Utrecht University

A Taxonomy of Event Log Preprocessing Techniques in Process Mining

[Master Thesis for Business Informatics]

Ying Liu (2203464)

First Supervisor: Dr. Ir. X. (Xixi) Lu
Second Supervisor: Dr. I.M. (Iris) Beerepoot
Third Supervisor: MSc. V. (Vinicius) Stein Dani

Submitted on Jul.11.2023

Acknowledgement

The end of my thesis marks the end of my two-year postgraduate journey at Utrecht University. It was definitely an unforgettable time. I went across the continent from China to the Netherlands two years ago out of curiosity for another completely different culture. When I started this program Business Informatics, I was one of the few girls and the only Asian; I frequently didn't get the jokes, but I still smiled as if I did; Sometimes I was isolated and excluded since I couldn't speak Dutch.

Even though assimilating and adapting to a new culture is never easy, I still benefit greatly from it. I've had a more comprehensive view of the world, my English has significantly improved, and I've gradually gained a lot of great friends. The previous two years have been fantastic.

I am very appreciative of Dr. Ir. X. (Xixi) Lu, my first supervisor. She not only gave me thesis advice but also taught me how to control my stress. She helped me see a more positive side of myself. Best wishes for her. I also like to express my gratitude to Dr. I.M. (Iris) Beerepoot and MSc. V. (Vinicius) Stein Dani for their support and assistance with my thesis.

Last but not least, I would also like to thank all of my friends for their support, inspiration, and company. Thank you to my family and myself. I still remember every moment of crying. Without the wonderful and outstanding you who are full of tolerance, encouragement, and strength, it would be difficult for me to get to where I am today.

Abstract

Context - Process Mining is a discipline that intersects business process management and data science. Event logs are extracted from information systems and analyzed through various techniques in data science to optimize business processes and improve user experience or process efficiency. Data preprocessing is an important part of data analysis, which often occupies most of the time in the analysis process.
Purpose - In most of the existing studies, we could see the descriptions of the event log preprocessing techniques used in specific case studies. However, only one research classified event log preprocessing techniques, but this taxonomy is not task-oriented and operation-based and does not include all preprocessing techniques. So this thesis project aims to propose a new taxonomy while exploring the relationship between the choice of data preprocessing techniques and domain, data domain, analysis purpose, and process mining task.
Method and result - A two-phase research method, systematic literature review (SLR), and taxonomy development are applied in this thesis project. A taxonomy of event log preprocessing techniques with six high-level categories and twenty low-level categories is proposed. The high-level category consists of log filtering, log transformation, log abstraction, log integration, log enriching, and log reduction. The results of SLR shows that the selection of log filtering is not data domain-dependent, whereas the choice of other log preprocessing techniques is. Data domain and analysis purpose affects the log preprocessing techniques selection, and the intended PM tasks do not have a significant impact on it.
Contribution - For scientific contribution, this research will fill in the gaps in academic research and update existing basic operations of event log preprocessing; For practical contribution, this research will provide insights to analysts in making preprocessing technique selection decisions in different task contexts.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Process mining (PM) is a discipline that intersects business process management (BPM) and data science. Process discovery, conformance checking, and process enhancement are the three types of process mining [116]. The basic element of process mining is event logs that can reflect the real process. By collecting and analyzing event logs from information systems, the actual process can be discovered and monitored. In the task process discovery, a process model is mined based on an event log, representing the activities and the ordering between them. Furthermore, fact-based insights such as deviation analysis can be gained to improve the process with the help of process mining tools and data mining techniques [116].

To complete more analysis and obtain more insights through detailed information, more and more event data is recorded in practice. With the increase in data volume, the quality of data has become a problem that needs to be solved [100]. Low-quality data can lead to inaccurate or misleading analysis results [125]. Thus, like any other data analysis, data preprocessing is necessary in process mining as well. This research focuses on event log preprocessing in process mining, aiming to give a taxonomy by using methods of systematic literature review and coding. In this chapter, the research context, research questions, and expected contributions are explained in detail.

1.1 Problem Statement

With the advancement of social technology and the popularization of digital applications, every aspect of human life generates data all the time. For example, logging in to a website, scanning a shopping credit card, booking a seat in a library, swiping a card to take a bus, etc. This massive amount of data is collected and stored, and it is considered an important asset for digital service providers. Huge amounts of data provide more possibilities for data analysis but also increase the difficulty of data processing. Several facts about data processing in reality are: 1) There are various reasons such as complex processes or user

carelessness leading to data loss and inconsistency; 2) Data is usually stored in different data tables, databases, and even different storage medium, different subsystems. The above can be summarized as there are two main reasons for conducting data preprocessing: 1) the error of the data itself, 2) the preparation for data analysis [41]. The quality of the data that was chosen for analysis has a significant impact on the results [79]. Therefore, data preprocessing is a very important and time-consuming part of data analysis, usually occupying most of the time of the entire analysis project [81].

According to different domain event log characteristics, log complexity, and expected event log processing results, analysts will adopt appropriate event log preprocessing methods, for example, log removal, log transformation, and log detection. In most of the existing studies, especially the case studies, descriptions of the event log preprocessing methods used in specific domains and conditions can be found. However, few studies have classified event log preprocessing techniques in process mining, especially those combined with business context.

In addition, in the existing literature, [74] proposed the first taxonomy of event log preprocessing techniques in process mining, but this taxonomy is not task-oriented, which refers to analysis task (purpose) orientation and PM task orientation, including process discovery, conformance checking and performance enhancement. For the data analyst, it cannot guide the analyst to make the selection decision of the preprocessing technology. Finally, we also believe that the basic operations on event log preprocessing in process mining proposed in [39] are not comprehensive and could be updated. The result of literature study can be found in Chapter 3.

1.2 Research Aim and Objectives

By defining a case notion and event classes, event data is converted into event logs [34]. We define the event log preprocessing stage that this paper focuses on as: from after event data extraction to before event log analysis. Since existing process mining tools such as Disco and ProM already support the data extraction stage and import data from different data sources [27], we regard the data extraction stage as a stage independent of event log preprocessing.

This study aims to propose a new taxonomy of event log preprocessing techniques in process mining, which is task-oriented and operation-based, providing insights to analysts in making preprocessing techniques selection decisions. The term "operation-based" refers to the preprocessing operations such as filtering, adding, converting, reducing and so on. More specifically, we aim to explore the most commonly used event log preprocessing techniques in different domain contexts, and accordingly, the focus of event log preprocessing techniques in different task contexts. In the field of process mining, we intend to find the connection between event log preprocessing techniques selection and task context by carrying out systematic literature review (SLR) and coding the literature to obtain qualitative data.

1.3 Research Question

To achieve the above objectives, we propose the following main research question:

- ***RQ*** *What is a task-oriented and operation-based taxonomy of event log preprocessing techniques in process mining?*

The term 'task-oriented' refers to analysis task (purpose) orientation and PM task orientation, including process discovery, conformance checking and process enhancement. While the term 'operation-based' refers to the preprocessing operations such as filtering, adding, converting, reducing and so on. For different concerns, following sub research questions are addressed to answer the research question from different perspectives.

- ***SRQ1*** *What is the existing taxonomy of event log preprocessing techniques in process mining?*

Understanding the characteristics of existing taxonomy and its limitations is the basis for proposing a new taxonomy. This sub research question will be answered by reviewing related work.

- ***SRQ2*** *What factors (e.g., data domain, analysis purpose...) will affect the selection of event log preprocessing techniques in process mining analysis?*
- ***SRQ3*** *Is the selection of event log preprocessing techniques data domain-dependent?*
- ***SRQ4*** *How would analysis task affect the selection of event log preprocessing techniques?*

From the domain (data domain) and task context perspective, domain characteristics will determine data characteristics, such as data scale and complexity. Is there a preprocessing technique that is widely used or generally applicable to specific domain? For different analysis task, what kind of event log preprocessing technology is widely used, why such event log preprocessing techniques are chosen, what are the factors? and how to affect? The above questions will be answered through SLR and coding analysis.

1.4 Expected Contributions

For scientific contribution, this research will fill in the gaps in academic research. After reviewing the existing literature, we only found one review for event log preprocessing techniques in process mining [74], but they did not pay attention to the domain and the task context where the technology is applied. Additionally, we found the basic event log preprocessing operations are not complete and

could be updated [39]. Therefore, this research will link event log preprocessing techniques and tasks for the first time and update the basic operations in event log preprocessing.

For practical contribution, for data analyst, this research will provide insights on event log preprocessing techniques that need to be adopted in different task contexts.

1.5 Proposal Structure

The rest of the proposal is structured as follows. In Chapter 2, the background knowledge about event log and data preprocessing is given. In Chapter 3, we summarize the existing literature review in the context of process mining, especially those addressed event log preprocessing techniques. The research methods of this study, systematic literature review and qualitative data coding, are explained in detail in Chapter 4. In Chapter 5, the coding analysis result is described. Discussion and conclusion are presented in Chapter 6 and Chapter 7, respectively.

Chapter 2

Backgrounds

This chapter explains the basic compositions of the event log and the three main tasks of process mining, namely process discovery, conformance checking, and performance enhancement.

2.1 Event Log

An event log is a start point for process mining. Each event in such a log refers to a case (i.e., a process instance), an activity (i.e., a well-defined step in some process), and a point in time [118]. For example, as a event log showed in Table 2.1, each row of data is an event, and the attribute Order No. is the unique identifier of the case, so there are 5 events and 3 cases (9901, 9902, 9903). In addition to the three basic necessary columns of case id (Order No.), activity, and timestamp, in fact, event logs may also collect and store additional information, such as resources (User and Product) and other data (Quantity) [118]. An event log can be seen as a collection of cases. A case can be seen as a trace or sequence of events. An ordered case's events can be thought of as one "play" of the process, for example, <register order, check order, ship order>.

Table 2.1: An example of event log.

Order No.	Activity	Timestamp	User	Product	Quantity
9901	register order	2022-01-22 09:15	Sara	iPhone5S	1
9902	register order	2022-01-22 09:18	Sara	iPhone5S	2
9903	register order	2022-01-22 09:27	Sara	iPhone4S	1
9901	check stock	2022-01-22 09:49	Pete	iPhone5S	1
9901	ship order	2022-01-22 10:11	Sue	iPhone5S	1

2.2 Process Mining Tasks

There are three main tasks in process mining: process discovery, conformance checking and process enhancement.

Process discovery techniques take selected event logs as input and conduct a process model describing the observed activities automatically [120]. As showed in Figure 2.1, given an event log $L_1 = [\langle a, b, c, e\rangle^{10}, \langle a, c, b, e\rangle^{5}, \langle a, d, e\rangle]$, three different process model examples can be found. The models also show frequencies. In addition to Directly-Followed Graph (DFG), Accepting Petri Net (APN) and Process Tree (PT), there are also many other process model representations or notations such as Business Process Model and Notation (BPMN) diagrams.

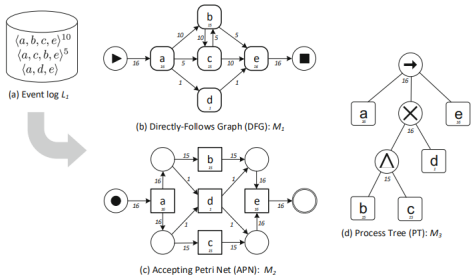


Figure 2.1: Three example process models [120].

Conformance checking compares both the actions in the process model and the events in the event log [116]. The goal of conformance checking is to find commonalities and discrepancies between the modeled behavior and the observed behavior [116]. On top of the process model, it is possible to replay the event log to look for unfavorable variances that might indicate inefficiencies. Moreover, conformance checking approaches can also be utilized for monitoring the performance of process discovery algorithms and to fix models that are not aligned well with reality [116]. There are four measurement criteria: fitness, precision, generalization, and simplicity. However, there is no such a thing as "the best process model".

Performance enhancement aims to extend or improve an existing process model using information about the actual process [119]. One type of enhancement is process extension [119], which focuses on high precision and aims to incorporate different perspectives [28]. Another type is process improvement [119], which focus on the other measurement criteria. Improvement ensure process models a) to better reflect reality, and/or b) to only enable executions that are valid from a domain standpoint and/or are connected to higher performances [28]. During this task, some tools such as Disco, Celonis and ProM can be used to visual the process model automatically and do analysis from different perspectives.

Chapter 3

Related Work

In this chapter, we summarized the existing works that focused on data and process mining methodologies, event log preprocessing techniques, and related literature review and surveys. Since this thesis aims to give a taxonomy on event log preprocessing techniques combined with task context in process mining, systematic literature review is adopted as a method and lead to results. Therefore, chapter 3.3 mainly focus on related literature review by other researchers rather than papers that described event log preprocessing in case studies.

3.1 Data and Process Mining Methodologies

Over the last decades, many researchers proposed data mining and process mining methodologies to carry out related tasks and applications. We could see that data preprocessing is one necessary part of them. In the field data mining, we mainly introduce three main popular methodologies: KDD, CRISP-DM and SEMMA. The Knowledge Discovery Databases (KDD) model is an iterative and interactive model aiming to extract knowledge from data. It has five steps in total, which are (1) selection, (2) preprocessing, (3) transformation, (4) data mining, (5) interpretation/evaluation [42]. On top of it, CRISP-DM (CRoss Industry Standard Process for Data Mining) was built. It is a comprehensive process model for carrying out data mining projects and includes six phases: (1) business understanding, (2) data understanding, (3) data preparation, (4) modeling, (5) evaluation, and (6) deployment [123]. Meanwhile, SAS institute developed the SEMMA (Sample, Explore, Modify, Model, Assess) methodology consisting of five phases: (1) sample, (2) explore, (3) modify, (4) model, and (5) assess [7].

In order to compare the above three methodologies more clearly and address the data preprocessing stage, Table 3.1 aligns the stages of the three methodologies. We can see that data preprocessing is an independent stage in KDD. In methodologies CRISP-DM and SEMMA, data preprocessing is included in data preparation and modify respectively. In SEMMA, the modify stage espe-

cially points out that modification is achieved through creating, selecting, and transforming the variables.

Table 3.1: Alignment of three methodologies KDD, CRISP-DM and SEMMA in data mining.

Model	KDD	CRISP-DM	SEMMA
Phases		business understanding	
		data understanding	sample
			explore
	selection	*data preparation*	*modify*
	preprocessing		
	transformation		
	data mining	modeling	model
	interpretation/evaluation	evaluation	access
		deployment	

However, all of the methodologies mentioned above are high level and independent of both the industry applied and the technology used, thus provided little support in practice [117, 123]. Moreover, they are generally for data mining projects rather than process mining projects.

Wil M.P. van der Aalst [117] proposed L^* life-cycle model, as seen in Figure 3.1. A life cycle of a typical process mining is described by this five stage model. The goal of this model is improving lasagna process (a structured process) specially. A structured process is one in which each activity is repeatable and has clear inputs and outputs, whereas specifying the prerequisites and outcomes of activities in an unstructured process can be challenging [117].

There are five stages in L^* life-cycle model: plan and justify (Stage 0), extract (Stage 1), create control-flow model and connect event log (Stage 2), create integrated process model (Stage 3), and operational support (Stage 4). All preparatory work before the official start of the project is included in stage0, such as time and activity schedule, milestones definition, monitoring standards.

Specially, in stage 1 *extract*, event data, models, objectives, and well as questions will be extracted from systems and stakeholders. Van der Aalst [117] indicated that finding the relevant data and scoping these data will be paid the most efforts. Also, the requirements that event logs need to satisfied to do data extractions are listed: be ordered in time and correlated. The stage 2 aims to product a control-flow model closely linked to the event log. In stage 3, additional perspectives are added to enhance the control-flow model. Both the outputs of stage 2 and 3 can be used to answer questions listed in stage 1 and take appropriate actions. After getting the integrated process model, three operational support activities of stage 4 of the L^* life-cycle model will be considered, which are detect, predict, and recommend.

