UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2021/2022, Quarter 1B

DATE
November 15, 2021

TEACHERS
Faizan Ahmed
Nacir Bouali
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Nicola Strisciuglio
Estefania Talavera
Brenda Voorthuis
Shenghui Wang

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

COURSE COORDINATOR
Maurice van Keulen (quartile 1A)
Karin Groothuis-Oudshoorn (quartile 1B)
Faizan Ahmed (quartile 2A)

Topic 7

Process Mining [PM]

7.1 Introduction

Nowadays, the amount of data generated is growing enormously. In general, there are four sources of event data: Internet of content (e.g. Google), Internet of people (e.g. Facebook), Internet of things (e.g. products from Phillips), Internet of places (e.g. Android). The exponential growth of data is challenging for scientists. The big question is how to extract real value from data by keeping 4V's of data: volume, velocity (rapid changes), variety(a lot of types of data), veracity(cannot be sure data is accurate, uncertainty). A data scientist is able to collect, analyze, and interpret data from a variety of sources by posing four generic data science question,

- what happened?
- why did it happen?
- what will happen?
- what is the best that can happen?

In order to interpret data and get answers to such questions a lot of different skills are necessary. Process mining is one of the skills that help to analyze, visualize and optimize these big datasets. In a dataset, in the end, it is the process that matters (not the data or the software) and not just patterns and decisions, but end-to-end processes. Improvement of these end-to-end processes by interplaying between event data and process models is the focus of this topic.

We will discuss three dimensions of process mining: process discovery, conformance checking, and enhancement. It is not about collecting data but analyzing business processes. Process mining bridges the gap (= missing link) between process model analysis and data-oriented analysis. The starting point for process mining is event data. Event data has different properties such as case id, activity name, timestamp, but also the resource and other related data. Mainly three types of relations between event data and processes exist: **play-out** is from (process) model to software system and the real world, **play-in** is discovery from event logs to the model and the conformance and enhancement of the model is done in **replay**.

7.1.1 Global description of the practicum and project

In the practicum, we will focus on how process mining really works? We will use event logs for an objective reconstruction of the process flows. The main focus is on discovering "why" instead of "what" happened in

the business processes.

7.1.2 Study material and tools

1. Mining, P. (2011). Discovery, Conformance and Enhancement of Business Processes. Springer-Verlag, 8, 18. https://ut.on.worldcat.org/search?queryString=bn%3A3642193455#/oclc/728098360
2. Van der Aalst, W. M. (2016). Process mining: data science in action. Springer. https://ut.on.worldcat.org/search?queryString=au:Wil%20van%20der.%20Aalst&databaseList=2375,3218,233,1875,3448,3535,2897,1697,3336,3313,3909,638,1847#/oclc/946935914
3. www.processmining.org

7.1.3 Deliverables and obligatory items

Topic teacher: Faiza A. Bukhsh

The practicum requires individual files including images and results. Please combine them in a ZIP-archive and upload that to Canvas.

7.2 Description of the practical assignments

7.2.1 Installation

Instructions In the following exercises, we will use one of the well-known process mining tool, Disco, and ProM. Disco is a commercial tool for process mining. It has a demo mode, which is free, but in the demo, we can only use up to 100 events. This is a big limitation for big data analysis. Therefore, we will use ProM. There is a lite version and a normal version of the tool since the normal version has unstable plug-ins in it. Besides that, the offer of plug-ins is overwhelming, which makes it not easy to use for the course Data Science. Following assignments are based on ProM Lite version 1.2. ProM lite1.2 can be downloaded from the website www.processmining.org. It is an open source framework for process mining algorithms. It can be used for free. Detailed guidelines can be found on http://www.promtools.org..

Note: After installation of ProM lite1.2, it needs packages from the Internet for initial set-up and if your Internet connection is weak, packages cannot be downloaded and you will receive an error ”No import plug-ins are available ”. The solution for such a problem is to use the wired Internet for setting up ProM lite1.2 first time. □

7.2.2 Data set(s)

For Assignments we will use following datasets,

- companyA.xes
- teleclaim.xes

7.2.3 Assignment 1: Process discovery from event logs (on paper)

Expected effort 4 hours.

The assignment is based on process discovery concepts. Process discovery starts with cleaning data for a process-mining tool. Sometimes process flows can be identified by interviewing process stakeholders. From event logs real process flow can be observed, rather than how it is perceived by its actors. Exception scenarios are mostly not known to the stakeholders. Initial process model generated from event log provides the basis for further analysis, such as process compliance, or a prediction of process paths based on historical data. Consider event log of CompanyA available on Canvas:

☞ **7.1**

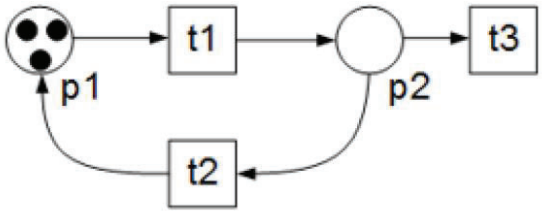


Figure 7.1: A PetriNet

(a) Explain what is this event log about, explain the process. Describe the observed business scenario in detail.
(b) Produce event traces/casual footprints from event log from cases 5 till 10.
(c) Given the case above, name if there is any concrete starting and ending event?
(d) Produce a generalized process model by hand based on given event log.
(e) Create a petri-net for (d) based on `http://pages.di.unipi.it/ferrari/CORSI/SISD/Lezioni/WFModel.pdf`.

