Fusing Rule and Process Mining

Ludwig Zellner

München 2024

Fusing Rule and Process Mining

Ludwig Zellner

Dissertation zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von
Ludwig Zellner
aus Traunstein

München, den 24.09.2024

Erstgutachter: Univ.-Prof. Dr. Thomas Seidl
Zweitgutachter: Distinguished Prof. Philippe Fournier-Viger, Ph.D.
Drittgutachterin: Univ.-Prof. Dr. Agnes Koschmider

Tag der Einreichung: 24.09.2024
Tag der mündlichen Prüfung: 17.01.2025

Eidesstattliche Versicherung

Hiermit erkläre ich, Ludwig Zellner, an Eides statt, dass die vorliegende Dissertation eigenständig angefertigt worden ist.

München, den 24.09.2024

Ludwig Zellner

Abstract

In many complex systems, the overall process is influenced by its sub-processes, which can be understood both vertically and horizontally. Vertically, individual components — parts of sequences — contribute to shaping the efficiency, flexibility, and outcomes of the larger process. However, sub-processes also exist horizontally, where a group of sequences collectively impacts the overall process. Both perspectives are crucial: individual parts drive the internal workings of a sequence, while the collective behavior of multiple sequences can reveal patterns, trends, and variations that influence the process on a broader scale. By understanding sub-processes in both dimensions, organizations can optimize not only internal operations but also adapt to the behavior of sequence groups, resulting in a more comprehensive improvement of the entire system.

Moreover, an event log representing a certain process consists of a spectrum of sequence frequencies. In particular, these gradations of frequencies can follow an arbitrary distribution. Hereby, we refer to frequent sequences and their associated attributes and behavior as the main process. Main processes have been the center of the research community's attention for a long time. However, the borderline of relevant sequential data and outliers is decisive, as well. Its importance stems from the fact that some deviations from the main process are large enough that their integration leads to a significant decrease in costs for process executions. Deviating process instances may occur in various scales and with various frequencies as well. Nevertheless, and up to a certain point of frequency, these instances may cause significant difficulties in the process execution. Therefore, it is essential for companies to categorize deviating behavior, hereafter referred to as micro-clusters or sub-processes, for further investigation.

Similarly, by now, the high complexity of their systems requires many companies and researchers to uncover their business processes and leverage event log analyses to prevent faulty behavior from occurring and to improve execution. Modern information and entertainment systems, for example, strongly increase their potential by recommending a course of action for users. Arising process insights from aforementioned event log analyses are a good basis to shape recommendations for further actions. It may happen that the recommendation system is inadequately trained, resulting in a restricted recommendation spectrum, which can potentially lead to a so-called "filter bubble". Considering that, recommender systems, thus, restrict the broad spectrum of articles on news websites or consciously limits the user's freedom of choice, this shows a negative development. In this context, it is crucial to use a representative data profile to capture the diverse data spec-

trum and to extend the focus beyond the data core by highlighting hidden dependencies. Moreover, not only large companies should benefit from providing recommendations for their workflow and product assortment. Given that the majority of companies in Germany are small and medium-sized enterprises (SMEs) with distinct requirements, we pursue this crucial research direction.

Yet again, hidden dependencies can be found in sub-processes and may serve as a reinforcement for recommender systems providing diversification. Additionally and with the assumption that time and order, in general, plays a vital role regarding the relevancy of sub-processes, we carry out research on similar temporal dependencies of sub-sequences, and the question if proximity of actions has an impact on an increased diversity of recommender systems is investigated.

Thus, this thesis focuses on the disclosure of sub-processes with a two-fold objective: First, to uncover deviating process instances and second with the goal of creating recommendation system from minimal sequence information. We find out that temporal information is a relevant source to reveal both the conformance of event sequences and to grasp to what extent such a sequence belongs to a group of anomalies. Additionally, we rely on the spectrum of sequence frequencies and proximity to recommend actions and support increased diversity in recommendations where sub-processes serve as a basis.

Zusammenfassung

In jedem komplexen System wird der Gesamtprozess sowohl durch vertikale als auch horizontale Teilprozesse beeinflusst. Vertikal tragen individuelle Komponenten – Teile von Sequenzen – zur Effizienz, Flexibilität und zum Ergebnis des übergeordneten Prozesses bei. Doch auch horizontal existieren Teilprozesse, bei denen eine Gruppe von Fällen gemeinsam den Gesamtprozess beeinflusst. Beide Perspektiven sind entscheidend: Einzelne Teile treiben das interne Geschehen einer Prozessinstanz voran, während das kollektive Verhalten mehrerer Sequenzen Muster, Trends und Variationen offenbaren kann, die den Prozess auf einer größeren Ebene beeinflussen. Durch das Verständnis von Teilprozessen in beiden Dimensionen können Organisationen nicht nur interne Abläufe optimieren, sondern auch das Verhalten von Sequenzgruppen berücksichtigen, was zu einer umfassenderen Verbesserung des gesamten Systems führt.

Darüber hinaus besteht ein Event-Log, das einen bestimmten Prozess darstellt, aus einem Spektrum von Sequenzhäufigkeiten. Insbesondere können diese Häufigkeitsabstufungen einer beliebigen Verteilung folgen. In dieser Arbeit beziehen wir uns auf häufige Sequenzen und deren zugehörige Attribute und Verhaltensweisen als den Hauptprozess. Hauptprozesse standen lange Zeit im Mittelpunkt der Aufmerksamkeit der Forschungsgemeinschaft. Die Grenze zwischen, für den Hauptprozess, relevanten sequentiellen Daten und Ausreißern ist jedoch ebenfalls entscheidend. Ihre Bedeutung ergibt sich daraus, dass einige Abweichungen vom Hauptprozess so signifikant sind, dass deren Integration zu einer erheblichen Senkung der Kosten für die Prozessdurchführung führen kann. Abweichende Prozessinstanzen können in verschiedenen Größenordnungen und mit unterschiedlichen Häufigkeiten auftreten. Diese Instanzen können bis zu einem gewissen Grad der Häufigkeit erhebliche Schwierigkeiten bei der Prozessdurchführung verursachen. Daher ist es für Unternehmen wichtig, abweichendes Verhalten, im Folgenden als Mikro-Cluster oder Teilprozesse bezeichnet, zur weiteren Untersuchung zu kategorisieren.

Ähnlich erfordert eine hohe Komplexität von Prozessen heutzutage, dass viele Unternehmen und auch Forscher Ereignisprotokollanalysen ihrer Geschäftsprozesse nutzen, um Fehlverhalten zu verhindern und die Durchführung der Prozesse zu verbessern. Moderne Informations- und Unterhaltungssysteme erhöhen ihr Potenzial erheblich, indem sie den Nutzern Handlungsempfehlungen geben. Die gewonnenen Prozesserkenntnisse aus den genannten Ereignisprotokollanalysen bieten dafür eine gute Grundlage.

Es kann vorkommen, dass das Empfehlungssystem nicht ausreichend trainiert ist, was zu einer Einschränkung im Empfehlungsspektrum führen kann. Das erzeugt möglicherweise

eine sogenannte „Filterblase“. Dass Empfehlungssysteme das breite Spektrum an Artikeln auf Nachrichtenwebseiten einschränken oder sogar bewusst die Entscheidungsfreiheit des Nutzers begrenzen, stellt im Allgemeinen eine negative Entwicklung dar. In diesem Zusammenhang ist es wichtig, ein repräsentatives Datenprofil zu verwenden, um das vielfältige Spektrum der Daten zu erfassen und den Fokus über den Datenkern hinaus zu erweitern, indem versteckte Abhängigkeiten hervorgehoben werden. Darüber hinaus sollten nicht nur große Unternehmen von der Bereitstellung von Empfehlungen für ihren Arbeitsablauf und ihr Produktsortiment profitieren. Da die Mehrheit der Unternehmen in Deutschland kleine und mittelständische Unternehmen (KMU) mit ihrerseits spezifischen Anforderungen sind, verfolgen wir diese wichtige Forschungsrichtung.

Gleichwohl können versteckte Abhängigkeiten in Teilprozessen gefunden werden und als Verstärkung für Empfehlungssysteme dienen. Damit kann die Diversifizierung erhöht. Zusätzlich und unter der Annahme, dass Zeit und Reihenfolge eine entscheidende Rolle hinsichtlich der Relevanz von Teilprozessen spielt, erforschen wir ähnliche zeitliche Abhängigkeiten von Teilsequenzen und untersuchen die Frage, ob die Aktualität von Handlungen Auswirkungen auf eine erhöhte Diversität von Empfehlungssystemen hat.

Aufgrund der oben genannten Ansatzpunkte, konzentriert sich diese Arbeit auf die Offenlegung von Teilprozessen mit einem zweifachen Ziel: Erstens, abweichende Prozessinstanzen aufzudecken und zweitens Empfehlungssysteme aus minimalen Sequenzinformationen zu erstellen. Wir stellen fest, dass zeitliche Informationen eine relevante Quelle sind, um sowohl die Konformität von Sequenzen aufzudecken als auch zu erfassen, inwieweit eine Sequenzen zu einer Gruppe von Anomalien gehört. Darüber hinaus stützen wir uns auf das Spektrum der Sequenzhäufigkeiten und die Aktualität zwischen Events, um Handlungen zu empfehlen und die Diversität in Empfehlungen zu unterstützen, wofür Teilprozesse wiederum als Basis dienen.

Acknowledgements

I feel fortunate to have been able to navigate challenging times with the help of people who closely accompanied me during this period and had a significant impact on my life. This position and the exceptional support I received are not to be taken for granted, which is why I want to express my deep gratitude to all of you:

I would like to express my deep gratitude to my doctoral advisor, Prof. Dr. Thomas Seidl, who has been consistently supportive and always lent a sympathetic ear.

I am profoundly thankful to my former colleagues in our research area, Florian and Janina, who introduced me to the field of process mining and guided me in all aspects of the life as a research assistant, a Jack of all trades.

Beyond that, I would like to thank all my colleagues who supported me throughout my journey, whether they joined earlier or later, especially Andrea and Simon, whose collaboration always brought me joy. Also, without Susanne's organizational skills and the technical advice from Franz and Joao, my journey would not have been possible in this way.

I am very happy to have friends who always stand by me with a listening ear and encouraging words. A special thanks belongs to Kassandra, who accompanied me especially in the beginning of my journey. Difficult times require special support, and I have relied on my parents for this as well.

Last but certainly not least, Viktoria. We began finalizing our theses simultaneously, but you far outpaced me. I am deeply grateful to you for sharing the weight with me, even though it was not yours to bear. Having someone who truly understood what I was experiencing was and is a gift.

Table of Contents

List of Figures

List of Tables

Acronyms

BPM Business Process Management. 9

BPMN Business Process Model and Notation. 1

HATC Hierarchical Agglomerative Trace Clustering. 16

IALD Intra-List Diversity. 26

IELD Inter-List Diversity. 26

KDE Kernel Density Estimator. 6

KPI Key Performance Indicator. 9

nDCG Normalized Discounted Cumulative Gain. 24

OCPM Object-Centric Process Mining. 8

POSR Partially-Ordered Sequential Rules. 20

RQ Research Question. 10

SeqDiv Sequence Diversity. 27

SME Small and Medium-Sized Enterprises. 4

TOAD Trace Ordering for Anomaly Detection. 16

TOSR Totally-Ordered Sequential Rules. 19

Chapter 1

Introduction

We have two ears and one mouth,
so we should listen more than we say.

Zeno of Citium –
Hellenistic Philosopher
and Founder of the Stoa

1.1 Motivation

One possible interpretation of this ancient quote is quite literal. Gathering information with our ears and spending a reasonable time to process and analyze it, leads to well-founded and sophisticated insights. This quote still fits exceptionally well to present times: As the amount of data that is available to us rises unimaginably fast, it is vital to take the time and analyze, derive and infer with reason.

1.1.1 Process Mining and Order

An essential area in which data is gathered and reflexively utilized is the area of Process Mining. In general, data is categorized differently and various perspectives lead to various insights especially in this area. It focuses on operational processes, i.e., processes requiring the repeated execution of activities to deliver products or services [3]. Extracting a representative model from process data not only proves to be difficult as it may contain multiple types of data quality problems, e.g., data from defective sensors or even missing entries [4]. Process Mining also has the obligation to find an appropriate balance between the four quality dimensions: fitness, simplicity, precision, and generalization. [5]. This not only applies to classical process models when using Petri Nets or BPMN to capture the process control-flow. Any kind of latent representation can be utilized to capture process characteristics including, e.g., graphs and sequential rules.

On a ground level, the order in which entities are shown is an abstract but vital aspect

to processes. Abstract, because the information has to be derived from the actual data, e.g. with timestamps or sequence analysis and vital, because processes are designed in a way to intrinsically contain a notion of order as the execution of activities follows a predefined succession. This leads to process data being an optimal candidate to apply methods based on order and succession.

1.1.2 Frequency and Deviations

There is yet another essential aspect influencing many algorithms dedicated to finding a model representation: frequency. Independent of focusing on frequency on sequence or on event level, it is a leading point when it comes to assessing relevancy. However, this does not mean that infrequent entities do not contribute significantly to a process [6]. In the following, we use the term *trace* as a sequence of activities that represents the execution of a specific process instance. For example, a set of traces can be most relevant if they represent anomalous process behavior. In cases where the frequency of such deviating sequences differs widely from the frequency of instances in the main process the number has to be analyzed in context of other deviations. In many cases, it is still worth and recommendable to reintegrate such micro-clusters of processes (hereafter referred to as *sub-processes*). On the one hand, handling a set of deviations at once reduces costs. On the other hand, it enhances the main process. Deviations occurring with a certain frequency cannot be overlooked, as they may indicate unmet demands or other significant conditions that require attention. Thus, it is most important to treat frequency as an aspect not only regarding the overall process but, especially, in the context of an appropriate comparison.

