Aligning Process Mining Workflows using Case-based Reasoning

Eleftherios Bandis[1], Stelios Kapetanakis[1], Miltos Petridis[2] and Andrew Fish[1]

[1] School of Computing, Engineering and Mathematics, University of Brighton, Moulsecoomb Campus, Brighton NB2 4GJ, UK
{E.Bandis, S.Kapetanakis, A.Fish}@brighton.ac.uk

[2] Department of Computer Science, Middlesex University London The Burroughs, Hendon, London NW4 4BT
M.Petridis@mdx.ac.uk

Abstract. Rail transportation improvements have always been considered of high impact to society due to their tangible improvement to quality of life in modern cities. Both public and private companies are highly concerned on how travel patterns, vehicle-passenger behaviours and other relevant phenomena such as weather affect their performance since usually any travel network can be remark-ably expensive to build and swiftly saturated after its public release. We propose suitable workflow similarity metrics for developing efficient performance measures in rail industry using extensive business process workflow pattern anal-ysis based on Case-based Reasoning. We use meta-heuristic features and extend our similarity measures to capture relevant-to-the-industry granular features and apply this work to an industrial case study. Preliminary results of this work seem promising since they cope with the complexity of the industry and can scale on demand.

Keywords: Case Based Reasoning, Process Mining, Business Process Work-flows, Workflow Monitoring

1 Introduction

Rail transportation industry experiences substantial growth over the last decades in operational method advancement (wayside detectors, wheel profile monitors, extended sensor network), processes, software and hardware equipment (Rail Defect Test Facil-ity, Asset Health Strategic Initiative, and others). The modernisation of the industry has led to increasing usage of computer systems for logistics, tactical, performance and maintenance reasons. Therefore, significant amounts of data have been accumulated and by all involved parties (e.g. private rail companies and government entities). Such data is monitored and, at later stage, analysed with the aim to improve the performance of the industry and the customer satisfaction. Such systems have recently been increas-ingly used to improve several experienced bottlenecks in the industry.

Information Systems (ISs) have been equipped to monitor business process work-flows. ISs can generate significant volume of monitoring data in the form of event logs.

Log files contain information about the execution of workflows. Event logs usually follow a temporal sequence of events and actions also known in the literature as "trans-action logs", "history" or "audit trails" [1]. Based on the extent of the captured data, event logs can include information regarding tasks and their execution time, involved actors and resources associated with tasks and relationships among any tasks. Log files enable workflow process managers to gather detailed information regarding executed process workflows and to be able to understand what is really happening during the execution of a process workflow.

Recently, rail companies have started incorporating audit systems. The Remote Con-dition Monitoring (RCM) is an example to those which allows trains to enrich both the volume and the quality of their information systems and be able to monitor their process workflows. RCM systems consist of multiple sensors attached as/to train components to identify their status and monitor their behaviour. RCM systems were initially used for tracking faulty mechanical components but since they offer continuous monitoring they can fully record any onboard process execution, useful both for engineering and possibly performance measurement and improvement.

In the UK, railway transportation across the country is operated by multiple private organizations usually called "Rail Operators" (ROs). ROs own the trains and manage all services whereas the whole rail network infrastructure is managed by the Network Rail [2]. Each Rail Operator can have as a bespoke business model for their processes or in this case best known as routes. However, any operational model must be approved from Network Rail to form a unified and functional timetable. The unified timetable shows the routes, timeframes and other relevant information that should be followed and respected by all operators. ROs must comply to the timetable timeframes to avoid disruption to other companies' routes and to sustain the desirable level of performance.

Rail industry can experience severe reduction in performance when it comes to un-expected disruptions in service. Such disruptions are experienced by the public as de-lays since a "delay" in service is a well understood term across all relevant stakeholders. Any cause of delay may be attributed to a train malfunction, temporary crew shortage and other reasons. In many cases the reasons behind a delay are difficult to identify, as it may have several contributing factors. For example, ROs can be aware that a specific route is delay prone during peak hours, however the cause of delay may not be easily identified and attributed to tangible causes. Delays should be quantified as whether they took place at station departure, arrival, due to crowd congested platforms, etc. How-ever, delays can have a cascading effect, triggering further delays or cancellations, etc. Domain experts usually have a strong indication of what went wrong and in case of ambiguity refer to Network Rail infrastructure for a deep cause analysis.

