DATA STREAM HARMONIZATION FOR HETEROGENEOUS WORKFLOWS

Eleftherios Bandis
Nikolaos Polatidis
Maria Diapouli
Stelios Kapetanakis
University of Brighton
Moulsecoomb Campus, Brighton NB2 4GJ, UK
{e.bandis, n.polatidis, m.diapouli, s.kapetanakis}@brighton.ac.uk

KEYWORDS
Data stream workflows, Graph Reasoning, Monitoring

ABSTRACT

Transport infrastructure relies heavily on extended multi sensor networks and data streams to support its advanced real time monitoring and decision making. All relevant stakeholders are highly concerned on how travel patterns, infrastructure capacity and other internal / external factors (such as weather) affect, deteriorate or improve performance. Usually new network infrastructure can be remarkably expensive to build thus the focus is constantly in improving existing workflows, reduce overheads and enforce lean processes. We propose suitable graph-based workflow monitoring methods for developing efficient performance measures for the rail industry using extensive business process workflow pattern analysis based on Case-based Reasoning (CBR) combined with standard Data Mining methods. The approach focuses on both data preparation, cleaning and workflow integration of real network data. Preliminary results of this work are promising since workflow integration seems efficient against data complexity and domain peculiarities as well as scale on demand whilst demonstrating efficient accuracy. A number of modelling experiments are presented, that show that the approach proposed here can provide a sound basis for the effective and useful analysis of operational sensor data from train Journeys.

INTRODUCTION

The modernisation of Rail industry has led to increasing usage of computer systems for logistics, tactical, planning, performance and maintenance reasons. Rail industry has experienced substantial growth over the last decade in terms of operational method advancement (wayside detectors, wheel profile monitors, extended sensor network), processes, software and hardware equipment (Rail Defect Test Facility, Asset Health Strategic Initiative, and others). These systems generate millions of records per day that are constantly monitored, enhanced and analysed with the aim to improve industry capability, reduce cost and ultimately increase customer satisfaction.

Most rail operations, such as scheduled train services can be treated as business workflows, since they comprise event trails of spatio-temporal data. Techniques developed and tested for monitoring workflow operations can also be used in the context of live train journey auditing and performance measurement.

An example of such systems that fit well workflow orchestration and choreography is Remote Condition Monitoring (RCM) systems. RCM comprise multi-sensor systems per any running vehicle that can offer the full picture of a how a locomotive performs within a pre-determined time span (minute, hour, day, etc.). Its captured information is very low level and can reproduce a train journey with all relevant mechanical data. RCM is primarily used for technical -incident- monitoring, however it has also been observed as an accurate indicator of performance malfunctioning over a period of time.

Rail networks are prone to delays since order has to be maintened with emphasis to driver and passenger safety, cost and performance. Workflow techniques based on data streams and process mining can be increbibly valuable to Train Operator Companies (TOCs) to understand bottlenecks, increase capacity and minimize cost throughout the networks. This paper presents a data harmonization approach for spatio-temporal data using graph representation and general time theory (Ma, 1994) which enable data harmonization across multi-provenance sensor streams. This work, although quite recent in inception, has been proven reliable for heavy volume data (Agorgianitis, 2016)systems and effective in real time TOC data. This paper is structured as follows: Literature section witll refer to state of the art work in the field, Methodology will present the rationale and foundation principles of this work, Evaluation will presents real life data integrations with TOC Data. Finally, Conclusion will describe results as well as next steps for this work.

