Documentation for Event Compliance over Streaming Event Data

Jingjing Huo 376323
Sabya Shaikh 384606
Zheqi Lyu 378653

Content

0 Project Overview

Background:

Process mining is widely used to visualize processes, analyze processes performance based on event logs. Unfortunately, the majority techniques handle the offline events data. Moreover, some techniques are applied on live event streams under the assumption, that the input stream has no spurious events which leads to the degradation of results quality on the real data. In order to release the assumption, we build probabilistic automata to filter out these spurious events.

Goal:

- Our project will be able to effectively check event compliance,i.e when anomalous pattern is detected on the streaming data, alerts will be triggered and spurious events will be filtered out using probabilistic automata with different prefix length
- Our system will handle 1000,000 streaming events in 1 hour.
- Eventually, the user must be able to access web-service via the client using http post request to provide the streaming events in order to check event compliance

Assumptions:

- The order of event arrival corresponds to the order of execution.
- The event is spurious, if any of the automata alerts that event to be spurious.

Constraints:

- The filtering of event will be done taking into account only recent history.
- Validity of real life data is hard to determine because of the lack of ground-truth.

1 Project Initiation

1.1 Related Business Cases

Event logs are important evidence to see if a company or organization works good. Until now a mass of event logs are recorded by information systems. In order to make it more useful, some visualization and evaluation should be done. Recently, some techniques are used for analysing the underlying processes in many different fields. Such as:

Service - Insurance [6]

Organization: (Netherlands)
Process: *Dental care process analysis, 2017*

Description: Solving of many operational problems much quicker by combining Lean tools with process mining. Using process mining, VGZ was able to visualize the flow of the dental care process within weeks. This directly pointed out bottlenecks and it demonstrated that there were long waiting times when the work was handed over from medical advisors to experts and vice-versa. By applying the traditional Lean tools, such as 5x Why, CZ was able to pinpoint the actual root causes.
Results:

- Reduction of the throughput time by 40%.

Service - Public Sector[6]

Organization: Copyright mediator company(Italy)
Process: *Event Licence Approval, 2016*
Description: His case study applied process mining techniques to event licence approval process to expose deviations and performance issues. Specifically, the study involved process discovery of the "as is" model and the conformance checking of the "as is" process to the expected process.
Aim: finding out the root of the problem that was affecting company's core processes.
Results:

- Anomalies and bottlenecks were clearly detected.
- It was found that core processes didn't perform well because of the lack of quality data and transparent communications.

Healthcare[6]

Organization: AMC Hospital(Netherlands)
Process: *Conformance analysis from Billing system, 2015*
Description:

- Mining the complex hospital processes giving insights into the process;
- Deriving the understandable models for large groups of patients;
- Comparing results with a flowchart for the diagnostic trajectory of the gynaecological oncology healthcare process.

Recent applications use process mining techniques in the offline setting, but sometimes the immediate responses of deviations are needed, especially relating to the live streaming data. With the advent of big data, velocity of data flow has increased rapidly and the need of systems that are able to handle this huge in-flow of data has arisen. Furthermore, presence of the spurious events in these streaming event logs leads to the degradation of online process mining techniques results. Filtering out this noise and processing the requests in real-time scenario with increased speed using Probabilistic Automata can boost effectiveness of these online process mining techniques(e.g. Conformance Checking) in terms of quality and time. We aim at building a project that realizes this(above) system

1.2 Feasibility Study

1.2.1. Theoretical Point of View:

The main idea of event-filtering using probabilistic automata is based on the paper [1]. The work proposed in this paper uses Java as the implementation language and we assume the similar idea could be implemented in python.

Test Data:

For the experiments relating to filtering accuracy, we can use 21 variations of the loan application process model from the paper [2] to generate 21*5 different random event logs.

1.2.2. Technical Point of View:

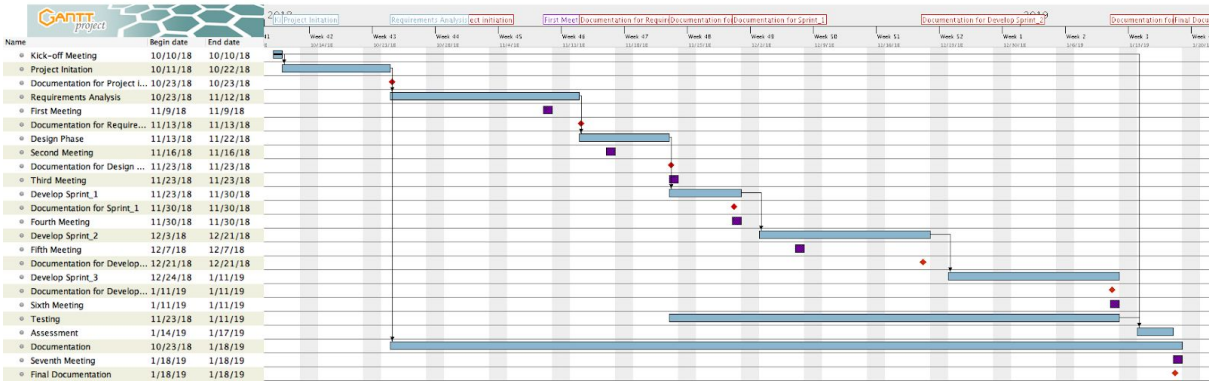
Programming language: Python
Web Service framework: Flask

Many companies use Python and Flask to construct their product and provide services. The combination of them has many benefits.
For eg, Netflix makes use of Flask[3] to build Rest Apis because of its adaptability and ease of implementation. To achieve the goal of building an API heavy product which enables 27,000+ developers to use and extend, WakaTime is therefor built with Flask [4].

