A Method to Build a Production Process Model prior to a Process Mining Approach

An Illustration through the Detection of Incidents

Britta Feau and Cédric Schaller

Altran Est
Parc d'Innovation,
Boulevard Sebastien Brandt,
67404 Illkirch-Graffenstaden, France

Marion Moliner

Altran Research, MEDIC@
Parc d'Innovation,
Boulevard Sebastien Brandt,
67404 Illkirch-Graffenstaden, France
Email: marion.moliner@altran.com

***Abstract*—While manufacturing plants are becoming more digital, the large amount of data they produce remains under-exploited. In particular, event logs generated by production lines are usually monitored only in case of unusual events, such as production incidents. Process mining, a recent research field at the interface between data mining and process modelling, is specialized in extracting knowledge from event logs about the real process followed by a system. While the process model is usually delivered by the provider of the automaton, specific situations can lead to a lack of global view. This adds extra difficulties to carry out a root cause analysis of the incidents. In this short paper, we propose a simple method to overcome that issue. We built an artificial log from the manual root cause analysis performed after series of production incidents and we then applied process mining techniques to recover an accurate view of the process model.**

Keywords–Manufacturing system; incident detection; process model; event logs; process mining.

I. INTRODUCTION

Modern manufacturing environments have become more digital, which increases, for example, the accuracy of the tracking of the production batches. This digitization results in a large amount of data being generated by manufacturing plants. While some of these data are monitored and widely used, some other, especially data generated directly by the automatons of the production lines are very little exploited, or even deleted a short time after being generated. Event logs are one particular type of data in which series of actions performed at specific timestamps are described. Event logs generated by manufacturing facilities are usually checked by stakeholders only in case of specific needs, for instance in order to investigate the cause of a production incident. A production process describes the series of steps to perform to transform an initial product or set of products into a final state. Various standard semantics exist to model a process, such as the Unified Modelling Language (UML) [1] or Business Process Model and Notation (BPMN)[2]. The functional documentation supplied by the providers of the industrial automatons is intended to provide the users a mean to visualize the process. However, it is common that in practice the process model is not known accurately. For example, at a macroscopic level, the application of many consecutive customizations can lead to a lack of complete and updated documentation and thus to a loss of the global view of the resulting process. Besides, customizations can lead to unexpected consequences and deviations of the process. As detailed below, at the automatons' level, logs can contain very detailed pieces of information that are not described in the general process model but that are relevant to carry out a root cause analysis of incidents. Building a model out of this content can therefore be helpful for stakeholders to diagnose causes of incidents. Production shut-downs are a major issue for manufacturing plants as they generate losses of productivity, extra costs and often require on-call provided by on-site teams. Reducing the occurrences of incidents is thus a challenging topic that we intend to tackle by using data generated by production lines. A global view of the possible deviations of the process is useful in order to carry out a root cause analysis of production incidents. Investigations are generally carried out "manually" by the production-related teams. Stakeholders collect event logs from the automatons and inspect them directly. This method lacks both efficiency and effectiveness due to the size and dimensionality of the logs. The diagnosis is based on the content the functional documentation and eventually multiples exchanges with the provider of the industrial automatons.

In this short paper, we present a method to build a global view of the process, based on the conclusions of the manual analysis of the event logs. Although our study is tested on data provided by a pharmaceutical manufacturing plant, the method is general and can be applied to other industrial contexts.
In Section II, we present related work of analysing production data from monitoring and control systems, in Section III we introduce the problem considered, in Section IV we introduce process mining and explain how our approach uses this technique. We finally conclude in Section V and give an overview of our roadmap to continue this work.