1.1.3 Temporal Aspects

One key aspect in determining succession and order is time. Additionally, the combination of time and frequency are popular means to define the main process and distinguish anomalous traces. Time and order in processes are strongly linked as there are sub-forms of traces. Since processes often consist of sub-processes both horizontally (across cases) and vertically (within case parts), managing time and order is crucial at different levels of the process hierarchy. Hence, it stands to reason to break open the trace as a whole and analyze parts by focusing on different intensities of order [7]. Let us have a look at an example: When configuring a custom computer, many components must be selected, each with dependencies on others. The casing has to have a distinct size such that the motherboard with a certain form factor and the cooling system fits. Additionally, the type of graphics card, again, depends on the form factor of the main board and so on. We may deduce that motherboard, CPU and graphics card are selected first, while casing and power supply unit are selected last after the size and power demand of each part of the computer is decided. While the majority of customers may select the parts in this order there yet could be another group that prioritizes design instead of computing power. Here, behaviour and interests have an impact on the order. The task is now to sort out which

items are selected in an arbitrary order and which items are selected because there actually exists a significant dependence.

1.2 Objectives of the Thesis

This thesis explores two main areas. First, it focuses on process mining, particularly the interplay between activities and their relevance in relation to deviating behaviors (cf. Figure 1.1). Second, it examines the integration of rules in process mining, using frequency and order as key indicators of relevance. Going into detail regarding the former key point,

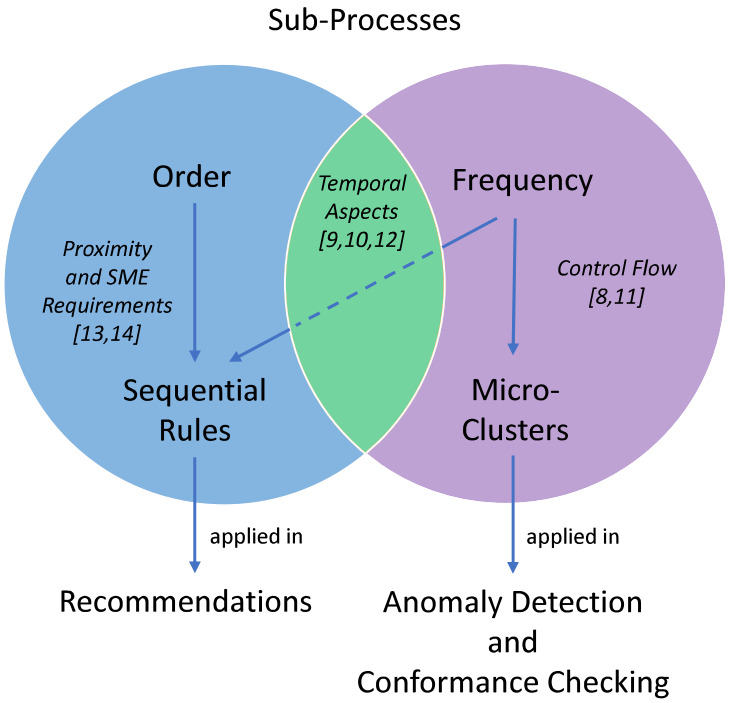


Figure 1.1: Venn diagram illustrating the main topics and their interrelations within this thesis. Overarching concepts are placed inside the circles, with utilized properties italicized, and downstream tasks positioned at the bottom. The assignment of reference numbers to publication titles is addressed in Section 1.3.

one way to assess relevancy is by using the distance to the main process and cluster traces to create groups of deviations. In our work, we aim at this idea two-fold: On the one hand, we want to detect these groups on a static process. On the other hand, we additionally want to apply this approach to continuously incoming traces. This area is represented in Figure 1.1 on the right hand side labeled "Control Flow".

Additionally, we want to incorporate the temporal perspective into the areas of anomaly detection and conformance checking. Here, we develop methods that, on one side, include temporal aspects to find deviating sub-processes and, on the other side, test traces for

conformance based on the temporal behavior of the main process. This area is characterized by an overlap of the overarching concepts of order and frequency, and marked with "Temporal Aspects" in Figure 1.1.

Not only the order of activities but also the frequency, previously mentioned as the second of two key points, directly impact the the behaviour of the process we are examining. In this context, we aim at searching for possibilities to enhance recommendation systems. Our goal is to balance precision and diversity to be able to fulfill user demands while not leaving out parts of the process. We focus on challenges such as small amount of provided data, sparsity of data, lower computational capabilities and lack of additional data which frequently occur in the area of small- and medium-sized enterprises (SMEs). Figure 1.1 illustrates this on the left-hand side, labeled "Proximity and SME Requirements".

In summary, we want focus on sub-parts of processes, which represent essential units themselves and analyze them principally regarding frequency and order. Based on these attributes we target downstream tasks such as anomaly detection, conformance checking and recommending events.

1.3 Overview of Contributions

After marking out the objective of this thesis we go into detail regarding the concrete set of contributions. Additionally, the main topics and the corresponding publications are outlined in Table 1.1. This table highlights the various areas addressed in combination across the publications.

This thesis begins by focusing on process mining, specifically analyzing traces that deviate from the main process. A significant area in process mining is conformance checking, which involves validating whether traces align with the boundaries of the main process. These boundaries can be based on either predefined or mined process models. For predefined models, often used in normative settings, deviations may indicate fraud or inefficiencies. In contrast, for mined models, deviations could represent exceptional behavior [3]. Understanding deviations is especially important given that most processes follow a Pareto distribution: approximately 20% of process variants account for 80% of the cases. This pattern continues, with the remaining 20% of cases covering 80% of the remaining variants [3]. This highlights the need for careful analysis of deviations to fully understand process performance and behavior.

We state that deviating traces should not be forcibly avoided since they may contain additional value concerning the process. This additional value can manifest in a way which can not be, or at least is not, represented in the main process, yet. With the increasing amount of deviating traces of the same type the probability of added value increases upon their reintegration. This assumption is driven by the aim of achieving a more cost-effective handling of these traces. We begin by addressing the issue of multiple cases exhibiting similar deviation characteristics, as discussed in [8]. To achieve this, we use the process model as a baseline to group similar deviating cases by referring to it as a map-like ground distance.

Table 1.1: Overview of the topics covered by each publication (as listed below), irrespective of the proportion in which each topic appears.

	Traditional Process Mining					Rule Mining	
Topics / Publ.	Sub-Processes	Conformance Checking	Anomaly Detection	Temporal Perspective	Concept Drift Detection on Trace Streams	Rule Mining and Gap Constraint	Recommendation Systems
[8]	■	■	■				
[9]	■		■	■			
[10]	■	■		■			
[11]	■	■			■		
[12]	■		■	■			
[13]	■					■	
[14]	■					■	■

Table 1.2: Publications included in the thesis, referenced by aforementioned tabular overview. The order has been adjusted for clarity and to distinguish between the main topics, and may not correspond to the order presented in the preceding or following text.

[8] Model-Aware Clustering of Non-Conforming Traces (s. Appendix B)

[9] TOAD: Trace Ordering for Anomaly Detection (s. Appendix A)

[10] TADE: Stochastic Conformance Checking Using Temporal Activity Density Estimation (s. Appendix C)

[11] Concept Drift Detection on Streaming Data with Dynamic Outlier Aggregation (s. Appendix D)

[12] Process Mining Techniques for Collusion Detection in Online Exams (s. Appendix E)

[13] SCORER-Gap: Sequentially Correlated Rules for Event Recommendation Considering Gap Size (s. Appendix F)

[14] On Diverse and Precise Recommendations for Small and Medium-Sized Enterprise (s. Appendix G)

With our approach TOAD [9], we go a step further by organizing cases in a vector embedding based on their mutual distances, allowing us to identify clusters of anomalies. This method again focuses on detecting multiple cases with similar deviation characteristics.

A third endeavour refers to the task known as conformance checking. With TADE [10] we test cases for conformance in a time-based manner. On that account, we utilize a kernel density estimator (KDE). We split up cases into events and calculate the occurrence probability at a specific point in time.

Another contribution in this regard, is the extension of [8] to the area of streaming data. This extension allows us to analyze the process further by detecting concept drifts in the data. Concept drifts have different manifestations such as sudden, gradual, incremental and recurring [15, 16, 17]. This work focuses on incremental and recurring drifts. In addition to broad applicability, this approach provides a progressively deeper understanding of the process over time (cf. [11]).

With an additional work, we adapt process mining techniques to online exam data of a university and solve the task of detecting colluding students in these exams. Our contribution includes adapting hierarchical agglomerative clustering and comparing it to the application of our proposed approach, TOAD. It yields promising results by revealing collective anomalies, here colluding students (cf. [12]).

The research field of frequent pattern and rule mining contains similar questions. The approaches and solutions from this area pose great opportunities to be adapted and extended. Grouping similar items to patterns and creating rules from them which imply the succession of patterns is another way to assess relevancy based on frequency. Rule mining approaches frequently suffer from the issue of producing a vast amount of rules. The next contribution addresses this with the assumption that recency and proximity have a high impact on the relevancy of rules. By penalizing rules based on the gap between antecedent and consequent we manage to decrease the amount of rules significantly while maintaining accuracy (cf. [13]).

Rule mining is used in different areas from which a popular one is the field of recommendation systems. The implication of one pattern based on another is naturally suitable to make recommendations. Here, again, the vast amount of matching rules in a certain state makes it difficult to select the most appropriate one. Moreover, varying companies necessitate distinct solutions contingent upon a different set of conditions. Given these conditions, we contribute by developing and comparing variants of recommender systems tailored to these specific circumstances. These variants encompass a ranking algorithm grounded in proximity as well as other methods designed to enhance diversity (cf. [14]).

The remainder of this thesis is structured as follows: Chapter 2 introduces fundamental knowledge required to follow the contributions. This information is divided in different sections such as Section 2.1 which provides basics regarding process mining, Section 2.2 is about defining and detecting deviating traces. It includes information about concept drifts and reflects upon catalysts for ill-fated processes. Section 2.3 addresses the area of rule mining and the importance of order and proximity in this regard. Based on that Section 2.4 takes up this information and extends it for further use in recommendation systems and its evaluation.

Chapter 2

Fundamentals and Background

The focus of this thesis is the fusion of rule and process mining. To understand the similarities, differences, advantages and limitations this chapter provides an overview and classifies subject-related terminology thematically.

2.1 Events, Event Logs and Object-Centric Process Mining

To follow this thesis, it is essential to understand what processes are. According to the Britannica Dictionary

> "[A process is] a series of actions that produce something or that lead to a particular result." [18]

In the process mining community this definition frequently tailored towards business processes:

> "The focus of process mining is on operational processes, i.e., processes requiring the repeated execution of activities to deliver products or services." [19]

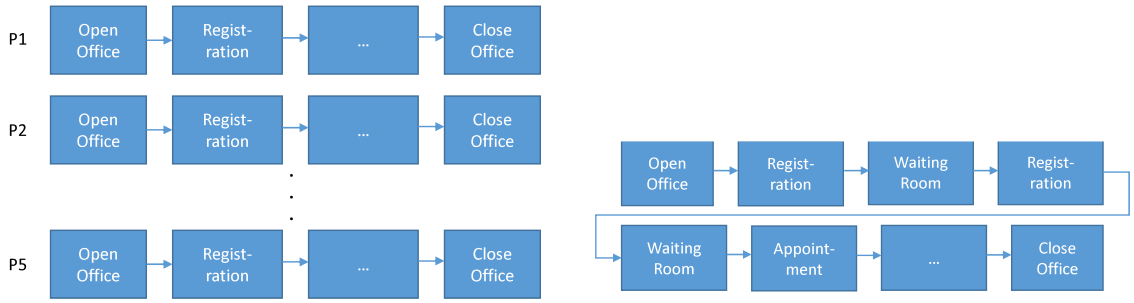
In any case, processes can vary greatly and occur in most fields. Many series of actions can be understood as a process ranging from cooking by following a recipe to conducting an examination.

Events are the center point of processes and, thus, process mining. They can be imagined as the smallest unit of a process consisting of different attributes. A case identifier, an activity label and in most circumstances a timestamp are fixed attributes. When events are grouped by their case identifier, we retrieve cases consisting of a sequence of events. The order of the sequence is either intrinsically given by the occurrence in the event log or it can be found out with the help of given timestamps. The event log, again, is the database that contains the set of cases. We can strip each event in a case from all attributes except the activity label which contains information about the action this event

represents. Thus, we are left with the raw sequence of actions. It stands to reason to apply methods from data mining which are designed to deal with sequences of entities such as items or sets of items. These sequences already contain plenty of additional information such as recurring patterns with or without including the order, implication between items, itemsets or patterns and the gap between the occurrence of one pattern (antecedent) and the triggering of another (consequent). Hence, aforementioned sequences represent a raw version of process instances with order and gap information, events are an enriched version of items and cases are the extended pendant to sequences on process mining level.

Since 2016, traditional process mining is gradually extended by Object-Centric Process Mining (OCPM) [5, 1]. As mentioned before an event gets assigned a case id that associates this event with a certain case. Actually, there are many possibilities to relate an event to a case depending on which attribute to be selected. Hence, event logs represent an already fixed view on a process. This leads to the possibility of shifting focus between entities of a process by deriving the case id from another attribute. Under this circumstance multiple problems may arise. We take a simplified process of patients having an appointment at a doctor's office with activities *open office, registration, waiting room, appointment*, and *close office* in one day as an example (refer to Figure 2.1).

