PM4KNIME: Process Mining Meets the KNIME Analytics Platform

Humam Kourani
Fraunhofer FIT
Sankt Augustin, Germany
humam.kourani@fit.fraunhofer.de

Sebastiaan van Zelst
Fraunhofer FIT
Sankt Augustin, Germany
sebastiaan.van.zelst@fit.fraunhofer.de

Barry-Detlef Lehmann
Fraunhofer FIT
Sankt Augustin, Germany
barry-detlef.lehmann@fit.fraunhofer.de

Gabriel Einsdorf
KNIME GmbH
Konstanz, Germany

Stefan Helfrich
KNIME GmbH
Konstanz, Germany

Fabian Lße
KNIME GmbH
Konstanz, Germany

***Abstract*—Process mining allows organizations to transform the data recorded during the execution of their processes into meaningful insights. These insights can help to detect problems and to improve the processes. Various process mining solutions have been developed, both for industrial and academic purposes. However, most of these solutions do not support the creation and execution of analytics workflows. The KNIME Analytics Platform (KNIME in short) is an open-source workflow-based analytics platform that supports various techniques in the field of data science. KNIME is widely used in numerous industries across many countries. This paper presents the process mining extension for KNIME, which integrates many powerful process mining algorithms into KNIME. Using the process mining extension of KNIME, process mining can be combined with other types of data science techniques available in KNIME.**

***Index Terms*—process mining, data science, workflow**

I. Introduction

Process mining helps to analyze and monitor processes based on the events recorded during their execution. The goal of process mining is to extract information from these events to allow organizations to detect problems in their processes and improve decision-making. The field of process mining [1] covers all techniques for discovering process models, checking conformance between event data and process models, and recommending process enhancements.

The growing interest in process mining led to the development of numerous process mining tools. ProM (https://www.promtools.org/) is one of the most powerful (academic) process mining tools available, i.e., it contains hundreds of plugins that implement numerous process mining algorithms. However, its academic nature hampers integration in other applications, and it does not support the creation and execution of analytical workflows. To bring process mining into a user-friendly workflow-based environment, we present the open-source process mining extension of KNIME: PM4KNIME.

KNIME [2] is an open-source workflow-based analytics platform that supports various techniques in the field of data science, e.g., machine learning, data mining, modeling, etc. Workflows are built in KNIME by sequentially connecting different nodes where each node is dedicated to performing a specific task based on the results of the preceding nodes. The KNIME Hub (https://hub.knime.com/) contains thousands of workflows ready to be applied to data sets. KNIME provides extensions and nodes for integrating many projects, systems, web services, and databases. For example, it supports the integration of Python (https://www.python.org/), Apache Spark (https://spark.apache.org/), MongoDB (https://www.mongodb.com/), and many cloud storage systems. *KNIME Server* is commercial software that enables collaboration between users and supports automated and distributed executions of workflows, deployment options, workflow management, and monitoring functionalities.

Thanks to its ease of use and high scalability (distributed executors on the KNIME Server, big data, and cloud integration), KNIME software is used by hundreds of companies in numerous industries. PM4KNIME integrates process mining algorithms implemented in ProM into KNIME. This allows for creating analytics workflows that combine process mining with the other types of data science techniques available in KNIME in a scalable, user-friendly environment. Instruction on how to install PM4KNIME can be found under https://pm4knime.github.io/userDoc/guides/installation.

II. Tool Overview

In this section, we provide an overview of PM4KNIME. A screen recording corresponding to this overview is available under https://pm4knime.github.io/userDoc/guides/demo.

A. KNIME Workflows

KNIME stores data in table-based objects called *DataTables*. Algorithms in KNIME are implemented as *nodes*. A node can have multiple *input ports*, *output ports*, *views*, and *dialogs*. The input ports should be connected to the input objects required for executing the underlying algorithm of the node. The dialogs are used to set the parameters of the algorithm. After the successful termination of an algorithm, the output objects can be accessed through the output ports. A *workflow* in KNIME is a directed graph connecting multiple nodes through their input and output ports.

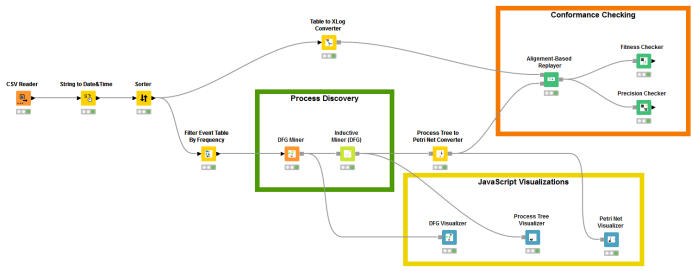

Fig. 1. Example process mining workflow in KNIME.

