User-Friendly Data Extraction and Event Log Building for Process Mining (Extended Abstract)

Shameer K. Pradhan[1,*]

[1]Hasselt University, Agoralaan Building D, 3590 Diepenbeek, Belgium

Abstract

Data extraction and event log building are crucial steps in process mining. To effectively utilize process mining algorithms, it is necessary to have process data available in a suitable event log format. However, the current process of extracting data and building event logs demands considerable time and effort. The objective of this Ph.D. research is to improve the support for process mining practitioners in extracting data from information systems and building event logs from the extracted data. Furthermore, we would like to facilitate interactive support with the minimum amount of input from the perspective of the process mining expert.

Keywords

Process mining, Data extraction, Event log building, Data preparation, Relational database

1. Introduction and Motivation

Data extraction and event log building from information systems has been recognized as a crucial step during the pre-analysis stage among the various stages of process mining, [1]. Those steps account for approximately 80% of the time invested in process mining [2]. Consequently, reducing the time and effort required for these activities would greatly benefit process mining practitioners.

An approach to build event logs from data stored in the SAP system has been devised by Berti et al. [3]. A *graph of relations* is created, representing the tables and relationships. However, designating the central and associated tables is still manual. Similarly, a framework, which includes relevant steps such as identifying requirements and constructing logs, has been developed to build event logs semi-automatically [4]. The approach also includes manual steps that need to be performed by different stakeholders. Calvanese et al. [5] devised an ontology-based data extraction approach called *onprom*. Similar to previous approaches, this method requires a series of manual steps to create a conceptual schema, map specification, and annotate the conceptual schema. Manual tasks are time-consuming and prone to errors.

This Ph.D. research will first aim to understand the current landscape of data extraction and event log building for process mining and identify knowledge gaps. Afterward, we will create methods to enhance the support for interactive data extraction and event log building

ICPM'23: Doctoral consortium, October 23–27, 2023, Rome, Italy

*Corresponding author.

shameer.pradhan@uhasselt.be (S. K. Pradhan)

0000-0001-8969-8772 (S. K. Pradhan)

for process mining, thus helping to reduce the time and effort required. We would incorporate the human-in-the-loop concept, wherein we involve the process experts, the system experts, and process mining experts to fine-tune our solution. Primarily, we will be focusing on the following challenges:

- **C1:** There is a lack of a comprehensive understanding of the activities that may need to be performed on raw process data stored in information system databases to make it suitable for process mining algorithms.
- **C2:** There is a lack of a comprehensive understanding of the challenges experienced by stakeholders with different expertise (process experts, system experts, and process mining experts) during data extraction and event log building for process mining.
- **C3:** Process experts understand the activities that are performed in a business process. System experts know the structure of the database. Process mining experts have the skills to perform process mining analysis. However, there needs to be more alignment in these stakeholders' knowledge. The current interaction between the stakeholders is a labor- and time-intensive process prone to errors and miscommunication.

2. Research Questions

In order to guide the Ph.D. research, we have formulated the following research questions.

- **RQ1:** What activities need to be performed during the pre-analysis stage of process mining?
- **RQ2:** What challenges exist in data extraction and event log building for process mining for different stakeholders, and how critical are these challenges?
- **RQ3:** How can (a subset of) the identified challenges be solved to enhance the support for labor- and time-intensive interactive data extraction and event log-building processes?

RQ1 allows us to address C1 by helping us understand the current landscape vis-à-vis the activities performed during the pre-analysis stage of process mining. RQ2 helps us identify the challenges in data extraction and event log-building steps in process mining, thus addressing C2. RQ2 also allows us to rank the identified challenges by their criticality. Finally, RQ3 guides our quest to devise solutions for the identified challenges. If more automated solutions for the identified challenges can be devised, the time and labor needs of the pre-analysis stage could be reduced, thus enabling us to address C3.

3. Research Methodology and Project Roadmap

We use the design science (DS) research process by Peffers et al. [6] as the overarching methodology in this Ph.D. research. DS research includes stages such as problem identification, solution objective definition, solution development, demonstration, evaluation, and communication [6], which is suitable for creating and evaluating novel solutions and artifacts to address problems. With DS as the primary methodology, we plan to conduct specific research projects to answer the research questions.

