1 Introduction

Digitalization has influenced nearly all areas of life and is an unstoppable phenomenon for upcoming times. Whether it is in the communication, shopping, entertainment, or mobility area, digitalization has changed business models and how these actions are carried out. Former paper-based processes are now possible only using a digital device or have become entirely obsolete. Furthermore, new business models and processes were just not possible before digitalization had entered the stage [1].

Digitalization comprises a multi-dimensional concept whose capture is not trivial. For this purpose, research has developed various digital maturity models [TMB20] that try to capture the current digital status of an organization, whereby business processes and data were discovered to be of prime importance.

Processes characterize the creation of value for any business. Traditionally, Business Process Management (BPM) is the field that deals with all aspects of processes in organizations. It focuses on understanding and modeling existing processes and identifying the differences between desired process executions. BPM has its origins early in the 1990s, with the work of Hammer [Ham90] introducing Business Process Reeingeneering (BPR), claiming that companies should rather completely rework their work procedures instead of automating them. Today, processes are often supported by at least one information system and, therefore, produce large amounts of data [MSW11]. Processes consume data to fulfill their purposes and produce new data based on their executions. Exploiting

data sources is crucial for a large majority of business models today. The data mining discipline, nowadays often associated with the *Big Data* term, focuses on exploiting all kinds of data sources to understand or predict relationships in a domain [TSK16].

Recently, data mining techniques have been applied to process data in the form of so-called event logs. Process mining [Wil16] has arisen, which uses the traces of process executions in IT systems to discover real-world processes. With that technology, process modeling is now a semi-automated task that does not require manual modeling activities. In general, process mining connects the BPM and data mining disciplines by exploiting data sources for all kinds of process analyses.

Research progress in process mining directly contributes to advancements in BPM which can be seen as one dimension of digitalization. Like other research efforts in the information systems discipline, process mining is closely linked between research developments and the technology application in practice. New findings in research are applied in practice to verify whether they hold in reality. Reversely, observations in the industry lead to research initiatives to explain observations or come up with solutions for a problem observed. As process mining is a relatively new technology and its application in practice is only at the beginning in many environments, questions arise of how to integrate it in the existing BPM environment and which benefit it can deliver in what kind of circumstances.

1.1 Motivation

Process Mining allows companies and other organizations to get an overview of the actual status of their business processes. The foundations for process mining were laid in 2012 with the Process Mining Manifesto [vdAAdM+12] with van der Aalst being one of its principal founders

who published a work that summarizes a large share of the knowledge about this technology in 2016 [Wil16]. As a subset of BPM, process modeling tries to capture and visualize the flow of processes happening in any organization. Such activity typically results in a process model or diagram in a notation like Petri nets or BPMN. Typically, process mining applications in larger companies consider Order-to-Cash (O2C), Purchase-to-Pay (P2P), production or administrative processes. Such processes are often supported by various information systems, whereby some may even communicate with external systems from customers or suppliers [24]. In many cases, modeling such processes is not feasible [DGDMM15] because, depending on the abstraction degree, they consist of many activities and have many variants, meaning that the activities can be gone through in various ways. There are no or only a few conditions and dependencies that enforce a particular order of the process. The flow of the process is relatively free as long as it adheres to the logical structures (e.g., XOR-gateways) and starts in a start and ends in an end state.

For this reason, process mining, where the process model is automatically generated based on the input of process execution data, may be a suitable option. Van der Aalst et al. define process mining as a technique "to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's (information) systems" [vdAAdM+12]. The challenge is to collect all relevant event data from the various information systems and integrate and transform them into one event log. Assuming this hurdle is cleared, the file can be imported into a process mining tool of choice. Then, a process discovery routine generates a process model underlying the event data input. In some cases, this results in a large, complex, spaghetti-like diagram that is confusing and does not deliver any concise overview.

Still, modelers and other process stakeholders would like to know the process's most critical procedures and relations. Declarative model-

ing [DGDMM15] notations can help here as they introduce constraints and conditions on the process without defining the exact flow of the process in detail. Such constraints and conditions can originate from either process mining activities taking the event log as the source or manual modeling activities by process modelers using their domain knowledge about the process or conducting interviews with employees who are part of the process under consideration. In this way, even though traditional process modeling and mining cannot deliver satisfying analysis results, process analysts could nevertheless deliver possibly meaningful insights that help to improve the process performance.

