ASM Sc. J., 19, 2024

https://doi.org/10.32802/asmscj.2023.1452

Exploring Event Data Pre-processing Approaches for Business Process Mining

Z. Lamghari*

[1]*Department of Computer Science, Faculty of Sciences, Mohammed V University, Rabat, 10000, Morocco*

[2]*Laboratory of Sciences, Engineering, and Management (LSEM), High School of Technology (EST), Sidi Mohamed Ben Abdellah University, Immouzer Road, Fez, 2427, Morocco*

Process mining empowers companies to optimise operational processes by deriving insights from event logs. It involves assessing event logs or the resultant process models against existing models and identifying challenges within executed processes to enhance their efficiency. However, a critical prerequisite for effective process mining is data cleaning, which simplifies the complexities inherent in real-world event data, making it amenable to comprehension, processing, and leveraging with process mining techniques. Consequently, new methodologies and approaches for event data pre-processing have garnered significant attention within the business process scientific community. This paper presents a systematic literature review summarising relevant techniques for event data pre-processing in the context of business process mining. The objective of this study is to categorise approaches and methodologies related to event data pre-processing while shedding light on the crucial issues associated with these techniques.

Keywords: process mining; filtering technique; clustering; pre-processing; event log quality

I. INTRODUCTION

Process mining (PMg) is a research subject that has sparked considerable interest in the computer science and Business Process (BP) modelling sectors (Maddah & Roghanian, 2021). It is a strong instrument for companies to acquire genuine models for simply understanding how their BPs operate and improving their decision making. PMg approaches enable the autonomous recognition, compliance, and refinement of process models deployed by different companies using knowledge retrieved from event logs in addition to the process model's relevant documentation (Durojaiye *et al.*, 2022). An event log, in this context, is a bundle of time-stamped event data generated when the BP is being operational (see Figure 1).

The quality of event data has a noteworthy influence on the resultant model. Thus, a low-quality event log (missing, erroneous or noisy numbers, duplicates, etc.) might result in a complicated, poorly structured, and complicated model; or a model that does not accurately consider the true BP's behaviour. Consequently, event data preparation is regarded as a task that may significantly increase PMg performance.

According to (Cappiello *et al.,* 2022), the quality of event data and processing durations might severely restrict PMg operations. This has increased the importance of pre-processing procedures.

Mans *et al.* (2012) define the log quality as a vision of double dimensions. The first dimension addresses the abstraction level of activities included in the process model. The second focuses on the varied precision levels and validity of the timestamps recorded in the event data.

Emamjome *et al.* (2022) tackle the data quality from a philosophical perspective, addressing the core reasons as well as the political, physical, and interpersonal aspects that result a data of mediocre quality. These factors are missed by available data cleaning approaches. Other authors (Scannapieco, 2006; Wand & Wang, 1996) define data

*Corresponding author's e-mail: zineb_lamghari2@um5.ac.ma

quality as a holistic paradigm that includes precision, reliability, thoroughness, clarity, interpretability, and responsiveness. In this study, we correlate event log quality to the detection, adjustment, or elimination of faulty or noisy events, incomplete, redundant, and inappropriate events. In real-world circumstances, several PMg activities depend on the premise that behaviour linked to the current process's execution is appropriately recorded in the vent log. Also, every saved instance related to the process in the event log is actually completed. Event logs, on the other hand, contain chaotic or incorrect data. It can be caused by a number of different circumstances: cases are redundant, insufficient, incorrect, or represent inappropriate actions. These issues are compounded by a variety of circumstances, including data transmission mistakes, storage faults, technological limits, or transcript inaccuracies when events come in the incorrect sequence. Also, noise is related also with occurrence of unusual events as a result of handling of exceptional circumstances, erroneous recording of chosen tasks during process execution, or even inaccurate interpretation of timestamps. It is hard to discriminate between noise and infrequently accurate activities in an event log, leading in a mining model with lower accuracy to the true model.

Numerous forms of noisy event data have a detrimental influence on quality issues, causing the PMg algorithm to deliver complicated, confusing, or even incorrect results. As a result, event log preparation is a required activity in many PMg methods to mitigate these detrimental impacts. The event log pre-processing aims to discover and eliminate noisy events, traces, or activities including such undesirable behaviour characteristics.

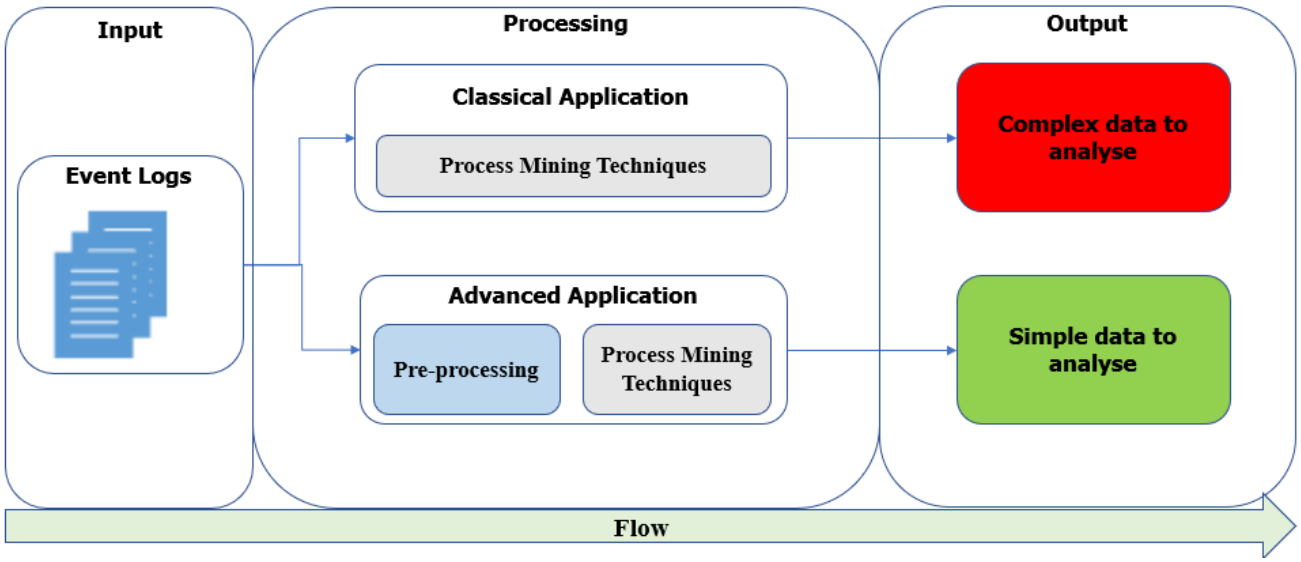


Figure 1. Process Mining Classical Application vs. Advanced Application

This scientific paper presents a thorough examination of the most typical methods for pre-processing event data, which is critical for the accuracy of PMg activities. First, we give an innovative review of event log pre-processing techniques such as heuristics (Tax *et al.*, 2019), visualisation algorithms, and integrative approaches. Second, we give a thorough qualitative and quantitative analysis of the works assessed.

In general, our study gives broad responses to three major challenges: 1. How can the various event log pre-processing approaches be classified? 2. What are the issues with establishing data quality? 3. How can we guarantee if a pre-processing strategy can significantly enhance data quality?

For instance, grouping and identifying matched issues with pre-processing approaches can help PMg implementers understand the numerous ways accessible and offer them additional components to choose the best suited methodology using suitable algorithms, regarding the matched algorithms, the type of quality concerns tackled, or specific challenges in the application area. This study makes three major improvements:

1. First, we give a review of event log pre-processing approaches, commonly known as data cleaning or data preparation strategies.
2. Second, we give a collection of event log pre-processing and repair approaches needed to develop more robust process models.
3. Last, we give research of significant variables related with pre-processing approaches that are considered when deciding whether or not to adopt a certain technique.

The reminder of this paper is organised as follows:

Section 2 presents the fundamental ideas of event log pre-processing and PMg. Section 3 describes the research technique used to create this survey. Also, this section proposes grouping event log pre-processing procedures based on ways documented in the state-of-the-art. Moreover, section 3 describes the tools that were used in illustrated case studies. Furthermore, Section 3.4 describes the representation techniques used for event log processing and transformation. Section 3.5 describes the many issues discovered in the event logs. Section 3.6 goes through the duties that are connected to pre-processing. Section 3.7 describes attributes and types that may be used to improve the event log's quality. Finally, Section 4 summarises this work.

II. PRELIMINARIES

As PMg algorithms operate on an event log, which is a collection of historical records from each BP instance. A trace relates to each event created during the execution of a BP instance (a case). The collection of all traces that correspond to the event log. This section introduces fundamentals for comprehending the basics of event log pre-processing.

GKW1 = (“filtering” OR “cleaning” OR “repairing” OR “clustering” OR “refinement” OR "preprocessing") “event log”
GKW2 = (“filtering” OR “cleaning” OR “repairing” OR “clustering” OR “refinement” OR "preprocessing") “trace”
GKW3 = (“ordered” OR “aligning”) “event log”
GKW4 = (“anomalous detection” OR “infrequent behavior” OR “noisy” OR “imperfection”) “event log”
GKW5 = (“anomalous detection” OR “infrequent behavior” OR “noisy” OR “imperfection”) “trace”

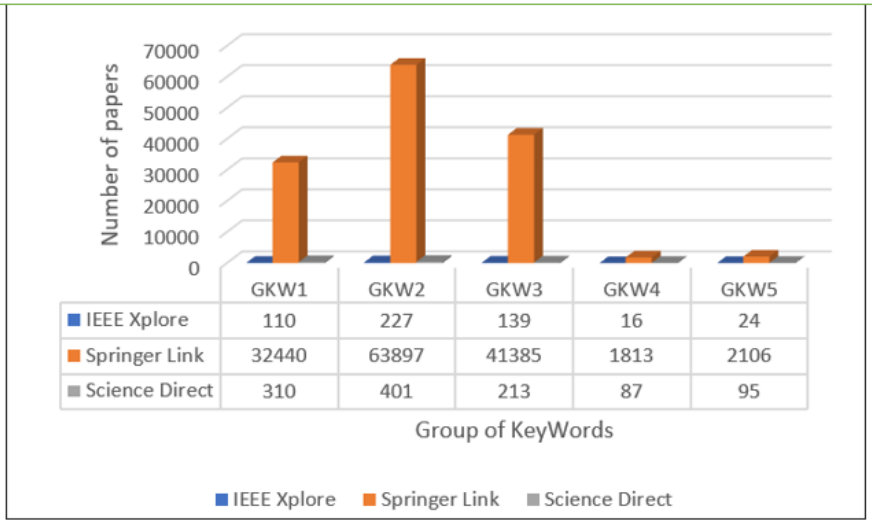


Figure 2. Data source

Postulate one. An event is defined as a case, an activity, and a moment in time. The event is defined by a series of attributes such as ID, timestamp, cost, and resource (Van der Aalst, 2016).

