EasyChair Preprint

№ 1111

Mining Process Mining Practices: An Exploratory Characterization of Information Needs in Process Analytics

Christopher Klinkmüller, Richard Müller and Ingo Weber

June 8, 2019

Mining Process Mining Practices: An Exploratory Characterization of Information Needs in Process Analytics

Christopher Klinkmüller[1], Richard Müller[2], and Ingo Weber[1]

[1] Data61, CSIRO, Eveleigh, NSW, Australia
{christopher.klinkmuller,ingo.weber}@data61.csiro.au
[2] Leipzig University, Leipzig, Germany
rmueller@wifa.uni-leipzig.de

Abstract. Many business process management activities benefit from the investigation of event data. Thus, research, foremost in the field of process mining, has focused on developing appropriate analysis techniques, visual idioms, methodologies, and tools. Despite the enormous effort, the analysis process itself can still be fragmented and inconvenient: analysts often apply various tools and ad-hoc scripts to satisfy information needs. Therefore, our goal is to better understand the specific information needs of process analysts. To this end, we characterize and examine domain problems, data, analysis methods, and visualization techniques associated with visual representations in 71 analysis reports. We focus on the representations, as they are of central importance for understanding and conveying information derived from event data. Our contribution lies in the explication of the current state of practice, enabling the evaluation of existing as well as the creation of new approaches and tools against the background of actual, practical needs.

Keywords: Process Mining · Visual Analytics · Qualitative Content Analysis

1 Introduction

Many activities in phases of the business process management life-cycle, including process discovery, analysis and monitoring [4], benefit from the investigation of event logs that were generated during the execution of a business process. Such event data can be used to answer questions like "Does the process behave as expected?" or "Are there any bottlenecks that negatively impact process performance?". Commonly, those high-level *domain problems* are too complex to be straightforwardly answered by applying a single analysis technique, and thus analysts divide them into more fine-grain questions, leading to lower-level *information needs* that can be satisfied through the application of analysis techniques. While this divide-and-conquer strategy enables experts to iteratively form a mental picture of the business process, analysts also "[...] often do not know what

they do not know" [19, p.43]. Consequently, the information needs are rarely predetermined, but arise from insights gained during the analysis process [7].

Research, predominantly in the field of process mining, has developed a plethora of approaches, e.g. [9,17,18] that enable analysts to satisfy specific types of information needs. Commercial and academic tools (like Apromore, Celonis, Disco, Everflow, Lana, myInvenio, ProM, QPR, TimelinePI, etc.) offer bundles of readily available analysis techniques. Moreover, project methodologies such as [21,3,23] provide universal, problem-independent guidelines for the application of such techniques in process mining projects. Due to the maturity of those research outcomes, they are increasingly adopted in real-world analysis projects, enabling us to examine those projects and elicit insights into the analysts' work practices. So far, reviews of such projects have focused on categorizing re-occurring problems [1,20], but lack insights into strategies that analysts choose to find answers to the domain problems. Yet, such insights would provide a foundation for further refining and enhancing the available approaches and tools.

On this basis, we aim to refine our understanding of the *relationship between the domain problems and the information needs* that arise in analysis projects. To this end, we conduct a systematic study as per [6] and analyze a corpus of 71 project reports that resulted from the problem-driven analysis of real-world event data in the context of the annual business process intelligence challenge (BPIC). While the significance of such studies was in general highlighted in [12,13], our particular contributions to process mining, visual process analytics, and business process management are twofold. First, the schema that we use to examine work practices can serve as a general reference point for assessing existing or for ideating advanced analysis approaches. Second, we take a first step towards a shared and refined understanding of work practices in process mining projects and present a consolidated overview of such practices from a large number of analysis projects. In future work, researchers can rely on these insights to orient the design of techniques towards actual, practical needs. We also hope that our work stimulates further analysis of work practices.

Specific findings from our study show that discovery of control flow is often conducted by analysts to establish a basic understanding of the business process, whereas other problems like the investigation of the time, case or organizational perspectives constitute the actual goal of the project. Moreover, for discovery analysts heavily utilize process mining algorithms to obtain descriptive process models, indicating that the low-level analysis techniques match the domain problem well. By contrast, for other domain problems analysts rely on general-purpose techniques or tables, pointing to situations where the analysis techniques do not match the domain problems. We also derive a set of eight frequent work practice patterns to provide direction for future work.

Following, we describe our methodology including the analyzed material and discuss limitations of our study in Section 2. In Section 3, we outline the annotation schema used to systematically describe the information needs and domain problems. In Section 4 we present the insights from our analysis. We conclude with a summary of related work in Section 5 and of our findings in Section 6.