All process stakeholders, including management people, must understand the resulting process models to ensure the analyses are beneficial. One widely-used notation for modeling processes in a declarative manner is Declare [PSVdA07]. It consists of a set of constraint templates restricting the existence of single activities or the relation of two or more activities between each other. Besides, a multi-perspective extension of Declare can also capture data-related conditions for their associated activities. As declarative notations for process modeling are nowhere near as widely used as imperative notations like the industry-standard Business Process Model and Notation (BPMN), employees that are not professional process modelers may have difficulties grasping the process depicted in such models or at least need some time to familiarize themselves [FLM+09]. Therefore, another format with which non-experts in modeling are more familiar but can still represent the process coherences may be desirable.

For a long time, data mining has used association rules [ZB03] and sequential patterns [16] to describe relations in all kinds of datasets, whereby the market basket analysis is the most prominent application scenario. Association rules create a statement that whenever a customer buys product A, he or she is likely to buy product B, whereas sequential patterns show frequent sequential orders in which customers buy products. Because

of their conciseness, both association rules and sequential patterns are intuitively understandable for most employees with various backgrounds. Transferring their application from market basket to process analysis means that the activities of a process are the new products and a process instance is one customer transaction describing the products a customer buys. In this way, they are an intermediary representation of process relations but not yet translated to model elements of the declarative process model notation. Ideally, they are derived directly from the event log input that is the same as for "traditional" process mining resulting in process graphs.

Consider the following motivating example that may happen similarly in practice based on a sample event log from the internet. It was used for the BPI Challenge 2019 [18] and describes the purchase order process of a larger multi-national company based in the Netherlands in one year. While the scenario is fictive, the goal is to illustrate the benefits of combining the declarative process modeling paradigm and process mining in such application settings.

Motivating Example

The setting is a company *Paint & Co.* producing coatings and paints for industrial applications. Controller C detects a notification at her Business Intelligence (BI) dashboard indicating that recently revenue numbers have dropped. After a chat with her colleagues in the production department, she finds out that production had to reduce the production speed as the company lacks raw material. Furthermore, she hears that supply companies complained about the unreliability of the purchasing department of *Paint & Co.* C wants to get to the bottom of things and talks to a process analysis department colleague.

The purchase order process is currently not part of the department's analyses as it was found to be very hard to be modeled because of its

large number of variants. Furthermore, the focus of the process analysis team was set on the production processes as higher cost-saving and performance improvement potentials were expected there. The process analysis team recently included process mining into their skills portfolio and toolbox as the team thinks it could benefit the company substantially. However, the application of process mining in *Paint & Co.* is still at the beginning. There have been successful pilot projects with more straightforward and smaller processes that have led to changes and optimizations. One reason why process mining often has not yet been established for larger, more complex processes is that data collection and integration requires significant effort with the introduction of ETL (Extract, Transform, Load) pipelines. C and her colleague agree that they should apply process mining on the purchase order process to clarify the reason for the shortage of raw material.

They contact the purchasing department of *Paint & Co.* and explain their project. The purchasing department provides them with an extract of all ERP system tables that come in contact with the purchase order process. With these extracts, BI and process analysts, with the help of a data science expert, build an event log that describes the executions of the purchase order process in the last five years. After importing the file in a process mining tool (Disco [4] in this example), they start the process discovery process. It results in the process model shown in Figure 1.1.

C and her colleague from the process analysis team are surprised because the model is confusing even though it only includes 50 percent of the paths. They can identify the happy path of the process through the color coding and varying thickness of the output, indicating which of the activities and paths are frequent. First, *Paint & Co.* creates a purchase order item (activity *Create Purchase Order Item*) which includes the products the company would like to buy and leads to the buying contract between the company and the suppliers. Then, the vendor creates an invoice, and

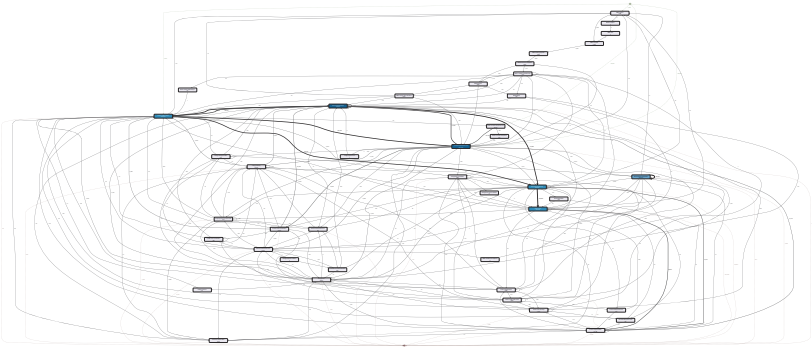

Figure 1.1: Discovery of Purchase Order Process in Disco [4].