To answer RQ1, we have already conducted a systematic literature review (SLR) [1] focusing on the pre-analysis stage of process mining. The SLR has helped us outline the current state-of-the-art pre-analysis stage of process mining. For instance, through the SLR, we have identified fifteen activities that can be performed on raw process data in the pre-analysis stage, such as data extraction, event log building, abstraction, log cleaning, and log merging.

We plan to answer RQ2 by conducting an exploratory case study to understand the intersection between the knowledge of different stakeholders, the vocabulary they use for their work, and how their work is passed on to another expert. Ideally, we will conduct the study in multiple organizations with different contexts. We also strive to identify the challenges stakeholders face with different expertise and responsibilities in data extraction and event log building. We will be interviewing various stakeholders with roles such as the process expert, the system expert, and the process mining expert.

Although we will further refine the scope of RQ3 after answering RQ2, we have explored some avenues for making data extraction and event log building more user-friendly. One such avenue is the application of natural language processing (NLP). Specifically, NLP can be applied to the data dictionaries of information systems to map the knowledge between process experts and system experts. For example, it could be used to identify a collection of tables for a specific document used in a process. NLP has already been employed in process mining for process querying during the analysis stage [7] [8] and for constructing event logs from unstructured text data [9] [10]. However, NLP has not been applied to extract data from information systems, such as SAP, which employs relational databases and builds event logs from the extracted data.

During artifact development, which answers RQ3, we will strive to ensure compatibility between the methods developed in this Ph.D. research and existing artifacts in data extraction and event log building. For instance, we plan to facilitate automation of certain parts of *OnProm*, a tool designed for ontology-based data extraction from relational databases [11, 12]. A critical step in the operation of OnProm is the annotation of activities by the process mining expert on a UML diagram that represents the underlying structure of the information system's database. However, for the process mining expert to accurately annotate the activities, they must possess prior knowledge of the activities stored in the process data. This necessary information can be provided to the process mining expert by a process expert, who, in this case, must also be well-versed in the structure of the underlying database by utilizing artifacts like *erprep* [13], a guided approach for stakeholder collaboration during data extraction. In this pipeline, we would like to provide automation to aspects of developing the erprep artifact and their mapping to the requirements of OnProm. Figure 1 shows the potential positioning of our research within the steps of the OnProm tool.

During the solution development phase of this Ph.D. research, we plan to use sample data based on the SAP ERP system. Similarly, to validate the developed artifact, we would be conducting user acceptance testing [14]. We will invite several process mining experts and process experts to test our proposed solution to extract data from their information systems. Likewise, we can request the users to build an event log with existing approaches and then build the event log with our technique. We can compare the two approaches' accuracy, required time, and complexity.

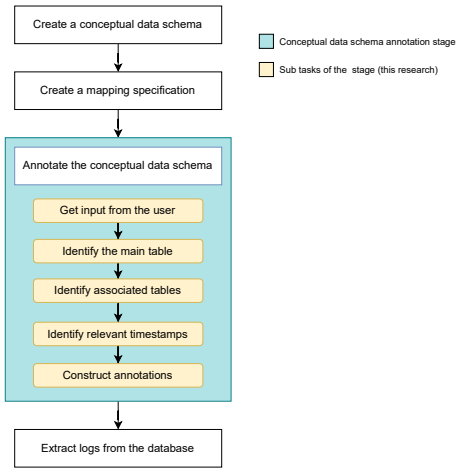

Figure 1: Our research positioned within the OnProm steps

4. Conclusion

Process mining is a set of methods and tools utilized in businesses with immense potential for revealing process-related insights. However, a significant obstacle to its widespread adoption lies in the pre-analysis stage's complex data extraction and event log-building steps. In an ideal scenario, the process mining expert, who would be the end user of the event log, could request an event log based on specific input criteria and receive the log relatively effortlessly. Our research aims to serve as a preliminary step toward achieving that goal by facilitating interactive support for the data extraction and event log-building steps.

Acknowledgments

This study was supported by the Special Research Fund (BOF) of Hasselt University under Grant No. BOF21OWB22, Belgium.

This Ph.D. thesis is supervised by Prof. dr. Mieke Jans (supervisor) and Prof. dr. Niels Martin (co-supervisor).

References

[1] S. K. Pradhan, M. Jans, N. Martin, Getting the data in shape for your process mining analysis: A review of the pre-analysis stage (under review).