If we take the perspective of each patient's procedure as a case and we have five patients, this would leave us with five cases each starting and ending with the activities *open* and *close*. Hence, the issue of having to replicate an event, namely the opening i.e. closing of the doctor's office to represent different cases in a log is called *convergence* (see Figure 2.1a).



(a) Simplified example showing that the events *open office* and *close office* both are related to multiple cases.

(b) Divergence leads to loops in process models because repeating events can not be distinguished.

Figure 2.1: Convergence and divergence are considered problems under the assumption that there is only a single process view and each event refers to exactly one case [1].

Another issue that arises with traditional process mining is called *divergence*. After shifting the focus to the doctor, we obtain multiple events in one case which refer to the activities *registration* and *waiting room* because patients not only arrive just before their appointment but arrive in random order especially when we deal with short appointments of ten or 15 minutes (see Figure 2.1b). As a consequence, thereof, we lose dependencies and causality in the worst case between events as, for example, sub-

processes including any orderings are possible. Thus, we may perceive parts of a sequence like $\langle registration, appointment, appointment, registration \rangle$. The reason is that, from the doctor's perspective, these events do not correspond to a specific patient. Object-centric process mining is tailored towards solving these issues and preventing from convergence and divergence [20]. Certainly, there are further benefits of applying an object-centric perspective such as single-time data extraction, analysis of multi-type object relationships and three-dimensional perspective in object-centric process models [21]. Nevertheless, this thesis focuses on traditional process mining, as it is not entirely distinct from OCPM. Traditional process mining still allows for the inclusion of various perspectives, even though these perspectives may not revolve around a specific object. In fact, the flexibility of traditional process mining enables the integration of multiple viewpoints, making it a versatile approach that can address complex processes without the need for object centrism. This highlights the complementary nature of both methods.

2.1.1 Key Performance Indicators as a Leverage Point

Every software developer, early in their career, likely encounters the concept of refactoring. This involves restructuring and redesigning the code base to improve readability and comprehensibility. Similarly, in Business Process Management (BPM), processes must be regularly reconsidered, evaluated, and, when necessary, redesigned [22]. Just as refactoring in software ensures cleaner code, refining business processes ensures greater operational efficiency and effectiveness.

In BPM, evaluating process performance is essential for companies to assess current conditions and derive actionable insights. Key Performance Indicators (KPIs), also known as *process performance measures* in Process Mining, serve as critical metrics that reflect the capacity of a process in specific dimensions. These metrics allow companies to track progress, define targets, and set improvement goals across various perspectives [23].

The effectiveness of process management depends not only on the overall flow but also on the details of individual events within the process. Many KPIs are heavily influenced by the specific attributes of these events. Consequently, understanding KPIs is crucial for assessing the current state of a process and identifying opportunities for improvement in associated entities. As discussed earlier, activities within processes offer a strong basis for recognizing areas in need of optimization.

Among the many dimensions that KPIs measure, time plays a particularly critical role in evaluating business processes. Time-based metrics are fundamental to understanding and optimizing process performance. For companies seeking to improve their operations, efficiency is often a key priority, and process efficiency is frequently measured by time. By analyzing temporal metrics, organizations can pinpoint bottlenecks, streamline operations, and enhance overall performance. Thus, incorporating the temporal perspective is essential for a comprehensive assessment of process effectiveness and for driving meaningful improvements in efficiency [24]. Since many leverage points are linked to anomalous behaviour regarding time, the temporal perspective becomes a crucial focus in our analysis of process deviations.

With this foundational understanding of KPIs and the importance of temporal factors in process performance, we now turn to the specific research areas that underpin this work. In the following sections, we briefly summarize related publications and highlight the research questions (RQs) that have shaped our investigation.

2.2 Anomalies and Deviations

Information in general can be treacherous. It is a crucial component to start analyzing an issue but one has to be aware of the context it emerges from. Figure 2.2 shows how different angles on a problem can be pivotal on how a conclusion is drawn. Additionally, it reveals how one solution does not automatically exclude another.

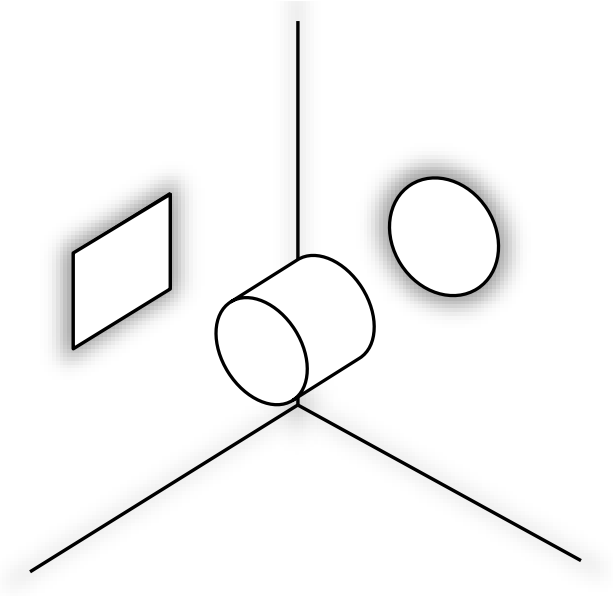


Figure 2.2: The same entity possesses different sometimes seemingly contradictory information based on the perspective from which we look at it.

In traditional process mining, the event log is a representation, a mere snapshot of a process during a fixed amount of time. It can be used to take up different perspectives from which the underlying process is analyzed. It consists of various information comprised in different attributes across the same level, e.g. activity and case level. Based on this information significant perspectives have been determined such as control-flow, data, resource time and function perspective, as Mannhardt states [25]. The author also mentions that these five perspectives do not encompass all possible perspectives. He notes that *costs* or *risks* are additional views that can be analyzed. While it can be claimed that *costs* are part of the data perspective since this information might be required for control-flow decision in a process instance, *risk* is a far more broad term that describes the occurrence of events that contain a negative impact. The stance of this perspective can be taken by combining

multiple attributes from the event log. Since these perspectives continue on the activity level and interplay between activities inevitably exists, perspectives should be measured in these groups.

In a similar manner, the concept of perspective is also relevant in the study of data deviations. In our work, we categorize anomalies not primarily based on their content or semantic value, but rather according to the environments in which these deviations occur. A seminal survey about this topic was written by Chandola et al. [26]. Especially, the division of anomaly types into *point*, *contextual* and *collective anomalies* reflects the issue that anomalies are manifold. Additionally, it shows that certain types only occur from specific views. In this work, we want to focus on the two latter types. We start with the following example: Table 2.1 exemplarily shows the average daily temperature (°C) over

Table 2.1: Average daily temperature (°C) for two weeks in August comparing Berlin and Rome.

City	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Berlin	19.7	20.9	17.3	17.6	18.2	22.0	22.7	22.7	22.7	22.8	23.1	25.4	27.1	25.6
Rome	20.1	27.5	23.2	5.1	22.4	23.6	24.1	24.5	20.8	24.5	26.5	22.8	26.3	23.1

two weeks in August 2023. Regarding Berlin, the sequence of temperatures does not attract attention, but if we take a closer look, the three equal successive temperature values on days seven to nine may be considered a collective anomaly. The sequence concerning Rome exhibits deviating behavior on day four. This is considered rather a contextual anomaly than a point anomaly. The reason behind is that 5.1°C would not stand out in the winter month but having this temperature in August is uncommon. We take one step further and compress these data via run-length encoding. This yields the following:

```
Germany 1*19.7 1*20.9 1*17.3 1*17.6 1*18.2 1*22.0 3*22.7
        1*22.8 1*23.1 1*25.4 1*27.1 1*25.6
Italy   1*20.1 1*27.5 1*23.2 1*5.1 1*22.4 1*23.6 1*24.1
        1*24.5 1*20.8 1*24.5 1*26.5 1*22.8 1*26.3 1*23.1
```

With this data representation the three successive temperature values can be considered a point anomaly because any other temperature value only occurs once most probably because of its precision with one decimal place. The gist of this remark is that anomalies and deviations are not unambiguously identifiable but mostly depend on other factors such as data representation, perspective and the comparison with other entities.

Regarding our work [9] we develop a method to detect collective anomalies while still considering the data context. We focus on the temporal perspective and put traces of a process at the heart of the analysis. This approach is motivated by the ability to aggregate similar deviations and handle them at once as a cost-efficient reaction. In our context, we call them *micro-cluster* anomalies, since many manifest themselves as batches of traces that are significantly smaller than the main process. These clusters can be regarded as a type of anomalous sub-process.

Our approach consists of four steps: First, the relation duration set between each activity pair over the whole dataset is calculated. Second, the arithmetic mean and the standard deviation of these duration sets are extracted. Afterwards, we apply z-scoring to align short- and long-term activities in the same model. Then, a temporal deviation signature in the format of a vector is derived from the results on which we apply OPTICS [27]. Lastly, we make use of the Savitzky-Golay filter to detect collective deviations by the size and steepness of their trough in the reachability plot.

In this work we address the following research questions: How does the approach perform on an anomaly detection task regarding fitness and precision scores (**RQ 1**)? Moreover, how does the performance of the approach change when the parameters of the underlying methods are adapted (**RQ 2**)?

In conducting experiments on the BPIC 2017 [28] and BPIC 2015 [29] datasets, we found that the MinPts parameter of the OPTICS algorithm [27] significantly impacts the results. Specifically, smaller MinPts values tend to highlight anomalies containing only a small amount of traces, whereas larger values often result in many anomalies being overlooked. Additionally, TOAD demonstrates the greatest potential when trace lengths are homogeneous. This is because shorter traces exhibit smaller distances, making them more likely to be classified as anomalies due to their size. Another key insight is the performance difference between precision and recall. In this context, precision is of greater importance because while missing an anomalous trace is acceptable, misidentifying and analyzing false anomalies can be time-consuming and resource-intensive. Therefore, even when TOAD occasionally fails to capture the entire collective anomaly but can identify the region, it is crucial to investigate the areas surrounding the detected dense cores.

We thus recommend thorough pre-processing before applying TOAD and a thoughtful examination of the results and their boundaries.

2.2.1 Non-Conforming Traces

Koschmider et al. [30] extend this understanding of anomalies with a differentiation between noise and outliers. This is especially crucial in the area of process mining as the source of errors needs to be identifiable. The reason is that errors in data recordings present a different challenge than errors accurately recorded by sensors but reflecting unexpected or unknown process behavior. In this thesis, we focus on the latter, specifically on anomalies arising from correctly functioning sensors that capture unforeseen process behavior.

Popular quality measures for relating observed processes and the corresponding models are *Recall*, *Precision*, *Generalization* and *Simplicity* [2, 5, 19]. Discovering models from observed process data requires a continuous trade-off between these dimensions. Equally, checking cases for conformance on the process model and its result depends heavily on the representation and validity of the model. Buijs et al. [2] also displayed the relation between system behavior, process model, and event log as a Venn diagram. According to the authors, a *system* can refer to both a specific information system implementation and the broader process context, where deviations from the system's intended behavior by individuals involved in operations may occur for legitimate reasons. We adopt and

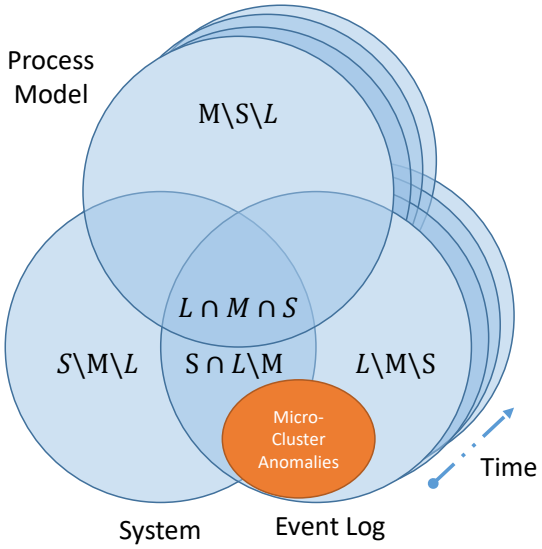


Figure 2.3: Extended Venn diagram showing the relation between system S, process model M and event log L [2]. We focus on detecting and clustering anomalies in the realm consisting of recorded process behaviour only $L \setminus S \setminus M$ and in intersection with real system behaviour $S \cap L \setminus M$.

extend this notion by the temporal aspect in Figure 2.3. We notice that the target area encompasses both recorded behavior that is inherent to the system and recorded behavior that is not inherent to the system.

The reason is that, on the one hand, we want to detect process deviations which are unintentionally conducted with the goal of reaching a certain target more quickly but in an unintended way. This could be the case, when employees have to act quickly and deviate from a given protocol, which is also called a "Breaking the Glass"-motive [5] in the context of hospitals. Detecting these types of deviations is crucial because integrating them into the current main process may present a decisive opportunity for improvement.

On the other hand, deviations which are recorded but not inherent to the system are not less vital to disclose. For example, external attacks on the process can cause these types of deviations that pose a potentially existential threat [31]. Thus, in this area, we are interested in a novel approach to cluster collective anomalies. Hence, in [8], we solely focus on deviations and investigate how suitable a process model is to cluster collective anomalies in comparison to traditional trace clustering with unigrams and bigrams (**RQ 3**).

In this approach, we begin by using the main process in the form of a process model, which is either created using a process discovery algorithm or directly provided by the process owner. Next, we check the event log for conformance and focus on traces that do not conform. The reason for this is that we want to group similar anomalous traces together based on the type of anomaly they exhibit. We group these traces by applying hierarchical agglomerative clustering with the average linkage criterion. For that, we combine every

non-conforming trace with each other. In detail, we compute the distance between each transition that is marked problematic by the conformance checking procedure. Eventually, the average of the accumulated distances yields our result of one combination of anomalous traces. The hierarchical clustering approach then is able to group all anomalous traces based on the structure of their anomaly by using the process model as a reference.