the goods and invoice receipts are recorded, i.e., stored in the system. In some cases, the invoice creation is skipped or executed after the receipt of the goods has been recorded. Finally, the invoice is closed and archived (activity *Clear Invoice*). C and her colleague do not identify any significant deviations from what is expected and cannot find the problem that leads to the missing raw materials as relevant statistics like throughput time are in the norm.

Therefore, they decide to split the event data and only load the purchase order process cases of the last two months into the process mining tools because this is roughly the time where the complaints of the suppliers and the lack of raw material started. After generating the process model, it can be seen that the frequency of activities in the cases has changed. The happy path remains unchanged, but another path from *Vendor creates invoice* to *Create Purchase Requisition Item* back to the *Create Purchase Order Item* activity grows substantially regarding frequency. In a significant amount of cases, *Paint & Co.* requests new items after the vendor has already created the invoice.

Such process behavior is unusual and represents an exceptional case that requires further investigation. Now that C and her colleague have an indication of why the purchase order process may currently not be optimal and cause the problems with the procurement of the raw materials, they still do not know why the purchasing department executes the purchase orders in this way. A small interview session with the employees does not lead to new conclusions because they state that nothing has changed and the purchase orders are executed in the same way like they have been in the previous years. The BI and process analyst conferred to discuss the next steps with this unsatisfying situation.

The process analyst proposes to apply association rule and sequential pattern mining on the event log to derive frequent relations on the process. Thereby, these fundamental data mining techniques shall be enriched with data, i.e., attribute values associated with each activity execution. In this way, the execution path of the process is linked with attribute values. Still, it requires some domain knowledge to interpret the resulting rules and patterns and decide whether they represent behavior outside the happy path that should generally be avoided or one should try to keep infrequent. The data science of team of *Paint & Co.* supports the process of applying the two data mining techniques on the event log and develops a Python script that produces association rules and sequential patterns based on minimum support and confidence values. Again, the script uses the data from the last two months. After playing around with these settings for some time, C and her colleague come across the following sequential pattern, with relatively high support of 0.5, raising their interest:

{Vendor Creates Invoice, Create Purchase Requisition Item/user_005, Create Purchase Order Item}

The *user_005* value originates from the *User* attribute of the event log, which is a so-called event attribute that specifies the user who was

responsible for executing a particular activity (if known and meaningful, can also be empty). A slash between *Create Purchase Requisition Item* and *user_005* indicates that this specific value for the *User* attribute and the activity appeared together. The pattern shows that the rework of the purchase order items is significantly associated with one particular user. C and her colleague arrange a confidential talk with the head of the procurement department. The de-anonymization of the *user id* shows that the employee associated is relatively new in the company. Here, it is crucial always to be cautious of sensitive user-related that could lead to accusations of individual employees. Privacy issues are an important (research) topic for the process mining field.

The problem solution is found after a private face-to-face talk of the procurement manager and the employee, including a live walkthrough through the purchase order process. Instead of creating new purchase orders for every product *Paint & Co.* wants to get delivered, the employee created a purchase requisition item based on the same invoice number when he believed that the additional products have a content-related connection to an already existing order. In this way, the supply company receives a purchase order associated with an invoice number that already exists and that was possibly already closed and archived, interrupting the automatic workflow at the supplier's side. An employee of the sales department has to manually inspect the case and initiate the shipment of the products, causing delivery delays. This explains why *Paint & Co.* lacks some essential raw materials as the company applies Just-in-time production with minimal storage capacity from which it could draw on.

Still, the possibility to create a purchase requisition item based on an existing invoice is intended and was designed for cases where the supply company could not deliver parts of the order. In some of these cases, *Paint & Co.* falls back for other suppliers to obtain the products. However, a supplier may generally be the only one selling one type of material or

others do not have it in stock as well. The purchase requisition shall explicitly state that *Paint & Co.* asks the supplier to restock these products to deliver them in the future even though this could take longer than usual and therefore violate delivery time agreements. In total, the supply company could also use the findings of the sequential pattern analysis to make their sales processes more robust against such deviations so that they do not lead to the loss of automation.