II. RELATED WORK ON SCADA LOG MINING

The manufacturing plant that provides us with their data is equipped with a Supervisory Control And Data Acquisition (SCADA) system to control and monitor the production processes. A general SCADA system is organized in layers and consists of components, such as Programmable Logic Controllers (PLCs), Human Machine Interfaces (HMI), Manufacturing Execution Systems (MES), networks, data bases. The first layer of the SCADA system consists in PLCs that are localized at the production level. The data they produce are sent and processed in the second layer, that contains in particular the MESs where we collect the event logs we are interested in. Finally, the third layer consists in work stations, where authorized users can access the information. If accurately tuned, the SCADA logs contain valuable pieces of information about the process to carry out a root cause analysis of unusual events, such as incidents or threats. Since manufacturing plants are becoming more connected, detecting intrusions has become a very relevant subject. Recent research work were published about SCADA log mining in the context

 ISBN: 978-1-61208-518-0

of process-related threats [3], [4], [5], [6], [7]. Data mining approaches, based on frequent pattern mining methods [4], [8], [9] or on semantic support [10], were successfully tried to detect intrusions but extensions strongly depend on the manufacturing system. In our use case, root cause analysis of incidents is currently made manually, not only because of the difficulty to put in practice an intelligent reading of the logs, but also because other complementary sources of data provided by the MES need to be taken into account to have a good understanding of the process deviations that result in a incident.

III. Problem statement: System Analysis by Manual Treatment

Let us first describe how we analyse the system by manual inspection. Then, we introduce process mining and present the issues one has to deal with when mining industrial logs. For the purpose of building and testing the method, we focus on a small piece of the full production process.

Two types of users interact with the SCADA system: operators and engineers. Operators monitor the production line and take actions in case of alarm to make sure the process works properly. In case of incident, engineers make a diagnosis and perform additional steps to restart the production. They are also in charge of: carrying out a root cause analysis, documenting the faults and eventually communicating with the provider of the automatons to take action and prevent the same incidents from occurring again. Production incidents have been carefully documented over more than one year. Three types of incidents were clearly identified and the percentage of each type over one year was calculated.

Progressive modifications and customizations of the production line implies that the functional documentation that was originally provided with the automatons might no longer be accurate. Updates should be documented but (i) there might be delays before all concerned teams receive the last version of the documentation (ii) even if specific modifications are documented global views of the processes are not always drawn. Besides, even if the process model is already accurate, automatons' logs can contain information at a finer level of granularity. Consider for example a process in which a rising edge, triggered during step_i, induces step_{i+1} to be performed. Suppose the process allows the rising edge to be repeated n times in case it is not acknowledged directly. If it is acknowledged before the $(n+1)^{\text{th}}$ try, at the lowest level, the process went fine. However, the detail inside the log will indicate that it was not acknowledged directly. Even though this does not lead to an incident, it may be interesting to be informed about recurrence of such events in the context of root cause analysis (for instance to study eventual slowness of the network). The manual investigation of the process is based on a careful analysis of the events logs after production incidents combined with the knowledge of the process model, as described by the provider of the automaton. The SCADA system, together with additional devices, records events at different places in the network. Accessing these logs allows the on-site teams to diagnose the incidents and decide what actions to take to restart the production. This approach has many disadvantages:

1) The large size of the files.
2) Some data are missing because files are deleted due to their large sizes and to the lack of use.
3) The search for patterns is biased. Manual inspection is based on searching for particular strings, which can be improved with small scripts. However, unusual patterns one is not explicitly looking for will not be found. The choice of patterns to look for is guided by some idea of what should be a possible cause of incident. If this pre-diagnosis is not correct, the real cause of incident might not be found.

All these motivate our research on intelligent methods to obtain a global view of the process. Process mining is a very promising option that we introduce in the next subsection.

IV. Solution Approach via Process Mining

In this work, our objective is to restore a model of the process based on the manual system analysis, in order to have an accurate theoretical description to compare with. In order to obtain this global view, we are using a process mining approach. Process mining [11], [12] is an emerging field at the interface between data mining and model-driven approaches. It allows to get a global, complete and objective view of the processes that were actually performed by the system. It aims to discover automatically processes from the events logs, check their conformance by monitoring the deviations between the model and the event logs, analyse the performance (e.g., find bottlenecks), make predictions and eventually improve the processes by making recommendations. In this work, we apply a process mining approach to the logs provided by the full SCADA system in order to diagnose faster and with more accuracy the incidents. We expect to discover new types of incidents that were not correctly diagnosed due to the difficulty for a human to read the logs. To the best of our knowledge, very few research work are dedicated to the application of process mining techniques to fault detection in industrial context [13], [14], [15]. In order to be treated by a process mining tool, a log needs to contain specific categories of information: a case ID (e.g., a batch number), timestamps and descriptions of the activities performed within one case. Extra information, such as the resources that realize each activity and their role also allows to extract valuable information (graphs of interactions, etc.).

