Process Mining on Distributed Time-Series Data (PhD Proposal)

Frederik Fonger[1]

[1]Kiel University, Group Process Analytics, Hermann-Rodewald-Str. 3, 24118 Kiel, Germany

Abstract

Process mining techniques are used for the discovery of process models from recorded events, to analyze the conformance of a specification derived from recorded events and a process model, and for predictive analytics. However, mostly the recorded events come from (business) data of IT systems and process mining techniques have been developed to process structured data that is at a high (business) level of abstraction. Plenty of scenarios exist with low-level data and where process mining could give valuable insights when analyzing these kinds of data. The purpose of this PhD proposal is to design a process mining pipeline for time-series data in distributed settings. The challenges for process mining in such a scenario are that usually no ground truth exist to learn and optimize against nor techniques exist to efficiently process the high volume of time-series data, which is typical in such scenarios. To address these challenges, we suggest a process analytics pipeline relying on the generation of synthetic data and data sampling.

Keywords

PhD Proposal: Process Mining, Time Series, Data Sampling

1. Introduction

Process mining techniques are used for the discovery of process models from recorded events, to analyze the conformance of a specification derived from recorded events and a process model, and for predictive analytics. However, mostly the recorded events come from (business) data of IT systems and process mining techniques have been developed to process structured data that is at a high (business) level of abstraction. In plenty of scenarios (e.g., IoT settings) and disciplines (e.g., natural or life sciences) low-level data is produced. One example for the analysis of sensor data is in oceanography. With the rising importance of climate change, seaweed has been identified as a large natural carbon storage possibility. By analysing time series data, we can better understand the growth process of seaweed.

Process mining could give valuable insights when being capable of analyzing these kinds of data [1]. For instance, time-series data is recorded in sensors and the data is analyzed with the purpose to identify trends and patterns over time and make forecasts. The analysis of time-series data could benefit from process mining. Its combination would allow to discover a process and to analyze causal effects through the simulation of the process. However, time-series data

13th International Workshop on Enterprise Modeling and Information Systems Architectures (EMISA), May 11–12, 2023, Stockholm, Sweden

ffo@informatik.uni-kiel.de (F. Fonger)
0009-0000-8445-8104 (F. Fonger)

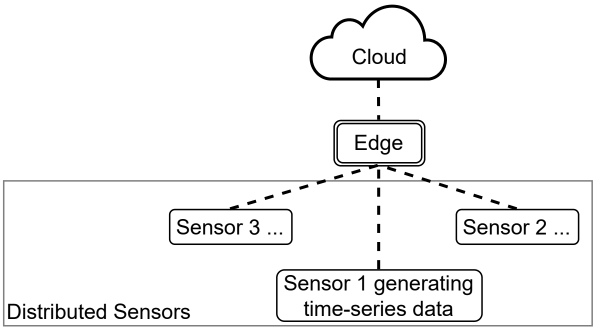


Figure 1: Setting with distributed sensors and central processing units like Edge and Cloud.

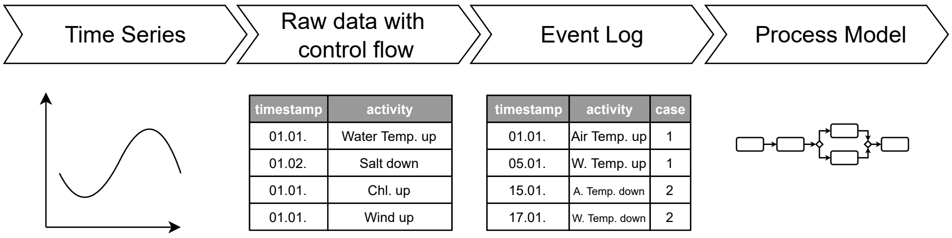
is unstructured data and at a low level of abstraction, while process mining techniques were developed for structured data that is at a high (business) level of abstraction. Therefore, time series data needs to be processed appropriately to extract an event log.
Additionally, in IoT scenarios time-series data is generated distributed and mostly in a large volume. This calls for an approach to efficiently process distributed time-series data for process mining. To address this, we present a process analytics pipeline relying on the generation of synthetic data and data sampling. Sampling allows to reduce the data size while preserving the information within the data. The challenge of data sampling is to identify a representative sample of the original dataset while reducing the volume of data.