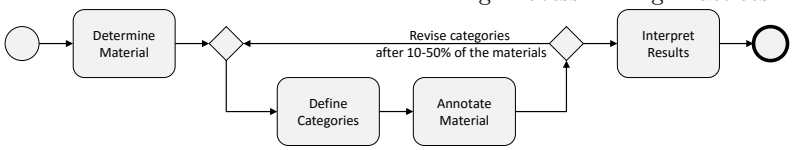


Fig. 1: The qualitative content analysis process (cf. [11])

2 Research Methodology

In this work, we adopted a qualitative research approach, which is suitable in situations like ours where a deeper understanding of a phenomenon is developed by investigating information material [16]. To this end, we followed guidelines for qualitative content analysis [11] and applied the analysis process depicted in Fig. 1. Following, we outline each of the activities and discuss limitations.

2.1 Step 1 - *Determine Material*

As source material we used all BPIC reports available to date. The annual BPI Challenge has been organized in conjunction with the international workshop on business process intelligence[3] since 2011. Every year the challenge publishes a dataset containing real-world event logs. The dataset is provided by an organization from industry or government which asked questions related to the underlying business process (except for the first year). Upon publication of the dataset, the organizers invite analysts from academia and industry who are given a few months time to answer the questions by analyzing the dataset and to submit a report. Frequently, the analysts were invited to express any other interesting insights they obtained. Finally, a committee examines the reports and awards the best submissions. At the time of writing, eight BPIC editions were conducted and a total of 71 reports were published with 213 contributors co-authoring at least one report. The reports cover a broad range of scenarios and involve an extensive number of analysts, both from industry and academia, and therefore form a solid basis for obtaining insights into business process analysis practices.

In the study, we focused on analyzing the visual representations from those reports, including amongst others process models, charts, network diagrams, and tables. The reason is that those representations are the major means to convey information related to the underlying business process. Hence, we regard them to be representative of the low-level information needs that arose during the analysis project. Resulting from the application of specific analysis techniques they also provide an overview of those techniques' capabilities. Yet, not all representations were relevant to our study, as some do not reflect a low-level information need. For example, some representations are about the applied methodology, algorithms or tools, or the quality of a prediction model. We thus defined the following inclusion criterion: *a visual representation must be generated from the provided event data and it must be used for explaining aspects of the underlying business process.* In total, we yielded a set of 2021 visual representations.

[3] `https://www.win.tue.nl/bpi/`, Accessed: 12/02/2019

2.2 Steps 2 and 3 - *Define Categories* and *Annotate Material*

We next needed to describe the visual representations. As we wanted to analyze the descriptions and derive patterns of work practices from them, it was important that they rely on a consistent vocabulary. Thus, we followed guidelines for qualitative content analysis [11] and determined a set of categories that refer to the dimensions of the representations that we wanted to examine. The dimensions refer to the information need associated with the representations as well as the high-level questions that representations contribute to. Here, we abstract from the applied categories (details are provided in Section 3) and focus on the applied methodology. For each category, we then needed to define the set of codes which we used to encode the characteristics that the visual representations show with regard to the respective dimension. These sets must be *exhaustive* and *mutually exclusive* [8], so that (i) the codes cover all relevant aspects, (ii) all visual representations can be annotated appropriately, and (iii) the codes refer to distinct concepts, in order to guarantee that each representation can be described clearly and that there are no two ways of describing a visual representation.

We applied the following procedure to infer the category codes. First, we determined the categories and derived initial code sets from the literature. Then, we began to annotate the visual representations using these categories and codes. While the categories remained unchanged during the study, our code definitions occasionally underwent conceptual changes. That is, when we encountered representations that could not be described appropriately using the code set, we introduced new codes. Additionally, we sometimes experienced that our perception of a certain code changed during the annotation procedure. Due to those conceptual changes, we needed to consolidate the sets of category codes from time to time. Moreover, after a consolidation we revisited previous annotations to ensure consistency with the new schema. These updates occurred during the annotation of the first 50% of the visual representations. After that the schema was mature and could be applied without further changes. Finally, the questions posed in the challenge were annotated as well.

The annotation of visual representations itself was primarily conducted by one author of the paper, and the annotation of the challenge questions was done by another author independently. To ensure high quality of the annotations, we implemented the following procedures. First, the definition of the categories was frequently discussed by all authors. Second, the respective other authors of the paper conducted random sample checks to validate the annotations. Third, annotations that were challenging were discussed among all authors.

2.3 Step 4 - *Interpret Results*

Lastly, we derived descriptions of work practices from the annotations by summarizing and relating them, in order to identify trends in the work practices. In this context, we mostly analyzed the annotations by means of frequency distributions, and pattern mining. The results are presented in Section 4.

2.4 Limitations

To any study like ours, a number of limitations and threats to validity are inherent. We discuss the main factors and our approaches to mitigation below.

First, there could be personal bias: the annotation process relies on our subjective perception, and the interpretation was driven by insights relevant to us. We aimed to mitigate this issue as discussed above, but a residual risk remains.

Second, the representativeness of the data and results might be limited. Our source data stems from the BPI Challenge and might differ from process analytics practices in industry. This point is, to a degree, mitigated by the data and challenge questions stemming directly from real-world organizations, as well as by the large numbers of co-authors (>200) and visual representations (>2000).