We compare this approach to traditional trace clustering on unigrams and bigrams, also by using hierarchical agglomerative clustering and the manhattan distance. To achieve that, we first vectorize traces by means of counting the occurrences of each activity (unigram) and counting the occurrences of each activity relation (bigrams). As the structure of the activities is neglected when using unigrams the approach results in one large cluster. Regarding the usage of bigrams the traditional trace clustering yields many small clusters as the control-flow is only partially contained causing the dimensions to only represent local behaviour. These small cluster variants are mainly caused by concurrency in the process and by a missing reference to the process model. Hence, this approach proves to be advantageous over traditional trace clustering by being able to focus the anomalous trace parts.

Another study of ours in this context, called TADE [10], examines the time perspective for conformance checking as a downstream task. We employ a Kernel Density Estimator (KDE), using time data as input to determine the probabilities of events occurring at specific timestamps against which traces can be checked. This means that once our KDE model has been trained, we apply it to the actual trace execution. In detail, we check if a case is conform by considering all events in this case and apply the corresponding probability density function for them. We calculate the arithmetic mean of all estimated probability values for one case to derive the temporal stochastic conformance fitness. This represents the likeliness of a case to belong to the specified process. Specifically, in this work we analyze the effect of different kernels and corresponding bandwidth parameter (**RQ 4**), assess different aggregation methods for the event probabilities (**RQ 5**), and examine two variants of the approach regarding classification accuracy and runtime (**RQ 6**). The two variants on the one hand comprise left-aligning the temporal perspective of all cases, i.e. the time difference between the first event and all subsequent events. On the other hand, we use the full Cartesian activity set, meaning that we design kernel functions for all relations of two temporally succeeding activities.

Our results demonstrate that both TADE and TADE with the full Cartesian approach (TADE-FC) outperform token-based replay as a conformance checking method in both the classification task and runtime comparison. E.g. we achieve an F1-Score of 0.46 and 0.73 on the first sublog of BPIC 2015, respectively. Comparing runtime with one minute TADE is 60 times faster than TADE-FC followed by token-based replay with around 1 hour and 20 minutes. However, since TADE primarily focuses on the temporal perspective, this comparison is vulnerable as token-based replay prioritizes control-flow. The main goal of TADE is not to outperform workflow-based conformance checking methods. However, TADE excels in filtering out non-conforming traces based on their temporal fitness using a straightforward stochastic technique, without the need to derive complex workflow models. For the remaining cases, other resources can be employed for a more precise analysis.

2.2.2 Concept Drifts

Figure 2.3 yet shows another aspect when it comes to relating system behaviour, event logs and process models. The system is subject to external impacts, which transitively affect the event log. This is especially true in online settings, where there is a constant flow of input data, also called streams. Here, the temporal aspect among others [16] plays a crucial role because most processes change over time underlying different factors. Examples for this can be legal or regulatory requirements as it is the case with the BPI Challenge dataset of 2015 [29]. In general, the term concept drifts is defined early in the area of machine learning and data mining as a state where the relation between the input data and the target variable, that has to be predicted by some model, changes unexpectedly over time [32]. In process mining, there are two categories to be differentiated when we deal with concept drifts during the analysis of even logs: Offline analysis and online analysis [33]. The former handles data from a process that already has been concluded. The latter deals with data in stream-like settings where sudden reactions are required to adapt the underlying process. Bose et al. [33] also identified four types of drifts, namely *sudden*, *recurring*, *gradual*, and *incremental* drifts. We provide coarse explanations for the four types: In a sudden drift an old version of a process gets fully replaced and all ongoing cases are handled by the new process version. A gradual drift includes the coexistence of two process versions for particular time range. A recurring drift incorporates two or more process versions that intermittently get activated. It is important to understand that the drift between processes can take place gradually as well as suddenly. The incremental drift is defined by minor changes in the process that led to many versions over time. Here, too, the changes can represent a sudden or a gradual drift.

A recent survey by Sato et al. [16] highlights additional aspects of defining concept drift, beyond the familiar factors such as duration, online or offline analysis, and dynamics, with particular emphasis on the different perspectives in which a concept drift can occur. The latter aspect is closely related to Section 2.2 since anomalies are ubiquitous and not only based on different types of processes but also different dimensions within. With our paper [11], we take this up and take it one step further by narrowing down the area in which concept drifts occur. We focus on anomalous traces in which we expect to detect concept drifts. The reason behind is the classification of certain anomalies into groups which can be analyzed in the aftermath. From certain attack patterns on servers to unexpected symptoms of a disease, each anomaly can be categorized over time. Here, we analyze how reliably collective deviating traces can be detected and subsequently categorized into types of concept drifts in a setting where new traces arrive over time. This includes the investigation of what impact different inter-drift distances on detection accuracy have (**RQ 7**) and how various sliding window sizes influence execution time (**RQ 8**). The number of parameters allows the approach to be fine-tuned but limits its application.

In more detail, we begin by creating an initial reference model. This model can be either normative or declarative, containing only conforming traces. In the subsequent iteration phase, as we process the data batch by batch, incoming traces are checked against the reference model, with a focus on the non-conforming ones. For each incoming trace, we

compute the local outlier factor [34]. For the distance computation between traces, we use the same approach as in [8]. Once a number K of traces have a local outlier score below a threshold T, a new process model is created from these traces to represent a novel process variant. This new model is then included in the conformance checking step.

Our results show that recurring drifts in a synthetic log, as well as incremental drifts in the BPIC 2015 dataset, can be detected with different inter-drift distances. In our experiments the F1-Score ranges between 0.86 and 0.96 for inter-drift distances of 100 up to 750. However, the number of variables involved makes the configuration challenging and requires a certain level of intuition about the data. For example, when a specific inter-drift distance is expected, the sliding window size must be adjusted accordingly to cover the entire concept drift. Therefore, we recommend applying this approach to data where certain drift distances are anticipated, such as seasonal drifts.

2.2.3 "Ill-Fated Processes" and Catalysts

Taking another perspective at anomalies further increases its apprehension. There are processes in place that entail a higher risk of non-compliance with their regulations. For example, in certain academic settings, the examination process represents an instance where the risk of non-compliance is heightened, particularly when students perceive a potential for greater rewards through illicit means. Engaging in dishonest practices in exams can be fueled by the wish or need to achieve good grades for future academic or professional pursuits. This can be traced back to the inherent competitiveness of the process. It cannot be ruled out that this competitive mindset might give rise to comparable issues in different sectors.

We encountered a similar issue when conducting a study with online exam data from the COVID-19 pandemic period. We solve the task of detecting colluding students by applying two approaches, namely Trace Ordering for Anomaly Detection (TOAD) [9] and Hierarchical Agglomerative Trace Clustering (HATC), an adaption of [35]. First, we address the question of whether collusion detection of online exams can be framed as a process mining problem (**RQ 9**). This question can be answered by choosing appropriate attributes for the mapping of the data to an event log. Hence, we create cases by selecting the user ID as the case ID, the task ID as the activity label and the corresponding timestamps.

We compare TOAD with HATC and explore how different parameter configurations impact the collusion detection results (**RQ 10**). The results show that TOAD only yields a small overlap (38%) with the set of students known to have colluded. More promising is the application of HATC, as the time window in which students submit their tasks can be adjusted. Here, we achieved a precision of around 0.68. Another major advantage of HATC is the use of a dendrogram as a visualization tool, which directly involves the teacher in the decision-making process of determining whether a student has colluded.

Additionally, we explore the advantages and drawbacks associated with the process mining framework, as well as the challenges of evaluation, considering that this approach may only serve as a supportive tool for teachers when determining if a student is involved in collusion (**RQ 11**). For this research question, we conclude the following: On the one

hand, colluding students must be included in the dataset, which cannot be taken for granted since students must give their consent. On the other hand, the outcome is limited by the number of tasks in which the students colluded; less collusion leads to greater distance. In this context, the concept of the *teacher in the loop*, combined with a highly adjustable tool, supports the decision-making process significantly.

As aforementioned, an important aspect is that this task can be classified as *positive unlabeled learning* [36]. This implies that the foundational data originating from processes like exam administration lacks negative instances and access to positive or labeled samples is barely possible. On the one hand, this stems from the fact that, e.g., students who obtain the maximal amount of points by cheating can not be distinguished from students who did not use unfair practices. On the other hand, only with proper observation it is certain that a student did not cheat which is not always feasible, especially in exams with a large number of participants. This underscores the importance of tools designed to identify collusion a posteriori, as they provide essential support in ensuring academic integrity.

This leads to the conclusion that there are processes that are prone to produce anomalies and, hence, inherently act as catalysts. It even permits the conclusion that both inherent dynamics and external factors contribute to a process's tendency to shift.

2.3 Sequential Patterns and Rules

This section introduces preliminaries which are required to understand the fusion of rule and process mining. It addresses the usefulness of rules for recommendations. Additionally, it draws the line from concurrency in process mining to partially-ordered rules.

2.3.1 A Matter of Order

When we deal with raw versions of traces, which are sequences of activity labels as explained in Section 2.1, insights into intrinsic information is limited. In that case, the order of elements is an excellent source to derive additional information. Consider a scenario where a user is configuring a personal computer. The user, uncertain about the necessary performance specifications, has a vague idea but lacks clarity on which components to purchase. After researching the core components such as CPU, GPU, and others it only then can be identified which case and power supply to purchase, since they depend on each other. This additional information about the order can then be used to enhance predictive models for, e.g., recommendation systems. This assumption should be made cautiously because an item sequence inherently possesses some sort of order. Hence, it does not automatically follow that items are correlated. Therefore, a method has to be applied that both relates connected items and ignores uncorrelated ones.

To understand the significance of order in pattern mining approaches we present three strategies ranging from methods devoid from any order to techniques with total order.

From Frequent Patterns to Sequential Rules

We begin with frequent patterns which exist for decades and are a popular tools in data mining [37]. Its use lead to many improvements in different data mining areas such as data indexing, clustering and classification [38]. Adaptions and extensions reach across areas such as graph pattern mining [39], streaming data mining [40] and parallel and distributed frequent pattern mining [41].

A traditional field of application for frequent pattern mining is *market basket analysis*. Here, the task is to gather sets of items that occur frequently together in a theoretical market basket. This is done by simply counting the occurrences of each possible itemset in the database and providing a threshold at which the occurrence of an itemset meets the frequency condition. Since this is a extremely inefficient way, the Apriori algorithm was introduced early [42]. The idea was to exclude itemsets from further calculation if they consist of infrequent itemsets themselves. This approach still contains the costly count of candidates which is why FP-Growth has been proposed which does not require candidate generation as it relies on an FP-Tree [43].

The mere counting of pattern occurrences is a popular measure of interestingness of an itemset, called *support*. This measure simply represents the aforementioned absolute or alternatively relative amount of occurrences of an itemset. To put this information to use, e.g., in a recommendation setting we can convert itemsets into Association Rules that consist of an implication in the form of $r_1 : \mathcal{X} \rightarrow \mathcal{Y}$. Here, $\mathcal{X}$ and $\mathcal{Y}$ each represent an itemset. Given $\mathcal{DB}$ as the database, s as a sequence of the database, and $|\mathcal{DB}|$ as the total number of sequences in the database, for r_1 we define the relative *support* as:

$$support(\mathcal{X}, \mathcal{Y}) = \frac{|\{s \in \mathcal{DB} \,|\, (\mathcal{X}, \mathcal{Y}) \in s\}|}{|\mathcal{DB}|} \tag{2.1}$$

This information alone is not sufficient since $\mathcal{Y}$ may frequently appear together with $\mathcal{X}$ but there may also be many transactions where this is not the case. In this case, inferring that $\mathcal{Y}$ is a good candidate for recommendation if $\mathcal{X}$ appears could be a wrong conclusion. For this reason other measures were proposed such as $confidence$. It quantifies how many times $\mathcal{Y}$ occurs given $\mathcal{X}$ occurs:

$$confidence(\mathcal{X} \rightarrow \mathcal{Y}) = \frac{support(\mathcal{X}, \mathcal{Y})}{support(\mathcal{X})} \tag{2.2}$$

There are many other measures of interestingness introduced which are not mentioned here [44, 45]. Further metrics that are used in our approaches are mentioned in Section 2.4.2. Beyond that, we focus on $|\mathcal{X}| = 1$ regarding the number of elements in an itemset.

So far, we completely exclude the order of items. This is appropriate in contexts such as the aforementioned market basket analysis, where it allows retailers to optimize product placements and corresponding promotions. Another example is the area of medical diagnosis, where associations between medical conditions, symptoms and treatments play an importance role for appropriate support of patients.

By including the order of items, we are located in the field sequential pattern mining [46] and sequential rule mining, respectively. We define $\prec$ as the symbol denoting that all items in $\mathcal{X}$ are predecessors of all items in $\mathcal{Y}$. We adapt *support* and $confidence$ for sequential rules as follows:

$$support(\mathcal{X}, \mathcal{Y}) = \frac{|\{s \in \mathcal{DB} \,|\, (\mathcal{X} \prec \mathcal{Y}) \in s\}|}{|\mathcal{DB}|}, \quad confidence(\mathcal{X} \rightarrow \mathcal{Y}) = \frac{support(\mathcal{X} \prec \mathcal{Y})}{support(\mathcal{X})}, \tag{2.3}$$

Several algorithms for mining sequential rules exist. These can be loosely divided into three categories [47]:

1. Algorithms that mine rules in a single sequence [48, 49]
2. Algorithms that mine rules that are common to several sequence excluding multiple occurrences in one sequence [50]
3. Algorithms that combine these and mine frequent rules in both single and across several sequences [51]

Just like this categorization, there are many gradations regarding the incorporation of order information in the generation of sequential rules. The entities within the sets included in a rule do not have to be limited to items; they can also be substituted with activities, particularly in a process mining context. We commenced without any order information as previously mentioned and progress to the opposite end of the spectrum: total order in rules. We wrap up this subsection by discussing partial order information.