B. Functionalities

PM4KNIME currently supports:[1]

- Importing and exporting different objects (e.g., Petri net).
- Exploring event logs (e.g., dotted chart).
- Converting objects (e.g., XES logs into DataTables).
- Event data manipulation (e.g., filtering).
- Process discovery (e.g., inductive miner).
- Conformance checking (e.g., alignment-based replay).
- JavaScript visualizations (e.g., for Petri nets).

Most implemented nodes work on DataTables. Internally, we wrap around the implementations of the underlying process mining techniques from the plugins available in ProM.

C. Example Workflows

Fig. 1 shows a typical workflow in the field of process mining. It contains nodes for importing data from a CSV file, preprocessing, process discovery, JavaScript visualizations of the discovered models, model and data transformation, and conformance checking. We applied this workflow to a real-life data set that records the execution of a ticketing management process [3]. Further workflow examples are available on the KNIME Hub under: https://kni.me/s/VJqKc-EypN7Jkrl2.

D. Tool Novelty

In [4], RapidProM was introduced as an extension of RapidMiner. It integrates process mining algorithms from ProM into the workflow-based platform RapidMiner. The idea of [4] is similar to our contribution, but PM4KNIME provides some features that differentiate it from RapidMiner.

We adapted some process mining techniques not supported in RapidMiner (e.g., hybrid Petri net miner). Moreover, most implemented algorithms in PM4KNIME work on DataTables (not XES logs). We wrapped around the implementations of the underlying process mining techniques in ProM. In data science, data is often stored in table-based files (e.g., CSV files) that can be easily imported as DataTables in KNIME. Applying process mining algorithms directly on DataTables improves the time performance because KNIME uses powerful caching strategies that ensure high scalability when processing large DataTables [2].

The KNIME Server provides many valuable features for organizations, such as automated and distributed executions of workflows, deployment options, workflow management, and monitoring functionalities. PM4KNIME provides JavaScript visualizations for the different types of supported process models. This allows for building interactive web-based applications using the deployment options on the KNIME server.

Both RapidProM and PM4KNIME allow for saving workflows to be reused later. However, PM4KNIME additionally supports the serialization of intermediate results. Each node in a KNIME workflow processes its entire input data and permanently stores its output before forwarding it to the successor nodes. By saving a workflow, the settings of all nodes and all already generated (intermediate) objects are stored together with the workflow structure. Therefore, it is possible to stop the execution of a KNIME workflow at any node. The workflow can be modified and saved to be resumed later without needing to re-execute already executed nodes that are not affected by any modifications. For all implemented (intermediate) objects in PM4KNIME, we created internal importers and exporters to support the serialization of results.

III. Conclusion

In this paper, we introduced the process mining extension of KNIME (PM4KNIME). PM4KNIME integrates process mining algorithms that are implemented in the academic process mining tool ProM into a workflow-based data science analytics platform that is widely used in industry. The process mining extension of KNIME supports many techniques for process discovery, conformance checking, event data manipulation, and visualization of process models.

As future work, we aim at adapting further algorithms to work directly on DataTables instead of XES logs (e.g., conformance checking algorithms). Moreover, we aim at supporting more types of process models (e.g., BPMN models) and integrating more process mining algorithms from ProM and/or other academic tools like PM4Py (http://pm4py.org/).

Acknowledgment

The authors would like to thank Kefang Ding and Ralf Riesen for their contribution to PM4KNIME.

References

[1] W. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[2] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: the konstanz information miner," in *Data Analysis, Machine Learning and Applications - Proceedings of the 31st Annual Conference of the GfKl*, ser. Studies in Classification, Data Analysis, and Knowledge Organization, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker, Eds. Springer, 2007, pp. 319–326.

[3] M. Polato, "Dataset belonging to the help desk log of an Italian Company," 2017.

[4] R. Mans, W. M. P. van der Aalst, and H. M. W. Verbeek, "Supporting process mining workflows with RapidProM," in *Proceedings of the BPM Demo Sessions Co-located with the 12th International Conference on Business Process Management*, ser. CEUR Workshop Proceedings, L. Limonad and B. Weber, Eds., vol. 1295. CEUR-WS.org, 2014, p. 56.

[1] See https://hub.knime.com/pm4knime/extensions/org.pm4knime.feature/latest for a complete overview of all available functionalities.