Filtering Toolkit: Interactively Filter Event Logs to Improve the Quality of Discovered Models

Mohammadreza Fani Sani[1], Alessandro Berti[1], Sebastiaan J. van Zelst[1,2], and Wil van der Aalst[1,2]

[1] Process and Data Science Chair, Lehrstuhl für Informatik 9 52074 Aachen, RWTH Aachen University, Germany

[2] Fraunhofer Gesellschaft, Institute for Applied Information Technology (FIT), Sankt Augustin, Germany

Abstract Process discovery algorithms discover process models on the basis of event data automatically. These techniques tend to consider the entire log to discover a process model. However, real-life event logs usually contain outlier behaviour that lead to incomprehensible, complex and inaccurate process models where correct and/or important behaviour is undetectable. Hence, removing outlier behaviour thanks to filtering techniques is an essential step to retrieve a good quality process model. Manually filtering the event log is tricky and requires a significant amount of time. On the other hand, some work in the past is focused on providing a fully automatic choice of the parameters of the discovery and filtering algorithms; however, the attempts were not completely successful. This demo paper describes an easy-to-use plug-in in the ProM process mining framework, that provides a view where several process discovery and outlier filtering algorithms can be chosen, along with their parameters, in order to find a sweet spot leading to a 'good' process model. The filtered log is easily accessible, and the process model is shown inside the view, in this way the user can immediately evaluate the quality of the chosen combination between process discovery and filtering algorithms, and is effectively assisted in the choice of the preprocessing methodology. Some commonly used metrics (fitness, precision) are reported in the view provided by the plug-in, in order to ease the evaluation of the process model. With the options provided by our plug-in, the difficulties of both fully-manual and automatic choice of the filtering approach are effectively overcome.

Keywords: Process Mining · Process Discovery· Outlier Detection · Interactive Filtering · Quality Improvement

1 Introduction

Process Mining bridges the gap between traditional data mining and business process management analysis. The main subfields of process mining are 1) *process discovery*, i.e, finding a descriptive model of the underlying process, 2) *conformance checking*, i.e, monitoring and inspecting whether the execution of the process in reality conforms to the corresponding designed (or discovered) reference process model, and 3) *enhancement*, i.e, the improvement of a process model, based on the related event data. With process mining we discover knowledge from event data, also referred to as *event logs*,

readily available in most modern information systems. In all these subfields event logs are used as a starting point.

Process discovery algorithms extract a process model out of the event log, with the aim to get a description of the reality. Some discovery techniques like the Alpha miner [3] aim to depict as much as possible behaviour in the event log with the assumption that all the information related to the execution of the underlying process is stored correctly. However, real event data often contains inaccurate or corrupt behaviour that should not be a part of the process model [9]. The infrequent behavior could be due to:

- Logging errors, e.g., mistakes and the inaccuracy of measurements.
- Behaviour that is rare due to the handling of exceptional situations.

Both types of infrequent behaviour- we call them **outliers**- make most process discovery algorithms return incomprehensible or even inaccurate process models. To reduce these negative effects, we can benefit from outlier filtering algorithms that try to detect and remove traces that contain such undesired behaviour [4]. By preprocessing event logs using these outlier filtering methods before the process discovery phase, we can improve the quality of discovered process models [10].

Process discovery is often an explorative approach. It means that we usually need to apply different process discovery algorithms with several parameters to generate different process models and evaluate them. To measure the quality of a model we usually use *fitness*, *precision* and *simplicity* [7]. Fitness measures how much behaviour in the event log is also described by the process model. On the other hand, precision computes how much of behaviour, that is described by the process model, is also presented in the event log. Moreover, *simplicity* measures how much the process model is simple to interpret. After discovering a process model, the end user needs to apply several unintegrated quality metrics to measure the quality of it and get a good quality process model.

In this paper, we propose an easy-to-use interactive filtering toolkit that permits the evaluation of several process discovery algorithms and outlier filtering mechanisms simultaneously, and let the user see the discovered process model (beside its quality measures). Unlike the most common trend that just focus on the most general behavior, we also let the users to discover a process model on infrequent behavior, to support deviation detection applications. After finding the desired process model, the process instances that are fit/unfit as an output to this process model can be retrieved for further analysis. The tool is implemented in the ProM framework [2], that is currently one of the most widely used open-source process mining platforms. Using the proposed toolkit, a good quality process model can be discovered in an easier way.