Postulate two. A trace may be thought of as a case, that is, a limited number of event logs $\boldsymbol{\sigma} \in \boldsymbol{E}^{*}$in which each event is recognised only once (Van der Aalst, 2016).

Postulate three. An event log is made up of a series of cases, and each case is made up of events. In the whole log, each occurrence occurs just once (Van der Aalst, 2016).

A trace, or a sequence of distinct occurrences, is used to represent the events in a case. Cases, like events, can also have qualities. An event log's structure consists of:

• An event log is made up of cases.

• A case is made up of occurrences, each of which is related to just one case.

• The events in a case are ordered.

• Attributes can be assigned to events: activity name, time, expenses, and resource.

Postulate four. A BP model is a graphical and analytical depiction of the business operations of a company.

Several visual approaches or notation languages, such as UML, Petri nets, and BPMN, are commonly used to describe a BP model. In the context of event log pre-processing, it is critical to detect difficulties that are intimately allied to the quality of the data gathered in the event log. As a result, some of the most common data quality challenges in event logs are covered in this section.

The noise problem refers to the scenario, in which the event log data contains mistakes or nonsensical data that deviates from the intended behaviour. Incorrect or noisy data can be caused by inconsistency or disparity in naming standards or data codes used, or by inconsistent formats for input fields such as timestamps. As a result, several strategies must be used to delete or replace the noisy data.

Moreover, the missing data issue might arise when various pieces of information are missing from the event log, even though it should be recorded on a regular basis. This happens, for instance, when an attribute in an event is missing owing to issues with transmitting, registering, or storing events from an information system.

In the situation of irrelevant data, there may be event records that are unrelated to the model under experiment's study, but they might be to obtain a record of a significant occurrence using transformation and filtering procedures.

However, redundant data happens when the same event is recorded in the event log many times, by the same resource, and on the same timestamp. Similarly, when an action is recognised more than once by the information system, the problem might occur, leading the process model to grow complicated (Chen *et al.*, 2022).

Data diversity occurs when an information system is highly generic and enables for events with varied precision levels, rendering process models confusing and difficult to depict. Many of the previously described data quality concerns have been treated in the examined preparation strategies in this study.

III. OUR PROPOSED METHODOLOGY

This section discusses the methodology for the literature review offered in this article, as well as the inclusion and exclusion criteria used to choose the publications examined.

A. Meta-analysis

The technique for searching and selecting research articles for this review was carried out in two steps. The first stage involved retrieving relevant publications from four prominent digital sources, such as Science Direct, IEEE Xplore, Springer Link, and Google Scholar, focusing on diverse areas such as PMg. We specifically collected papers published since 2005 (the year that automatic algorithms for mining processes, such as the alpha algorithm, were first proposed) that contained in their title or abstract the following terms: fine-tuning, mending, cleaning, refinement, filtering, clustering, pre-processing, ordered, aligning, abstraction, anomaly detection, rare behaviour, noisy, inaccuracy, traces, event log, process mining.

These keywords have been merged to create expressions, which are used as input queries to the four electronic sources. Considering the diversity of the scientific papers obtained from the designated electronic sources, a search and selection method were used in the inclusion or exclusion criteria in the second stage to choose which scientific papers would considered for inclusion in the final assessment. Indeed, we gathered all prior studies using the

following keywords combined with the "process mining" keyword (see Figure 2).

1. Qualification

The scientific papers evaluated and tackled in this study are chosen according to the following qualifications:

1. English language.
2. Published research papers in dissertations (theses) and international conferences and journals.
3. Scientific papers that have been released since 2006.
4. Scientific papers expressly focusing on event log pre-processing or cleaning.

2. Elimination

The scientific papers that are not relevant to this summary are as described as follows:

1. Scientific papers that is unrelated to PMg.
2. Scientific papers in the field of PMg that do not specialise in specific industries.
3. Scientific papers excluding assessment and experimental outcomes.

Following a topic-specific filter, i.e., topic of interest, delete duplicated articles, a total of 180 scientific papers were collected and examined in accordance with the inclusion criteria. Also, by applying the elimination conditions to this set, we obtained an outcome of seventy publications. All of these are taken into account in our subjective examination.

Figure 3 displays the qualified scientific papers from 2006 to 2022, according to the year of publication. Moreover, figure 3 shows a modest rising trend, it is impossible to predict if that trend will continue.

Figure 4 shows that 30% of the selected scientific papers were published in international journals, 72% were observed for conferences, and the remainder were dissertations. Previously reported works (2006–2008) had a significant impact on the community, as seen by Figure 5, which shows a total of 2440 citations. Furthermore, the most recent works (2020–2022) exhibit almost 104 citations.

In this study, we depict a network of phrases that are intricately connected and explain the many themes handled by event log pre-processing. These phrases are classified by importance:

- Pre-processing
  - Tools: ProM, RapidMiner, TimeCleaner, Apromore.
  - Structure: Sequence, Graph, Automation.
  - Manually: Incomplete, Infrequent, Missing, Outlier and Duplicated event log.
  - Automatic: Incomplete, Infrequent, Missing, Outlier and Duplicated event log.
  - Techniques
    - ✓ Embedded
    - ✓ Classification: Supervised Learning, Classifier Rules, Bayesian.
    - ✓ Filtering: chaotic, event, activity, trace,
    - ✓ Rules
    - ✓ Timestamp information: Order Anomaly, varied precision Level.
    - ✓ Clustering: Log Profiles, K-gram Model, Agglomerative
    - ✓ Alignment
    - ✓ Patterns
    - ✓ Metrics: Distance by Levenshtein, Distance by Euclidean, Distance by Generic

This classification might help scientists comprehend and organise the wide range of existing pre-processing approaches.

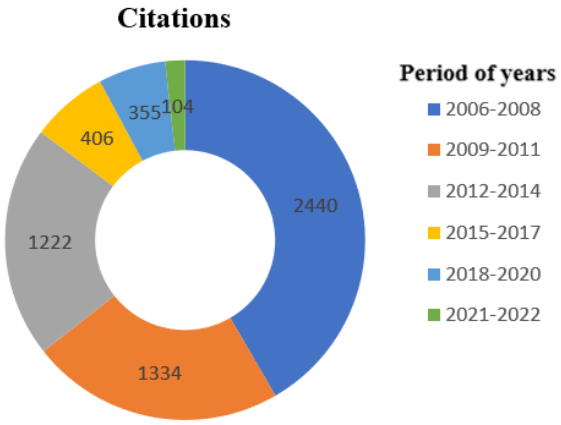


Figure 3. Periodical publication of scientific papers

Content research was carried out as part of the literature evaluation. We recognised and categorised the common and prominent features observed in the examined publications in this study. The following Terms are considered as the

most prominent features related to the PMg discipline (F1-techniques to transform, detect and visualise. F2-tools to realise PMg techniques, we focus here on ProM, Disco, RapidProM, Celonis, Apromore and RapidMiner. F3-representation schemes using standard representation, F4-imperfection kinds ignore non-useful behaviour, F5-related tasks to abstract and align, and F6-information types represent event logs attributes as timestamp, activity name, ID, etc.

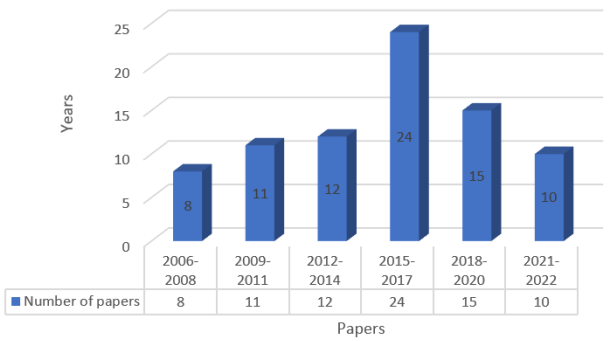


Figure 5. Number of citations for each of the surveyed publications between 2006 and 2022

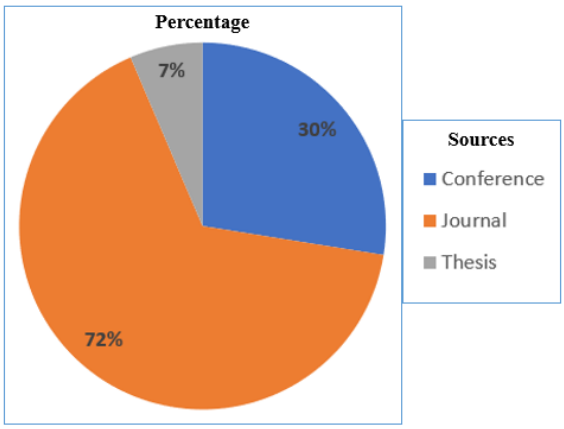


Figure 4. Scientific papers related to the dissimilar sources included in this survey

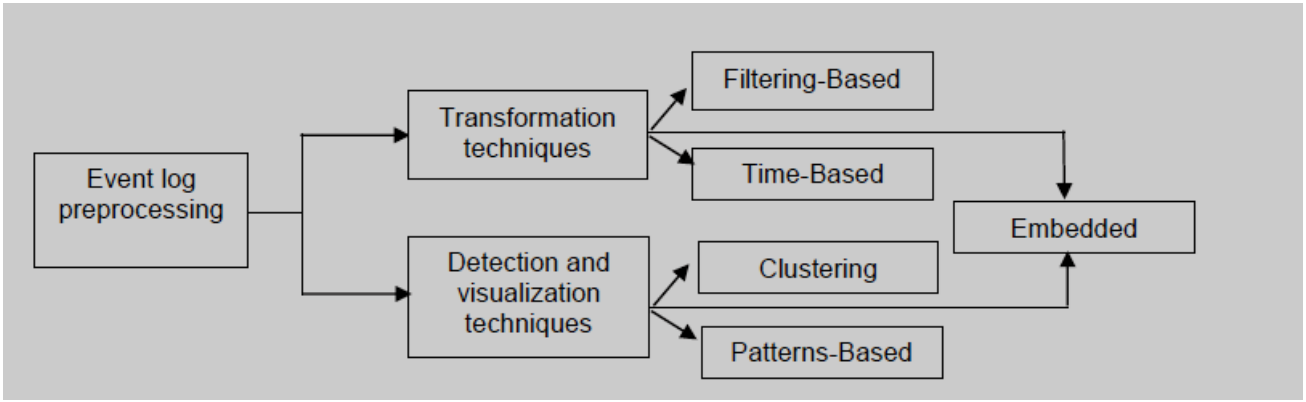


Figure 6. Two main groups of data pre-processing related to process mining

B. F1 Techniques

Related to the PMg discipline, different criteria may result in diverse taxonomies of data pre-processing approaches. We divide the existing event log pre-processing approaches into two categories based on the works reviewed: transformation techniques and detection-visualisation techniques. The strategy used by pre-processing techniques to filter the data from deficiencies that encompasses error identification, isolation, and correction, is the major classification criterion. Figure 6 depicts a suggested taxonomy for the works assessed. The suggested taxonomy organises the diversity of current pre-processing approaches and aids in the identification of traits that they may share. Our categorisation also aids to determine which data quality concerns require the usage of certain approaches.