[2] J. De Weerdt, M. T. Wynn, Foundations of process event data, in: Process Mining Handbook, Springer, 2022.

[3] A. Berti, G. Park, M. Rafiei, W. M. P. Van Der Aalst, An event data extraction approach from SAP ERP for process mining, in: Lecture Notes in Business Information Processing, Springer International Publishing, Cham, 2021, pp. 255–267. doi:10.1007/978-3-030-98581-3_19.

[4] J. D. Hernandez-Resendiz, E. Tello-Leal, U. M. Ramirez-Alcocer, B. A. Macías-Hernández, Semi-Automated Approach for Building Event Logs for Process Mining from Relational Database, Applied Sciences 12 (2022) 10832. doi:10.3390/app122110832.

[5] D. Calvanese, T. E. Kalayci, M. Montali, A. Santoso, OBDA for Log Extraction in Process Mining, in: G. Ianni, D. Lembo, L. Bertossi, W. Faber, B. Glimm, G. Gottlob, S. Staab (Eds.), Reasoning Web. Semantic Interoperability on the Web: 13th International Summer School 2017, London, UK, July 7-11, 2017, Tutorial Lectures, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2017, pp. 292–345. doi:10.1007/978-3-319-61033-7_9.

[6] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A Design Science Research Methodology for Information Systems Research, Journal of Management Information Systems 24 (2007) 45–77. doi:10.2753/MIS0742-1222240302.

[7] L. Barbieri, E. Madeira, K. Stroeh, W. van der Aalst, A natural language querying interface for process mining, Journal of Intelligent Information Systems 61 (2023) 113–142. doi:10.1007/s10844-022-00759-9.

[8] H. Yeo, E. Khorasani, V. Sheinin, I. Manotas, N. P. An Vo, O. Popescu, P. Zerfos, Natural Language Interface for Process Mining Queries in Healthcare, in: 2022 IEEE International Conference on Big Data (Big Data), 2022, pp. 4443–4452. doi:10.1109/BigData55660.2022.10020685.

[9] J. C. d. A. Goncalves, F. M. Santoro, F. A. Baiao, Business process mining from group stories, in: 2009 13th International Conference on Computer Supported Cooperative Work in Design, 2009, pp. 161–166. doi:10.1109/CSCWD.2009.4968052.

[10] C. Kecht, A. Egger, W. Kratsch, M. Röglinger, Event Log Construction from Customer Service Conversations Using Natural Language Inference, in: 2021 3rd International Conference on Process Mining (ICPM), 2021, pp. 144–151. doi:10.1109/ICPM53251.2021.9576869.

[11] D. Calvanese, M. Montali, A. Syamsiyah, W. M. P. van der Aalst, Ontology-Driven Extraction of Event Logs from Relational Databases, in: M. Reichert, H. A. Reijers (Eds.), Business Process Management Workshops, Lecture Notes in Business Information Processing, Springer International Publishing, Cham, 2016, pp. 140–153. doi:10.1007/978-3-319-42887-1_12.

[12] D. Calvanese, T. E. Kalayci, M. Montali, S. Tinella, Ontology-Based Data Access for Extracting Event Logs from Legacy Data: The onprom Tool and Methodology, in: W. Abramowicz (Ed.), Business Information Systems, Lecture Notes in Business Information Processing, Springer International Publishing, Cham, 2017, pp. 220–236. doi:10.1007/978-3-319-59336-4_16.

[13] D. Calvanese, M. Jans, T. E. Kalayci, M. Montali, Extracting Event Data from Document-Driven Enterprise Systems, in: M. Indulska, I. Reinhartz-Berger, C. Cetina, O. Pastor (Eds.), Advanced Information Systems Engineering, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 193–209. doi:10.1007/978-3-031-34560-9_12.

[14] K. Ganesh, S. Mohapatra, S. P. Anbuudayasankar, P. Sivakumar, User Acceptance Test, in: K. Ganesh, S. Mohapatra, S. P. Anbuudayasankar, P. Sivakumar (Eds.), Enterprise Resource Planning: Fundamentals of Design and Implementation, Management for Professionals, Springer International Publishing, Cham, 2014, pp. 123–127. URL: https://doi.org/10.1007/978-3-319-05927-3_9. doi:10.1007/978-3-319-05927-3_9.