SCADA log mining is however a long term task for various reasons:

1) The format of the SCADA logs is not suited for process mining. Major data cleansing is necessary [16]. Even after that step, the eventual lack of case ID (i.e., batch number not printed) makes it impossible to use the content of the log with a process mining tool. A possible cure to that issue consists in making recommendations to the stakeholders. After showing the added value of process mining to support root cause analysis, one can recommend modifications of the content of the logs.
2) One needs to mine more than one type of log. Events collected from various sources then need to be properly connected which is a complex task. A possible solution we are investigating is applying frequent patterns mining techniques, as performed in [4].

 ISBN: 978-1-61208-518-0

3) In relation with the previous point, the network introduces delays from one log to another, even within the same activity.
4) Data is missing. Examples of causes are: data are deleted after a short time or data are erased after a reboot. Stakeholders could take action on the first point by providing suitable servers.

While applying process mining techniques directly from the SCADA logs is still under investigation, we built an artificial log to replay the process, such as understood from the manual inspection. Applying process mining software to this log allows us to mimic process incidents.

1) Building a process model from the manual analysis: In order to build an artificial log, iterations together with the engineers who know about the process and investigate the incidents are required. Preparation of the model is done after manual investigation of the process. We proceeded the following way:

1) Identify what are the steps of the regular process. This matches the functional documentation and can be checked in logs outside unusual events.
2) For each incident, get a picture of the different activities that were performed. This task is crucial but very cumbersome due to the fact that one need to read more than one source of data, those sources are large files and they were not designed to be user-friendly.
3) Classify incidents and try to see if categories can be found. In our case, three types of incidents were identified.
4) Document over a long time (in our case over the full year 2015) and keep record of incidents to get a picture of the percentage of occurrences.

From this manual study, we built an artificial log that describes reality. For this, we wrote a programming code able to generate series of cases that include incidents with the same proportion as the one observed in reality. A case is defined by a series of activities. For example in our case, the regular case (i.e., no incident) contains six activities all of them are performed within seconds. Each type of incidents involves either new activities or unusual repetitions of a given activity. Moreover, bottlenecks appear, they are due to complex necessary intervention, such as rebooting the system in order to restart the production line. While the percentage of incidents cases is known, we want them to take place at random time. For the moment an uniform random sorting is performed to spread these unusual case within the log. In case the manual study finds out that some correlations seem to exist, this sorting could easily be improved to mimic reality better.

After an artificial log containing n events is generated, it can be used directly in a process mining tool. We used an evaluation licence of the Disco process discovery software [17], developed by Fluxicon. Fig. 1 shows results highlighting frequency (top chart) and performance (bottom chart) performed with a fuzzy miner algorithm [18]. In the top chart (frequency), thick paths correspond to the most frequent paths, i.e., the regular process. We clearly see the six activities (dark blue boxes) and three extra activities (clear blue boxes) that take place only during incidents (acknowledgement failure that leads to either SCADA reboot or manual acknowledgement).