The first category includes strategies that use event log modifications to fix poor behaviours (missing, useless, redundant data, etc) before applying a PMg algorithm. The second category includes strategies for detecting or diagnosing flaws in an event log. While the second group of approaches just detects foreseeable problems with data quality in the event log, the first category of techniques directly corrects the defects discovered in the event log.

C. F2 Tools

Tools for event log preparation are frequently incorporated as part of PMg tools. Also, several PMg tools (Perceptive PMg by Lexmark (Workflow and Case Management, 2009; Van Cruchten and Weigand, 2022), Interstage BP Manager Analytics by Fujitsu Ltd. 2009, Minit by Gradient ECM. 2015, myInvenio by Cognitive Technology (Massimiliano, 2016), and others do not support event log pre-processing tasks that support event log quality improvement. There are specific tools, apps, or frameworks designed for certain event log pre-processing tasks (Evermann *et al.*, 2016; Kong *et al.*, 2019; Xu and Liu, 2019; Folino *et al.*, 2011; Ghionna *et al.*, 2008; Chapela-Campa *et al.*, 2017; Lu *et al.*, 2016).

The majority of these solutions are restricted to a single process modelling language and include some sort of data deployment or modification. Furthermore, specialist tools such as ProM (Van Dongen *et al.*, 2005), Apromore (T.A. Foundation, 2011), Celonis (S.E, 2009), and RapidProm (Mans *et al.*, 2014) contain various filters, routines, and algorithms for pre-processing the event log in order to enable PMg activities.

Will van der Aalst (2016) distinguishes different distribution of PMg tools that include event log pre-processing. The first distribution defines PMg tools, which are designed primarily to address ad hoc inquiries concerning event log pre-processing. Disco (BV. F, 2011) is an example of this sort of technology. It allows the user to group objects dynamically and immediately project that data on a freshly acquired process model.

The analytic approach is made clear in the second distribution of PMg tools; that is, the user may view and determine which items to isolate or exclude from the event log. RapidProM is an example of this sort of tool. Finally, the third distribution of PMg tools are designed to answer specified questions repeatedly in a given situation. These technologies are commonly used to generate "process dashboards," which are standard views of process models. Celonis PMg, for example, facilitates the building of such process-centric dashboards.

Following that, we will go through different tools that feature pre-processing or event log repair procedures as part of their functionality. These tools were chosen based on

their popularity in the PMg field (as stated in multiple articles) and the presence of pre-processing techniques.

For cleaning event logs, the ProM framework (Van Dongen *et al.*, 2005) includes many events log filters (Filter event log using: selection, attribute value, basic heuristics, high-frequency trace). These filters are extremely useful when looking at real-world logs since they allow for not just visualising data in the log, but also modify data by adding information to the log, eliminating process instances (cases), avoiding and changing events. There are available plug-ins in ProM that can be used to delete or repair activities, attributes, and events (for example, delete unused activities, delete value-empty attributes, delete events without timestamps, and refine labels globally).

Because many of the research suggestions are available through ProM, it is the most common PMg tool that generally contains pre-processing approaches. The majority of existing pre-processing approaches, however, are mainly used event filtering and trace grouping. ProM supports a wide range of languages and formats, including Petri nets, BPMN, EPCs, social networks, and others. A broad variety of models, from a Petri net to LTL formulae, may be imported via plug-in import.

The ProM framework supports interaction among plug-ins, which include deployment of algorithms and declarative techniques for BP analysis, PMg, organisational mining, social network analysis, clustering, decision mining, prediction, and recommendation.

Apromore (TA Foundation, 2011) is an open-source framework for sophisticated BP modelling. It enables the use of a range of filtering algorithms to reduce and slice an event log. Apromore supports case filters and event filters. These techniques enable users to create a filter based on certain requirements on cases or occurrences. A case filter enables users to slice a log, retaining just a portion of the process cases. An event filter enables slicing a log or retaining a portion of the process across numerous situations. Other criteria, such as timeline, allow you to keep or remove instances chronologically. Another filtering approach is the rework repetition filter, which may be used to eliminate process sequences that include particular repeats.

RapidProM (Van der Aalst & van Zelst, 2017) is a RapidMiner expansion in which the PMg framework ProM is incorporated into RapidMiner to represent a hybrid tool. Complex PMg processes may be designed, performed, and reused for different data sets with RapidProM. This tool contains data cleaning and filtering methods for filtering cases using throughput time, with the option of selecting a different performance annotation.

The RapidProM is responsible on the examination of event data and the development of process models.

1. Transformation

Transformation techniques perform operations and actions to record changes in the raw event log's original structure to enhance the log's quality. There are two basic approaches within this group: filtering and time-based procedures. On the one hand, filtering approaches seek to predict the recurrence of events or traces based on their surroundings. The events or traces that occur seldom are eliminated from the first event log. Filtering approaches are designed to remove logging errors and avoid propagating to process models. The goal of time-based approaches, on the other hand, is to preserve and adjust the sequence in which the events were recorded in the log based on the timestamp information.

Filtering methods are primarily concerned with the detection and removal of noisy events or traces with missing values. Their key distinguishing feature is the filtering of abnormal behaviour found in the event log that may impact the execution of next PMg jobs. These strategies mimic the contexts of actions that occur frequently and filter away the contexts of occurrences that occur seldom in the log.

Several publications (Dixit *et al.*, 2018; Fani Sani *et al.*, 2018a; Fani Sani *et al.*, 2018b; Lamghari *et al.*, 2022; Neerumalla & Parvathy, 2022; Rai *et al.*, 2022; Vathy-Fogarassy *et al.*, 2022; Wang *et al.*, 2013; Zelst *et al.*, 2018) have been published in the literature that suggest the development of filtering methods. Conforti *et al.* (2016) described a strategy for identifying abnormalities in a log automaton. In this context, as an automaton, the approach creates an abstraction of the process activity recorded in the log (a directed graph). This automaton records the event

log's direct follow dependencies. Then, infrequent transitions are deleted from the log using an alignment-based replay approach while reducing the number of events eliminated.

van Zelst *et al.* (2018) suggested an online or real-time event stream filter for detecting and removing erroneous events from event streams. The key notion behind this technique is that dominant behaviour has a greater occurrence probability within the automaton than spurious activity. This filter was created as a free and open-source plugin for the ProM (Van Dongen *et al.*, 2005) and RapidProM (Van der Aalst *et al.*, 2017) applications.

Wang *et al.* (2013) provided research of approaches for recovering lost events, resulting in a more comprehensive list of possibilities. The authors employed backtracking to eliminate the duplicate sequences associated with simultaneous occurrences. Then, a branching architecture was added, in which each branch may directly apply the backtracking. To further speed the calculation directly, the authors created a branching index and established reachability testing and lower bounds of recovery distances. Lamghari *et al.* (2022) developed techniques for recognising chaotic activities, which are defined as activities that do not have partial fixed positions in the process model's event sequence with a fixed chance of occurrence (or changes little) because of the occurrence of other activities.

Dixit *et al.* (2018), Fani Sani *et al.* (2018a) and Fani Sani *et al.* (2018b) employed trace sequences as a framework for handling the event log in pre-processing techniques based on event-level filtering. In majority of these studies, this structure enables the sorting and measuring the probability of event occurrence for detecting noisy event log.

Other publications, such as (Bezerra & Wainer, 2008; Bezerra & Wainer, 2012; Bezerra & Wainer, 2013; Jalali & Baraani, 2010), offer methods for detecting and eliminating abnormal traces of process-aware systems, where an erroneous trace is defined as an event log trace with a conformance value less than a criteria given as input to the algorithm Anomaly traces, if found, must be assessed to determine whether they are incorrect executions or suitable but infrequent executions.

Cheng and Kumar (2015) planned to train a classifier on a small segment of the log and then use the classifier rules to eliminate noisy traces from the log to produce inaccurate logs from standard process models, and to mine process models by implementing PMg algorithms to both the noisy log and the cleaned version of the same log, and then evaluating the discovered models by the guideline model. The next suggestion involved using structural and behavioural parameters to compare the models derived before and after cleaning the log.

Sani *et al.* (2017) suggested a filtering method based on conditional probabilities between activity sequences. Their method calculates the conditional likelihood of an action occurring depending on the number of preceding activities. If this likelihood is less than a certain level, the behaviour is termed an outlier. Outliers were defined by the authors as both noise and uncommon behaviour (Fox *et al.*, 2022). They also employed a conditional occurrence probability matrix (COP-Matrix) to store dependencies between present and already happened actions at greater distances. These papers (Bezerra & Wainer, 2009; Bezerra & Wainer, 2012; Böhmer & Rinderle-Ma, 2016; Cheng & Kumar, 2015; Jalali & Beraani, 2010; Sani *et al.*, 2019) provide more strategies for filtering aberrant events or traces.