LITERATURE

Modern organisations use Business Process Workflows (BPW) to coordinate their processes, tasks, roles and manage resources with the aim to improve efficiency, efficacy and profitability. Workflows can automate processes, make them more agile and increase monitoring for obscure, erroneous or complex events to

Communications of the ECMS, Volume 35, Issue 1,
Proceedings, ©ECMS Khalid Al-Begain, Mauro Iacono,
Lelio Campanile, Andrzej Bargiela (Editors)
ISBN: 978-3-937436-72-2 / 78-3-937436-73-9(CD) ISSN 2522-2414

company managers to increase productivity (Workflow Management Coalition, 2021; BPMI, 2021). BPW management differs across organisations. The size, sector and strategic orientation of an organization plays a key role on how they adopt, analyse and practice BPWs (Van der Aalst, 2003). A common taxonomy includes the phases of: Design, Implementation, Enactment, Monitoring and Evaluation as the workflow life cycle in BPW management (Muehlen, 2004). Among those the Monitoring phase enables the supervising of business processes in terms of management (e.g. performance, accuracy) and organization (e.g. utilization of resources, length of activities etc.) (Reijers, 2003). Monitoring is key operation informing process managers and workflow designers necessary adjustments to improve their processes.

In the case of using Business process Modelling techniques to monitor train jour-ney operation there is a need to integrate various data from different rail systems, as well as the timetable to provide a detailed insight into real train journeys. RCM data are key to provide the basis of this analysis, but there is a considerable challenge to associate, workflow execution trails with the expected business process instances (i.e. timetable). This has proven to be a complicated task as several problems exist within the Railway data collection systems. For example:

- RCM systems are independent enough, installed on several trains at contrasting times. They generate data that denote a workflow process execution, however, there is no available information (linkage) between monitored workflow traces and their corresponding workflow on a seasonal timetable.
- Data monitoring has several phases. Firstly, telemetric sensors are used to gather data as "low level events". Then data is filtered by a processing system to produce workflow processes. Finally, the extracted workflows are stored on persistence lay-ers of variant formats. Each phase represents a single entity since it is created at various times and by different architectures. Consequently, the data transformation along each phase allow margin for error which leads to partially inconsistent, in-complete and ultimately faulty data. Through data analysis which has been con-ducted on real RCM datasets we found that such percentage can vary but it ultimately can affect crucial attributes making workflow generation and workflow alignment to business process extremely difficult.
- Transport industry has many similar processes. For instance, the same route might run multiple times within a few minutes interval. It is difficult to distinguish identical processes since most of their attributes having significant similarity.
- RCM data can contain missing and erroneous values -due to different clocks, analogue sensors and error-prone data transmission systems and areas (such as tunnels)-.
- TOCs have several fleets of similar trains that may employ several dif-ferent RCM systems. Several processes can be stored in different da-tasets which make workflow operations substantially complex.
- Data format can follow several popular or bespoke formats, hardening a universal workflow monitoring approach.

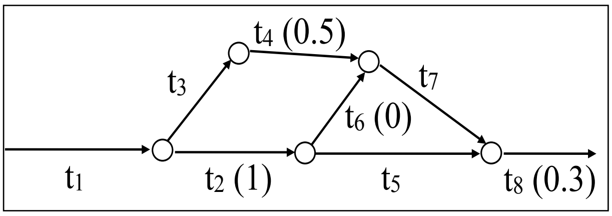
Workflow experts can use various methods to evaluate their processes, however, large or extended volumes of data can make the analysis of event logs extremely difficult. Process Mining (PM) is the technique used to extract knowledge and insights by discovering and analysing processes from event logs (Van der Aalst, 2011). By applying process mining, domain experts can use the derived information as feedback to design new processes or revise and enact predefined ones. In the literature, several algorithmic techniques have been introduced to solve the process mining problem. Algorithms like Alpha miner and alpha+ have been used extensively but other heuristics, genetic and fuzzy algorithms have also been applied (Tiwari, 2008). Each algorithm has its limitations on a different aspect of the process discovery such as fitness, simplicity and precision, and they may be unfit to areas where uncertainty, inconsistency and fuzziness is present. In such cases a CBR approach (Alshammari, 2017) may be more appropriate. CBR has been proven effective in monitoring business process workflow instances under uncertainty (Kapetanakis, 2009; 2010; 2011; 2012; 2013; 2014) in different interdisciplinary domains (Adedoyin, 2017), (Al Murayziq 2015, 2017), (Amin, 2019, 2020), (Ekpenyong, 2019), (Lansley,2019), (O'Connor, 2018) by retrieving similar solutions for similar problems.