The remainder of this paper is structured as follows. In Section 2, we discuss the related work and the motivation of having our toolkit. Section 3 describes the main features that are provided in the implemented toolkit. Section 4 concludes the paper and presents some directions for extending the implementation.

2 Related Work and Motivation

Many process discovery algorithms are proposed in the literature. Some of them as the Alpha Miner [3], and the ILP Miner [14] depict as much as possible behaviour of the

event log in the process model. Some others as the Split Miner [5] and the Inductive Miner [11] have internal filtering mechanisms to deal with outlier behaviour. For each of these algorithms, there are different variations and settings that can be used. The combination of different settings usually result in completely different process models. However, according to [4] we could not have a discovery setting that always results in the best process model. Note that the best setting is usually not known and a suitable process model should be found by several try and error attempts.

In [8–10], it is shown that by automatically removing outlier behavior in event logs, process discovery algorithms are able to discover process models with higher quality. Moreover, [12] uses data attributes to filter out noisy behavior. The filtering algorithms proposed by research are, unfortunately, scarcely used in most of commercial process mining tools, where only basic filtering techniques are provided, such as variants, attributes, paths, timeframe and performance filters. Note that, the removal of outlier behaviour is a very important step in order to get an accurate process model.

The usual process of discovering a good quality process model can be summarized as an iterative process including the following steps (in order):

1. Preprocessing the event log.
2. Applying a discovery algorithm.
3. Perform several analysis, with different plug-ins, on the quality of discovered process models (fitness, precision, simplicity).

The quality of values for the current iteration affects the settings of the algorithms in the next iteration. For example, if the fitness of a discovered model is high but its precision is low, we may need to consider less infrequent behaviour in the model. These iterations will continue until the desired process model is found. All these steps wastes a lot of time on the human side. This problem motivates us to provide a solution that makes it possible to filter out infrequent behaviour and discover process models in an interactive way. So, the aim is to provide a toolkit to let the user see the results of each change in thresholds or method on the discovered process model and its quality measures interactively.

There are few available interactive discovery plug-ins, e.g., the *Inductive Visual Miner* and the *Interactive Data-Aware Heuristic Miner*, that are widely used to discover process models with some decorations. However, both of these tools just focus on specific discovery techniques, without providing pre-processing features. Also, these plug-ins do not provide quality measures on the output process model.

The plug-in proposed in this paper aims to provide an holistic approach of preprocessing and discovery, that is able to display in a single representation both the process model, the choice of discovery and filtering algorithms, and the quality measures. This helps to discover a process model with suitable quality in a systematic way, by leading the user to the best choice of discovery and filtering approaches. For example, user can see by increasing the filtering threshold, the fitness will be decreased and the precision will be increased.

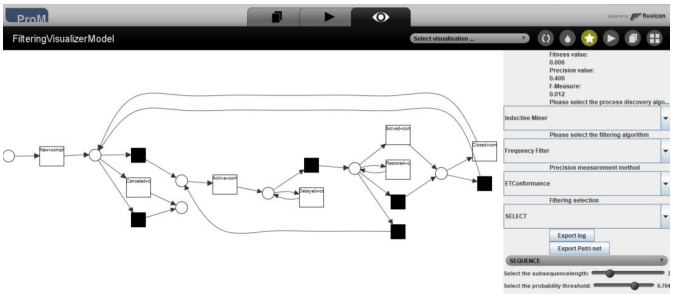

Figure 1: The screen shot of the discovered process model.

3 Interactive Event Log Filtering

We developed our toolkit in the ProM framework to increase its integration with other process mining plug-ins. This open-source toolkit is called Interactive Filtering[3]. The input of this toolkit is an event log and the desired process model is obtained along with a completely fit/unfit event log. A video demo, that shows how our toolkit can help users to find their desired process model intractively on a real-life event log is provided[4].

Figure 1 shows the general view of our toolkit. In the right panel, the settings of process discovery and outlier filtering parameters can be adjusted. In the left panel, the resulting process model are shown as a Petri net. Also, quality values of the discovered process model are shown. Here, we specify the options that our toolkit provides in different aspects.

Process discovery algorithms: To discover the process model, the Alpha miner [3], the Inductive Miner with infrequent behaviour filtering [11], the ILP miner [14], and the Split Miner [5] can be used.