Other sorts of transformation techniques for data preparation in event logs include time-based strategies. Many studies on event log preparation have concentrated on data quality challenges linked to timestamp information and its implications for PMg (Dixit *et al.*, 2018; Song *et al.*, 2016). Incorrect event ordering might have a negative impact on the results of PMg analysis. According to the studies, time-based strategies produce superior outcomes in data preparation. The authors proved in (Dixit *et al.*, 2018; Suriadi *et al.*, 2017; Van der Aalst & Santos, 2021) that one of the most hidden and common issues in the event log is one linked with anomalies related to data variety (precision levels) and the sequence in which events are logged in the logs. As a result, tactics based on timestamp information are of tremendous interest in the forefront.

Dixit *et al.* (2018) developed an iterative method for addressing event order imperfection by interacting the context of information immediately into the event log and assessing the impact of the corrected log. This method is

based on the creation of three types of timestamp-based indicators to recognise ordering-related errors in an event log to highlight actions. These actions may be erroneously ordered, as well as a strategy for rectifying found faults using domain knowledge.

Hsu *et al.* (2017) suggested a k-nearest neighbour technique for identifying abnormal process instances systematically using a collection of activity-level durations, notably execution, transmission, queue, and procrastination durations. The necessary duration time to perform an activity and contextual information, such as staff information and customer transactions, derived from a medium-sized organisation company's ERP system is referred to as activity-level duration. The distances between occurrences were computed by subtracting the modified durations.

Tax *et al.* (2019) provided a methodology for the automatic production of label refinements focusing on time characteristics, allowing behaviourally distinct instances of the same event type to be distinguished using time parameters. The events that occurred by a single sensor were grouped using a hybrid model composed of components of different components, which is the circular counterpart of the normal distribution. On three event logs from the human behaviour domain, four methodologies for multiple label refinement were used.

Song *et al.* (2016) proposed a method based on the minimal change principle for repairing timestamps that do not correspond to temporal restrictions, such as finding a repair that is as near to the first observation as feasible. The challenge is addressed by finding a succinct collection of potential qualified applications applying an algorithm to compute the best repair from the produced ones and a heuristic solution through the selection of repairs from the qualified applications.

Rogge-Solti *et al.* (2013b) used stochastic Petri nets, alignments, and Bayesian networks to reconstruct timed event records. The technique divides the problem into two parts: mending the time and restoring the structure for each trace. This study takes into consideration all the observable data and generates efficient estimates for activity durations and route probabilities.

Fischer *et al.* (2020) suggested a method for identifying and measuring timestamp errors in event logs based on fifteen quality criteria organised over four categories of data quality and log levels.

2. Detection and visualisation

The goal of detection-visualisation approaches is to discover, organise, and isolate those events or traces that potentially cause difficulties with the event log's quality. Two methodologies are mentioned within this group: clustering and pattern-based strategies. Clustering methods separate the event log into numerous subsets, making each member of the subsets easier to comprehend and analyse. The next step is to identify noise or anomaly elements within the studied subsets. Clustering is one of the most often used approaches for data pre-processing in PMg, and it has mostly been used to identify quality concerns related with noisy values, as well as data variety. It is feasible to find imperfection patterns associated to noisy data in the different properties of the event logs based on the creation of comparable clusters.

Several strategies for trace clustering have been presented in the recent decade. They are classified as vector space techniques (Bose & van der Aalst, 2009; Evermann *et al.*, 2016; Song *et al.*, 2008; Xu & Liu, 2019), context aware approaches (Bose & Van der Aalst, 2009; Wang *et al.*, 2011; Bose, 2012; Hompes *et al.*, 2015; Sun *et al.*, 2017), and model-based approaches (Greco *et al.*, 2006; Ferreira *et al.*, 2007; Medeiros *et al.*, 2007; De Weerdt *et al.*, 2013; Nguyen *et al.*, 2016; Folino *et al.*, 2011). Most of the clustering algorithms simply take the event log as input and generate clusters using various internal representations. These methods have been used without regard for the availability of a process model. In contrast, recent publications (Boltenhagen *et al.*, 2019; Chatain *et al.*, 2017) give a distinct perspective on traces clustering in an event log. In contrast, recent publications (Boltenhagen *et al.*, 2019; Chatain *et al.*, 2017) give a distinct perspective on traces clustering in an event log. The authors presuppose the existence of a process model, which is utilised to construct simplified groups of homogenous traces.

Some detection-visualisation systems execute event log preparation from pattern recognition using the formulation

and implementation of heuristic guidelines. These guidelines are recognised by skilled analysts in PMg from the examination of various event logs in various contexts based on observed behaviours or gained experiences. Many pattern-based algorithms argue that if a specific pattern is not discovered in the log, the event log is not valid (Suriadi *et al.*, 2017). These approaches are typically used in conjunction with clustering, abstraction, or alignment techniques to identify patterns linked to noisy data or data variety.

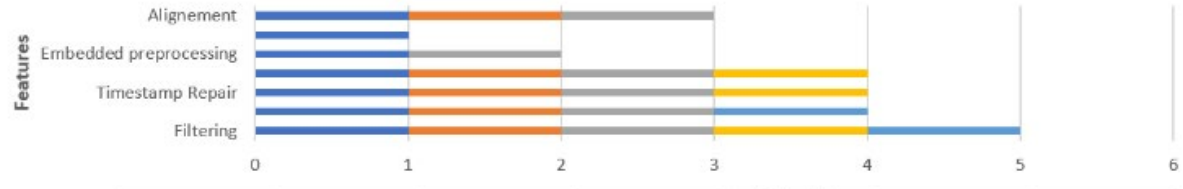
Suriadi *et al.* (2017) suggest assessing event log quality by describing a set of 11 log irregularity patterns derived from their experience generating event logs. A pattern is defined as an abstraction from a physical form that recurs in specified non-arbitrary circumstance. Ghionna *et al.* (2008) offer a method that combines frequent execution pattern identification with a cluster-based anomaly detection mechanism. Special methods are utilised to reduce the number of bogus activities and to code a technique that concurrently clusters a log and its related S-patterns (patterns and clustering).

WoMineI (Chapela-Campa *et al.*, 2017) retrieves features from the model definition in logs (activities, series, choices, similarities, loops, etc). WoMineI does an a priori search, beginning with the smallest patterns and narrowing the search space by removing rare patterns.

Jagadeesh *et al.* (2009) present an iterative technique for altering traces that recognises looping structures and sub-processes and replaces duplicate occurrences with an abstracted object. The authors (Cappiello *et al.*, 2022; Günther *et al.*, 2009; Hsu *et al.*, 2017; Leemans *et al.*, 2013; Lu *et al.*, 2016) propose other pattern-based techniques. Furthermore, several PMg methods (Folino *et al.*, 2009; Günther *et al.*, 2007; Gu *et al.*, 2007; Leemans *et al.*, 2013; Leemans *et al.*, 2014; Mannhardt *et al.*, 2017; Weijters & Ribeiro, 2011) use event log pre-processing processes (embedded approaches) as part of their methodology.

This is demonstrated by several research (Mannhardt *et al.*, 2017; Vanden Broucke & De Weerdt, 2017), which demonstrate that it is feasible to develop more solid and resilient models through the detection of noisy data as well as the flexible tuning of parameters in pre-processing procedures.

In this work, we provide a considerable overview of some of the most popular event log preparation approaches previously covered. Then, we cite main techniques that influenced the event pre-processing technique: event or trace level filtering, clustering, pattern-based approaches, Embedded techniques, time-based techniques, alignment, and abstraction. Moreover, we display the specific goal (discovery, conformance, or enhance) that is meant to be enhanced by using a pre-processing approach in that same sequence.



	Filtering	Clustering	Timestamp Repair	Remove	Embedded preprocessing	Abstraction	Alignement
ProM	1	1	1	1	1	1	1
Apromore	1	1	1	1	0	0	1
RapidProM	1	1	1	1	1	0	1
Disco	1	0	1	1	0	0	0
Celonis	1	1	0	0	0	0	0

Figure 7. Comparison study of prominent process mining tools with event log pre-processing integration

The key issues discovered in the event log are also indicated as missing data, noise data, diversity data (Bezerra & Wainer, 2008), irrelevant data, and duplicate data. In this context, we observe that the trace clustering approach and event or trace filtering are the two most commonly utilised strategies in PMg for the pre-processing task. Time-based pre-processing algorithms have lately demonstrated promising results in data pre-treatment by studying, correcting, and eliminating data linked with the timestamp characteristic.

Furthermore, we reveal that the vast majority of pre-processing algorithms have been developed to enhance the quality of discovered process model, as well as to reduce the model's complexity by managing and clean event log.

Furthermore, around 66% of the approaches analysed are accessible in PMg tools, such as the ProM tool, while a tiny fraction relates to applications that combine pre-processing techniques separately. Last, we depict the two most common issues in event logs are: noise and data variety related to different precision levels.

Disco (BV, 2011) is a commercial PMg tool that offers non-destructive filtering features for exploratory drill-down and analysis focus. These filters are available from any view and are simple to set up. They enable filtering by case performance, time, variation, characteristics, event relationships, or endpoints.

Celonis (S.E. 2009) proposes a method that identify the precise reasons of observed nonconformities of a BP. Through grouping and filtering techniques, this version, treats difficulties associated with noisy events or missing information.

Figure 7 outlines the key features of the previously stated tools. These are the most well-known and extensively utilised software. They incorporate pre-processing techniques to apply on event logs investigated in this study and enables PMg tasks (discovery, conformance, and augmentation) to be performed alongside pre-processing algorithms inside the same tool.

D. F3 Event Log Representation Schemes Used in Pre-processing Techniques

For many years, information representation has been a basic requirement in every sector, including PMg. Even if overall storage capacity is no longer an issue, because external memory can store massive numbers of data and is inexpensive, the time required to retrieve the event logs remains a significant blockage in the algorithm at hand. An adequate event log structure or representation scheme will enable efficient handling of large event logs by providing algorithms that process events straight from the representation.

The vector space model (or bag-of-events) (Greco *et al.*, 2006) is one of the most frequent event log representations used in pre-processing approaches, where each trace is represented as a vector and each dimension belongs to an event category. The similarity of traces in this form of representation is assessed using standard techniques such as Euclidean distance or Cosine similarity. Because it is easier to filter, aggregate, or eliminate new events or traces on this structure, several suggested systems for event log pre-processing employ traces or event sequences as data structures for representing and manipulating of event logs. Other structures, such as automatons, directed graphs, and trace arrays, have also been examined.