Finally, the insights into work practices are restricted by the method of sourcing data from the *results* of these practices only. In particular, visual representations in the reports were exclusively two-dimensional and static; in contrast, analysts can interact with tools and data. Also, the reports cannot be assumed to show the full analysis process, e.g., for some information needs the analysts might not have found satisfactory results, and hence did not include any representations in the report. However, in the challenge setting with multiple teams addressing each question, this issue is partly mitigated: as long as *any* team has answered an information need, the data was included in our study. Next, visual representations were annotated based on the respective report's content and structure, which might not cover all influences that a representation had on the analysis process. Further, the choice of visual representations might be based on personal preference or tool access. To mitigate the risk of overemphasizing the visual aspects, we did not only focus on how data was presented, but we also investigated what and why data was analyzed (see Section 3).

While some of these limitations and threats could not be mitigated in the chosen study design, we believe the insights gained and described in the following to be of high relevance to advancing the fields of process mining and analytics.

3 The Annotation Schema: Categories and Codes

During the annotation, we focused on describing information needs and domain problems that are associated with the visual representations. According to [13], understanding these two aspects is a prerequisite for the development of data visualization tools. Hence, we defined the categories shown in Fig. 2.

The first category that we considered is the *domain problem*. It refers to the general question that was posed by the dataset provider or that the analysts found interesting to explore. The argumentation related to such a question is commonly not backed up by one, but by multiple visual representations. As a consequence, the first step in annotating the representations within a report was to identify the domain problems that this report examined. For each of the questions, we then introduced a conceptual section and assigned all visual representations that are related to the respective problem to that section. We

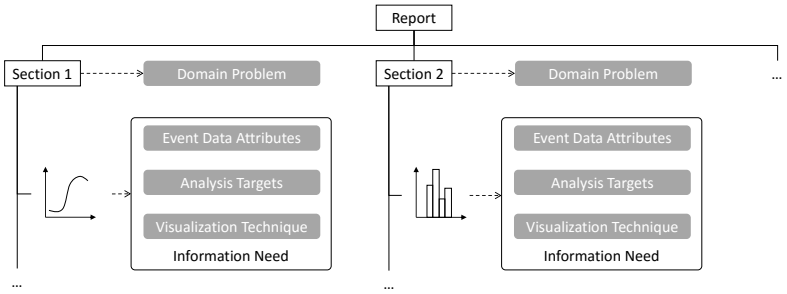


Fig. 2: Categories for the annotation of visual representations

also annotated the sections and thus by extension the representations with the code for the respective domain problem. The resulting conceptual document structure is oriented towards, but does *not* necessarily represent the structure of the report itself, as, e.g., some visual representations were listed in the appendix and referenced in the text, an executive summary outlined basic findings that were presented in more detail in separate sections, or the logical section structure was very fine-grained and divided visual representations by irrelevant aspects. Further, we only assigned representations to one section based on the context in which they were referenced. We hence might ignore their relevance to other sections. Yet, without further inquiry the assignment to other sections reflects our subjective interpretation, but unlikely the representations' actual influence.

We then annotated the visual representations, focusing on the information needs that are linked to them. To this end, we followed the guidelines from [13] that suggest to define a visual representation in terms of what, why, and how data is analyzed. First, we examined what part of the *event data* was used to generate the visual representation. Second, with regard to the why-dimension we focused on the *analysis target*. This category is related to the relationship in the data that is expressed by the visual representation. Finally, we captured how the data was represented by annotating the *visualization technique*. Note that some visual representations might serve multiple information needs; especially tables contained different types of data which needed to be distinguished. Consequently, we obtained 2085 information needs for the 2021 visual representations. In the following, we introduce the specific codes for each of the categories.

Domain Problem. The purpose of this category is to provide an abstract encoding for the specific domain problems that are investigated in the report. In this regard, we derived our initial set of five codes from the process mining use cases [1] and the more general BPM use cases [20]. This set included the problems of *process discovery* where a process model describing the control flow is inferred from the data and of *conformance checking* which deals with verifying that the behavior in the event log adheres to a set of business rules, e.g., defined as a process model. While these two use cases focus on the control-flow perspective, there are three enhancement use cases which refer to other perspectives. Domain problems related to the *time perspective* deal with understanding the performance of the process such as throughput times, working times or waiting times. The *organizational perspective* focuses on the utilization of resources and

their dependencies and the *case perspective* deals with the influence of other process attributes, e.g., related to the customer, on the behavior.

During the annotation process, we identified three additional domain problems. First, there are *prediction* problems where analysts aimed to create models that can forecast the development of process instances. This type is strongly related to the case perspective, as it is about comprehending the influences of attributes on the process behavior. However, given its explicit focus on prediction, we decided to capture it separately. Second, *drift detection* aims to recognize points in time at which the underlying behavior of a process changed and to provide details regarding this change. Finally, *familiarization* is an activity that helps experts to understand basic characteristics of the business process and the event data. While not necessarily related to a specific business question, we included it in our study due to its significance for the analysis process.