To be able to identify the reasons behind delays, we propose process mining [5] techniques based on workflows to domain experts to assist in deviation measurements from scheduled processes (i.e. timetable routes) against the workflows logged by the information system. Such an approach can enable the process managers to identify patterns and possible bottlenecks within workflow processes. To achieve that, work-flow executions should be associated with the expected business process instances (i.e. timetable). However, this has proven to be a complicated task as several bottlenecks exist within the Railway system. 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 layers 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, incom-plete and ultimately faulty data. Through data analysis which has been conducted 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.
- In transportation industry is expected to have many identical 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.
- ROs usually have several “families” of similar trains that may employ several different RCM systems. As a result, several processes can be stored in different datasets which make workflow operations substantially complex.
- Data format can follow several popular or bespoke formats, hardening a universal workflow monitoring approach.

We introduce a multi-level Case-based Reasoning (CBR) approach to achieve work-flow alignment between monitoring data and business processes by considering the railway domain unique characteristics and challenges as described above.

The rest of this paper is organised as follows: Section 2 presents the relevant literature in terms of CBR, Workflows, Process mining and hybrid models, Section 3 formu-lates our proposed methodology for effective process mining in rail workflows, Section 4 shows our preliminary evaluation results and finally we discuss our findings overall and our future research steps.

2 Related Work

2.1 Business Process Workflows

Modern organisations use business process workflows to coordinate their processes, tasks, roles and synchronise their resources with the aim to improve efficiency, efficacy and profitability. Workflows can automate processes, make them more agile and can increase visibility of obscure, erroneous or complex events to company managers to increase productivity [3] [4]. The evolution of technology inside modern organizations and the widely usage of web and semantics allowed the construction and management of workflow based processes. Workflows can be graphically represented using the Business Process Modelling Notation (BPMN), which is the standard notation model

brought by Business Process Management Initiative (BPMI) and Object Management Group (OMG) [5]. Other standards for workflow management and execution have emerged through the past years such as the OASIS Business Process Execution Language (BPEL) [6] and the XML Process Definition Language (XDPL) which allows transferability of business process definitions across different systems and software fa-cilities [7].

Business process workflow management differs across organizations. The size, sec-tor and strategic orientation of an organization play a key role on how they adopt, ana-lyse and practice Business workflows [8]. A common taxonomy includes the phases of: Design, Implementation, Enactment, Monitoring and Evaluation as the workflow life cycle in Business process management [8] [9] [10]. Among those the Monitoring phase enables the supervising of business processes in terms of management (e.g. perfor-mance, accuracy) and organization (e.g. utilization of resources, length of activities etc.) [10]. Therefore, the Monitoring phase is a crucial operation which indicates to process managers and workflow designers what amendments are required to improve their processes.

2.2 Case Based Reasoning

CBR is an approach based on the assumption that: “problems tend to (re)-occur”. Thus, problems occurred in the past tend to re-appear in the future in a similar form. Respec-tively, any solutions that managed to solve previous problems may be recycled to solve currently experienced problems [11].

A requirement for CBR to work is the availability of cases. Cases are usually stored in a Case base along with their associated solutions. Based on this knowledge, CBR can produce a solution for a new problem by following the CBR process cycle defined in [12]. The four main (R) phases of CBR are those of:

Retrieve: This process of cycle is responsible to retrieve the most similar cases which are stored in the case base. The queried cases are measured with past ones using simi-larity metrics

Reuse: The solutions of retrieved cases are being reused to provide solutions for a new case

Revise: The proposed solutions are being evaluated

Retain: This process provides the option to update the case base with the experience extracted from the new case

2.3 Process Mining

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 [13] [14]. By applying process mining, domain experts can use the derived information as feedback to design new processes or revise and enact predefined ones [15].

Three main types of process mining can be identified and summarised as:

PM-Discovery: used when a predesigned process does not exist. In this case the event logs can be used to produce a model.

PM-Conformance checking: Aims to compare event log processes with process models. Conformance checking can detect and highlight any possible deviations among process models and their execution.

PM-Enhancement: This type is required when trying to improve an existing process model. The event logs can demonstrate that other perspectives could enrich the model activities. Such perspectives can be information regarding time, resources and actors.

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 [24] [25]. 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, therefore a CBR approach may be more appropriate.

2.4 Workflow Monitoring and CBR

The literature shows several related researches attempts to address problems around workflows. Van der Aalst et al. [16] proposed an approach that compares process mod-els. This approach shows how the degree of similarity between process models can be measured. Also, it is being considered the fact that distinct parts of a process might have "stronger" notion than others. The results are presented on a Petri nets structure.