This work (Wang *et al.*, 2015) proposes a graph-fixing strategy for detecting unsound structure and correcting incorrect event names. This method restores event data with inconsistent labelling but sound structure, employing the least change principle to preserve as much of the original information as feasible. Then, an algorithm detects and repairs filthy event data at the same time, reporting unsound structure or providing the least amount of reparation for inconsistent event names. Furthermore, in (Wang *et al.*, 2015), an approximation technique that is introduced to repair one transition at a time, which is continually executed until all violations are erased or no further mending can be performed.

Mueller and Schultz (2013) describe a four-step pre-processing strategy for reconstructing process instance graphs to event log with a sequentially ordered set of activities by creating a directed sequence flow between instance graph activities. In this method, instance graphs are divided into separate pieces that may then be mapped into a sequential event log. Firstly, it is important to explore data using the financial PMg (FiPM) algorithm in order to provide process instances illustrated as graphs. Secondly, it

is primordially to convert these graphs into directed activity graphs. The third step is to count all the available pathways that constitute sub-sequences of the mined process instances. Last, in the fourth stage, all pathways indicating the logical sequence of events are stored in an event log.

According to the works reviewed, the best structure for representing, manipulating, and transforming event sequences that may be defined as a vector of data and are simple to build and use. This style also makes entering and removing events or traces easier, especially when working with long cases or instances in the event log.

E. F4 Event Logs Imperfection Type

According to Chandra Bose *et al.* (2009), most real-world that describe event logs are: granular, diverse, voluminous, incomplete, fine, and loud, which can present significant challenges in various PMg activities. Precision levels, on the other hand, is connected to the amount of detail with which events are stored. Without taking into account the intended degrees of analysis, this might vary greatly. This form of imperfection is directly connected to the quality issue of data variety because of a lack of sufficient logging rules and norms.

Due to the varied precision levels of the event log, the models created by the discovery process are sometimes spaghetti-like and difficult to understand. Heterogeneity in the event log indicates that numerous genuine processes occur in varied and unstructured contexts, resulting in a heterogeneous blend of these environments in the created event log.

Heterogeneity is caused by operational procedures that evolve through time in order to adjust to changing conditions. Trace clustering algorithms have been demonstrated to be an efficient way of dealing with diversity.

There are also imperfections in the event logs related to particulate granular timestamps, which implies that the series of events within the log may diverge from the sequence in which the events occurred in reality; mashed granular timestamps, which indicates that there are events for which the level of abstraction of their timestamps is distinct; and inaccurate timestamps, which signifies that the recorded timestamp of (some or all) events in the log is completely inaccurate.

Furthermore, there are additional typical flaws in the event log that are related to the data quality problems, such as lacking attribute values or events that are omitted from the trace despite the fact that they happened in fact There could also be errors in the event log as a result of event ambiguity, such as when many events have the same activity name. Another challenge is activity conflict, which occurs when one instance of an activity is begun and before the previous instance of the same activity is completed, a new instance of the same activity is begun. Furthermore, the occurrence of noisy data or extremes is typical, i.e., unusual, exceptional, or aberrant execution behaviour.

The authors of (Suriadi *et al.*, 2017) discovered a set of flaws patterns that are typically observed when pre-processing raw source logs. These patterns are determined by the issues discovered while converting. The raw data source logs into a 'clean' event log and useful for PMg research. Some of the imperfection patterns presented in (Suriadi *et al.*, 2017) are relevant to the challenge of omitted or noisy event values, such as when events in a log are not clearly linked to their proper case identifiers, or when critical process steps are omitted in the event log being evaluated but captured elsewhere. The majority of event for trace-level filtering algorithms discussed in Section 3 aim to find omitted events and repair them, or to delete aberrant events from the event log., to address this sort of imperfection pattern.

Another form of imperfection pattern discussed in (Suriadi *et al.*, 2017) is one connected to issues with the timestamp property. This flaw happens when mistakes in the timestamp are recorded, i.e., in timestamp values that are stored in various format than intended (data diversity), or when events are stored from digital forms (discrepancy) in the sequence of the events that were conducted. Techniques for dealing with timestamp issues are primarily focused on assessing the influence that timestamp information has on improving the quality of the event log.

There are correlations associated with difficulties in event labels, including the existence of a set of values that are grammatically distinct but conceptually related, or the presence of different values that do not have an excellent

match with certain other values but have robust syntactic and semantic commonalities. Abstraction methods and clustering appear as the most suited for translating event labels to a higher accuracy level, allowing them to serve as a bridge between a low-level event log and an acceptable high-level view of the log.

However, Dixit *et al.* (2017) have discovered that there are time-related indications that can reveal flaws in the sequence of occurrences in a log. Among many of the detected markers are: (1) the presence of either coarse timestamp granularity or heterogeneous timestamp precision from many systems, each of which maintains timestamps uniquely. For instance, consider the case where an event x is recorded at varied precision levels. Another event within the same case may have varied precision levels. The sequence of these two occurrences will be erroneous; (2) recognising events with extraordinary chronological arrangement (for example, recreate input of the same event; 3) beginning to learn the chronological place of one action in relation to other activities, or the dispersion of timestamp values over all occurrences in a log, may confirmed the presence of timestamp-related difficulties. For instance, we suppose that a log includes events from multiple services. Thus, timestamps can be represented in different ways, resulting in the 'miss fielded' or 'unmovable' timestamp problem.

Regardless of the variety of flaws that may exist in the event log, as shown in a survey of the state-of-the-art, prevalent issues are linked to the existence of unbalanced data, but also data variety in the event log that that differs from expected behaviour.

E. F5 Related Tasks

We identified two tasks strongly related to data pre-processing in PMg. The first is event abstraction, and the second is alignment. Both activities have the ability to increase the event log or process model quality, in addition to the reliability of several PMg approaches.

1. Event logs abstraction

The bulk of PMg approaches presume that event data is gathered at the same precision levels level. Nevertheless, in the real world, information systems record events at varied precision levels (Mueller-Wickop & Schultz, 2013). In many cases, events are captured in a single event log are combined and presented at a fine-grained level, resulting in models of unexplained processes or models that do not fit event log produced by process techniques and process discovery algorithms. In these cases, event abstraction techniques transform the event log to a higher precision level, find the relationship between the unique low-level event log and the required high-level approach on the log, allowing for the discovery of more comprehensible process models.

When a segment of the sequences has elevated explanations for reduced occurrences, some proposed event abstraction techniques use supervised learning. These expressions make information on how to categorise higher-level occurrences as well as information on the appropriate level of abstraction A broad strategy for controlled event abstraction uses two input information: (1) a series of annotated trace amounts in which for all poor events in the trace, the high-level event to which it corresponds (the label property of the low-level event) is determined, and (2) a collection of unlabelled data traces in which low-level events are not plotted to high-level events.

Tax *et al.* (2016) employ supervised learning in conjunction with a condition generalised additive learning phase. to abstract events from a low-level of event log. A supervised learning model on a collection of traces with defined high-level target labels can be used to produce a high-level analysis of a low-level event log, and the model may be used to categorise further low-level traces.

High-level event label recognition may be thought of as a sequence mentioning a problem where each event is labelled as a higher-level event from a high-level event nomenclature. This study presents a sequence-focused measure for assessing supervised event abstraction outcomes, which is relevant to process discovery and conformance checking activities.

On the annotated traces, conditional random fields are generated to construct a randomised transformation of low-level events into high-level occurrences. Once obtained, this mapping may be used to estimate the comparable high-level event in unlabelled data traces.

Sun and Bauer (2016) suggest a process model abstraction mechanism to optimise the quality of the treated high-level model while also taking into account the quality of the created sub-models, where each sub-model is utilised to display the specifics of its essential high-level activity.

Other methods investigated in the PMg field address the challenge of abstracting low-level events to higher-level events (Baier & Mendling, 2013; Günther *et al.*, 2009; Van Dongen & Adriansyah, 2009; Sun & Bauer, 2016; Mannhardt *et al.*, 2016). Existing event abstraction approaches (Alharbi, 2019; Bezerra & Wainer, 2013; Mannhardt *et al.*, 2017) depend on unsupervised learning techniques to combine low-level events into a single high-level event. Available methods need the user or process analyst is expected to supply high-level event labels based on domain expertise, or to construct long labels by appending the labels of all low-level events in the ensemble. Many well-known unsupervised event abstraction techniques include one or more parameters that evaluate the quantity of events classified into higher-level events. Determining the correct degree of abstraction to yield significant results is sometimes a risky endeavour.

2. Alignment

Correlating the events in the log with the activities in the process model is required for alignment, as is finding the gaps and degrees of conformance between a log and a model. Alignment is important not just for conformance testing, but also for event log and model repair. Model fixing, on the other hand, focuses on mending the essential portions of the process model in order to playback event sequences in the log (Song *et al.*, 2016).

The sequential alignment of a trace and a process model is defined as a series of moves, each of which connects an event in the trace to an activity in the model. Each possible move is assigned a cost by a cost function. An optimal alignment is a sequential alignment with the lowest cost according to the cost function. Unfortunately, finding the optimal alignment is an NP-hard problem (Leoni & Van Der Aalst, 2013). This indicates that determining the optimal alignment is difficult when the process is large. The alignment is optimal when a trace in the log and a recurrence sequence in the model have the minimal distance measure.

For correcting missing log entries, Rogge-Solti *et al.* (2013b) propose a cost-based alignment technique. The authors employ probabilistic filtering models to identify the most probable timing of lacking events using route likelihood and stochastically enhanced process models. A stochastic Petri net captures the relationships between arbitrary durations using Bayes' theorem.

Song *et al.* (2015) proposed a method for retrieving omitted events in process logs that involves breaking the process down into sub-processes and utilising heuristics to remove inadequate sub-processes that promise to implement the minimal recovery. To decrease duplicate recoveries in concurrent routings, the authors employ an approach that utilises trace replaying to quickly determine a minimal recovery.