Event Data Attributes. This category refers to the parts of the data that the visual representation examines and is thus used to capture the attributes in the data that are investigated to satisfy the information need. The codes for this category are not based on a categorization from the literature, but were developed in the context of our study. A first set of codes refers to the entities that are examined in a visual representation. These entities include *cases* representing single process instances and *activity instances* within those cases representing the execution of a certain *activity*. An activity can belong to a *subprocess*. A case often processes an *item*, e.g., a claim, a product, or a diagnosis, and involves *external partners*, e.g., customers or suppliers, as well as *organizational entities* which perform activities or who oversee a case. Types of organizational entities include resources, departments, branches, and locations. Analysts are also interested in relationships between these entities. The *control flow* refers to constraints on the ordering of activities at the process level. The *conformance* to such a control flow definition can be examined at the individual or the aggregated case level. Similarly, *execution patterns* are related to whether a case shows a certain type of behavior or not. With regard to the organizational units, *responsibilities* are often investigated, i.e., the activities that resources work on. Additionally, analysts are interested in the *organizational hierarchy* to identify teams and they evaluate *work practices* which focus on combinations of resources that frequently work on the same cases. The last set of analysis attributes is related to timing. Here, *durations* are examined with regard to the individual or groups of cases as well as to resources and their performance. The data can also be clustered or narrowed down by focusing on certain *time points*, such as years, months, weeks, weekdays, mornings, etc. In this context, the *execution status* of a case at a certain point is a specific derived attribute. Finally, *drift scores* provide information on how well the behavior in a case is aligned with the behavior in cases that were handled in a given time window.

Analysis Targets. There are different ways in which the attributes can be examined. In this regard, we capture the analysis targets. Here, the analysis targets specified in [14] served as a basis for our annotation. There are targets that refer to the entities within the dataset. In this context, *trends* describe

overall characteristics of the entities, *outliers* are entities that do not adhere to these characteristics, and *features* are patterns that outline interesting structures within the data. Attribute-specific targets include those that are focused on single attributes: its *distribution* or its *extremes*, i.e., the minimum and maximum values. Relationships between attributes can be quantified based on their *correlation*, i.e., the degree to which their values are related. A *dependency* between attributes exists if the values of one attribute determine values of the other. Additionally, the *similarity* is a quantitative measure that is based on all values of an attribute. Finally, data might be represented as a graph to inspect its *topology*. We also recognized one additional target: *meta-information* is important for analysts to understand the attributes' meaning.

Visualization Technique. The last category refers to the visualization technique that is applied, to make the data interpretable. In this regard, we used the terminology from the data visualization catalogue[4] which specifies general-purpose techniques. The techniques applied in the reports are *bar chart* (including column charts and multi-set versions), *box and whisker plot*, *chloropleth map*, *chord diagram*, *heatmap*, *line graph*, *network diagrams*, *pie chart*, *radar chart*, *scatter plot*, *table*, *tree diagram*, *treemap*, *venn diagram* and *word cloud*. Detailed information on each of these techniques can be found in the catalog.

As can be expected, the source data included process-specific visualization techniques. Following our methodology, we added these to our vocabulary during annotation. Specifically, there are two types of specialized network diagrams. The *process model* depicts the control-flow of a process and the *social network* the relationships between organizational units. The *dotted chart* is a specific scatter plot used to visualize the correlation of attributes of activity instances such as timestamps, activities, resources, and cases. Finally, the *trace alignment* is a table-based technique that shows the sequences of activity instances for a set of cases and how their sequential ordering is aligned with a default ordering.

4 Analysis of Mining Practices

We now evaluate the information needs and domain problems. In particular, we describe patterns of mining practices that we detected based on our annotations. In Section 4.1, we provide an overview of all domain problems. We then use the insights to prioritize the domain problems and present a detailed analysis of the most important problems in Section 4.2.

4.1 Holistic View

Our first analysis focuses on the importance of the domain problems to the analysts. As an importance indicator we computed the absolute frequencies of information needs for each combination of domain problem and BPIC edition. For better comparability, we normalized the frequencies per edition, i.e., based

[4] `https://datavizcatalogue.com`

Table 1: Distribution of the domain problems per year

	2011	2012	2013	2014	2015	2016	2017	2018	Avg.
Discovery	**55.6%**	**28.4%**	5.5%	4.8%	1.5%	0%	11%	7.3%	14.3%
Conformance	0%	3.4%	32.3%	0.9%	0%	0%	0.6%	0%	4.7%
Time Pers.	0%	20.5%	0%	5.1%	19.5%	2.9%	23.5%	0%	8.9%
Org. Pers.	8.3%	13.6%	3.1%	4.5%	**37.9%**	0%	8.7%	13%	11.2%
Case Pers.	13.9%	6.3%	**54.4%**	**60.7%**	19.9%	**80.3%**	**44.3%**	24.6%	**38.1%**
Prediction	0%	1.1%	0%	3%	0%	0%	0.8%	1.5%	0.8%
Drift Detection	0%	0%	0%	6.9%	8.8%	6.6%	0.3%	23.2%	5.7%
Familiarization	22.2%	26.7%	4.7%	14.1%	12.3%	10.2%	10.7%	**30.4%**	16.4%

on the total number of information needs within an edition. Table 1 shows these frequencies and their averages, per domain problem.