Another methodology to guide the execution of process mining projects is PM^2 (a Process Mining Project Methodology), aiming to improve process performance or compliance to rules and regulations [34]. Compared to L^* life-cycle

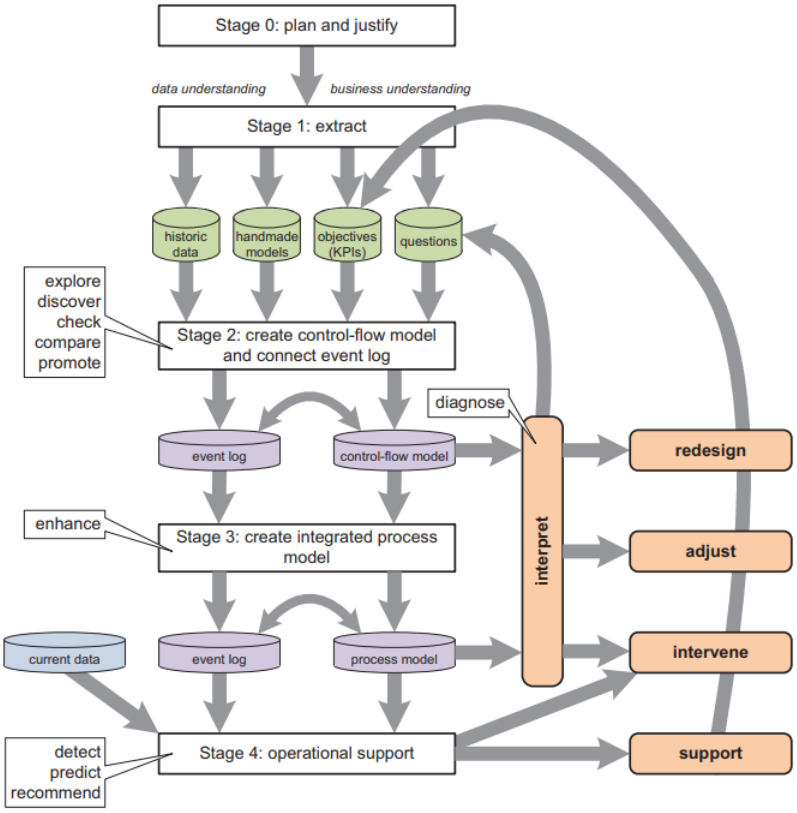


Figure 3.1: L^* life-cycle model [117].

model, PM^2 is more applicable for large and complex projects because its covering of numbers of process mining and other analysis techniques. Moreover, PM^2 does not only aim at structured process, but also unstructured process. There are six stages in PM^2 methodology: (1) planning, (2) extraction, (3) data processing, (4) mining & analysis, (5) evaluation, and (6)process improvement & support. An overview of the methodology see Figure 3.2.

Similar to the L^* life-cycle model, the PM^2 methodology also includes the stage *extract*, but the PM^2 methodology makes a more specific identification for this stage. Three activities are listed in stage 2 *extract*: determining scope, extracting event data, and transferring process knowledge. Firstly, the scope of data extraction is determined. Then, event data is created by collecting the selected process and joining them into a single collection of events. Lastly, process knowledge such as written process documentation or handmade process models is made and shared to drive after-stages more effectively.

Event data extraction is considered as an independent stage in both two

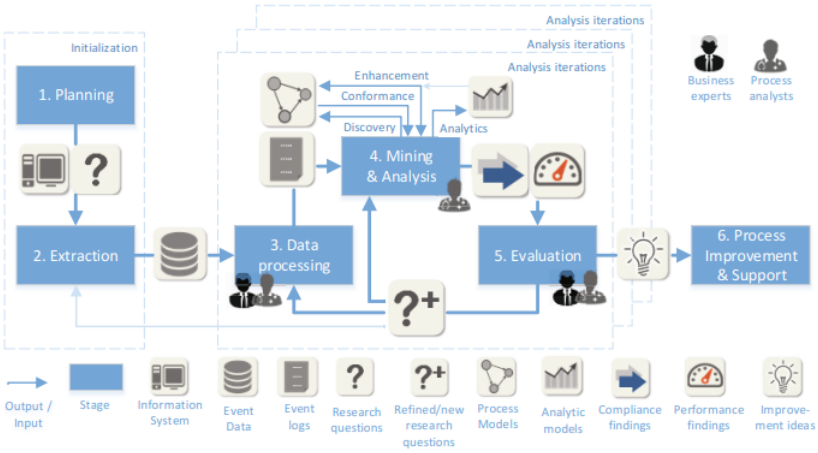


Figure 3.2: An overview of PM^2 [34].

process mining methodologies. We define that event log prepossessing stage does not include data extraction in the context of this paper. The event log preprocessing stage defined in this paper corresponds to the third stage of data processing in the PM^2, that is, after data extraction and before data mining and analysis. The input to the event log preprocessing stage is event log that usually extracted from different data sources, and the output is event logs that are ready for analysis.

3.2 Data and Event Logs Preprocessing

In data mining and process mining methodologies discussed above, data preprocessing operations and techniques are highly abstracted in specific stages. This subsection introduces them in detail. Moreover, through an initial comparison of data preprocessing and event log preprocessing techniques, we found that there are certain differences between the two, which is mainly caused by the characteristics of event logs.

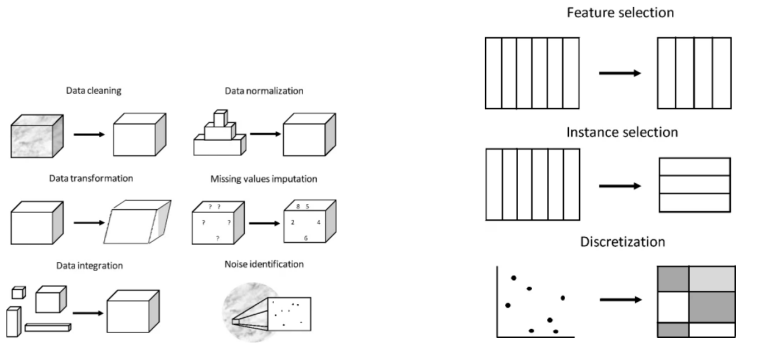
3.2.1 Data Preprocessing

In [44], data preprocessing techniques are divided into two disciplines, data preparation and data reduction. Data preparation includes data cleaning, data normalization, data transformation, missing values imputation, data integration, noise identification, while data reduction includes feature selection, instance selection and discretization, as seen in Figure 3.3. General explanations of these techniques are listed:

- **Data cleaning:** usually consists of two stages, qualitative error detection and error repairing, aiming to solve data quality problems, such as duplicates, missing values, integrity constraints violations, and outliers [24].
- **Data normalization:** A method of converting the data to a particular range, such as 0 to 1. When there are significant discrepancies between the ranges of several attributes, it is necessary [86].
- **Data transformation:** Data transformations are the application of a mathematical modification to the values of a variable [82].
- **Data integration:** The challenge of merging data from several sources and giving the user a consistent view of these data is known as data integration [68].
- **Missing values imputation:** One method is to discard cases that having a missing value. Another example is statistics-based method that sample the approximate probabilistic models to fill in the missing variables by using maximum likelihood approaches [44].
- **Noise identification:** Mainly consists of two methods: (1) Data polishing: to correct the noise especially if it interferes with an instance's labeling, (2) Noise filtering: identify and remove the noisy instances [44].
- **Feature selection:** The process of identifying and removing as much irrelevant and redundant information as possible [51], aiming to extract a subset of the original problem's attributes and still adequately describes it [49].
- **Instance selection:** Selecting appropriate instances from a very large number of instances so that they can be prepared as input for a data mining algorithm [44].
- **Discretization:** Through the division of the numerical features into a finite number of non-overlapping intervals, it converts quantitative data into qualitative data. Every numerical value is mapped to each interval by using the boundaries, thus making the values discrete [44]. Benefits are data simplification and reduction, and readability, but may have information loss [70].

However, within the data mining technology classification above, we could see some overlap between the categories. For instance, *noise identification* can be considered as one part of *data cleaning*. *Discretization* can reduce the complexity of the data analysis process, but it will not reduce the amount of raw data. At the same time, both *discretization* and *data normalization* can be regarded as a way of *data transformation*.

Han et al. han2022data defined a clearer and more concise classification of data preprocessing techniques in data mining, which consists of four categories:



(a) Data Preparation Tasks (b) Data Reduction Approaches

Figure 3.3: Data Preprocessing Tasks [44].

- **Data cleaning:** It is considered as a process. It focus on handling missing values, identifying noise or outliers, and repairing errors.
- **Data integration:** It includes data join and match, data consistency check, and redundancy removal.
- **Data transformation:** It mainly includes normalization, discretization, data compression and sampling.
- **Data reduction:** data reduction can be divided into dimensionality reduction such as attribute subset selection, and numerosity reduction such as data partitioning, clustering, and data cube aggregation. Data compression and sampling methods mentioned in data transformation are also common data reduction techniques depending on data processing purpose.

Since event logs are still a type of data in essence, the above high-level classification of preprocessing technology in data mining provides a powerful reference for the taxonomy of process mining preprocessing technology we intend to build. But it is not exactly equal to the preprocessing technique in process mining. The log emphasizes the time sequence of events to generate a business process model in actual operation, so there are preprocessing technologies for specific attributes such as timestamp, which are not specifically highlighted in the general preprocessing technology in data mining. The discipline of process mining combines business process management and data science, so the preprocessing of logs has its own special features that distinguish it from data preprocessing in data science.

3.2.2 Event Logs Preprocessing

As the specific data analysis in the context of process mining, event log analysis requires data preprocessing steps similar to any other type of data analysis

[39]. First of all, in reality, event logs is not recorded in perfect way either. An example is that due to possible mistakes in the iterative update process of products and codes, the values in a certain attribute may be confusing and inconsistent. For instance, before a certain moment, the values stored in the gender attribute were "female", "male", "others", but after that, it were "0", "1", "2". Additionally, event logs can come from a variety of sources, including a database system, a comma-separated values (CSV) file or spreadsheet, a transaction log, an ERP system (SAP, Oracle, etc.), a business suite, a message log (from IBM middle-ware, for example), an open API that provides data from websites or social media, and more. After data extraction to create the event log, the data quality is not sufficient enough to permit direct analysis in most cases. Therefore, the event log is necessary to be preprocessed.

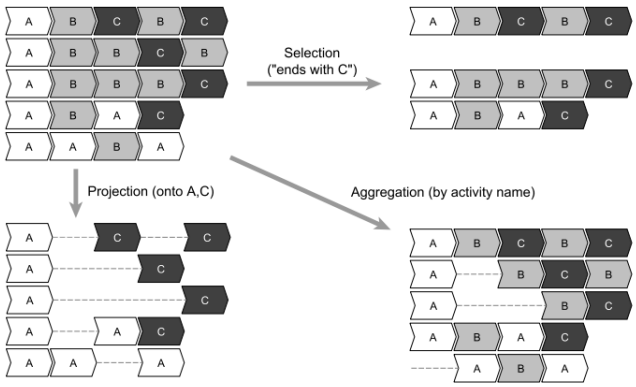


Figure 3.4: An example to explain three basic preprocessing operations on event logs [39].

Fahland Fahland2022 indicated that there are three basic preprocessing operations on event logs, which are *selection*, *projection* and *aggregation*. The majority of additional event log preprocessing techniques can be thought of as a combination of these three basic operations. As shown in Figure 3.4, *selection* refers to selecting a subset of traces that meet specific requirements in the original trace set, for example, selection filters out all traces that end with event C. *Projection* refers to removing events that do not meet specific requirements from all traces, for example, all event B is removed in Figure 3.4. *Aggregation* refers to synthesizing and replacing consecutive and repeated events in all traces with a single event. In Figure 3.4, the continuous event B is synthesized into a single event B.

By corresponding these three basic event data preprocessing operations [39] with the four traditional data mining technology classifications [52], as shown in Table 3.2, we could see that indeed each type of data preprocessing technology

contains at least one basic operation. However, some techniques cannot be described by these three basic operations. For example, data cleaning focus on handling missing values, identifying noise or outliers, and repairing errors, but repairing errors in data cleaning is neither selection, nor projection, nor aggregation.

In addition, the traditional classification of data preprocessing techniques is not complete enough. For example, enriching logs is a relatively common processing method in log data preprocessing. However, no matter in data cleaning, data integration, reduction, or transformation, enriching data is not included. [34] discussed two ways of enriching logs, which are (1) extrapolating or computing new events and data attributes from the log itself, and (2) adding external attribute. The difference between enriching data and data integration is that enriching data is operated based on a single event log file such as a csv file, which essentially adds additional information. However, data integration is to integrate two or more files, such as join or match, and essentially does not add more attributes. Therefore, for process mining, the taxonomy of log preprocessing techniques is a further discussion on the taxonomy of data mining preprocessing technology. At the same time, the definition of basic operations by [39] also has room for improvement and update.

Table 3.2: Alignment of basic event log preprocessing operations and data preprocessing high-level techniques.

Data preprocessing high-level techniques by [52]	Basic event log preprocessing operations by [39]		
	selection	projection	aggregation
Data cleaning	X	X	X
Data transformation	X		X
Data integration	X		
Data reduction	X	X	X

3.3 Literature Review and Surveys

In the existing literature, there are already some literature reviews and surveys in the context of process mining. We introduce and summarize them in this subsection. By discussing existing reviews, the preliminary taxonomy result is purposed to provide coding basement for phase 2.

3.3.1 Literature Review in Event Log Preprocessing

Marin-Castro and Tello-Leal [74] reviewed 70 related papers that were published from year 2005 to 2020 and explicitly mentioned the event logs preprocessing or cleaning. As shown in Figure 3.5, this literature review grouped preprocessing techniques into transformation techniques and detection-visualization

techniques. Transformation techniques mark modifications made towards the original structure of event log. While the events or traces that can lead to issues with data quality are identified, grouped, and isolated using detection-visualization techniques.

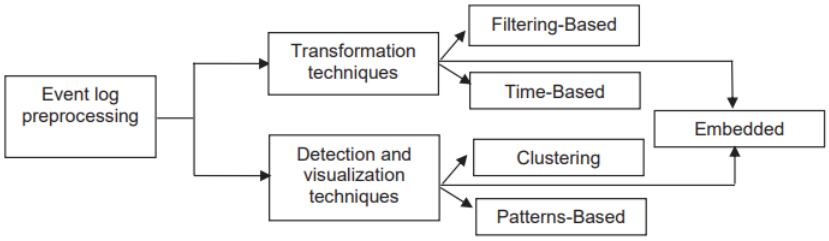


Figure 3.5: A grouping way for event log preprocessing in process mining [74].

It is worth mentioning that this review is the first literature review about the main approaches applied in data preprocessing for process mining. In this review, the results were conducted according to the characteristics of the data preprocessing techniques, such as time-based or filtering-based, or pattern-based. However, it did not mention in which business context or domain these techniques were used. This is exactly where our research will focus on.

Our study differs in the following aspects. Firstly, in different domains, (1) the obtained logs present different characteristics, and, (2) there are specific purpose analyzes in specific domains. The nature of data itself and the purpose of data analysis will affect the choice of data preprocessing technology. Therefore, we propose that domains are related to data analysis and preprocessing techniques, and add the aspect of domain to the literature review in the second stage.

In addition, from the perspective of analysts, this taxonomy is not readable and not task-oriented. It fails to guide analysts in making choices about which preprocessing technique to use when performing specific event log preprocessing tasks. For business-oriented analysts, this taxonomy is too technical, and they may not even be able to accurately distinguish whether a certain preprocessing technology is pattern-based or filtering-based.

Another difference between our research and this review is that we will include more articles, for example, those published in the years 2021 and 2022. In summary, we expect to propose a new taxonomy. It will combine the data preprocessing taxonomy in traditional data mining with the log analysis features in process mining, and is task-oriented and analyst-friendly, that is to say, it can support and guide data analysts in making event log preprocessing technology selection decisions.

A literature review and taxonomy on event abstraction, which can be considered as a part of event logs preprocessing, are conducted in [122]. This paper

focused on the abstraction from fine-granular events to coarse-granular events. They only include papers that addressed an underlying model. Papers that focused on event correlation and potential techniques that described incompletely were excluded. Finally, 28 publications published from years 2009 to 2020 were reviewed. As a result, a taxonomy with nine dimensions was conducted, those are (1)Supervision strategy, (2)Unsupervised techniques, (3)Supervised techniques, (4)Fine-granular event interleaving, (5)Probabilistic nature of outcome, (6)Data nature, (7)Alternative perspectives, (8)Relation between event classes and activity classes, and (9)Relation between event instances and activity instances. For the similarities and differences between data abstraction and data preprocessing, [122] also addressed that some data preprocessing techniques do not aim to change data granularity level and do aim to find (in)frequent patterns that can lead to data preprocessing basic operations instead. A suggestion that the combination of data abstraction techniques with other data preprocessing techniques might lead to a better result was given in this review. Corresponding to our research, we consider event log abstraction as an indispensable part of preprocessing stage and add log abstraction as a high-level category when coding papers. Also, still, like the event log preprocessing review mentioned above, the data abstraction taxonomy in this paper does not mention domain or industry and we will consider them as one important dimension.