De Leoni *et al.* (2012) provide a technique for verifying conformity that is predicated on the idea of identifying an alignment with an event log and its related process model. For each trace in the event log, the A* algorithm is utilised to identify an optimal combination, that is, an alignment that minimises the cost of the deviations. To deal with the wide search spaces produced by expressive models' intrinsic flexibility, the authors alter alignment-based techniques. They give inspections at the trace level based on such alignments, explaining why events in a trace need to be added or eliminated, and colouring constraints and activities in the model depending on their degree of compliance at the model level.

Song *et al.* (2016) use process model structural and behavioural features. To decrease the search space for the ideal alignment, efficient algorithms based on process decomposition and trace replaying are used. They use a divide-and-conquer tactic. For the alignment to happen, the events in the log must be connected with the activities in the process model. In order to discover the best alignment between process models and event logs with omitted, redundant, or dislocated events and activities, a generic framework is created. This framework is still used to not only align event logs with process models, but also to repair event logs. Their idea is realised in the Effa tool, which acts as a ProM plugin. Other alignment works can be found in

(Bose *et al.*, 2012; Jagadeesh Chandra Bose *et al.*, 2010; Lu *et al.*, 2014).

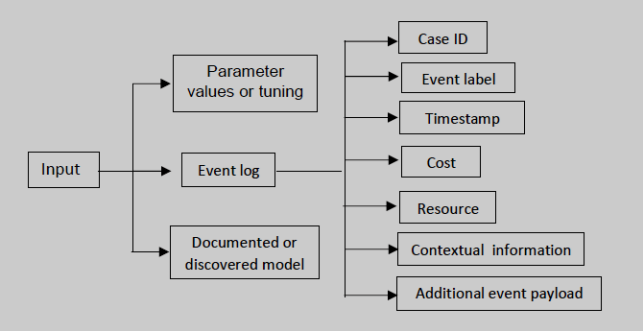


Figure 8. Popular input data of pre-processing techniques

F. F6 Information Types

Pre-processing approaches make use of a variety of information resources to enhance the quality of event logs generated by mined or already developed models. Also, the event log is viewed as the foundation element to be used for PMg activities, namely the automatic discovery of process models, reference source in terms of conformity with the previously established model. Many data pre-processing solutions in PMg revolve around working with a set of instances or traces created from a collection of events registered for process executions.

Therefore, various techniques based on flaws found in these scientific articles have been suggested. Some studies have concentrated on some of the features appeared in the set of event logs (timestamp that specifies the sequence of events inside the trace). As a result, erroneous event sequencing might have a detrimental influence on PMg analysis results. In some cases, the timestamp property has been leveraged in PMg to identify unusual processes by combining activity-level durations with relevant information. CaseID, event label, timestamp, cost, resource, relevant information, and additional event are used during the pre-processing task (see Figure 8). If any of these attributes' information is undeclared, incomplete, disordered, noisy, or irregular, the pre-processing techniques clean and replace it.

Lamghari *et al.* (2022) uses chaotic activity detection to recognise deficiencies from event logs and enhance their quality. These are activities that have no fixed position in the process model and may happen at any time during the workflows. In this context, it is difficult to model the process representing this later. Other works on event log pre-processing techniques focus on the existing process model that the organisation's specialists have previously built (Rogge-Solti *et al.*, 2013a; Chatain *et al.*, 2017; Boltenhagen *et al.*, 2019; Song *et al.*, 2015; Goel *et al.*, 2022; Martin *et al.*, 2022).

Furthermore, for filtering or accepting events, most pre-processing techniques require specific parameter values or tuning. This enables the establishment of specific decision thresholds and, as a result, the determination of whether any features can be considered inaccurate or noisy.

Many noisy event identification and visualisation approaches, notably clustering algorithms, take a collection of traces as input, allowing the original event log to be divided more easily. Thus, distances between traces are therefore estimated using this set, and instances with high similarity are grouped together, while instances with low similarity are separated. Some clustering strategies use event-related (data, resources, etc), to enhance event log segmentation. Indeed, pattern-based pre-processing approaches, on the other hand, typically employ the raw event log to find structural forms that maintain repeating non-arbitrary contexts, with the timestamp attribute being the most often adapted by these techniques.

To uncover problems related with omitted or inaccurate data included in the distinct characteristics in the event log, it is typical to apply a collection of traces inside transformation techniques (filtering).

Therefore, the links between the various features of the pre-processing techniques studied in this work approve that the filtering-based techniques are available in the majority of PMg tools. The pattern-based techniques, on the other hand, are uniquely used with the ProM tool. To readily execute transformations on the records, the bulk of the processing techniques of the various classes handle sequences of traces or events as their event log encoding scheme. Traces are information resources that are typically employed in the pre-processing work. Furthermore, all pre-processing strategies consider the detection, isolation, and eradication of noisy data, as well as, to a limited extent, the resolution of incomplete, redundant, and inappropriate data issues.

IV. CONCLUSION

For the first time, we published a literature overview on the primary methodologies utilised in data preparation for PMg in this study. A description of approaches and algorithms, tools, commonly challenges, views, and data kinds were all covered in the review. Representative works were rigorously revised to identify the main components in pre-processing strategies that contribute to improved process model quality.

Indeed, this publication presented a grouping of the many available pre-processing strategies. This section is divided into two sections: transformation techniques and detection-visualisation approaches. Transformation techniques perform activities to record changes in the raw event log's original structure to enhance the log's quality. While detection–visualisation approaches discover, aggregate, and isolate those events or traces that may cause difficulties with the event log's quality.

We also discussed the issues that these strategies must overcome. Moreover, this survey highlights important factors to take into account during the data preparation phase in PMg: 1. grouping available techniques of event log pre-processing; 2. available in the literature pre-processing tools in the context of PMg; 3. the suitable data structures to represent and manipulate event data according to the pre-processing algorithms; 4. the issues and internal defects frequently recognised into event data; 5. the tasks commonly associated with event log in the pre-processing stage; and 6. the type of attributes or information to treat.

This study may be used as a reference tool to suggest the types of pre-processing procedures and their characteristics. Furthermore, it aims to emphasise the characteristics that should be considered in order to generate process models that are basic and simple to read.

As a result of our study, we can state unequivocally that data preparation in the context of PMg still a primordial issue. With the emergence of complex event data, it is required to build new pre-processing mechanisms to cope with these new difficulties none previously discovered or treated in the generation of massive event logs. While clustering and data filtering approaches have made noteworthy progress in PMg, other strategies, use the detection of flaw patterns, have yet to adequately address them with the automatic detection.

A set of metrics identifying the presence of noise patterns in event logs should be established as part of future work to be considered. The noise patterns may influence an event log at several levels. Hence, the different pervasiveness measures could represent the number of attributes, events, and cases impacted by noisy patterns. Few studies (Suriadi *et al.*, 2017; Fischer *et al.*, 2020) have attempted to detect and measure the number of flaws in an event log. In this sense, exact measures are note established to evaluate the aforementioned level.

As further work, we aim at creating frameworks that compute not only the accuracy or fitness gained by the model to examine the influence of the pre-processing performed, as well as the computational expenses, memory requirements, and cyclical complexity with data cleaning. In future work, we hope to develop frameworks that calculate the computing costs, memory, and behavioural complexity of events logs matched with the prep-processing phase.

V. REFERENCES

Alharbi, AM 2019, 'Unsupervised abstraction for reducing the complexity of healthcare process models', PhD thesis, University of Leeds, UK.

Baier, T & Mendling, J 2013, 'Bridging abstraction layers in process mining by automated matching of events and activities', Business Process Management, vol. 8094, pp. 17-32.

Bezerra, F & Wainer, J 2008, 'Anomaly detection algorithms in logs of process aware systems', in Computing Machinery: Proceedings of the ACM symposium on Applied computing, 16 March 2008, Ceara, Association for Computing Machinery, Brazil.

Bezerra, F, Wainer, J & van der Aalst, WM 2009, 'Anomaly detection using process mining' Business Information Processing, vol. 29, pp. 149-161.

Bezerra, F D L & Wainer, J 2012, 'A dynamic threshold algorithm for anomaly detection in logs of process aware systems', Journal of Information and Data Management, vol. 3, no. 3, pp. 316–331.

Bezerra, F & Wainer, J 2013, 'Algorithms for anomaly detection of traces in logs of process aware information systems', Information Systems, vol. 38, no. 1, pp. 33-44.

Böhmer, K & Rinderle-Ma, S 2016, 'Multi-perspective anomaly detection in BP execution events', in the Move to Meaningful Internet Systems: Proceedings of the OTM Confederated International Conferences, 24 October 2016, Rhodes, Springer, Greece.

Boltenhagen, M, Chatain, T & Carmona, J 2019, 'Generalized alignment-based trace clustering of process behavior', in Theoretical Computer Science and General Issues: Proceedings of the International Conference on Applications and Theory of Petri Nets and Concurrency, 23 June 2019, Aachen, Springer, Germany.

Bose, RJC & Van der Aalst, WM 2009, 'Context aware trace clustering: Towards improving process mining results', in Theory and practice of data mining: Proceedings of the SIAM International Conference on Data Mining, 2009, Philadelphia, Society for Industrial and Applied Mathematics, USA.

Bose, RP & van der Aalst, WM 2009, 'Trace clustering based on conserved patterns: Towards achieving better process models', in Business Process Management Workshops: Proceedings of the International Conference on BP Management, pp. 170-181, Springer, Ulm, 7 September 2009, Springer, Germany.

Bose, RJC & van der Aalst, WM 2012, 'Process diagnostics using trace alignment: opportunities, issues, and challenges', Information Systems, vol. 37, no.2, pp.117-141.

Bose, RJC 2012, 'Process mining in the large: preprocessing', discovery, and diagnostics, PhD thesis, Eindhoven University of Technology, Netherlands.

BV, F 2011, 'Discover Your Processes', Fluxion Process Mining for Professionals, viewed on 12 March 2022, <https://fluxicon.com/disco/>.

Cappiello, C, Comuzzi, M, Plebani, P & Fim, M 2022, 'Assessing and improving measurability of process performance indicators based on quality of logs', Information Systems, vol. 103, no. 101874.

Celonis, S E & Munich, G, Celonis, Process Mining, CELONIS, viewed on 12 Marsh 2022, <https://www.celonis.com/>.

Chapela-Campa, D, Mucientes, M & Lama, M 2017, 'Discovering infrequent behavioral patterns in process models', in Business Process Management: Proceedings of the International Conference on BP Management, Barcelona, 10 September 2017, Springer, Spain.