In the first edition in 2011, discovery was the dominating domain problem; it also was the problem that the analysts focused on the most in 2012, although the other domain problems started to receive increased attention. In the remaining editions the case perspective is the most frequently investigated problem. In this regard, 2018 is an exception where many information needs arose during familiarization and the case perspective ranked second. On average, the case perspective was the most important problem. A large share of the information needs also emerged during familiarization and discovery. Moreover, while conformance checking, prediction, and drift detection only played minor roles, the time and organizational perspectives were moderately important.

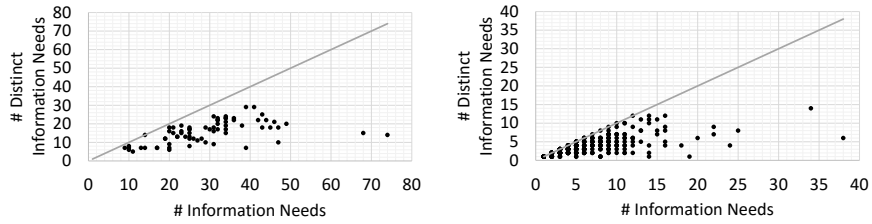
Next, we compared the importance of the domain problems assigned by the analysts to the importance assigned by the organizations that provided the datasets. To this end, we determined the problem frequencies based on the domain problems that we assigned to these questions. However, about 10% of the questions asked for any interesting insights beyond those addressed by the other questions without providing further direction; for these, we did not assign any problem. Additionally, familiarization was not present as a domain problem, as it is a task that analysts conduct to prepare for the examination of the domain problems. Similar to the reports, in the questions perspective-related problems ranked first, with the case perspective being associated with 29.8% of the questions, the organizational perspective with 14.9% and the time perspective with 10.7%. The group of conformance checking, drift detection and prediction were the subject of 5.3% to 10.7% of the questions. Interestingly, discovery was only posed as a domain problem by the organizations in three years and hence only 8% of the questions were related to it. We hypothesize that the mismatch between the importance of discovery for organizations and for analysts can be traced backed to the relevance of discovery for establishing a basic understanding of the underlying business process. That is, in accordance with the L* life-cycle model [21] analysts rely on the insights from this activity for the investigation of the other domain problems. Consequently, for analysts discovery often played a role similar to familiarization and supported analysts in their preparation efforts.

Table 2: Correlation between visualization techniques and domain problems

	Discovery	Conformance	Time Pers.	Org. Pers.	Case Pers.	Prediction	Drift Detection	Familiarization
Bar Chart	6.4%	14.8%	15%	10.3%	14.3%	14.1%	14.2%	13.3%
Chord Diagram	0%	0%	0%	0%	1.7%	0%	0%	0.8%
Line Chart	2%	5.8%	7.7%	5.4%	11.5%	26.9%	6.4%	9%
Network Diagram	0%	0%	0.4%	3.1%	0.5%	0%	2.8%	1%
Pie Chart	0%	0.6%	0.4%	0.4%	2.3%	1.3%	0.7%	1.3%
Scatterplot	0%	0%	1.8%	1.8%	2.6%	10.3%	1.1%	2.1%
Tree	1.5%	0%	0.7%	1.8%	1.4%	0%	1.8%	1.3%
Other	1%	0%	0%	1.8%	1.6%	0%	1.1%	1.1%
General-purpose	10.9%	21.3%	25.9%	24.6%	36%	**52.6%**	28%	29.8%
Heatmap	0.5%	0%	0%	0.4%	1.6%	0%	0.7%	0.9%
Table	20.3%	41.3%	52.2%	41.1%	41.5%	34.6%	55%	42.3%
Tables	20.8%	**41.3%**	**52.2%**	**41.5%**	**43.1%**	34.6%	**55.7%**	**43.2%**
Dotted Chart	5%	0%	0%	6.3%	0.3%	0%	7.1%	2.1%
Process Model	60.4%	34.8%	21.2%	8%	15.6%	12.8%	9.2%	20.1%
Social Network	1%	0.6%	0.7%	19.6%	4.7%	0%	0%	4.3%
Trace Alignment	2%	1.9%	0%	0%	0.2%	0%	0%	0.4%
Process Mining	**68.3%**	37.4%	21.9%	33.9%	20.9%	12.8%	16.3%	27%

To obtain first insights into the analysis process, we next investigated the use of visualization techniques with respect to each domain problem. We focused on the techniques, as we distinguished between general-purpose techniques, tables and those specific to process mining: dotted charts, process models, social networks, and trace alignments. Thus, the techniques provide a rough estimation for the application of process mining-specific analysis techniques. Note however that the general-purpose techniques might display event data attributes and analysis targets that were obtained from the application of process mining techniques. For each combination of domain problem and visualization technique, we computed the absolute frequencies with regard to the information needs, and normalized the frequencies with respect to the overall number of information needs per domain problem. Table 2 summarizes the results.

