring of urban infrastructure, using both Web images and Google Street View imagery to train a CNN model. Pursuing the goal of analysing road traffic and pollution data for the city of Aaruhs (Denmark), in [12] big data technologies ease the calculation and visualisation of the least polluted route. Despite multi-dimensional data organisation is not envisaged in any of the former work, we share with [8] the introduction of metrics to identify relevant data. Furthermore, only [8, 10] foster summarisation techniques, both of them relying on clustering. Nevertheless, in [8] several clustering algorithms from the literature are cited, but none of them is conceived to be applied incrementally on a stream of data, whilst in [10] details on how the algorithm is applied are not provided. Regarding the formulation of exploration scenarios to support data exploration, only [12] sketches scenarios targeted to Smart Mobility, but details are coarsely given.

7. Concluding remarks

In this paper, we described our contribution in the scope of the MoSoRe research project, presenting an approach based on big data exploration techniques to support road maintainers in analysing and assessing road surface conditions. The approach includes three precise components: (i) a multi-dimensional model apt to represent the road network and to enable data exploration; (ii) data summarisation techniques, in order to simplify exploration of massive data streams collected by vehicles; (iii) a measure of relevance aimed at attracting the attention of the road maintainers on relevant data only. Moreover, the paper illustrates the application of the approach in three exploration scenarios. Future research efforts regard the formalisation of an exploration methodology rooted on exploration scenarios, taking into account also personalisation aspects for road maintainers (e.g., in order to explore relevant snapshots more related to their analysis interests) and the setup of an extensive campaign of usability experiments, to be performed in the last phases of the MoSoRe project on the prototype GUI used for exploration purposes. Additionally, a generalisation of the proposed approach will be also investigated for other domains permeated by big data (e.g., healthcare, robotics).

References

[1] C. Lim, K.-J. Kim, P. P. Maglio, Smart cities with big data: Reference models, challenges, and considerations, Cities 82 (2018) 86–99.

[2] S. Paiva, M. A. Ahad, G. Tripathi, N. Feroz, G. Casalino, Enabling technologies for urban smart mobility: Recent trends, opportunities and challenges, Sensors 21 (2021) 2143.

[3] S. E. Bibri, The anatomy of the data-driven smart sustainable city: instrumentation, datafication, computerization and related applications, Journal of Big Data 6 (2019) 59.

[4] S. Campos-Cordobés, J. Del Ser, I. Laña, I. I. Olabarrieta, J. Sánchez-Cubillo, J. J. Sánchez-Medina, A. I. Torre-Bastida, Big data in road transport and mobility research, in: Intelligent Vehicles, 2018, pp. 175–205.

[5] D. Bianchini, M. Garda, Big data exploration techniques for road surface conditions assessment, in: 7th Italian Conf. on ICT for Smart Cities and Communities (I-CiTies), 2021.

[6] A. Bagozi, D. Bianchini, V. De Antonellis, M. Garda, A. Marini, A relevance-based approach for big data exploration, Future Generation Computer Systems 101 (2019) 51 – 69.

[7] C. Aggarwal, J. Han, J. Wang, P. Yu, A framework for clustering evolving data streams, in: Proc. of 29th Int. Conf. on Very Large Data Bases (VLDB), 2003, pp. 81–92.

[8] S. Kumar, D. Toshniwal, A data mining framework to analyze road accident data, Journal of Big Data 2 (2015) 1–18.

[9] J. Dong, W. Meng, Y. Liu, J. Ti, A framework of pavement management system based on iot and big data, Advanced Engineering Informatics 47 (2021) 101226.

[10] C.-T. Yang, S.-T. Chen, Y.-Z. Yan, The implementation of a cloud city traffic state assessment system using a novel big data architecture, Cluster Computing 20 (2017) 1101–1121.

[11] M. Alipour, D. K. Harris, A big data analytics strategy for scalable urban infrastructure condition assessment using semi-supervised multi-transform self-training, Journal of Civil Structural Health Monitoring 10 (2020) 313–332.

[12] J. Zenkert, M. Dornhofer, C. Weber, C. Ngoukam, M. Fathi, Big data analytics in smart mobility: Modeling and analysis of the aarhus smart city dataset, in: 2018 IEEE Industrial Cyber-Physical Systems (ICPS), 2018, pp. 363–368.