3.3.2 Literature Review in Process Mining

In the field process mining generally, there are also many literature reviews focused on specific domain. We mainly introduce related review in healthcare, business and education domains here. In the healthcare domain, Rojas et al. Rojas2016 conducted a literature review which covered 74 publications. Eleven main aspects, which are: process and data types; frequently posed questions; process mining techniques, perspectives and tools; methodologies; implementation and analysis strategies; geographical analysis; and medical fields, were concerned to analyse those publications. They concluded that due to the noisy and incomplete data characteristics and the low-structured process characteristics in the healthcare domain, the techniques or algorithms Trace Clustering, Fuzzy Miner and Heuristics Miner are often applied.

In the same year 2016, Ghasemi and Amyot Ghasemi2016 conducted a systematised literature review (SdLR) on the domain healthcare that was based on existing eleven literature reviews rather than a wide collection of original sources. Therefore, a Systematic Literature Review (SLR) is more complete and rigorous than a SdLR [60]. However, Ghasemi and Amyot indicated that this SdLR was more healthcare-specific than the majority of other assessments and offered more precise information regarding the state of process mining in the healthcare industry. Generally speaking, this SdLR is not updated enough, and only reviewed the latest literature up to 2016. Even a future work mentioned in this article is conducting a systematic literature review of process mining in the healthcare, which has been done in the study mentioned above.

Erdogan and Tarhan [36] presented a systematic mapping study, covering

172 studies published between 2005 and 2017. These studies were classified by several attributes, for example, application context, healthcare specialization, mining activity and algorithm, process modeling type and notation. It showed that healthcare process mining is expanding quickly and open for more researches. Especially for process mining in oncology, [66] analysed 37 papers to conduct a thematic review. The potentiality of process mining for enhancing cancer treatment procedures was highlighted.

In business process mining, Dakic et al.[26] conducted a literature review focusing on applications of business process mining in industry from year 2009 to 2018. They discussed the process mining tasks, industries, tools and algorithms in the selected 36 papers. The most noteworthy finding is that process mining is always a very profitable discipline, applicable in multiple industries and on various process types. Eggers and Hein [35] analysed 58 papers to contribute to the field of business value realization from process mining. In the context of organization, organizational practices and organization's capabilities for process mining influence each other. They proposed that the development of technological capabilities is amplified by realized value potentials, and values realized from process mining will have an impact on how organizational practices are developed going forward.

In educational process mining, event logs specifically from educational environments such as students' online learning activities are collected. [16] conducted a survey on educational process mining and introduced the representation models, techniques, applications. By using raw event data, educational process mining enables a deeper comprehension of the underlying educational process. It has lots of applications already, for example, computer-supported collaborative learning, computer-based assessment, student registration, and 3D educational virtual worlds.

Corresponding to our research project, the papers included in our research project will be cross-domain rather than focusing on a certain domain. Observing the relationship between domains and data characteristics, general event log preprocessing expectations will be an important aspect of our research.

Chapter 4

Research Methods

A two-step approach is followed to answer research questions. Firstly, a systematic literature is conducted. Based on selected papers which are the results of step one, paper coding definition is applied to product qualitative data in step 2. As a result of step 2, a taxonomy of log preprocessing techniques for process mining can be drawn. Both two steps are described in detail in this section.

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted, as seen Appendix A. It classified this research as low-risk with no further ethics review or privacy assessment required.

4.1 Systematic Literature Review

Literature review is a scientific method to test a specific hypothesis and/or develop new theories through summarizing, analyzing, and synthesizing a collection of related material [85, 126]. There are mainly three processes in a successful literature review, which are planning, conducting, and reporting the review [17], as shown in Figure 4.1. The first stage includes two steps that are formulate the problem, and develop and validate the review protocol. The following majority steps are presented in the second stage, conducting the review. The results of this stage will lead to the last stage reporting findings.

To plan the review and get the intended paper pool, on the website Scopus (https://www.scopus.com/), keywords ("process mining") AND ("case study" OR "case studies") were used to search within article title, abstract and keywords. 4565 documents could be found on the date of Dec.20.2022. Next, in order to narrow the scope of the review, only papers meeting the inclusion criteria listed below were selected.

- Published in 2021, 2022, 2023
- Published in conference, journals and under review
- Explicitly mentioned "process mining" in keywords

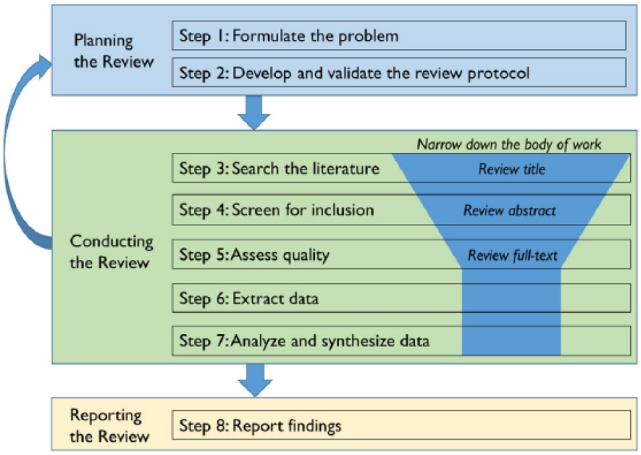


Figure 4.1: Process of systematic literature review [126].

- And in English

As a result, 355 papers were available. Since this thesis focuses on log preprocessing techniques, it is not only required to include case study in the aiming paper, but also requires the application of real data in the specific case study. Therefore, we read the abstracts of all 355 papers, and we filtered the articles that did not mention collecting data from the real world or using public datasets for case studies. Finally, 159 papers were left to do the following step paper screen.

The 159 papers were downloaded in pdf format and imported into the software Nvivo. During the coding process, the papers that did not mention the data preprocessing steps in the text were screened out, and finally 86 papers were evaluated as valid papers. The complete paper screening process refers to Figure 4.2.

4.2 Coding and Taxonomy Derivation

The following codes were defined to code papers and then obtain qualitative data.

- **Domain:** The domain of the *institution or organization* conducting the case study. The following table 4.1 shows the four most mentioned domains and their definitions in the paper coding stage.

  For example, if the event log comes from a factory that produces medicines or ventilators, then the domain of that paper will be defined as manufacturing instead of healthcare, because the organization that generates the data is a manufacturing factory.

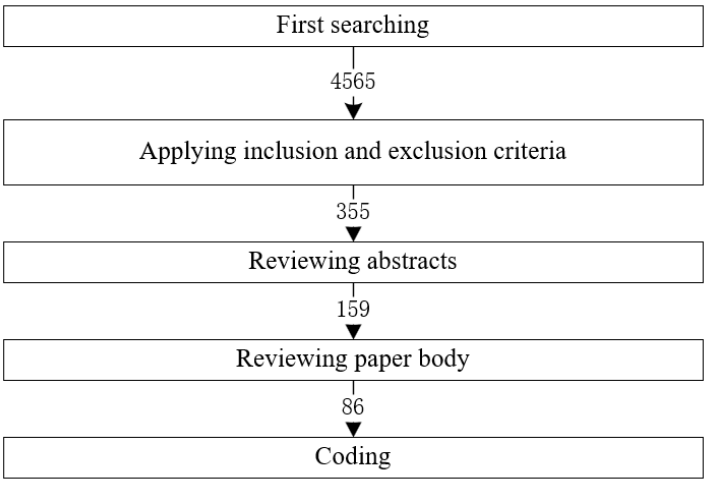


Figure 4.2: Paper screen process.

Table 4.1: Four common domains and their definitions.

Domain	**Definition and example**
Healthcare	Systems or institutions established to meet human health needs [59], for example, hospitals and electronic health (eHealth) information systems.
Education	Systems or institutions that focused on e-learning or online learning [45].
Manufacturing	Factories and information systems that support manufacturing activities.
Finance	Financial institution including insurance companies, and information systems that support financial activities.

- **Data domain:** The type of process that collected log represents. For example, loan application process, patient journeys in a hospital, and course evaluation process.
- **Data quality issue:** Data characteristics that create data quality issues. For example, redundancy, missing data, inconsistent data.
- **High-level category:** Six high-level categories were defined based on related work. It consists of log integration, log reduction, log abstraction, log filtering, log enriching and log transformation. The details are explained in Chapter 5, especially in Figure 5.1.
- **Low-level category:** Subdivision of high-level category. In the process of coding, the value of low-level category is gradually revealed and sum-

marized. Also see Figure 5.1 and more explanations are in Chapter 5.

- **Aiming data level:** The data level that the preprocessing technique works on. It includes cases, events, relationships, case attributes, positions, activity names, timestamps, resources, and event attributes.
- **Analysis purpose:** The purpose of log analysis in case study. For example, aiming to analyse student behavior by analysing online learning logs, or aiming to discover human habits by analysing IoT logs in a smart house.
- **Preprocessing purpose:** The purpose of a specific preprocessing step. For example, converting log format from CSV to XES is to suit PM tools.
- **PM task:** Process discovery, conformance checking, performance enhancement and prediction.
- **Preprocessing tool:** The tool used to perform preprocessing operations. For example, Disco, Celonis, ProM and even non-PM tools.
- **Analysis tool:** The tool used to analyse logs. For example, Disco, Celonis, ProM and even non-PM tools.
- **Year:** Paper publication year.
- **Country:** The country in which the first author was located.

Chapter 5

Result

In this chapter, we elaborate on the analysis of the results of coding 86 papers, including the taxonomy of complete log preprocessing techniques, and the relationship between domain, data domain, analysis purpose, PM tasks, and event log preprocessing techniques.

5.1 Taxonomy of log preprocessing techniques

Through the discussion of the references in Chapter 3, we propose a high-level taxonomy of event log preprocessing techniques to provide basis for further coding and complete taxonomy derivation, as shown in Figure 5.1. The solid boxes in the middle indicate the six preliminary high-level categories, including **log integration**, **log reduction**, **log abstraction**, **log filtering**, **log enriching**, and **log transformation**. The blue blocks in Figure 5.1 are the corresponding low-level categories. Its sources of reference are listed on the left and right, with each color representing one resource. Double arrows link the sources and preliminary categories. The low-level categories under each high-level category are described separately in subsections.

5.1.1 Log filtering

Log filtering is the most commonly used category of preprocessing techniques, with 55 out of 86 papers applying this technique. In these 55 papers, we have read a lot of nouns or adjectives describing data quality, such as noise, outliers, redundant, duplicate, missing values, useless values, blank values, irrelevant values and so on. Filtering itself is a simple operation, but the filtered objects are complex and diverse. According to the characteristics of the filtered object, the category log filtering is subdivided into 9 detailed low-level categories, see Figure 5.2. In the figure, the numbers in parentheses represent the number of papers that used the techniques in the specific category. Since a single paper may filter out many types of data, such as filtering out incomplete data and

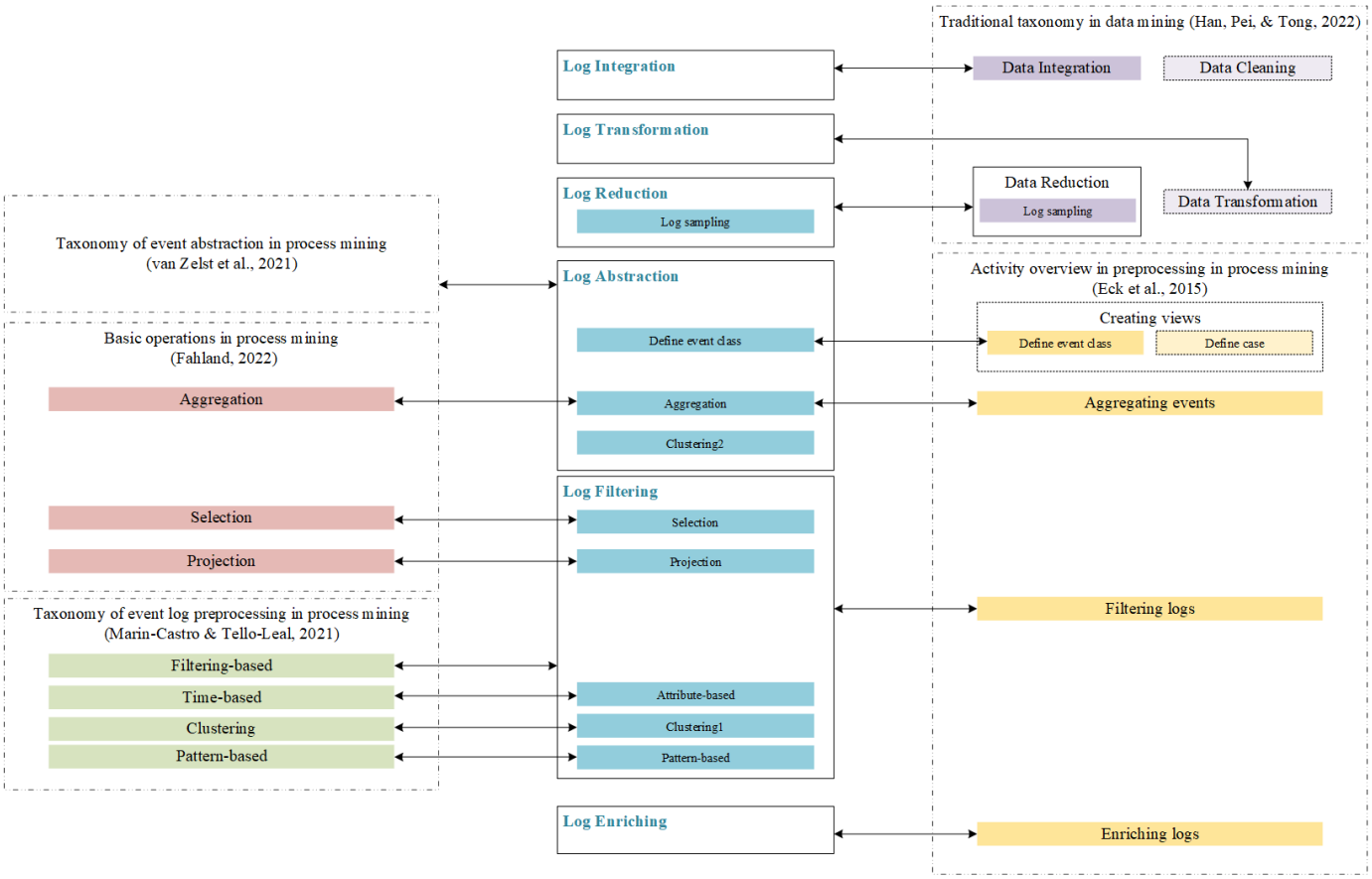


Figure 5.1: Preliminary result of high-level taxonomy of event log preprocessing techniques.

infrequent data at the same time, the sum of the numbers followed by low-level categories is greater than 55.

Filtering irrelevant data

The definitions of "irrelevant" in Collins Dictionary [1] are "if you describe something such as a fact or remark as irrelevant, you mean that it is not connected with what you are discussing or dealing with" and "If you say that something is irrelevant, you mean that it is not important in a situation". Therefore, in the data analysis context of process mining, we define irrelevant data as "Those resources, activities, attributes, events, traces that are not concerned or not important in the specific analysis".