The process mining-specific techniques and especially the process models are the most important means for discovery, providing experts with important insights into the control-flow perspective. However, with regard to the other domain problems these techniques are less important. Indeed, process models are used across all problems and satisfy 17.4% of the information needs on average. Moreover, social networks play a key role for the organizational perspective. Yet, the majority of information is represented using general-purpose techniques and tables. Especially tables, as a flexible visualization technique suited for displaying high-dimensional data, are used very frequently and cover 41.6% of all informa-



(a) Per report (b) Per section

Fig. 3: Information needs in total and distinct information needs

tion needs on average across all problems. The general-purpose techniques are applied to 28.6% of the information needs on average, with bar and line charts being the most widely adopted techniques.

The interpretation of these results must be treated with care, as they are insensitive to cases were general-purpose techniques and tables summarize the results of process mining analysis techniques. Nevertheless, the widespread use of general-purpose techniques and tables does indicate a lack of standardized approaches at the domain problem level. That is, while there are invaluable techniques that address issues at the level of information needs, there is limited support for analysts in orchestrating these techniques to understand specific domain problems. For example, discovering process models from logs is indispensable for understanding the control flow; however *discovery at the problem level* is addressed with a broader spectrum of representations than process models.

Lastly, we assessed the diversity of the analysts' information needs. To this end, we conducted the following analysis once for each report and once for each section. First, for a given section or report, we counted the information needs contained in it. Among those information needs we also determined the number of distinct information needs, i.e., where the annotations for visualization techniques, event data attributes, and analysis targets are identical. Fig. 3 outlines the results. The grey line in the figure marks the equality between both measures, i.e., dots on the line are reports (a) or sections (b) where each information need is unique. The trend in the figure shows that the analysts tend to reuse certain types of visual representations. There are two possible explanations for this observation. First, analysts might be interested in certain aspects and re-apply the same technique to analyze different snapshots of the data. Here, they might benefit from dashboard-like tools, enabling them to configure views that can dynamically be updated with different subsets of the data. Second, analysts might be familiar with only a few analysis techniques. In this case, advanced guidance approaches might help analysts to explore data from various perspectives. Yet, in order to arrive at a final conclusion further experimentation is warranted.

4.2 Details for Frequent Domain Problems

So far, we have looked at the importance of domain problems and general work practices. We now focus on the analysis of specific domain problems and the

mining practices associated with them. In particular, we identify and describe frequent information needs. The explication of these needs constitutes important input for assessing and designing analysis techniques. In this regard, we focus on the two most frequent domain problems. First, we examine how analysts familiarize themselves with the data. Here, we also consider discovery problems, as our analysis revealed that discovery is often linked to the familiarization problem. Second, we focus on the case perspective as the most frequent problem.

Familiarization & Discovery. A first result stems directly from our annotation process, during which we inductively developed the codes describing the event data attributes. At the level of technique development the data model that is generally applied is a logical data model comprising *log*, *trace*, and *event* entities, relationships between them as well as a set of continuous and discrete attributes describing the entities. While this level of abstraction ensures that the developed techniques are reusable, it is also free of semantics. Yet, analysts typically view the data from the conceptual standpoint and think about the data in terms of entities including activities, organizational entities, and items, as well as relationships between them including responsibilities, work practices, or the control flow dependencies. With regard to the development of analysis tools, it might thus be valuable to enable analysts to map the physical data model to a conceptual model and to conduct the analysis based on the conceptual model. Moreover, entities and attributes in this data model might be the result of a specific analysis, e.g., a social network visualization might be used to identify groups of resources within the hierarchy whose performance is later on investigated as well. Thus, tools could also support analysts in incorporating analytical results into the domain model.

To identify analysis patterns specific to familiarization and discovery, we extracted frequent pairs of annotated codes from the information needs associated with these two problems. We only considered pairs and codes that occurred in at least 5% of the information needs. Fig. 4 summarizes these pairs using a parallel