The PhD proposal is structured as follows. The next section describes the research problem. Section 3 summarizes a solution and challenges that have to be addressed. Related works are discussed in Section 4, while the paper concludes with a summary.

2. Research Problem

The aim of the PhD proposal is to discover process models from distributed time-series data. To achieve this, a pipeline to efficiently process and analyze the data has to be designed. As mentioned above, distributed time-series data is generated in high volume, mostly with inappropriate data quality and in scenarios with low latency. Therefore, we suggest a pipeline relying on the generation of synthetic data and data sampling, while satisfying the requirements in distributed settings. The synthetic data is generated on a low level by simulating the individual sensors. The primary objective is to generate data for the development that is similar in characteristics like the format, time intervals and synthetic dependencies with noise. The synthetic dependencies are only for validating the developing methods and do not have to mirror the real data. Real data will be used on the developed methods at a later point in time. Fig. 1 shows a scenario with distributed sensors that a centrally processed at e.g., the edge and cloud.

Existing process mining algorithms are not capable to process time-series data from such scenarios. Activity recognition in time series is still an unsolved problem in time series analysis [1]. In this respect, the mapping of activities onto a time series is still a challenge. In this PhD proposal we will in particular focus on the volume and data quality aspect of time-series data



timestamp	activity
01.01.	Water Temp. up
01.02.	Salt down
01.01.	Chl. up
01.01.	Wind up

timestamp	activity	case
01.01.	Air Temp. up	1
05.01.	W. Temp. up	1
15.01.	A. Temp. down	2
17.01.	W. Temp. down	2

Figure 2: Process analytics pipeline for time-series data for process mining

for process mining. This means that we also have to focus on the data perspective in distributed settings to deal with the data quality aspect. Particularly, we suggest to generate synthetic time-series data in order to bridge the gap. For this, we developed a structured approach to map time-series data on control-flow patterns that we annotated for our purpose. Based on the simulation of the patterns it is possible to generate synthetic data in varying quality, which is again a crucial step for accurate results from machine learning techniques [2].

Fig. 2 shows the process analytics pipeline for distributed time-series data on an abstract level, which we plan to implement in the PhD project. First, we map time-series data on control-flow patterns in order to increase data quality through synthetic data[1]. Next, we plan to apply sampling strategies and to abstract activities from the data and finally to use process mining. When generating an event log, the large uncertainty between measuring points in some datasets is also essential to be considered.

3. Activities of the PhD proposal

To provide a solution several challenges have to be addressed: first, data sampling is necessary to reduce the data size. Then, activity recognition for time-series is essential, too. To address this, we plan to generate synthetic data with our tool, insert different levels of noise and then to design a technique for activity abstraction. Depending on the use case, there may be long time intervals between measurement points in the time series. This leads to uncertainty between measurement points, which must also be taken into account. Finally, we have to evaluate our techniques of mapping time-series data on process patterns. Particularly, we have to evaluate and to quantify the occurrence of the patterns in real time-series data.

Table 1 shows an example of the time series data used in the approach. The real and the synthetic data have the same data structure. Depending on the dataset, there are one or more entries for each timestamp. In this example, the air temperature and the salinity are measured weekly.

The next steps to implement the pipeline are improving the generating of synthetic data, testing different approaches to identify activities in time-series data and designing a method for dealing with uncertainty within the time-series data. Furthermore, a visualisation for the synthetic time-series data generation will be implemented. We plan to evaluate different data quality levels (i.e., less or more noise) to provide a solution for activity recognition. Also, we

[1] This step is already completed. We refer to [2].

Timestamp	Air Temperature	Salinity	...
2020-03-18	6.507	32.363	...
2020-03-25	7.484	32.313	...
2020-04-01	8.461	32.263	...
2020-04-08	9.291	32.251	...

Table 1
Example of time-series data from the marine use case.

will evaluate the sampling methods for two different scenarios. This research is relevant for uncovering phenomena and underlying processes in natural and life sciences. Ziolkowski et al. have already shown a first application of process mining on times series for data from oceanography [3]. My work will continue in the same direction and build on it.

This PhD proposal is conducted within the Marispace-X project, which in this way presents the scenario and real-time data. The purpose of the project is to develop a cloud-based platform to improve data exchange and efficient processing of maritime data. The technical cloud foundation relies on the GAIA-X framework. The data includes time-series data acquired from distributed sensors from underwater locations, alongside data from single sensors mounted on research vessels and stations.