Dijkman et al. [17] attempts to rank business process models according to their sim-ilarities. Four distinct types of graph matching algorithms were compared to solve the similarity search problem. The produced results by the algorithms were based on a trade-off between computational complexity inherited from graph matching and the comparison accuracy. Weber et al. [18] presents a tool that is based on conversational case-based reasoning which complements an adaptive workflow management system. The tool provides knowledge to a management system and enables the adaptiveness of predefined workflow models based on confronted circumstances. Workflow manage-ment systems produce more accurate results over time since it builds experience on the knowledge gained previously. Minor et al. [19] presented a case-based reasoning ap-proach that allows the reuse of previous adaptations of workflow instances on the on-going ones.

We used a graph based system to retrieve previous cases of adaptations for each part of the workflow structure. Therefore, previous modification that occurred on a similar case can be evaluated to be applied again. Kapetanakis et al. [20] [21] provided expla-nations to the intelligent monitoring of business process workflows. This approach showed how a similarity measure between workflow instances can be establish consid-ering intervals and temporal relationships using CBR. The fundamental assumption in this approach is that a workflow structure is not met during execution. Therefore, the workflow instances are identical but not same. Consequently, workflow instances

marked as problematic, that seem to be similar with other instances, they probably share the same problems and require similar solutions.

3 A CBR approach for aligning workflow executions

CBR has been shown effective in monitoring workflow instances under uncertainty [22] [21]. Utilising CBR's fundamental principle of "similar problems" have usually "similar solutions" we investigated several rail data instances to model route cases appropriately.

CBR retrieves past solutions from a case-base matching workflow instances to route-processes. In our industrial scenarios a workflow is a route stored in event log sequences and a business process is the scheduled route as planned and showed on a public time-table. In our CBR model, we treat routes as cases and their related business processes as solutions for those cases. Based on temporal and spatial data our case representation is formulated as in Figure 1.

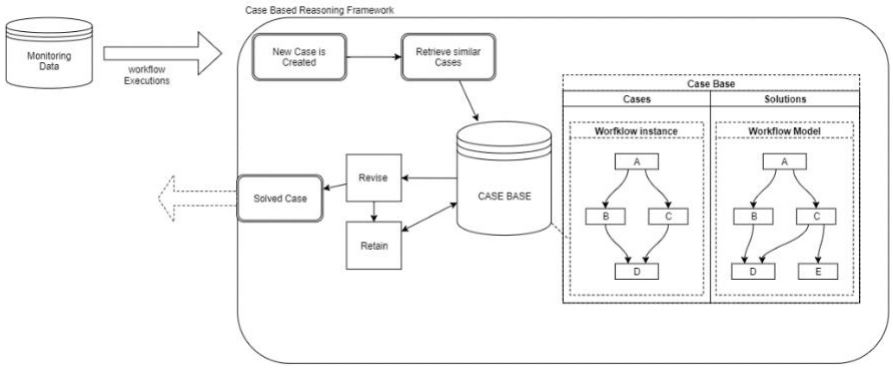

Fig 1. Case representation as generated from temporal/ spatial workflow data

Delay patterns are often related to location and time e.g. rail platforms during peak hours can be overcrowded and this may lead to delays. Another common example are busy junctions during certain hours causing overheads to any related services. Therefore, it is presumed that same or similar routes share similar bottlenecks (delays).

This section will present our case representation, the formulated similarity metrics for our investigated domain and their formal relevance to temporal logic.

3.1 Case 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

When CBR is adopted to provide solutions, new cases are created enclosing workflow data within cases. Therefore, a new query case will have the following structure:

{ UnitNumber_q, $\textit{StartDay}_q$, JourneyTimeq, StartTime_q, EndTime_q, $\textit{StationList}_q$, $\textit{Activ-ityList}_q$, }

And for each activity:

{StationName_q, StopDuration_q, NextStation_q, $\text{TimeUntilNextStation}_q$}

3.2 General Time Theory

Our workflow data follow a sequential temporal vs. spatial pattern since they represent a variety of activities (as presented in Section 3.1) over time. To represent their se-quence in a formal way we use the General Time Theory (GTT) [23] as it can be seen in Figure 2. 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

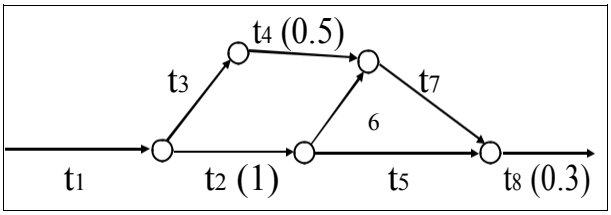

Fig 2. Graphical representation of a log temporal inference using the GTT

In our 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 rep-resentation (Section 3.3) presents a foundation for adequate CBR workflow cases.

