Event Log Extraction: How to Minimize the Effort of the Human-in-the-Loop? (Extended Abstract)

Vinicius Stein Dani
Department of Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands
v.steindani@uu.nl

***Abstract*—To conduct process mining an event log is required. However, extracting event logs is often time-consuming, especially when the respective IT systems do not store events in a process-oriented way. The main reason is that tasks have to be performed manually, such as identifying entries in a transactional database that relate to the activities of the considered process. In this PhD project, we address this problem of the human-in-the-loop during event log extraction. Our main goal is to minimize, through automation, the manual effort involved in the extraction of event logs.**

***Index Terms*—process mining, event log extraction, automation, human-in-the-loop**

I. Introduction

Process mining is widely used in a plethora of fields including healthcare, banking, and production [1], [2]. The general applicability and the value of process mining in these fields has been demonstrated in the context of various case studies [3]. Process mining requires an event log [4], which is not always readily available in practice [5]. An event log is composed of at least three different attributes related to the execution of the considered process: (i) a case identifier, to identify the instances of the process; (ii) an activity identifier, to represent the activities pertaining to the process; and, (iii) a timestamp, to represent when the activity was executed [4].

Different approaches exist to extract event logs and partially reduce the manual effort associated with this extraction [6]–[8]. However, to the best of our knowledge, there is no event log extraction approach available that fully automates this task. Besides, in current process mining methods the data extraction stage is described minimally [9], [10]. Therefore, the goal of this PhD project is to minimize the manual efforts during the event log extraction through automation.

In the remainder of this work, Section II presents the motivation for our research. Section III elaborates on our research goal and questions. Section IV discusses how we intend to address our research goal. Finally, in Section V, we report on the current stage of this PhD project.

II. Motivation

Event log extraction is time-consuming and requires a significant amount of domain knowledge [8]. It is a costly part of any process mining project, draining resources that could have been used during other stages. Several approaches address the problem of automating event log extraction, with their own assumptions and requirements. Some assume the existence of redo-logs [6], while others require the interaction of a domain expert [7]. A recent approach [8] proposes to link additional information to the event log and, in this way, integrate the process and data perspective. However, all these approaches are far from being fully automated.

Based on the findings from an exploratory literature review we conducted, we identified that one recurring time-consuming stage of a process mining project is related to tasks that require domain knowledge and a domain expert interaction [1]–[3], [5]–[8], [11]. Often, a domain-expert is required to point relations between, for example, the process model that is to be analyzed and the data model (or data schema) that represents the data that is to be used as a source for the event log extraction. This so-called correspondence definition is, among others, a manual preparation effort that may be feasible to address through automation.

III. Research Goal and Research Questions

Considering this, we address the problem of the event log manual preparation effort. Our main goal is the following:

- To minimize, through automation, the manual preparation steps that are associated with the event log extraction phase from an operational database.

With this purpose in mind, we consider answering the following research questions as the basis of our project:

- *What are the techniques currently being used to extract event logs?* The objective here is to identify existing techniques, their inputs, outputs, assumptions, the extent to which these are automated, and the necessary steps for extracting an event log;
- *What are the manual tasks performed during the event log extraction?* By answering this, we aim to acquire a clear view of which manual tasks are required during the event log extraction;
- *Can such tasks be object of automation?* With this, we expect to build a list of tasks that are feasible to automate, which can be used to prioritize and focus our efforts.

Netherlands Organisation for Scientific Research.

IV. Research Method

To solve our goal, we will conduct this project based on an iterative, seven-step approach, where we will: (i) gather from the literature - and practitioners' experience - a set of currently used techniques to extract event logs; (ii) identify the tasks which still demand a human-in-the-loop; (iii) map the findings from the literature and practitioners' perspectives to see where they overlap or differ; (iv) build a list of feasible tasks to automate; (v) validate this list with experts; (vi) prioritize, also with experts, the tasks to be tackled first; (vii) build a script to automate the task with higher priority in the list; (vii) validate the automation with experts.

To address our research questions, as well as steps (i) and (ii) specifically, we will perform a systematic literature review [12] and a survey [13]. We want to gain a two-sided perspective on the event log extraction task: both from the literature and from practitioners. Hence, we will bring these findings together to see if the tasks that the practitioners mention they perform during the event log extraction are related to the ones found in the literature. Thereafter, we will advance to the next steps of this approach towards our main goal.

V. Current Stage of this Research Project

This project is still in an early phase. So far, we manually extracted an event log to better understand, from a practitioner's perspective, how such a task is performed. We used sample data from a SAP P2P process provided by an industry partner.