4. Related Work

Herbert et al. proposed a methodology for generating synthetic time series data for process analytics [4]. However, this approach lacks the ability to specify the effects present in the data. To address this limitation, my proposed approach provides more flexibility in configuring the effects represented in the data, as well as the complexity level. An approach that uses process mining for time series data from smart products was presented by Eck et al. [5]. The approach applies human activity recognition on the data collected by the smart products and subsequently event logs are generated for process mining.

A challenge when using process mining methods can be the uncertainty in the data [6, 7]. Pegoraro et al. introduced a concept and a tool for not deleting and losing data, but rather incorporating the uncertainty into a resulting model [6, 7]. Another challenge stemming from the uncertainty in time series data is the presents of imprecise timestamps. The evaluation of such partially ordered events has already been addressed by Lu et al. [8]. Process mining for marine time series was introduced by Ziolkowski et al. by using a clustering algorithm for generating a event log from the time series [3]. Subsequently, a process mining algorithm was used to mine a process model.

5. Conclusion

This PhD proposal presents a pipeline for discovering process models from distributed time-series data. For the development and evaluation of novel methods, synthetic data is generated in varying complexity and noise intensity. Ultimately, this will be used for developing a method for mapping activities onto distributed time-series data. Furthermore, sampling approaches will be evaluated for two different scenarios within the pipeline. In the end, we want to be able to apply sampling and activity recognition on real marine time series data in order to generate event logs and subsequently use this for process mining.

6. Acknowledgement

I would like to express my gratitude to my PhD supervisor, Agnes Koschmider, for her invaluable guidance and support in writing my PhD proposal. I would also like to thank Milda Aleknonytė-Resch, for her insights and feedback on my PhD proposal. Furthermore, this project has received funding from the German Federal Ministry for Economic Affairs and Climate Action under the Marispace-X project grant no. 68GX21002E.

References

[1] A. Koschmider, N. Oppelt, M. Hundsdörfer, Confidence-driven communication of process mining on time series, Informatik Spektrum 45 (2022) 223–228.

[2] F. Fonger, M. Aleknonytė-Resch, A. Koschmider, Mapping time-series data on process patterns to generate synthetic data, in: CAiSE Workshops 2023, Lecture Notes in Business Information Processing, Springer, 2023. To appear.

[3] T. Ziolkowski, R. Schubert, M. Renz, A. Koschmider, Process Mining for Time Series Data, Technical Report, 2022. doi:10.1007/978-3-031-07475-2.

[4] T. Herbert, J. Mangler, S. Rinderle-Ma, Generating Reliable Process Event Streams and Time Series Data based on Neural Networks, volume 421, 2021, pp. 81–95. URL: http://arxiv.org/abs/2103.05462. doi:10.1007/978-3-030-79186-5_6, arXiv:2103.05462 [cs].

[5] M. L. van Eck, N. Sidorova, W. M. P. van der Aalst, Enabling process mining on sensor data from smart products, in: 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), IEEE, Grenoble, France, 2016, pp. 1–12. URL: http://ieeexplore.ieee.org/document/7549355/. doi:10.1109/RCIS.2016.7549355.

[6] M. Pegoraro, Probabilistic and non-deterministic event data in process mining: Embedding uncertainty in process analysis techniques (2022). arXiv:2205.04827.

[7] M. Pegoraro, M. S. Uysal, W. M. P. van der Aalst, PROVED: A Tool for Graph Representation and Analysis of Uncertain Event Data, in: D. Buchs, J. Carmona (Eds.), Application and Theory of Petri Nets and Concurrency, volume 12734, Springer International Publishing, Cham, 2021, pp. 476–486. URL: https://link.springer.com/10.1007/978-3-030-76983-3_24. doi:10.1007/978-3-030-76983-3_24, series Title: Lecture Notes in Computer Science.

[8] X. Lu, D. Fahland, W. M. P. van der Aalst, Conformance Checking Based on Partially Ordered Event Data, in: F. Fournier, J. Mendling (Eds.), Business Process Management Workshops, volume 202, Springer International Publishing, Cham, 2015, pp. 75–88. URL: http://link.springer.com/10.1007/978-3-319-15895-2_7. doi:10.1007/978-3-319-15895-2_7, series Title: Lecture Notes in Business Information Processing.