Università di Camerino

University
University of Camerino

School of Science and Technology
MSc in Computer Science
Process Mining Course

Process Mining project

Student
Tommaso Catervi - 121049

Professor
Prof. Barbara Re

Academic Year 2021/2022

Abstract

Processes are all around us. Any series of tasks that together achieve an objective can be called a Process. Thanks to the digital revolution copious amounts of data related to diverse processes are being generated and accumulated. In the field of data science, analysis and drawing insights from the operational processes is of particular importance.

Modelling the process allows us to perform conformance checks and even provide us with the capability to improve the processes. This kind of extraction of insights from event data is called process mining.

The objective of this project is to apply some of the most common process mining algorithms seen during the course on a given input log, experimenting and becoming familiar with some of the most widely used tools in the field today.

Contents

Introduction

The purpose of this chapter is introducing the main objective of the project and the data used to carry it out.

1.1 Project Description

The objective of this project is to apply some of the most common process mining techniques and algorithms seen during the course using an arbitrary input log, experimenting and gaining knowledge of some of the most widely used tools in the field today.

For the purpose of this project, a log sourced from 4TU.ResearchData was used.

4TU.ResearchData is an international data repository for science, engineering and design. Its services include curation, sharing, long-term access and preservation of research datasets. These services are available to anyone around the world. In addition, 4TU.ResearchData also offers training and resources to researchers to support them in making research data findable, accessible, interoperable and reproducible (FAIR).

In particular, the chosen log consists of an event log obtained from a Loan Application Process. It is dated 23.04.2012 and comes from Eindhoven University of Technology. It contains 13087 traces and was sourced in a XES format (eXtensible Event Stream), that is the standard format for process mining, supported by the majority of process mining tools.

Tools

The purpose of this chapter is presenting the tools and technologies studied and used during the realization of the project, briefly introducing their functionalities and employment.

2.1 PM4Py

PM4Py (Figure 2.1) is a software product developed by the Fraunhofer Institute for Applied Information Technology (FIT). It consists of a Python library that supports (state-of-the-art) Process Mining algorithms in Python. It is completely open-source and intended to be used in both academia and industry projects.

Figure 2.1: PM4Py logo.

Some of the hot feautures included are multiple-format event data handling, custom event data filtering, process discovery, Petri Nets management, conformance checking, log-model evaluation and simulation.

PM4Py does not provide any kind of graphical user interface, but is in fact a low-level tool that provides the user with a rich library of functions and tools.

For the purpose of this project, the tool was used for the anlysis of the Loan Application events.

2.2 ProM Tools

ProM (Figure 2.2) is an extensible plugin-based framework for process mining supporting a wide variety of process mining techniques. It is platform independent as it is implemented in Java, and can be downloaded free of charge.

Figure 2.2: ProM Tools logo.

The framework is centered around three basic concepts:

- Data Objects which live the workspace tab;
- Plugins which you can access via the run tab;
- Visualizers which you can access via the visualizer tab.

For the purpose of this project, ProM was used exclusively for the final visualization of the Petri Net obtained by applying the Alpha algorithm.

Project development

The purpose of this chapter is presenting the development of the project and the experimentation of some of the process mining tools and techniques.

3.1 Process discovery

The Process Discovery algorithms are used to find a suitable process model based on an events log, in this case the Loan Application log. The used process discovery algorithms are the Alpha Algorithm and the Heuristic Miner, both of them were executed using the functions provided by the tool PM4Py.

The Alpha algorithm focus on control flow such as the ordering of the activities and it is the first process discovery algorithm ever created. It scans the event log for particular patterns. The goal of the Alpha algorithm is to take an event log in input and convert it into a Petri Net. To do that, it considers the footprint of the log, that is based on direct successions, causality successions, parallel successions and choice successions. The algorithm has some kind of problem, for example, it cannot handle loops of lengths one and two, it cannot discover an invisible and duplicated task and it is weak against noise.

The Heuristics Miner is an improvement of the Alpha algorithm. We consider the frequency of the traces, we count the cardinality of the direct successions on the traces, and we calculate the dependency measure. This provides the basis for identifying the dependencies and to build the resulting Heuristic Net. Thanks to the Heuristic net, it is possible to see the number of times each activity and path is executed. The algorithm also provides a way to handle noise by applying threshold values on activity and arcs to consider only the strong relations excluding the weaker. Eventually, there are tools that can convert this dependency graph into a Petri Net.

3.2 Results

All the final Nets are the result of function provided by the tool PM4Py, that provides a lot of functionalities about Process Mining. This Heuristic graphical result is really clear and the dependencies inside the graph itself are well shown, thanks to a function of PM4Py the resulting Dependency graph has been converted into a Petri Net. Both the Petri Nets and the Heuristic Net, coming from Heuristic Miner and from Alpha Algorithm, will appear as images when you run the script.py, thanks to the functionalties of the graphical tool GraphViz. The two Petri Nets can also be found in the two files with the extension ".pnml", that can be exported in any other Process Mining tool that supports that format, for example I used the tool ProM to visualise it.

Conclusions

In this chapter I will present my personal conclusions on the work done.

The project can be intended as a playground to understand the basic process mining algorithms and techniques in practice and to try out some of the process mining tools available today. In particular, PM4PY happens to be a truly powerful piece of software providing a lot of basic and advanced features, and it is indeed the one I liked the most. PM4Py allows for complex operations such as Petri Net and Process Tree management, feature selection and prediction for Decision Mining, evaluation of quality metrics, simulations, Social Network Analysis, Streaming Process Mining, and more.

The final results obtained are satisfactory and two different algorithms have been implemented, namely the Alpha algorithm and the Heuristic Miner, both with the same input log. Because of the large size of the log used as input, it is unfortunately not particularly easy to compare the two resulting Petri Nets. However, the dependency graph was sufficiently clear and well readable.