Whether the data is relevant to the analysis task is determined by experts based on domain knowledge and analysis requirements. For example, in [19], the

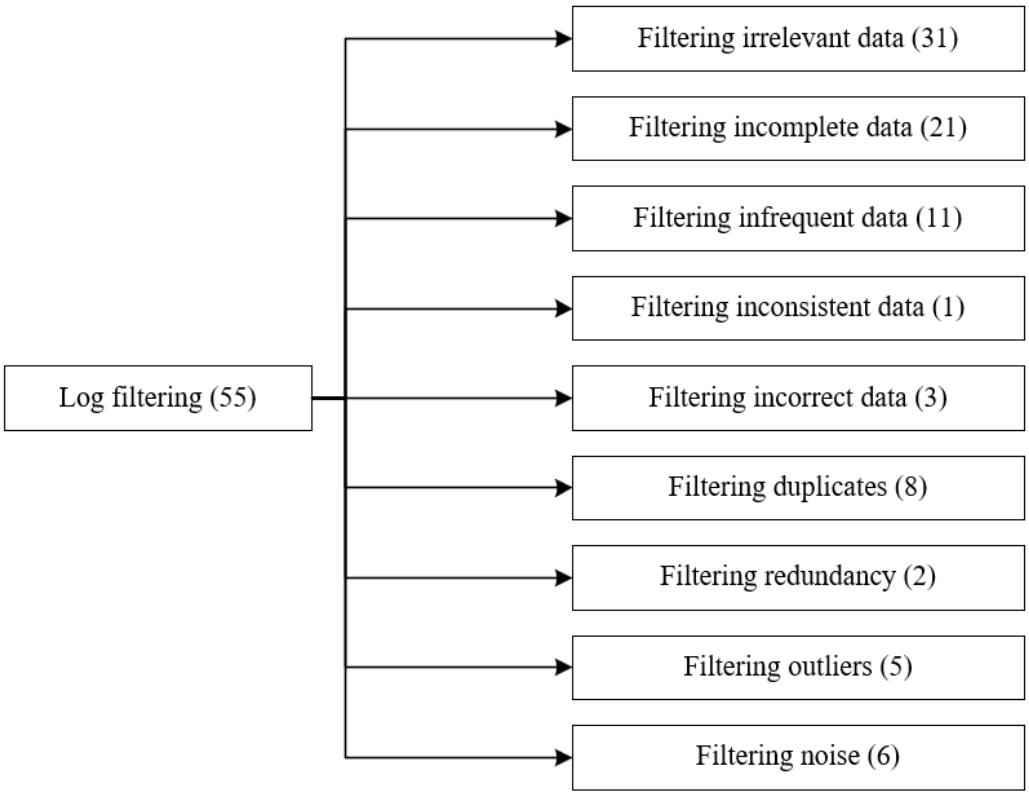


Figure 5.2: Taxonomy of the high-level category log filtering.

analysis only focused on the students who participated in the class (resource), so the events generated by other resources were defined as irrelevant data and filtered. The paper [46] intended to analyze the activities of PHD students and improve their journeys, so a filtering condition was given after discussion by analysts and stakeholders, that is, only kept the traces of full-time students who completed PHD and withdrew (case status).

The term "useless data" is also used in many papers to describe irrelevant data. For example, "filtering useless information such as links and marker symbols" [107], since the links and marker symbols (attributes) cannot make any contribution to the intended analysis and are regarded as useless data, so these data can also be considered irrelevant to this analysis.

Filtering incomplete data

Incomplete data can be subdivided into incomplete attribute value in event and incomplete case.

The incompleteness of event is usually described as missing values. Incomplete events include missing case id [67], missing timestamps [31, 46, 84], and missing activities [23, 31], missing other attribute values that are relevant to this analysis [43].

The incompleteness of case is usually described as unrepresentative data. It means the lack of event, for example, "remove any record that may create only one event per case as it will not depict the sequence of activities and hinder the performance analysis of the model" [89] and "removing cases that did not cover the whole steps" [23]. A case is sequence of events. Cases containing only one event or less than a complete trace cannot represent the real-life process and are therefore removed during the data preprocessing stage.

Filtering infrequent data

Infrequent data is the case with a low frequency of occurrence from the obtained log. Filtering infrequent data is to "prevent the PM tool from returning incomprehensible or inaccurate results" [97], and "to improve the quality of results, and to avoid low precision and highly complex results" [110].

Filtering inconsistent data

A simple example of inconsistent data is values in different format,"2023-01-01" and "2023/01/01", in the attribute timestamps. This inconsistency in data format may be due to errors caused by manual input, or it may be that different data sources have different definitions of data formats. Inconsistent event labels make it difficult to assign clear semantics to the activities of a discovered process model [116], and may also bring about a dimensional explosion of the process model.

Filtering incorrect data

Incorrect data is wrong or unreliable data that violates the logic of reality. For example, in the real process, activity A should be executed earlier than activity B, but in the log, the timestamp of A in a specific case is later than activity B [84].

Filtering duplicates

Duplicates refers to repeated data. In process mining, the case id needs to be a unique identifier, and the data represented by different case ids must be different, so as to ensure the accuracy of the data. However, in real life, duplicate data is usually generated due to system bugs or other reasons. For example, in [30], repeated events with the same Call ID were excluded.

Filtering redundancy

Only two papers mentioned redundancy [22, 23]. In [23], redundant events were included in data error. "we conducted some data preprocessing, including handling data error (e.g., removing redundant events and eliminating multiple yield values)", while there was no further definition and explanation in [22].

Filtering outliers

In [14, 15, 69], the papers only mentioned "removing outliers" without any explanation or definition, while in [21], "we noticed the existence of outliers, i.e., cases that take too long, or incomplete", too long trace and incomplete data are considered outliers; in [96], "if lecture activities in the short semester are included, it will be an outlier because it has activities that are far more than short than activities in the semester in general", traces that are too short are also considered outliers.

Han et al. [52] defined outlier as follows: "Assume that a given statistical process is used to generate a set of data objects. An outlier is a data object that deviates significantly from the rest of the objects, as if it were generated by a different mechanism".

We also argue that outliers in process mining can be subsumed into low frequency data. A data that is obviously different from the others is also obviously infrequent data.

Filtering noise

Noise is an overused word. People even usually describe any data that is not conducive to the analysis task as noise. An interesting point is that among the 86 papers, more than one paper mentioned noise, but only one paper described what is noise and how to filter it, "In the original log the noisy activities were conveniently named as "Noise", so they were removed using a filter on the activity name [29]".

Han et al. [52] provided a definition of noise, "Noise is a random error or variance in a measured variable". A noise can be irrelevant data, incomplete data, infrequent data, inconsistent data and any data that analyst considers noisy.

5.1.2 Log transformation

The techniques in the category of log transformation were employed in 38 of the 86 papers. According to the objects being transformed, this category is separated into four subcategories: transforming format, transforming values, reordering, and transition matrices and encoding, see Figure 5.3.

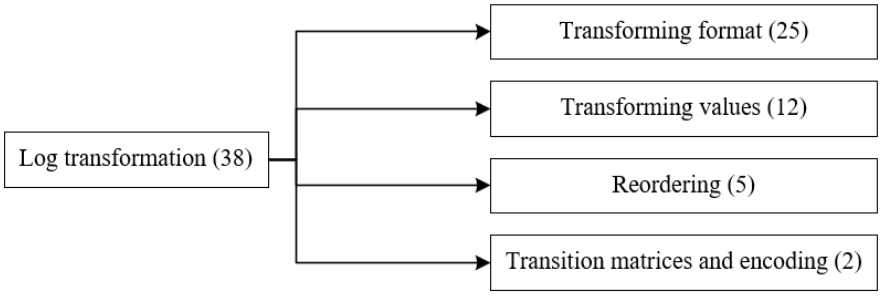


Figure 5.3: Taxonomy of the high-level category log transformation.

Transforming format

Among the format transformation, the transformation of the log format from CSV to XES was mentioned the most (14/25 papers), so that the logs can be used in PM tool. Because the log format after extraction is usually CSV, and PM tool requires the log format to be imported as XES.

The remaining format transformation is to determine which columns are key columns (such as case ID, activity name and timestamps) after importing the log into PM tools.

Transforming values

The difference between transforming values and transforming format is that transforming values means the change of specific value in event. For example, replace infrequent values with the value 'other' to avoid dimension explosion, replace missing values, replace NaN values with 'zero', capture data, and encrypt.

Reordering

Reordering is the process of sorting the log by a particular timestamp. When the original log is out of order, it is essential to reorder it so that the process model displays the activities' proper execution sequence.

Transition matrices and encoding

In particular, transition matrices and encoding are used in prediction tasks. This thesis just lists it as a different category without providing any further justification.

5.1.3 Log enriching

In 16 out of 86 papers, the log-enriching techniques were applied. Log enriching is split into the following five categories, adding calculation metrics, labelling, adding case id, and adding noise, in accordance with the various properties of the data added on the basis of the original log, as shown in Figure 5.4.

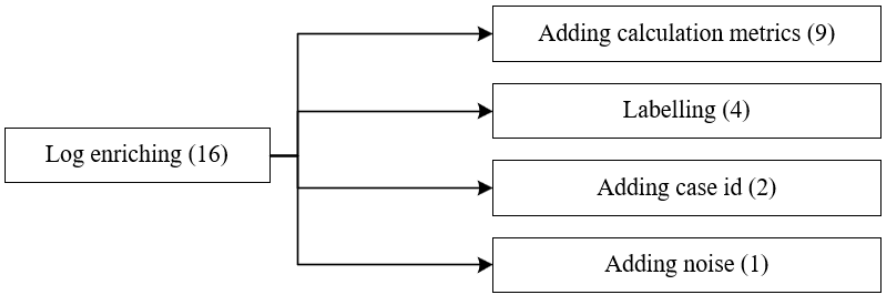


Figure 5.4: Taxonomy of the high-level category log enriching.

Adding calculation metrics

In this low-level category, the calculation metrics are computed from existing attributes in the log. For example, in the paper [30], call center processes of a company was examined. In the original event log, each call only had attributes Start and Call Duration, but process analysis expected the end time of the call. Therefore, the attribute End was obtained by adding Call Duration to Start, see Figure 5.5.

Labelling

Labeling is giving event or trace a state value or class. In the paper [108], "the cases are labelled as either successful or failed, depending on how they have been executed and their outcome", to further divide the log into two logs. In the paper [83], for recording differences over time between the intended operation and the actual execution, one label was developed to determine if the event is carried out on time or not.

Adding case id

Case id is a unique attribute in event logs. The data collected in some case studies did not have the attribute of case id, then the case id was created artificially in the data preprocessing stage. For example, in [106], "the caseid is created by combining the three-digit client number (MANDT) with a ten-digit document number and a five-digit item number".

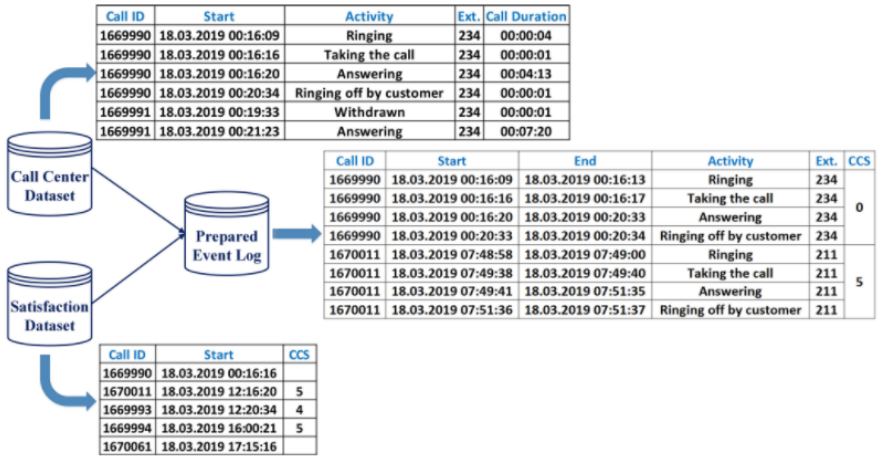


Figure 5.5: An example of adding calculation [30].

Adding noise

The case of "adding noise" is quite special, and just one publication described it. The paper [104] evaluated privacy assurance of healthcare metadata. Noise-adding plugins in the tool ProM were used to make the original event logs more privacy-preserving [78].

5.1.4 Log reduction

In the 86 papers, 12 papers used log reduction to do log preprocessing. The techniques for log reduction can be summarized into 3 low-level categories, which are dividing into sub-logs, sampling and cutting traces, as shown in Figure 5.6. Examples of 3 different log reduction operations see Figure 5.7.

Dividing into sub-logs

In the example Figure 5.7, the original log is divided into two logs by the date in timestamp. In [29, 38], IoT logs were collected in a smart house and the aim was to explore human habits. They firstly divided logs into smaller pieces by timestamps to analyse the time distribution of the activities (user habits) within a day [29].

Resource could also be a common attribute for division. The paper [92] divided the traces into subsets to model different profiles of users. Dividing original logs according to specific attributes is usually for more in-depth analysis [46].

In addition, in order to test the proposed algorithm or approach, the log was divided into training data and test data according to a certain proportion [23, 53].

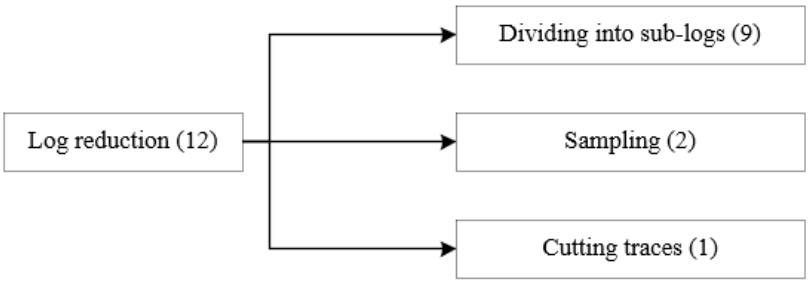


Figure 5.6: Taxonomy of the high-level category log reduction.

Sampling

The most notable characteristic of sampling is randomness. The reduction here is to reduce the trace, that is to say, the data processing needs to be in the unit of trace. In the example in Figure 5.7, there are 4 traces, $[\langle A, B, C, D, E\rangle, \langle A\rangle, \langle A, C\rangle, \langle B\rangle]$. After randomly sampling 50% of the traces, the log $[\langle A\rangle, \langle A, C\rangle]$ in the lower right corner is obtained.

Cutting traces

In the example in Figure 5.7, compared to other traces, the trace $\langle A, B, C, D, E\rangle$ is obviously longer and contains more events. Cutting off the event at the end of the trace will get the processed log in the lower left corner. The purpose of this technique is to avoid bias from very long traces [40].

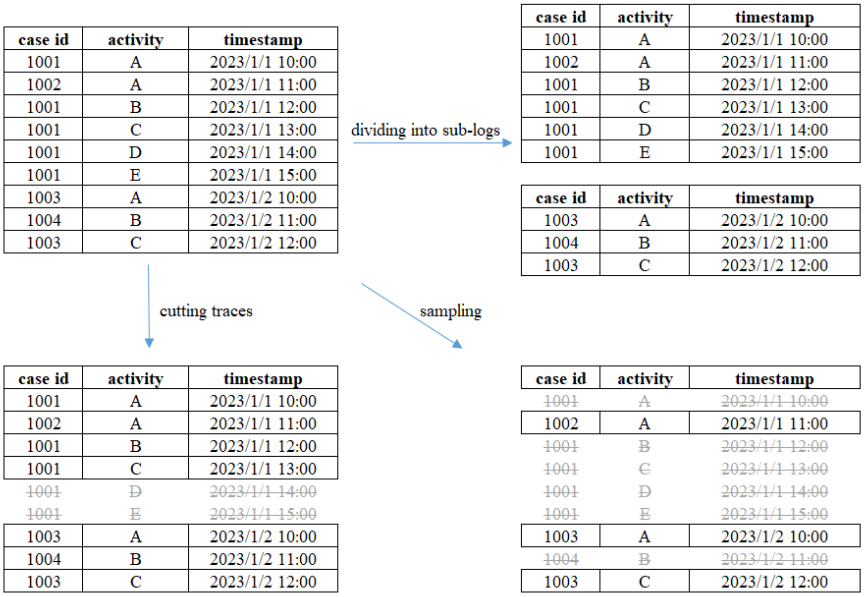


Figure 5.7: Simple example of log reduction.

5.1.5 Log integration

Among these 86 papers, 14 papers used log integration to combine multiple data tables. Since the essential operation and object-oriented of log integration are obviously single, that is, using the join command in the database language SQL to combine different data tables, therefore, in the high-level category of log integration, there is no further subdivision category.

As mentioned in 5.1.3, a simple example is shown in 5.9. A new log is given by matching two data tables by the common attribute "student_id". Another example worth mentioning is that some papers stated that additional data was added to the original log data without indicating the source, but we believe that the combination of these data is realized by log integration. For example, "Besides the attributes shown in Table 4, we included the educational level of the nurses executing the activity, as well as their nursing experience/organisational role, the hospital shift and weekday on which the activities were performed, and the ward in which the shift took place" [121]. It is reasonable to speculate that this additional information actually comes from a separate data table that stores information about all nurses.

5.1.6 Log abstraction

37 of the 86 papers used preprocessing techniques in log abstraction, which is the most widely used technique after log filtering and log transformation among

the six high-level categories of log preprocessing techniques.