Next, we conducted an exploratory literature review to build up initial knowledge regarding existing approaches for the event log extraction and its preparation efforts. For instance, current tools such as ProMimport [14] and XESame [15] help users to integrate event log data from heterogeneous sources, although they are not able to automatically identify the relevant database tables. From another side, in [16], the relevant database tables may be identified and related to the process model; however, new adapters need to be implemented for every different data source format and structure. To approach this problem from another perspective may be only one of the building blocks for minimizing the human-in-the-loop effort in the event log extraction task. And, to achieve this, some scientific challenges need to be addressed:

- Partial alignment: the relationship between the tables of a database and a process model is partial, i.e., only a small number of tables relate to process model activities. Mechanisms to differentiate between relevant and irrelevant tables are required;
- Naming differences: process models and operational databases often use completely different, or even cryptic, names to refer to similar entities. This means that identifying relevant tables cannot exclusively rely on textual similarity measures.

Considering these challenges, one possible direction is to work on a technique to automate the identification of correspondences between a process model and a database schema to identify the database tables that contain information relevant to the process execution. Such identification represents an alignment problem. Acknowledging this, we intend to characterize the relations between a process model, a database schema, and its textual descriptions; and, afterward, to compute the most likely alignment. Based on this, we may use a Markov Logic formalization to specify how an optimal alignment can be obtained. Based on the alignment generated by our technique, the extraction of the event log could be achieved.

Finally, we will prepare a research paper to report on the exploratory literature review and the manual event log extraction we performed. We will discuss what could be done differently and automatically as to minimize the manual efforts during the event log extraction task.

Acknowledgment

Part of this research was funded by NWO (Netherlands Organisation for Scientific Research) project number 16672.

References

[1] C. d. S. Garcia, A. Meincheim, E. R. Faria Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, and E. E. Scalabrin, "Process mining techniques and applications – A systematic mapping study," *Expert Systems with Applications*, vol. 133, pp. 260–295, 2019.

[2] A. Corallo, M. Lazoi, and F. Striani, "Process mining and industrial applications: A systematic literature review," *Knowledge and Process Management*, no. January, pp. 1–9, 2020.

[3] G. B. Pereira, E. A. P. Santos, and M. M. C. Maceno, "Process mining project methodology in healthcare: a case study in a tertiary hospital," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 9, no. 1, pp. 1–14, 2020.

[4] W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*, 2011.

[5] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," in *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, 2020.

[6] W. M. P. van der Aalst, "Extracting Event Data from Databases to Unleash Process Mining," pp. 105–128, 2015.

[7] M. Jans, "From relational database to valuable event logs for process mining purposes : a procedure," Tech. Rep. January, 2017.

[8] E. González López de Murillas, H. A. Reijers, and W. M. van der Aalst, "Connecting databases with process mining: a meta model and toolset," *Software and Systems Modeling*, vol. 18, no. 2, pp. 1209–1247, 2019.

[9] M. Bozkaya, J. Gabriels, and J. M. Van Der Werf, "Process diagnostics: A method based on process mining," *Proceedings - International Conference on Information, Process, and Knowledge Management, eKNOW 2009*, no. February 2009, pp. 22–27, 2009.

[10] M. L. Van Eck, X. Lu, S. J. J. Leemans, and W. M. P. van der Aalst, "PM2: A Process Mining Project Methodology," 2015.

[11] C. Rodríguez, R. Engel, G. Kostoska, F. Daniel, F. Casati, and M. Aimar, "Eventifier: Extracting process execution logs from operational databases," in *CEUR Workshop Proceedings*, vol. 936, 2012, pp. 17–22.

[12] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University Joint Report, Tech. Rep. EBSE 2007-001, 2007.

[13] A. Bryman, "Social research methods," *Journal of Chemical Information and Modeling*, vol. 53, no. 9, pp. 1689–1699, 2013.

[14] C. W. Günther and W. M. Van Der Aalst, "A generic import framework for process event logs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4103 LNCS, 2006, pp. 81–92.

[15] H. M. Verbeek, J. C. Buijs, B. F. Van Dongen, and W. M. Van Der Aalst, "XES, XESame, and ProM 6," in *Lecture Notes in Business Information Processing*, vol. 72 LNBIP, 2011, pp. 60–75.

[16] E. González López De Murillas, "Extracting Event Data from Real-Life Data Sources Process Mining on Databases," Tech. Rep., 2019.