C and her colleague are content with the outcome of their joint project. They could find the cause of the short raw materials. For the future, they agree that applying association rule and sequential pattern mining to process event data seems promising and should be continued with other processes, especially those that were not touched before by process analysis due to their complexity and deep integration in various company parts. The association rules and sequential patterns were easily understandable, also by non-process-experts, whereby sequential patterns were found to be more beneficial for understanding the process steps chronology.

Some questions remain open, for instance, how this new procedure of applying fundamental data mining techniques on event data should be included in the current process analysis cycle. Currently, the process analysis team of *Paint & Co.* starts with context analysis of the process and conducts interviews with employees who work with the process afterwards, which results in process models. Should process mining, in general, happen after a context analysis, and should, if possible, both modeling and mining be performed for one process? In case the process analysis team performs both techniques, how should *Paint & Co.* deal with non-matching results? Related to that is how to represent the association rules and sequential patterns in a global process model.

It is already clear that a declarative modeling notation with the opportunity to model data-related relations fits this purpose best here. Still, there is the question of which rules and patterns lead to which model

elements in the declarative modeling notation. C and her colleague raise the question of whether the notations currently described in the literature are sufficient or they have to be extended to make them more fitting to represent the mining output. Furthermore, both agree that general guidelines providing indications about parameter settings and the question of which rules and patterns should be selected for including them in a declarative model could be immensely helpful.

1.2 Research Objectives

The research foci of this thesis are two-fold. One part concentrates on rather technical issues of the implementation. It shows how a general implementation of association rule and sequential pattern mining could look like in programming languages dealing with datasets. Thereby, the thesis expands on all preprocessing steps and parameter settings in detail to finally apply the two fundamental data mining techniques. Additionally, it illustrates how various types of resulting association rules and patterns can be transformed to a Declare constraint to be included in a declarative process model based on a compact sample process event data set. The following questions capture the main contributions of this thesis.

- **RQ1** How to develop a general implementation of applying association rule and sequential pattern mining to event data in the form of event logs for use in practice?
- **RQ2** How can resulting rules and patterns be translated to declarative model elements, and are the existing (multi-perspective) notations appropriate and fitting to capture their meaning?

Another aspect focuses on integrating the declarative process mining approach presented in this thesis and the declarative paradigm in general

into a company's structure. It addresses how and for whom a rule/pattern and declarative model representation may be beneficial, compared to existing declarative process mining approaches whose algorithmic details may not be transparent to non-technical users. Both parts are supported by applying the approach to sample event logs from web repositories; however, real-world applications are not included and left for future research interests. Therefore, the following two questions are not explicitly answered but touched in the course of the thesis.

- Is the association rule and sequential pattern representation beneficial for users, especially those without knowledge about algorithmic details and programming languages?
- How should the application of association rule and sequential pattern mining be integrated into the BPM life cycle to maximize a company's benefit for process understanding?

1.3 Thesis Structure

The remaining thesis is structured as follows: Part 1 encompasses the foundations that play a substantial role throughout the thesis. Chapter 2 introduces a maturity model to assess a company's degree of digitalization concerning a set of dimensions. It identifies processes as the most critical aspect of digitalization and motivates the joint consideration of Business Process Management (BPM) and process mining. Both concepts or research fields are introduced in more detail in Chapter 3. Process mining enriches the traditional BPM field with data-driven analyses of real-world processes. Together, their application could result in findings for future process analysis initiatives that were not possible before. Next, Chapter 4 recalls association rule and sequential pattern mining as two fundamental data mining techniques.

Part 2 of this thesis combines the data and process mining techniques. Starting from their traditional, typical application settings like market basket analysis, Chapter 5 transfers them to the process analysis field and sets up a process to apply them to event logs for process mining. Furthermore, the chapter shows how the resulting rules and patterns can be translated to model elements in the (multi-perspective) Declare notation and how these notations can be adapted to represent them even more precisely. Chapter 6 evaluates the approach by applying it to sample event logs of different sizes from web repositories. This results in full declarative models and assesses whether and how process analysis and understanding benefit from this representation.

At last, Part 3 relates the contribution of this thesis with the research context and compares it with existing declarative process mining approaches. The goal is to demonstrate that the straightforward generation of association rules and sequential from event logs benefits the common understanding of a process model and a transformation to declarative model elements and backwards is practical. Finally, Chapter 8 concludes the thesis and points out possible directions of future research.