In [74], a review and taxonomy of event abstraction was presented. An example of log abstraction is shown in Figure 5.8, the low-level recorded events were abstracted into high-level business activities to increase the level of event granularity. Aggregation, clustering and defining event class can be the operations used in log abstraction.

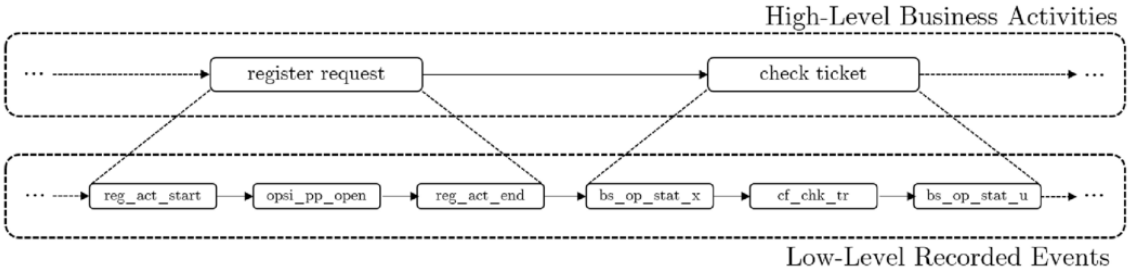


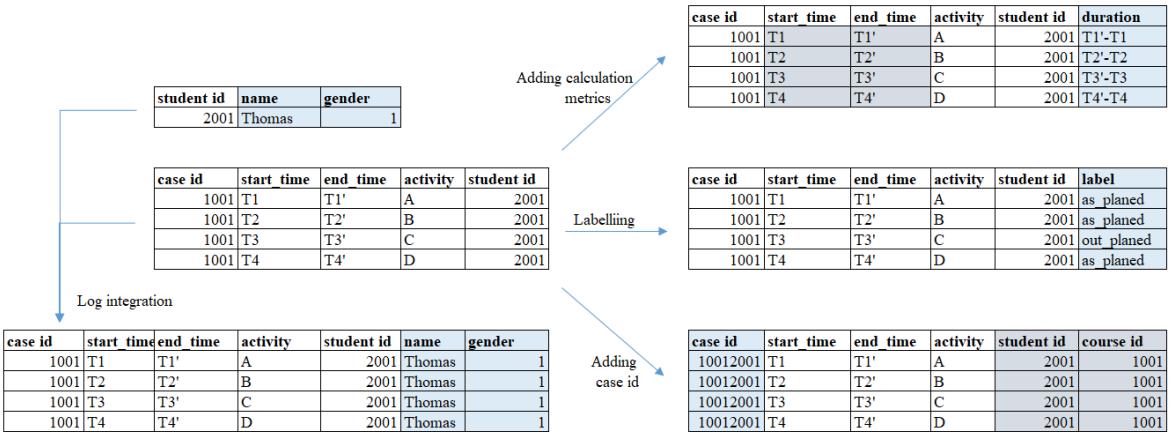
Figure 5.8: An example of log abstraction [74].

5.1.7 Discussion

Log enriching and log integration Log integration is making a new log by joining two or even more logs, while log enriching is adding more attributes based on the one log itself. Examples are shown in Figure 5.9. The left side of the figure represents log integration, which combines two separate data tables into one table through a common attribute (student id). The right side of the figure shows three types of log enriching. Adding calculation metrics is adding the attribute "duration", whose value is equal to "end_time" minus "start_time". Labelling is giving each event a tag "as_planned" or "out_planed". Adding case id is creating the value of the attribute "case_id" by combining "course_id" and "student_id".

Labelling and log abstraction Labeling is essentially defining a class for event or trace, which is similar to some techniques in log abstraction such as define event class, but the purpose of the two is completely different. Labelling is to differentiate between events or traces for further classification analysis, or to differentiate between logs for comparison analysis such as conformance checking. Defining event class in the log abstraction is to abstract log, making log more simplified and structured. Labelling can be used for log abstraction, but not exclusively. The operations that the two preprocessing techniques are performing are fundamentally different.

Log filtering, log abstraction, and log reduction Log filtering and log abstraction can also reduce logs, but there are still differences between the techniques in these two categories and the log reduction here. Their motivations are different. ***Log filtering*** is due to the quality issues of the original data. It obtains



student id	name	gender
2001	Thomas	1

case id	start_time	end_time	activity	student id
1001	T1	T1'	A	2001
1001	T2	T2'	B	2001
1001	T3	T3'	C	2001
1001	T4	T4'	D	2001

case id	start_time	end_time	activity	student id	duration
1001	T1	T1'	A	2001	T1'-T1
1001	T2	T2'	B	2001	T2'-T2
1001	T3	T3'	C	2001	T3'-T3
1001	T4	T4'	D	2001	T4'-T4

case id	start_time	end_time	activity	student id	label
1001	T1	T1'	A	2001	as_planed
1001	T2	T2'	B	2001	as_planed
1001	T3	T3'	C	2001	out_planed
1001	T4	T4'	D	2001	as_planed

case id	start_time	end_time	activity	student id	name	gender
1001	T1	T1'	A	2001	Thomas	1
1001	T2	T2'	B	2001	Thomas	1
1001	T3	T3'	C	2001	Thomas	1
1001	T4	T4'	D	2001	Thomas	1

case id	start_time	end_time	activity	student id	course id
10012001	T1	T1'	A	2001	1001
10012001	T2	T2'	B	2001	1001
10012001	T3	T3'	C	2001	1001
10012001	T4	T4'	D	2001	1001

Figure 5.9: Simple example of log integration and enriching.

higher-quality logs by filtering out incorrect, incomplete, inconsistent, and irrelevant data. ***Log abstraction*** is due to the complexity of the original data. It groups logs through aggregation, defining event class, and clustering to reduce the complexity of logs. ***Log reduction*** is due to the data volume of the original data. It reduces the amount of data processed in a single analysis by random sampling, dividing or cutting, but still makes the data representative.

5.1.8 Summary

The complete taxonomy of event log preprocessing techniques see Figure 5.10. The detail for log preprocessing technique category and the corresponding references are shown in Table 5.1. The Excel file for coding results see Excel file link in onedrive.

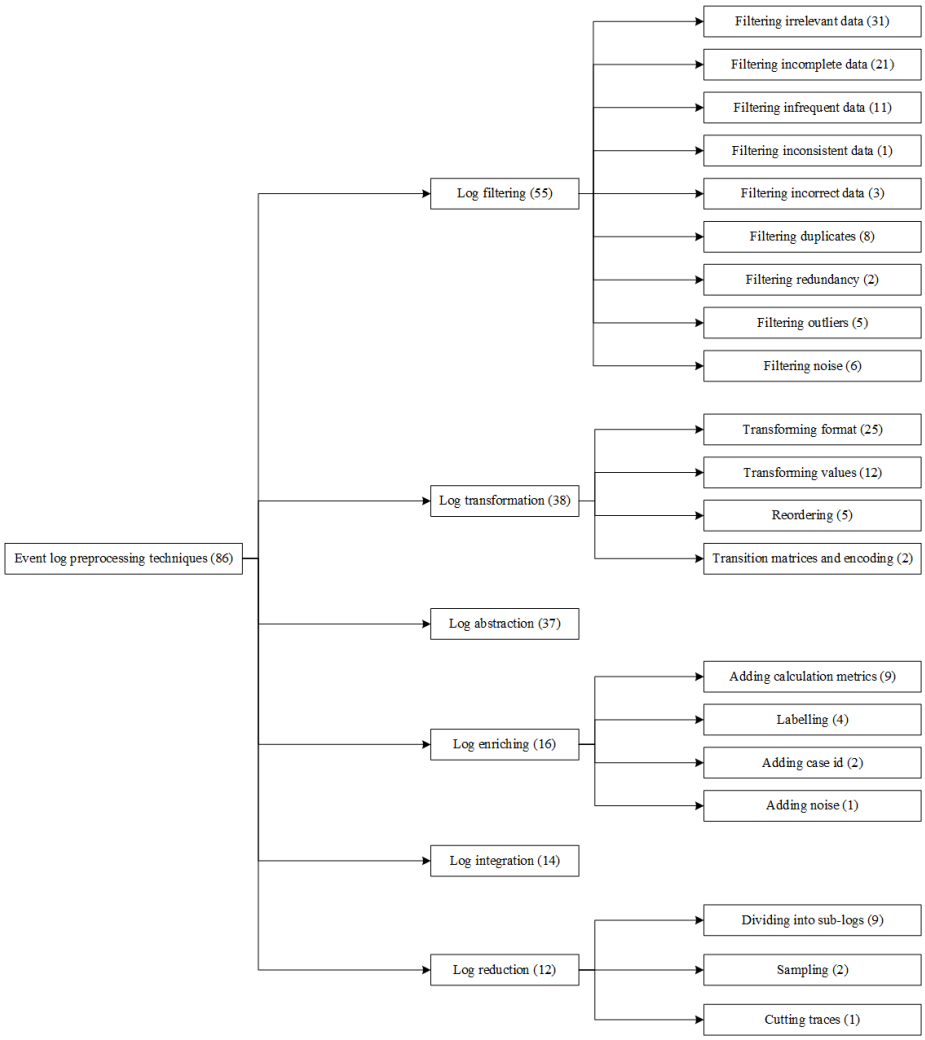


Figure 5.10: Taxonomy of log preprocessing techniques.

Table 5.1: Category citation details.

High-level category	Low-level category	References
Log filtering (55)	Filtering irrelevant data (28)	[2, 6, 10, 15, 19, 20, 21, 25, 33, 46, 47, 50, 55, 56, 58, 71, 87, 88, 91, 94, 98, 99, 101, 102, 106, 111, 112, 124]
	Filtering incomplete data (16)	[9, 23, 31, 33, 43, 46, 53, 58, 67, 84, 87, 89, 97, 98, 99, 111]
	Filtering infrequent data (13)	[8, 12, 22, 33, 121, 55, 57, 67, 89, 97, 109, 108, 110]
	Filtering duplicates (8)	[15, 30, 31, 33, 40, 89, 98, 101]
	Filtering outliers (5)	[14, 15, 21, 69, 96]
	Filtering unreliable data (3)	[43, 64, 105]
	Filtering redundant (2)	[22, 23]
	Filtering useless data (1)	[107]
	Filtering inconsistent data (1)	[20]
	Filtering incorrect data (1)	[84]
	Filtering noise (3)	[20, 29, 92]
Log transformation (38)	Transforming format (25)	[6, 11, 13, 15, 19, 20, 90, 31, 33, 37, 48, 50, 54, 55, 58, 62, 75, 83, 84, 89, 91, 96, 113, 114, 115]
	Transforming values (12)	[23, 33, 40, 64, 67, 76, 77, 83, 87, 98, 101, 102]
	Reordering (5)	[10, 31, 33, 71, 84]
	Transition matrices and encoding (2)	[33, 127]
Log abstraction (37)	-	[3, 4, 12, 14, 19, 22, 21, 25, 90, 32, 33, 37, 38, 47, 54, 61, 63, 64, 67, 29, 71, 72, 73, 75, 76, 77, 84, 88, 93, 95, 101, 107, 109, 108, 113, 124, 127]

Log enriching (16)	Adding calculation metrics (8)	[30, 32, 53, 57, 61, 80, 95, 103]
	Adding extra separate data (7)	[18, 25, 37, 121, 76, 77, 96]
	Labelling (4)	[5, 56, 83, 108]
	Adding case id (2)	[96, 106]
	Adding noise (1)	[104]
Log integration (14)	-	[18, 25, 90, 30, 37, 46, 121, 65, 77, 80, 89, 96, 101, 103]
Log reduction (12)	Dividing into sub-logs (9)	[23, 38, 46, 53, 62, 29, 92, 103, 112]
	Sampling (2)	[40, 105]
	Cutting traces (1)	[40]

5.2 Domain, data domain and preprocessing techniques

This subsection analyses the distribution of domains throughout 86 papers as well as the connections between domain, data domain, and event log preprocessing techniques.

5.2.1 Domain and data domain

Table 5.2 shows the domain distribution among the 86 papers. The top four most popular domains, which account for 71% of all papers, are healthcare, education, manufacturing and finance. For definitions of the four most common domain, see Table 4.1 in Chapter 4. The definition of the domain "business" is especially clarified here: in the case study of the paper, only "business company" or "business process" was mentioned, without any clear reference to other domains. The "unknown" domain includes 1) public datasets such as BPIC, and 2)the case study was implemented in multiple domains, so no exact domain could be found.

Table 5.2: Domains distribution.

Domain	**Number of papers**	**Proportion**
Healthcare	28	33%
Education	20	23%
Manufacturing	8	9%
Finance	5	6%
Construction	3	3%
Energy	3	3%
E-Commerce	3	3%
IoT	2	2%
Insurance	2	2%
Business	2	2%
Telecom	2	2%
IT	1	1%
Tourism	1	1%
Agriculture	1	1%
3D modeling	1	1%
Unknown	4	5%
Total	**86**	**100%**

The paper [30] analysed a call process in an Energy sector and aimed to reveal the relationship between customer satisfaction and other key performance indicators (KPIs). The calling process is the data domain in this instance, and energy is the domain, as defined by the definitions of domain and data domain. We could see that both the analysis purpose and log preprocessing techniques

have a greater connection with the data domain than the domain. As a result, we propose that the data domain should be prioritized above the domain, but we are still very happy to see researchers actively exploring the use of process mining in various domains, especially in some uncommon domains such as agriculture and 3D modeling.

The data domain defines the type of process being analyzed. Table 5.3 shows examples of the six most common data domains and the number of papers.

Table 5.3: Data domain examples.

Data domain	**Examples**	**Number of papers**
Healthcare-related process	nurses' process and treatment process	29
Education-related process	course evaluation process and lecture preparation process	19
Manufacturing-related process	production process in factory and printer's process	8
E-Commerce process	purchase-to-pay process and online sales transaction process	6
Service-related process	calling process, the chat-bot's executed steps in customer service conversations and information service request process	5
Finance-related process	loan application process	5

5.2.2 Data domain and preprocessing techniques

Table 5.4 lists the frequencies of event log preprocessing techniques used in the six most common data domain and others. Log filtering is still the most popular log preprocessing technique in most data domains except manufacturing-related process and E-Commerce process. The second-most popular technique differs between each data domain.

Table 5.4: Data domain and log preprocessing techniques.

Data domain	**log filtering**	**log transformation**	**log abstraction**	**log enriching**	**log reduction**	**log integration**
Healthcare-related process	19	11	**12***	4	1	5
Education-related process	13	**13***	9	1	1	3
Manufacturing-related process	3	3	1	2	2	**4***
E-Commerce process	3	4	2	2	2	
Service-related process	5	1	2	3		1
Finance-related process	4	1	3	1		
Others	8	5	8	3	5	1
Total	55	38	37	16	11	14

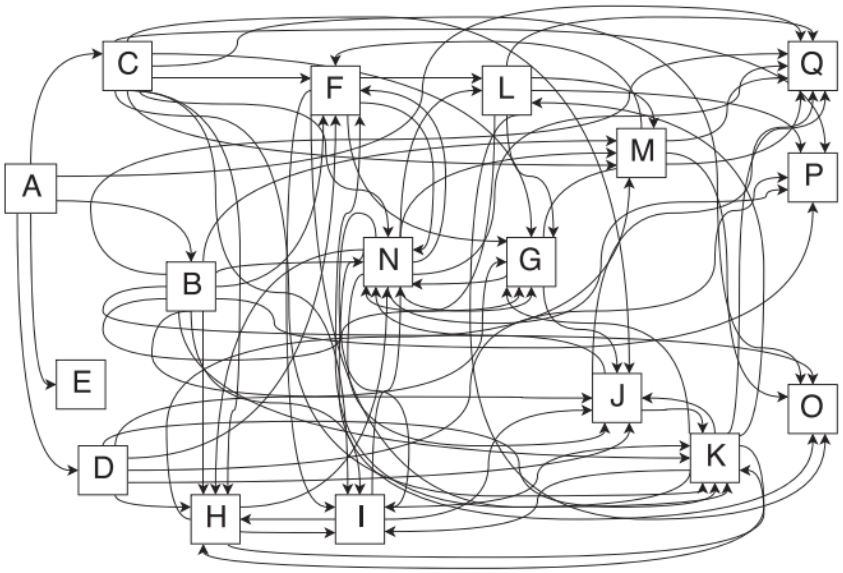


Figure 5.11: Examples of a spaghetti process model discovered from healthcare-related process [32].