3.3 **Similarity Functions**

We define a set of multi – level similarity functions relevant to the problem domain. Since elements of temporal information are present throughout a log journey, a GTT representation as shown in Section 3.2 allows for vectorised workflow mapping. Simi-larity measures are split into two levels (Level 1 and Level 2) based on the workflow structure.

Level 1: Identifies 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 sta-tions. For C_1 and C_2 if Start date is the same (Binary equal) && Start time relies within mins fluctuation && C_{1L} is like $C_{2L'}$ based on an μ string threshold.

$$(1, 2)= | \quad 1 = \quad 2\leq \gamma |* \quad 1 +$$

(equation 1)

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

(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:

$$(1, 2) = | [1, \quad 1, \quad 2, 2] \ 1 - [\quad 1, \quad 1, \quad 2, 2] \ 2 | * \ 1 + | \qquad 1= \qquad 2|* 2+ | \quad 1= \quad 2| *$$

w_3

(equation 4)

Where UN_1 and UN_2 are actor related identification numbers

4 Experiments & Results

Our case representation, as presented in Section 3, allows for a rigid problem definition. A key challenge presented from the application domain is the lack of "solutions" due to the following reasons:

1. Constant changes at business process level. A Rail timetable changes every 6 months (seasonal). Variations in any normal operation can vary among a week before the actual service, a few days in advance, or even hours. In several cases any of the above amendments could be chained e.g. a disruptive change a few days between the DTR – LBG route may also be changed hours before the actual service. This raised significantly the fuzziness within the data.
2. Incomplete data. Our provided data although very rich in volume had substan-tial degree of repetition and severe incompleteness at cases.
3. Our data corpus was coming from variant datasets that posed heavy uncer-tainty due to their technical compatibility and inconsistencies.

To overcome the above challenges, we had several sessions with senior business process experts, analysts and industry engineers that elaborated extensively in several cases vs. right solution matches. With their help a case base was formulated containing several past workflow executions and their solutions as corresponding process models.

The CBR cycle was modified in the following way to suit the domain:

1. Retrieve: Similarity functions were defined based on the 2-level similarity model (section 3) based on multiple perspectives (such as time, resources, or-der flow, relations between activities, etc.)
2. Revise: when a queried case couldn't converge in finding a similar case with > 60% relevance, the case was tagged as a "newly" encountered pattern. A low similarity score on cases indicated an incomplete or an "just in time" (JIT) amended services which had no previous process model.
3. Retain: Indicated no modifications to existing cases. After every "new" case encounter it updated the case base and pushed any case with similarity lower than 60% to a new case base for further investigation.

For our evaluation we used an (hourly) data sample of one 238MB of workflow data, achieving a performance accuracy of 76% on cases vs. ranked business process from industry experts with a 10-fold validation on 30%(test)-70%(training) split.

Motivated by the success of the initial experiment we used the trained case-base on substantially larger dataset of 480GB which was a seasonal corpus (3 months) of work-flow data. The ranked case-base performed adequately to similar routes however a sub-stantial number of "newly" seen cases emerged which could not be adequately at-tributed.

To benchmark our approach a probabilistic approach was adopted, comparing any new investigated case with possible nearest neighbour "paths". An example may be appropriate to explain the concept: Let's assume an imaginary path of letter-labelled stations: A, B, C, D, E and F. Our case base may contain ABC and ACF. If ABDE comes in the 2-level similarity cannot produce convincing results whereas a

probabilistic approach can rank ABDE as an AB variant with 50% probability and ignore ACT since the probability to fit, there is less than 33%.

The probabilistic approach seemed to work better than CBR on the large dataset due to sheer volume characteristics which were very hard to address. On a variant experi-ment using 2 different datasets (one ranked by experts and the other unknown) our methodology did achieve similar performance results: 70% accuracy on the ranked da-taset and substantially lower (not decent enough to present) to the other. By applying a similar probabilistic approach as to the one above the gained results seemed better com-pared to a CBR approach.

5 Conclusions

This work presents transport friendly approach to break down the complexity of temporal spatial data and attempt to identify workflow patterns and trends over time. We propose a new multi-level similarity approach that can elicit meta-heuristic features and can assist in capturing relevant-granular features. We presented some preliminary re-sults from our work on real industry case study, although our results were affected from the investigated dataset(s) bias, limited ranking and very large volume and variety. In our future work we plan to improve substantially our model towards automatic detec-tion of workflow differences, mine patterns efficiently and work on ways to tackle large data volumes and dataset discrepancies. We will also work on establishing the right benchmark tools to enhance the accuracy, precision and recall of our proposed meth-odology.