□

7.2.4 Assignment 2:The PROM Tool

Expected effort 2 hours.

Following assignments are based on ProM Lite version 1.2.

☞ **7.2**

(a) Import companyA.xes into ProM and obtain an overview of the data set. What is the most concurrent event in the dataset?
(b) Create and explain a Petri-net from the transition system using (any)plug-in in ProM.

□

7.2.5 Assignment 3: Petri Nets

Expected effort 2 hours.

The input and output of many actions in ProM are a Petri-net. Because of the usage in the tools and because it is a good way to present the outcome of process discovery, this assignment is based on concepts of states, transitions, arcs and tokens. Specific concepts of Petri-nets like reachability graphs, liveness and bounded are introduced as well.

☞ **7.3**

(a) Consider Figure 7.1, which of the following properties hold and explain why,
   a. All places are safe
   b. All transactions are dead
   c. All places are bounded
   d. The marked Petri-net has reachable dead marking
   e. All transactions are live
(b) Produce a Petri Net from the CompanyA's Dataset. Which of the above properties hold for this Petri Net?

□

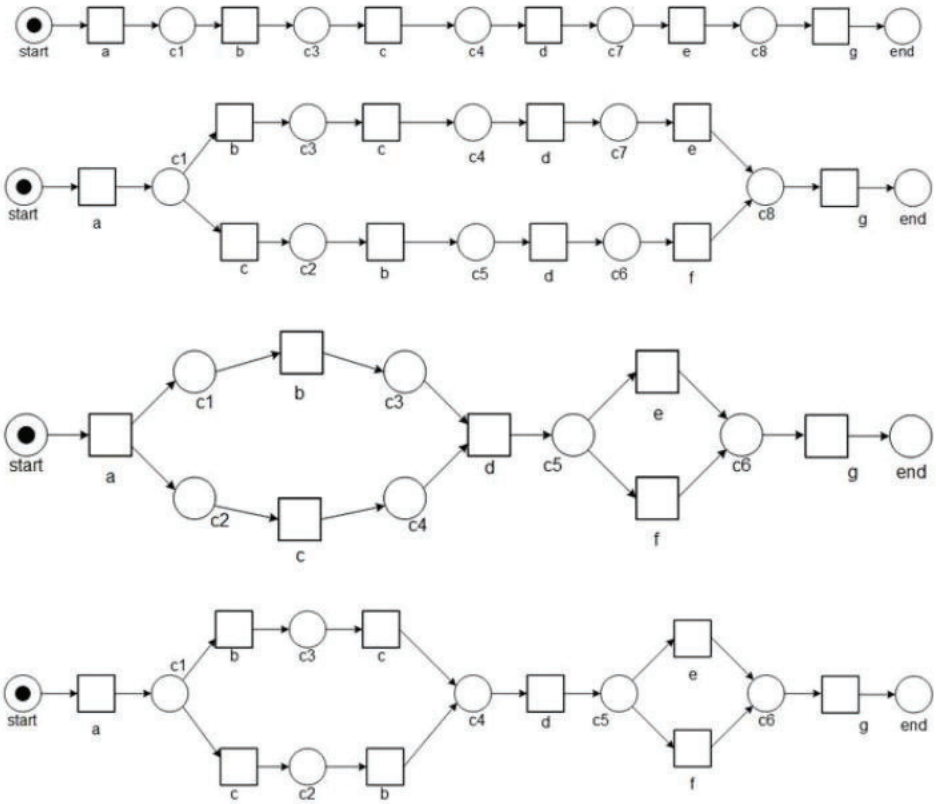
7.2.6 Assignment 4: Process Discovery by using Alpha Algorithm (on-paper)

Expected effort 2 hours.

There are different algorithms to discover a process from event data. The most initial and basic is the alpha algorithm. The alpha algorithm is good in discovering basic processes. The alpha algorithm consists of eight steps. Before these eight steps, a causal footprint must be created. You have created a casual footprint in Assignment 7.1(b) already.

☞ **7.4**

(a) Given the event log $L = [< a,b,c,d,e,g >^8, < a,c,b,d,e,g >^1, < a,c,b,d,f,g >^1]$, which result does the Alpha Algorithm produce?



(b) Given the event log $L = [< a,b,c,d,f >, < a,b,d,c,f >, < a,c,d,b,f >, < a,e,f >, < a,d,c,b,f >]$ execute the Alpha algorithm manually to produce the Petri Net for L.

□

7.2.7 Assignment 5:Process Discovery and Enhancement in PROM

Expected effort 8 hours.

This exercise is about discovering the process and analyzing the process based on the replaying of data on the discovered processes in order to do more extensive analysis. The dataset **teleclaim.xes** is about the handling of claims in an insurance company. The exercise is about comparing the results of different miners on the teleclaim.xes data set. Your task is to explore which miner works the best and why.

☞ **7.5**

(a) Execute the miners below and compare the produced process models. Which one do you find best? Explain why. Include screen shots of the process models discovered by each miner.
   (i) Alpha miner
   (ii) Heuristic Miner (must be followed by a transition from a Heuristics Net to a Petri Net)
   (iii) Inductive Miner
   (iv) IPL-Based Process Discovery

(b) Suggest at least two improvements in your chosen (best) process model.

□