In the healthcare-related data domain, the second most used technology is log abstraction, which is associated with the high granularity and high complexity of the healthcare logs. In [32], it mentioned "the ED (emergency department) process has the characteristics of a spaghetti process that is an unstructured process", including "more than 7.800 possible patient paths involving 17 different ED activities". Figure 5.11 shows an example of spaghetti process model, which is disorganized, complicated. To make the process model simple and structured, then the patients having similar pathways were clustered into groups in accordance with their characteristics such as main symptom, urgency code, age. Another detailed example is that in [77], the activities of 'Request an echo', 'Request a CT' and 'Request an MRI' were grouped in an activity 'Request a scan', which was then combined with 'Request lab test' to generate the activity 'Request diagnostic test'.

In the education-related data domain, the most used technology is log transformation, tied for first place with log filtering, indicating that it is frequently necessary to modify the education log format in order to employ process mining. The log of education-related process is usually collected from Learning Management System (LMS) [10, 19, 20, 115] or other online learning platform [6, 50, 55, 75, 90, 91, 96, 113], and in CSV format. However, the log imported into PM tools should be in XES format. Thus, a format transformation is usually necessary in education-related process.

The performance of the manufacturing-related process in the application

of log preprocessing techniques is entirely different from the other types of process. Log integration is the technique most frequently utilized in manufacturing-related processes, demonstrating that there are frequently numerous sources of manufacturing-related data and that it is crucial to merge data from numerous sources before data analysis. For example, the data used in [80] was from various departments (digitalization, logistics, and production) in a manufacturing factory, and in [18] the data was from multiple devices (printer).

5.3 Analysis purpose, data domain and preprocessing techniques

In this section, we analyze the used log preprocessing techniques from the perspective of analysis purpose. Each paper employed process mining technology for a certain purpose. Some papers desired to understand user behavior, whereas some aimed to improve process effectiveness. The analysis purpose was coded and grouped into seven class, pattern detection, process understanding, process improvement, process evaluation, process analysis, PM applicability and prediction. Table 5.5 shows the seven class of analysis purpose, examples, and the number of papers for that purpose.

Table 5.5: Analysis purpose examples.

Purpose	**Examples**	**Number of papers**
pattern detection	to discover human habits, to discover the common patient pathways	19
process understanding	to capture a more comprehensive view within a company's audit, to understand patients' journeys	19
process improvement	to improve care processes, to improve process discovery results	18
process evaluation	to quantify hospital adaptation to the pandemic, to assess the status of EMR (electronic medical record) adoption	12
process analysis	to discover deadlocks, statistical analysis, bottleneck analysis	11
PM applicability	to study the applicability of process mining	4
prediction	to predict Hospital-Stay Duration	3

5.3.1 Analysis purpose and data domain

Table 5.6 shows the distribution of analysis purposes of papers in each data domain. In healthcare-related processes, most papers (9/29) aimed at improving the processes. Whereas education-related processes focus more on pattern detection, that is, identifying the habits or patterns of students in a particular course or when using a learning platform. Manufacturing-related processes pay more attention to the understanding of the process, for example to capture a more comprehensive view or to gain more knowledge.

Table 5.6: Analysis purpose and data domain.

Data domain	process understanding	pattern detection	process improvement	process evaluation	process analysis	PM applicability	prediction	Total
Healthcare-related process	6	5	**9***	5	1	2	1	29
Education-related process	4	**6***	3	2	4			19
Manufacturing-related process	**3***		2		1	1	1	8
E-Commerce process	1	2	1	1	1			6
Service-related process	1	1		2	1			5
Finance-related process	1	1	1	1	1			5
Others	3	4	2	1	2	1	1	14
Total	19	19	18	12	11	4	3	86

5.3.2 Analysis purpose and preprocessing techniques

Table 5.7 shows the total number of times preprocessing techniques were used in different groups of analysis purposes. For example, 19 papers aimed at process understanding and used log filtering techniques for 12 times in total. We divide 12 by 19 to get that in the analysis purpose group of process understanding, each paper used an average of 0.63 log filtering techniques. All the data are processed in turn, then Table 5.8 is obtained. In Table 5.8, the last row represents the average frequency of each paper using different techniques across all papers. Comparing the data in each row with the last row, we can find out in which groups of analysis purposes, a log preprocessing technique is used significantly more or less than average.

The comparison shows that in the process evaluation purpose group, the use frequency of log filtering is significantly higher than the average, while in the process understanding and pattern detection purpose group, the technology with a significantly higher use frequency than the average is log abstraction.

Table 5.7: The total number of times preprocessing techniques were used in different groups of analysis purposes.

Analysis purpose	log filtering	log transformation	log abstraction	log enriching	log integration	log reduction	Number of papers
process understanding	12	8	**10***	3	5	2	19
pattern detection	9	9	**10***	5	1	3	19
process improvement	13	8	5	3	5	2	18
process evaluation	**10***	5	3	3	2	2	12
process analysis	8	5	4	1		1	11
PM applicability	1	1	2	1	1	1	4
prediction	2	2	3				3
Total	55	38	37	16	14	11	86

Table 5.8: The average frequency of preprocessing techniques used per paper in different analysis purpose groups

Analysis purpose	log filtering	log transformation	log abstraction	log enriching	log integration	log reduction
process understanding	63%	42%	**53%***	16%	26%	11%
pattern detection	47%	47%	**53%***	26%	5%	16%
process improvement	72%	44%	28%	17%	28%	11%
process evaluation	**83%***	42%	25%	25%	17%	17%
process analysis	73%	45%	36%	9%	0%	9%
PM applicability	25%	25%	50%	25%	25%	25%
prediction	67%	67%	100%	0%	0%	0%
Total	**64%**	**44%**	**43%**	**19%**	**16%**	**13%**

5.4 PM task, data domain and preprocessing techniques

In this section, we analyze the used log preprocessing techniques from the perspective of PM task. PM task refers to process discovery, conformance checking, performance enhancement and prediction.

5.4.1 PM task and data domain

Table 5.9 shows the total number of different PM task implementations in papers in different data domains. Process discovery is the most frequently performed PM task in any data domain, especially in healthcare-related processes and education-related processes.

Table 5.9: PM task and analysis purpose

Analysis purpose	process discovery	conformance checking	performance enhancement	prediction
Healthcare-related process	**16***	7	2	1
Education-related process	**10***	4		
Manufacturing-related process	4	1	2	2
Service-related process	4	4		1
IT-related process	3		1	
Construction-related process	3	1		
E-Commerce process	3	1		
Finance-related process	3	1	1	
IoT process	2			
Business process	1			
Agriculture-related process	1			1
Energy-related process	1	1		
Total	51	20	6	5

5.4.2 PM task and analysis purpose

Table 5.10 shows the number of times different PM tasks were executed in the papers of different analysis purpose groups. It is obvious that regardless of the analysis purpose group, process discovery is the most frequently carried out PM task.

Table 5.10: PM task and analysis purpose

Analysis purpose	process discovery	conformance checking	performance enhancement	prediction
pattern detection	12	5		1
process understanding	11	1	2	1
process improvement	10	4	2	1
process analysis	9	4	2	
process evaluation	6	4		
PM applicability	2	2		
prediction	1			2
Total	51	20	6	5

5.4.3 PM task and preprocessing techniques

Table 5.11 shows the distribution of the total number of log preprocessing techniques used by papers that perform different PM task combinations.

According to the results of section 5.1, in all 86 papers, log filtering is the most used technology, log transformation is the second most used, and log abstraction is the third most used. We noticed that in papers that only implemented process discovery, log abstraction is the second most used technique, while log transformation is the third most used technique, but the gap is very small.

In the PM task combination of process discovery and performance enhancement, four [32, 84, 88, 95] of the five papers used log abstraction. However, these four papers come from different data domain and have different analysis purposes, as shown in Figure 5.12. Perhaps due to the limitation of the small number of papers, we could not find the regularity.

Domain	Data domain	Data domain group	Analysis purpose	Analysis purpose group
Healthcare	emergency department (ED)	Healthcare-related process	to improve the ED performance, alleviating the overcrowding.	process improvement
Education	IT ticket handling process	IT-related process	to analyze business process execution complexity	process analysis
Finance	loan process	Finance-related process	to discover problems and areas for improvement	process improvement
Healthcare	acute care process	Healthcare-related process	for understanding acute care	process understanding

Figure 5.12: The four papers that used log abstraction and implemented process discovery and performance enhancement.

Table 5.11: The average frequency of preprocessing techniques used per paper in different analysis purpose groups

PM task	log filtering	log transformation	log abstraction	log enriching	log integration	log reduction	Number of papers
process discovery	14	**9***	**12***	3	2	5	24
process discovery, conformance checking	14	11	4	5	1	1	18
process discovery, prediction	2	1	1			1	2
process discovery, performance enhancement	2	1	**4***	2	1		5
conformance checking	1		1				1
process discovery, conformance checking, prediction	1		1	1			1
process discovery, performance enhancement, prediction	1						1
prediction		1	1				1
-	20	15	13	5	10	4	33
Total	55	38	37	16	14	11	86
Total	**64%**	**44%**	**43%**	**19%**	**16%**	**13%**	

Chapter 6

Discussion

In this chapter, insights summary, the papers that are screened out during coding stage, limitations, and future work are discussed.

6.1 Summary

Six high-level categories, including log filtering, log transformation, log abstraction, log enriching, log integration, and log reduction, are used to summarize a taxonomy of log preprocessing techniques by conducting SLR. Since [74] has proposed a taxonomy of log abstraction, we don't do any further detailed explanation on abstraction. There are also no longer any classifications in log integration due to the singleness of operation and object.

We discovered that the data domain is the most influential factor in the choice of log preprocessing techniques by analyzing the relationships between log preprocessing techniques and data domain, analysis purpose, and PM task. This is because the log collected in a particular data domain typically has similar characteristics. For instance, 1) The log is only accessible at various levels of granularity in healthcare-related process. So log abstraction is typically chosen to give the process model structure. 2) In education-related processes, the data from the online learning platform is usually in CSV format, while PM tools require the imported data format to be in XES format. Therefore, log transformation is usually a necessary step. 3) In manufacturing-related processes, data is usually scattered in different departments and devices, so log integration is usually selected.

However, the selection of log preprocessing techniques cannot simply be determined based on the data domain. Correlation does not imply necessity, nor does it imply causation. The motivation and purpose behind the selection should be more concerned. For example, given logs from a healthcare-related process, an analyst would need to understand the granularity of activities and use cases in the log as well as the necessary process model to decide whether to perform log abstraction, instead of simply understanding that all logs related

to healthcare processes must implement log abstraction.

Therefore, in addition to the taxonomy, we propose a task-oriented checklist related to the selection of data preprocessing technology for analysts.

1) Check the data source If the data comes from different sources (departments, equipment, etc.), and to integrate the data into a log file for analysis, log integration is a recommended step.

2) Understand analysis requirements Experts and stakeholders can decide which data is relevant based on domain knowledge and analysis requirements, thereby filtering out irrelevant data. For example, if an analysis only needs to focus on the logs of department A that occurred in the year 2022, then the data of other departments or that occurred in other years needs to be filtered out.

3) Understand the analysis purpose In most cases, analysts often focus on frequent traces, but there are also analyses aiming to understand unhappy paths and extreme paths. Therefore, whether to screen out infrequent data, incomplete cases, outliers, and noise need to be determined according to the analysis purpose.

4) Understand aimed data quality Incomplete events, inconsistent data, incorrect data, duplicates, redundancy, outliers, and noise are usually data errors. To obtain high-quality logs, these data errors usually need to be screened out. However, in some cases, noise is even added to the logs to improve data privacy. Therefore, aimed data quality determines whether to screen out, transform, or even add the above data.

5) Understand the aimed process model According to the structure and simplicity of the aimed process model, log abstraction and filtering infrequent data are decided on whether to choose.

6) Check if additional data is required If additional information is required, then depending on the source of the information, log integration and log enriching are selected.

7) Check the data format Since PM tools require the imported data format to be in XES format, analysts need to check the format of the final data file to determine whether format transformation is required.

6.2 Papers without mentioning preprocessing techniques

In coding process of the paper screen stage, 73 papers were excluded because of no preprocessing techniques mentioned. These papers cover the following three situations: 1) no process model exists; 2) a process model exists but no data preprocessing steps are documented; and 3) there are data preprocessing steps but we are not convinced it is a preprocessing technique.

In the second situation, an interesting finding is that some of the papers used public datasets (such as BPIC). Therefore, it's conceivable that certain publicly accessible datasets are preprocessed datasets.

The third situation is mainly because of the manual mentioned in the paper. In our opinion, the term "manually" refers to processing data by hand, line by line, instead of using any tools, such as PM tools or programming.

6.3 Limitations

One limitation is imposed by the number of papers that were subjected to the final analysis. Even though we believe 86 publications is a valid quantity, coding and analysing more papers might provide new insights, particularly with taxonomy derivation.

Another limitation is that taxonomy is only derived from text analysis without incorporating more diverse sources of information, such as interviews or questionnaires to industry professionals to obtain qualitative data or evaluate the taxonomy.

6.4 Future work

Code the order between preprocessing steps If the execution sequence between preprocessing techniques can be coded, then we can regard the preprocessing stage in data analysis as a process, and the preprocessing technique selections are the activities in the process. In this way, we can analyze what the most frequent preprocessing process is for different analysis purpose in each data domain. Due to time constraints, it is a pity that this analysis, which may have yielded interesting findings, could not be completed.

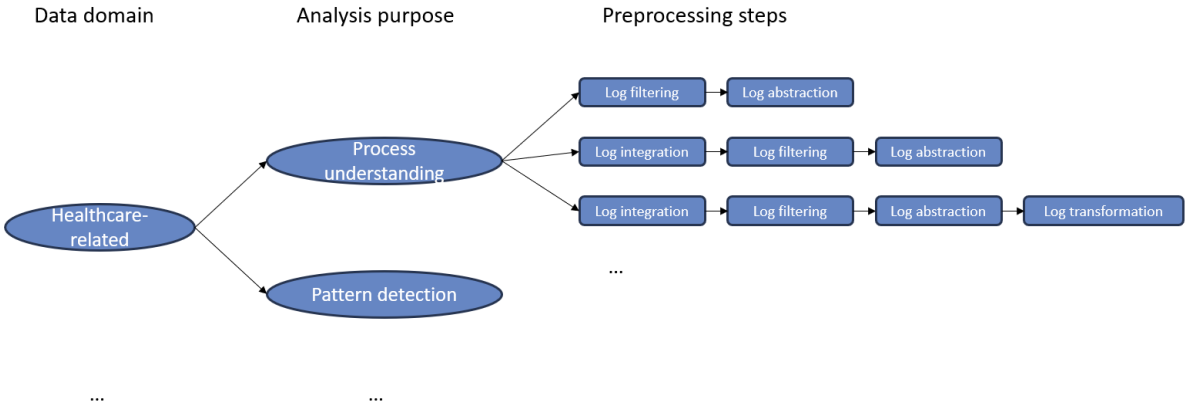


Figure 6.1: An analysis prototype after coding the order between the preprocessing steps.

Evaluation Interviews or questionnaires could be done in the future to evaluate the results.

Tool development In the category log transformation, many case studies convert the log in CSV format to log in XES format, so that the log can be analysed in process mining tools. Then, the development of a plugin to facilitate format transformation in the PM tool may be a future work.

Robotic Process Automation (RPA) Automated or semi-automated log preprocessing may also be an interesting direction.

Chapter 7

Conclusion

In this thesis, we studied the preprocessing techniques used in the case study of process mining-related papers published in the past two years. Through coding and analysis of 86 papers, a taxonomy with six high-level categories and twenty low-level categories was proposed. Based on this taxonomy, we analyzed the relationship between log preprocessing techniques and data domain, analysis purpose, and PM task. In the following, we summarize the answer to each research question.

RQ *What is a task-oriented and operation-based taxonomy of event log preprocessing techniques in process mining?*

The taxonomy refers to Figure 5.10, including six high-level categories, log filtering, log transformation, log abstraction, log enriching, log reduction and log integration. Four of these, log filtering, log transformation, log enriching and log reduction, were further divided into a total of twenty low-level categories.

SRQ1 *What is the existing taxonomy of event log preprocessing techniques in process mining?*

The only existing taxonomy of event log preprocessing techniques was developed by [74], see Figure 3.5. However, we expect to propose a new taxonomy from another perspective, which is task-oriented and operation-based, see more arguments in Section 3.3.1.

