Theses of PhD dissertation

Frequent pattern and process mining-based process operation analysis

László Bántay

University of Pannonia
Chemical Engineering and
Material Sciences Doctoral School

Supervisor
János Abonyi DSc.

Department of Process Engineering
Veszprém
2025

1 Introduction and goals

This research aims to explore potential data and process science tools that support the analysis and modelling of process-related events in industrial control systems. The appearance of the Industry 4.0 approach has raised the importance of stored process data with the increasing need for deep analysis of processes. With a reach literature in the field of continuous data handling, the topic of discrete process event analysis and modelling with process mining techniques has room for improvement. The goal of my work was to collect the available techniques and develop new supporting methods to process industrial log files, with alarm management in the focus.

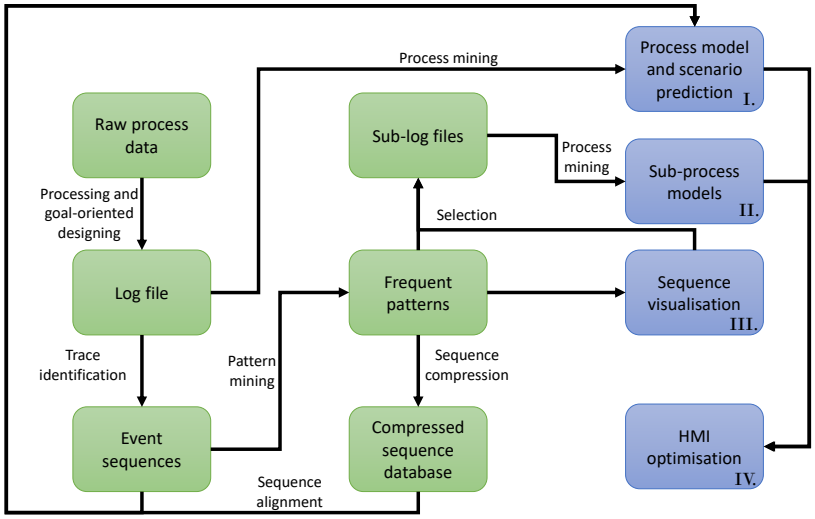

Figure 1: Roadmap of the dissertation. The input of process mining is the stored raw data, that has to be formatted in a standardized way. The standard format allows to gain process models and frequent event patterns as well, supporting an event scenario prediction method. Frequent event patterns also can be used to generate sub-process models, a network-like visualisation of the patterns supports the selection step. The mentioned techniques form a machine learning-based method to optimise human-machine interfaces.

The collection and storage of industrial data became easy and inexpensive at the beginning of the twenty-first century. Adding the exponential increase in computational speed resulted in a process-mining breakthrough. The input of process mining is a properly designed log file that contains all relevant information about the stored events. In addition to the mandatory event properties (event ID, timestamp, case ID), the log file can be enriched with specific data (resource, organisational role, cost, etc.) enabling targeted or hierarchical process mining resulting in specific ***process model***s. A properly designed log file has other benefits too, for example, frequent sequential event patterns can be gained to ***predict*** event ***scenarios***. Frequent event sequences can be stored in a database using sequence compression. If actual events are transformed into sequences, they can be compared with the database with the help of sequence alignment. Based on the similarity of the actual sequence to the stored ones, a probability event prediction can be made.

One potential issue with industrial log files may be that parallel processes are stored in the same log, resulting in biased and hard-to-understand process models. By assuming that the specific processes contain typical event groups, frequent pattern mining techniques can help to identify the sub-logs of the sub-processes, supporting the generation of ***sub-process models***. The challenge in this kind of log file partitioning is to choose the right frequent patterns. In the case of huge logs, an excessive number of patterns will be found that are hard to understand. There are several methods for sequential pattern post-processing; one of them is visualisation. By handling the frequent event sequences as elements of a network, an information-rich visualisation can be made. With the developed method, the pattern selection step in creating sublog files can be faster and easier.

In addition to the aspect of the targeted exploration of processes by creating sublog files, the gained models can be used to develop or ***optimise*** level two or three Human Machine Interface (***HMI***) displays. Typically, HMI displays are drawn based on the Piping and Instrumentation Diagrams (PIDs), which is good for level one displays, but on higher-level displays, a different approach may be necessary. To support the work of operators, that is, to develop human-centric HMI, process-aware layouts are one solution. The results of process mining combined with network theory can identify the groups of actors that are essential parts of a given operation, and shall be put on the same display.

2 Experimental tools and methods

During my research, based on a comprehensive literature overview, I have collected the existing data and process mining techniques and identified their shortcomings with regard to industrial process analysis tasks using distributed control system (DCS) data. To answer those shortcomings, new process analysis support tools were developed:

- a targeted event log transformation method to have log files in a standardised format,
- a sequence alignment-based event prediction method,
- a frequent pattern-based log file partitioning method,
- a similarity- and network theory-based sequence visualisation method,
- and a machine learning-supported HMI optimisation method.