Total Order in Rules

Totally-ordered sequential rule mining was first proposed by Zaki et al. [52]. They mine the set of all frequent sequences first and then create totally-ordered sequential rules (TOSR) from them. When speaking of frequent sequential patterns the creation of rules in a totally-ordered manner stands to reason. The reason behind is that sequential patterns inherently consist of an order. By splitting the pattern and calculating the corresponding *confidence* value, rules can be created (cf. RuleGen in [52]). We define the Totally-ordered Sequential Rules in Definition 1.

However, several significant drawbacks exist when dealing with TOSR [53]. First, items that occur concurrently can lead to an unsubstantiated distinction in rules. This is best explained with an example. Imagine the aforementioned case of a personal computer configuration. The sequence $\langle CPU, GPU, Mainboard, PowerSupply \rangle$ appears in different orders, where the *PowerSupply* is always selected last, e.g.

- $\langle CPU, GPU, Mainboard, PowerSupply \rangle$
- $\langle GPU, CPU, Mainboard, PowerSupply \rangle$
- $\langle CPU, Mainboard, GPU, PowerSupply \rangle$

- $\langle Mainboard, GPU, CPU, PowerSupply \rangle$

This leads to four different TOSR as each instantiation has a *support* = 1. Hence, given a minimum support threshold of 2, none of the rules would appear in the resulting rule set. This directly has influence on the second issue: measures of interestingness such as *confidence* can be affected and yield misleading numbers when these patterns are split up into multiple rules.

A third disadvantage influences the application in a recommendation setting. Since the rules are formulated very restrictively, either the probability of them matching a novel input sequence is low, or the number of rule candidates must be kept extremely high.

Definition 1 (Totally-Ordered Sequential Rule (TOSR))**.** *A totally-ordered sequential rule* $r : \mathcal{X} \rightarrow \mathcal{Y}$ *links two non-empty sequences* $\mathcal{X} = \langle x_1, x_2, ..., x_n \rangle$ *and* $\mathcal{Y} = \langle y_1, y_2, ..., y_n \rangle$ *as an implication, where* $\mathcal{X}$ *is called the* ***antecedent*** *and* $\mathcal{Y}$ *the* ***consequent*** *and* $\mathcal{X} \cap \mathcal{Y} = \emptyset$*. For all* $x_i \in \mathcal{X}$ *and* $y_i \in \mathcal{Y}$ *holds that* $x_i \prec y_i$*. For totally-ordered sequential rule mining, a TOSR means that all items in* $\mathcal{X}$ *occur before all items in* $\mathcal{Y}$ *whereby the items in the antecedent are restricted to a specific order. This also holds for the consequent, analogously.*

Partial Order in Rules

A possible solution for this where multiple rules are combined into a single one are partially-ordered sequential rules (POSR). These are defined as follows:

Definition 2 (Partially-Ordered Sequential Rule (POSR))**.** *A partially-ordered sequential rule* $r : \mathcal{X} \rightarrow \mathcal{Y}$ *links two non-empty sets* $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ *and* $\mathcal{Y} = \{y_1, y_2, ..., y_n\}$ *as an implication, where* $\mathcal{X}$ *is called the* ***antecedent*** *and* $\mathcal{Y}$ *the* ***consequent*** *and* $\mathcal{X} \cap \mathcal{Y} = \emptyset$*. For all* $x_i \in \mathcal{X}$ *and* $y_i \in \mathcal{Y}$ *holds that* $x_i \prec y_i$*. For partially-ordered sequential rule mining, a POSR means that all items in* $\mathcal{X}$ *occur before all items in* $\mathcal{Y}$ *whereby the items in the antecedent are not restricted to a specific order. This also holds for the consequent, analogously.*

With this approach it is possible to gather the four aforementioned sequential patterns in one rule: $r_2 : \{CPU, GPU, Mainboard\} \rightarrow \{PowerSupply\}$ because it is irrelevant in which order the items in the antecedent occur.

There is an essential relation to one of the major characteristics in Process Mining: *Concurrency*. This term describes the circumstance where it is unclear whether an activity happens before or after another. Both sequences are possible and do not necessarily happen in parallel. Modeling each manifestation of this characteristic in a process model would contradict the simplicity criteria of the four essential criteria of traditional process mining [5]. Thus, a partially-ordered representation of process paths proves to be useful. This enables the association of multiple items while also disregarding their order partially.

Regarding OCPM there are additional benefits to partially ordered events [1]. First, in terms of event logs, it may occur that the entries of the log are too coarse-grained, e.g. the timestamps only show the date by day, resulting in unclear sequencing.

Another reason is to intentionally avoid sequential occurrences of events if they lack correlation. Otherwise, assuming dependency might be erroneous.

2.3.2 Temporal Similarity and Proximity

Regardless of whether TOSR or POSR are used, rule mining results in a model that comprises rules encompassing implications between items. As mentioned in Section 2.3.1 based on different measures such as support, clusters emerge where the similarity of the clustered entities is always maximal, i.e. only equal entities are counted. In this regard, rule mining can be seen as a clustering approach that possesses a strict similarity condition. As soon as we extend the realm of information this notion of similarity also broadens.

For instance, this is the case when temporal information is additionally extracted. In general, a sequence is a strict total order on a specific set of objects. This order is arranged by the context in which the objects exist and occur. In terms of process mining the temporal information is frequently used and determines this order. Hence, mining frequent patterns and rules can be extended by other minimal information comprised in a sequence of events before additional event attributes are required. This proves to be especially useful in cases where companies can not provide meta-information about their process and there is only minimal information available.

Rule mining algorithms frequently suffer from a high amount of rules in the resulting rule set [54]. This drawback of a model, which also incorporates criteria such as generalization and precision, serves as a leverage point for improvement. Hence, it is reasonable to include additional information such as a time-based notion to restrict the rule mining process and also find more relevant rules. This type of information can be utilized in different ways.

Cüppers et al. [55] focus on discovering sequential patterns by analyzing the time delays between an event X and a prediction sequence Y, which indicates that a significant event is likely to occur at that specific moment. They introduce the concept of reliability, incorporating the duration until the predicted event occurs as a key piece of relevant information.

Another option is to focus on distant rules, that means rules, where the antecedent and the consequent occur far apart in a sequence. This approach is useful in the context of managing supplies as an early purchase prediction is helpful to stock up [56].

An alternative approach focuses on *proximity*, also known as *closeness* or *recency*, in rule generation. Accordingly, rules can be filtered by how close the antecedent and the consequent of a rule occur in a sequence. This requires a certain threshold at which the occurrence counts for the rules support value or not. The relevancy assumption for close items arises from the recency effect which refers to the phenomenon where individuals are more likely to recall recent events more accurately than earlier ones. As a result, recent information tends to carry greater weight in decision-making [57].

It should be noted that an occurrence of a rule in a sequence splits it into chunks. Then, the proximity is represented by the gap between the relevant elements from the antecedent and the consequent. This concept of proximity can either be satisfied by the quantity of events within the gap or the time elapsed between the occurrence of the last event in the antecedent and the occurrence of the first event in the consequent (s. Figure 2.4). In our work, SCORER-Gap [13], we examine how a decay factor influences the number of mined

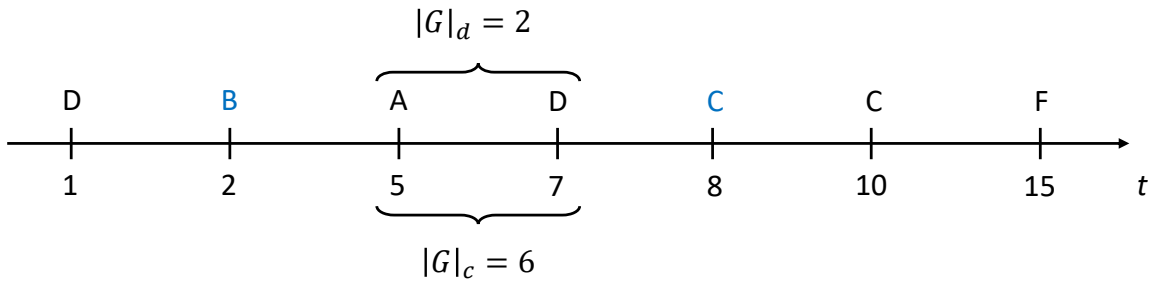


Figure 2.4: Example of a rule $B \rightarrow C$ having either a gap of 2 according to the number of items in between or a gap of 6, analogously for time.

rules by assigning a lower weight to rules that occur with larger gap sizes between them (**RQ 12**). Essentially, the RuleGrowth algorithm by Fournier-Viger et al. [47] is used as a basis, and it is adapted by weighting each occurrence with a decay factor. For example, given a rule $A \rightarrow B$ that occurs twice with gap sizes of 2 and 6, and with a decay factor of 0.7, we calculate $(0.7^2 + 0.7^6)/2$. If the result is greater than the minimum support threshold, this rule is included in the resulting rule set.

By applying this method to three real-world datasets, we find that this approach significantly reduces the number of resulting rules, while accuracy does not deteriorate in a recommendation task. This approach proves particularly effective in sparse datasets, where the number of rules tends to be quite high. However, memory consumption may limit its application, as the a priori principle is not applied in our approach, leading to a significant increase in runtime depending on the complexity of the dataset.

This novel and promising perspective on partially-ordered sequential rules considers the proximity of events when selecting rules.

2.4 Recommendation Systems

One of the primary areas where rules find extensive application are recommendation systems. The inherent implication that the consequent follows the antecedent makes it particularly suitable for this purpose. Over time and with the rise of big data including not only specific items but additional meta-data it has been superseded by deep learning approaches in many areas [58, 59].

Different types of recommendation systems exist. The most popular ones are based on *collaborative filtering* and *content-based filtering*. The former deals with finding similarities between user behaviors which means that it aggregates user interactions in form of ratings and other meta-data and recognizes commonalities between users on this basis. Then, recommendations are formed based on the comparison between users. In the latter approach, items are analyzed based on their attributes. Similar items are then used as candidates to be recommended to users which have interacted with on of them in the past. Also other types of recommendation systems exist that are, e.g., based on utility [60] or

hybrid recommendation systems combining multiple approaches.

Especially, deep learning techniques have been used to further enhance the performance of these systems [61, 62, 63]. While they manage to outperform traditional recommendation systems in many areas they are not superior in every aspect [64]. The area of application might play an important role as neural-network based systems can easily handle complex data and a large amount of data after an appropriate training time with ease. This can be traced back to the most compact model representation. Popular benchmarks are given by big tech companies like Netflix [59], YouTube and Spotify which benefit highly from advances [65, 66]. Hence, state-of-the-art recommendation systems are tailored towards the data provided by these benchmarks.

However, there are certainly many other companies that benefit highly from an appropriate tool to enhance performance like increasing sales. An example for that are SMEs which have to deal with challenges arising from less financial resources among others. Frequently, the data they can provide is limited in quantity, is highly sparse, and these companies may suffer from low computational capabilities. Beyond that, they may lack of additional user data since accounts may be not required on their platforms. Traditional approaches like collaborative filtering or modern approaches incorporating deep learning methods are not applicable [67, 68].

Additionally, a new concept for recommendation systems emerged in recent years, called session-based recommendation systems. It is based on the assumption that recent interactions have more value to the user regarding a next event prediction than interactions that lie far in the past [69]. Moreover, this aligns with the previously mentioned condition in which a user may appear anonymously, resulting in limited availability of historical data and the absence of metadata.

This circumstance leads to sequence-based recommendation algorithms, which are more appropriate to meet aforementioned conditions.

2.4.1 Rule Mining as a Foundation

As already mentioned in the beginning of Section 2.4 rules are suitable by design for a recommendation task. This is the result of its implicatory nature, where a consequent occurs in succession to a certain antecedent. This type of recommendation system is often called traditional or grammar-based. It comes with the advantage of a transparent and interpretable reasoning as the origin of rules is clear. Beyond that, the well-understood and established methods which lead to resulting rule sets contribute to its comprehensibility. Another advantage is that the requirement for computational power is comparably low.