Chatain, T, Carmona, J & Dongen, BV 2017, 'Alignment-based trace clustering', in Conceptual Modeling: Proceedings of the International Conference on Conceptual Modelling, November 2017, Hal science, Spain.

Cheng, HJ & Kumar, A 2015, 'Process mining on noisy logs-can log sanitization help to improve performance?', Decision Support Systems, vol. 79, pp. 138-149.

Chen, Q, Lu, Y, Tam, CS & Poon, SK 2022, 'A Multi-View Framework to Detect Redundant Activity Labels for More Representative Event Logs in Process Mining', Future Internet, vol. 14, no. 6, pp. 181.

Conforti, R, La Rosa, M & ter Hofstede, AH 2016, 'Filtering out infrequent behavior from BP event logs', IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 2, pp. 300-314.

Denisov, V, Fahland, D & van der Aalst, WM 2020, 'Repairing event logs with missing events to support performance analysis of systems with shared resources', in Petri Nets: Proceedings of the International Conference

on Applications and Theory of Petri Nets and Concurrency, 24 June 2020, Paris, Springer, France.

De Weerdt, J, Vanden Broucke, S, Vanthienen, J & Baesens, B 2013, 'Active trace clustering for improved process discovery', IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 12, pp. 2708-2720.

Durojaiye, A, Fackler, J, McGeorge, N, Webster, K, Kharrazi, H & Gurses, A 2022, 'Examining Diurnal Differences in Multidisciplinary Care Teams at a Pediatric Trauma Center Using Electronic Health Record Data: Social Network Analysis', Journal of medical Internet research, vol. 24, no. 2.

Dixit, PM, Suriadi, S, Andrews, R, Wynn, MT, ter Hofstede, AH, Buijs, JC & van der Aalst, WM 2018, 'Detection and interactive repair of event ordering imperfection in process logs', in Advanced Information Systems Engineering: Proceedings of the International Conference on Advanced Information Systems Engineering, Tallinn, 11 June 2018, Springer, Estonia.

Emamjome, F, Andrews, R, ter Hofstede, A & Reijers, H 2020, 'Alohomora: Unlocking data quality causes through event log context', in Computer Science: Proceedings of the 28th European Conference on Information Systems, Association for Information Systems, 15 June 2017, Marrakech, AIS, Morocco.

Evermann, J, Thaler, T & Fettke, P 2016, 'Clustering traces using sequence alignment', in Lectures Notes in Business Information: Proceedings of the International Conference on BP Management, Innsbruck, 31 August 2015, Springer, Austria.

Fani Sani, M, Zelst, SJV & van der Aalst, WM 2018a, 'Repairing outlier behaviour in event logs', in Business Information Systems: Proceedings of the International Conference on Business Information Systems, Berlin, 18 July 2018, Springer, Germany.

Fani Sani, M, Zelst, SJV & van der Aalst, WM 2018b, 'Applying sequence mining for outlier detection in process mining', in Confederated International Conferences: Proceedings of the OTM Confederated International Conferences, On the Move to Meaningful Internet Systems, Valletta, 22 October 2018, Springer, Malta.

Ferreira, D, Zacarias, M, Malheiros, M & Ferreira, P 2007, 'Approaching process mining with sequence clustering: Experiments and findings', in Business Process Management: Proceedings of the International conference on BP management, Brisbane, 24 September 2020, Springer, Australia.

Fischer, DA, Goel, K, Andrews, R, Dun, CGJV, Wynn, MT & Röglinger, M 2020, 'Enhancing event log quality: detecting and quantifying timestamp imperfections', in Business Process Management: Proceedings of the International Conference on BP Management, Seville, 13 September 2020, Springer, Spain.

Folino, F, Greco, G, Guzzo, A & Pontieri, L 2009, 'Discovering expressive process models from noised log data', in IDEAS: Proceedings of the international database engineering & applications symposium, Calabria, September 2009, ACM, Italy.

Folino, F, Greco, G, Guzzo, A & Pontieri, L 2011, 'Mining usage scenarios in BPes: Outlier-aware discovery and run-time prediction', Data & Knowledge Engineering, vol. 70, no. 12, pp. 1005-1029.

Fox, F, Whelton, H, Johnson, OA & Aggarwal, VR 2022, 'Dental Extractions under General Anesthesia: New Insights from Process Mining', JDR Clinical & Translational Research, vol. 8, no. 3, pp. 267-275.

Ghionna, L, Greco, G, Guzzo, A & Pontieri, L 2008, 'Outlier detection techniques for process mining applications', in Foundations of Intelligent Systems: Proceedings of the International symposium on methodologies for intelligent systems, Toronto, 20 May 2008, Springer, Canada.

Greco, G, Guzzo, A, Pontieri, L & Sacca, D, 2006, 'Discovering expressive process models by clustering log traces', IEEE Transactions on knowledge and data engineering, vol. 18, no. 8, pp. 1010-1027.

Gschwandtner, T, Aigner, W, Miksch, S, Gärtner, J, Kriglstein, S, Pohl, M & Suchy, N 2014, 'TimeCleanser: A visual analytics approach for data cleansing of time-oriented data', in i-KNOW'14: Proceedings of the 14th international conference on knowledge technologies and data-driven business, 19 September 2014, Graz, ACM, Austria.

Günther, C. W & Van Der Aalst, WM 2007, 'Fuzzy mining–adaptive process simplification based on multi-perspective metrics', in Business Process Management: Proceedings of the International conference on BP management, Brisbane, 24 September 2007, Springer, Australia.

Gu, C Q, Chang, HY & Yi, Y 2008, 'Workflow mining: Extending the αalgorithm to mine duplicate tasks', in LNCS: Proceedings of the International Conference on Machine Learning and Cybernetics, IEEE.

Günther, C W, Rozinat, A & Van Der Aalst, WM 2009, 'Activity mining by global trace segmentation', in LNPIB:

proceedings of the International Conference on BP Management, Ulm, 7 September 2009, Springer, Germany.

Goel, K, Leemans, SJ, Martin, N & Wynn, MT 2022, 'Quality-Informed Process Mining: A Case for Standardised Data Quality Annotations', ACM Transactions on Knowledge Discovery from Data (TKDD)

Hsu, PY, Chuang, YC, Lo, YC & He, SC 2017, 'Using contextualized activity-level duration to discover irregular process instances in business operations', Information Sciences, vol. 391, pp. 80-98.

Hompes, B, Buijs, JCAM, Van der Aalst, WMP, Dixit, PM & Buurman, J 2015, 'Discovering deviating cases and process variants using trace clustering', in Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC), November, vol. 5, no. 6.

Huang, Y, Zhong, L & Chen, Y 2020, 'Filtering infrequent behavior in BP discovery by using the minimum expectation', International Journal of Cognitive Informatics and Natural Intelligence, vol. 14, no. 2, pp. 1-15.

Interstage BP Manager Analytics by Fujitsu Ltd., 2009, Interstage, viewed on 12 Marsh 2022, <https://www.fujitsu.com/>.

Jagadeesh Chandra Bose, RP & Van der Aalst, WM 2009, 'Abstractions in process mining: A taxonomy of patterns', in Business Process Management: Proceedings of the International conference on BP management, Ulm, 8 September 2009, Springer, Germany.

Jagadeesh Chandra Bose, RP & Aalst, WVD 2010, 'Trace alignment in process mining: opportunities for process diagnostics', in Business Process Management: Proceeding of the International Conference on BP Management, Hoboken, 13 September 2010, Springer, USA.

Jalali, H & Baraani, A 2010, 'Genetic-based anomaly detection in logs of process aware systems', International Journal of Computer and Systems Engineering, vol. 4, no. 4, pp. 692-697.

Kong, L, Li, C, Ge, J, Li, Z, Zhang, F & Luo, B 2019, 'An Efficient Heuristic Method for Repairing Event Logs Independent of Process Models, in Proceedings of the 4th International Conference on Internet of Things, Big Data and Security IoTBDS, Creta, 2 May 2019, SCITEPRESS, Grece.

Lamghari Z, Saidi R, Radgui M & Rahmani, MD n.d., 'Chaotic activities recognizing during the pre-processing event data phase', International Journal of Business Intelligence and Data Mining (IJBIDM), vol. 20, no. 4, pp. 413-239.

Leemans, SJ, Fahland, D & Van Der Aalst, WM 2013, 'Discovering block-structured process models from event logs containing infrequent behaviour', in International conference on BP management, pp. 66-78, Springer, Cham.

Leemans, SJ, Fahland, D & van der Aalst, WM 2014, 'Discovering block-structured process models from incomplete event logs', in International conference on applications and theory of petri nets and concurrency, pp. 91-110, Springer, Cham.

Leoni, MD, Maggi, FM & van der Aalst, WM 2012, 'Aligning event logs and declarative process models for conformance checking', in Business Process Management: Proceedings of the International Conference on BP Management, 3 September 2012, Tallinn, Springer, Estonia.

Leoni, MD & Van Der Aalst, WM 2013, 'Aligning event logs and process models for multi-perspective conformance checking: An approach based on integer linear programming', Business Process Management, vol. 8094, pp. 113-129.

Li, G & van der Aalst, WM 2017, 'A framework for detecting deviations in complex event logs', Intelligent Data Analysis, vol. 21, no. 4, pp. 759-779.

Lu, X, Fahland, D & van der Aalst, WM 2014, 'Conformance checking based on partially ordered event data', in Business Process Management: Proceedings of the International Conference on BP management, 7 September 2014, Eindhoven, Springer, Netherlands.

Lu, X, Fahland, D & van der Aalst, WM 2016, 'Interactively Exploring Logs and Mining Models with Clustering, Filtering, and Relabeling', BPM, vol. 1789, pp. 44-49.

Maddah, N & Roghanian, E 2021, 'Data-driven performance management of business units using process mining and DEA: case study of an Iranian chain store', International Journal of Productivity and Performance Management, vol. 72, no. 2, pp. 550-575.

Mannhardt, F, Leoni, MD, Reijers, HA, Van Der Aalst, WM & Toussaint, PJ 2016, 'From low-level events to activities-a pattern-based approach', in Lectures Notes Computer Science: Proceedings of the International conference on BP management, Rio de Janeiro, 18 September 2016, Springer, Brazil.