The log data used were from two sources:

- DCS data from different parts of the oil refinery plant of MOL in Százhalombatta,
- and click data from a publicly available data repository.

The data processing and analysis steps were performed partially in ProM. To ensure that the developed methods can be used in a flexible way with respect to the source data, the whole framework was put into Python programming language. Python environment offers a wide range of solutions, log transformation to XES (eXtendable Event Stream) standard, the SPMF library for frequent pattern mining, the pm4py library for process mining, and other libraries for visualisation tasks (graphviz, networkx). The developed method and Python package can be used for other discrete data analysis projects, independently from the data source.

3 Theses

Thesis I.

I have proved that with goal-orientated trace identification and log file design, process mining is a suitable tool for analysing, discovering and predicting industrial processes. [1,2]

The exploration of industrial processes is to get information from related logged events. Related events in log files form traces. Trace identification is essential to explore industrial processes. However, it is not a trivial task, as usually the logged data is not labelled from this point of view. There are several methods to group events; one way is to define a minimum time interval between two events. Structuring and labelling the log file is also inevitable. Proper data labelling enables goal-oriented and hierarchical process mining, and every aspect of the processes can be interpreted. In alarm management the alarms, operator, and return to normal actions can be evaluated together and separately as well, showing different layers of the same process.

A well-designed log file allows to gain frequent sequential patterns containing useful information about process evolvement. If the most frequent event streams are collected in a compressed sequence database, comparing them with actual event sequences by using sequence alignment, the most probable event scenario can be predicted.

The developed trace identification and log file designing method was applied in the alarm management rationalisation project of an industrial hydrofluoric acid alkylation plant. I have identified the key events and generated different log files for different process discovery purposes. The resultant log files were analysed with process mining tools. The project demonstrated that with the help of process mining, alarm signals could be rationalised; therefore, the work of the operators will become safer, as well as more effective and, last but not least, the workload can be decreased. The sequence compression and alignment method has been examined and characterized using real-life data originating from a delayed coker, and its usability and limitations have been determined. The results show that the method has a very effective pattern mining capability that extended with the sequence alignment method can recognize an operational state just after a few typical alarms and match it with historical patterns in less than a second. High-confidence predictions could be obtained easily.

Thesis II.

I have developed a frequent sequential pattern-based log file partitioning method to gain specific models of subprocesses stored in the same log file. [3]

The suggested method of filtering out unnecessary events while simultaneously creating sub-logs combines frequent pattern mining and traditional process cube operations. Frequent itemset and sequential pattern mining were applied on the original log file, and the resultant itemsets and sequences were the basis of the log file partitioning. If the structure of the log file is proper (Thesis I.), a hierarchical partitioning is possible as well.

The method was applied to a log file of an industrial Hydrofluoric Acid Alkylation plant. The resultant process models in the case study were evaluated using the performance metrics introduced. The results proved that the method is effective in partitioning log files, regrouping events for targeted process discovery tasks, and handling the problem of parallel processes. The benefits, requirements, and limitations of the method were identified and are the following:

- Benefits
  - Sequential pattern mining and process mining use the same source. No extra preparation of the log-file is required.
  - The method is efficient from a computational demand point of view, as the size of the log-file to be processed has been significantly reduced.
  - Prior knowledge can be transferred to identify the relevant frequent patterns, facilitating iterative work. Process-relevant information can be efficiently included in process-mining.
  - The defined metrics enable an objective evaluation of the results. The evaluation step can facilitate the automation of the method.

- Requirements - Limitations
  - Although the method needs a pattern mining tool, this is not a critical issue as open-source tools are widely available.
  - A certain amount of knowledge is required to select the right pattern-mining algorithm. Suggestions are made regarding this topic.
  - Prior knowledge of the process is essential. Similarly to the *Knowledge Discovery Databases* (KDD) process-model, the iterative and interactive character of the method eases this issue.

Thesis III.

I have developed a network-based visualisation of frequent sequences (NBVFS) that supports quick understanding of event scenarios. [4]

The developed method transforms frequent event sequences to ego-networks that contain sequences having a common starting event, which is placed in the centre of the network. As every sequence may have different event expansions, the branches result in a tree structure. The key point in the method is how to place the nodes (the n-long sequences) within the network. I used the length of the sequences, the frequency (*support*) of the sequences, and the similarity between the different chains of events. Three visualisation methods were developed:

- **NBVFS-WN** (**N**etwork-**B**ased **V**isualization of **F**requent **S**equences - **W**eighted **N**etwork): this method uses a confidence-based adjacency matrix and the calculated metrics of the sequences. It results in a weighted network where those sequences are connected that are direct extensions of each other. The weight is the confidence value of the transition between the two connected sequences. This kind of visualisation helps to understand the conditions and consequences of occurring events.
- **NBVFS-CM** (**N**etwork-**B**ased **V**isualization of **F**requent **S**equences - **C**onfidence-based **M**DS projection): this method uses a similarity calculation, using transition confidence values for similarity measurement. The adjacency matrix must be enriched regarding the nondirectly connected sequences using transitive distance calculation. The positions of the nodes on the network are calculated with Multidimensional scaling, and the more similar the two sequences are, the closer they will be presented on the network. This kind of network representation contains more information about the relationship of sequences.
- **NBVFS-TM** (**N**etwork-**B**ased **V**isualization of **F**requent **S**equences - **T**ransaction-based **M**DS projection): the positions of the nodes are calculated with MDS identically to the NBVFS-CM method. The process is more or less the same, but the similarity calculation is based on the overlap of their common supporting transactions. This approach is closer to process mining, as transactions can be taken as traces and can provide feedback to the mining process.

This network-like visualisation of time-series-type event databases has many possible applications. The introduced method is a goal-oriented analysis tool to extract useful knowledge from the event database by visualising the event sequences. Due to its interactive character, it can successfully support iterative data analysis tasks. For example, it enables the adjustment of the frequent sequence pattern mining parameters. The proposed visualisation method allows for quick recognition of relevant event chains and key events with their most important attributes, such as support and confidence. This kind of information interpretation besides the support in cognitive information processing, can speed up the parameter identification step of machine learning model building, for example. It can also validate trace rules and other parameters used in process discovery tasks with respect to process mining. It has to be customised for the exact purpose of the task.

Thesis IV.

I have proposed a machine learning-based method to develop human-centric and process-aware HMI displays. [5]

The developed method aims to support the application of the *Navigation and Layout* principle of the ISA 101 standard, namely the *grouping related elements together* aspect. The principle of the iterative method developed is to use the combination of proper frequent pattern and process mining techniques. The proposed HMI displays and functionalities are based on the combined interpretation of P&I diagrams, process deviation models based on typical event/alarm sequences, and process control behaviour models based on typical operator responsive action sequences.

Alarm signals, operator actions, and display actions from a hydrofluoric acid alkylation plant were analysed. The log file was partitioned based on frequent sequential patterns of the stored events. Process control models were explored with process mining tools. A network-like visualisation of a typical process enabled a comparison of the number of workflows and the number of displays where the workflows are represented. A new, higher-level display was recommended based on the models gained and the existing structure of the HMI displays.

Even without a deep understanding of the system, the presented method proved effective. However, manual work is needed to evaluate the displays, resulting in subjective decisions. The method can be upgraded to define objective key performance indicators that support solving the minimalisation problem.

4 Future application of the results

Although the developed methods were successfully applied in alarm and process analysis tasks, independently from the data source there are additional application areas - for example, EDU-mining -, and potential future research directions are given.

By comparing the actual online process data with formerly defined event scenarios, a semi- or fully automated online HAZOP analysis tool can be built. The gained process patterns can be compared to a reference model that allows the evaluation of the work quality of the operators and supports the development of the Operator Training System (OTS).

Another potential future research direction is the implementation of machine learning in SCADA systems, which is an actual topic in the industry, as more and more solutions will be assisted by artificial intelligence. The proposed method is capable of processing acquired data and finding patterns among the event chains that lead to failures and the response actions of the operators. Using ML solutions enables the continuous dynamic and adaptive development of industrial process control systems.

The topic of trend charts can also be an interesting area. With the help of ML-supported solutions, process-variable dependencies can be gained from historical data, adding a predictive function to the system. For example, if the operator calls a trend chart of a specific process variable, an intelligent system can offer other trend charts, which may be useful in the actual control activity.

Publications

1. L. Bántay, G. Dörgő, F. Tandari and J. Abonyi, "Simultaneous Process Mining of Process Events and Operator Actions for Alarm Management", *Complexity, Frontiers in Data-Driven Methods for Understanding, Prediction, and Control of Complex Systems*, Vol. 2022, 2022.

2. L. Bántay and J. Abonyi, "Frequent pattern mining-based log file partition for process mining", *Engineering Applications of Artificial Intelligence*, Vol 123., Part A, 2023.

3. L. Bántay, N. Sas, G. Dörgő and J. Abonyi, "Sequence Compression and Alignment-Based Process Alarm Prediction", *Industrial & Engineering Chemistry Research*, Vol. 62(27), 10577-10586, 2023.

4. L. Bántay and J. Abonyi, "Machine Learning-Supported Designing of Human–Machine Interfaces", *Applied Sciences, New Insights into Human-Computer Interaction*, Vol. 14(4), 2024.

5. L. Bántay and J. Abonyi, "Network-based visualisation of frequent sequences", *PLOS ONE*, Vol. 19(5), 2024.