As rule sets are created based on various parameters like *support* and *confidence* (cf. Section 2.3.1) among others, this constitutes a drawback as they have to be manually adapted to the underlying data. Additionally, rule mining approaches are frequently applied to transactional and categorical data such as retail transactions, data from the area of e-commerce or clickstream data. Especially, companies selling their products online have much more opportunities to satisfy customer need for niche products since the rise of the internet [70, 71, 72]. This development, where the importance of blockbuster products

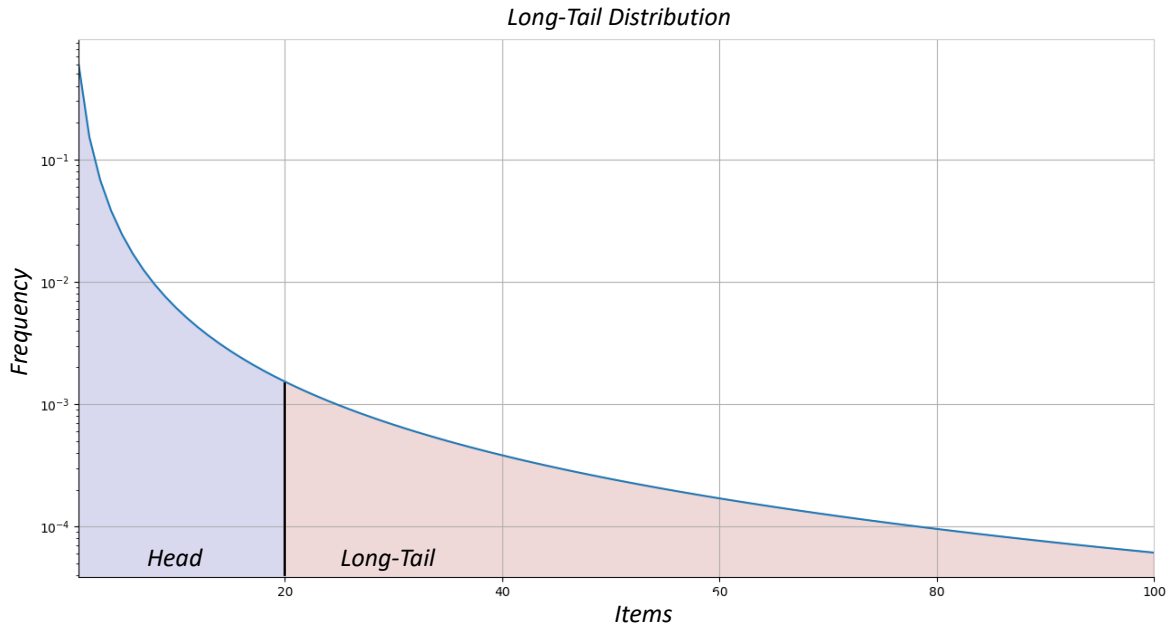


Figure 2.5: Example of a long tail plot with a split into head and tail. Head describes the frequently bought blockbuster items and tail represents the long distribution of niche items.

diminishes as niche products capture a substantial portion of overall demand is called the long-tail of a product assortment [73] (s. Figure 2.5).

Another drawback emerges from this observation: When applying the minimum support threshold, only transactions with a particular occurrence frequency can be captured. This is an indication that a dynamic adaption of thresholds is beneficial (cf. [13]).

In this regard, it is of utter importance to examine if approaches achieve the objective they have been set. Therefore, quality metrics are deployed.

2.4.2 Quality Metrics

To measure the performance of recommendation systems various quality metrics have been introduced over time. While the field initially focused on accuracy [74] and precision metrics such as *Hit-Ratio*, *Mean Reciprocal Rank* and *Mean Average Precision* one of the most popular is *Normalized Discounted Cumulative Gain (nDCG)*. All of these methods operate using a predefined list of recommendations, from which the performance value is calculated. Taking nDCG as an example, this metric also penalizes relevant items which are placed at the bottom of the recommendation list. It puts more emphasis on the ranking with which potential recommendations are presented to the user. Thus, a relevant item is weighted according to its position in the recommendation list. The calculation is performed

using the following formula:

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)} \tag{2.4}$$

The discounted cumulative gain gets normalized by using the ideal discounted cumulative gain. This value is calculated analogously to DCG@k but with the assumption that each of the items in the recommendation list is relevant.

$$nDCG@k = \frac{DCG@k}{IDCG@k} \tag{2.5}$$

Typically for evaluation purposes a specific number of recommendations is assumed. The included number of items is defined by k. Both the premises that recommendations are provided in a ranked or ordered format as well as unordered are valid. The ordered format provides the opportunity to also evaluate approaches on a scale of relevancy of its items. In general, recommendation systems should be evaluated from different perspectives as relying on a single metric typically leads to a one-sided assessment and fails to capture the overall performance of the system [75].

Eli Pariser coined the term "filter bubble" in his book "The filter bubble: What the Internet is hiding from you" [76]. It describes the hypothesis that recommendation systems try to satisfy a user's needs by predicting its interests. By that, it excludes information which does not match with the user's goals or intentions. Hence, it narrows down the provided information and influences the context in which a user operates.

A similar term from the field of cognitive psychology is "confirmation bias". Here, it is taken one step further as the notion does not only describe the search and selection of information that meets the expectation but also comprises the interpretation of such.

However, while there is work stating the existence [77] it is not yet proven that recommendation systems actually have this kind of impact [78]. Interestingly, there are cases where a narrowing effect arises, especially, when recommendation systems are not in use which would be the opposite of the "filter bubble" thesis. Nevertheless, whether these tools are used or not the potential consequence should be kept in mind. It is beyond discussion, that limiting or even excluding information from a user's range is an intervention into the user's autonomy. The amount of data and information is rigorously increasing which leaves no alternative but to apply these tools, regardless.

As well as there are different sub-tasks regarding the application of recommendation systems [74] there are different perspective with which these systems can be evaluated. These perspectives are intertwined with the corresponding task as it defines the value it can generate to the user. Silveira et al. [79] identified six further concepts beyond precision metrics with which recommendation systems can be evaluated, namely *utility, novelty, diversity, unexpectedness, serendipity* and *unexpectedness*. While each of these concepts has its validity and lead to its own field of research, in our following work, we primarily focus on diversity. The reason behind this is that diversity, in particular, aims at covering the full range of interest for the user and precision- and accuracy-based metrics are tailored towards

learned representation missing out on unknown and diverse entities. Ricci et al. [80] state that recommendation lists with low variety lowers the user's interest. As recommendation systems without focus on diversity frequently recommend similar or equal items compared to the user's past interactions, diversity is required to cover the whole range of a user's interest [81]. Diversity frequently is defined as the opposite of similarity s.t. the items in a recommendation list must be different in some way. It suffices that the label of an item is different but it can also be defined on a more detailed level like the diversity of the item's attributes. Since there are different variants of the diversity metric, in the following we cover some of which we use in our work. Note, that we apply the difference in a binary manner, meaning that either an item is equal or different. In the following k represents the amount of highest ranked items in a recommendation list, also known as *top-k*. The symbol $r \in \mathcal{R}$ denotes a relevant item out of the set of relevant items, as validated by the ground truth. For the sake of completeness of quality measures, we begin with Recall and Accuracy.

Recall Recall@k (also known as Hit Rate) is implemented as the mean number of hits over all recommendation lists.

$$Recall@k = \frac{\sum_{i=1}^{k} r_i}{k} \tag{2.6}$$

Accuracy It is defined as the ratio of all correct recommendations. Let n_{rec} be the number of all recommendations proposed in the experiment.

$$Accuracy = \frac{\sum_{i=1}^{k} r_i}{n_{rec}} \tag{2.7}$$

Intra-List Diversity (IALD) $IALD@k$ is defined as the amount of variety in a single recommendation list.

$$IALD@k = \frac{2}{N(N-1)} \sum_{i \neq j} (1 - \frac{|i_i \cap i_j|}{|i_i \cup i_j|}). \tag{2.8}$$

Here, $\frac{2}{N(N-1)}$ is the number of possible item combinations within a recommendation list, and i_i, i_j is an item at position i, i.e., j.

Inter-List Diversity (IELD) In the following, we distinguish between $IELD_p@k$ and $IELD_s@k$, i.e., a pairwise comparison of the recommendation list and the set-theoretic difference between two recommendation lists. For $IELD_p@k$ we calculate the mean non-overlap ratio between two recommendation lists R'_1 and R'_2 of successive recommendation steps.

$$IELD_p@k(R'_1, R'_2) = \frac{1}{k} \cdot \sum_{\substack{i=1, \\ r_1 \in R'_1, r_2 \in R'_2}}^{i=k} (1 - \frac{|r_{1_i} \cap r_{2_i}|}{|r_{1_i} \cup r_{2_i}|}), \tag{2.9}$$

We achieve this pairwise, meaning that an item of the first recommendation list at position i is compared with the item of the second recommendation list at the same position i.

Furthermore, we adopt the variant from Lathia et al. [82]. We treat each successive recommendation list as a set of items R_1'' and R_2''. With these sets, For $IELD_s@k$ we calculate the set-theoretic difference between each successive recommendation list as a set of items R_1'' and R_2'' and define the inter-list diversity based on the set-theoretic difference.

$$R_2'' \setminus R_1'' = \{x \in R_2'' | x \notin R_1''\}, \quad IELD_s@k(R_1'', R_2'') = \frac{|R_2'' \setminus R_1''|}{k}. \tag{2.10}$$

Hence, we define the inter-list diversity based on the set-theoretic difference as

$$IELD_s@k(R_1'', R_2'') = \frac{|R_2'' \setminus R_1''|}{k} \tag{2.11}$$

Sequence Diversity (SeqDiv) Given the function $len()$ that yields the number of items in a sequence and $set()$ which returns the sequence without duplicates, we define $SeqDiv$ as follows:

$$SeqDiv = \frac{len(set(s))}{len(s)}, \tag{2.12}$$

Referring to our approach, in [14] we investigate how six variants of session-based recommendation systems perform regarding four diversity and three precision metrics (**RQ 13**). We especially focus on the question how conditions derived from small and medium-sized companies can be incorporated in the design of such recommendation systems (**RQ 14**). Hence, all of these variants are based on minimal information such as a mere trace of activities without additional attributes. Moreover, these systems can cope with the limited amount of provided data, data sparsity, lower computational capabilities, and the lack of additional or metadata, as users typically do not have accounts on the provided platforms. Additionally, one-time users are more likely, as SMEs typically offer a smaller product portfolio.

Our six variants of recommendation systems are based on the ER-Miner by Fournier-Viger et al. [83] and are incorporated as a post-processing step. The *DGap* variant utilizes the discrete gap between the antecedent and the consequent of a rule, as described in Figure 2.4. Here, "discrete" means that the number of events in the gap is counted. For each rule, a histogram is created by counting its occurrences with a certain gap size in the database. Based on this histogram, we create a ranking of rules for the recommendation step at hand. *CGap*, another variant, functions analogously but uses the time between the antecedent and the consequent. For these two approaches, there are additional variants called *DGap-Acc* and *CGap-Acc*, respectively. In these variants, instead of looking up the specific gap size in the histogram, we accumulate every occurrence frequency up to the required gap size. Additionally, *UniqueC* and *UniqueC-TW* represent recommendation systems where either the consequents of rules are permitted only if they did not occur yet in the entire sequence, or if they did not occur within a specific time window. We conducted experiments using the aforementioned quality metrics, excluding *nDCG*.

The results show that *DGap* and *UniqueC* are strong competitors, with *UniqueC* eventually outperforming the other methods. E.g. the F1-Score of *UniqueC* on all datasets ranges between 0.43 and 0.73. It is notable that the limited amount of provided information significantly influences the potential of all variants. Additionally, a fundamental challenge with rule mining approaches is the restricted item coverage due to parameters such as minimum support and minimum confidence thresholds. Especially, in datasets with a long-tail item frequency distribution (see Figure 2.5), only a small portion of the item space can be covered.

We assess the problem posed by the specific conditions of small and medium-sized enterprises from different angles. Nevertheless, these approaches provide a beneficial starting point for SMEs to incorporate recommendations into their still-growing systems.

Chapter 3

Conclusion

In the following discussion, we will present a summary of our work in three key parts. First, we will outline the results of our research within the broader context, highlighting the significance of our findings. Next, we will address potential limitations of the methods we developed, which naturally suggest avenues for future research. These limitations in combination with the experience we gained during our research give rise to new questions and challenges, which we will explore in the second section, focusing on the subjects that emerge as priorities for further investigation.

3.1 Summary

As we reach the closing of this thesis, it is essential to reflect on the journey undertaken, the insights gathered, and the implications for the broader field. Table 3.1 summarizes the research questions discussed, providing a concise overview for quick reference. Each question is paired with the format in which it has been addressed.

This thesis explores two main objectives: The first objective focuses on process mining, particularly exploring how the relevance of activities interacts with deviating behavior. The second objective involves integrating rules with process mining by evaluating frequency and sequence as indicators of relevance.

To delve deeper into the first aspect, relevance can be assessed by measuring the distance from the main process and clustering traces to identify deviation groups. Our approach is twofold: we aim to detect these groups in a static process, and we also seek to apply this methodology to continuously incoming traces. On one hand, we found that detecting groups of anomalies by aligning their structure with the main process model offers a distinct advantage over traditional trace clustering methods. In our procedure, we reference the main model based on the control flow of each trace, which allows us to utilize the model's structure to guide the clustering process, with a specific focus on the structure of the anomalies.

This concept is also applicable to trace streams: In our additional work, we investigate the benefits and limitations of using a local outlier factor to detect various types of concept

Table 3.1: Summary of research questions. Due to space constraints, the questions are presented in a shortened, nominal form.

ID	Section	Research Question	Format
RQ1	Section 2.2	Assessment of performance on an anomaly detection task.	Line and reachability plot
RQ2	Section 2.2	Assessment of performance when adapting parameters.	Line plot
RQ3	Section 2.2.1	Suitability of process model to cluster collective anomalies in comparison to trad. trace clustering.	Dendrogram
RQ4	Section 2.2.1	Analysis of the effect of different kernels and bandwidth.	Line plot
RQ5	Section 2.2.1	Assessment of different aggregation methods.	Table
RQ6	Section 2.2.1	Examination of two approach variants regarding classification accuracy and runtime.	Table and bar chart
RQ7	Section 2.2.2	Investigation of the impact of varying inter-drift distances on detection accuracy.	Line plots and Gantt charts
RQ8	Section 2.2.2	Influence of various sliding window sizes on execution time.	Bar chart
RQ9	Section 2.2.3	Investigating the possibility of framing collusion detection in exams as a process mining problem.	Textual explanation
RQ10	Section 2.2.3	Exploring how different parameter configurations impact the collusion detection results.	Dendrogram and Text
RQ11	Section 2.2.3	Exploring the advantages and drawbacks as well as the challenges of evaluation.	Discussion
RQ12	Section 2.3.2	Examining a decay factor and its influence on the number of mined rules.	Line plots and tables
RQ13	Section 2.4.2	Investigation of six variants of Session-based Recommendation Systems regarding performance on four diversity and three precision metrics.	Line plot and table
RQ14	Section 2.4.2	Examining the incorporation of conditions derived from SMEs in the design of recommendation systems.	Textual explanation

drifts. We reveal that this approach has significant potential, especially with its detailed configuration options, although understanding the underlying drift is essential for effective application.