For our compliance checking system, we aim at building a scalable system, which can better integrate with company software, accept the streaming data, return the JSON and is easy to implement and maintain.
Our project demands adaptability, ease of implementation, quick JSON-responses[4] and scalability to reach our goals which can be achieved using Python and Flask.

1.3 Project Charter

For the project management we use the Gantt Chart, which can show the duration of each phase and milestones of our project. The small red diamond in the chart indicates 7 milestones, and the purple rectangle show the time of meetings.



1.4 Project Team

The responsibilities of team member:

Tasks	**Team Member**
Designer, Developer, Tester, Documentation	Z. Lyu, J. Huo, S. Shaikh
Assign tasks, Set up group meetings, Communicate with tutor	Z. Lyu
Update Trello dashboard, Documentation reviewer	S. Shaikh
Update gantt diagram, Code reviewer	J. Huo

Zheqi Lyu (Team leader): My major is Master Computer Science and now it's my third semester. I have some project experiences from my bachelor project and I'm familiar with Java, C, C++. And I have also taken the courses of programming during the study in RWTH, such as, Matlab, Mathematics and this semester learning Python. Besides I have learned the web developer technics like PHP, JS, HTML, CSS but I am not good at them. I can also use the tools for database and big data, for example, SQL, Hadoop, Spark.
Since I study at RWTH, I force on the theory learning, and have taken more course in theoretical, like "Foundation of Data Science""Seminar in Theoretical concepts of Data Science" and also the course "Business process intelligence". Now taking this lab, in my opinion, will bring me more python programming experiences and deep aspects in process mining.

Jingjing Huo(Team-member): I'm in the third semester of computer science. My undergraduate degree is software engineering and I have learned C/C++, java programming

language during that time and also did some projects and a short-term internship in a software company. After I entered in RWTH-Aachen I have learned matlab, mathematica in some lecture, but don't have much programming experience. The most I have learned in my master program until now are theoretical things about learning theory, data science or business process intelligence and so on. Besides in this semester I also take a basic python programming class. So I think this practical is a good choice to improve my programming capability, I will cherish it.

Sabya Shaikh (Team-member): I am currently studying Masters in Software Systems Engineering-3rd Semester. I have taken courses like Advanced Data Model, Social Computing, Seminar in Theoretical concepts of Data Science. I have 3 years of professional experience from "Tesco Bangalore" in Data-warehouse Management. I have published paper "Classification of Facebook News Feeds and Sentiment Analysis" in IEEE conference (ICACCI, 2014). I developed multiple projects during my Bachelors study and familiar with programing Languages like C, C++, Java, SQL, HTML, PHP, JavaScript, Shell-scripting, Vue.js, Spring Boot and currently learning Python. Tools I have worked on - ETL processing in Abinitio, Teradata, Hadoop.
This lab shall be a great opportunity to learn about process mining techniques and spread my roots in data mining field.

1.5 Project Office

We use ***Trello*** to track our project and ***Github*** to manage our code. The dashboard of Trello and codes are available at https://trello.com/streamingeventcompliance and https://github.com/lvzheqi/StreamingEventCompliance

1.6 Phase Review

Review from Zheqi Lyu:
- About Lab: I like the organisation of this lab. It's bring me a clear structure of what we need to do and it's helpful for us to contact with tutor and clear the mission step by step.
- About groups: I feel comfortable to work with my teammates. They also help me to clear the doubts in the paper. Although we had differences in opinion of the project at the beginning, it is running well currently. I'm really thankful of their trust, that they choose me as the team leader.
- About me: So far english is my challenge. I'll try my best to improve it and we can do the lab more effectively.

Review from Jingjing Huo:
From my part, everything goes very well until now. Even my oral english is bad, sometime it's a little difficult to express my thoughts, but my teammates are quite patient. I feel comfortable to work with them. Up to now we have figured out what we are going to do and what is the most important thing. If we keep regular group meetings like in the first week, and ask our tutor for help in time when we have some problems, we are able to complete this project.

Review from Sabya Shaikh:
- Meetings with team members were planned well and hence we could manage time and work well before the deadlines.
- I did not have previous knowledge in this field as I did not take up the BPI course. Thanks to my team members for being really patient and helpful for clearing my doubts.
- It was really helpful meeting our tutor to clear our doubts regarding the problem statement. We had a different understanding about the problem statement after the Lab official meeting on Friday. But it was cleared by meeting tutor personally with the team.

Reference:

[1] S.J. van Zelst, M. Fani Sani, A. Ostovar, R. Conforti, M. La Rosa: Filtering spurious events from event streams of business processes. In: Proceedings of the CAISE (2018)
[2] M. Dumas, M. La Rosa, J. Mendling, H.A. Reijers: Fundamentals of Business Process Management. Springer (2013)
[3] F. Allanabanda, V. Veeraraghavan: Automation as a Service -Introducing Scriptflask (11.2017)
<https://medium.com/netflix-techblog/automation-as-a-service-introducing-scriptflask-17a8e4ad954b>
[4] H. Alan: Pirates use Flask, the Navy uses Django(04.2015)
<https://wakatime.com/blog/14-pirates-use-flask-the-navy-uses-django>
[5] A. Ostovar, A. Maaradji, M. La Rosa, A.H.M ter Hofstede, B.F. van Dongen: Detecting Drift from Event Streams of Unpredictable Business Processes. In: Proc. of ER'16. LNCS, vol. 9974, pp. 330–346. Springer (2016)
[6] HSPI Management Consulting, Process mining: a database of applications (2017)
<http://www.hspi.it/wp-content/uploads/2017/11/HSPI_Process_Mining_Database_v1.1-Nov_17.pdf>