Mannhardt, F & Tax, N 2017, 'Unsupervised event abstraction using pattern abstraction and local process models', vol. 1859, pp. 1-9.

Mannhardt, F, Leoni, M D, Reijers, HA & van der Aalst, WM 2017, 'Data-driven process discovery-revealing conditional infrequent behavior from event logs', in Information system application: Proceedings of the International conference on advanced information systems engineering, Essen, 12 June 2017, Springer, Germany.

Mans, RS, Van der Aalst, WM, Vanwersch, RJ & Moleman, AJ 2012, 'Process mining in healthcare: Data challenges when answering frequently posed questions', Process support and knowledge representation in health care, vol. 7738, pp. 140-153.

Mans, R, van der Aalst, WM & Verbeek, HMW 2014, 'Supporting Process Mining Workflows with RapidProM', BPM, vol. 56, pp. 1-6.

Martin, N 2021, 'Data quality in process mining', eds Carlos Fernandez-Llatas, in Interactive Process Mining in Healthcare, Springer, pp. 53-79.

Martin, N, Van Houdt, G & Janssenswillen, G 2022, 'DaQAPO: Supporting flexible and fine-grained event log quality assessment', vol. 191, no. 7.

Massimiliano, D 2016 myInvenio By Cognitive Technology, viewed on 12 March 2022, <www.my-invenio.com>.

Medeiros, AKAD, Guzzo, A, Greco, G, Van der Aalst, WM, Weijters, AJMM, Dongen, BFV & Sacca, D 2007, 'Process mining based on clustering: A quest for precision', in Lecture Notes in Computer Science: Proceeding of the International conference on BP management, Brisbane, 24 September 2007, Springer, Australia.

Minit By Gradient ECM, 2015, viewed 12 March 2022, <https://golden.com/wiki/Minit-5NNVAR>.

Mueller-Wickop, N & Schultz, M 2013, 'ERP event log preprocessing: timestamps vs. accounting logic', in Design Science at the Intersection of Physical and Virtual Design: Proceedings in the International Conference on Design Science Research in Information Systems, Helsinki, 11 December 2013, Springer, Finland.

Neerumalla, S & Parvathy, LR 2022, 'Improved invasive weed-lion optimization-based process mining of event logs', International Journal of System Assurance Engineering and Management, vol. 2022, pp. 1-11.

Nguyen, P, Slominski, A, Muthusamy, V, Ishakian, V & Nahrstedt, K 2016, 'Process trace clustering: A heterogeneous information network approach', in Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, 5 May 2016, Society for Industrial and Applied Mathematics, Society for Industrial and Applied Mathematics Publications, United States.

Rai, S, Geetha, M & Kumar, P 2022, 'Preprocessing of Datasets Using Sequential and Parallel Approach: A Comparison', Expert Clouds and Applications, vol. 209, pp. 311-320.

Rogge-Solti, A, Mans, RS, van der Aalst, WM & Weske, M 2013a, 'Repairing event logs using timed process models', in Lecture Notes in Computer Science: Proceedings of the OTM Confederated International Conferences on the Move to Meaningful Internet Systems, Graz, 9 September, Springer, Austria.

Rogge-Solti, A, Mans, RS, van der Aalst, WM & Weske, M 2013b, 'Improving documentation by repairing event logs', in LNBIP: Proceedings of the IFIP Working Conference on The Practice of Enterprise Modeling, Riga, 6 November 2013, Springer, Latvia.

Sani, MF, Zelst, SJV & van der Aalst, WM 2017, 'Improving process discovery results by filtering outliers using conditional behavioural probabilities', in Business Process Management Workshops: Proceeding of the International Conference on BP Management, Barcelona, 10 September, Springer, Spain.

Sani, MF, Berti, A, van Zelst, SJ & van der Aalst, WM 2019, 'Filtering Toolkit: Interactively Filter Event Logs to Improve the Quality of Discovered Models', Business Process Improvement, vol. 2019, pp. 134-138.

Sani, MF, van Zelst, SJ & van der Aalst, WM 2019, 'Repairing outlier behaviour in event logs using contextual behaviour', Enterprise Modelling and Information Systems Architectures, vol. 14, pp. 1-5.

Scannapieco, M 2006, 'Data Quality: Concepts, Methodologies and Techniques', Data-Centric Systems and Applications, vol. 2006.

Song, M, Günther, CW & Van der Aalst, WM 2008, 'Trace clustering in process mining', Business Process Management, vol. 17, pp. 109-120.

Song, W, Xia, X, Jacobsen, HA, Zhang, P & Hu, H 2015, 'Heuristic recovery of missing events in process logs', in Proceedings of the IEEE International Conference on Web Services, June 2015, Chicago, IEEE, USA.

Song, W, Xia, X, Jacobsen, HA, Zhang, P & Hu, H 2016, 'Efficient alignment between event logs and process

models', IEEE Transactions on Services Computing, vol. 10, no. 1, pp. 136-149.

Song, S, Cao, Y & Wang, J 2016, 'Cleaning timestamps with temporal constraints', VLDB Endowment, vol. 9, no. 10, pp. 708-719.

Sun, Y & Bauer, B 2016, 'A Graph and Trace Clustering-based Approach for Abstracting Mined BP Models', in Software Methodologies: Proceedings of the first International Conference of Enterprise Information Systems, 8 December 2016, SCITEPRESS.

Sun, Y, Bauer, B & Weidlich, M 2017, 'Compound trace clustering to generate accurate and simple sub-process models', in Service Oriented Computing: Proceedings of the International Conference on Service-Oriented Computing, 13 November 2017, Malaga, Springer, Spain.

Suriadi, S, Andrews, R, ter Hofstede, AH & Wynn, MT 2017, 'Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs', Information systems, vol. 64, pp. 132-150.

Tax, N, Sidorova, N, Haakma, R & van der Aalst, WM 2016, 'Event abstraction for process mining using supervised learning techniques', in SAI Intelligent Systems: Proceedings of the SAI Intelligent Systems Conference, London, 21 September 2016, Springer, UK.

Tax, N, Alasgarov, E, Sidorova, N, Haakma, R & van der Aalst, WM 2019, 'Generating time-based label refinements to discover more precise process models', Journal of Ambient Intelligence and Smart Environments, vol. 11, no. 2, pp. 165-182.

TA Foundation, Apromore-Advanced Process Analytics Platform, 2011, The University of Melbourne, viewed on 12 March 2022, < https://apromore.org/>.

Ukkonen, E 1995, 'On-line construction of suffix trees', Algorithmica, vol. 14, no. 3, pp. 249-260.

Van Cruchten, R & Weigand, H 2022, 'Towards Event Log Management for Process Mining-Vision and Research Challenges', in Research in Information Systems: Proceedings of the International Conference on Research Challenges in Information Science, Barcelona, 17 May 2022, Springer, Spain.

Van der Aalst, WMP 2016, 'Data Science in Action', Process Mining: Data science in Action, 2nd ed, Berlin Heidelberg, Springer, pp. 3-22.

Van der Aalst, WM, Bolt, A & van Zelst, SJ 2017, 'RapidProM: mine your processes and not just your data', Data Mining and Knowledge Discover, vol. 2, pp. 1-25.

Van der Aalst, WM & Santos, L 2021, 'May I Take Your Order? On the Interplay Between Time and Order in Process Mining', ArXiv, vol. 1703.

Van Dongen, BF & Adriansyah, A 2009, 'Process mining: fuzzy clustering and performance visualization', in Business Information Processing: Proceedings of the International Conference on BP Management, Ulm, 7 September 2009, Springer, Germany.

Van Dongen, BF, de Medeiros, AKA, Verbeek, HMW, Weijters, AJMM & van Der Aalst, WM 2005, 'The ProM framework: A new era in process mining tool support', in Applications and Theory of Petri Nets: Proceedings of the International conference on application and theory of Petri nets, Miami, 20 June 2005, Springer, FL.

Vanden Broucke, SK & De Weerdt, J 2017, 'Fodina: a robust and flexible heuristic process discovery technique', decision support systems, vol. 100, pp. 109-118.

Vathy-Fogarassy, Á, Vassányi, I & Kósa, I 2022, 'Multi-level process mining methodology for exploring disease-specific care processes', Journal of Biomedical Informatics, vol. 125.

Van Zelst, SJ, Mannhardt, F, de Leoni, M & Koschmider, A 2021, 'Event abstraction in process mining: literature review and taxonomy', Granular Computing, vol. 6, no. 3, pp. 719-736.

Vidgof, M, Djurica, D, Bala, S & Mendling, J 2020, 'Cherry-picking from spaghetti: Multi-range filtering of event logs', Enterprise, Business-Process and Information Systems Modeling, vol. 387, pp. 135-149.

Wand, Y & Wang, RY 1996, 'Anchoring data quality dimensions in ontological foundations', Communications of the ACM, vol. 39, no. 11, pp. 86-95.

Wang, X, Zhang, L & Cai, H 2011, 'Using Suffix-Tree to Identify Patterns and Cluster Traces from Event Log', in Signal Processing and Information Technology: Proceedings of the International Joint Conference on Advances in Signal Processing and Information Technology, Amsterdam, 1 December 2011, Springer, Netherlands.

Wang, J, Song, S, Zhu, X & Lin, X 2013, 'Efficient recovery of missing events', VLDB Endowment, vol. 6, no. 10, pp. 841-852.

Wang, J, Song, S, Lin, X, Zhu, X & Pei, J 2015, 'Cleaning structured event logs: A graph repair approach', in Proceedings of the 31st International Conference on Data Engineering, May 2015, IEEE.

Weijters, AJMM & Ribeiro, JTS 2011, 'Flexible heuristics miner (FHM)', Symposium on Computational Intelligence and Data Mining, vol. 2011, pp. 310-317.

Workflow and Case Management 2009, viewed on 12 March 2022, <online: www.lexmark.com>.

Xu, J & Liu, J 2019, 'A profile clustering-based event logs repairing approach for process mining', IEEE Access, vol. 7, pp. 17872-17881.

Zelst, SJV, Fani Sani, M, Ostovar, A, Conforti, R & Rosa, ML 2018, 'Filtering spurious events from event streams of BPs', in Advanced Information Systems Engineering: Proceedings of the International Conference on Advanced Information Systems Engineering, Tallinn, 11 June 2018, Springer, Estonia.