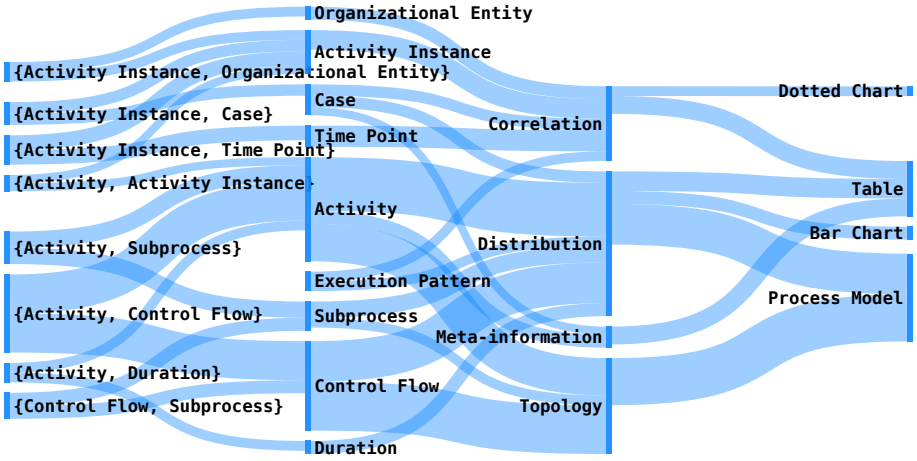


Fig. 4: Frequent analysis patterns related to familiarization and discovery

sets visualization. In this visualization there are four columns of nodes. Starting from the left, sets of event data attributes are depicted in the first column, event data attributes in the second, analysis targets in the third, and visualization techniques in the last. An edge depicts the frequency of a code pair or, in case of the sets of event data attributes, the frequency of attribute containment. Note that the size of the nodes is also proportional to the frequencies of the codes.

The figure shows four main types of analysis. First, process models are used to visualize the topology of the process or the control-flow, respectively. In this regard, the frequency of activities and their connections is displayed as well. Second, meta-information primarily regarding activity and case attributes is captured in tables. Third, the major category of information needs is related to understanding the distribution of cases, activities, execution patterns, and durations, and is visualized using bar charts, tables or other techniques. Fourth, analysts also investigate the correlation between a broad range of attributes including execution patterns, items, durations, time points and organizational entities. This type of information is displayed in tables, dotted charts or other types of general-purpose techniques. Additionally, Fig. 4 shows which data attributes were often examined in combination, e.g., activities and durations, activity instances and time points, etc.

Case Perspective. We repeated the above analysis for the case perspective and obtained the parallel sets visualization in Fig. 5. Here, we identified three main use cases. First, process models including the frequency of activities, their dependencies, or execution times are inspected. Process models are also used to identify execution patterns and to put them into context. Second, the distribution of subprocesses, activity instances, and execution patterns is represented using tables and various other types of general-purpose techniques. Finally, the third and main use case deals with examining the relationships between attributes. In this context, a large portion of information needs is linked to correlating exe-

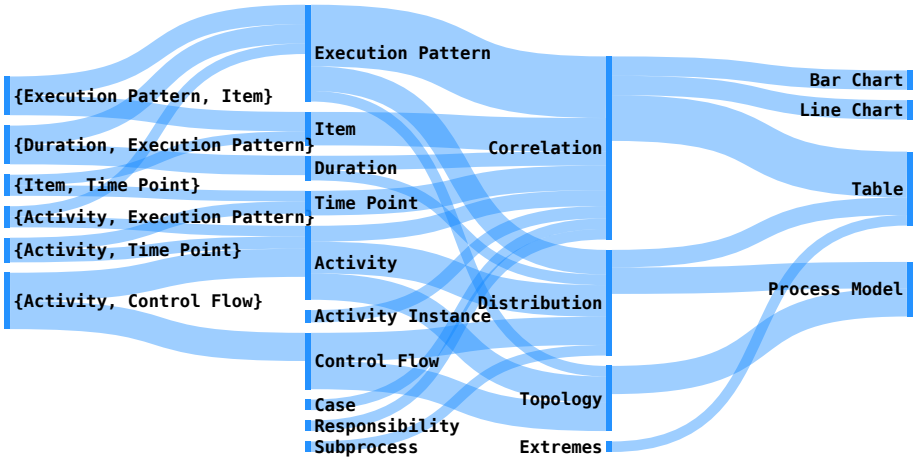


Fig. 5: Frequent analysis patterns related to the case perspective

cution patterns to items, durations, and responsibilities, amongst others. Here, bar charts, line charts, and tables are mainly utilized for visualization.

5 Related Work

There are two streams of research that are relevant to our study. First, there are analysis techniques and visual idioms which support analysts in the analysis of specific sub-questions. The development of visual idioms is subject to the field of *visual process analytics* and examples include the dotted chart which provides an overview of the events in an event log [18]; a technique to replay cases on top of process models [22]; or confusion matrices to compare process variants with respect to different perspectives [15]. The idioms often make use of *process mining* [21] techniques that extract knowledge from event logs, including, amongst others, the process' actual control flow (e.g., [2,9]) and its conformance to the intended behavior (e.g., [5,17]). In this paper, we focused on understanding how these techniques are applied in the context of process mining projects.

More relevant to our work are those works that focus on the work practices of analysts. On the one hand, there are methodologies for systematically approaching analysis projects, e.g., PM^2 [23], the L* life-cycle model [21], and the Process Diagnostics Method [3]. These methodologies comprise high-level processes including generic activities like data collection, data cleaning, and data analysis. Additionally, they provide anecdotal and exemplary evidence to outline their intended use. In contrast, we focus on explicating and analyzing the actual work practices based on empirical data. In this context, there are a few empirical studies that provide insights into the work practices. This includes catalogs of business process management [20] and process mining use case [1]. Additionally, Martens and Verheul [10] categorized the techniques applied in the first four editions of the BPIC. Yet, these studies focus on the categorization of problems or techniques, but do not provide details insights into their relationship.