SRQ2 *What factors (e.g., data domain, analysis purpose. . .) will affect the selection of event log preprocessing techniques in process mining analysis?*

Through the previous analysis, we found that 1) Data domain affects the selection of data preprocessing techniques the most. 2) Different analysis purposes have different performances in the use of specific log processing technologies. Process evaluation employs more techniques in log filtering, whereas process understanding and pattern detection employ more techniques in log abstraction when compared to the work in other analysis purpose. 3) The analysis

of the intended PM task does not see a significant impact on the choice of log preprocessing techniques.

SRQ3 *Is the selection of event log preprocessing techniques data domain-dependent?*

Through the preceding analysis, we discovered that while log filtering is a common technique across all data domains, other preprocessing techniques have completely different frequency distributions. As a result, we conclude that log filtering is a widely used approach and is NOT data domain-dependent, however, the choice of other log preprocessing techniques is.

The fundamental reason why log filtering is NOT data domain-dependent is that log quality is the cornerstone of log analysis. Log filtering produces higher-quality data sets by filtering out various types of data that are not conducive to data analysis.

The fundamental reason why other preprocessing technologies are data domain-dependent is that different data domains have varied data characteristics. For instance, log abstraction is a frequently used technology in the manufacturing data domain because manufacturing data is often spread across multiple sources. Similarly, log integration is a frequently used technology in the healthcare data domain because it has high granularity and complexity.

SRQ4 *How would analysis task affect the selection of event log preprocessing techniques?*

1) *data resource* If the data comes from many data sources, then log integration is a necessary step. Due to the multi-source nature of the data generated by manufacturing-related processes, log integration is commonly used in the analysis of manufacturing-related processes.

2) *analysis requirement and domain knowledge* Analysis requirements first need to be evaluated and analyzed by stakeholders and experts based on domain knowledge, to determine which data is relevant and needs to be collected, and which data is irrelevant and needs to be filtered.

3) *analysis purpose* Since some analyses focus on unhappy paths or even very extreme paths, data that is likely to be screened out in common analysis is retained, such as infrequent data. Therefore, the analysis purpose determines the object of log filtering to some extent.

4) *aimed data quality* As introduced in section 5.1.1, several data quality issues might exist in the collected log such as incomplete data (unrepresentative case and missing event), inconsistent data, incorrect data, duplicates, and redundancy. Depending on the aimed data quality, these data can be filtered or transformed.

5) *aimed process model* To make aimed process model simpler and more structured, log abstraction might be selected to group specific activities or traces.

6) *whether additional information needed* The selection of log enriching is depending on whether the key attribute (case id) lacked and additional infor-

mation is needed. If the case id is missing or additional attributes such as calculation and label are required, log enriching should be selected.

7) *data format and tools* To suit PM tools such as Disco, Celonis, and ProM, the imported data format should be XES. Thus, data format checking is necessary before applying PM tools. If the data is not in XES format, log transformation is needed.

The taxonomy proposed in this paper provides a detailed classification of the preprocessing techniques used in the case study of 86 articles published in the past two years. By analyzing the chosen preprocessing technique and different aspects (data domain, analysis purpose, and PM task) of the analysis, we find a correlation between technique choice and data domain. Correlation, however, does not imply inevitability and causation. Therefore, we discuss in more depth the motivation and reasons for the selection of some preprocessing techniques. Listing the factors (data resource, analysis requirement and domain knowledge, analysis purpose, aimed data quality, aimed process model, if additional information needed, data format and tools) that may affect the selection of preprocessing techniques in the analysis task, this paper provides analysts with insights into the selection of preprocessing techniques during data analysis to some extent.

Bibliography

[1] Definition of irrelevant. https://www.collinsdictionary.com/zh/dictionary/english/irrelevant, 2023.

[2] Acacio-Claro, P. J., Estuar, M. R. J., Villamor, D. A., Bautista, M. C., Sugon, Q., and Pulmano, C. A micro-analysis approach in understanding electronic medical record usage in rural communities: Comparison of frequency of use on performance before and during the COVID-19 pandemic. *Procedia Computer Science 196* (2022), 572–580.

[3] Adams, J. N., van Zelst, S. J., Rose, T., and van der Aalst, W. M. Explainable concept drift in process mining. *Information Systems 114* (mar 2023), 102177.

[4] Ahmad, N. D., Mat, H., Shahuddin, A. Z., Shaffiei, Z. A., Elias, S. J., Hatim, S. M., and Ahmad, S. Process mining of cardiovascular diseases trajectories in malaysia public hospital: A feasibility study. In *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)* (sep 2021), IEEE.

[5] Araghi, S. N., Fontanili, F., Lamine, E., Okongwu, U., and Benaben, F. Stable heuristic miner: Applying statistical stability to discover the common patient pathways from location event logs. *Intelligent Systems with Applications 14* (may 2022), 200071.

[6] Ardimento, P., Bernardi, M. L., and Cimitile, M. Using process mining to understand students' and teams' dynamics. In *Higher Education Learning Methodologies and Technologies Online*. Springer International Publishing, 2022, pp. 63–73.

[7] Azevedo, A., and Santos, M. F. KDD, SEMMA and CRISP-DM: A Parallel Overview. *IADS-DM* (2008).

[8] Bahaweres, R. B., Amna, H., and Nurnaningsih, D. Improving purchase to pay process efficiency with RPA using fuzzy miner algorithm in process mining. In *2022 International Conference on Decision Aid Sciences and Applications (DASA)* (mar 2022), IEEE.

[9] Bahaweres, R. B., Trawally, J., Hermadi, I., and Suroso, A. I. Forensic audit using process mining to detect fraud. *Journal of Physics: Conference Series 1779*, 1 (feb 2021), 012013.

[10] Bajo, O. E., Amarasinghe, I., Gutiérrez-Páez, N. F., and Hernández-Leo, D. Using process mining techniques to discover the collective behaviour of educators in a learning community platform. In *Collaboration Technologies and Social Computing*. Springer International Publishing, 2022, pp. 175–189.

[11] Battineni, G., Chintalapudi, N., and Amenta, F. Model discovery, and replay fitness validation using inductive mining techniques in medical training of CVC surgery. *Applied Computing and Informatics 18*, 3/4 (jul 2020), 245–255.

[12] Battineni, G., Chintalapudi, N., and Zacharewicz, G. Process mining in clinical practice: Model evaluations in the central venous catheter installation training. *Algorithms 15*, 5 (apr 2022), 153.

[13] Beerepoot, I., Lu, X., Van De Weerd, I., and Alexander Reijers, H. Seeing the signs of workarounds: a mixed-methods approach to the detection of nurses' process deviations.

[14] Benevento, E., Aloini, D., and van der Aalst, W. M. How can interactive process discovery address data quality issues in real business settings? evidence from a case study in healthcare. *Journal of Biomedical Informatics 130* (jun 2022), 104083.

[15] Birk, A., Wilhelm, Y., Dreher, S., Flack, C., Reimann, P., and Gröger, C. A real-world application of process mining for data-driven analysis of multi-level interlinked manufacturing processes. *Procedia CIRP 104* (2021), 417–422.

[16] Bogarín, A., Cerezo, R., and Romero, C. A Survey on Educational Process Mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8*, 1 (2018), e1230.

[17] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software 80*, 4 (apr 2007), 571–583.

[18] Brockhoff, T., Uysal, M. S., Terrier, I., Göhner, H., and van der Aalst, W. M. P. Analyzing multi-level BOM-structured event data. In *Lecture Notes in Business Information Processing*. Springer International Publishing, 2022, pp. 47–59.

[19] Cenka, B. A. N., S. H. B. . J. K. Analysing student behaviour in a learning management system using a process mining approach. *Knowledge*

Management &amp E-Learning: An International Journal (mar 2022), 62–80.

[20] Chanifah, S., Andreswari, R., and Fauzi, R. Analysis of student learning pattern in learning management system (LMS) using heuristic mining a process mining approach. In *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)* (jul 2021), IEEE.

[21] Chen, L., and Klasky, H. B. Six machine-learning methods for predicting hospital-stay duration for patients with sepsis: A comparative study. In *SoutheastCon 2022* (mar 2022), IEEE.

[22] Chen, Q., Lu, Y., Tam, C. S., and Poon, S. K. A multi-view framework to detect redundant activity labels for more representative event logs in process mining. *Future Internet 14*, 6 (jun 2022), 181.

[23] Cho, M., Park, G., Song, M., Lee, J., Lee, B., and Kum, E. Discovery of resource-oriented transition systems for yield enhancement in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing 34*, 1 (2020), 17–24.

[24] Chu, X., Ilyas, I. F., Krishnan, S., and Wang, J. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data* (jun 2016), ACM.

[25] Cuendet, M. A., Gatta, R., Wicky, A., Gerard, C. L., Dalla-Vale, M., Tavazzi, E., Michielin, G., Delyon, J., Ferahta, N., Cesbron, J., Lofek, S., Huber, A., Jankovic, J., Demicheli, R., Bouchaab, H., Digklia, A., Obeid, M., Peters, S., Eicher, M., Pradervand, S., and Michielin, O. A differential process mining analysis of COVID-19 management for cancer patients. *Frontiers in Oncology 12* (dec 2022).

[26] Dakic, D., Stefanovic, D., Cosic, I., Lolic, T., and Medojevic, M. Business Process Mining Application: A Literature Review. In *Proceedings of the 29th International DAAAM Symposium 2018*. DAAAM International Vienna, 2018, pp. 0866–0875.

[27] Dakic, D., Stefanovic, D., Lolic, T., Narandzic, D., and Simeunovic, N. Event Log Extraction for the Purpose of Process Mining: A Systematic Literature Review. In *International Symposium in Management Innovation for Sustainable Management and Entrepreneurship* (2020), Springer, pp. 299–312.

[28] de Leoni, M. Foundations of Process Enhancement. In *Process Mining Handbook*. Springer, 2022, pp. 243–273.

[29] de Leoni, M., and Pellattiero, L. The benefits of sensor-measurement aggregation in discovering IoT process models: A smart-house case study. In *Business Process Management Workshops*. Springer International Publishing, 2022, pp. 403–415.

[30] Dogan, O. A process-centric performance management in a call center. *Applied Intelligence 53*, 3 (may 2022), 3304–3317.

[31] Du, L., Cheng, L., and Liu, C. Process mining for wind turbine maintenance process analysis: A case study. In *2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2)* (oct 2021), IEEE.

[32] Duma, D., and Aringhieri, R. Real-time resource allocation in the emergency department: A case study. *Omega 117* (jun 2023), 102844.

[33] Dupuis, A., Dadouchi, C., and Agard, B. Predicting crop rotations using process mining techniques and markov principals. *Computers and Electronics in Agriculture 194* (mar 2022), 106686.

[34] Eck, M. L. v., Lu, X., Leemans, S. J. J., and Van Der Aalst, W. M. P. PM2: A Process Mining Project Methodology. In *International conference on advanced information systems engineering* (2015), Springer, pp. 297–313.

[35] Eggers, J., and Hein, A. Turning Big Data into Value: A Literature Review on Business Value Realization from Process Mining. In *ECIS* (2020).

[36] Erdogan, T. G., and Tarhan, A. Systematic Mapping of Process Mining Studies in Healthcare. *IEEE Access 6* (2018), 24543–24567.

[37] Erdogan, T. G., and Tarhan, A. K. Multi-perspective process mining for emergency process. *Health Informatics Journal 28*, 1 (jan 2022), 146045822210771.

[38] Esposito, L., Leotta, F., Mecella, M., and Veneruso, S. Unsupervised segmentation of smart home logs for human habit discovery. In *2022 18th International Conference on Intelligent Environments (IE)* (jun 2022), IEEE.

[39] Fahland, D. Extracting and Pre-Processing Event Logs, 2022.

[40] Fahrenkrog-Petersen, S. A., Tax, N., Teinemaa, I., Dumas, M., de Leoni, M., Maggi, F. M., and Weidlich, M. Fire now, fire later: Alarm-based systems for prescriptive process monitoring. *Knowledge and Information Systems 64*, 2 (dec 2021), 559–587.

[41] Famili, A., Shen, W.-M., Weber, R., and Simoudis, E. Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis 1*, 1 (jan 1997), 3–23.

[42] Feyyad, U. M. Data Mining and Knowledge Discovery: Making Sense Out of Data. *IEEE expert 11*, 5 (1996), 20–25.

[43] Gao, W., Wu, C., Huang, W., Lin, B., and Su, X. A data structure for studying 3d modeling design behavior based on event logs. *Automation in Construction 132* (dec 2021), 103967.

[44] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., and Herrera, F. Big Data Preprocessing: Methods and Prospects. *Big Data Analytics 1*, 1 (nov 2016).

[45] Ghazal, M. A., Ibrahim, O., and Salama, M. A. Educational Process Mining: A Systematic Literature Review. In *2017 European Conference on Electrical Engineering and Computer Science (EECS)* (nov 2017), IEEE.

[46] Goel, K., Leemans, S., Wynn, M. T., ter Hofstede, A., and Barnes, J. Improving phd student journeys with process mining: Insights from a higher education institution. In *Proceedings of the Industry Forum (BPM IF 2021) co-located with 19th International Conference on Business Process Management (BPM 2021)* (2021), Sun SITE Central Europe (CEUR), pp. 39–49.

[47] Grüger, J., Geyer, T., Kuhn, M., Braun, S., and Bergmann, R. Verifying guideline compliance in clinical treatment using multiperspective conformance checking: A case study. In *Lecture Notes in Business Information Processing*. Springer International Publishing, 2022, pp. 301–313.

[48] Grüger, J., Kuhn, M., and Bergmann, R. Reconstructing invisible deviating events: A conformance checking approach for recurring events. *Mathematical Biosciences and Engineering 19*, 11 (2022), 11782–11799.

[49] Guyon, I., and Elisseeff, A. An Introduction to Variable and Feature Selection. *Journal of machine learning research 3*, Mar (2003), 1157–1182.

[50] Hachicha, W., Ghorbel, L., Champagnat, R., Zayani, C. A., and Amous, I. Using process mining for learning resource recommendation: A moodle case study. *Procedia Computer Science 192* (2021), 853–862.

[51] Hall, M. A. *Correlation-based Feature Selection for Machine Learning*. PhD thesis, The University of Waikato, 1999.

[52] Han, J., Pei, J., and Tong, H. *Data Mining: Concepts and Techniques*. Morgan kaufmann, 2022.

[53] Huda, S., Aripin, Naufal, M. F., and Yudianingtias, V. M. Identification of fraud attributes for detecting fraud based online sales transaction. *Indian Journal of Computer Science and Engineering 12*, 5 (oct 2021), 1409–1424.

[54] Husin, H. S., and Ismail, S. Process mining approach to analyze user navigation behavior of a news website. In *2021 The 4th International Conference on Information Science and Systems* (mar 2021), ACM.

[55] Ivanka, M. D., Andreswari, R., and Fauzi, R. Bottleneck analysis of lectures grades input process at information system academic using inductive miner. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)* (jan 2022), IEEE.

[56] Jonk, J., Schaller, M., Netzer, M., Pfeifer, B., Ammenwerth, E., and Hackl, W. Process mining of nursing routine data: Cool, but also useful? In *Studies in Health Technology and Informatics*. IOS Press, may 2022.

[57] Kecht, C., Egger, A., Kratsch, W., and Röglinger, M. Quantifying chatbots' ability to learn business processes. *Information Systems 113* (jan 2023), 102176.

[58] Khaosanoi, L., and Limpiyakorn, Y. Conformance checking and discovery of information service request process. In *2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (oct 2021), IEEE.

[59] Khatri, S., Alzahrani, F. A., Ansari, M. T. J., Agrawal, A., Kumar, R., and Khan, R. A. A Systematic Analysis on Blockchain Integration With Healthcare Domain: Scope and Challenges. *IEEE Access 9* (2021), 84666–84687.

[60] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., and Linkman, S. Systematic Literature Reviews in Software Engineering – A Systematic Literature Review. *Information and Software Technology 51*, 1 (jan 2009), 7–15.

[61] Koçi, R., Franch, X., Jovanovic, P., and Abelló, A. Web API evolution patterns: A usage-driven approach. *Journal of Systems and Software 198* (apr 2023), 111609.

[62] Kołakowska, A., and Godlewska, M. Analysis of factors influencing the prices of tourist offers. *Applied Sciences 12*, 24 (dec 2022), 12938.

[63] Kropp, T., Bombeck, A., and Lennerts, K. An approach to data driven process discovery in the cost estimation process of a construction company. In *Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC)* (nov 2021), International Association for Automation and Robotics in Construction (IAARC).