Furthermore, we aim to incorporate a temporal perspective into anomaly detection and conformance checking. We develop methods that, on one hand, use temporal aspects to identify deviating sub-processes and, on the other hand, evaluate trace conformance based on the temporal behavior of the main process.

Here, we focus on sets of traces with high internal density that are well-separated from other traces. With this procedure we identify sub-processes holding significant economic potential for process optimization, either by incorporating positive behaviors or by mitigating negative failures. Additionally, this temporal perspective can be leveraged for conformance checking, where the average fitness, based on the probability of occurrences at specific timestamps, plays a decisive role. The model's simplicity is beneficial, as applying the temporal perspective can often serve as an early indicator of impending changes.

In exploring a new application area — collusion detection in online exams — we assess the performance of trace clustering in comparison to TOAD, an anomaly detection method grounded in the temporal perspective. Trace clustering demonstrates its strength by allowing users to fine-tune the configuration according to the underlying data, thereby significantly reducing the risk of false positives, which is a critical factor in this context.

The frequency and sequence of activities, highlighted as the second key point, have a direct impact on the control flow of the process under examination. In this context, our goal is to explore ways to enhance recommendation systems.

In this context, we first develop a rule mining algorithm that weights the occurrence of a rule based on its gap size within the sequence, emphasizing the importance of proximity on relevance. This approach allows us to significantly reduce the number of resulting rules while maintaining accuracy in the recommendation process. Furthermore, we strive to balance precision and diversity in a recommendation task to meet user demands without excluding parts of the process. This involves addressing challenges such as limited data availability, data sparsity, lower computational resources, and the absence of additional data, which are common in small- and medium-sized enterprises. Concerning this, we examine six different variants of recommendation systems regarding their suitability. Simple measures, such as ranking rules based on their frequency of occurrence with a specific gap size or prohibiting the application of a rule whose consequent has already occurred, prove to be helpful as a starting point for SMEs.

In summary, our focus is on analyzing specific sub-processes, which are essential units, primarily in terms of their frequency and sequence. These attributes are used to address downstream tasks such as anomaly detection, conformance checking, and event recommendation.

While this research answers several important questions, it has also opens up new avenues for inquiry. The limitations acknowledged in the study pave the way for future research to build upon and refine the existing knowledge.

3.2 Identification of Far-Reaching Research Questions

This section provides an outlook on specific areas for future research, building on the insights gained from the current studies. These proposed directions aim to address the limitations encountered and explore untapped potential in the field. The discussion will end in a set of key research questions that could serve as a foundation, guiding future scientists in their pursuit of advancing this domain.

Our approaches (cf. [8, 10, 9]) that focus on conformance checking and anomaly detection are jointly addressed in the following: Regarding [8], one limitation is the narrow scope that results from relying solely on the control-flow of the model. The aforementioned approaches extend this by incorporating another highly relevant perspective — namely, time.

Our work, TADE [10], has drawbacks in handling parallelism and loops. A leverage point would be to incorporate workflow models into the temporal model or replace the provided kernel functions with more complex kernels to fully evaluate the capabilities of this approach. However, the inclusion of process models might increase the complexity of the conformance checking procedure as additional checks are necessary. Another question that would have to be solved is the model representation of these properties that are inherent to the control-flow.

As far as [12] is concerned the sensitivity regarding parameter configuration is also apparent. Here, an important point for future work is to revisit the topic of defining an anomaly or the affiliation to a cluster. Here, the more similar a submitted solution of an exam is to a sample solution the less potential there is to differentiate between well-prepared and colluding student. This leads to a high difficulty proving a collusion emphasizing the importance of refining given parameter ranges for new datasets and including new attributes.

Grammar-based recommendation system approaches as they are investigated in [14] are certainly a good basis for small- and medium-sized enterprises. Nevertheless, the success of these approaches hinges with the information that can be used to train the model, i.e., create the rule set. A key area for further investigation is whether there is additional inherent information in SME data and how it might be used to improve recommendation systems.

Evaluating approaches in Process Mining can be difficult at times. One of the more frequently occurring issues with data is the lack of ground truth and reliable benchmarks. In this regard, general benchmarking frameworks are helpful to establish an evaluation basis. During the past years, we frequently wished for realistic data augmentation and simulation tools that already have a strong standing in the community. Van der Aalst identified this requirement as well, for which the research in OCPM can be an important step [84, 85]. We formulate the following research question: What methodologies can be developed for realistic data augmentation and simulation in process mining, and how can these tools improve the validity and reliability of benchmarking frameworks?

Another direction that proved to be highly relevant is the inclusion of multiple perspectives. As OCPM is able to include multiple objects and the perspectives they are covering, it appears to be a viable direction (e.g. cf. Li et al. [86]). Based on this requirement, we

formulate the following research question: How can the integration of multiple perspectives enhance solutions in already established fields such as trace clustering, conformance checking, anomaly detection or concept drift detection?

Additionally, the current research in the field of recommendations seems to be mainly tailored towards large companies that provide specific information like meta-data such as ratings, behavioral data and user account information to derive recommendations. However, there are many other companies such as small and medium-sized enterprises (SMEs) that also have the demand for recommendation systems but provide other information as input. In fact, 99.3% of companies in Germany belong to SMEs in 2022 [87], hence, there lies demand and research potential. We formulate the following research question: What specific requirements do small and medium-sized enterprises articulate, what information do they provide, and how can these needs be systematically translated into effective recommendation tools?

To extend the scope further and for future research among the improvement of our existing implementations and approaches regarding runtime optimization and research on different variants we propose a look at OCPM regarding sub-process analysis and a detailed evaluation of anomalies and deviations. Especially as processes are complex and intertwined around objects, the notion of sub-processes becomes vivid once more. These gradations of processes across different scales highlights the nested patterns that can be found within object-centric processes and their underlying structures.

Finally, another area that highly benefits from process mining methods is Educational Data Mining and Learning Analytics. Transforming underlying procedures into the process space enables the effective application and adaptation of process mining methods, enhancing their utility and relevance in practical scenarios.

In conclusion, this thesis advances the exploration of micro-clusters, i.e. sub-processes, that can be revealed by applying various process mining perspectives. These perspectives, including the integration of time and control-flow, enable enhanced analysis in areas such as deviation detection, concept drift identification, and the development of recommendation systems. The findings open up new avenues for future research, encouraging further investigation into their potential applications in real-world scenarios.

Bibliography

[1] Wil MP van der Aalst. Object-centric process mining: Dealing with divergence and convergence in event data. In *Software Engineering and Formal Methods: 17th International Conference, SEFM 2019, Oslo, Norway, September 18–20, 2019, Proceedings 17*, pages 3–25. Springer, 2019.

[2] Joos CAM Buijs, Boudewijn F van Dongen, and Wil MP van der Aalst. Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. *International Journal of Cooperative Information Systems*, 23(01):1440001, 2014.

[3] Wil MP van der Aalst. Process mining: a 360 degree overview. In *Process Mining Handbook*, pages 3–34. Springer, 2022.

[4] RP Jagadeesh Chandra Bose, Ronny S Mans, and Wil MP Van Der Aalst. Wanna improve process mining results? In *2013 IEEE symposium on computational intelligence and data mining (CIDM)*, pages 127–134. IEEE, 2013.

[5] Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

[6] Yifeng Lu. *Pattern mining under different conditions*. Ph.D. Thesis, LMU Munich, 2021.

[7] Niek Tax, Laura Genga, and Nicola Zannone. On the use of hierarchical subtrace mining for efficient local process model mining. In *7th International Symposium on data-driven process discovery and analysis (SIMPDA 2017)*, pages 8–22. CEUR-WS.org, 2017.

[8] Florian Richter, Ludwig Zellner, Janina Sontheim, and Thomas Seidl. Model-aware clustering of non-conforming traces. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings*, pages 193–200. Springer, 2019.

[9] Florian Richter, Yifeng Lu, Ludwig Zellner, Janina Sontheim, and Thomas Seidl. Toad: trace ordering for anomaly detection. In *2020 2nd International Conference on Process Mining (ICPM)*, pages 169–176. IEEE, 2020.

[10] Florian Richter, Janina Sontheim, Ludwig Zellner, and Thomas Seidl. Tade: Stochastic conformance checking using temporal activity density estimation. In *Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18*, pages 220–236. Springer, 2020.

[11] Ludwig Zellner, Florian Richter, Janina Sontheim, Andrea Maldonado, and Thomas Seidl. Concept drift detection on streaming data with dynamic outlier aggregation. In *Process Mining Workshops: ICPM 2020 International Workshops, Padua, Italy, October 5–8, 2020, Revised Selected Papers 2*, pages 206–217. Springer, 2021.

[12] Andrea Maldonado, Ludwig Zellner, Sven Strickroth, and Thomas Seidl. Process mining techniques for collusion detection in online exams. In *Process Mining Workshops: ICPM 2023 International Workshops*. Springer, 2023.

[13] Ludwig Zellner, Janina Sontheim, Florian Richter, Gabriel Lindner, and Thomas Seidl. Scorer-gap: Sequentially correlated rules for event recommendation considering gap size. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 925–934. IEEE, 2021.

[14] Ludwig Zellner, Simon Rauch, Janina Sontheim, and Thomas Seidl. On diverse and precise recommendations for small and medium-sized enterprises. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–130. Springer, 2024.

[15] Indrė Žliobaitė. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*, 2010.

[16] Denise Maria Vecino Sato, Sheila Cristiana De Freitas, Jean Paul Barddal, and Edson Emilio Scalabrin. A survey on concept drift in process mining. *ACM Computing Surveys (CSUR)*, 54(9):1–38, 2021.

[17] Supriya Agrahari and Anil Kumar Singh. Concept drift detection in data stream mining: A literature review. *Journal of King Saud University-Computer and Information Sciences*, 34(10):9523–9540, 2022.

[18] Encyclopædia Britannica. Process. In *britannica.com dictionary*. Encyclopædia Britannica, Accessed: August 6, 2024.

[19] Wil MP van der Aalst and Josep Carmona. *Process mining handbook*. Springer Nature, 2022.

[20] Alessandro Berti and Wil MP van der Aalst. OC-PM: Analyzing object-centric event logs and process models. *International Journal on Software Tools for Technology Transfer*, 25(1):1–17, 2023.

[21] Wil MP van der Aalst. Object-centric process mining: An introduction. In *International Colloquium on Theoretical Aspects of Computing*, pages 73–105. Springer, 2021.

[22] Marlon Dumas, L Marcello Rosa, Jan Mendling, and A Hajo Reijers. *Fundamentals of business process management*. Springer, 2018.

[23] Adela del Rıo-Ortega, Manuel Resinas, and A Ruiz-Cortés. Towards modelling and tracing key performance indicators in business processes. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, 3(3), 2009.

[24] Saoussen Cheikhrouhou, Slim Kallel, Nawal Guermouche, and Mohamed Jmaiel. The temporal perspective in business process modeling: a survey and research challenges. *Service Oriented Computing and Applications*, 9:75–85, 2015.

[25] F. Mannhardt. *Multi-perspective process mining*. Ph.D. Thesis, Eindhoven University of Technology, 2018.

[26] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[27] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

[28] Boudewijn van Dongen. BPI Challenge 2017. Eindhoven University of Technology, 2017.

[29] Boudewijn van Dongen. BPI Challenge 2015. Eindhoven University of Technology, 2015.

[30] Agnes Koschmider, Kay Kaczmarek, Mathias Krause, and Sebastiaan J van Zelst. Demystifying noise and outliers in event logs: review and future directions. In *International Conference on Business Process Management*, pages 123–135. Springer, 2021.

[31] Simone Coltellese, Fabrizio Maria Maggi, Andrea Marrella, Luca Massarelli, and Leonardo Querzoni. Triage of iot attacks through process mining. In *On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings*, pages 326–344. Springer, 2019.

[32] Gregory Ditzler, Manuel Roveri, Cesare Alippi, and Robi Polikar. Learning in non-stationary environments: A survey. *IEEE Computational Intelligence Magazine*, 10(4):12–25, 2015.

[33] RP Jagadeesh Chandra Bose, Wil MP van der Aalst, Indrė Žliobaitė, and Mykola Pechenizkiy. Handling concept drift in process mining. In *Advanced Information Systems Engineering: 23rd International Conference, CAiSE 2011, London, UK, June 20-24, 2011. Proceedings 23*, pages 391–405. Springer, 2011.

[34] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[35] Soo Jeong Ingrisone and James N Ingrisone. Hierarchical agglomerative clustering to detect test collusion on computer-based tests. *Educational Measurement: Issues and Practice*, 42(3):39–49, 2023.

[36] Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: A survey. *Machine Learning*, 109(4):719–760, 2020.

[37] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[38] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data mining and knowledge discovery*, 15(1):55–86, 2007.

[39] Philippe Fournier-Viger, Ganghuan He, Chao Cheng, Jiaxuan Li, Min Zhou, Jerry Chun-Wei Lin, and Unil Yun. A survey of pattern mining in dynamic graphs. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6):e1372, 2020.