Outlier Filtering: Different outlier filtering techniques can be applied such as common variant filtering, probabilistic methods [8, 9], and sequential mining based method [10]. There are also some other techniques that work based on different abstractions (i.e., set, multi-set, and sequence) and frequency of them in an event log. Also, the user can decide if he/she is interested in the main stream or infrequent behaviour.

Quality Measures: To evaluate the quality of process models, the precision [13], fitness [1] and F-Measure of process models are reported and the user can measure the simplicity by herself. In all of these measurements the original event log is used.

In all the mentioned algorithms, the settings can be adjusted. Moreover, the toolkit is developed in a way that it could easily be extended by any other ProM plug-ins. Any other process discovery algorithm that receives an event log as an input, and returns a Petri net along with an initial marking could be added to this toolkit. This is the same for filtering plug-ins when they take into account only event logs.

The plug-in is intended to be easy to use that means no detailed knowledge about the techniques described in the outlier filtering literature is required.

[3] The last version is accessible through `svn.win.tue.nl/repos/prom/Packages/LogFiltering`, and a previous version is available in ProM 6.9 via the package manager.

[4] `https://youtu.be/T31sLvfQD0E`

4 Conclusion

Here, we present an interactive event log filtering toolkit that enables the discovery of a good quality process model with an easy and systematic approach. Using this Toolkit that is implemented in ProM, users can filter interactively infrequent behavior using many different algorithms, and apply several process discovery algorithms. It is also possible to focus on infrequent behaviour, this may help in understanding deviations. Finally, the toolkit can return the desired process model along with an event log containing the process instances that are perfectly fit according to the model (or the ones containing deviations). We aim in future to extend this toolkit with more process discovery and filtering algorithms and to add this toolkit to other platforms like PM4Py [6].

References

1. Van der Aalst, W., Adriansyah, A., van Dongen, B.: Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **2**(2), 182–192 (2012)
2. van der Aalst, W.M.P., van Dongen, B., Günther, C.W., Rozinat, A., Verbeek, E., Weijters, T.: ProM: The Process Mining Toolkit. BPM (Demos) **489**(31) (2009)
3. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow Mining: Discovering Process Models From Event Logs. IEEE Trans. Knowl. Data Eng. **16**(9), 1128–1142 (2004)
4. Andrews, R., Suriadi, S., Ouyang, C., Poppe, E.: Towards Event Log Querying for Data Quality: Let's Start with Detecting Log Imperfections (2018)
5. Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Polyvyanyy, A.: Split miner: Automated Discovery of Accurate and Simple Business Process Models from Event Logs. Knowledge and Information Systems pp. 1–34 (2019)
6. Berti, A., van Zelst, S.J., van der Aalst, W.: Process Mining for Python (PM4Py): Bridging the Gap Between Process-and Data Science pp. 13–16 (2019)
7. Buijs, J.C., van Dongen, B., van der Aalst, W.M.P.: On the Role of Fitness, Precision, Generalization and Simplicity in Process Discovery. In: OTM, " On the Move to Meaningful Internet Systems". pp. 305–322. Springer (2012)
8. Conforti, R., La Rosa, M., ter Hofstede, A.: Filtering Out Infrequent Behavior from Business Process Event Logs. IEEE Trans. Knowl. Data Eng. **29**(2), 300–314 (2017)
9. Fani Sani, M., van Zelst, S.J., van der Aalst, W.M.P.: Improving Process Discovery Results by Filtering Outliers Using Conditional Behavioural Probabilities. In: Business Process Management BPM Workshops, Barcelona, Spain. pp. 216–229 (2017)
10. Fani Sani, M., van Zelst, S., van der Aalst, W.M.P.: Filtering Outliers Using Sequence Mining. In: COOPIS. pp. 216–229. Springer (2018)
11. Leemans, S.J., Fahland, D., van der Aalst, W.M.P.: Discovering Block-Structured Process Models from Event Logs Containing Infrequent Behaviour. In: Business Process Management Workshops, pp. 66–78. Springer International Publishing (2014)
12. Mannhardt, F., de Leoni, M., Reijers, H.A., van der Aalst, W.M.P.: Data-driven process discovery-revealing conditional infrequent behavior from event logs
13. Munoz-Gama, J., Carmona, J.: Enhancing precision in process conformance: Stability, confidence and severity. In: IEEE CIDM. pp. 184–191. IEEE (2011)
14. van der Werf, J., van Dongen, B., Hurkens, C., Serebrenik, A.: Process Discovery using Integer Linear Programming. Fundam. Inform. **94**(3-4), 387–412 (2009)