6 Findings & Recommendations

In this work, we presented a systematic study in which we examined the work practices in process mining projects based on reports that resulted from these projects. In our study, we observed that the most frequently examined problems are those referring to the analysis of perspectives other than the control-flow perspective, especially the case perspective. In this regard, our analysis revealed that the problems are largely explored via visualization techniques *not* specific to process mining, pointing to areas that might benefit more sophisticated analytical support. Additionally, the data revealed that discovery is a domain problem that organizations need to explore. Moreover, discovery is also often analyzed as part of the familiarization with the data in order to establish a basic understanding of the underlying process. Finally, we noticed that analysts rely on similar sets of visual representations when addressing different information needs. This indicates that analysts apply a work practice of defining an analysis technique

and re-applying it to different data snapshots. We also presented a set of eight work practice patterns that can guide the development of advanced tools.

In future work, it would be interesting to extend the investigation of work practices by assessing the usefulness of a visual representation in the overall analysis process, as well as its contribution towards actually answering a domain question. Doing so would require interviews with analysts and business stakeholders as well as observations in laboratory settings; relying on the reports for these purposes would be too speculative.

References

1. Ailenei, I., Rozinat, A., Eckert, A., van der Aalst, W.: Definition and validation of process mining use cases. In: BPM Workshops. pp. 75–86 (2012)
2. Augusto, A., Conforti, R., Dumas, M., La Rosa, M.: Split miner: Discovering accurate and simple business process models from event logs. In: ICDM. pp. 1–10 (2017)
3. Bozkaya, M., Gabriels, J., van der Werf, J.: Process diagnostics: a method based on process mining. In: eKNOW. pp. 22–27 (2009)
4. Dumas, M., La Rosa, M., Mendling, J., Reijers, H.: Fundamentals of Business Process Management. Springer, Heidelberg (2013)
5. García-Bañuelos, L., van Beest, N., Dumas, M., La Rosa, M., Mertens, W.: Complete and interpretable conformance checking of business processes. IEEE Trans. Softw. Eng (2017)
6. Isenberg, P., Zuk, T., Collins, C., Carpendale, S.: Grounded evaluation of information visualizations. In: Workshop on Beyond Time and Errors: Novel Evaluation Methods for Information Visualization. pp. 6:1–6:8 (2008)
7. Keim, D., Andrienko, G., Fekete, J., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. In: Information Visualization: Human-Centered Issues and Perspectives, pp. 154–175. Springer, Berlin (2008)
8. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology (second edition). Sage Publications, Thousand Oaks, CA, USA (2004)
9. Leemans, S., Fahland, D., van der Aalst, W.: Discovering block-structured process models from event logs - a constructive approach. In: Petri Nets. pp. 311–329 (2013)
10. Martens, J., Verheul, P.: Social performance review of 5 dutch municipalities: Future fit cases for outsourcing? In: BPI (2015)
11. Mayring, P.: Qualitative content analysis. Forum Qualitative Social Research **1**(2) (2000)
12. Meyer, M., Sedlmair, M., Munzner, T.: The four-level nested model revisited: Blocks and guidelines. In: BELIV. pp. 11:1–11:6 (2012)
13. Munzner, T.: A nested model for visualization design and validation. IEEE Transactions on Visualization and Computer Graphics **15**(6), 921–928 (2009)
14. Munzner, T.: Visualization Analysis and Design. CRC Press, Boca Raton, FL, USA (2014)
15. Nguyen, H., Dumas, M., La Rosa, M., ter Hofstede, A.: Multi-perspective comparison of business process variants based on event logs. In: International Conference on Conceptual Modeling. pp. 449–459 (2018)
16. Recker, J.: Scientific Research in Information Systems: A Beginner's Guide. Springer, Berlin, Germany (2013)

17. Rozinat, A., van der Aalst, W.: Conformance checking of processes based on monitoring real behavior. Inf. Syst **33**(1), 64–95 (2008)
18. Song, M., van der Aalst, W.: Supporting process mining by showing events at a glance. In: WITS'07. pp. 139–145 (2007)
19. Spence, R.: Information Visualization – An Introduction. Springer, Switzerland (2014)
20. van der Aalst, W.: Business process management: a comprehensive survey. ISRN Software Engineering (2013)
21. van der Aalst, W.: Process Mining: Data Science in Action. Springer, Berlin (2016)
22. van der Aalst, W., de Leoni, M., ter Hofstede, A.: Process mining and visual analytics: breathing life into business process models. BPM reports, BPMcenter.org (2011)
23. van Eck, M., Lu, X., Leemans, S., van der Aalst, W.: PM^2: A process mining project methodology. In: CAISE. pp. 297–313 (2015)