RESEARCH METHODOLOGY

Our workflow data follow a sequential temporal and spatial pattern since they represent a variety of activities over time. Information about workflows can be encoded as events (points in time) or states (time intervals). In order to combine the two representation primitives and retain the full information and its provenance, there is a need for a formal underlying theory and representation that captures both temporal information and temporal relations (order, concurrency etc). To represent effectively workflows and their sequence and relationships in a formal way we use the General Time Theory (GTT) (Bandis, 2017; 2018), (Petridis, 2014). The general time theory takes both points and intervals as primitive. It consists of a triad (T, Meets, Dur), where:

- T is a non-empty set of time elements;
- Meets is a binary order relation over T;
- Dur is a function from T to R0+, the set of non-negative real numbers.

A time element t is called an interval if Dur(t) > 0; otherwise, t is called a point



Graphical representation of a log temporal inference using the GTT

In a graph representation each node represents a station whereas any edge represents the duration from station A to station B. A GTT workflow representation allows for a unified log interpretation which in conjunction with the multi-level similarity representation presents a foundation for adequate CBR workflow cases (Kapetanakis, 2014).

REPRESENTATION

A workflow process consists of multiple activities. Activities involve tasks such as "start of a journey", "departure from a station", "arrive on a station" or "end of a journey". The tasks contain multi-perspective information such as:

1. Time-related information: The start and the end of each activity is marked with a timestamp. The duration of an activity is also given.
2. Location: The station of which the activity takes place
3. Relationships: One activity holds which activity follows as well as the time duration between them

General information about the workflow is also available:

1. The total duration of all activities
2. The train unit responsible to undertake all the workflow activities
3. The day of the week the workflow took place
4. The workflow start and end time

Workflows are represented as GTT event-duration graphs with spatial information as node-specific tags. Every node can be represented as:
{$StationName_q$, $StopDuration_q$, $NextStation_q$, $TimeUntilNextStation_q$}

Similarity among graphs is represented using multi-level representation based on the workflow structure. This can be annotated as:
Level 1: Relevant timestamps from workflow data.
For example, Let case 1, C_1 and case 2, C_2 as workflow representations and C_{1L}, $C_{2L'}$ their respective list of stations. For C_1 and C_2 if Start date is the same (Binary equal) && Start time relies within γ mins fluctuation && C_{1L} is like $C_{2L'}$ based on an μ string threshold.

$$distance\,(C_1,C_2) = |\; StartTimeC_1 = StartTimeC_2 \leq \gamma| * w_1 + |EndTimeC_1 - EndTimeC_2 =< \gamma| * w_2 + |StationListC_1 - StationListC_2| * w_3 \quad \text{(equation 1)}$$

Where w_1, w_2, w_3 are empirically (expert-based) derived domain constants and

$$w_1 + w_2 = w_3 \quad \text{(equation 2)}$$

Upon successful relevance on similarity 1, a Level 2 similarity can be defined as:

p_1: create relationships => {[S_1, Dur(S_1), Dur(S_2), Meets S_2] ...}

(equation 3)

Where S_1 is a starting point, Dur(S_1) is the time spent on the station, Dur(S_2) the time till the next station, and Meets S_2 the station that follows. A Level 2 similarity is based on equation 3 quadruplets as:

$$distance\,(C_1,C_2) = |\,[S_1, Dur_{S1}, Dur_{s2}, S_2]C_1 - [sS_1, Dur_{S1}, Dur_{s2}, S_2]C_2| * w_1 + |\; StartDayOnlyC_1 = StartDayOnlyC_2| * w_2 + |UN_1 = UN_2| * w_3$$

(equation 4)

Where UN_1 and UN_2 are system identification numbers

EVALUATION

For the needs of evaluation we used data from 159000 trail records approximately over the period of ten months. Workflows were represented as graphs using GTT. Moving windows using level 1 and 2 similarities repsecitively, were used to combine together relevant workflows. Four types of datasets were used including:

1) RCM data from live train journeys
2) Performance data from planned / expected, already ran journeys
3) Timetabling data indicating planned, long-term planned and emergency routes across all networks
4) Spatio-temporal data for any assets (stations, signals, depots) and train location data available from sensors

GTT enabled workflow representation for all datasets starting from structured ones, like: Timetabling and Locations as well as free form ones: Performance and RCM. Level 1 and 2 similarities enabled workflow alignment and match of segments with complementary data provenance and information. Every performance journey was ranked with an indicator of delay which could be

1. Type A: No delay
2. Tybe B: Sub-threshold delay between 1-3'
3. Type C: Recorded Delay between 3-15'
4. Type D: Severe Delays of more than 15'

These classification scale was available just to one type of workflows and not the others. With the workflow unification, industry experts were able to see the journey classification as well as retrace back what happened on that specific case, see relevant information for the underlying family of services, routes as well as any available information on a daily basis. Based on the combined multiple provenance workflow data machine learning tehchniques were used to verify the accuracy of the system in numerical prediction e.g. given a specific trail of data can this be attributed to the right family of workflows and can it be classified accurately against delays of type A-D.

For the first part of the evaluation the aggregation results using GTT enabled graphs and level 1, 2 similarity were encouraging with 93.89% success rate.

Table 1 summarises the results in terms of successful vs. unsuccessful cases.

	Accurate Match	Total records
Workflow records	100%	159000
Matched successfully	93.89%	149282
Unsuccessful match	6.11%	9718

Table 1: Workflow match accuracy

Workflow matching had a high match ration, however still a high number of cases was not able to be connected dud to data inconcistencies, duplicate records and hardware peculiarities that required further processing and filtering. The results from this initial phase were treated as encouraging from industry stakeholders and requested the emphasis of the evaluation work to be placed on delay prediction given partial visibility of real time datasets. For this phase BPW mining techniques in workflow numerical prediction were used by applying generalized linear model, regression and a neural network classifier trained from existing workflow. Target was set as predicting whether a service will experience delay using early available data from the beginning of each route. A typical route can contain any number of stop between the range of 18 - 50 stations approximately. The first three nodes for each workflow graph where used as predictors for a combined workflow journey. For the needs of the evalution just week working days were selected as well as peak times where most delays take place usually.

	Generalised Linear Model	**Regression**	**ANN**
Min Error	-878	-1025	-476
Max Error	1754	1831	1907
Mean Absolute Error (MAE)	56	58	68
Standard Deviation	102	106	96
Linear Correlation	0.756	0.787	0.863
Occurrences	96,671	96,671	96,671

Table 2: Predicution results, journey times in seconds

As shown in Table 2, neural network predictors were shown most accurate in predicting delay. Results were interpreted positively from rail experts, however they expressed views for further workflow segmentation, special cases identification and filtering (for abnormal events) as well as the need for further explainability which will be the focus for further work.

CONCLUSION

This work presents a workflow harmonization approach in a real industrial environment. This work has been promising to domain experts since it is able to collate together workflows originating from different origins and present them under a common ground. There is substantial amount of improvement that can be applied in this field. Further work will focus explicitly on specialized workflow segmentation, algorithmic explanation and enhancement of the workflow auditing results. This approach seems generic and reusable to other domains, work which will be pursued in the future phases of this work.

REFERENCES

Adedoyin, A., Kapetanakis, S., Samakovitis, G., Petridis, M. (2017) Fraud Detection in Mobile Payment Transfer, In proceedings of the 22nd UK CBR workshop, Peterhouse, December 2017,(Ed M. Petridis), Brighton press, pp. 41-44

Agorgianitis, I., Petridis, M., Kapetanakis, S., Fish, A. (2016) Evaluating Distributed Methods for CBR Systems for Monitoring Business Process Workflows. In proceeding of ICCBR 2016, Workshop on Reasoning about time in CBR, Atlanta, GA, October 28-November 2, 2016, pp.122-131