[64] Kropp, T., Faeghi, S., and Lennerts, K. Evaluation of patient transport service in hospitals using process mining methods: Patients' perspective. *The International Journal of Health Planning and Management 38*, 2 (nov 2022), 430–456.

[65] Kumbhar, M., Ng, A. H., and Bandaru, S. Bottleneck detection through data integration, process mining and factory physics-based analytics. In *Advances in Transdisciplinary Engineering*. IOS Press, apr 2022.

[66] Kurniati, A. P., Johnson, O., Hogg, D., and Hall, G. Process Mining in Oncology: A Literature Review. In *2016 6th International Conference on Information Communication and Management (ICICM)* (2016), pp. 291–297.

[67] Lamghari, Z. Process mining: A new approach for simplifying the process model control flow visualization. *Transdisciplinary Journal of Engineering &amp Science 13* (jul 2022).

[68] Lenzerini, M. Data Integration: A Theoretical Perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02* (2002), ACM Press.

[69] Lim, J., Kim, K., Song, M., Yoo, S., Baek, H., Kim, S., Park, S., and Jeong, W.-J. Assessment of the feasibility of developing a clinical pathway using a clinical order log. *Journal of Biomedical Informatics 128* (apr 2022), 104038.

[70] Liu, H., Hussain, F., Tan, C. L., and Dash, M. Discretization: An Enabling Technique. *Data mining and knowledge discovery 6*, 4 (2002), 393–423.

[71] López-Pernas, S., Saqr, M., and Viberg, O. Putting it all together: Combining learning analytics methods and data sources to understand students' approaches to learning programming. *Sustainability 13*, 9 (apr 2021), 4825.

[72] Macak, M., Kruzelova, D., Chren, S., and Buhnova, B. Using process mining for git log analysis of projects in a software development course. *Education and Information Technologies 26*, 5 (may 2021), 5939–5969.

[73] Macak, M., Oslejsek, R., and Buhnova, B. Process mining analysis of puzzle-based cybersecurity training. In *Proceedings of the 27th ACM Conference on on Innovation and Technology in Computer Science Education Vol. 1* (jul 2022), ACM.

[74] Marin-Castro, H. M., and Tello-Leal, E. Event Log Preprocessing for Process Mining: A Review. *Applied Sciences 11*, 22 (nov 2021), 10556.

[75] Martinez, P., Montañes, O., Serralta, J. M., and Tansini, L. Modelling computer engineering student trajectories with process mining. In *LALA* (2021), pp. 48–57.

[76] Mehraby, N., Neysiani, B. S., Nogorani, M. Z., and Ataabadi, P. E. Abnormal behavior detection in health insurance assessment process. In *2022 8th International Conference on Web Research (ICWR)* (may 2022), IEEE.

[77] Mertens, S., Gailly, F., Sassenbroeck, D. V., and Poels, G. Integrated declarative process and decision discovery of the emergency care process. *Information Systems Frontiers 24*, 1 (oct 2020), 305–327.

[78] Mivule, K. Utilizing noise addition for data privacy, an overview.

[79] Munk, M., Kapusta, J., and Švec, P. Data Preprocessing Evaluation for Web Log Mining: Reconstruction of Activities of a Web Visitor. *Procedia Computer Science 1*, 1 (may 2010), 2273–2280.

[80] Oberdorf, F., Schaschek, M., Weinzierl, S., Stein, N., Matzner, M., and Flath, C. M. Predictive end-to-end enterprise process network monitoring. *Business &amp Information Systems Engineering 65*, 1 (dec 2022), 49–64.

[81] Ordonez, C. Data Set Preprocessing and Transformation in a Database System. *Intelligent Data Analysis 15*, 4 (jun 2011), 613–631.

[82] Osborne, J. Notes on the Use of Data Transformations. *Practical assessment, research, and evaluation 8*, 1 (2002), 6.

[83] Pan, Y., and Zhang, L. Automated process discovery from event logs in BIM construction projects. *Automation in Construction 127* (jul 2021), 103713.

[84] Pang, J., Xu, H., Ren, J., Yang, J., Li, M., Lu, D., and Zhao, D. Process mining framework with time perspective for understanding acute care: A case study of AIS in hospitals. *BMC Medical Informatics and Decision Making 21*, 1 (dec 2021).

[85] Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. Synthesizing information systems knowledge: A typology of literature reviews. *Information &amp Management 52*, 2 (mar 2015), 183–199.

[86] Peshawa J. Muhammad Ali, R. H. F. Data Normalization and Standardization: A Technical Report. *Mach Learn Tech Rep 1*, 1 (2014), 1–6.

[87] Pieters, A. J., and Schlobach, S. Combining process mining and time series forecasting to predict hospital bed occupancy. In *Health Information Science*. Springer Nature Switzerland, 2022, pp. 76–87.

[88] Porouhan, P. Optimization of overdraft application process with fluxicon disco. In *2022 20th International Conference on ICT and Knowledge Engineering (ICT&ampKE)* (nov 2022), IEEE.

[89] Pradana, M. I. A., Kurniati, A. P., and Wisudiawan, G. A. A. Inductive miner implementation to improve healthcare efficiency on indonesia national health insurance data. In *2022 International Conference on Data Science and Its Applications (ICoDSA)* (jul 2022), IEEE.

[90] R., D. S., and Patil, M. M. Study of learners behaviour in virtual learning environment using process mining. In *2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (jul 2021), IEEE.

[91] Rahmawati, R., Andreswari, R., and Fauzi, R. Analysis and exploratory of lecture preparation process to improve the conformance using process mining. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (jan 2022), IEEE.

[92] Ramos-Gutiérrez, B., Varela-Vaca, Á. J., Galindo, J. A., Gómez-López, M. T., and Benavides, D. Discovering configuration workflows from existing logs using process mining. *Empirical Software Engineering 26*, 1 (jan 2021).

[93] Rashid, K. M., and Louis, J. Integrating process mining with discrete-event simulation for dynamic productivity estimation in heavy civil construction operations. *Algorithms 15*, 5 (may 2022), 173.

[94] Real, E. M., Pimentel, E. P., and Braga, J. C. Analysis of learning behavior in a programming course using process mining and sequential pattern mining. In *2021 IEEE Frontiers in Education Conference (FIE)* (oct 2021), IEEE.

[95] Revina, A., and Ünal Aksu. An approach for analyzing business process execution complexity based on textual data and event log. *Information Systems 114* (mar 2023), 102184.

[96] Ridwanah, R. D., Andreswari, R., and Fauzi, R. Analysis and implementation of TELKOM university lecture business processes evaluation on heuristic miner algorithm: A process mining approach. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)* (jan 2022), IEEE.

[97] Rismanchian, F., Kassani, S. H., Shavarani, S. M., and Lee, Y. H. A data-driven approach to support the understanding and improvement of patients' journeys: A case study using electronic health records of an emergency department. *Value in Health 26*, 1 (2023), 18–27.

[98] Rojas, G. S. P., and Armas-Aguirre, J. Integration method to protect the privacy and security of information in process mining projects: a case study on surgery block. In *2021 IEEE Sciences and Humanities International Research Conference (SHIRCON)* (nov 2021), IEEE.

[99] Ruschel, E., de Freitas Rocha Loures, E., and Santos, E. A. P. Performance analysis and time prediction in manufacturing systems. *Computers &amp Industrial Engineering 151* (jan 2021), 106972.

[100] Salih, F. I., Ismail, S. A., Hamed, M. M., Mohd Yusop, O., Azmi, A., and Mohd Azmi, N. F. Data Quality Issues in Big Data: A Review. In *Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018)* (2019), Springer, pp. 105–116.

[101] Sanchez-Segura, M.-I., González-Cruz, R., Medina-Dominguez, F., and Dugarte-Peña, G.-L. Valuable business knowledge asset discovery by processing unstructured data. *Sustainability 14*, 20 (oct 2022), 12971.

[102] Saralaya, V., Saralaya, S., Kotian, L., Miranda, A., Bekal, I., and Jyothi, Y. Application of process mining for tuberculosis testing process. In *2022 IEEE 7th International conference for Convergence in Technology (I2CT)* (apr 2022), IEEE.

[103] Schuh, G., Gützlaff, A., Schmitz, S., Kuhn, C., and Klapper, N. A methodology to apply process mining in end-to-end order processing of manufacturing companies. In *Lecture Notes in Mechanical Engineering*. Springer Singapore, oct 2021, pp. 127–137.

[104] Sohail, S. A., Bukhsh, F. A., and van Keulen, M. Multilevel privacy assurance evaluation of healthcare metadata. *Applied Sciences 11*, 22 (nov 2021), 10686.

[105] Song, W., Chang, Z., Jacobsen, H.-A., and Zhang, P. Discovering structural errors from business process event logs. *IEEE Transactions on Knowledge and Data Engineering 34*, 11 (nov 2022), 5293–5306.

[106] Stephan, S., Lahann, J., and Fettke, P. A case study on the application of process mining in combination with journal entry tests for financial auditing.

[107] Tang, J., Liu, Y., yi Lin, K., and Li, L. Process Bottlenecks Identification and Its Root Cause Analysis Using Fusion-based Clustering and Knowledge Graph. *Advanced Engineering Informatics 55* (jan 2023), 101862.

[108] Tariq, Z., Charles, D., McClean, S., McChesney, I., and Taylor, P. Anomaly detection for service-oriented business processes using conformance analysis. *Algorithms 15*, 8 (jul 2022), 257.

[109] Tariq, Z., Charles, D., McClean, S., McChesney, I., and Taylor, P. Time efficient end-state prediction through hybrid trace decomposition using process mining. In *2022 14th International Conference on*

Computational Intelligence and Communication Networks (CICN) (dec 2022), IEEE.

[110] Tavakoli-Zaniani, M., Gholamian, M. R., and Hashemi-Golpayegani, S. A. Improving heuristics miners for healthcare applications by discovering optimal dependency graphs. *The Journal of Supercomputing 78*, 18 (jun 2022), 19628–19661.

[111] Tavazzi, E., Gerard, C. L., Michielin, O., Wicky, A., Gatta, R., and Cuendet, M. A. A process mining approach to statistical analysis: Application to a real-world advanced melanoma dataset. In *Lecture Notes in Business Information Processing*. Springer International Publishing, 2021, pp. 291–304.

[112] Theis, J., Galanter, W. L., Boyd, A. D., and Darabi, H. Improving the in-hospital mortality prediction of diabetes icu patients using a process mining/deep learning architecture. *IEEE Journal of Biomedical and Health Informatics 26*, 1 (2021), 388–399.

[113] Thiyagarajan, G., and S, P. A process mining approach to analyze learning behavior in the flipped classroom. In *2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4)* (dec 2021), IEEE.

[114] Tridalestari, F. A., Warsito, B., Wibowo, A., and Prasetyo, H. Analysis of e-commerce process in the downstream section of supply chain management based on process and data mining.

[115] Valensia, L., Andreswari, R., and Fauzi, R. Implementation of process mining to discover student learning patterns using fuzzy miner algorithm (case study: Learning management system (LMS) telkom university). In *2021 3rd International Conference on Electronics Representation and Algorithm (ICERA)* (jul 2021), IEEE.

[116] van der Aalst, W. M. P. *Process Mining*. Springer-Verlag Berlin Heidelberg, 2011.

[117] van der Aalst, W. M. P. Process Mining: Discovering and Improving Spaghetti and Lasagna Processes. In *2011 IEEE symposium on computational intelligence and data mining (CIDM)* (2011), IEEE, pp. 1–7.

[118] van der Aalst, W. M. P. Process Mining: Overview and Opportunities. *ACM Transactions on Management Information Systems 3*, 2 (jul 2012), 1–17.

[119] van der Aalst, W. M. P. *Process Mining Data Science in Action*. Springer London, Limited, 2016.

[120] van der Aalst, W. M. P. Foundations of Process Discovery. In *Process Mining Handbook*. Springer, 2022, pp. 37–75.

[121] van Hulzen, G. A., Li, C.-Y., Martin, N., van Zelst, S. J., and Depaire, B. Mining context-aware resource profiles in the presence of multitasking. *Artificial Intelligence in Medicine 134* (dec 2022), 102434.

[122] van Zelst, S. J., Mannhardt, F., de Leoni, M., and Koschmider, A. Event Abstraction in Process Mining: Literature Review and Taxonomy. *Granular Computing 6*, 3 (2021), 719–736.

[123] Wirth, R., and Hipp, J. CRISP-DM: Towards A Standard Process Model for Data Mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (2000), vol. 1, Manchester, pp. 29–39.

[124] Wisudiawan, G. A. A., and Kurniati, A. P. Process mining on learning activities in a learning management system. In *2022 24th International Conference on Advanced Communication Technology (ICACT)* (2022), IEEE, pp. 476–482.

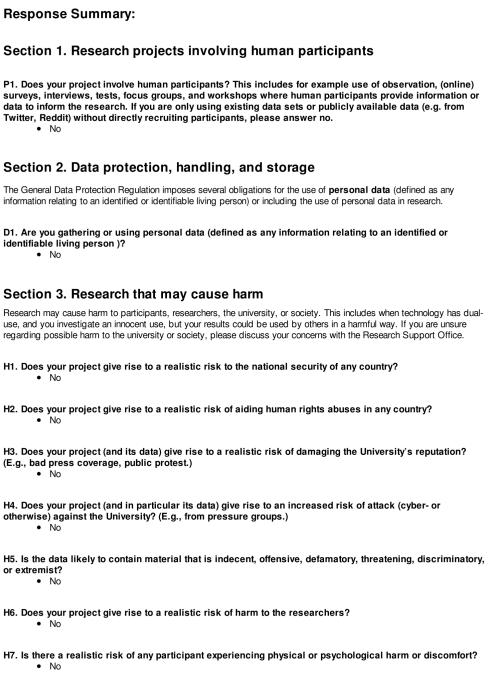
[125] Wynn, M. T., and Sadiq, S. Responsible Process Mining - A Data Quality Perspective. In *Business Process Management: 17th International Conference, BPM 2019, Vienna, Austria, September 1–6, 2019, Proceedings 17* (2019), Springer, pp. 10–15.

[126] Xiao, Y., and Watson, M. Guidance on Conducting a Systematic Literature Review. *Journal of Planning Education and Research 39*, 1 (aug 2017), 93–112.

[127] Yang, M., Moon, J., Jeong, J., Sin, S., and Kim, J. A novel embedding model based on a transition system for building industry-collaborative digital twin. *Applied Sciences 12*, 2 (jan 2022), 553.

Appendix A

Ethics and Privacy Report



Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No

H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

Figure A.1: Ethics and Privacy Quick Scan Report Page 1/3.

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

- Information and Computing Science

Z1. Your full name:
Ying Liu

Z2. Your email address:
y.liu29@students.uu.nl

Z3. In what context will you conduct this research?

- As a student for my master thesis, supervised by::
  Xixi Lu

Z5. Master programme for which you are doing the thesis

- Business Informatics

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):
x.lu@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):
g.wagenaar@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:
A Taxonomy of Data Pre-processing Techniques in Process Mining

Figure A.2: Ethics and Privacy Quick Scan Report Page 2/3.

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

Introduction

Process Mining is a discipline that intersects business process management and data science. It collects event logs data in the information system and analyzes the data through various techniques in data science. Data preprocessing is an important part of data analysis, which often occupies most of the time in the analysis process. According to different industry data characteristics, data complexity, and expected data processing results, analysts will adopt different data preprocessing methods. In most of the existing studies, especially the case studies, we can see the authors' descriptions of the data preprocessing methods used in specific domains and conditions. However, few studies have classified data preprocessing techniques in process mining, especially those combined with business context. Therefore, this study aims to give a taxonomy of data preprocessing techniques in different business contexts in the process mining discipline.

Research questions

* What is the suitable data pre-processing technique in different domains such as healthcare?
* What is the application focus of the same data preprocessing technique in different domains?

Methods

This study will adopt the method of systematic literature review. By defining inclusion and exclusion criteria, the literature is screened, and representative studies are selected for summarization. Through qualitative and quantitative analysis, research conclusions can be drawn.

Expected results

The expected result of this research is a taxonomy description of data preprocessing techniques in different business contexts. For example, the different adopted data preprocessing techniques for different analysis targets in a certain domain; the differences between different domains that applied the same data preprocessing technique.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

- Not applicable

Scoring

- Privacy: 0
- Ethics: 0

Figure A.3: Ethics and Privacy Quick Scan Report Page 3/3.