[40] Ruoming Jin and Gagan Agrawal. Frequent pattern mining in data streams. *Data streams: Models and algorithms*, pages 61–84, 2007.

[41] Syed Khairuzzaman Tanbeer, Chowdhury Farhan Ahmed, and Byeong-Soo Jeong. Parallel and distributed frequent pattern mining in large databases. In *2009 11th IEEE International Conference on High Performance Computing and Communications*, pages 407–414. IEEE, 2009.

[42] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB*, volume 1215, pages 487–499. Santiago, 1994.

[43] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[44] Michael Hahsler, Bettina Grün, and Kurt Hornik. arules-a computational environment for mining association rules and frequent item sets. *Journal of statistical software*, 14(15):1–25, 2005.

[45] Sergey Brin, Rajeev Motwani, Jeffrey D Ullman, and Shalom Tsur. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, 1997.

[46] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, and Rincy Thomas. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1):54–77, 2017.

[47] Philippe Fournier-Viger, Roger Nkambou, and Vincent Shin-Mu Tseng. Rulegrowth: mining sequential rules common to several sequences by pattern-growth. In *Proceedings of the 2011 ACM symposium on applied computing*, pages 956–961, 2011.

[48] Heikki Mannila, Hannu Toivonen, and A Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data mining and knowledge discovery*, 1:259–289, 1997.

[49] Don-Lin Yang, Yuh-Long Hsieh, and Jungpin Wu. Using data mining to study upstream and downstream causal relationship in stock market. In *9th Joint International Conference on Information Sciences (JCIS-06)*, pages 528–531. Atlantis Press, 2006.

[50] Usef Faghihi, Philippe Fournier-Viger, Roger Nkambou, and Pierre Poirier. A generic episodic learning model implemented in a cognitive agent by means of temporal pattern mining. In *Next-Generation Applied Intelligence: 22nd International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2009, Tainan, Taiwan, June 24-27, 2009. Proceedings 22*, pages 545–555. Springer, 2009.

[51] Sherri K Harms, Jitender Deogun, and Tsegaye Tadesse. Discovering sequential association rules with constraints and time lags in multiple sequences. In *International symposium on methodologies for intelligent systems*, pages 432–441. Springer, 2002.

[52] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42:31–60, 2001.

[53] Philippe Fournier-Viger, Usef Faghihi, Roger Nkambou, and Engelbert Mephu Nguifo. Cmrules: Mining sequential rules common to several sequences. *Knowledge-Based Systems*, 25(1):63–76, 2012.

[54] Farah Hanna AL-Zawaidah, Yosef Hasan Jbara, and AL Marwan. An improved algorithm for mining association rules in large databases. *World of Computer science and information technology journal*, 1(7):311–316, 2011.

[55] Joscha Cüppers, Janis Kalofolias, and Jilles Vreeken. Omen: discovering sequential patterns with reliable prediction delays. *Knowledge and Information Systems*, 64(4):1013–1045, 2022.

[56] Lina Fahed, Philippe Lenca, Yannis Haralambous, and Riwal Lefort. Distant event prediction based on sequential rules. *Data Science and Pattern Recognition*, 4(1):1–23, 2020.

[57] Hakkyu Kim and Dong-Wan Choi. Recency-based sequential pattern mining in multiple event sequences. *Data Mining and Knowledge Discovery*, 35:127–157, 2021.

[58] Paul Covington, Jay Adams, and Emre Sargin. Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198, 2016.

[59] James Bennett, Stan Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, 2007.

[60] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Tin Truong-Chi, and Roger Nkambou. A survey of high utility itemset mining. *High-utility pattern mining: Theory, algorithms and applications*, pages 1–45, 2019.

[61] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

[62] Guijuan Zhang, Yang Liu, and Xiaoning Jin. A survey of autoencoder-based recommender systems. *Frontiers of Computer Science*, 14:430–450, 2020.

[63] Le Wu, Xiangnan He, Xiang Wang, Kun Zhang, and Meng Wang. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4425–4445, 2022.

[64] Yushun Dong, Jundong Li, and Tobias Schnabel. When newer is not better: Does deep learning really benefit recommendation from implicit feedback? In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 942–952, 2023.

[65] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[66] Carlos A Gomez-Uribe and Neil Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):1–19, 2015.

[67] Marius Kaminskas, Derek Bridge, Franclin Foping, and Donogh Roche. Product recommendation for small-scale retailers. In *E-Commerce and Web Technologies: 16th International Conference on Electronic Commerce and Web Technologies, EC-Web 2015, Valencia, Spain, September 2015, Revised Selected Papers 16*, pages 17–29. Springer, 2015.

[68] Marius Kaminskas, Derek Bridge, Franclin Foping, and Donogh Roche. Product-seeded and basket-seeded recommendations for small-scale retailers. *Journal on Data Semantics*, 6:3–14, 2017.

[69] Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38, 2021.

[70] Eric K Clemons, Guodong Gordon Gao, and Lorin M Hitt. When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of management information systems*, 23(2):149–171, 2006.

[71] Chris Anderson. *The long tail : why the future of business is selling less of more*. Hyperion, New York, 2008.

[72] Oliver Hinz, Jochen Eckert, and Bernd Skiera. Drivers of the long tail phenomenon: an empirical analysis. *Journal of management information systems*, 27(4):43–70, 2011.

[73] Erik Brynjolfsson, Yu Hu, and Duncan Simester. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management science*, 57(8):1373–1386, 2011.

[74] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[75] Eva Zangerle and Christine Bauer. Evaluating recommender systems: survey and framework. *ACM Computing Surveys*, 55(8):1–38, 2022.

[76] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[77] Qazi Mohammad Areeb, Mohammad Nadeem, Shahab Saquib Sohail, Raza Imam, Faiyaz Doctor, Yassine Himeur, Amir Hussain, and Abbes Amira. Filter bubbles in recommender systems: Fact or fallacy—a systematic review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(6):e1512, 2023.

[78] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686, 2014.

[79] Thiago Silveira, Min Zhang, Xiao Lin, Yiqun Liu, and Shaoping Ma. How good your recommender system is? a survey on evaluations in recommendation. *International Journal of Machine Learning and Cybernetics*, 10:813–831, 2019.

[80] Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pages 1–35, 2021.

[81] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32, 2005.

[82] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

[83] Philippe Fournier-Viger, Ted Gueniche, Souleymane Zida, and Vincent S Tseng. Erminer: sequential rule mining using equivalence classes. In *Advances in Intelligent Data Analysis XIII: 13th International Symposium, IDA 2014, Leuven, Belgium, October 30–November 1, 2014. Proceedings 13*, pages 108–119. Springer, 2014.

[84] Wil MP van der Aalst. Toward more realistic simulation models using object-centric process mining. In *ECMS*, pages 5–13, 2023.

[85] Benedikt Knopp, Mahsa Pourbafrani, and Wil MP van der Aalst. Discovering object-centric process simulation models. In *2023 5th International Conference on Process Mining (ICPM)*, pages 81–88. IEEE, 2023.

[86] Tian Li, Gyunam Park, and Wil MP van der Aalst. Checking constraints for object-centric process executions. In *International Conference on Process Mining*, pages 392–405. Springer, 2023.

[87] Statistisches Bundesamt Destatis. Kleine Unternehmen, Mittlere Unternehmen. `https://www.destatis.de/DE/Themen/Branchen-Unternehmen/Unternehmen/Kleine-Unternehmen-Mittlere-Unternehmen/aktuell-beschaeftigte.html`, 2022. Accessed: August, 12 2024.

Appendix A

TOAD: Trace Ordering for Anomaly Detection

This chapter comprises the following publication:

Richter, F., Lu, Y., **Zellner, L.**, Sontheim, J., & Seidl, T. (2020, October). TOAD: trace ordering for anomaly detection. In 2020 2nd International Conference on Process Mining (ICPM) (pp. 169-176). IEEE.
DOI: 10.1109/ICPM49681.2020.00033

Declaration of Authorship The presented idea of this work was conceived by Florian Richter who also performed experimental computations. The following substantial contributions regarding elaboration and evaluation were incorporated by Ludwig Zellner: The benefits of OPTICS, e.g. in detecting different levels of densities when applied in comparison to DBSCAN, was carved out in assistance. Following this, a filtering mechanism is proposed to detect these troughs. Finally, in the evaluation, a connection is drawn between the MinPts parameter of OPTICS and fitness scores. All co-authors discussed the results periodically and polished the final manuscript.

Appendix B

Model-aware Clustering of Non-conforming Traces

This chapter comprises the following publication:

Richter, F., **Zellner, L.**, Sontheim, J., & Seidl, T. (2019). Model-aware clustering of non-conforming traces. In On the Move to Meaningful Internet Systems: OTM 2019 Conferences: Confederated International Conferences: CoopIS, ODBASE, C&TC 2019, Rhodes, Greece, October 21–25, 2019, Proceedings (pp. 193-200). Springer International Publishing.
DOI: 10.1007/978-3-030-33246-4_12

Declaration of Authorship This manuscript is a shared main authorship between Ludwig Zellner and Florian Richter. A significant scientific contribution by Ludwig Zellner includes refining the methodology and both high- and low-level design decisions during the implementation of the approach, e.g. the measure for cluster linkage, the usage of an undirected graph and shortest paths as a distance unit. The idea was further developed in close collaboration with Janina Sontheim. Ludwig Zellner and Florian Richter performed final computations and evaluations. The first version of the manuscript was written by Ludwig Zellner and finalized by Florian Richter. All authors discussed the results regularly and polished the final manuscript.

Appendix C

TADE: Stochastic Conformance Checking Using Temporal Activity Density Estimation

This chapter comprises the following publication:

Richter, F., Sontheim, J., **Zellner, L.**, & Seidl, T. (2020). TADE: Stochastic conformance checking using temporal activity density estimation. In Business Process Management: 18th International Conference, BPM 2020, Seville, Spain, September 13–18, 2020, Proceedings 18 (pp. 220-236). Springer International Publishing.
DOI: 10.1007/978-3-030-58666-9_13

Declaration of Authorship The presented idea of this work was conceived by Florian Richter who also performed experimental computations. The subsequent fundamental contributions in both development and evaluation were included by Ludwig Zellner: Based on the identification of drawbacks when left-aligning timestamps of events another version of the approach called TADE-FC was proposed where all pairs in a case are considered. The proposal comes with the drawback of a higher runtime, so the evaluation in terms of runtime comparison was also included. All co-authors discussed the results periodically and finalized the manuscript in collaboration.

Appendix D

Concept Drift Detection on Streaming Data with Dynamic Outlier Aggregation

This chapter comprises the following publication:

Zellner, L., Richter, F., Sontheim, J., Maldonado, A., & Seidl, T. (2021). Concept drift detection on streaming data with dynamic outlier aggregation. In Process Mining Workshops: ICPM 2020 International Workshops, Padua, Italy, October 5–8, 2020, Revised Selected Papers 2 (pp. 206-217). Springer International Publishing.
DOI: 10.1007/978-3-030-72693-5_16

Declaration of Authorship Ludwig Zellner proposed the idea and discussed it with Florian Richter and Janina Sontheim. The implementation and experiments were performed by Ludwig Zellner. The results were discussed periodically with all authors. Ludwig Zellner wrote the manuscript and all authors assisted in finalizing the manuscript.

Appendix E

Process Mining Techniques for Collusion Detection in Online Exams

This chapter comprises the following publication:

Maldonado, A., **Zellner, L.**, Strickroth, S., & Seidl, T. (2024). Process Mining Techniques for Collusion Detection in Online Exams. In Process Mining Workshops: ICPM 2023 International Workshops, 2nd International Workshop Education Meets Process Mining (EduPM), Rome, Italy, October 23-27, 2023, Revised Selected Papers (Vol. 503, p. 336). Springer Nature.
DOI: 10.1007/978-3-031-56107-8_26

Declaration of Authorship The idea was proposed by Ludwig Zellner and developed by both Andrea Maldonado and Ludwig Zellner. Ludwig Zellner implemented the first version of the approach which was refined by Andrea Maldonado. The evaluation was performed in close collaboration with Andrea Maldonado. The results were discussed periodically with all authors. The manuscript was written by both Andrea Maldonado and Ludwig Zellner and finalized in assistance with all authors.

Appendix F

SCORER-Gap: Sequentially Correlated Rules for Event Recommendation Considering Gap Size

This chapter comprises the following publication:

Zellner, L., Sontheim, J., Richter, F., Lindner, G., & Seidl, T. (2021, December). Scorer-gap: sequentially correlated rules for event recommendation considering gap size. In 2021 International Conference on Data Mining Workshops (ICDMW) (pp. 925-934). IEEE.
DOI: 10.1109/ICDMW53433.2021.00121

Declaration of Authorship The research idea was conceptualized by Ludwig Zellner. Ludwig Zellner implemented the approach assisted by Gabriel Lindner. The results and evaluation were discussed with Janina Sontheim, Florian Richter and Thomas Seidl. The manuscript was written by Ludwig Zellner and finalized in assistance with all authors.

Appendix G

On Diverse and Precise Recommendations for Small and Medium-Sized Enterprises

This chapter comprises the following publication:

Zellner, L., Rauch, S., Sontheim, J., & Seidl, T. (2024, April). On Diverse and Precise Recommendations for Small and Medium-Sized Enterprises. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 118-130). Singapore: Springer Nature Singapore.
DOI: 10.1007/978-981-97-2262-4_10

Declaration of Authorship The idea was proposed by Ludwig Zellner and discussed with Janina Sontheim and Thomas Seidl. Ludwig Zellner implemented the approach and conducted the experiments. The evaluation was performed in close collaboration with Simon Rauch and Janina Sontheim. The manuscript was written by Ludwig Zellner and finalized in assistance with all authors.