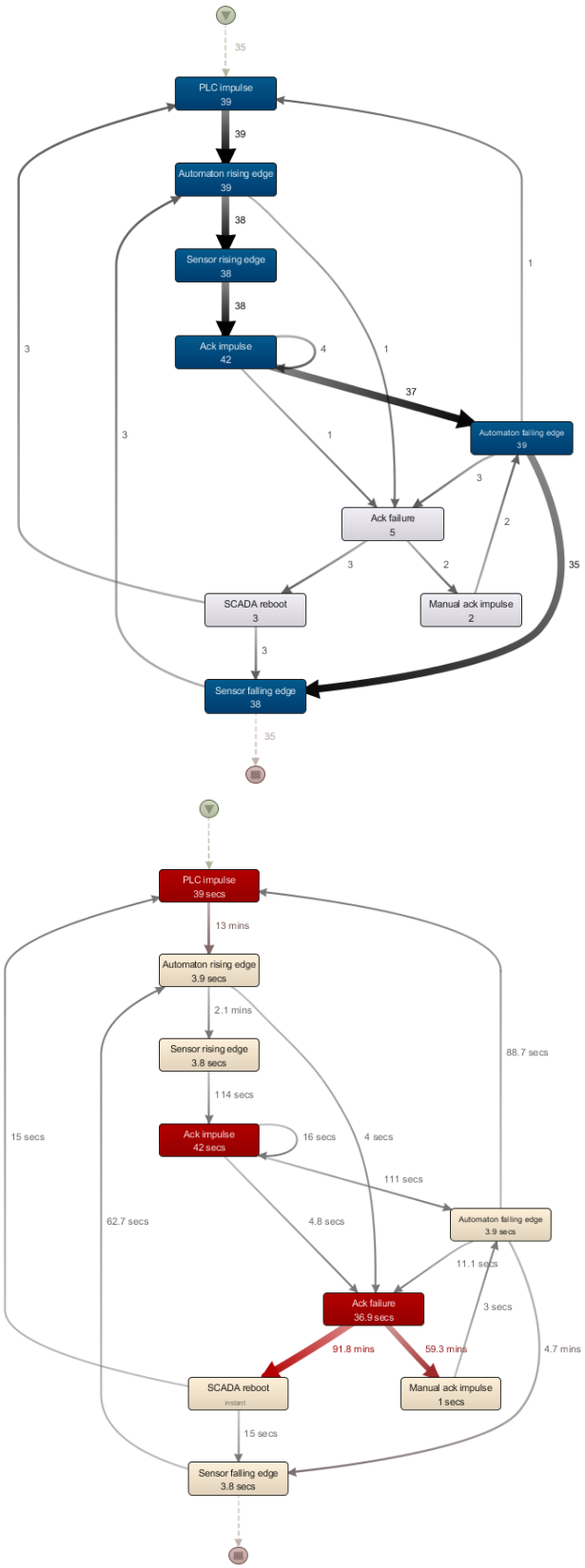

Figure 1. Model of the process and its deviations obtained by the fuzzy miner algorithm from the Disco software [17].

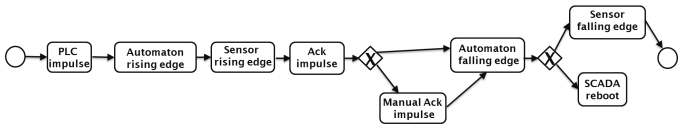

Figure 2. BPMN model corresponding to the process from Fig. 1 extracted automatically with ProM [19] from the artificial log.

The bottom chart (performance) shows the durations of the transitions between activities. The thick red lines correspond to the manual operations, i.e., when an engineer needs to act on the production line. These are the most time-consuming tasks, i.e., the bottlenecks of the process.
Note that, due to confidentiality issues, the names of the activities were anonymized. Of course, no unknown process deviation is to be found since all the deviations were implemented from the understanding we got from the manual approach. This view of the process however offers a better visualization of the outcome of the manual study for stakeholders who need a global picture of the various types of productions incidents. We then used another process mining tool, ProM [19], to extract a process diagram. Fig. 2 shows the BPMN representation [2].

V. Conclusion and future research

We are working on an automatic approach to support root cause analysis for production shut-downs. The data used are production event logs. Process mining seems to be a very promising approach to tackle these issues. We have identified barriers that prevent us from using process mining straight forward with our data:

- Major and complex data preprocessing: cleansing, formatting, merging various logs to get the complete required information.
- Missing data: all the process information is not recorded in the logs, logs are erased after a very short time.
- Lack of case ID number and network delays that make it difficult to connect properly the activities within one case.

Solving these issues is still on-going work and requires iterations with the stakeholders.

Before being able to drawn a process diagram directly from the logs, we have presented in this paper a method to build a process model from the manual approach that is performed to support root cause analysis of the incidents. Our method allowed us to provide a global view of the identified process deviations. If the manual analysis is correct, the artificial log that we have built to mimic reality should match quite accurately the real logs with a similar proportion of incidents. In this case, a conformance analysis will thus evaluate the accuracy of the manual analysis.