References

1. Reijers, H. A., Weijters, A. J. M. M., Dongen, B. F. V., Medeiros, A. K. A. D., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. Information Systems, 32.
2. Network Rail, https://www.networkrail.co.uk/who-we-are/about-us/, last accessed 28/10/2017.
3. Workflow Management Coalition. Workflow management coalition glossary &terminology. http://www.wfmc.org/standars/docs/ TC-1011_term_glossary_v3.pdfS, 1999.
4. Accountants, I. o. M. (2000). Implementing Automated Workflow Management. USA, Institute of Management Accountants 10 Paragon Drive Montvale, NJ 07645.
5. Business Process Management Initiative (BPMI): BPMN 1.1: OMG Specification, February 2008, http://www.bpmn.org/, accessed April 2017.
6. OASIS: BPEL, The Web Services Business Process Execution Language Version 2.0. http://www.oasis-open.org/apps/org/workgroup/wsbpel, May 2006.
7. Workflow Management Coalition (WfMC): XPDL 2.1 Complete Specification (Updated Oct 10, 2008), http://www.wfmc.org/xpdl.html.
8. Van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M.: 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 (2003)

9. Zur Muehlen, M.: Workflow-Based Process Controlling: Foundation, Design and Application of Workflow-driven Process Information Systems. Logos (2004)
10. Reijers, H.A.: Design and Control of Workflow Processes: Business Process Management for the Service Industry. Springer, Heidelberg (2003)
11. Leake, D. (1997). Case Based Reasoning. Experiences, Lessons and Future Directions. AAAI Press. MIT Press, USA, 1997.
12. Aamodt, A., Plaza, E. (1994) Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(i), 1994.
13. Van der Aalst. (2011) Process Mining: Discovery, Conformance and Enhancement of Busi-ness Processes. Springer-Verlag, Berlin, 2011.
14. Van der Aalst, W. M. P., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., & Weijters, A. J. M. M. (2003). Workflow mining: A survey of issues and approaches. Data & Knowledge Engineering, 47, 237–267.
15. Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer Verlag, Berlin (ISBN 978-3-642-19344-6).
16. Van der Aalst, W., Alves de Medeiros, A. K., Weijters, A : Process Equivalence: Comparing two Process Models Based on Observed Behavior, In Proc. Of BPM 2006, vol 4102 of LNCS, pp 129-144, Springer, (2006)
17. Dijkman, R.M., Dumas, M., Garcia-Banuelos, L. Graph matching algorithms for business process model similarity search. In U. Dayal, J. Eder (Eds.), Proc. of the 7th Int. conference on business process management. (LNCS, Vol. 5701, pp. 48-63). Berlin: Springer. (2009)
18. Weber, B., Wild, W. and Breu, R. "CBRFlow: Enabling Adaptive Workflow Management Through Conversational Case-Based Reasoning", in Proceedings of ECCBR04, Advances in Case-Based Reasoning, LNCS, Vol. 3155, 434-448, Springer (2004)
19. Minor, M., Tartakovski, A. and Bergmann, R.: Representation and Structure-Based Similar-ity Assessment for Agile Workflows, in Weber, R., O. and Richter, M., M.(Eds) CBR Re-search and Development, Proceedings of the 7th international conference on Case-Based Reasoning, ICCBR 2007, Belfast, NI, UK, August 2007, LNAI 4626, pp 224-238, Springer-Verlag, (2007)
20. Kapetanakis, S., Petridis, Ma, J., Bacon, L.: 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 Workshop on expla-nation- Aware Computing Exact (ECAI 2010), Lisbon, Portugal (2010)
21. Kapetanakis, S.,Petridis, M.,Knight, B.,Ma, J.,Bacon,L. : A Case Based Reasoning Approach for the Monitoring of Business Workflows, 18th International Conference on Case-Based Reasoning, ICCBR 2010, Alessandria, Italy, LNAI (2010)
22. Kapetanakis, S., Petridis, M.: 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. 2014, Springer Berlin Heidelberg. p. 43-54.
23. Ma, J., Knight, B.: A General Temporal Theory, the Computer Journal, 37(2), 114-123 (1994).
24. 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
25. Van der Aalst, W. M. P., de Medeiros, A. K. A., & Weijters, A. J. M. M. (2005). Genetic process mining. Applications and Theory of Petri Nets.