Al Murayziq, T. S., Kapetanakis, S.,Petridis, M. (2015). Towards successful prediction of Dust Storms using Case-based Reasoning and Artificial Neural Networks. In proceedings of the 20th UK CBR workshop, Peterhouse, December 2015,(Ed M. Petridis), Brighton press, pp. 58-67

Al Murayziq, T., S., Kapetanakis, S., Alshammari, G., Petridis, M. (2017) Identifying and Predicting the Dust Events by Using Case Based Reasoning (CBR) , In Proceedings of the 22nd UK CBR workshop, Peterhouse, December 2017,(Ed M. Petridis), Brighton press, pp. 45-46

Alshammari, G., Jorro Aragoneses, J. L., Kapetanakis, S., Petridis, M., Recio-Garcia, J. A., Diaz-Agoudo, B. (2017) A hybrid CBR approach for the long tail problem in recommender systems. In proceedings of The International Conference in Case-based Reasoning (ICCBR 2017), pp. 35-45

Amin, K., Kapetanakis, S., Althoff, K., Dengel, A., Petridis, M. (2019) Building Knowledge Intensive Architectures for heterogeneous NLP workflows, In proceedings of the AI 2019

Amin, K., Kapetanakis, S., Polatidis, N., Althoff, K., Dengel, A. __(2020)__DeepKAF: A Heterogeneous CBR & Deep Learning Approach for NLP Prototyping, International Conference on Innovations in Intelligent Systems and Applications: INISTA 2020. IEEE

Bandis, E, Kapetanakis, S, Petridis, M, Fish, A. 2017. Effective Similarity Measures for Process Mining Using CBR on Rail Transport Industry, *in Proceedings of the 22nd UKCBR worksho*p, Cambridge UK

Bandis, E, Petridis, M, Kapetanakis, S,: Predictive Process Mining Using a Hybrid CBR Approach for the Rail Transport Industry, in RATIC 2018, *Proceedings of the 26th International Conference in Case Based Reasoning*, Stockholm, Sweden 9-12 July 2018

Business Process Management Initiative (BPMI): BPMN 1.1: OMG Specification, February 2008, http://www.bpmn.org/, accessed Feb 2021

Ekpenyong, F., Samakovitis, G., Kapetanakis, S., Petridis, M. (2019) An ensemble method: Case-Based Reasoning and the Inverse Problems in Investigating Financial Bubbles, in Proceedings of the International Conference on Cognitive Computing (ICCC 2019)

Kapetanakis, S., Petridis, M., Ma, J., Bacon, L. (2009). Workflow Monitoring and Diagnosis Using Case Based Reasoning on Incomplete Temporal Log Data. Proceedings of the Workshop on Uncertainty, Knowledge Discovery, and Similarity in Case Based Reasoning UKDS, in Workshop proceedings of the 8th International Conference on Case Based Reasoning, Seattle, USA, 2009

Kapetanakis, S., Petridis, Ma, J., Bacon, L. 2010. Providing explanations for the intelligent monitoring of business workflows using case-based reasoning. In: Roth-Berghofer, T., Tintarev, N., Leake, D. B., Bahls, D. (eds.) *Proceedings of the 5th International Work-shop on explanation-aware Computing Exact* (ECAI 2010), Lisbon, Portugal

Kapetanakis, S.,Petridis, M.,Knight, B.,Ma, J.,Bacon,L. 2010. A Case Based Reasoning Approach for the Monitoring of Business Workflows, *18th International Conference on Case-Based Reasoning,* ICCBR 2010, Alessandria, Italy, LNAI

Kapetanakis, S., Petridis, M., Ma, J., Knight, B., Bacon, L. (2011). Enhancing Similarity Measures and Context Provision for the Intelligent Monitoring of Business Processes in CBR-WIMS, In: Process-oriented Case-Based Reasoning workshop (PO-CBR), ICCBR2011

Kapetanakis, S., Samakovitis, G., Gunasekara, B., Petridis, M. (2012). 'Monitoring Financial Transaction Fraud with the use of Case-based Reasoning', Seventeenth UK Workshop on Case-Based Reasoning (UKCBR 2012), 11th December 2012, Cambridge, UKa