Acknowledgment

The authors would like to thank the on-site teams for providing data and for fruitful discussions about the system analysis.
This work is supported by Altran Est.

References

[1] Unified Modeling Language, Object Management Group Std. [Online]. Available: http://www.uml.org

[2] Business Process Model and Notation, Object Management Group Std. [Online]. Available: http://www.bpmn.org/

[3] D. Hadziosmanovic, "The process matters: cyber security in industrial control systems," Ph.D. dissertation, Enschede, the Netherlands, January 2014, iPA Dissertation Series No. 2014-02. [Online]. Available: http://doc.utwente.nl/88730/

[4] D. Hadziosmanovic, D. Bolzoni, and P. H. Hartel, "A log mining approach for process monitoring in scada," International Journal of Information Security, vol. 11, no. 4, 2012, pp. 231–251. [Online]. Available: http://dx.doi.org/10.1007/s10207-012-0163-8

[5] R. E. Kondo, E. de F. R. Loures, and E. A. P. Santos, "Process mining for alarm rationalization and fault patterns identification," in Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies Factory Automation (ETFA 2012), Sept 2012, pp. 1–4.

[6] M. Ficco, A. Daidone, L. Coppolino, L. Romano, and A. Bondavalli, "An event correlation approach for fault diagnosis in scada infrastructures," in Proceedings of the 13th European Workshop on Dependable Computing, ser. EWDC '11, 2011, pp. 15–20. [Online]. Available: http://dl.acm.org/citation.cfm?doid=1978582.1978586

[7] B. Lamas, A. Soury, B. Saadallah, A. Lahmadi, and O. Festor, "An experimental testbed and methodology for security analysis of scada systems," INRIA, Technical Report RT-0443, Dec. 2013. [Online]. Available: https://hal.inria.fr/hal-00920828

[8] D. Hadiosmanovic, D. Bolzoni, P. Hartel, and S. Etalle, "Melissa: Towards automated detection of undesirable user actions in critical infrastructures," in Computer Network Defense (EC2ND), 2011 Seventh European Conference on, Sept 2011, pp. 41–48.

[9] G. Grahne and J. Zhu, "Fast algorithms for frequent itemset mining using fp-trees," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 10, Oct 2005, pp. 1347–1362.

[10] A. Venticinque, N. Mazzocca, S. Venticinque, and M. Ficco, "Semantic support for log analysis of safety-critical embedded systems," in Tenth European Dependable Computing Conference - EDCC 2014, 2014. [Online]. Available: http://arxiv.org/abs/1405.2986

[11] W. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, S.-V. Berlin and Heidelberg, Eds., 2011.

[12] W. van der Aalst et al., "Process mining manifesto," in Business Process Management Workshops (BPM 2011), Lecture Notes in Business Information Processing, Springer-Verlag, Ed., vol. 99, 2011, pp. 169–194.

[13] N. Khajehzadeh, "Data and process mining applications on a multi-cell factory automation testbed," Master's thesis, Tampere University of Technology, 2012.

[14] S. Dasani, "Developing industrial workflows from process data," Master's thesis, University of Alberta, 2013.

[15] S. Dasani, S. L. Shah, T. Chen, and J. F. R. W. Pollard, "Monitoring safety of process operations using industrial workflows," in Preprints of the 9th International Symposium on Advanced Control of Chemical Processes, 2015, pp. 451–456.

[16] L. T. Ly, C. Indiono, J. Mangler, and S. Rinderle-Ma, "Data transformation and semantic log purging for process mining," in Proceedings of the 24th international conference on Advanced Information Systems Engineering (CAISE 12), S.-V. Berlin, Ed., 2012, p. 238.

[17] Disco, Fluxicon. [Online]. Available: https://fluxicon.com/disco/

[18] C. W. Günther and W. M. P. Van Der Aalst, "Fuzzy mining: Adaptive process simplification based on multi-perspective metrics," in Proceedings of the 5th International Conference on Business Process Management, ser. BPM'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 328–343. [Online]. Available: http://dl.acm.org/citation.cfm?id=1793114.1793145

[19] Prom, process mining workbench, Process Mining Group, Eindhoven Technical University. [Online]. Available: http://www.promtools.org/doku.php