Kapetanakis, S., Samakovitis, G., Gunasekara, B., Petridis, M. (2013). The Use of Case-Based Reasoning for the Monitoring of Financial Fraud Transactions, Journal of Expert Update Vol. 13 (1) pp.75-83

Kapetanakis, S., Petridis, M. 2014. Evaluating a Case-Based Reasoning Architecture for the Intelligent Monitoring of Business Workflows, *in Successful Case-based Reasoning Applications-2*, S. Montani and L.C. Jain, Editors, Springer Berlin Heidelberg. p. 43-54

Lansley, M., Polatidis, N., Kapetanakis, S., Amin, K., Samakovitis, G., Petridis, M. (2019) Seen the villains: Detecting Social Engineering Attacks using Case-based Reasoning and Deep Learning, In Proceeding of the Deep Learning Workshop in Case-based Reasoning, ICCBR 2019

Ma, J., Knight, B. 1994. A General Temporal Theory, the Computer Journal, 37(2), 114-123

O' Connor, D., Kapetanakis, S., Samakovitis, G., Floyd, M., Ontañon, S., Petridis, M. (2018) Autonomous Swarm Agents using Case-based Reasoning In Proceedings of the Thirty-eighth SGAI International Conference on Artificial Intelligence, AI 2018, pp. 210-216

Petridis, M., Kapetanakis, S., Ma, J., Burlutskiy, N. (2014). Temporal Knowledge Representation for Case Based Reasoning Based on a Formal Theory of Time. In: Gundersen, O. E., Montani, S. (eds) Proceedings of RATIC: Reasoning about Time in CBR, ICCBR 2014, pp. 154-164, Springer, Heidelberg(2014)

Reijers, H.A. 2003. Design and Control of Workflow Processes: *Business Process Management for the Service Industry*. Springer, Heidelberg

Tiwari, A., Turner, C. J., & Majeed, B. 2008. A review of business process mining: State-of-the-art and future trends. *Business Process Management Journal*, 14(1),5–22

Van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. 2003. Business Process Management: A Survey. *In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) BPM 2003. LNCS*, vol. 2678, pp. 1-12. Springer, Heidelberg

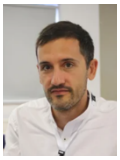
Van der Aalst. 2011. Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, Berlin

Workflow Management Coalition. Workflow management coalition glossary & terminology. http://www.wfmc.org/standars/docs/ TC1011_term_glossary_v3, 2021

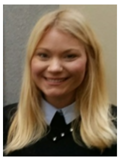
Zur Muehlen, M. 2004. Workflow-Based Process Controlling: Foundation, Design and Application of Workflow-driven Process Information Systems. Logos

AUTHOR BIOGRAPHIES

Eleftherios Bandis has a BSc in Software Enginering and a PhD in Machine Learning from the Univesity of Brighton, UK. He worked for several years in the Transportation Industry as Data Engineer and Machine Learning Expert. His expertise resides in Real time systems, Spatiotemporal Workflows and Graph Theory. His e-mail address is : e.bandis@brighton.ac.uk



Nikolaos Polatidis has a BSc in Computer Science from Heriot-Watt University, an MSc in Internet Software Systems from the University of Birmingham, UK and a PhD from University of Macedonia. His background is in Artificial Intelligence, Machine Learning and Cyber Security. His e-mail address is : n.polatidis@brighton.ac.uk



Maria Diapouli has a BSc in Software Enginering and an MSc in Enterprise Systems from the Univesity of Greenwich, UK. Her background is in Advanced Databases and Distributed Systems. She has 10 years of experience in the Transportation and Online Marketing Industry. Her e-mail address is : m.diapouli@brighton.ac.uk

Stelios Kapetanakis has a PhD in Artificial Intelligence and an MBA in Knowledge and Innovation Management. He has been a Principal Lecturer in the University of Brighton as well as technical consultant for several startups in Europe , Australia and the US. His work focuses on machine learning solutions in enterprise environments. His e-mail address is : s.kapetanakis@brighton.ac.uk